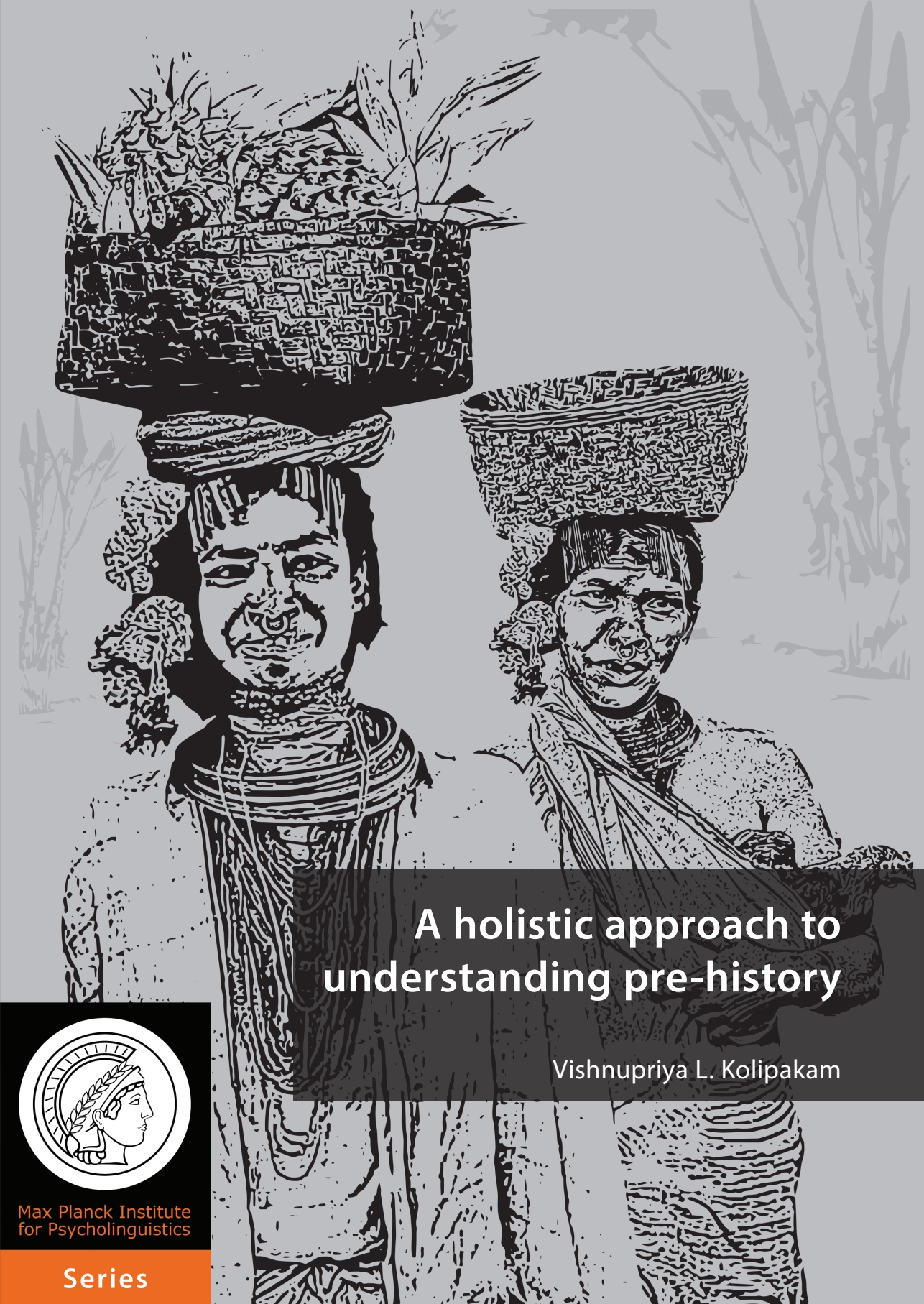


A holistic approach to understanding pre-history

Vishnupriya L. Kolipakam

Max Planck Institute for Psycholinguistics

Series

A holistic approach to understanding pre-history

The cover depicts two women of the Khond tribe of central India, who speak the Dravidian language Kuwi. Women of this tribe have a distinctive hair style and love dressing up with flowers and ornaments made of beads. They are heavily dependent on forest and forest produce for sustenance. Their survival is now threatened due to heavy mining and development projects.

ISBN: 978-90-76203-58-4
Cover design: © Paritosh Bharti
Printed and bound by Ipskamp Printing, Enschede, The Netherlands

A holistic approach to understanding pre-history

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen
in het openbaar te verdedigen op maandag 7 mei 2018
om 14.30 uur precies

door

Vishnupriya Lakshman Kolipakam
geboren op 4 september 1985
te Chennai, India

*Promotor*en:
Prof. dr. Stephen C. Levinson
Prof. dr. Fiona M. Jordan (University of Bristol, Verenigd Koninkrijk)
Prof. dr. Michael J. Dunn (Uppsala Universitet, Zweden)

Manuscriptcommissie:
Prof. dr. Asifa Majid
Prof. dr. Pieter C. Muysken
Prof. dr. Jamie Tehrani (Durham University, Verenigd Koninkrijk)
Dr. Dan Dediu (Collegium de Lyon, Frankrijk)
Dr. Adam Powell (Max-Planck-Institut für Menschheitsgeschichte, Jena, Duitsland)

The research reported in this thesis was supported by a grant from the Max-Planck-Gesellschaft zur Förderung der Wissenschaften, München, Germany, as part of the Max Planck Research Group "Evolutionary Processes in Language and Culture".

कुछ तो जोड़ता है रवानी-ए-खून को रूह से

ज़बानो ने तो बनाली सरहदे अपनी

- अनम करा

There is something in blood and spirit that bridges all our souls

while languages have created frontiers of their own

- Anam Cara

TABLE OF CONTENTS

Acknowledgement

This thesis could not have been completed without the support, encouragement, kindness and patience of many people in my life. I would like to thank everybody from the Max Planck Institute, Radboud University and Wildlife Institute of India, who were with me during this sometimes very stressful and frustrating, but mostly happy times.

First and foremost, I would like to thank my supervisors, Michael Dunn and Fiona Jordan. Thank you both for giving me the opportunity to start this journey. You both nurtured a wonderfully stimulating work environment. You taught me to explore the unknown without fear, and for that I cannot thank you enough. Thank you Michael, firstly for believing in me, for teaching me that different perspectives enrich the way we look at science. The joy and enthusiasm you have for research was contagious, and motivational for me, even during tough times in the Ph.D. pursuit. Thank you for your patience and for not losing faith in me and encouraging me every step of the way, even when I thought of giving up. I owe a lot to you. I am forever grateful for the emotional support you extended during my stay at Nijmegen and beyond. Fiona, thank you for your guidance, encouragement and constructive criticism. Thank you for the hard questions which encouraged me to work harder and widen my research from various perspectives. You moulded me into a better researcher than I ever thought was possible, and made me push boundaries to realise my potential. Along the way, you made sure that I not only took science seriously, but also learnt other skills that would enable me to function as a good scientist. Thank you so very much. I am also extremely grateful for the moral support through my stay at Nijmegen.

Steve, thank you for making me see this through. I am grateful that you agreed to be my promotor. Your constructive feedback has improved the manuscript tremendously. I wish to thank members of my reading committee, Asifa Majid, Peiter Muysken, Jamie Tehrani, Dan Dediu and Adam Powell for their time and constructive feedback. Thank you Saskia and Ashwini for being my paranymphs and standing by my side on this very important day.

This thesis would definitely have not been completed without the constant support, encouragement and facilitation by my current project investigators: Dr YV Jhala and Prof Qamar Qureshi. Dr Jhala, thank you sir for being a constant source of inspiration and for teaching me to ask the right questions. Thank you for pushing and facilitating me to finish this journey I started. I could not have asked for a better mentor to work with. Your insights into science leave me amazed, and make me strive harder in the hope of achieving that level of scientific acumen, one day. Thank you for leading by example on being an excellent scientist. Qamar sir,

thank you for all the wonderful discussions on scientific and non-scientific matters. You have helped me push boundaries and made me think out of the box. After every single conversation with you, I've always walked away feeling excited about science and hoping (wishfully thinking rather) that I would someday gain the perspective you have. Thank you for all your support and encouragement through out. Thank you for moulding me into a better person.

Annemarie, I would like to specially thank you for being with me every step of the way. I could not have asked for a better office mate. From "*Ik wil een kopje thee*", introducing me to Dutch traditions, to braving on to try out new Indian traditions. You were my family every step of the way. Even after leaving, your love and affection for my well-being, and your constant encouragement has made this journey not only very tolerable, but also possible. I will never forget 715 and our afternoon tea tradition. My life is better for having known you. Thank you for being there. Thank you also for translation of the Dutch summary, I appreciate you doing this over such a short time.

Els, you played the most important role in getting me here. Thank you for your honest, impartial and unconditional support at the IMPRS. Your door was always open for me to come talk to you, and most importantly, thank you for facilitating to bring this journey to completion. I owe this to you. Rachel, Dirkje and Kevin, thank you for your constant support. Thanks to Edith Sjoerdsma for taking care of all administrative issues (especially complicated by my being not there). Thanks Edith for your patience. Thanks to Angela heuts, for helping me settle in and making sure I was okay though out my stay, the MPI librarians (especially Karin Kastens and Meggie Uijen), Tobias van Valkenhoef and Paul Trilsbeek for computer and server support without which many of the simulation runs or phylogenetic work would not have been possible.

Thanks to Dan Dediu, Nicolas Brucato, and Oscar Lao for your inputs with genetic data analysis. Dan & Nicolas, thank you for some very stimulating conversations, and for guiding me in the right direction. I would like to thank Gail Coehlo for unhesitatingly giving me data on the Kurumba tribe. Thanks to many people who facilitated my data collection efforts on Dravidian languages, especially Abdul Raziq from Balochistan, Mohan Raj from WWF-India, Jagdeesh, Foundation for Ecological security Team at Mandla, Dr Mishra, Mr Baxla, Kishore, Chubakki, and Pawan.

My time at the MPI has been wonderful, and I have learned much from the academic and non-academic discussions. Thank you Melissa and, Asifa for showing me the brighter side of every situation. For those stimulating work discussions, wonderful lunch time discussions, fun borrels and parties, that made work just that bit more fun, thank you Melissa, Michael, Anne, Dan, Rachel, Saskia, Jeremy, Rebecca, Gunter, Fiona, and Aarthy. Thank you to my fellow IMPRS and midi-plankers for being such good company, Ruti, Sarah, Edwin, Jeremy, Dejan,

Rebecca, Sylvia, Tyko (Sir Humphrey), Helen, Inge, Gertie, Sarah, Marijt, Christina and Joost.

This journey of mine has been marked with screams, manic laughter, sombre tragedy, the sense of awe, agony and unbridled joy that underpins any doctoral life span. If not for the company of many loving souls, this journey would not be as special as it feels now. Spending long periods away from home, I formed many friendships that made living in an alien environment not only bearable, but immensely enjoyable. Anne, Eric, Ashwini, Amat, Leike, David, Julian, Aravind, Subhra, Pragnya, Aarthy, Venkat, Ashim, Anil, Asha and Chandan, thank you for being my surrogate family and support system. For all the exciting, fun evenings and weekends, the delicious food we all made together, thank you being part of my life. Rachel Sheer and Angela Terril, thank you so much for your moral and emotional support during my time at Nijmegen. For my friends here in India, just have to say, thank you for putting up with me! Shikha and Stotra, thank you for wordlessly doing the painful job of proof-reading my thesis. Shikha, thank you for being there for me, your support and belief and constant positive perspective to things has made this journey more liveable (ironically because of all the zombies we created!). Swati and Madhura, our long drives, girls' night outs and impromptu cooking sessions have kept me sane, or atleast I can thank you for trying! Manjari, Sutirtha, evil child, Sudip, Bipin and Stotra thank you for being there for me, always supporting me and giving me hope, listening to my midnight rambilngs, and panic attacks and for just being you guys. Stephen and Rupinder, thank you both for the wonderful weekend trips and marathon weekend parties, you provided a great respite from the arduous task of getting this to completion. Samrat and Pamela, thank you for encouraging me and always pushing me towards better things in life. Thanks to Shweta, Bhawana, Farha and other labmates at the Conservation genetics lab, for a fun working atmosphere and your unconditional support. Thank you Teju, Jinta, Lachu pisachi, Namitha, Sneha, Sandhya, Hari, Vijay, Jagdeesh, Ayan, Shweta, Ujjwal, Neha, Ninad, Sabuj, Shazia, Kausik da, our friendship has helped me pull through many tough situations.

On a more personal note, I have to thank my family, without whom none of this would have been possible. For their belief in me, encouraging me to take up a career in science, standing by me with unwavering support and trusting my decisions all through, thank you mom (Lavanya) and dad (Lakshman). You both are my pillars of strength and confidence. I am lucky to have you both, any kid would be. I am forever grateful for your sacrifices to make sure I am happy. I love you both. To my sister Vasudha, thank you is a word that is not enough. Vasudha akka, this PhD journey was as much yours as it was mine. You were there with me every step of the way, constantly guiding me, giving me strength and support, pulling me up from the lows and making sure I stay alive! The one sane voice in this insane world I could depend on, even if it was in the middle of the night Thank you for consoling me, cajoling me and kicking me, when I needed it. The round

pokey man definitely did his job well! You deserve not only the “Ph” but the “D” as well. I would also like to thank Arvind, my brother-in-law for his love, support and encouragement. And to my lovely niece Adya, thank you for your unconditional love, and for being the most loving, innocent little thing you are. You are a constant source of joy in my life. To my in-laws Sunanda and Ponnappa, thank you for believing in me, and for your love, encouragement and support. To my extended family, Somu thank you for showing me the lighter side of things, and bringing a smile on my face, and Pallavi thank you for the support and understanding. To my husband, Bopanna, a very special thank you (for lack of a better word), for always being there when I needed you and for keeping me grounded and calm. Without the support, love, encouragement and most importantly laughter you brought into my life, this thesis would have been an insurmountable task.

In this splendid journey of mine, the list of people who offered their help, support and friendship are numerous. I thank everyone who has intentionally and unintentionally contributed towards completion of this work. And finally, for giving me a glimpse of their world, and for letting me into their lives, even if only for a moment, I must thank the people of Polynesia and Dravidian societies.

1 Introduction

"Each with its own beauty, and each with a story to tell" - Stephen Jay Gould

1.1 A holistic approach to understanding pre-history

The very essence of science lies in the quest to understand the world we live in and the diversity it encompasses, through the eternal process of problem solving. Much of this depends and builds upon the knowledge of the past. While written records and archaeological evidence are some of the tools that aid in our understanding of history, understanding pre-history is a much more challenging endeavour where we have to depend on inferential rather than direct evidence. Understanding human history has heavily relied on evidence from archaeological data as well as linguistic and ethnographic records to build a timeline of human evolution. Unfortunately, however, little historical material can be gleaned from non-literate cultures and anthropologists must either proceed ahistorically or construct history by inference. Innovations and development in molecular techniques have facilitated making these inferences regarding pre-history with relative ease. However, molecular data alone cannot contribute to a holistic understanding of pre-history. For example, the origins and spread of Etruscans, a pre-classical society in Europe, has been a subject of debate amongst historians and archaeologists. While ancient historians posit the Etruscans origin to lie in Asia Minor, modern archaeologists believe that the Etruscan civilization emerged from local Villanovan culture. Genetic data modelling (Belle *et al*, 2006) suggested that Etruscans showed a strong genetic affinity to Anatolian populations, with no relationship to the modern day Tuscans. However, when other likely models of population history were tested, after assimilating information from cultural and archaeological data, another plausible model of history emerged. Models revealed that Etruscans were most likely a socially elite stratum, only representing a small portion of the society which did not contribute to the genes of the modern day Tuscans. This theory was supported by archaeological evidence, where the burial sites of the Etruscans, from which the samples were collected, were characterized by extravagant tombs, with broad collection of artefacts and this could be attributed to the burials of the rich. It was then inferred that social stratification and lack of representation of the entire Etruscan society in the genetic material could also have given rise to the genetic pattern observed in present day modern Tuscans. If conclusions were drawn on molecular evidence or archaeological evidence alone, this dimension of investigation would have been lost and a less

likely population history scenario would have been inferred. This study demonstrates the uncertainty in deciphering pre-history by using molecular data alone, and how by using it in conjunction with information from other sources, we could increase the confidence in our interpretation of the past. This aspect forms one of the main driving ideas behind my thesis and the several hypotheses I proceed to test.

Drawing inferences from different fields together is not as straightforward as it seems. For example, sex-linked DNA (Y chromosomes and mitochondrial DNA), sometimes present different and contrasting patterns (Wilkins and Marlowe, 2006), and this is linked to social constructs like marriage and residence rules (Kobben *et al*, 1967; Lansing *et al*, 2008; Watkins, 2004). Since marriage and residence determines the movement of sexes in society, it should also determine the patterns of sex-linked DNA. Heyer *et al* (2011), found that the pattern expected (post-marital residence vs sex-linked DNA) did not hold good, and put forth a theory stating, apart from residence norms, effective population sizes, variance in reproductive success and social factors like descent rules must also play an important role in shaping the diversity of SSM markers. However, there has been little work to tease apart the effect of these social rules on genes, and to understand their relationship with each other (Ségurel *et al*, 2008; Verdu *et al*, 2013). The reason for this could be that population genetics has to reach beyond simple analytical models, and look towards a framework incorporating information from social processes and genes, to understand their interaction with each other, and thereby incorporating them in testable population models. In my thesis, I also focus on addressing this gap through appropriate methods discussed in the following sections.

Several pioneering American anthropologists like Boas and Sapir, as early as 1916 recognized the advantages of a consilient approach to understand human history and put forward the "four fields of anthropology" approach and a more rigorous form of "historical particularism" (Kirch, 2010; Sapir, 1916). The lead up to this thinking was the understanding of how cultural and biological traits were intricately linked in the evolution of human society. The idea of understanding the diachronic perspective of cultural and biological traits for a more holistic approach to deciphering pre-history was in vogue before Boas' work. Augustus Pitt-Rivers set into motion this need for a holistic approach to interpretations of pre-history, and established the Pitt-Rivers museum in Oxford with an innovative manner of displaying archaeological and ethnographic items based on their evolutionary relationships, thereby drawing attention to the co-evolutionary nature of different aspects of human history.

1.2 Gene-culture coevolution

Hypotheses of gene-culture co-evolution postulate that cultural traits play a significant role in shaping genetic variation. Socially shared information and norms are an integral part of human societies, and it has recently been shown that these cultural norms can positively influence selection on the human genome (Durham, 1991; Laland *et al*, 1995). This is a fundamental example of the niche construction theory posited by Odling-Smee *et al* (2003). The niche construction theory states that our behaviour, for example through cultural traits, is far more influential in shaping our external environment than previously thought. Consequently, it influences the positive/negative selection that acts upon us and this effect persists and percolates down to future generations (Kendal *et al*, 2011; Laland *et al*, 2000; Laland *et al*, 2010; Laland *et al*, 2011). Cultural traits such as social norms and linguistic traits influence the selection of human alleles and therefore influence human evolution. For example, the trait of lactase persistence in a population is related to the cultural practice of dairying (Mace *et al,* 2003). Variation in the genes that code for lactase persistence arose in the Middle East and African populations around 7,500 – 9,000 years ago and in Europe within the past 2000 years, in response to a strong selective advantage for the ability to digest milk as adults (Mace *et al*, 2003). This shows that cultural traits have induced a change in the molecular make up of humans. In essence, every aspect, whether cultural or biological, associated with human societies is inextricably intertwined.

The process of language inheritance, and the process of cultural transmission through rituals and practices, follow processes akin to transfer of biological information encoded in genes. While the idea of cultural transmission following a biological trait like inheritance and evolution was hypothesised as early as when the theory of evolution was proposed by Darwin (1859), we lacked the methodological advances to understand how cultural traits change over time, and how to incorporate this understanding in our interpretation of human history. While cross-cultural studies have sought comparisons between cultures to tease apart the dynamics of evolution of a cultural trait, it is wrought with problems of confounding effects of ancestry. Adapting methods from population genetics, early pioneers like Cavalli-Sforza and Feldman (1981), Boyd and Richerson (1985) and Durham (1991) made substantial efforts in developing methods to track cultural trait evolution, by drawing parallels between the evolution of biological traits and cultural traits. While they made great headway in moving closer to a holistic approach to pre-history, the progress of incorporating this evolutionary thinking, and adapting evolutionary methods to integrate and test findings regarding biological trait evolution with those of cultural traits is still in its nascent stage. The potential for integrating these fields is tremendous, especially with the development of a suite of evolutionary methods in the field of genetics.

The idea of cultural traits "evolving" started long before in linguistics than in biology. Schleiger, Hensleigh, Wedgwood and other linguists were already constructing relationships between languages based on the principle of sharing a common ancestry (Van Wyhe, 2005), based on the idea of descent with modification, which would give rise to variation in languages. This was later applied to the evolution of species by Darwin. Darwin, and subsequently many other scientists like Cavalli-Sforza (1975), (Boyd *et al*, 1997), Jones (2003), Gray and Jordan (2000), Mace and Pagel (1994); Pagel and Harvey (1988), and Atkinson and Gray (2005), have drawn our attention to the similarities between the characteristics of biological evolution and that of cultural or linguistic evolution. Dual-inheritance theory incorporates the knowledge that cultural inheritance acts akin to biological inheritance and aims at providing a framework under the umbrella of Darwinian evolution to understand the evolution of both traits and integrate that knowledge to comprehend human history (Boyd *et al*, 1997; Cavalli-Sforza and Feldman, 1981; Durham, 1991; Laland *et al*, 1995; Laland *et al*, 2010).

1.3 Methodological advancements to aid holistic approach

1.3.1 Cultural and biological trait evolution

The basic traits defined by Darwinian evolution – variation, selection and inheritance also can be applied to linguistic and cultural evolution. For the purpose of this chapter, both cultural and linguistic evolution are conceived as cultural evolution.

Variation in biological evolution is brought about by mutations and descent with modification. It is obvious that language and culture of human societies are inherently diverse, with over 7,000 different languages spoken by approximately 10,000 different ethnic groups in the world (Jordan, 2007). The human genome accumulates mutation at varying rates bringing about a change in the organism. For example, parts of the mitochondrial DNA, maternally inherited, accumulate mutations at the rate of 1 mutation every 5000 years (Soares *et al*, 2009). Some other parts of the genome are more susceptible to mutation, like the Y chromosome, where the mutation rate is approximately ten fold more when compared to the mitochondrial DNA mutation rate (Zhivotovsky *et al*, 2004). Mutations are analogous to innovations in cultural evolution. Different language families are evidence in themselves on how innovations can lead to new languages (Friedlaender *et al*, 2009; Hunley *et al*, 2008). Innovations are the basis of change in both cultural and biological evolution.

Heritability of genetic information from one generation to the next can be compared to the transmission of linguistic syntax or words in each language, with almost perfect conformity, to the next generation. It is observed that as closely

related groups are similar in their genetic makeup, in culture too there are traits, observed through cross-cultural studies, which are heritable and behave in the same way as genes do. Kinship and social organization traits are thought to be heritable at a group level, resulting in closely related groups sharing similar kinship and social organization (Boyd *et al*, 1997; Boyd and Richerson, 1996). Scholars have reiterated how concepts from evolutionary biology have proven useful in explaining many cultural phenomena, such as drift (Neiman, 1995) and selection (Boyd and Richerson, 1992) to name a few amongst many others (Atkinson and Gray, 2005; Boyd *et al*, 1997; Cavalli-Sforza, 1975; Gray and Jordan, 2000; Laland, 1993; Laland *et al*, 1995; Mace and Pagel, 1994; Mesoudi *et al*, 2004; Mesoudi *et al*, 2006). These parallels have clearly established that the fundamentals of biological evolution, inheritance, variation and selection, are also the basic attributes of cultural evolution. This demonstrates the evolutionary nature of cultural evolution and that cultural traits can and should be treated as evolutionary traits. Also, another important facet of this discovery is that if cultural data is to be constructively linked to archaeological or genetic evidence, then it is necessary that an independent investigation of the cultural data be performed.

These parallels between biological and cultural evolution mean that methods from evolutionary biology could be applied to study transmission of cultural traits (Mace and Pagel, 1994; Pagel, 1994; Pagel, 1997) or test hypotheses regarding the history of a population by combining data from biological, linguistic and socio-anthropological data. A growing body of scholarly work has attested to being able to use these approaches developed for evolutionary biology on cultural data to test hypotheses effectively (Fortunato *et al*, 2006; Jordan *et al*, 2009; Mace *et al*, 2003; Tehrani and Collard, 2002). This gives us an opportunity to understand the different lines of evidence in a comparable framework. It is also evident from these studies, that given the complex nature of interactions of genes, culture and language, we require a range of methods from molecular tools to computation modelling and comparative and cross-cultural studies to understand diversity in human societies. Modern evolutionary anthropology takes such a consilient approach by using phylogenetic methods adapted from evolutionary biology to understand the evolutionary processes that have shaped cultural variation (Gray *et al*, 2007; Harvey and Pagel, 1991; Mace and Holden, 2005). With the adaptation of evolutionary methodology, the focus of scholarly work has shifted from investigation of within population mechanisms of cultural macroevolution, to reconstruction of evolutionary relationships between cultural traits, languages, and thereby enabling us to derive a holistic perspective to population history (Gray *et al*, 2007; Lipo, 2006; Pagel, 2009).

1.3.2 Adapting methods from evolutionary biology to understand cultural trait variation

How is evolutionary methodology helping bring together disciplines on a common framework? Whether in biology or culture, understanding variation is the key to understanding evolution. In anthropology, cross-cultural comparison (Brown, 1963; Heath, 1958; Kobben *et al*, 1967; Murdock, 1940; Murdock and White, 1969) has been a key focus of research in deciphering the patterns of cultural variation and the drivers of evolution and diversity across human societies. However, since human societies are inherently related, there exists the problem of historical non-independence for comparative analysis. This problem of non-independence was first expounded by Galton (Tylor, 1889; also refer to Chapter 2). In biology too, comparative studies are undertaken to explain when a character has evolved or what has caused evolution to occur. However, each sample—whether species or society—is not independent. There exists an evolutionary history that connects these samples. Similarity in traits could be due to inheritence from a common-ancestor or due to functional adaptation, (homoplasies and homologies in biological terms). Just as in biology, phylogenetic comparative methods help understand relatedness between societies while controlling for the non-independence of data points. However, it is only in the past two decades (Mace and Holden, 2005; Mace and Pagel, 1994), that phylogenetic comparative analysis has been used in cultural and linguistic studies. Mace and Pagel (1994) have put forth a strong argument on how population history can be modelled using linguistic or genetic data. Since then, population trees have been built not only using allele frequencies (Cavalli-Sforza *et al*, 1993), but also using lexicons from linguistic data in phylogenetic algorithms (Gray and Jordan, 2000; Holden, 2002). By using evolutionary methodology, we are able to tease apart the true evolutionary nature of cultural traits. These methods have progressed from using maximum parsimony to more sophisticated methods like Bayesian Markov Chain Montecarlo (MCMC) (Pagel, 2009). Using this data built from evolutionary models, or by integrating them into other evolutionary models with data on language or genetics, will help bring the different threads of history together on a common framework.

Studies using linguistic data have applied phylogenetic methods to understand population histories of the Austronesians (Gray *et al*, 2009; Gray and Jordan, 2000), Bantu societies (Holden, 2002; Rexova *et al*, 2006) and Indo-Europeans (Gray and Atkinson, 2003; amongst others). Similarly, comparative method and phylogenetics has been used to understand evolution of cultural traits in different societies, for example post-marital residence patterns in Austronesians (Jordan *et al*, 2009) and Indo-Europeans (Fortunato and Jordan, 2010), and the transmission of material culture amongst other traits (Jordan and Shennan, 2009; Tehrani and Collard, 2009). These studies have not only detailed

the status of traits in societies but also depicted how these traits are prone to change and the direction of change of these traits. By understanding these dynamics, one can truly understand and infer the social change that could and probably would have occurred in prehistoric societies, and with confidence. This aspect of cultural change cannot be captured by archaeological or anthropological studies that do not incorporate the evolutionary nature of cultural traits. The details of the comparative methods are discussed at length in subsequent Chapters. Taking advantage of evolutionary methods, hypotheses regarding linguistic evolution have been tested (Dunn *et al*, 2011b) and postulated (Levinson and Gray, 2012).

Methodological advancements in population genetic tools have aided in bringing confidence to our inference of the past. With progress in computational technology and the field of bioinformatics, mathematical modelling adapted to model evolution and trajectory of genes has helped develop population models. Bridging the gap between theory and application, population genetics has emerged as an inter-disciplinary science where hypotheses about population histories from the fields of anthropology and ecology can be examined quantitatively. Keeping pace with the dual-inheritance theory, and with the development of coalescent theory, reconstructions of past human migrations, tracking demographic changes in populations, admixture events could be made with some degree of confidence. More importantly, alternative hypotheses regarding population histories can not only now be tested using coalescent modelling, but these events could be detected, dated and quantified (Li and Durbin, 2011; Rosenberg and Nordborg, 2002; Shapiro *et al*, 2004; Thangaraj *et al*, 2005).

1.3.3 Looking beyond using methods – tying the story together

In the example above regarding the effect of marriage and residence rules on sex-linked markers, it is difficult to tease apart the interaction between cultural traits and genetic variation in order to understand the effect of social rules on genes (either co-evolution or correlation). The reason for this difficulty is that the patterns we observe in real world data cannot be solely accounted for by the interaction between genes and culture, but rather by a combination of interactive effects of several different phenomena (environment, biological, etc.). And if we make the effort to incorporate all these effects, the model may become intractable under realistic genetic scenarios, or we might have to rely on approximations which do not give a clear picture. Additionally, the effect we perceive might not always be true under different conditions, and understanding if and how the interactions vary under different conditions is practically impossible with existing real world data. However, simulation techniques can help avoiding such problems,

and as such have grown to become a popular analytical predictive tool in population genetics (Carvajal-Rodríguez, 2010; Hoban *et al*, 2012; Peng and Kimmel, 2005). With the repeated generation of pseudo-data that incorporates stochasticity and variation inherent in populations, analyses may test and predict the effects of interacting forces and thereby will contribute to inferring historical processes more robustly. Especially in ecology, simulation techniques are used to understand ecological patterns and processes, and in timing important demographic events (Hoban, 2014). These techniques are an excellent methodological advancement that enables us to understand the evolutionary and genetic consequences of complex interactions, like in the case of gene-culture co-evolution. Attempts have been made to develop tools that would address some of these questions (Excoffier and Foll, 2011; Guillot and Cox, 2014), and to understand dynamics of population histories like hunter-gatherer (Chikhi *et al*, 2010; Ray *et al*, 2003). However, their application and use in anthropology is still to reach its potential.

The possibility of using these phylogenetic and population genetic models with cultural traits means that we have a powerful tool at our disposal to investigate population history using data from language and culture. Recent studies have highlighted that constructing language genealogies can help us track cultures in a way that genes cannot (Friedlaender *et al*, 2009; Atkinson, 2010). It has been shown that linguistic features are less likely to diffuse across population boundaries than genes, and when populations speaking different or unrelated languages come into contact, then it is language that is more resistant to change when compared to genes (Friedlaender *et al*, 2009; Hunley *et al*, 2008; Lipson *et al,* 2018; Posth *et al,* 2018). For example, in Melanesia, it was found that where unrelated languages exist in the same landscape (Papuan, Austronesian), genetic boundaries become blurred, but linguistics features are less likely to diffuse across these socio-linguistic boundaries (Hunley *et al*, 2008). And when adequate data is available, analysis of linguistic tools can become useful in the reconstructions of population histories, especially when populations are in contact, admixture is high and genetic history is blurred.

Scholarly work has largely accepted the evolutionary nature of cultural traits and is well progressing in using methods developed in biology to understand cultural trait evolution and in drawing parallels on the gene-culture co-evolution front. The gap in this advancement lies in the lack of a robust model testing framework to quantitatively establish the relationship in some scenarios, and to triangulate information from different information sources.

1.4 Note of caution

Alongside the growing interest in the evolutionary analysis of culture, there has also been strong opposition to the basic premise of cultural evolution resembling

biological evolution (Ingold, 2000; Ingold, 2007) and how using evolutionary methods to analyse cultural data is not an acceptable/suitable approach. More constructively, Sperber and Claidiere (2006) argue that there is a difference in biological and cultural inheritance. They put forth the argument that while it is easier to separate conserved regions of the genome from regions that are adaptable to variation, in cultural evolution the differential success of cultural traits is dynamic and the distinction between targets of preservation and adaptive evolution can be blurred. This difficulty in differentiating could be problematic in determining whether a trait is adaptive or is prone to selection.

Another difference between cultural and biological evolution is the process of transmission of information. While biological transmission is generally vertical and tree-like, cultural traits can be transmitted horizontally and across lineages (Borgerhoff Mulder *et al*, 2006; Nunn *et al*, 2006), and this could distort the inference of phylogenetic analyses of cultural traits, which assumes vertical transmission.

There is also an argument that counters these deviations of cultural evolution from that of biological evolution. Mesoudi (2007) argues that if we considered a biological phenomenon such as genomic imprinting, where the expression of a gene is determined by the source of inheritance (whether it has come from the mother of the father), then we come across a similar blurring process between preservation and selection in cultural traits. Also, in the case of horizontal transmission, bacteria and plants frequently transmit genetic information in this manner and across lineages (Abbott *et al*, 2003; Doolittle, 1999; Rivera and Lake, 2004; Temkin and Eldredge, 2007) and hence, rather than a mammalian-focused evolution, a general biological model of evolution might be an appropriate parallel for cultural evolution. While it is evident that biological and cultural evolution are not identical, it is also safe to assume that they are similar to an extent where the same methodology can be used to make inferences in a variety of circumstances, with some degree of confidence. To understand gene-culture coevolution, models have been developed to incorporate continuous, non-discrete and non-gene like traits as well as with non-Mendelian blending inheritance and Lamarckian-like inheritance of acquired characteristics (Boyd and Richerson, 1985). In phylogenetics too, attempts have been made to test the effect of horizontal transmission in cultural datasets (Collard *et al*, 2006) and to develop methods that incorporate this mode of transmission (Forster and Toth, 2003; Riede, 2008). One conclusion to draw from this is that it is important to understand the differences between cultural and biological evolution and encourage the development of novel mathematical analyses that would help model cultural inheritance (Mesoudi *et al*, 2006). It is also important to realize that the similarities that these two traits (biological and cultural) share in terms of evolution, overshadows the differences between them, and we should make the

best use of the methodological framework available to us, in order to bring to understand the process of their evolution and interaction.

1.5 Focus of thesis

In my thesis, I focus mainly on exploring ways in which a multi-pronged approach to pre-history can a) interrogate some of our assumptions regarding underlying models, and b) reveal evolutionary processes that explain patterns of diversity.

As emphasized earlier, researchers can borrow sophisticated and statistically rigorous techniques developed for evolutionary biology to analyse cultural change. As Mayr (1982) developed a theoretical framework to integrate different disciplines into a coherent research program, evolutionary biology has the practical framework to do the same with cultural data. The triangulation technique pointed out by Kirch and Green (2001) and the use of evolutionary methods to track cultural trait evolution (Mace and Pagel, 1994), in combination would together prove to be the most useful tools in taking a holistic approach to understanding pre-history with a rigorous statistical framework.

I focus on using methods from population genetics, and evolutionary biology to understand and test hypotheses regarding human pre-history by taking a holistic approach to linguistic, cultural and genetic data. By using modern evolutionary methods for cultural and linguistic data, hypotheses about human social behaviour are tested. The co-evolutionary theory of genes and culture and their sphere of influence on each other are also tested, by determining and quantifying the dynamics of this relationship through the population genetic tools of the forward simulation framework. I draw into using population genetics by taking into account information from anthropology and linguistics (triangulation) and test it against existing hypotheses regarding the dispersal history of human societies to validate models on whether cultural practices are the main forces influencing genetic patterns. Finally, I demonstrate the value of linguistic data to decipher pre-history in a complex system where biological and cultural data cannot give a coherent picture, by employing it in a phylogenetic framework.

In the following sections, I introduce the background to the questions I have asked in my thesis, which are inspired by societies that are contrasting in their geography and anthropological context: Polynesian societies in the Pacific and Dravidian societies in India.

1.6 The Austronesian context

The pacific region plays a crucial role in our understanding of human pre-history and migrations, as it was witness to one of the last major human migrations in the world. This colonization event resulted in the peopling of previously uninhabited

islands of Polynesia, in the last 1000 years (Figure 1–1). The term "Polynesia" was initially coined by Dumont d'Urville, a French voyager and Naval commander, in his "Notice sur les îles du Grand Océan et sur l'origine des peuples qui le habitant (1832)"(Dumont, 1832). As part of his voyage on the "Astrolabe", mapping the Pacific, he classified the inhabitants of the Pacific islands into three major groups – a) "Polynesians" (meaning inhabitants of many islands) – living on the eastern Pacific islands, including Hawai'i, Rapa Nui (Easter Island), and New Zealand. Inhabitants of these islands are generally light-skinned and spoke similar languages, b) "Micronesians", for those living on the many little atoll islands in the western Pacific region, north of the equator, and c) People of darker-skin and inhabiting New Guinea, Solomons, Vanuatu, New Caledonia and Fiji were termed "Melanesians" (Melanesia - dark islands). Although based on apparent and superficial understanding of the Pacific Islanders, this tripartite classification is still used widely.

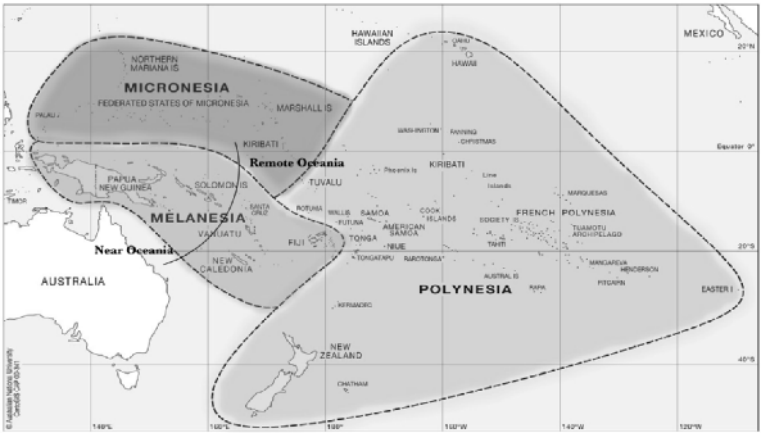


Figure 1–1: Map of Pacific, with areas of interest depicted, and their historical classification into "Micronesia", "Melanesia" and "Polynesia". Also depicted is the curved line demarcating Near Oceania and Remote Oceania. (adapted from : https://www.changemakers.com/geography-pacific-0).

Roger Green (1991b) argued that the terms Near and Remote Oceania provide a more meaningful description of the Pacific than the geographic markers of "Melanesia" and "Micronesia", which do not imply any cultural or historical unity. The only classification of Dumont d'Urville which was a meaningful unit for culture and historical analysis is "Polynesia", as probably this was formed on the basis of linguistic similarities rather than apparent racial affinities. Oceania refers to the region encompassing all the three groups described above, excluding Australia and Island Southeast Asia (ISEA). The classification of Near and Remote Oceania was on the basis of two human migrations, thousands of years apart and carrying people of different cultural units, peopling the region. Oceania was the endpoint of one of the earliest human migrations, the "Out of Africa" migration,

which carried humans to Australia and New Guinea approximately 50,000 years ago. It is believed that around 110,000 years ago, anatomically modern humans dispersed out of Africa through Asia, and colonised the greater pacific region around 55,000 KYA (Kirch, 2002). The sea levels were much lower during this period, and humans would have been able to walk through most of what is now ISEA. At this point in time, Australia and New Guinea were still connected to each other, and together called the "Sahul landmass", but were disconnected from the rest of the continents. To traverse present day ISEA and reach Sahul, humans would have needed the knowledge of building some form of watercraft, as they would have encountered their first water barrier of at least 70 kms near Java and Borneo (Summerhayes *et al*, 2010). We find the earliest archaeological records of anatomically modern humans in Sahul dating back to 45,000 years ago (O'Connell and Allen, 2004). Evidence of human occupation in New Britain and New Ireland, which were not part of the Sahul landmass, also dates to around the same time, ~ 40,000 years ago (Leavesley *et al*, 2002). The eastern most limits of this Pleistocene expansion is believed to be Greater Solomon Islands, where archaeological records indicate human presence at least 30,000 years ago (Wickler and Spriggs, 1988). This region is referred to as Near Oceania, to delineate it from the more recently colonized islands of the Pacific, termed as Remote Oceania (Green, 1991b). The earliest archaeological records of human presence in Remote Oceania dates back to only around 3000 years before present (Petchey *et al*, 2014). Near Oceania includes the large islands of New Guinea, the Bismarck archipelago and the Solomon Islands as far as Santa Ana. Remote Oceania encompasses all of the islands from the Reef/Santa Cruz islands in the Southeast Solomons, to the eastern most limits of Polynesia and includes Hawaii and Rapa Nui (Easter Island) and Micronesia. Therefore, the distinction between Near and Remote Oceania, is not just geographical, but is one that delineates two major events in human history. Our focus will be on the hypotheses that have emerged out of research till date explaining the colonization of Remote Oceania.

1.6.1 Hypotheses explaining the colonization of Remote Oceania

The first wave of modern human migration did not proceed beyond Near Oceania and the limits to this first expansion were generally attributed to lack of developed navigation skills (Di Piazza *et al*, 2007; Spriggs, 1984). It was proposed that a second wave of dispersal (Bellwood, 1991) spread Austronesian speakers, beyond Near Oceania and into Remote Oceania, as far east as Rapanui (Easter island), as far north as Hawaii and down south till New Zealand (See Figure 2-1). Austronesian languages are spoken throughout Taiwan, ISEA, part of Near Oceania, and almost exclusively in Polynesia. Initial work of Blust (1999), classified the Austronesian language family into 10 sub-families, with all the extra-

Formosan languages grouped in one sub family and with representatives of the remaining 9 spoken only in Taiwan. There are several hypotheses regarding the origin, pace and mode of these migrations.

The first scenario dubbed the "Express Train" hypothesis (ETH) or the "Out of Taiwan" hypothesis (Diamond, 1988) puts forward the notion of an expansion, involving people who spoke a different language than the resident Melanesians. The hypothesis states that the language spoken by the new colonisers arose from or near Taiwan, with a completely different culture and genetic makeup than the Melanesians. According to ETH, the Austronesian speakers moved rapidly through ISEA, the Philippines, Near Oceania and further into Remote Oceania, becoming the first people to colonize this area. Diamond(1988) proposes that the Austronesian speakers, also the bearers of the Lapita culture[1], originated somewhere near or in Taiwan and left Taiwan around 5500 years ago, colonizing the Philippines by 5000 years BP, and reached the coast of Papua New Guinea around 3600 YBP. Subsequently, they were able to proceed to Near Oceania by 3200 BP, and enter Remote Oceania at about 3000 BP, finally reaching Hawaii by 1500 BP and New Zealand about 1000 BP. This scenario places the Austronesian Homeland in Taiwan, and the Austronesians to be the bearers of the Lapita culture who introduced innovations in farming, fishing, agriculture, husbandry and more importantly long-distance sailing to the Pacific. They rapidly progressed to colonize ISEA, and Near Oceania with little or no genetic exchange with already existing societies, before their spread into Remote Oceania. There were no significant pauses in the Austronesian expansion to Polynesia in this hypothesis.

The "Entangled Bank" hypothesis (EBH), also called the Bismarck Indigenous Inhabitants scenario (Green, 2003) proposes a more indigenous origin of the Austronesians, in or around Bismarck Archipelago and suggests a slow moving colonisation process, that is at the other end of the spectrum from the ETH. Other hypotheses that lie between the two extremes of ETH and EBH are for example, the "Slow Train" (STH) hypothesis (Hurles *et al*, 2002). While STH is an extension of the ETH, it mainly differs from the ETH about the pace of the Austronesian colonisation. Also according to the STH, significant genetic exchange took place along their route of colonization with pauses along the way. Another difference is that STH suggests an earlier start of the Austronesian expansion than

[1] Lapita is a distinct ornate style of pottery found at several archaeological sites throughout Oceania (Spriggs, 1990). Sherds in this style have been unearthed on islands located in a wide arc of the southwestern Pacific, from Aitape on the Sepik coast of New Guinea and stretching all the way eastward to Fiji, Tonga and Samoa" (Green, 1994; Terrell, 1997). It is hypothesized to be associated with the Neolithic Austronesian expansion and soon came to be identified as the "ancestral cultural complex" from which Polynesian culture was derived (Green, 1937 & 1974).

as hypothesized by ETH. Another hypothesis termed, "The slow boat" model (SBM) (Oppenheimer and Richards, 2001b), puts the origins of the Austronesian speakers in Island Southeast Asia. The proponents of this model view the origin of the Austronesian speaker's homeland to be around Eastern Indonesia (Sulawesi or Maluku). This theory is primarily based on evidences which suggests the existence of a safe voyaging corridor from eastern Indonesia through Bismarck Archipelago, all the way up to the Solomon Islands (Irwin, 1994). They also draw support from evidences presented for the ETH or other theories, but have interpreted in a manner which would seem to support the SBM theory.

One of the more popular of the proposed hypotheses, is the "Voyaging Corridor Triple I (Intrusion, Innovation and Integration)" model (VC Triple I) by Green (2003). Green suggests that there is continued interaction between Eastern Indonesia, Bismarcks and the Solomons from as far back as 6 KYA up to 3.5 KYA. Through these constant interactions and intrusions, innovations were developed and adapted to become the Lapita cultural complex. The sequence of events is the same as those in the ETH, but with much more interaction and pauses in between these events. The first event occurred around 4,000 years ago from Taiwan to Philippines. The next stage was the development of the Lapita cultural complex and movement up until the coast of Papua New Guinea, which reached the Bismarcks and Solomon Islands around 3500 – 3300 years ago. Finally, around 3200 years ago, it moved into Remote Oceania, with the last two waves around 1000 and 800 years ago moving into Eastern and Southern Polynesia respectively. Other theories like South American origins of the Polynesians proposed by Norwegian explorer Thor Heyerdahl (Heyerdahl, 1950) did not gain momentum due to lack of evidence (like lack of any Lapita sites in the Americas) and was soon dismissed.

1.6.2 Evidence from linguistics

Linguistic evidence, pioneered by Ross (Pawley and Ross, 1995; Ross, 1988; Ross, 1996) and Gray (Gray and Jordan, 2000; Greenhill and Gray, 2005) strongly supports the hypothesis of a Southeast China/Taiwan origin for the Austronesian expansion (Gray and Jordan, 2000), which subsequently expanded into Oceania. Further, recent work by Ross (2009), based on lexical evidence, places the Malayo-Polynesian subgroup of the Austronesian language family within the Formosan sub-group, instead of as a sister branch. This evidence further substantiates a Taiwan/Mainland Asia origin for the Austronesian speakers. In line with a "holistic approach", Gray *et al* (2009) adapted methods available in evolutionary biology for linguistic data, to test the origin, spread and evolution of the Austronesian language family. They found evidence for pauses during the Austronesian expansion, and this points to at least a moderate amount of interaction of the

Austronesians with the non-Austronesians in ISEA and Melanesia. Based on the grouping of the linguistic sub-families, and recent evidence from phylogenetic methods, the Express Train, VC Triple I and the Slow Train models are the most supported. Evidence from linguistics has mainly substantiated and raised questions regarding the timing and speed of the Austronesian expansion.

1.6.3 Molecular evidence

Genetics has played a major role in elucidating our understanding of human dispersal in the pacific. Mitochondrial DNA (mtDNA) and Y chromosome haplotypes[2] clearly delineate signatures of early human migrations into Near Oceania from those of the later migrations carrying ancestral Polynesians.
Early genetic evidence investigating mtDNA, supported a Taiwanese origin for ancestral Polynesians. A combination of a mtDNA mutation, with a characteristic single nucleotide polymorphism at three nucleotide positions (16217, 16247 and 16261), and a nine base pair deletion in the COII/tRNALys intergenic region, defined the B4a1a1a haplogroup. This haplogroup was found throughout the Pacific, but reached high frequency and near fixation in some islands of Polynesia and therefore was termed as the "Polynesian Motif"(PM) (Hagelberg and Clegg, 1993; Melton *et al*, 1995). The ancestral haplogroup to this clade (a precursor haplogroup from which the current B4a1a1a evolved), B4a1a, is found in Taiwan and throughout ISEA and therefore lent support to the Taiwanese origins and "Express Train" or Voyaging corridor Triple-I hypothesis (Friedlaender *et al*, 2007; Trejaut *et al*, 2005).

Kayser *et al* (2006) found that over 60% of the Y chromosomes in Polynesia are of Melanesian origin (primarily comprising of the C2aM208 haplogroup). Comparable to the PM of mtDNA lineages, the Remote Oceanian populations show a substantial presence of the Near Oceanic derived Y chromosome haplotypes (Kayser, 2010; Kayser *et al*, 2006; Kayser *et al*, 2000). While evidence from maternally inherited data (mtDNA) posited a Taiwanese origin to the Austronesians (Lum *et al*, 1998), paternally inherited data (Y chromosome) showed a very high affinity of the present day Polynesians to Melanesians (Kayser *et al*, 2000). Therefore, molecular evidence actually points to different origins of ancestral males and females of present day Polynesians. This apparent contradiction between male and female histories was initially attributed to the Lapita societies having a matrilineal descent structure and a matrilocal residence pattern (Hage and Marck, 2003), thereby driving sex-biased dispersal. However, a

[2] Haplotypes are defined as a DNA sequence with a particular sequence of nucleotides. Sequences differing in even one nucleotide base pair are considered as a different haplotype.

recent study of Tongans and Samoans revealed the presence of high frequency of an Asian derived Y-haplogroup (O3a2c-P164), which is also found in Ami (Taiwanese aboriginal tribe), reaching to almost 53% frequency in Tongan males, thereby providing direct support to a Taiwanese homeland for ancestral Polynesians (Mirabal *et al*, 2012).

Using autosomal and nuclear markers, Wollstein *et al* (2010) calculated approximately 87% Asian and 13% Near Oceanic ancestry of Polynesian people, supporting an origin of Remote Oceanians in Asia, but with substantial genetic exchange with Near Oceanic peoples. Pan-Asian SNP (single nucleotide polymorphism) dataset, indicated a population migration originating in East Asia and moving through ISEA and New Guinea around 4000 years ago, corresponding to the Austronesian expansion proposed through ETH (Xu *et al*, 2012). Using Asian-derived markers, it was discovered that in Near Oceanians, admixture on the X chromosome was discovered to be much greater than across the genome generally, indicating a sex-biased admixture (Friedlaender *et al*, 2008). Asian admixture was found to occur at a low frequency, and only along the coast of New Guinea, through the islands of Bismarck Archipelago, indicating that the Austronesian expansion followed a coastal route, and any admixture was probably sex-biased and was limited to the societies living on the coast (Friedlaender *et al*, 2008).

In summary, while the mtDNA evidence supported an Asian/Taiwanese origin (lending support to "Express Train", "Voyaging Corridor Triple I" hypotheses), the Y chromosome evidence seemed to indicate a more insular origin of the Polynesians in Melanesia or Bismarck Archipelago (supporting the "Slow Boat" hypothesis). The nuclear and autosomal markers only indicate a sex-biased admixture, with a greater contribution of Asian genes to the Polynesian gene-pool in some populations, and does not necessarily reject/lend strong support to either school of thought.

1.6.4 Summary

To summarise, several studies using genetic, cultural and linguistic data explored the proposed plausible hypotheses of the origin of present day Oceanians. Using linguistic data, Gray and Jordan (2000), and later Gray *et al* (2009) have been able to put to rest the debate regarding the origin and timing of the Austronesian dispersal, both largely supporting the Out of Taiwan and Voyaging Corridor Triple I models. However, the analysis of sex-specific markers (mitochondrial DNA (mtDNA) – maternally inherited, and Y chromosome – paternally inherited) displayed contradictory patterns. While mtDNA data posited a Taiwanese origin to the Austronesians (Lum *et al*, 1998), Y chromosome (paternally inherited) data showed high affinity of the present day Polynesians to Melanesians (Kayser *et al*,

2000). The bi-parental markers show a varying amount of genetic admixture of the two hypothesized homelands (Kayser *et al*, 2008; Wollstein *et al*, 2010). Therefore, molecular evidence actually points to a different origin for the ancestral males and ancestral females of present day Polynesia. exist.

Hage and Marck (2003) and later on Kayser *et al* (2006) and Kayser (2010) attempted to explain the reason for the contradicting origins of male and female sex-specific markers, through the examination of of cultural traits like post-marital residence. They draw on the evidence of a predicted ancestral matrilocal trait of Austronesians (Jordan *et al*, 2009) to explain this contradiction in genetic patterns. Matrilocal post-marital residence is where, post-marriage, the couple reside with the kin of the bride (See Chapter 2 & Chapter 3 for more detailed explanations). And as Austronesian societies were hypothesized to be matrilocal, scholars have speculated that there was little exchange of women during the course of their dispersal to Polynesia. In contrast, such a system allows for a high gene flow of men from outside the population (i.e., men from outside settling with dispersing Polynesian societies), which would indicate the varying origins depicted by the sex-specific markers of mitochondrial DNA and Y chromosome. There has been no testing of whether post-marital residence can in fact influence sex-specific dispersal and if so, how long are these patterns persistent, and under what conditions (a more detailed explanation is available in Chapter 2). This mode of sex-specific dispersal has also not been incorporated into modelling population history and tested whether this could be a plausible scenario which gave rise to the present state of genetic diversity that we observe.

Another dimension to attributing post-marital residence norms as the predictor variable for sex-specific genetic diversity is that these norms themselves are variable across the world, undergo change from time to time, and are known to be influenced by several other factors like migration and depopulation events (Eggan, 1966; Murdock, 1949a), warfare/feuding (Ember and Ember, 1971; Otterbein and Otterbein, 1965), male-absence (Harris, 1980) and the sexual division of labour (Ember and Ember, 1971). Of these, specific hypotheses have been proposed on the influence of sexual division of labour norms on post-marital residence practises. As already explained, we now have the advantage of using evolutionary methods for testing the dynamics of cultural traits, and we could test hypotheses of the influence of sexual division of labour traits on post-marital residence traits. Given how it is considered that Lapita culture brought innovation in foraging strategies, and this would influence the sexual division of labour for subsistence, it is important to explore how the sexual division of labour could influence post-marital residence norms. This has direct consequences on the information we incorporate in our population models. This would also help understand the atmosphere of cultural change in Remote Oceania.

1.6.5 Questions

The Austronesian context is interesting to us because of the variety of testable hypotheses regarding its pre-history, and also the variation in social norms it presents. This would help in teasing out the co-evolutionary hypotheses of genes and culture, as well as the evolution of cultural traits themselves. Taking advantage of the variety of norms that exist in the Austronesian societies, I explore the gene-culture co-evolution hypothesis pertinent to Remote Oceanian colonization, i.e., effect of post-marital residence norms on genetic diversity of sex-specific markers. I test whether in fact post-marital residence norms influences genetic diversity in a predictable manner, and if yes, whether it is generalizable in all demographic conditions.

By taking advantage of the evolutionary methodology applicable to cultural and linguistic data, I also try to add to the existing knowledge of Austronesian pre-history. Given the importance of post-marital residence norms, I try to understand the drivers of this trait and test different hypotheses regarding its evolution in an evolutionary framework.

Given the speculation on post-marital residence norms driving the differential sex-specific origins of Remote Oceanians, I build testable models by incorporating information from the first step (the effect of post-marital residence on genes), and test different plausible hypotheses for the patterns we see, in a statistically robust framework. This investigation is about exploring the ways in which a multi-pronged approach to pre-history can (a) interrogate some of our assumptions underlying models (b) reveal evolutionary processes that explain patterns of diversity.

1.7 Dravidian context

It was only after the independence of India in 1947 that a substantial and continuous tradition of work on the Dravidian languages was established, when American linguists trained several Indian scholars in modern and historical linguistics that led to an increase in the descriptive work of the Dravidian languages. Several languages like Kota and Toda of the Nilgiri region (Emeneau, 1938; Emeneau, 1944), and Dravidian languages of central India (Burrow and Bhattacharya, 1953; Burrow and Bhattacharya, 1960; Burrow and Bhattacharya, 1963; Burrow and Bhattacharya, 1970), were described for the first time. Burrow and Bhattacharya (1961) was the first etymological dictionary published for the Dravidian language, and this work has proved indispensable for facilitating comparative work on the Dravidian languages (Krishnamurti, 1961). The history of Dravidian languages in India is highly contradictory and is plagued by studies with inadequate sample sizes or flawed classifications. Based on all the evidence, there is still no consensus on the origin and further spread of the Dravidian

languages, and it is difficult to answer this question given the scanty archaeological or literary evidence. Given that India lies at a cross-roads of human migration and has an extensive history of admixture it is even more difficult to delineate the population history of Dravidian speakers using genetic data.

The Indian subcontinent encompasses a variety of biogeographical, biological, ethnological and linguistic zones. Surrounded by the Himalayas on the North, the Iranian plateau on the west and Indian Ocean on the south, and with a population of 1.2 billion people, India is one of the most densely populated regions in the world. Through its history and pre-history, South Asia has been a melting pot of various ethnic groups, cultures and languages. While the written history of the sub-continent dates only to two millennia, this glimpse reveals numerous accounts of invasions and a multitude of cultural contacts with societies near and far. This leads us to believe that such interactions extend to a far longer prehistoric period of the region, emerging as the highly diverse Indian population we see today. Languages of the Indo-European, Dravidian, Austroasiatic and Sino-Tibetan language families are spoken in the country. Of interest to us, currently restricted to the Southern part of the Indian sub-continent, and with language isolates in Balochistan to the west and parts of northern India, is the Dravidian language family with about 85 languages (Lewis *et al*, 2009). It was only in 1816 that Ellis recognized and delineated the languages of this family as an independent language family. Following this, Caldwell in 1856 proposed a relationship between Dravidian languages and Elamite at Behistun and coined the term "Dravida" for this language family.

McAlpin (1974), after analyzing linguistic data proposed that Dravidian language family had split from the Elamite language family, and that proto-Dravidian speakers originated out of India. Whereas Campbell (1998 & 1999) proposed that the origin of this language family lies in the northern part of India, where we still find members of this language family in isolated pockets. He further hypothesized that the heartland of Dravidian language diversification was in this north-western part of India, with a subsequent migration of the remaining Dravidian languages into the rest of India. It was argued that the presence of Brahui in the north-west indicated an entry of Dravidian languages into India from the West and a diversification within India subsequently, with Brahui representing an ancient lineage. While the current spread of the Dravidian languages is restricted to South India, there are speculations that it was previously wide-spread and the current restricted distribution is due to the introduction of Indo-European languages in India (Krishnamurti, 2003). Another hypothesis put forward by Fuller (2003), based on well-argued archaeobotanical evidence suggested against an Elamite link or north Indian origin, and proposed an indigenous, south/central Indian origin of the Dravidian language family (discussed later). So, broadly speaking, there are competing hypotheses on an

indigenous origin of the Dravidian language family versus an out of India origin and subsequent entry into India and diversification of the language family.

To understand the origins of the Dravidian language family, several attempts using molecular techniques were undertaken. Rosenberg *et al* (2006), explored the genetic variation amongst different linguistic groups and discovered very low levels of genetic divergence, and very high admixture, contrary to the patterns found in countries like Nepal, where there is a strong gene-language correlation (Yngvadottir, 2007). Due to the lack of any genetic signal related to language groups, subsequently, most genetic studies focused on "Caste-tribe continuum", and used these findings to make assumptions on the origin and diversification of different linguistic groups (Chaubey *et al*, 2007). Caste is a social institution governed by sanctions from, and deeply entrenched with, the Hindu religion. Believed to have been composed between 1700 – 1100 B.C., the earliest known scripture of Hinduism, the Rig Veda, gives an insight into this social stratification into castes (Rao *et al*, 2009). This institution evolved to direct all social, economic and religious activities of people. Stratification was based on the "Chaturvarana doctrine", i.e., four levels of hierarchical divisions. These four divisions are based on the role of each person in the society, and the hierarchy is linear (the social status given to each caste reduces as one proceeds down the division). It is hypothesized that Indo-European speakers introduced the caste system in India, whereas the Dravidian and the Austroasiatic speakers (supposed to have already been present in India when Indo-European speakers arrived in India), existed in tribes. However, a large number of Dravidian speakers today also are merged into this caste-based system, while few still exist as tribal societies (e.g., Todas, Koya etc.). Given that language did not serve as a group marker genetically, in Indian societies, genetic work then shifted to exploring the origin of the caste-based system, and then tried to link these findings to infer the history of different linguistic groups in India.

Genetic work focusing on the origin of the caste-system mainly postulated male-mediated migration of the Indo-European speakers, who pushed the previously wide-spread indigenous Dravidians towards Southern India and Sri Lanka, while establishing the Indo-Aryan speakers as the upper castes of the society (Bamshad *et al*, 2001; Basu *et al*, 2003; Cordaux *et al*, 2004; Quintana-Murci *et al*, 2001; Sahoo *et al*, 2006; Sengupta *et al*, 2006; Thanseem *et al*, 2006; Wells *et al*, 2001). Further, the caste-populations were postulated to be closer to Central Asians, and Europeans, and they were genetically divergent from tribal populations (non-Indo European) (Bamshad *et al*, 2001; Bamshad *et al*, 1998; Cordaux *et al*, 2004). This indicated that the caste-populations were representative of a recent migration originating out of India, and tribal populations represented Dravidian and Austroasiatic speakers. Many of these studies were plagued by small sample sizes, incorrect assignments and flawed inferences. A review by Boivin (2007) & Endicott *et al* (2007) summarises the

limitations of the genetic studies based on the caste-system. Molecular work since, has revealed that while ancestral north Indians are closer to the higher castes in the caste based system, the ancestral south Indians are closer to the tribal communities (Reich *et al*, 2009). Based on this, it has been inferred that it was in Southern India, that Dravidian languages were more widely distributed, and these communities later adopted to the caste-system. The close links between "Ancestral South Indians" and Dravidian tribal societies has largely been taken to mean a support for an indigenous origin, i.e., a South/Central Indian origin for Dravidian languages.

Molecular evidence for understanding Dravidian history, is confusing at best. Given the highly admixed populations, where direction and timing of admixture is not clearly decipherable, it is difficult to test hypotheses regarding the spread and origin of a particular set of native language speakers, without anyway of clearly delineating those without a history of admixture. Strong arguments regarding the origin and spread of Dravidian emerged from archaeobotanical and archaeozoological data (Fuller *et al*, 2004; Fuller, 2003; Fuller, 2006; Fuller, 2007; Madella and Fuller, 2006; Saraswat, 2004; Thomas and Joglekar, 1994). Fuller argued against an out of India homeland for Dravidian based on the archaeological dating of the presence of crops in peninsular India, and the lexical evidence for Dravidian origin of some of the crops like millets and beans. Fuller (2007) states that existing hypotheses on proto-Dravidian origin are proposed on the assumption that Dravidian societies were agricultural. However, Fuller argues that there is substantial evidence to indicate that early Dravidians were essentially 'Mesolithic', but with the technology for threshing/de-husking, grinding and some degree of storage, implying that Dravidian societies may have already some domestic fauna before the spread of Neolithic agriculture into the country. Therefore, we do not need to look at Indo-Iranian borderlands for its origin or entry based on the assumptions of an agricultural society. Fuller suggests an indigenous origin of Dravidian language family and explained the presence of Brahui in the North-western India, as a westward expansion of this family. He proposes that the homeland of Dravidians lies in South/Central India, resulting in Brahui's presence in Balochistan and the other Eastward towards East India and Nepal, indicated by the presence of Kurukh and Malto branches. This is also supported by Y-chromosome evidence (Sengupta *et al*, 2006), where it was found that the distribution of Y chromosomes reveal a peninsular origin for Dravidian speakers, with significant input from demic diffusion associated with agriculture. This data also supports the hypothesis that there were two migrations originating from central or southern India, which took speakers eastward, and then west towards Gujarat or Rajasthan eventually. There is also a theory put forth by Winters on an African origin, but this has not gained much credit (Winters, 2010)

There is still no consensus on the origin of the Dravidian languages, and whether it was within India, followed by a diversification, or an out of India origin,

with a west-ward entry, and whether Brahui represents an ancient lineage. It also raises the question that if the Dravidians were previously widespread, why were they now restricted to the southern part of the sub-continent. As discussed earlier, we draw upon the strength of the evolutionary nature of language to investigate the unanswered questions regarding the history of the Dravidian language family. Gray and Jordan (2000), and subsequently others (Gray *et al*, 2009; Holden, 2002), have shown how population models using phylogenetic methods on linguistic data can also be used to investigate population history. While linguistic data has been used to explain relationships between the Dravidian languages, these have several drawbacks. Attempts have largely been using distance based methods, which do not correct for historical relatedness (Andronov, 1964). Secondly, other studies have been limited in their methodology or sampling (Rama *et al*, 2009) to make any robust inference. Wells *et al* (2001) discusses an overview of previous quantitative work, including various distance-based phylogenetic analysis of morphological and lexical data drawn from existing datasets. Taraka and Kolachina (2013) conducted a distance-based analysis of lexical data from the Dravidian Etymological Dictionary (DEDR), 2nd edition (Burrow and Emeneau, 1984), and lexical data from Krishnamurti (2003). However, neither of these datasets have been gathered with the aim of character-based phylogenetic inference for Dravidian family relationships, as has recently become the standard for similar investigations of other language families, including Austronesian (Gray *et al*, 2009) and Indo-European (Bouckaert *et al*, 2012). The current study remedies this gap by combining appropriate first-hand collection of lexical data with sophisticated Bayesian phylogenetic inference methods

I have attempted to address some of the outstanding questions regarding the Dravidian language family by employing modern phylogenetic methods to linguistic data (See Chapter 5 for details). The questions asked were: a) what was the historical distribution of the Dravidian language family? Was it spread across the North? and b) how did the current substructure of the Dravidian family emerge? and was it possible to fit a sequence to the evolution of the language family? This is the first attempt at building a robust phylogenetic tree of the Dravidian language family. This enquiry added another dimension to testing hypotheses using linguistic data, especially when information from molecular evidence is not useful to address any debates regarding pre-history.

1.8 Thesis Outline

In Chapter 2, I explore the premise of gene-culture co-evolution and correlation. Given the Austronesian context, where speculations have been made regarding the effect of post-marital residence on sex-specific marker genetic diversity, the hypothesis whether cultural traits have an influence on genetic data and if so, is the change, in the anticipated direction with predictable outcomes, is tested. This

holds huge implications for Austronesian pre-history, I also test whether this supposition is valid in all demographic conditions and whether we could make generalizations regarding the effect of cultural traits on genes. These tests are conducted through a rigorous simulation driven framework, by looking into the effects of post-marital residence on the genetic diversity of sex-specific markers. The results would also give us an insight into the dynamics of genes and culture.

Cultural traits like post-marital residence have been used as important precursors to understanding the dispersal of Polynesians and very little is known regarding the factors that influence this trait. In Chapter 3, by taking advantage of the existing evolutionary methods, hypotheses regarding the factors that influence post-marital residence in the pacific are tested. Sexual division of labour was hypothesized to drive a change in post-marital residence (Driver and Massey, 1957) and the postulated hypotheses are based on cross-cultural observations with no control for historical relatedness. There are also several debates regarding the validity of this hypothesis (White *et al*, 1981, See Chapter 3 for more details). I test the different hypotheses by using co-evolutionary analyses in comparative framework, to understand if sexual division of labour drives a change in post-marital residence traits and if yes, to tease out the mechanisms that drive this change.

In my fourth Chapter, I explore incorporation of the triangulation technique into inference of historical processes. The hypotheses of sex-biased dispersal in the pacific is tested within in a coalescent driven simulation framework, by incorporating information derived from cultural and linguistic history of the Austronesians. Most importantly, the plausibility of post-marital residence driving sex-specific marker patterns in Remote Oceanians versus other possible hypotheses is quantitatively tested in this population genetic framework. This also serves the purpose of understanding the importance of using a holistic approach to understand human evolution

In Chapter 5, I exploit the evolutionary nature of language to understand the evolution of the Dravidian language family. This is an improvised attempt at building a Dravidian phylogeny and to understand the evolution of the language family to its present structure, through the testing of different hypotheses. The other aspect to this Chapter is that it posed a completely different challenge, when compared to the pacific (in terms of admixture, existing data on population history), It highlighted the importance of being able to use linguistic data in unravelling information, where other evidence was not helpful in deciphering a coherent picture of pre-history.

2 Genes to Culture: Correlation of Sex-Specific Markers and Post-marital Residence

"Society works not because we have consciously invented it, but because it is an ancient product of our evolved predispositions. It is literally in our nature." - Matt Ridley

2.1 Abstract

Investigating evolutionary processes that underlie patterns of human genetic diversity that is seen today is crucial for building an integrative picture of human pre-history. Sex-linked genetic markers have been used to explain male and female histories in a population (Kayser *et al*, 2006). In recent studies, differing male and female histories have been linked to cultural processes like post-marital residence practices (Hage and Marck, 2003). There is a debate about whether cultural patterns can affect genetic patterns of a population (Heyer *et al*, 2011; Wilkins and Marlowe, 2006) and if they do affect genetic patterns, is there a "time-lag" before the genetic pattern can catch up to reflect the change in cultural patterns (Bolnick *et al*, 2006; Chaix *et al*, 2007; Gunnarsdottir *et al*, 2011). I take a novel approach to test these (sometimes post-hoc) hypotheses, using a forward simulation framework, for different demographic scenarios of a constant, growing and declining population size. I investigate whether a change in post-marital residence brings about a change in genetic diversity patterns. By quantifying the number of generations it takes for a switch in post-marital residence to be apparent in genes, I find support for the time-lag hypothesis, but this is dependent on demography, migration rates and drift.

2.2 Introduction

The advent of molecular tools has considerably increased our knowledge of human history. Information from genetic markers has added crucial details regarding the process of dispersal of humans, especially of the major migrations in Holocene (Cavalli-Sforza *et al*, 1993; Henn *et al,* 2010; Kayser, 2010; Stoneking and Krause, 2011). Molecular anthropology has significantly aided in testing hypotheses emerging from the fields of archaeology, anthropology and linguistics regarding human history (Diamond and Bellwood, 2003). The knowledge of population genetics has not only facilitated in gaining insights into the dispersal and demographic aspects of human migrations, but also the co-evolutionary relationship between genes and culture. Tracing the history of dairying through genes, which are associated with the ability to digest lactose is a very good

example for the co-evolutionary nature of genes and culture. The ability to digest lactose in adulthood is not species-typical, and it was discovered through genetic data that this ability is linked to single dominant gene present in societies with a history of dairying. This demonstrates the intricate bond between genes and culture and how population genetics can aid in elucidating significant cultural aspects of a society (Durham, 1992; Feldman, 1996).

In the excavations of burial sites at Eulau (Germany), there was a lack of archaeological evidence on social organization or kinship of the discovered late Stone Age society (Haak *et al*, 2008). However, the contrasting patterns of mtDNA and Y chromosome revealed a differential origin for males and females, i.e., using a combination of isotope analysis on diets and ancient DNA evidence, it was discovered that the men had not moved from the place of their birth and while the women had a different origin from that of men. This pattern of SSM markers was attributed to the society probably practicing patrilocality and exogamous marriage system, thus resulting in a differential origin for males and females. The ability to determine the dynamics of social interactions using SSM shows the potential of using the co-evolutionary relationship between genes and culture in gathering information regarding the cultural traits of many prehistoric societies. These insights into gene-culture co-evolution has facilitated in gaining information regarding the social processes governing a society, especially where adequate evidence is not available to make direct conclusions (Feldman, 1996).

Seielstad *et al* (1998) demonstrated convincing evidence for a co-evolutionary relationship between social systems and genes, by showing that the cultural processes of marriage and post-marital residence were highly correlated with patterns of mtDNA and Y chromosome variation. Different post-marital residence practices such as patrilocality (where after marriage, the couple resides with or near the family of the groom) and matrilocality (where after marriage, the couple resides with or near the family of the bride) left a detectable and unique genetic signature on SSM. In a global genetic analysis, Seielstad *et al* (1998) found that MSY had a highly localised structure and large between-population distances, whereas mtDNA patterns were of high diversity and low between-population genetic distances. The majority of the societies today, across the world are patrilocal, and in an intergenerational perspective, this means there is a higher migration of females in patrilocal societies. This higher migration of females, in principle, leads to a diverse mtDNA structure and a localized Y chromosome structure globally, congruent with the findings of by Seielstad *et al* (1998).

Heyer *et al* (2011) and Wilkins and Marlowe (2006) argued that the effect of female/male migration based on marriage and residence is a local phenomenon in societies, and thereby challenged the theory put forth by Seielstad *et al* (1998). By analysing a different global set from that of Seielstad *et al* (1998), Heyer *et al* (2011) found that the pattern expected from patrilocality did not seem evident. They postulated that apart from residence norms, effective population sizes,

variance in reproductive success and social factors like descent rules must also play an important role in shaping the diversity of SSM markers. The effect of post-marital residence norms on the diversity of SSM is a matter of debate. In particular, the extent to which it shapes the pattern of SSM is not clearly understood.

2.2.1 Deciphering cultural traits using SSM

Although other aspects of kinship can affect patterns of genetic variation, studies have focused on investigating the links between residence norms and genetic diversity, because post-marital residence is a central organizing aspect of kinship acting to localize related men and women. The first direct evidence to support the hypothesis that SSM and post-marital residence were linked in a predictable manner i.e., for example, low MSY and high mtDNA variability associated with patrilocality, and that social systems can be inferred through studying SSM patterns, was by Oota and colleagues (2001) from a population in Thailand. Following this study from Thailand (Oota *et al*, 2001), whenever SSM show completely contrasting diversity patterns, it has become common practice to attribute them to post-marital residence practices e.g.: Africa (Destro-Bisol *et al*, 2004), South America, (Mesa *et al*, 2000). Proponents of the different hypotheses on the Bantu expansion have relied on evidence from SSM (de Filippo *et al*, 2011; Salas *et al*, 2002; Scozzari *et al*, 1999; Veeramah *et al*, 2010; Wood *et al*, 2005) to demonstrate the strength of their propositions. The contrasting marker diversity patterns of mtDNA and Y chromosome was attributed to differential male and female origins (Pakendorf *et al*, 2011), citing the coevolution of residence pattern and SSM marker diversity. The significantly higher mtDNA diversity was attributed to the patrilocal tradition of the Bantu communities, where Bantu men married women from local indigenous populations along their route of migration. This was accompanied by very little or no male migration, thus explaining the restricted Y chromosome diversity. In some of the studies using SSM to understand history of a population, it was observed that the varying dynamics of population structure and demographic events affected the structure of the SSM genes to a great extent (Belledi *et al*, 2000; Chaix *et al*, 2007; Kittles *et al*, 1999; Tarazona-Santos *et al*, 2001). Accounting for these processes is particular crucial in scenarios where the patterns of SSM are used to argue a case regarding dispersal histories and prehistoric migrations, as in the Bantu expansion in Africa (Pakendorf *et al*, 2011) or in the Pacific (Kayser *et al*, 2006; Lum *et al*, 1998; Oppenheimer and Richards, 2001a). Hence, there is a need to quantify and understand the dynamics behind gene-culture co-evolution with respect to SSM and residence.

For certain cases, the reflecting genetic pattern did not always match the practiced residence norm as expected (Bolnick *et al*, 2006; Chaix *et al*, 2007;

Gunnarsdottir *et al*, 2011). One possible reason attributed for the genetic patterns of male-specific Y chromosome patterns to have deviated from expectation was immigration regulation: the control a society exerts over "non-locals" marrying into the society. Immigration regulation was generally discovered to be stricter in patrilocal rather than matrilocal societies (Hamilton *et al*, 2005). Another reason for the deviation from the expected pattern was attributed to a recent shift (5-6 generations ago) in residence patterns. Raff *et al* (2011) showed that the original pattern of the post-marital residence reflected at least 5-6 generations after a switch in residence and it disappeared after twenty generations (Chaix *et al,* 2007). For example, in a study on the Semende (matrilocal) and Besemah (patrilocal) groups from Sumatra, the MSY did not show any significant difference in diversity between the two groups (Gunnarsdóttir *et al*, 2011). In this particular case however, it seemed that matrilocality was more tightly regulated than patrilocality or that patrilocality was more loosely regulated than matrilocality, resulting in very similar MSY patterns in Semende and Besemah. A recent change in the residence norms of the Besemah, i.e., less than 20 generations, from matrilocality to patrilocality could also be one reason for a deviation from the expected MSY pattern in this group. Several studies have referred to this "time lag" when genetic patterns of the sex-linked markers did not correspond to the post-marital residence practise (Bolnick *et al,* 2006; Chaix *et al,* 2007; Gunnarsdóttir *et al,* 2011). However, there exists no quantification of this "time lag", or if it is generalizable in nature.

It is evident from these studies that an intricately complex system links social systems of a society to its genetic structure. It also shows that the mechanism driving the relationship between SSM patterns and residence norms is not simplistic, and the dynamics defining the relationship need to be tested before any inferences can be made regarding the hypothesized correlation between SSM patterns and residence practices. To date, there has not been a study quantifying the effects of post-marital residence under different demographic conditions, or the quantum of this effect under different conditions and whether this effect is discernible or not.

2.2.2 Genes and post-marital residence in Oceania

The Pacific, and in particular Oceania, is an apt study area to test questions regarding the coevolution of SSM and residence. It encompasses a large variation in post-marital residence (Jordan *et al*, 2009; Jordan and Shennan, 2009), which provides us with ample opportunities to test different residence scenarios leading to the genetic patterns in SSM. There has also been a large body of interdisciplinary interest in this region focussing on dispersal and colonisation history, providing information regarding the possible demographic mechanisms that shaped genetic and residence variation in this region (Gray and Jordan, 2000;

Kayser *et al*, 2006; Kirch and Green, 2001; Oppenheimer and Richards, 2001a; Terrell, 1988).

There have been several theories regarding the origin of the Austronesian expansion, which carried Austronesian speakers into near and Remote Oceania, colonising present day Polynesia. It was proposed that after the initial out of Africa migration that settled Near Oceania with non-Austronesian speakers, there was a second wave of migration carrying Austronesian speakers across ISEA, Near Oceania and the first colonizers of Remote Oceania (Diamond, 1988). Initial mtDNA results revealed a strong Asiatic affinity of the Austronesians (Lum *et al*, 1994). It was assumed that the origins of the Austronesian speakers lay somewhere on the coast of mainland Asia or Taiwan. However, there have been several studies since then suggesting a more geographically close, "Melanesian" origin for the Polynesians (Lum and Cann, 1998; Richards *et al*, 1998). Y chromosome studies also revealed a geographically closer, local, Melanesian origin for the Remote Oceanians (Kayser *et al*, 2000). In the study by Kayser *et al* (2000), the Polynesian Y chromosomes were found to have originated from the stock of Melanesian populations in the study. Genomic studies have also confirmed this contrasting pattern of substantially higher contribution of Asian ancestry to women than men (Wollstein *et al,* 2010). The contrasting Asian origin of the mtDNA (Kayser *et al*, 2006; Lum *et al*, 1994) and Melanesian origin of the Y chromosome (Kayser *et al*, 2000) were attributed to differing male and female histories of the Remote Oceanians (Hage and Marck, 2003; Kayser *et al*, 2006). Linguistic and genetic evidence thus suggested that the ancestral Austronesian societies were most likely matrilocal (Hage and Marck, 2003), and it was proposed that during the Austronesian expansion, men from the local Melanesian populations married into Austronesian communities with little or no female mediated gene flow. Therefore, the contrasting diversity patterns and origins of mtDNA and Y chromosome was attributed to the matrilocal practice of the migrating ancestral Remote Oceanians. Jordan *et al* (2009), using comparative phylogenetic methods on cross-cultural data, supported the plausibility of ancestral matrilocal residence in ancestral Austronesians; however, the Austronesians during the course of their dispersal to colonize Remote Oceania, would have undergone severe bottlenecks, serial founder effects and drift (Ramachandran *et al*, 2005). As seen in other studies (Belledi *et al*, 2000; Chaix *et al*, 2007; Kittles *et al*, 1999; Tarazona-Santos *et al*, 2001), these demographic and stochastic events could well have influenced the SSM marker patterns seen in the Remote Oceanic populations. Also, recent studies using advances in genomic techniques indicate that early colonisers of Remote Oceania in Vanuatu and Tonga did not carry any signatures of admixture with Near Oceanic non-Austronesians (Skoglund *et al*, 2016), whereas genomic data from contemporary populations showed a significant NO-NAN genetic association (Posth *et al,* 2018). Recent work also suggests that colonisation scenario is not as simple as a sex-biased admixture

and subsequent colonisation, but rather a series of colonisation events, by people of varying ancestry (Posth *et al*, 2018). It is therefore necessary to not assume that the observed patterns of mtDNA and MSY are solely due to social systems without investigating all plausible models of population history.

The SSM marker diversity and information on post-marital residence have played a major role in hypothesizing both the origins and social systems of present day Austronesians. However, no effort has been made to investigate whether these SSM marker diversities are also a result of stochastic demographic processes like drift, or other processes like bottleneck events. We do not even know if the effect of post-marital residence are discernible in Austronesian societies. Therefore, it was prudent to understand these phenomena: first, the effect of post-marital residence on SSM marker diversity under different demographic conditions and influence of stochastic events and second, whether we can make robust inferences, regarding post-marital residence traits, by just using SSM marker diversity patterns.

The studies inferring or attributing the discordancy between mtDNA and MSY genetic structure to residence were based on real populations, where factors other than social systems could have affected genetic diversity. Evolutionary and demographic processes like drift, bottlenecks, and population size change are known to contribute in shaping of genetic patterns in a population (Ramachandran *et al*, 2005). We cannot discount the importance of accounting for these effects in shaping SSM diversity. Here we use forward simulations as a unique tool with which to investigate evolutionary processes, and their effects on populations, under different conditions. With this method, we can tease apart the cause and effect relationship between different variables affecting genetic structure. We are also able to manipulate how different population parameters affect a population in a scenario of our choice.

A number of forward simulation approaches have been developed and used in the recent past (Hoban *et al*, 2012). For example, Bruford *et al* (2010) used forward simulations as a predictive tool to demonstrate the consequences of different management actions on orang-utan populations along with traditional ecological modelling. These populations inhabit a fragmented forest with low population connectivity, a factor that has led to population decline of the orang-utans. Forward simulations helped understand the roles different variables played in affecting the future of the orang-utan populations. Simulating the consequences of different scenarios like non-intervention, translocation and corridor establishment with both ecological and genetic data helped understand the critical need for intervention, without which the populations would have become extinct in a very short period of time. The use of genetic data along with ecological data, in a forward simulation framework revealed that the best management intervention was to establish corridors, but also to translocate certain individuals before the population went extinct due to inbreeding. This information could have

only come by incorporating genetic data in the models. Forward simulation, therefore, is an apt method to understand the effect of different and changing residence patterns on the diversity of SSM of a population, and to also test if there is evidence to indicate a time-lag. This would be an appropriate way of testing the effects of different demographic and social processes on a population's SSM diversity, and to apply it in a real world setting.

2.3 Objectives

In this Chapter, the mechanism shaping SSM diversity under different demographic and social conditions is investigated. The results are compared to real world populations of Oceania to verify if any of the generalizations from forward simulation models would be able to predict conditions in a real population. As information on the social and demographic conditions of Oceanic societies are relatively well documented, it would be a good system to interpret the results for these societies and test them against existing hypotheses.

In this study I first simulated a genetic dataset forward in time, under different demographic conditions relevant to what is known in Oceania, in an aim to find out:

a) How does post-marital residence affect SSM genetic diversity patterns under different demographic conditions?
b) Is there evidence for a time lag in the change of genetic pattern after a shift in residence practise?

The results of the simulated data set and patterns observed were then compared to information known from Oceanic societies to see

c) Whether genetic patterns actually correlate with post-marital residence patterns in Near and Remote Oceania and if not, what could be inferred regarding the discordant mtDNA and MSY results, based on the results from the forward simulations and known history of Oceania.

2.4 Materials and Methods

2.4.1 Forward simulations

Forward simulations using an individual based simulator – simuPOP were conducted (Peng and Kimmel, 2005). It was possible to simulate different practices of post-marital residence and sex-biased dispersal in these simulations by setting differential migration rates for mtDNA and MSY. A large number of mating schemes and operators are available to manipulate the evolution of populations. Different demographic processes that could potentially affect genetic patterns were then simulated.

Models involving two populations, A and B, where the individuals were modelled with sex specific chromosomes (mtDNA& MSY) were simulated. The duration of evolution for these populations was fixed at 700 generations (for generation times, refer to Fenner (2005)). The populations had a specific post-marital residence practise at the start of evolution, and during the course of evolution, either at the 100th, 250th or 400th generation, the populations underwent a split in their evolutionary trajectories. The generation times were decided keeping in mind the timeline of the start of dispersal and settlement of the Pacific. The Austronesian expansion leading to the colonisation of Remote Oceania began at maximum around 6000 years ago i.e., ~250 generations (Gray *et al*, 2009). Therefore, 400 generations is an ample interval to include any change in post-marital residence before the start of Austronesian expansion. As the mutation rates of mtDNA are much slower than MSY, any change will take longer to be visible for mtDNA. Therefore, we modelled a generation time of 700, to account for the differential mutation rates of mtDNA and MSY. In one trajectory, the initial form of residence is continued and in the second trajectory there is a change in post-marital residence. In the second trajectory, the populations A & B were named as A' & B' respectively. Endogamy is not considered as an option in the simulations, i.e., where marriages happen within the same community and the distance either male or female of a couple has moved after marriage does not change. The reason for this is that we would not be able to detect such a pattern with genetic sampling, as genes would be moving within a population. Therefore, populations where geographic boundaries are clearly established, and marriages and post-marital residences occur across these boundaries, are considered appropriate for simulations.

Human populations are dynamic, and real populations fluctuate between growth, decline and constant size during their course of evolution. The following variables were modelled to account for all possible fluctuations in a population.

- Nine different migration rates (m) (0.001, 0.0025, 0.005, 0.0075, 0.01, 0.025, 0.05, 0.075, and 0.1, per generation) were used to test the effect of rate of migration on changing genetic patterns.
- For each migration rate m, the switch occurred at generation (n): 100, 250 or 400. The change in timing of the switch was to test if the emerging pattern was dependent on the number of generations a society had practiced a particular form of residence pattern
- For each switch n, the population model (s), was modelled to be exponentially growing (from 100 to 100,000), declining (from 100,000 to 1000) or a constant size population of 1000 individuals. This was in order to determine what effect population size, and in turn demography and drift, had on detecting signatures of post-marital residence changes in genes.

- For each population model *s*, all six possible combination of residence changes (p_i *to* p_j), were modelled: ambilocal to matrilocal or patrilocal; matrilocal to ambilocal or patrilocal; and patrilocal to ambilocal or matrilocal.
- Under each combination of parameters, each residence change (p_i *to* p_j), was run five times. In the five runs, at each generation, statistics for genetic variability (mean number of pairwise difference and expected heterozygosity) and genetic distance (F_{ST} and G_{ST}) were calculated for each of the MSY and mtDNA.

In each run, populations A and B with population size *s* started by exchanging migrants (males and females) at the rate of *m,* based on their practise of residence p_i. They evolved until generation *n*, where the trajectory of populations A and B split (as described above). In the first trajectory, A and B retained their original residence form, while in the second trajectory, populations A and B, now referred to as A' and B', changed their residence pattern according to the model simulated. In both trajectories, the populations continued to evolve until 700 generations. At each generation, summary statistics (mean number of pairwise differences, expected heterozygosity, etc.) that measure genetic variability were computed by randomly sampling 100 individuals from the evolving population. In the same manner, the genetic distance measures F_{ST} and G_{ST} (between the two populations in a trajectory) were calculated. The genetic structure of A to A' and B to B' was compared to address the following questions.

- Was there a change in the genetic structure of A and B after the switch in residence pattern? If so, what was the direction of change? This was to help test and validate the expected genetic pattern correlating to residence.
- How long after the switch in residence patterns did the trajectories become significantly different from each other? i.e., at what generation after the change was the delta of the summary statistics significantly different from zero. This was to help us determine if there was a lag in the appearance of genetic patterns corresponding to residence after a change in the latter and to also determine the persistence of a particular genetic pattern.

When there was a switch in residence from matrilocality to patrilocality, we expected the heterozygosity measures of mtDNA to increase. This was because, after a change in patrilocality, the rate of female migration into the society increases. For MSY, we expected a decrease in variation, as the number of male migrants would decrease. In this scenario, the distance between the mtDNA of A' and B' was expected to decrease as the number of female migrants after a change to patrilocality was increased, whereas we expected the genetic distance between the MSY of A' and B' to decrease as in the new residence form, males are not exchanged. The expected direction of change in the genetic structure for each

possible combination of change in residence patterns is given in Table 2-1Table 2-1.

Table 2-1: Predicted direction of change of population genetic summary statistics for sex-linked markers after a change in post-marital residence.

Change ↓	**Statistic→**	mtDNA		MSY	
		Heterozygosity	Distance	Heterozygosity	Distance
Ambilocal to Matrilocal		Decrease	Increase	No change	No
Ambilocal to Patrilocal		No change	No	Decrease	Increase
Matrilocal to Ambilocal		Increase	Decrease	No change	No
Matrilocal to Patrilocal		Increase	Decrease	Decrease	Increase
Patrilocal to Ambilocal		No change	No	Increase	Decrease
Patrilocal to Matrilocal		Decrease	Increase	Increase	Decrease

The comparison of the two trajectories of evolution, i.e., one trajectory where the populations underwent a change in residence pattern and the other where they retained their original state of residence practise, was done by estimating the window where the summary statistics became significantly different from zero, after the two trajectories diverged. These were calculated for 5, 10 and 25-generation windows by using first fitting a local polynomial regression function, and calculating delta for each 5, 10 and 25 generation windows using the loess function in R (Cleveland *et al*, 1992). A t-test was performed within each window to see if the delta for each statistic was significantly different from zero (excluding statistics that were not meant to change, see Table 2-1). A high correlation (r=0.95) was found between F_{ST} and G_{ST}, and so, only G_{ST} was retained for further analyses. We then used Fisher's method for combining p-values (Quinn and Keough, 2002) and tested whether the combined p-value of all the statistics was significant. The first window where the delta becomes significantly different from zero for all the relevant statistics combined, for each of the 5 runs, was noted. The median, mean and the variance of the windows were then calculated for the 5 runs.

2.4.2 Genetic structure of Oceanic societies

Genetic data on SSM markers for nine societies in near and Remote Oceania (Figure 2-1) were collected from already published studies. The genetic data collated for this study was from a number of different sources (Table 2-2) and the sampling schemes were not designed with questions like ours in mind. For this study, questions regarding the history of a population were being addressed and

therefore, it was important not to have related individuals in the sample set. In a meta-analysis where we collate information from different sources, it would have been difficult to control for such conditions. But the studies from which these samples were taken were all used to understand population histories, and therefore, unrelated individuals were particularly sampled and hence this particular issue did not pose a problem in our data set. Another problem is the spread of representation of populations from the study area. Missing information i.e., lack of information for a particular region or marker, is likely to bias results. But by limiting the region to Polynesia and Near Oceania and, excluding Micronesia as we did not have a good representation of both the SSM markers, we were able to draw our inferences more accurately.

2.4.2.1 Genetic Data of SSM markers

2.4.2.2

mtDNA: We examined a 365 bp region of the hypervariable segment-I region (HVS-I) of the mtDNA (Table 2-1) for 377 individuals from nine societies. The HVS-I segment of mtDNA is commonly used in studies to trace population history, as it is polymorphic across populations. The mtDNA sequences were aligned using MEGA 4.0 software (Tamura *et al*, 2007). Samples with excessive missing data (>10%) were deleted from the analyses. For all the sequences, bases 16180-16183 were deleted because for more than 70% of the samples these bases could not be established with confidence by Sanger sequencing.

MSY: We examined the genetic information from 7 linked microsatellites (also known as short tandem repeats or STRs) for 643 individuals from nine societies. These 7 STRs (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393) were from the male-specific region of the Y chromosome (MSY), and hence represented only the male lineage of the population.

2.4.2.3 Coding of post-marital residence

The post-marital residence patterns of the nine societies were coded using information from the *Ethnographic Atlas* (Murdock, 1967) and followed the coding scheme used in the work of Jordan and colleagues (Fortunato and Jordan, 2010; Jordan *et al*, 2009). Uxorilocal populations were coded as matrilocal and virilocal populations as patrilocal. The reason for this clubbing of traits is that the implications for genetics are the same, i.e., Murdock (1967) refers to groups where the post-marital residence is established with or near the patrilineal kin as "patrilocal", if the patrikin are aggregated in patrilocal and patrilineal groups. If groups are not aggregated, they are termed as "virilocal". However, in virilocal communities the implications for genetics is the same as patrilocal, as the bride

moves to reside with the groom. Similarly, for matrilocal and uxorilocal communities, the implications for the movement of the groom to reside with the bride's family persists and genetically this means the same movement pattern of genes in both these communities, irrespective of whether the groups are aggregated matrilineally or not.

Table 2-2: Details of the geographic, post-marital residence and source of the populations sampled for SSM genetic data

S.no	Population	Geographical classification	Post-marital residence practice	No. of mtDNA samples	No. of MSY samples	Source References
1	Bereina	Near	Patrilocal	31	35	(Kayser *et al*, 2006a)
2	Vanuatu	Remote	Unknown*	40	44	Cox (2007); Lum *et al* (1998); Pierson *et al* (2006)
3	Fiji	Remote	Patrilocal	45	101	(Kayser *et al*, 2006a); Whyte *et al* (2005)
4	Tuvalu	Remote	Patrilocal	59	100	(Kayser *et al*, 2006a)
5	Tonga	Remote	Patrilocal	41	146	Cox and Lahr (2006); (Kayser *et al*, 2006a); Whyte *et al* (2005)
6	Ontong Java	Near	Matrilocal	32	32	Delfin *et al* (2012)
7	Guadalcanal	Near	Matrilocal	58	56	Delfin *et al* (2012)
8	Lavukaleve	Near	Matrilocal	26	34	Delfin *et al* (2012)
9	Trobriand	Near	Matrilocal	45	95	(Kayser *et al*, 2006a); Lum *et al* (1998); Pierson *et al* (2006)

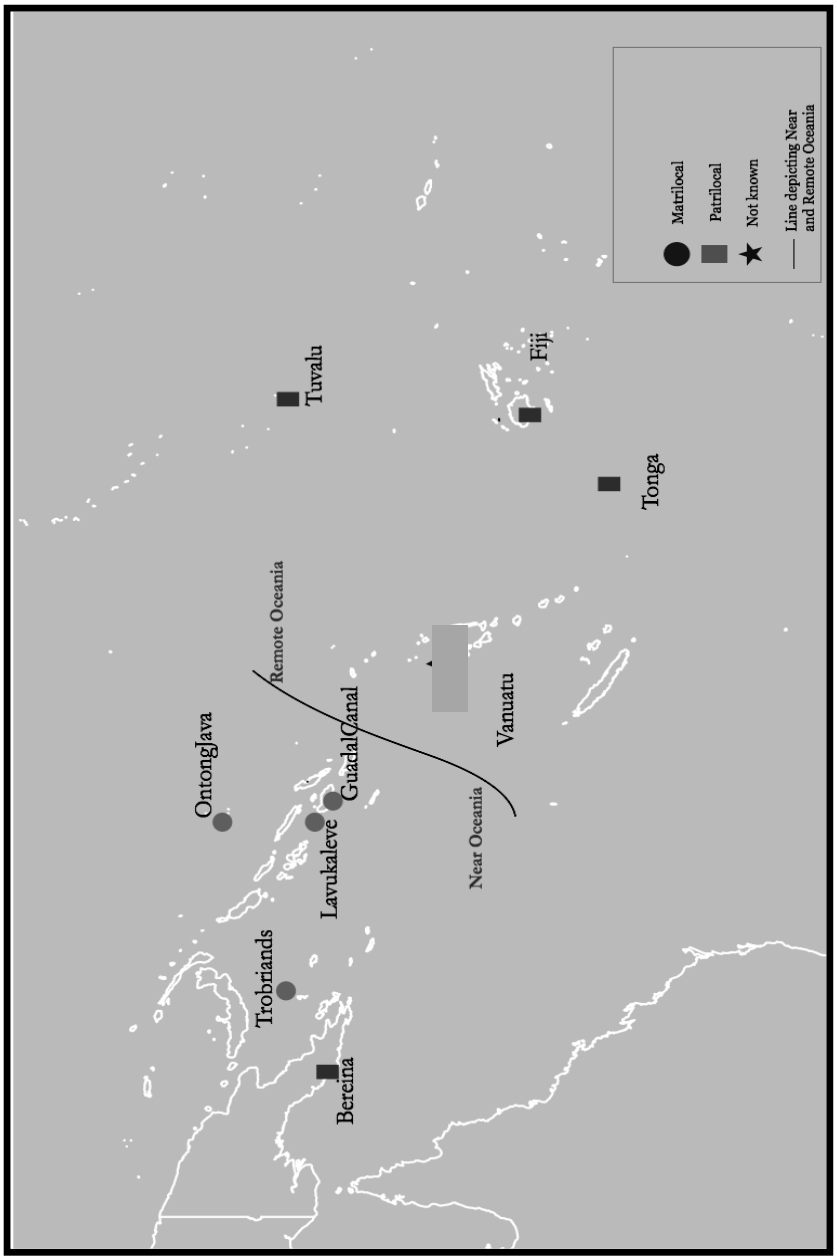


Figure 2–1: Map of the study area, with details of the matrilocal and patrilocal populations sampled for SSM data. The map also depicts the regions of Near and Remote Oceania

2.4.2.4 Methods

2.4.2.4.1 Indicators of demographic parameters

To estimate the population demographic processes, like trajectory of the population size, and whether there was evidence of sub-structure within the populations, Tajima's D and Fu's F indices were calculated using Arlequin software (Excoffier *et al*, 2005). These indices are indicators of population size change and selection. Tajima's D is calculated based on the number of segregating sites in a sample and on mean number of pairwise differences between haplotypes. Fu's Fs is based on the probability of observing alleles in a sample of given size on the observed average number of pairwise differences. This index has been shown to be sensitive to departure from population equilibrium, especially in cases of population expansion. A negative Tajima's D and Fu's F indices are indicators for either demographic growth or balancing selection. In human populations, negative values of these two indicators are attributed to low population sub-structure and a history of very rapid population growth, rather than widespread balancing selection (Tishkoff and Verrelli, 2003).

Mismatch distribution analyses, which are widely used to estimate demographic parameters of expansion or contraction, were also calculated. Mismatch distributions are the number of observed differences between pairs of haplotypes. If a population were in demographic equilibrium, we would find a multimodal distribution of the mismatches, whereas a unimodal distribution would result in populations which have undergone a recent expansion (Rogers and Harpending, 1992) or high levels of migration (Ray *et al*, 2003). This is measured through simulating data and comparing the sum of square deviations (SSD) between the observed and the expected mismatch as a test statistic. A raggedness index of the observed mismatch distribution is also computed. The raggedness index is indicative of whether the distribution is closer to a unimodal or multimodal distribution. For populations undergoing a recent expansion (unimodal), one would expect a lower value for the raggedness index than for when a population is in equilibrium. Simulated data is tested against observed data with the null hypothesis of demographic expansion. For example, from a combination of all these statistics, if a population was undergoing expansion, we would expect a negative Tajima's D and Fu's F indices, a significant SSD and a non-significant raggedness index.

2.4.2.4.2 Characterising genetic variation of SSM markers

The genetic structure of populations was estimated by mean pairwise difference of mtDNA and MSY, in each population, using Arlequin software (Excoffier *et al*, 2005). Structure was also estimated by partitioning a population's genetic

variation into its inherent variation and variation introduced by gene flow from other populations, using a Bayesian approach implemented in the software package BAPS (Corander and Marttinen, 2006; Tang *et al*, 2009).

By quantifying the amount of admixture in each population, the amount of inherent variation and the variation introduced by gene flow for mtDNA and MSY was quantified. This helped establish and compare the direction of gene flow for males and females in patrilocal and matrilocal societies. In a patrilocal society, the amount of admixture due to gene flow was expected to be less in MSY when compared to a matrilocal society, while the mtDNA gene flow was expected to be much greater in patrilocal societies. In a similar manner, in matrilocal societies we expected the mtDNA diversity present to be due to its own inherent variation and not due to gene-flow, and the MSY diversity to be largely explained through gene-flow.

2.5 Results

2.5.1 Forward simulations

2.5.1.1 Results

Forward simulations were used to test if there was a detectable pattern in the SSM of a population corresponding to its residence practice. After a switch in the residence practise of a population, it was tested if the summary statistics change in the direction predicted by the change in post-marital residence. If the change is in the direction expected, how long after the switch in post-marital residence is the change reflected in genetic data? The corresponding effect of migration rate, population demographic model (decline, growth and constant size) on the SSM pattern were also tested. There were a total of 180 models, and each of these was run with 5 replicates, giving rise to a total of 540 models. These 540 models were grouped to 5, 10 and 25-generation windows for further analysis (Table 2-3). For example, a 5-generation window model meant that 5 generations of evolution of one trajectory and corresponding 5 generations of the second trajectory were taken as one unit and were compared.

The first test was whether a change in residence showed a corresponding change in SSM marker pattern. This test also revealed whether marker patterns could be predicted based on the residence practised by a society. Only in growing populations did the change in summary statistics correlate with the SSM pattern change predicted by the post-marital residence. Also, in growing populations, this pattern remained consistent through the course of evolution and did not fluctuate. In contrast, in declining and constant size populations, the pattern was not consistent. Even when the summary statistics trended in the direction expected, after a change in residence, they were not consistent through the course of

evolution, and did not reflect the post-marital residence of the society. This was tested by observing the trend of delta values calculated between the two trajectories until the end of simulations. Delta values compare the difference between the relevant summary statistics of the two trajectories, so we used these to test if the values remained significantly different from each other and in the direction of expected change.

In growing populations, the direction of change of the summary statistic was as expected with the type of residence practise modelled. For example, when a population shifted from matrilocal to patrilocal, the heterozygosity of the mtDNA increased, and correspondingly the MSY diversity decreased. For constant and declining populations, it was difficult to quantify the trend with confidence, as the delta value fluctuated between significance and non-significance (refer to median of variance in Table 2-3). Migration rate did not have an effect on the delta value fluctuations.

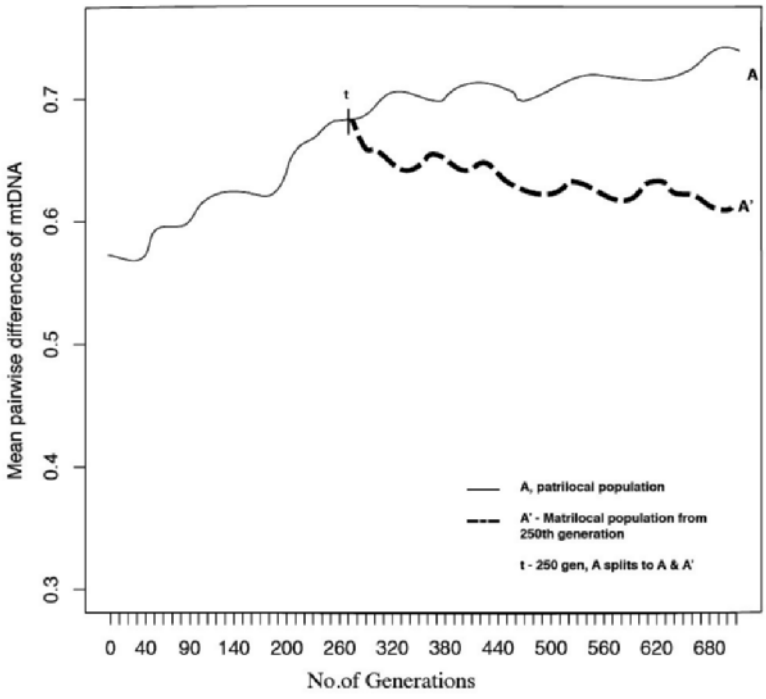


Figure 2–2: Figure 2–2: A smoothed curve (after fitting Loess function) representing the changes in the expected heterozygosity of MSY in a declining population A and A' during the course of evolution. A' (dashed line) has undergone a change from patrilocality to matrilocality at 250 generations, while A (solid) has remained patrilocal

Demography, especially with respect to population size change, emerged as the single most important factor in determining how populations responded to a change in residence pattern and whether this response was detectable and/or predictable by observing SSM variation. For example, in a growing population (Figure 2-2), A underwent a change from patrilocality to matrilocality after 250 generations of evolution, while A' continued its evolution in matrilocality. Here, the change of mtDNA diversity of A was clearly seen after 250 generations, and in the expected direction of change (decrease in mtDNA diversity of A) and this difference between the trajectory A and A' remained stable, and significant for the remainder of its evolution, devoid of any fluctuation. This can also be seen in the range and variation of the median of the first window, where the delta values between the two trajectories of evolution of a population become significant.
In a growing population, the variation was relatively low, and the median stayed consistent across runs. For example in Table 2-3, if we look at the 10-generation window, the growing population has the least variance (0.2) among the range of demographic models, for example, a declining population has a variance of 1.35, almost 10 times more than a growing population.

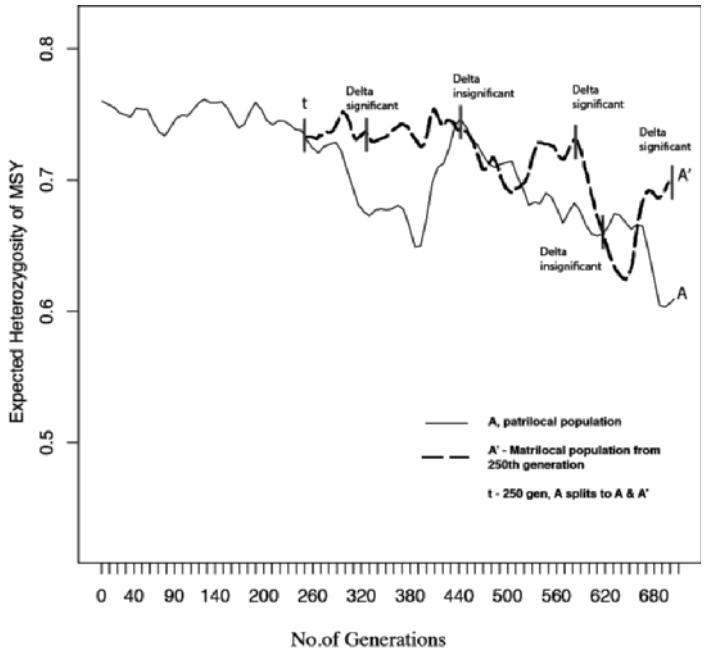


Figure 2–3: A smoothed curve (after fitting Loess function) representing the changes in the expected heterozygosity of MSY in a declining population A and A' during the course of evolution. A' (dashed line) has undergone a change from patrilocality to matrilocality at 250 generations, while A (solid) has remained patrilocal

The range and variation of the median was very high for declining populations and to a large extent in constant size populations as well. In declining populations, for example in Figure 2–3, patrilocal population A at the 250th generation, underwent a change from patrilocality to matrilocality. It continued its evolution in matrilocality as A'. According to the pattern predicted by post-marital residence, we expect to see an increase in MSY expected heterozygosity in A' as compared to A (i.e., after shifting to matrilocality, there is an influx of MSY diversity). Initially, after a change into matrilocality, an increase in MSY heterozygosity is seen as expected, but as time progresses, it is found that this shift is not stable. The delta values fluctuate between significance and non-significance through the course of evolution of the trajectories, till the end of evolution. This variance is also reflected in the range of the medians and the variance of the medians of the windows, where the delta statistic is first significant, and continues in this state of significance till the end of evolution. The model describe above in Figure 2-3 is representative of the majority (>90%) of the runs in the simulations, where the population was modelled to decline. There were some instances, where the models follow the predicted trajectory, where there is no fluctuation between significance and non-significance, and this could be accounted to chance.

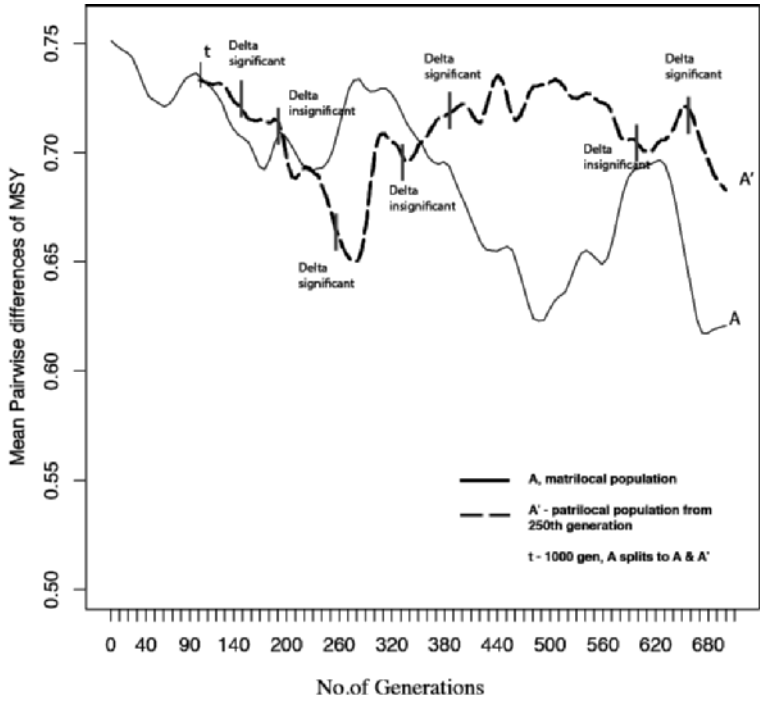


Figure 2–4: A smoothed curve (after fitting Loess function) representing the changes in the mean pairwise difference of MSY in a constant sized population A and A' during the course of evolution. A' (dashed line) has undergone a change from matrilocality to patrilocality at 250 generations, while A' (solid) has remained matrilocal

In constant sized population models, as is the case with declining population models, the SSM marker diversity does not consistently correlate with the corresponding post-marital residence practice. In Figure 2-4, a matrilocal population of a constant size, shifts from matrilocality to patrilocality and continues its evolution as A'. A decline in the mean pairwise difference of MSY marker is expected, as the number of incoming males into the population A' decreased, as compared to A. However, the delta values, i.e., the mean pairwise difference between A and A', fluctuated between significance and non-significance. The trend of the mean pairwise differences also fluctuated between increase and decrease and did not stay consisted till the end of the simulations.

Table 2-3: Range of median, median, and median of the variance in delta values across 5, 10 and 25-generation windows of the first instance where summary statistics become significantly different for the two trajectories of evolution in each population model.

Population Window - Model	Range of Median	Median	Median of Variance
5 Gen window - all	1-41	14	0.8
5 Gen – Constant	1-33	18	0.3
5 Gen – Decline	3-39	32	8.65
5 Gen – Growing	1-41	2	0.6
10 Gen window - all	1-31	11.5	0.3
10 Gen - Constant	1-17	9	0.2
10 Gen - Decline	2-21	17	1.35
10 Gen - Growing	1-31	1.5	0.2
25 Gen window - all	1-33	1	0
25 Gen - Constant	1-1	1	0
25 Gen - Decline	1-4	2	0.3
25 Gen - Growing	1-3	1	0

The difference in the delta values of the two trajectories is almost always non-significant in the first 5-generation window, in all the models. In only 0.01% of the runs, a significant delta between the two trajectories in the first 5 generations after the switch was observed, which could be accounted to chance. When the 10-generation window was analysed, there was low variance in the median and the delta became significant before the second window, i.e., 20 generations after the switch. In the 25-generation windows, the deltas became significant and the two trajectories were significantly different from each other in the first window of 25 generations after the change in residence. It can be concluded that, when a change in residence occurred in a growing population, the change is not immediately reflected in genes, but it takes between 5 and 25 generations of evolution before a predictable change in genes is observed.

The comparison between growing and constant size populations revealed that the median of the range of a constant size population was much higher across the 5, 10 and 25 generation windows (18, 9 and 1), when compared to the growing population (2, 1.5 and 1). This meant that across runs, a growing population consistently resulted in similar changes, while there were a lot of fluctuations in the constant size populations.

To test the hypothesis of a time-lag from the time a switch in residence occurred to its reflection in genes, the window (Table 2-3) at which the delta for relevant statistics between the two trajectories became significant was determined. The correlation between mean and median was 0.99 ($p<0.001$) and since the mean is easily influenced by extreme values in the range, the median values are reported. As there are several models for which this statistic is calculated, the median of the first window between the two trajectories for each model where the delta values became significant was calculated. The range, the median of the range and the variance of the medians across all the models for each 5, 10 and 25-generation windows was then calculated. The first signs of change in the genetic pattern correlated with the post-marital residence depended heavily on the type of population size model and the migration rate. There was a negative correlation between the migration rate and the time when there is a significant change in the genetic pattern ($r = -0.32$), i.e., as the migration rate increased, the patterns became pronounced and strong more quickly and therefore, the difference between the two trajectories became significant much earlier. Apart from migration rate, population size played a crucial role in determining when the change of genetic pattern corresponded with a change in social practices, and whether this change was stable and predictable.

2.5.2 Genetic Structure of Oceanic societies

2.5.2.1 Results

2.5.2.1.1 Indicators of demographic parameters

Tajima's D and Fu's F indices were calculated to understand the demographic processes that could influence the genetic structure of populations (Table 2-4). Results showed a highly negative Tajima's D and Fu's F for majority of the populations, which indicates population growth. While the results from Tajima's D index were significant for our populations of interest, they did not significantly indicate a population expansion with Fu's Fs index. Fu's Fs is not sensitive to recent demographic expansion and the lack of statistical evidence to demographic expansion could be attributed to the low power of the test to detect recent events and can be ignored in favour of a highly negative and significant Tajima's D.

Mismatch distribution analyses for demographic expansion showed non-significant SSD and raggedness indices for Fiji and Tuvalu, while it was significant for Tonga. Non-significant values of SSD and raggedness index for the mismatch distribution are a result of the model not being able to dismiss the null-hypothesis of demographic expansion.

Table 2-4 mtDNA Indicators of Demographic expansion.
*SSD = Sum of Squared deviations. * = p<0.05.*

Society	Geographical classification	SSD	Raggedness index	Tajima's D	Fu's F
Trobriand	Near Oceania	0.059	0.196*	-1.175	-0.935
Solomons - Lavukaleve	Near Oceania	0.013	0.037	-1.481	-0.108
Solomons - Ontong Java	Near Oceania	0.442*	0.223	0.717	3.901
Guadal Canal	Remote Oceania	0.040*	0.080	-0.345	-2.010
Fiji	Remote Oceania	0.043	0.041	-1.001*	-1.822
Tonga	Remote Oceania	0.159*	0.512*	-1.498*	-2.986
Tuvalu	Remote Oceania	0.060	0.212	-1.498*	-1.741
Vanuatu	Remote Oceania	0.0315	0.045	-1.498*	-1.512
Bereina	Near Oceania	0.0273	0.032	0.001	-0.175

2.5.2.1.2 Characterising genetic variation of SSM markers

Genetic structure was assessed by comparing mean pairwise difference of SSM (Figure 2–5) and by estimating admixture through gene flow using BAPS (Table 2-5). The Near Oceanic results revealed that the genetic patterns observed in these populations (Bereina, Trobriand, Lavukaleve, Ontong Java) correspond to their currently practised post-marital residence pattern.

In the patrilocal population of Bereina, the mean pairwise difference of the mtDNA was higher than that of MSY and the majority of mtDNA variation was introduced from other populations through gene flow, as expected. The MSY pattern in Bereina was inherent and ethnographically attested, and both the mtDNA and MSY pattern corresponds to the pattern expected from the practiced

post-marital residence. Trobriand and Ontong Java reflected their highly matrilocal patterns, with 100% of their mtDNA diversity stemming from local cultural processes with no variation introduced through gene flow (Table 2-5). For other matrilocal populations in Near Oceania, Lavukaleve and Gudadalcanal, the MSY and mtDNA diversity matched the expected pattern with MSY diversity being more than mtDNA (Figure 2–5)

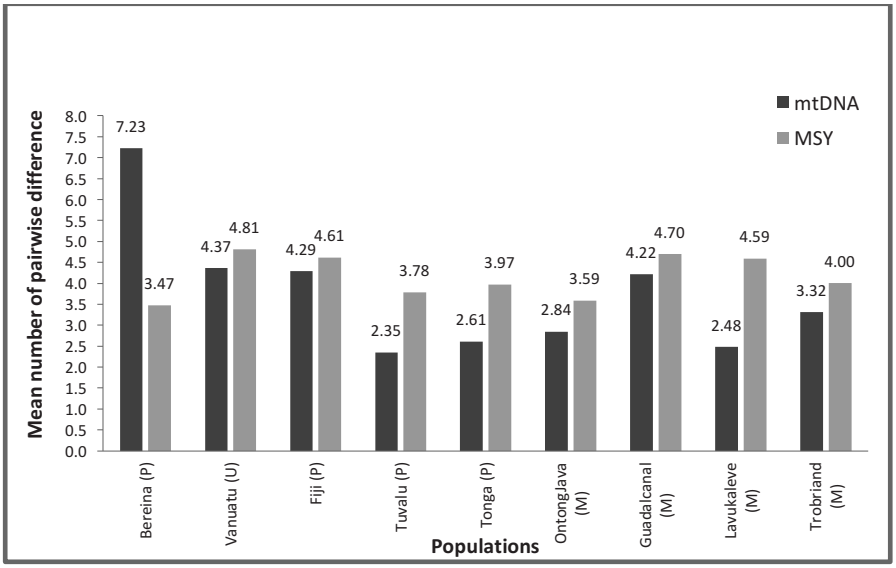


Figure 2–5 Mean pairwise difference of mtDNA and MSY. Bereina, Fiji, Tuvalu & Tonga are patrilocal populations, while OntongJava, Guadalcanal, Lavukaleve and Trobriand are matrilocal populations. The post-marital residence practice of the Vanuatu population was unknown. All differences are significant.

However, in three of the four Remote Oceanic populations (Fiji, Tuvalu & Tonga), the genetic patterns observed from the results of mean pairwise differences and estimation through admixture and gene flow indicate matrilocality, while these populations currently practise patrilocality. Fiji, Tonga, and Tuvalu showed more MSY diversity than mtDNA, a pattern generally taken to be indicative of matrilocal post-marital residence (Table 2-5). The diversity of mtDNA in Fiji, Tonga and Tuvalu stemmed from gene flow and admixture (96%, 93% and 97% respectively), rather than it being inherent as expected by patrilocal societies. The post-marital residence practice of the populations in Vanuatu was not known; however, the genetic pattern revealed a structure that indicated a patrilocal society.

Table 2-5: The proportion (in percentage) of genetic diversity that exists inherently (own) and the proportion (in percentage) derived from gene-flow (others) for mtDNA and MSY.

Society	Residence	mtDNA		MSY	
		Own	Others	Own	Others
Bereina	Patrilocal	48	52	91	9
Trobriands	Matrilocal	100	0	92	8
OntongJava	Matrilocal	100	0	78	22
Lavukaleve	Matrilocal	92	8	88	12
Guadalcanal	Matrilocal	98	2	96	4
Vanuatu	Unknown	19	81	95	5
Fiji	Patrilocal	96	4	74	26
Tonga	Patrilocal	92	8	72	28
Tuvalu	Patrilocal	97	3	87	13

2.6 Discussion

2.6.1 Forward Simulations

Forward simulations were employed to test whether there was a detectable change in the genetic signature of a population following a change in post-marital residence practise. Tests for evidence of a lag from the time of change in residence pattern, to its reflection in genes were also conducted. Only in the absence of drift could post-marital residence patterns influence the genetic structure of SSM genes in the expected pattern. In constant size and exponentially declining populations, as expected, we found that the effect of drift was much higher than in exponentially growing populations (Caballero, 1994). A change in the genetic pattern of the SSM related to the residence pattern was completely masked by this effect of drift, in populations with low or decreasing effective sample size. In these populations, the genetic pattern was not stable and fluctuated from one generation to another and across different replicates, with no discernible pattern.

Drift causes the loss of random alleles and in turn, loss of diversity, without a set pattern. The effect of this stochastic process is quite strong in populations of small effective sample size and those that have undergone recent decline due to bottlenecks or founder effects, and hence is important in the context of the Pacific. The Remote Oceanic populations have been recently colonised (between ~ 3500 and 770 YBP (See Green, 2003; Kirch and Green, 2001) and have undergone severe bottlenecks and founder effects, accompanied by population decline, during the process of their settling the Polynesian islands (Gray *et al*, 2009; Ramakrishnan *et al*, 2005). In this scenario, it is imperative to understand the extent of the effect of drift on genetic pattern, as it impacts the inferences that have

been made on the social systems in these societies, based mainly on evidence from sex-biased genetic patterns (Hage and Marck, 2003; Kayser *et al*, 2006). Based on these results, simulations demonstrate that it is important to understand the demographic and evolutionary processes acting on the populations before making inferences on the social and cultural processes of the pacific populations based on SSM patterns. Recent studies have also shown that apart from cultural processes, the effective sizes of males and females differs greatly and has a lasting, significant impact on sex specific markers (Ségurel *et al*, 2008; Underhill and Kivisild, 2007).

In models where the effect of drift could be discounted, i.e., the growing population models, genetic variation of the population corresponded with the practised post-marital residence norm and changes in the practise resulted in changes in the SSM genetic patterns in a predictable fashion. This was true for all demographic models tested in our study. The pattern of differentiation between types of post-marital residence patterns was clear, strong and consistent across different runs, indicating a tractable model of correlation between genes and culture when drift is absent.

Secondly, we explored the possibility of a time lag in the change of genetic diversity pattern following a change in residence pattern. This was tested by quantifying the number of generations after which the switch in residence resulted in a corresponding switch of genetic patterns and this change remained stable. An evidence for this time lag appeared in all the demographic scenarios we tested. The duration of this time lag varied among the different demographic models and between runs, but the variation was not high (see Table 2-3), indicating a constant trend. On average, the time lag existed for at least 5 generations, but not beyond 25 generations (unless in extreme cases of drift, where the relationship between residence and SSM patterns was not tractable). The quantification is an important step in understanding the dynamics of gene culture co-evolution in societies where the SSM show neither the expected pattern of change nor do they show homogeneity in the SSM that could be attributed to a problem with the scale of sampling.

Forward simulations provided us with a unique insight into the mechanistic processes underlying evolution of sex-specific genetic patterns. This study tested if we could make inferences regarding the social processes occurring in a population by observing genes. It was found that while the social practices of a society could strongly influence the sex-specific genetic pattern in an anticipated manner, the model did not hold good for populations affected by drift. A population under the force of drift (reduction in population size) did not show any coherent genetic correlation with residence, and drift had the ability to mask any other process occurring in the population. Conclusive evidence regarding the presence of a time lag (5-25 generations) for a population undergoing change in their residence to “recover” from the effects of their former residence practise was evident.

2.6.2 Genetic Structure of Oceanic Societies

The genetic diversity and composition of Near and Remote Oceanic societies with known post-marital residence was quantified. The genetic patterns of the SSM in the Near Oceanic populations were in line with expectations from previous studies. In the Near Oceanic populations, it was observed that the mtDNA structure of the matrilocal Near Oceanic populations (Trobriand, Lavukaleve, Guadalcanal, Ontong Java) was entirely due to inherent variation i.e., not due to gene flow and with very low diversity, whereas the MSY structure was shaped by admixture and gene flow from other populations, and showed a higher diversity than mtDNA. Similarly, the patrilocal Bereina showed a lower MSY diversity compared to mtDNA that was shaped by gene flow and admixture.

In Remote Oceania, ethnographically-recorded populations on Fiji, Tuvalu and Tonga currently practise patrilocality. Contrary to expectations, their SSM patterns resemble populations practising matrilocality. The mtDNA diversity is significantly lower than MSY and the mtDNA structure is shaped more by gene flow and admixture, while the MSY structure is more inherent and has very little gene flow. This contrasting patterns could be a result of three plausible causes a) scale of observation, which also includes variation between ethnographic observations and sociocultural reality b) a recent shift in the post-marital residence ("time-lag hypothesis), or c) the effect of demographic or stochastic evolutionary forces. We explore each of these reasons in the context of Remote Oceanic populations below.

A plausible explanation for the disparity in residence pattern and genetic patterns could be due to the scale of observation. If patrilocality were practised endogamously and/or at a local scale, the resulting pattern in SSM would be undetectable with population level analyses (Kumar *et al*, 2006; Wilder *et al*, 2004), and we should not see any difference between the mtDNA and MSY genetic diversity levels, i.e., the levels of diversity of both the markers should be close to equal. Instead, in Fiji, Tonga and Tuvalu, , we see a strong genetic pattern reflecting a matrilocal society. Therefore, we can convincingly deduce that scale/endogamy is not an issue for this analysis, and there are other mechanisms at play.

The inconsistencies observed between the social patterns and corresponding genetic patterns could also be attributed to a very recent shift in residence norms. According to the time-lag hypothesis, a society has recently shifted from matrilocality to patrilocality in these societies, and consequently the populations are still in the time-lag phase and the genetic pattern has not yet caught up to reflect the current residence practices. Evidence from comparative methods have shown that proto-Austronesian societies could have a long history of matrilocality (Jordan *et al*, 2009). Matrilocality is often linked to matrilineal

descent systems (Hage and Marck, 2003; Harris, 1985; Marck, 2008) and might be favoured in societies with long male absence (due to war, trade, or resource exploitation such as deep sea fishing), like in most early Oceanic societies.

The historical records of Fiji do not date very far back in time, and therefore it is challenging to determine the ancestral form of residence or descent practised concretely. Some of the older records depict a society with considerable diversity in social organization, with evidence for matrilineal, matrilocal as well as patrilineal and patrilocal practises on the island (Geraghty, 1983; Geraghty, 1996; Thomson *et al*, 1908). But according to Marck (2008), only the Maola of the Fiji were ancestrally patrilineal and this was an isolate. And that the Fijians, like the other Austronesian societies were ancestrally matrilineal (and potentially matrilocal). However, matrilineal forms tend to be unstable (Richards, 1950) and matrilocality often evolves towards other forms of residence (Fortunato and Jordan, 2010). Immediately following their occupation, the Polynesian islands may have offered little reasons for men to continue their long absences from home. The competition for resources was comparatively low, as the islands were mostly isolated from each other and these conditions did not favour the continuity of matrilineal descent or matrilocal residence (Hage and Marck, 2003). Therefore, the instability of matrilineality and matrilocality, coupled with the changing social environment, most likely encouraged the society into gradual transition to patrilineal and patrilocal forms. The switch in social systems was further fuelled by a possible wave of migration of the non-Austronesian Melanesians into Vanuatu, Fiji and the rest of western Polynesia (Posth *et al,* 2018), once they refined their maritime skills. The multitude of social systems recorded in Fiji is evidence for this transition phase, where several societies recently switched to the patrilocal social systems, while others were in the process of switching from their original state of matrilocality (as seen in the multimodal distributions of the mismatch analysis). The current state of patrilocality (Murdock, 1967) is most likely then a reflection of this recently acquired change and could well be an explanation for the genetic pattern still reflecting matrilocality.

Based on the knowledge and analysis of kinship terms, it was established that there was a more than likely chance of Tonga being matrilineal and matrilocal about 2500 years ago (Marck and Bostoen, 2010) but they currently practice ambilineality (Murdock, 1967). This current state of ambilineality with an ancestral state of matrilineality, and is concurrent with predicted model of change of a society from matrilineality to patrilineality, with an intermediate ambilineal state (Opie *et al,* 2014). Davis (1984) proposed a model of change for the unstable matrilineal descent system. Matrilineal forms first change to ambilineal and then to patrilineal with a corresponding change in residence from matrilocality to patrilocality along with the change in descent. In Fiji, a plausible explanation for the incongruence between the genetic pattern and the residence practise in Tonga could also be attributed to a time-lag in the change of genetic patterns to reflect

the recent change in residence norm. This is also reflected in the mismatch distribution test, where there was a signal for the complex processes at play given the multitude of signals from the Tongan genetic data. Based on the history of residence norms practised in these societies, it seems likely that the time-lag hypothesis is good candidate to explain the contrasting genetic and social patterns found in our analyses.

The reason for the apparent disparity could also be a strong effect of stochastic evolutionary forces such as drift, and/or demographic forces such as bottlenecks, both of which can overshadow the effect of residence on SSM. Remote Oceania consists entirely of island populations, colonized as recently as ~770 YBP (Green, 2003; Harris, 1985; Kirch and Green, 2001). The effect of drift, founder effects and bottlenecks are substantially high for island populations (Ramachandran *et al*, 2005) and these processes could essentially mask the signal of post-marital residence on genetic diversity. Tajima's D and Fu's Fs both indicate a recent demographic expansion in Tonga, Tuvalu and Fiji. While we could confidently conclude that Fiji and Tuvalu were undergoing a recent demographic expansion through mismatch distribution tests, Tonga seemed to be experiencing a more complex process that needs to be examined by looking at its history. The lack of a population decline or evidence of bottleneck suggested that there was a need to look elsewhere to understand why the SSM patterns of these populations did not conform to the expected pattern. A further test to understand the demographic pressures in these populations would be by means of tests like coalescent analyses, or Bayesian analysis to infer past population dynamics (Drummond *et al*, 2005). While demographic processes are known to affect genetic diversity, we do not understand the mechanism, magnitude or direction of these effects. It is prudent to investigate these effects before concluding their influence on SSM marker diversity.

Both groups of scenarios (stochastic evolutionary processes and the time-lag scenario) seem equally plausible. Due to the complexity of processes that act and interact in a natural population, it is difficult to predict and understand cause-effect relationships. In such cases, computer simulations, especially forward simulations, are excellent tools to understand the mechanism behind such complex systems. Eliciting information regarding the interaction between two or more parameters acting on a population would be intractable in a realistic setting. Forward simulations provide us with a unique environment, wherein we could test and observe the effects of evolutionary parameters of our interest, while keeping other forces that act on a real population constant (Calafell *et al*, 2001) and this would help predict outcomes under different scenarios.

2.6.3 Combined discussion

For the Remote Oceanic populations in the current dataset that did not follow the expected pattern of change–Fiji, Tonga and Tuvalu–tests for demographic expansion were positive. This indicates that even though the communities on these islands previously underwent a series of bottlenecks and founder effects, they were now in the process of a demographic expansion, indicating population growth. It was also observed that the effect of drift on growing populations is negligible, and therefore the deviation from the genetic pattern expected of these populations could be explained by a recent change in residence. If we consider that three populations from Remote Oceania were in fact in the time lag phase, that would mean that there should be evidence for a change in residence between 5-25 generations before the present. When we examine the events that occurred during that approximate time scale in the region, around AD 1300 and AD 1800, there was a rapid cooling of temperatures (known as the little ice age). While there have been several debates and theories about what this cold period meant for the communities in Remote Oceania (Anderson, 2002; Field and Lape, 2010), archaeological evidence shows that there was substantial resource depletion around 2100 YBP, even though colonization began about 2900 YBP (Anderson and Clark, 1999). The little ice age provided an opportunity for the development of intensive agriculture in the region, to compensate for the already rapidly declining resources, and thereby resulted in an increasing human density. The rise in sea levels during this period would have rendered long-distance sea voyaging for subsistence unprofitable, and societies would have had to resort to other modes of subsistence. In accordance with this, there is evidence of movement of people from the coast to inland during this time (Anderson and Clark, 1999) and change of subsistence to intensive agriculture (Enright and Gosden, 1992). This change in subsistence also meant that there was a reduction of long male absence from home, one of the main proponents for the stability of matrilocality (Divale, 1974b; Hage and Marck, 2002; Hage and Marck, 2003; Lévi-Strauss, 1969). It is possible that matrilocality then started evolving towards its present state of patrilocality, substantiated by records of high social and cultural change during this period (Jones, 2009). Another explanation for social change was a second wave of migration of the non-Austronesians into Polynesia. Blust (2008) proposed that the migration of Austronesians from Near Oceania into Remote Oceania was closely followed by a wave of migration of Papuan speakers, after developing their sailing technologies due to interaction with Austronesians, into Vanuatu and western Polynesia. This would also explain why in Fiji, there is a mixture of different societal practices (and also the multimodal distribution as seen in the mismatch analyses). Whether the change in post-marital residence was a cultural change induced by this wave of migration is not definitive, but the hypothesis is a worthy candidate for consideration. The cause for an induced change in post-marital

residence was not clear, but the results from this study show that if not the effect of strong drift, then a change in post-marital residence may have occurred 650-750 years ago, in the populations of Fiji, Tonga and Tuvalu. For the Remote Oceanic populations (Fiji, Tonga and Tuvalu), there was ample support to argue for a shift in residence in recent enough times, as explained in 2.6.2, and the population's SSM could well be in the lag phase of change.

2.6.4 Conclusion

In the present study I found that while change in post-marital residence does result in a predictable and detectable change in patterns of mtDNA and MSY, this is only true in cases where the effects of drift do not mask local social processes. Also, the existence of a time-lag of 5-25 generations for a change in post-marital residence to reflect on the patterns of SSM marker diversity is quantified and this lag is conditioned upon drift and migration effects. Often, a change in residence is correlated with elements like introduction of agriculture, trade and change in lifestyle (Marlowe, 2004). By quantifying the duration of the time lag, not only can we track the process of change in residence, but also garner information regarding other correlated elements like introduction of agriculture, trade, change in lifestyle all of which are said to be correlated with the change in residence (Fortunato, 2011; Wilkins, 2006). This information also provides us with a distinctive opportunity to add to the history of populations, where sex-linked markers do not exhibit similar patterns and also do not correspond to the existing residence practices.

Given these results, it is imperative to progress with caution to make a robust interpretation of history from genetic information. When using genetic data to understand the history of a population, accounting for stochastic and demographic evolutionary process like drift, is the single most important step to perform before making any inference.

3 Sexual Division of Labour and Residence

3.1 Abstract

Across the world, there is variation in the norms that specify where couples may reside after marriage. Newlyweds may live with the groom's kin (patrilocality), the bride's kin (matrilocality), or have some flexibility (ambilocality). Anthropologists have claimed that "the sex that stays put after marriage is the sex that contributes most to subsistence" (Driver and Massey, 1957:425). However, worldwide cultural analyses do not support this claim (Ember and Ember, 1971). I test this co-evolutionary hypothesis in Austronesian-speaking societies, where substantial variation in residence is demonstrated, and matrilocality is predicted to be the ancestral form (Jordan *et al*, 2009). Cultural data for 109 ethnolinguistic groups represented on Austronesian lexical phylogenies, fine-grained coding of residence, sexual division of labour (SDL) and marriage norms were collected. While phylogenetic signal tests revealed a complex pattern of phylogenetic clustering and non-clustering of traits, we found no evidence for a direct co-evolution of sexual division of labour and residence traits in polygynous societies.

3.2 Introduction

3.2.1 Background

Along with migration and colonization, culturally specific marriage practices and norms of post-marital residence constitute important factors for the displacement of genes from one population to another. The practise of certain types of post-marital residence norms can result in sex-biased migration in a population and this sex-biased migration can be tracked using sex-specific markers (See Chapter 2). The movement of women can be followed using the female linked mtDNA marker, while male movement can be monitored using male specific Y-chromosome (MSY). These SSM in a population have a specific genetic diversity and genetic structure representing the population of their origin. A significant fraction of variance found in SSM markers has been attributed to the trans-generational movement of males and females due to post-marital residence norms (Seielstad *et al*, 1998; Stoneking, 1998). Previous studies have established that different kinds of post-marital residence can influence this structure and variation of SSM in a predictable manner (Besaggio *et al*, 2007; Bolnick *et al*, 2006; Gunnarsdottir *et al*, 2011; Oota *et al*, 2001). In addition, we also found quantitative proof that this predictable effect of post-marital residence on SSM is seen both in growing populations and populations where the effect of drift is not significant (for discussions related to the mechanisms and significant consequences, please

refer to Chapter 2 and Chapter 4). However, the norms themselves are variable across the world, undergo change from time to time, and are known to be influenced by several other factors like migration and depopulation events (Eggan, 1966; Murdock, 1949a), warfare/feuding (Ember and Ember, 1971; Otterbein and Otterbein, 1965), male-absence (Harris, 1980) and the sexual division of labour (Ember and Ember, 1971). It is therefore important to understand how these factors affect the dynamics of post-marital residence, given its effect on the genetic structure of a society.

Cross-cultural scholars have long thought that the sexual division of labour is associated with post-marital residence (Murdock, 1949a). It has been hypothesized that the "sex that stays put after marriage is the sex that contributes most to subsistence" (Driver and Massey, 1957; Ember and Ember, 1971). For example, societies where women contribute more to subsistence, also called "matridominant" societies, are expected to be matrilocal, and where there is mostly equal division of labour between the sexes towards subsistence activities, societies are expected to be ambi-/neolocal. Similarly, societies where men contribute more to subsistence, also called "patridominant", are expected to be patrilocal. For example, for the Aleuts, who inhabit the subarctic chain of islands bordering the Alaskan peninsula for over 8000 years (McCartney and Veltre, 1999), fishing forms their major subsistence activity and contributes to over 90% of their food source (Marlowe, 2007). The Aleuts are patridominant, where men contribute most to subsistence activities, and they are also patrilocal, following the predicted pattern of division of labour and post-marital residence (Ember and Ember, 1971).

Why should gender specific contribution to subsistence drive social organisation? Evolutionary theory predicts that, like other organisms, humans should act in ways that maximise their fitness (Mace, 2000). For example, in the Himalayas, where resources are sparse, living conditions are tough and need for labour is high, the society tends to prefer polyandry, because it is beneficial for a man to join his male kin as a secondary husband rather than being monogamous (Borgerhoff Mulder, 1991), thereby driving social organisation. In terms of maximising fitness through increase in reproductive success or prolonging survival, the driving force is meeting energy requirements and thereby subsistence.

The strategies to maximise fitness have been debated widely, in terms of the mechanisms through which humans change their life history traits, and whether individuals' or collective interests play a more prominent role (Borgerhoff Mulder, 1991; Cronk, 1991; Winterhalder and Smith, 2000). To maximise returns of energy investments in addition to safeguarding offspring provisioning, gender differences in investment towards subsistence is important. Apart from adaptive individual behaviour, social factors like cooperation, marriage patterns and post marital residence play a major role in maximising

these energy returns (Codding *et al*, 2011). For example, Wood and Marlowe (2011) examined the life history strategies associated with post-marital residence choice. They tested Hamilton's kin selection theory, which proposes that human kin cooperation is directly proportional to the degree of relatedness, when the cost to one's self is comparatively less than the benefit to the other person. It was found that energy investment (of which resource garnering for subsistence forms one aspect) and its returns are indeed the main drivers of post-marital residence choice. Schlegel and Barry III (1986) collated information on female contribution to subsistence activities of 186 non-industrial societies of the world and investigated the consequences of increased female contribution to subsistence (matridominant) on social organization. They analysed the number of hours contributed to subsistence activities based on Murdock's Ethnographic atlas (1967). They found two major categories of consequences that women's contribution to subsistence could elicit, adaptive and attitudinal. With an increase in contribution by women to subsistence, the change in attitude is predicted to result in the elevated status of women in the society, as they are perceived to be more self-sufficient and less manipulable. This change in attitude is not a direct result of the actual increase in contribution to subsistence, but is rather a change in perception from being viewed as objects for reproductive needs as compared to societies where women's contribution to subsistence activities was low (Sanday, 1974). The adaptive correlates of this increased contribution includes bride price being more likely than dowry (due to a woman's elevated status), and polygyny more likely than monogamy (Heath, 1958). A man from a patridominant society, i.e., where men contribute more to subsistence, would not consider the possibility of having many wives, as it would only increase his burden of the number of people dependent on him for subsistence. In matridominant societies, by marrying polygynously, a man increases the number of people that would share the burden towards subsistence; polygyny is therefore a very attractive option in matridominant societies (Korotayev, 2003b) and a number of cross-cultural tests have shown a positive association between matridominant societies and polygyny (Burton and Reitz, 1981; Heath, 1958; Osmond, 1965; White *et al*, 1981). It has also been hypothesised that in a society where women contribute more to subsistence, it may be more beneficial for her to reside with her kin, where she is likely to receive cooperation in activities not only related to subsistence but also in care-giving and other household activities, especially where there is a background of warfare and male absence (Driver and Massey, 1957; Korotayev, 2003a; Panter-Brick, 2002). The sexual division of labour therefore has an important role in the dynamics of social organization. While warfare and migration also play a major role in influencing post-marital residence (Divale, 1974a; Divale *et al*, 1976; Ember, 1974; Ember and Ember, 1971 for details), in this chapter we concentrate on the effect of sexual division of labour towards

subsistence, as it presents itself as a factor that directly influences post-marital residence dynamics.

3.2.2 Previous research

3.2.2.1 Background to work on sexual division of labour

The first cross cultural study to test the labour-residence hypothesis found support for this link in aboriginal North American populations (Driver and Massey, 1957; Driver and Ulvestad, 1956). In their seminal work, Driver and Massey (1957) tested hypotheses regarding the drivers of change, correlations and sequence of evolution of social norms in North American societies. They found ample examples in the ethnographic literature and in quantitative analysis of significant correlation between sexual division of labour and residence in the Native Americans of North-America. For example, among the Iroquois a matrilineal descent system is followed where cultivated lands are parcelled out to maternal lineages and they practised matrilocal post-marital residence. The Iroquois women are heavily involved in the processing of maize, beans and squash and the oldest matron directed farm work, and the society is deemed matridominant. Similarly, in inland North-West American patrilocal societies like Tanainas and Kutchins the contribution of males in the main subsistence activity of hunting and fishing were more than that of the women. Due to an increase in demand for fur in trade, the dominance of males in these activities persisted as they could afford to buy wives from the economically forward coastal societies. Further, it has also been noticed in ambilocal societies like Puyallup-Nisqall, that men and women contribute equally to subsistence (bicentric) (Driver and Massey, 1957). The examples of these three different types of residence and division of labour substantiate the presence of a strong link between residence and sexual division of labour. Driver and Massey (1957) suggest that there is support for the idea that the sex that contributes most to subsistence is the sex that is localised, henceforth referred to as the "Classic Hypothesis" (CH). However, Alkire (1960) study of eight Micronesian societies (Ponape, Truk, Ifaluk, Yap, Palau, Ulithi, Mokil and Losap) found no consistent trend linking sexual division of labour with post-marital residence and suggested that the type of crop/subsistence might play a more crucial role in determining the division of labour between sexes. Ember and Ember (1971) tested the CH with 455 societies from across the world, using traditional analyses, and found no significant pattern. Of interest to us, they further tested this relationship within different geographic clusters to make the dataset comparable to Driver and Massey's (1957). Contradictory to expectations, Ember and Ember (1971) found a significantly opposite trend from that hypothesized relationship between the SDL and residence traits. Subsequently,

there were a number of studies using cross-cultural data that failed to find any correlation between the sexual division of labour and post-marital residence (Alkire, 1960; Divale, 1974a; Ember and Ember, 1971; Hiatt, 1970).

Recently however, Korotayev (2003b) used a worldwide cross-cultural sample to test for the inverse of the proposed CH, where instead of testing whether matridominant labour traits predict matrilocality, he tested if low female contribution to subsistence predicted non-matrilocality/patrilocality and found evidence to fully support this hypothesis. To explain the reason behind the lack of correlation between residence and sexual division of labour found by several studies (mentioned earlier), Korotayev (2003a) reasoned that marriage type played a major role and needed to be considered in this equation of division of labour and residence, henceforth referred to as the "Marriage-Mediated Hypothesis" (MMH). Korotayev (2003b) used standard correlation tests (Fisher's exact test), with the worldwide standard cross-cultural sample of 186 societies (Murdock, 1967), and found that with increasing female contribution to subsistence activities, there was a tendency for the population to be matrilocal. In a society where the contribution of women to subsistence exceeds that of men, it is understandable that matrilocality will prevail because women will receive more help in subsistence activities, where such co-operation yields returns by staying with kin. He also found the converse: in patridominant societies, residence was often patrilocal but the positive correlation between female contribution and matrilocality became negative once female contribution crossed a threshold, i.e., patrilocality was found to be preferred after this threshold. A likely factor for the curvilinear relationship between the female contribution to subsistence and post-marital residence, proposed by Korotayev (2003b), was the attraction held by non-sororal polygyny as female contribution to subsistence increased. Non-sororal polygyny is polygyny wherein a man marries many women who are not related as sisters. In a polygynous marriage system, low female contribution means that men carry the subsistence burden for large families, adding a cost to taking on more wives (Gibson and Mace, 2007). However if women contribute more than men, by taking on more wives a man gains increased labour contribution towards subsistence, which will enable a large family size (Holden and Mace, 2003). The adaptive outcome of this functional relationship, changes in social organisation leading to increased fitness of the society/family, is predictable through evolutionary theory. MMH, the relationship between increased female contribution of labour to subsistence and polygyny was supported in some societies (Heath, 1958; Schlegel and Barry III, 1986; White *et al*, 1981). In a society where sororal polygyny exists, matrilocality and matridominant sexual division of labour seems logical, like in the Arikara and Wichita Indian tribes in North America (Driver and Massey, 1957). Korotayev (2003b) showed that in matridominant, matrilocal societies, sororal polygyny is the most practised form of polygyny. Non-sororal polygyny/ general polygyny is positively correlated with

patrilocality, as it is practically impossible for non-sororal polygyny to exist in a matrilocal society (Murdock, 1949a). Thus, if not accompanied by polygyny, increase in female contribution to subsistence should not correlate with matrilocality (Korotayev, 2003b). Therefore, any understanding of the sexual division of labour and residence must take into account marriage strategies.

3.2.2.2 Methodological issues with previous studies

In all previous attempts at untangling the relationship between sexual division of labour (SDL) and post-marital residence, the problem of non-independence of data-points or autocorrelation has not been addressed. Autocorrelation means that data points that are close either temporally or spatially are very similar to each other and hence do not represent statistically independent points. Statistical methods traditionally employed in cross-cultural studies have assumed independence of the societies or populations under the study. However, human populations are related to each other through varying degrees of descent, and share cultural features due to geographical or cultural proximity. While testing for the correlation or co-evolution between two traits, it is imperative to determine if the change in one trait is dependent or influenced by the behaviour of another trait. Traditionally, one way to test associations was by simply comparing the two traits over a sample of societies to find if there existed a correlation between the two traits. However, if the societies were related, then the societies might have inherited both the traits from their common ancestor and a correlation between the two traits could simply be due to common ancestry. How would we then differentiate adaptation or co-evolution from shared ancestry? Francis Galton first identified this problem of autocorrelation in 1889 when Tylor (1889) presented a paper on the correlation between patrilocal societies and social complexity at the Royal Anthropological Institute. Tylor examined more than 300 societies with matrilocal or patrilocal residence, and the social complexity of these groups, and concluded that patrilocal societies were more likely to be complex. Galton pointed out that the correlation observed between these traits could be not only due to adaptation, but they could also exhibit similar features due to shared ancestry or through borrowing. If all the patrilocal societies in the sample shared a common ancestor, then the complexity might be a common trait inherited by all of the descending societies, and not a consequence of patrilocality. Also, if the societies were living in close proximity, then the resulting similarity could be due to borrowing and then again, the traits might not share a dependent relationship with each other. Galton argued that without controlling for shared history, one cannot infer evolutionary relationships between two cultural traits, and this problem in cross-cultural studies is now known as “Galton’s problem” (Stocking, 1968).

Scholars have since worked on many different solutions to address Galton's problem. Murdock, in his work on the Ethnographic Atlas (1967), tried to address the problem of autocorrelation by creating a sub-sample of 186 societies that had no apparent relationship, known as the Standard Cross Cultural Sample (Murdock and White, 1969). He proposed that by using societies that were completely unrelated to each other, the problem of shared ancestry would be addressed. Naroll (1961) identified two solutions to deal with the problem of autocorrelation. One solution was to sample societies that were geographically distant from each other and which had no known common ancestry, but this would have drastically reduced the sample size of available societies to test, and no meaningful hypotheses can be tested with such small samples. The other solution was a weighted regression, where societies that were related were penalized on the regression scale and this weighting taken into consideration while calculating the probability of the values observed being due to pure chance or if there was a significant correlation between the traits. One major flaw of all these solutions is the assumption of apparent relation or non-relation by observation. Assuming that two societies or a set of societies did not have a relationship only meant that there was no apparent relationship, but all human societies are related to some extent (Mace and Pagel, 1994). With a weighted regression, the method assumes that all relationships are equally distant. For example, if A, B and C are related, this method assumes that all three of them share an ancestor, and that the distance between them is equal. But generally, this is not the case. A and B might be more closely related to each other, than A and C. Therefore, by using a weighted regression, as Naroll (1961) proposed, or by controlling for linguistic or geographical proximity, it is assumed that there is no internal structuring.

To resolve this issue of discarding useful information regarding variation between cultural clusters and amongst closely related societies, Mace and Pagel (1994) suggested the use of biological comparative methods. Comparative analyses proceed with the underlying premise that cultural diversification follows biological diversification, i.e., as species diversify by descent and modification, cultures too evolve in a similar manner. Darwin proposed that an appropriate method for comparing biological trends would be able to account for instances of independent evolutionary change in species (Revell *et al*, 2008; Ridley, 1983). Mace and Pagel (1994) argue that for cultural evolution too, counting independent instances of cultural change, by constructing a phylogeny – the phylogenetic comparative methods, is an appropriate method for cross-cultural comparisons. A recent study showed that a paired t-test (a traditional statistical approach for testing correlated evolution of traits) overestimates the amount of correlation (Type-I error) by about 70% and thereby indicates a relationship of two variables where none exists (Lindenfors *et al*, 2010). By using phylogeny to control for non-independence of data, and by tracing trait evolution across the tree, interpretations regarding the relationship between traits can be robustly inferred.

To establish and understand the correlation between two cultural traits, cultural evolution scholars have (mainly) used linguistic phylogenies rather than genetic trees. Linguistic phylogenies present an appropriate tool with which to examine cultural evolution while accounting for relatedness, as language is more likely to reflect the speed at which cultural features evolve than genes. Language also acts as a group marker of identification more than genetics (Friedlaender *et al*, 2009; Hunley *et al*, 2008). The reason why we choose language trees to understand cultural traits and evolution is because even when cross-cultural migrations occur, cultural features can easily transform and migrants can adapt local traditions and human gene flow patterns cannot differentiate between adaptation and ancestry. This signal of adaptation is lost, when we take genes alone into consideration while accounting for cultural trait similarity. Whereas, since language trees are built with features that are known to be resistant to change, to a great extent, they would reflect the historical relationships between societies, and any functional cultural trait similarity due to adaptation or contact can be easily recognised (Mace and Jordan, 2011). For example, in Northern Island Melanesia (NIM), inhabitants belong to two different language families; Papuan and Oceanic. These two language families represent two different sets of migration into NIM. The Papuan speakers are considered to be the descendants of the wave of migration that occurred around 30,000 years ago, the Oceanic speakers are from a more recent migration dating to around 3000 years ago and the two migrations represent societies with different cultures and traditions. Hunley *et al* (2008) showed that due to cross cultural migrations and geographic proximities, the genetic identity of these two groups was blurred, whereas linguistic phylogenies still showed a more historical picture of relationships between the two groups, and were not as affected by contact as genes were and that language trees are preferable over gene trees to understand culturally transmitted traits. Also, while genes can cross over boundaries of cultures without much indication, it is very rare for language traits to cross over boundaries of groups where little or no cultural exchange has taken place. Several studies have examined cultural trait evolution and correlation of cultural traits using these methods, for e.g. the correlation between cultural traits of marriage and inheritance (Cowlishaw and Mace, 1996), fertility and mode of subsistence (Sellen and Mace, 1997), spread of cattle and matriliny (Holden and Mace, 2003), and evolution of cultural traits like post-marital residence (Fortunato and Jordan, 2010; Jordan *et al*, 2009), kinship (Opie *et al*, 2014), land tenure systems (Kushnick *et al*, 2014), the evolution of folktales and how they help understand population histories (da Silva and Tehrani, 2016), the coevolution of belief in moral high gods and social complexity (Watts *et al*, 2015a) and evolution of traditional knowledge (Saslis-Lagoudakis *et al*, 2014), amongst others.

3.2.2.3 Focus of this Chapter

In this Chapter, I focus on testing the hypothesized relationship between residence norms and the sexual division of labour traits in a subset of Austronesian societies. The large Austronesian language family contains a great deal of variation in residence norms (Jordan *et al*, 2009), presenting a unique opportunity to test the several theories described above. Given the pivotal role that post-marital residence has played in deciphering the genetic history of the Pacific, it is imperative to understand the dynamics of how residence norms change and evolve in the Austronesian context (see Chapter 2 & Chapter 4).
In this Chapter, I will address the following two questions:

1. Classic hypothesis: What is the correlation between post-marital residence and sexual division of labour? Do we find any support for the "classic" hypotheses?
2. Marriage-Mediated hypothesis: How does the relationship between female contribution to labour and post-marital residence work against the background of marriage type?

3.3 Questions

To answer the classic hypothesis of whether there is correlated evolution between the sexual division of labour towards subsistence and post-marital residence, I tested five models:

1. Does matrilocal residence correlate with greater female contribution to labour towards subsistence activities? If so, do changes towards matrilocality precede matridominant labour, or vice-versa?
2. Does ambilocal residence correlate with equal division of labour by both the sexes towards subsistence? If so, do changes towards ambilocality precede equal labour, or vice-versa?

I also wanted to examine the anti-thesis of the hypotheses tested above. Therefore, to test whether patrilocal residence is correlated with female contribution to subsistence activities, we investigated the following question:

3. Does patrilocal residence correlate with greater female contribution to labour towards subsistence? If so, do changes towards patrilocality precede matridominant labour, or vice-versa?

To understand the role of marriage ecology in influencing division of labour and post-marital residence, I asked the following questions by restricting the sample to only those societies that allow for polygyny:

4. Does patrilocal residence correlate with greater female contribution to labour towards subsistence activities in *polygynous* societies? If so, do changes towards patrilocality precede matridominant labour, or vice-versa?

5. Is patrilocality correlated with equal division of labour towards subsistence in polygynous societies?

3.4 Data

Data was collected on 109 Austronesian societies regarding their primary post-marital residence, marriage form, and information on sexual division of labour with reference to the primary subsistence activity. Information on subsistence types was also collected where possible. The data for post-marital residence was collected from the Ethnographic Atlas (EA) (Murdock, 1967), the Standard Cross Cultural Sample (SCCS) (Murdock and White, 1969) and (Jordan *et al*, 2009). As all the societies that were coded from the EA and SCCS were part of the sample used by Jordan *et al* (2009), the same coding scheme used by Jordan *et al*, 2009a was followed for post-marital residence societies in the current sample (See Appendix for data and coding).

3.4.1 Coding schemes

The primary subsistence activity was first identified from the Ethnographic Atlas (EA) (Murdock, 1967; Murdock and White, 1969), Ethnic groups of Island Southeast Asia (Lebar, 1972), the Encyclopaedia of World Cultures (Levinson, 1993), and from primary ethnographic descriptions where necessary. In the EA, the categories of "Gathering", "Hunting", "Fishing", "Horticulture" and "Agriculture", were examined for percentage dependence of each society for each of these subsistence activites, and these values determined the main subsistence of each society (See appendix for details of the societies and their coding).

Subsequently, the division of labour towards subsistence activity, for the primary mode of subsistence, was identified through the category of "Sex Differences" in the EA. If there were more than one main subsistence type for a society, then the main subsistence type was decided based on the "principal" or "major" subsistence activity mentioned in the Pulotu database, a database recently compiled of Austronesian beliefs and practices (Watts *et al*, 2015b).

3.4.1.1 Sexual division of labour trait coding

Societies were coded under three coding schemes. If women contributed more to subsistence, the code assigned was "F", otherwise known as matridominant (females contribute more). We categorised the categories of "Females only or mostly alone" and "Females appreciably more", in the Ethnographic Atlas as belonging to this matridominant category. If men contributed more, i.e., the categories of "Males only or mostly alone" and "Males appreciably more", it was

coded as patridominant, denoted as "M". When men and women contributed to subsistence equally, either with or without marked differentiation in their traits – the categories of "Equal participation, no marked differentiation" and "Differentiated but equal participation", they were coded as E", also termed equidominant traits. Both these categories represent equal division of labour, hence the codes for both these categories is combined. Of the 109 societies in the sample, we were able to garner information on sexual division of labour traits for 86 societies (equidominant (n=39), matridominant (n=15), patridominant (n=32). Missing data was coded as "-" or "NA".

3.4.1.2 Post-marital residence

The primary mode of residence, i.e., "Marital residence with kin: after first years", was taken as indicator of the post-marital residence status of the society. Societies were coded as "P" if their primary state of post-marital residence was either virilocal or patrilocal (See Chapter 2 for definitions), as both states involve residence with the groom's patrilineal kin. They were coded as "M", if they were matrilocal, avunculocal or uxorilocal, as all these states involve residence with the bride's matrilineal kin (avunculocal indicates residence specifically with maternal uncle, the mother's eldest brother). They were coded as "A" if they were ambilocal or neolocal, as both these residences mean the same when looked at in terms of the sexual division of labour hypothesis (and both states are henceforth referred to as ambilocal). Societies were categorised as "U" if the residence state was unknown or the description was ambiguous. Further, marriage type for each society was recorded as to whether polygyny was a main form of marriage. This division was in order to test for a correlation between sexual division of labour towards subsistence both with and without the background of polygyny. Residence traits were coded for 107 out of the 109 societies in our dataset (ambilocal (n = 25), matrilocal (n =27) and patrilocal (n = 55)).

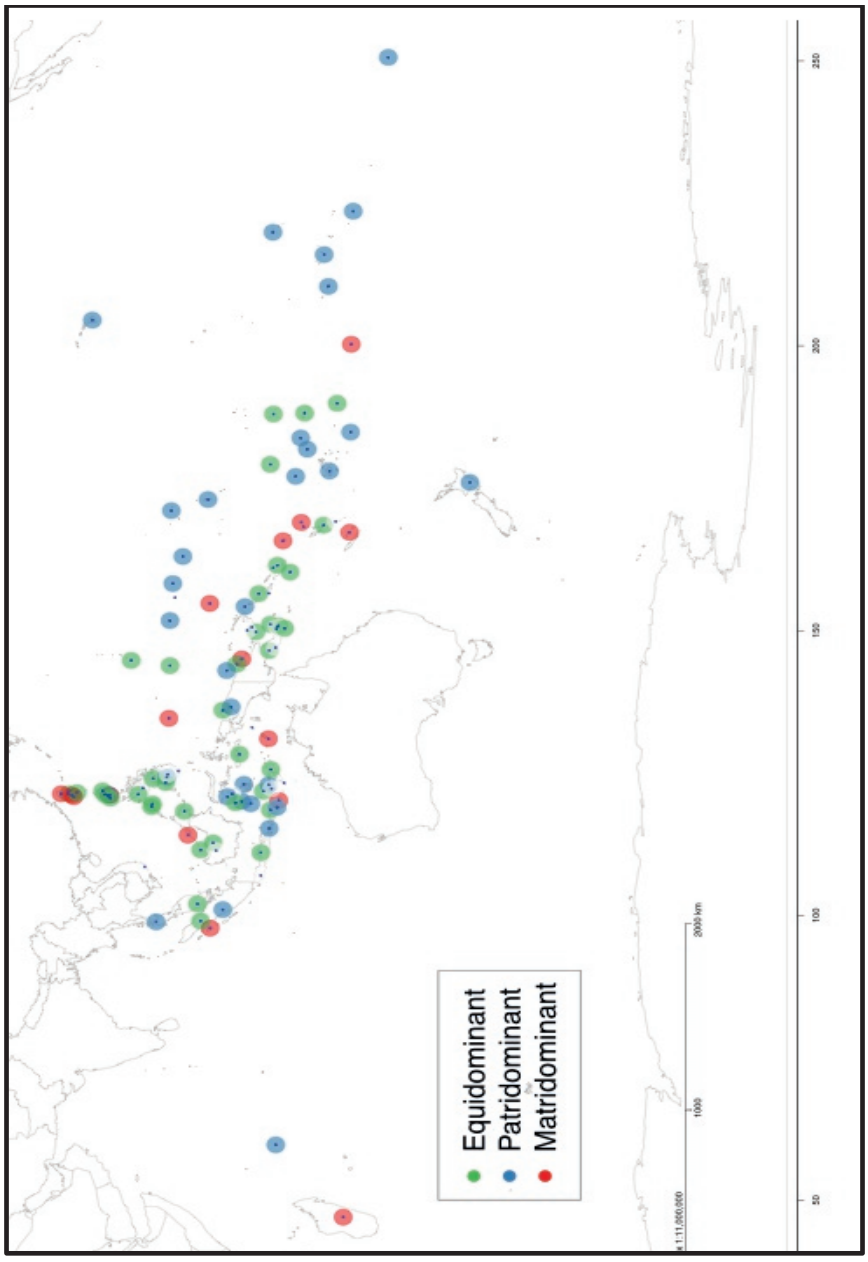


Figure 3–1: Distribution of sexual division of labour traits "Equidominant", "Patridominant" and "Matridominant" of Austronesian societies as coded in the Chapter

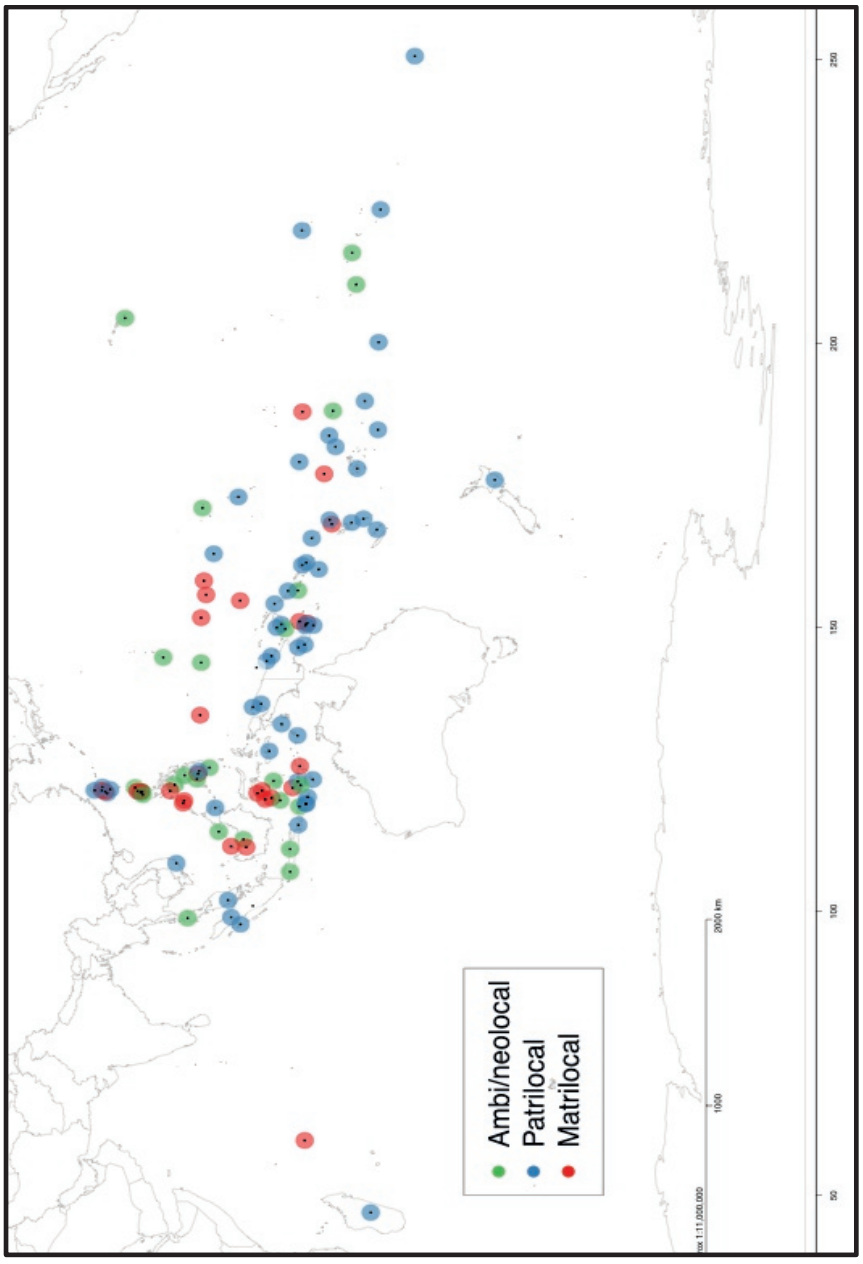


Figure 3–2: Distribution of residence traits "Ambilocal & Neolocal", "Patrilocal" and "Matrilocal" residence traits of the Austronesian societies as coded in our Chapter.

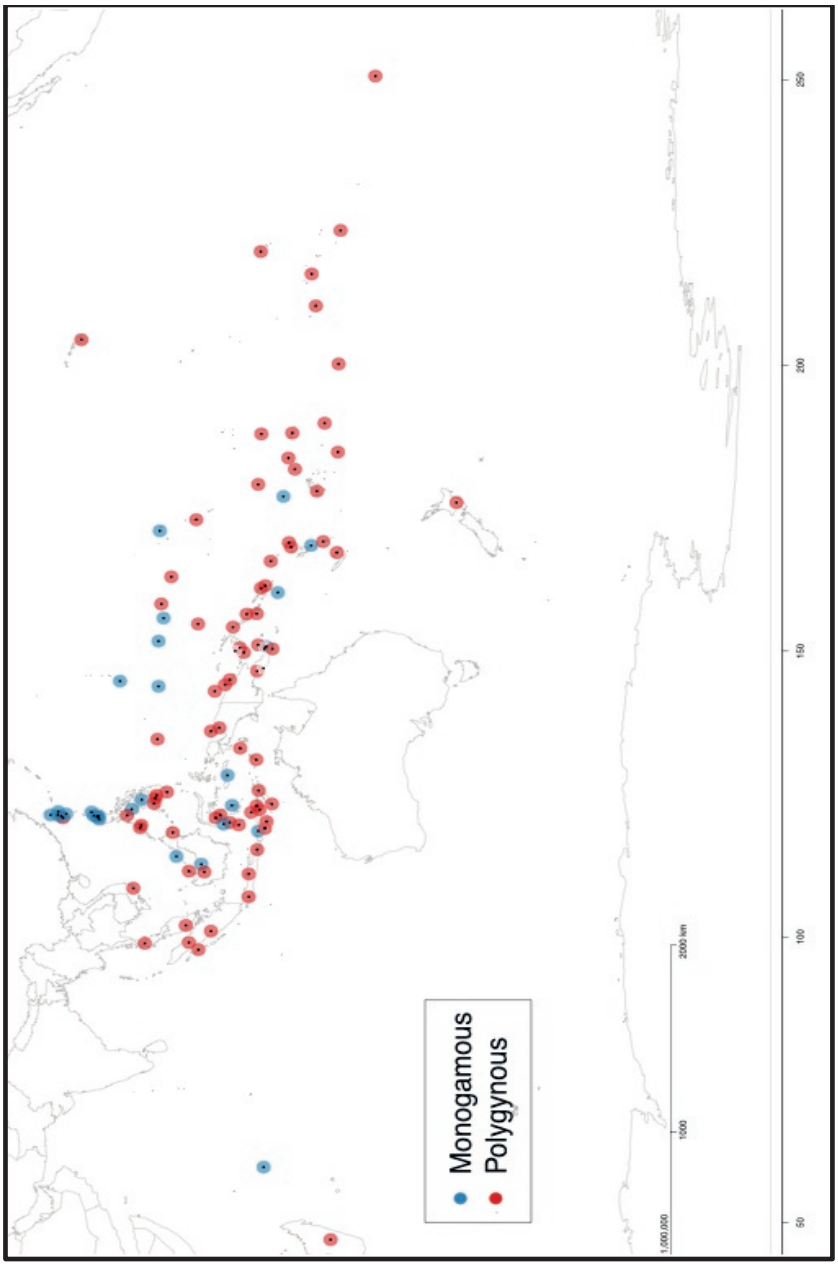


Figure 3–3: Distribution of "Polygynous" and "Monogamous" traits in Austronesian societies. Polygynous societies here are societies where Polygyny is allowed.

3.4.2 Models to be tested

To test the "Classic hypothesis" (CH), we used the coding schemes where data was categorised into

a) Matridominant division of labour and matrilocal residence vs. Others
b) Equidominant division of labour and ambilocal residence vs. Others
c) Matridominant division of labour and patrilocal residence vs. Others

To test the "Marriage mediated hypothesis" (MMH), we used

d) Matridominant and patrilocal residence vs Others, in Polygynous societies
e) Equidominant and patrilocal residence vs Others, in Polygynous societies

3.4.3 Language data

Language trees derived from the Austronesian Basic Vocabulary Database (ABVD), which has cognate-coded lexical items from the 210-term basic vocabulary set for over 400 Austronesian languages was used (Greenhill *et al*, 2008). The set of trees (1000) was inferred using Bayesian Markov-chain Monte Carlo (MCMC) methods (Huelsenbeck *et al*, 2000; Jordan, 2007; Jordan *et al*, 2009; Yang and Rannala, 1997).

3.5 Methods

3.5.1 Non-phylogenetic analysis

The distribution of each trait in combination of other traits was explored. Since sample size was limited and the distribution of traits was extremely skewed, it was prudent to scale these distributions according to their presence in our dataset. For example, there are 22 patridominant and 14 matridominant societies practicing agriculture, however, these two numbers cannot be compared directly to deduce that agricultural societies are largely patridominant. There is a need to scale both matridominant and patridominant societies in this case, to their total number in the sample. The number of patridominant societies are 32 while the number of matridominant societies are 15, therefore, by scaling, we see that 93% of the total matridominant societies in our dataset practice agriculture while 69% of the patridominant societies practice agriculture.

Also, for each coding scheme, traditional methods for correlation used in cross cultural studies, i.e., Fisher's exact test and chi-squared tests were performed to test the independence of these traits. Along with these, Bernard's coefficient was also performed, as it is said to be more sensitive in detecting underlying correlations.

3.5.2 Language Phylogenies

The 1000 posterior trees from the ABVD dataset were pruned to include only those languages for which we had data on sexual division of labour, marriage and post-marital residence norms (n = 109). For the models testing the MMH, the datasets were pruned further, as there were only 75 societies of the 109, where polygyny was allowed. The properties of trees, such as branching patterns and branch lengths, were retained in the resultant trees in proportion to their posterior probabilities. From this posterior distribution of pruned trees, a maximum clade credibility (MCC) tree was computed using the *TreeAnnotator* package of BEAST v.1.7.2 (Drummond *et al*, 2012). The MCC tree is the tree from the posterior distribution that has the highest overall score of clades appearing in all the trees in the posterior sample. The MCC tree is used to visually summarize the posterior, however for the comparative analyses, all 1000 trees were used. The advantage of this approach, in using a set of trees, is that it allows for the quantification of uncertainty in the phylogenetic tree model. This is particularly relevant for studying cultural evolutionary systems, as a single tree is unlikely to accurately represent human population history (Boyd *et al*, 1997; Gray *et al*, 2009; Pagel, 1997).

3.5.3 Phylogenetic Comparative Methods

3.5.3.1 Testing for phylogenetic signal of the sexual division of labour trait

One approach to understand if trait variation is correlated to phylogenetic relatedness of the societies in the current sample, was by quantifying the phylogenetic signal in the data. A strong phylogenetic signal is when the probability of related societies to possess similar traits is significantly higher than when two societies are drawn at random from a phylogenetic tree (Blomberg *et al*, 2003). Conversely, a low phylogenetic signal indicates a random distribution of traits in the tree. Estimating the phylogenetic signal will give us an insight into whether the use of phylogenetic methodology is justified to understand the evolution of a trait. In studies investigating animal societies, the use of phylogenetic methods to understand biological or behaviour trait evolution is becoming increasingly popular (Johnson and Stinchcombe, 2007). While some argue that closely related species or societies will always occupy similar environments (Swenson *et al*, 2007; Wiens and Graham, 2005), and that strong phylogenetic signal should be an apriori expectation while examining trait variation (Duncan and Williams, 2002), others argue that variation of behavioural traits exhibit strong signal for some traits and not others (Losos, 2008). Nevertheless, it seems prudent to test for the strength of phylogenetic signal in

our data, and if we discover a strong phylogenetic signal, it would reinforce the importance of controlling for phylogeny while understanding trait evolution and the use of phylogenetic comparative methods. Alternately, if we find a low phylogenetic signal, then the strength of correlation of traits by using traditional methods should be similar to the results from using phylogenetic comparative methods.

While there are several tests to estimate phylogenetic signal for continuous traits (traits which cannot be categorized into binary variables, for e.g. brain size, weight, etc.,), like Blomberg's K (Blomberg *et al*, 2003) and Pagel's λ (Freckleton *et al*, 2002; Pagel, 1997; Pagel, 1999b), we are interested in methods like the measure *D* (Fritz and Purvis, 2010) that can estimate phylogenetic signal for discrete traits. *D* is a measure based on the sum of differences between sister clades in a phylogenetic tree and is implemented in the caper R package (Orme *et al*, 2012). A phylogenetic signal would exist if traits evolve in Brownian-motion, i.e., in a random walk with constant variation over time (Felsenstein, 1985). In such a scenario, a linear relationship would be expected between phylogenetic relatedness (quantified as branch lengths between two taxa) and the degree of trait similarity i.e., the less amount of time since the divergence of two societies from a common ancestor, the less the expected trait difference between them (Blomberg *et al*, 2003; Harvey and Pagel, 1991; O'Meara *et al*, 2006; Pagel, 1999a; Pagel, 2002).

To calculate *D*, the mean sum of changes under the Brownian model of trait evolution and random model of trait evolution are simulated and used to calibrate the observed sum of changes for the trait of interest. A clumped trait will have very low values of sum of difference between sister clades, and a high value for the sum of difference between sister clades indicate a highly dispersed trait, all else being equal. At each internal node of the tree, the value for the node was estimated as the mean of the values at the descendant nodes inversely weighted to the length of the branches leading to them. This accounts for the Brownian model of evolution by taking into account the time since divergence of the two branches and variation in the trait, i.e., longer branches are more likely to have a chance of trait change from the ancestral node than short branches. Then the differences between each pair of sister clades were summed across the tree to give the observed sum of sister-clade differences, d_{obs}. Two datasets were then simulated, given the prevalence of the binary trait in consideration and the phylogeny of interest, under a random model of trait evolution and a Brownian model of trait evolution by shuffling the traits across the tree (see Figure 3-10). The *d* values for the random and Brownian model of trait evolution, across many simulations were then calculated to yield a distribution of d_r and d_b scores, respectively. The measure of phylogenetic signal *D* is then estimated by scaling the observed sum of sister-clade differences (d_{obs}), with the mean values of the expected distribution under random trait evolution, d_r, and brownian trait evolution, d_b. It is given by

$$D = [d_{obs} - \text{mean}(d_b)]/[\text{mean}(d_r) - \text{mean}(d_b)]$$

The *D* statistic is expected to be equal to 1 if the binary trait is phylogenetically randomly distributed across the phylogeny and equal 0 if the binary trait has a clustered distribution resembling a trait evolving under Brownian motion. Values of *D* are not necessarily limited between 0 and 1. A value of above 1 indicates an over-dispersion, where societies which are phylogenetically closer will be more dissimilar than two societies picked at random. Similarly, values less than zero indicate extreme clustering, where probably the trait is present in few closely related branches of the phylogeny and absent in others. Significance of the estimated *D* value is then calculated by testing if it is significantly different than 0 (indicating random distribution of traits) or if it is significantly different from 1 (indicating phylogenetic clustering) by permuting the data set a number of times.

The assumptions of the model deriving *D* indicate that the discrete trait needs to be coded as a binary trait, and the discrete trait must be a categorical variable based on a underlying continuous variable (Fritz and Purvis, 2010). The sexual division of labour trait is categorized based on the amount of energy and time invested by a particular sex in the task. Therefore, the variable is actually categorized based on an underlying continuous variable and upholds the assumption for this test. The distribution of residence trait cannot be tested for a phylogenetic signal as the trait is not based on a continuous variable.

The value of D was estimated over the sample of 1000 trees from the ABVD sample pruned for the dataset, for the five different coding schemes. The median and range for the D value across the trees was evaluated. Calculating the value of D across the tree sample accounts for the phylogenetic uncertainty. The proportion of the sample where the D value was either significantly different from 0 (indicating random evolution of traits) or 1 (indicating phylogenetic evolution of traits) was also estimated. The D value for the MCC tree and the corresponding p-values was calculated, to get an idea of the predominant pattern for the distribution of traits in the sample.

3.5.3.2 Testing for the co-evolutionary relationship between residence norms and sexual division of labour traits

To reconstruct the co-evolutionary trajectories of post-marital residence norms and sexual division of labour to subsistence, the phylogenetic comparative method (*Discrete)* implemented in *BayesTraits* was employed (Pagel, 1999b; Pagel and Meade, 2006; Pagel *et al*, 2004). This procedure tests the likelihood of co-evolution of two traits by comparing two models of trait evolution, an independent and a dependent model. An independent model works under the assumption that

changes in one trait do not affect or influence changes in the second trait, i.e., the two traits evolve independent of each other. In a dependent model, it is assumed that the change in one trait affects the state of the second trait, i.e., the traits co-evolve. The likelihoods of each model are calculated and compared. Along with the likelihoods, the probability of state change in a trait is calculated i.e., the probability that the state of a trait (given the two states: 0,1) will change from 0 to 1 or from 1 to 0: these are termed transition rate parameters. For an independent model since the rates of change of one trait does not influence the other trait, we will have a transition rate matrix consisting of 4 rates (Fig 3-4). For a dependent model, since the state of trait depends on the other, there are essentially four possible states 00, 01, 11 and 10. The transition rate matrix will be composed of eight rates, with instantaneous rate changes possible but also assuming that both the states cannot change in the same instance to account for the model (Figure 3-5). Given the comparative data of the two traits and the tree sample, the model uses a continuous Markov model to describe the evolution of the traits of interest along the branches of the phylogeny. The rate parameters define the probabilities of each change in the rate transition matrix (explained in detail below), and therefore both the character states at the internal nodes on a tree and the overall likelihood of the data (Pagel, 1994; Pagel, 1999b). The rate parameters of the maximum likelihood solution for each of the 1000 phylogenies was obtained.

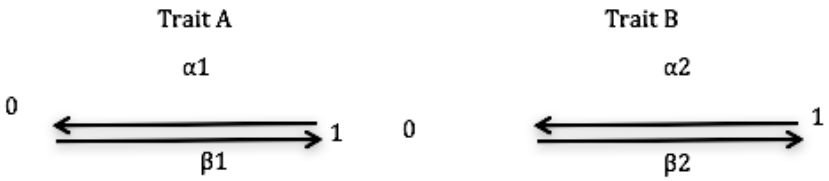


Figure 3–4: Independent model of trait evolution. Trait A evolves independently of Trait B. Each trait can take either state 0 or state 1. There are four possible transition rates: α1, α1, β1, β2. α represents the change in state of a trait from 1 to 0, while β represents the change in the state of a trait from 0 to 1

To infer which model best explains the assignment of the states of both traits on the tips of the phylogeny, we calculated the likelihoods of both the models by characterizing the probability of change of state. In an independent model, if we considered one of the states at the root, then the transition rate parameters (the probability of change from 0 to 1/1 to 0) for the two traits independently to produce the observed diversity was calculated. This probability was computed for

all possibilities at the internal nodes and over all the branches of the tree, for both traits. In a dependent model, instead of four probabilities for each branch, there were eight transition rate parameters, which were calculated in the same way across the phylogeny (refer to Pagel, 1999b for details on calculations).

One advantage of this likelihood method is that it accounts for the changes in branch lengths, i.e., changes in longer branches are more likely than changes in a shorter branch. Rather than making a definitive construction of ancestral state at each node, the likelihood approach integrates over all the possible values at all the ancestral nodes. Subsequently, to determine which model of trait evolution is more likely, dependent or independent, we use a likelihood ratio statistic (LR). This statistic tests the goodness of fit of both the independent and the dependent model.

$$LR = -2\,log_e \frac{lh(I)}{lh(D)}$$

Equation 3-1: Likelihood ratio test (LR). lh(I) represents the likelihood of the independent model and lh(D) represents the likelihood of the dependent model.

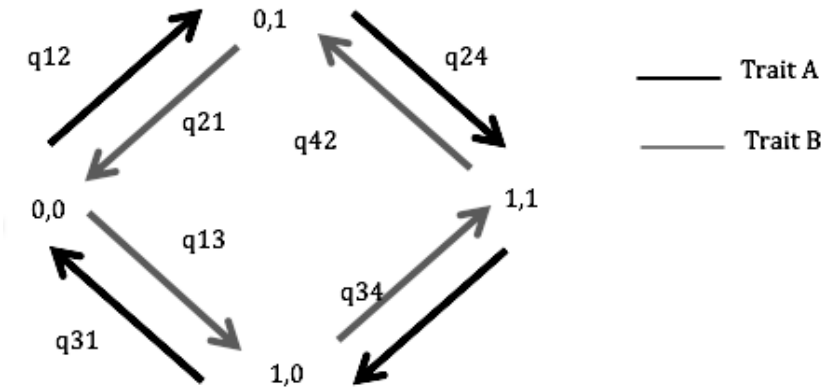


Figure 3–5: Dependent model of trait evolution. There are 4 transition states with 8 rate parameters represented by "q". The subscript of the rate parameters represents the starting and ending state of each transition, indicated by 1,2,3,4 corresponding to states (0,0), (0,1), (1,0), (1,1) [(first trait, second trait)].

The likelihood ratio statistic (LR) is quantified by using the formula in Equation 3-1, where lh(I) is the likelihood of the independent model and lh(D) is the likelihood of the dependent model. A null distribution of LR scores is simulated and a chi-square distribution was used to compare our value of LR (with the

degrees of freedom equivalent to the difference in model parameters) to determine if the LR score obtained is significant. If the value is significant, then the dependent model of evolution of traits is significant; else, the independent model fits significantly better for understanding the evolution of the traits in consideration. The dependant model is treated as the general model, as the independent model can be obtained by constraining some of the rates of the dependent model to zero and therefore, a significant LR score means that a dependent model of evolution is significantly more likely.

There is character uncertainty inherent in calculating the likelihood by integrating over all possible state changes in a tree (Ronquist, 2004) and as explained earlier, there is also uncertainty associated with the phylogeny (as will be seen in Chapter 5). In this maximum likelihood method, we approximate over a sample of trees by computing the likelihood of each state at all sites as the probability (data|model), and the solution with the highest overall likelihood is selected for each tree.

3.5.3.2.1 Maximum likelihood estimation of correlated evolution

To test whether the two traits of residence and division of labour were correlated while controlling for the shared history of these societies, we used methods outlined in section 4.3 and implemented in BayesTraits. 1000 trees from the posterior probability sample of language trees published by Gray *et al* (2009) was this used. Maximum likelihood (ML) analyses were performed over the sample of 1000 pruned trees with 10 ML tries per tree. The rate parameters were estimated for each tree and the medians are reported. These reported medians were used to compare the independent and dependent model through the likelihood ratio test.

3.6 Results

3.6.1 Non-Phylogenetic analysis

3.6.1.1 Exploratory analysis

A cursory glance at the distribution of the main subsistence types (Figure 3-6) reveals societies mainly dependent on agriculture for subsistence in the dataset (~80%).

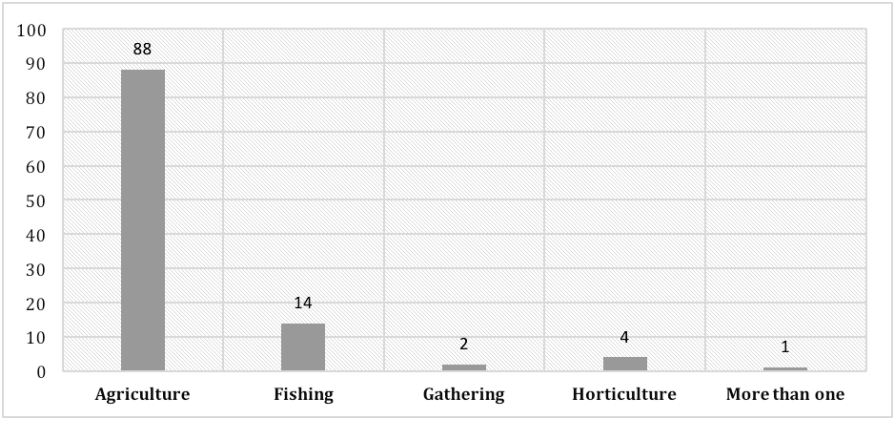


Figure 3–6: Chart depicting the distribution of the main subsistence types amongst the societies in our dataset of Austronesian societies. Each of the subsistence types indicated here are the main contributors to subsistence of a society. "More than one" category indicates any society where the main subsistence type is more than one activity.

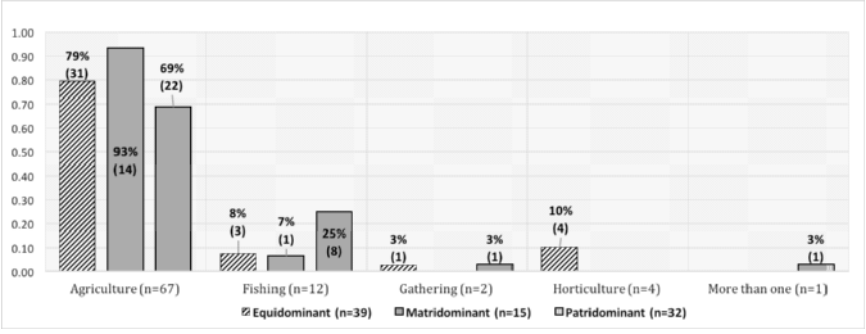


Figure 3–7: Chart depicting the scaled distribution(according to the number of societies present under each division of labour code) of sexual division of labour traits amongst the different subsistence modes. The "More than one" category of subsistence has 3 societies, but we have data for only one society. Y-axis represent the percentage of societies present amongst each division of labour trait present under each subsistence type.

The distribution of the division of labour traits amongst the different subsistence types (Figure 3–7) reveals fishing societies to be mainly patridominant in their division of labour and horticulture almost always seems to be equidominant. However, only 12 fishing societies and 4 horticultural societies are present in a sample of 86 societies with available division of labour traits, so the pattern has to be viewed with caution. Examining the distribution of the residence traits amongst the different residence types, we see predominantly ambilocal and patrilocal societies, however the number of fishing societies for which residence data is available, is only 13 out of 107, so the pattern might be a spurious correlation and not survive in a larger dataset. The scaled distribution of

residence traits amongst the sexual division of labour traits was examined next. For both traits, data was available for a total of 84 amongst the 109 societies.

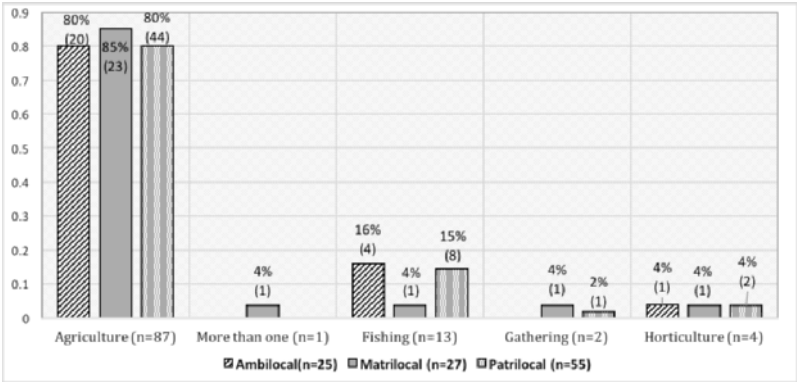


Figure 3–8: Chart depicting the scaled distribution (according to the number present in the sample under each residence type) of residence traits amongst the different sexual division of labour traits. The "More than one" category of subsistence has 3 societies, but we have data for only one society. Y-axis represent the percentage of societies amongst each residence type present under each subsistence type.

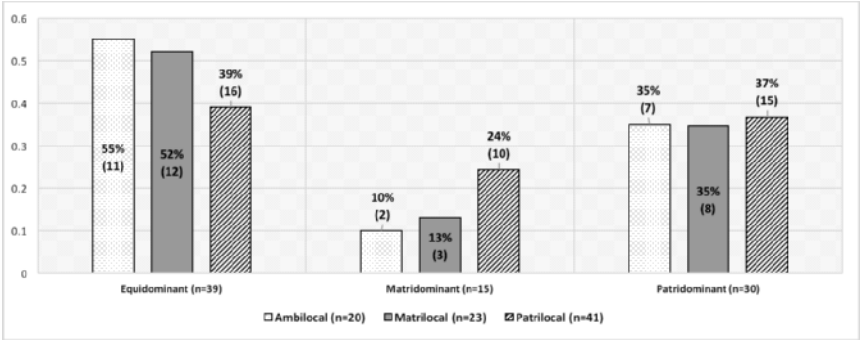


Figure 3–9: Chart depicting the scaled distribution(according to the number of societies present under each division of labour code) of residence traits amongst different sexual division of labour traits amongst the different subsistence modes. Y-axis represent the percentage of societies present amongst each residence trait present under each division of labour trait coding.

Matridominant societies largely seem to be practising patrilocality, however this should be viewed in light of only 15 matridominant societies present in the sample. However, if this relation were to hold true, then it would resemble the results found by Ember and Ember (1971), when they tested the CH within geographical clusters. The number of patrilocal societies is almost double that of ambilocal and matrilocal societies in our dataset, and the distribution of patrilocal societies present is double that of the other residence traits (ambilocal and matrilocal) in each division of labour trait categories. So, any pattern with patrilocality cannot be discounted due to chance. No discernible pattern is visible

in the distribution of residence traits amongst the different division of labour traits.

3.6.1.2 Tests of independence using traditional methods

The measure of correlation between sexual division of labour and post-marital residence was investigated next. The chi-squared values for all the models, followed by a Fisher's exact test (including Yates correction for small sample sizes) to test for significance and Bernard's test of independence was estiamted.

3.6.1.2.1 Model1: "Matrilocality and Matridominant division of labour"

Here tests for a correlation between matrilocal societies and women contributing more to subsistence i.e., matridominant societies (F) was performed (Table 3-1).

Table 3-1 Matrilocality versus matridominant division of labour. Values represent the sample sizes (number of societies) for each of the four states represented in fig 4.2.

Subsistence Coding	Matrilocal	Other	Chi squared (p-value)	p-value from Fisher's exact test	Bernards Test two-sided p-value
Matridominant (F)	3	13	0.625	0.5412	0.4615
Other (O)	20	50			

3.6.1.2.2 Model2: "Ambilocal & Equal division of labour"

Correlation between ambi-/neolocal with equal division of labour (E) was tested (Table 3-2).

Table 3-2 Ambilocality & Equal division of labour. Values represent the sample sizes (number of societies) for each of the four states represented in fig 4.2.

Subsistence Coding	Ambilocal	Others	Chi squared (p-value)	p-value from Fisher's exact test	Bernard's Test two-sided p-value
Equal (E)	11	30	0.622	0.61	0.58
Others (O)	9	36			

3.6.1.2.3 Model3: "Patrilocal & Matridominant division of labour"
Correlation of patrilocal societies with females contributing more to subsistence – matridominant (F) was tested (Table 3-3).

Table 3-3 Patrilocality & matridominant division of labour. Values represent the sample sizes (number of societies) for each of the four states represented in fig 4.2.

Subsistence Coding	Patrilocal	Others	Chi squared (p-value)	p-value from Fisher's exact test	Bernard's Test two-sided p-value
Matridominant (F)	11	5	0.1659	0.1647	0.1183
Other (O)	32	38			

3.6.1.2.4 Model4: "Polygynous – Patrilocal & Matridominant division of labour"

Correlation of patrilocal [P] societies with females contributing more to subsistence (F) in polygynous societies was tested (Table 3-4).

Table 3-4 Patrilocality & matridominant division of labour in Polygynous societies. Values represent the sample sizes (number of societies) for each of the four states represented in fig 4.2.

Subsistence Coding	Patrilocal	Others	Chi-squared (p-value)	p-value from Fisher's exact tests	Bernards Test two-sided p-value
Matridominant (F)	9	2	0.2204	0.1742	0.125
Other (O)	27	21			

3.6.1.2.5 Model5: "Polygynous – Patrilocal & Equal division of labour"

Correlation of patrilocal societies with equal division of labour (E) in polygynous societies was tested (Table 3-5).

Table 3-5 Patrilocality & Equal division of labour in Polygynous societies. Values represent the sample sizes (number of societies) for each of the four states represented in fig 4.2.

Subsistence Coding	Patrilocal	Others	Chi-squared (p-value)	p-value from Fisher's exact tests	Bernards Test two-sided p-value
Equal (E)	12	12	0.244	0.1814	0.1689
Other (O)	24	11			

No significant association was found in any of the tested models. As Phylogenetic relationships were not taken into account for these tests, there could be associations that were not detected. In a study of East African kinship and marriage Borgerhoff Mulder *et al* (2001), found at least two instances where "traditional" methods did not find any correlation between traits but when phylogenetic comparative methods were used, a significant relationship was detected. The reason we can expect such a correlation to emerge could be the lack of power in the "traditional" tests, where the instances of independent evolution, even if scarce, might not be weighed against those which have evolved due to common ancestry. Therefore, the next step was to test these associations against the backdrop of phylogeny.

3.6.2 Phylogenetic comparative method

3.6.2.1 Testing for phylogenetic signal

The phylogenetic signal was tested by estimating *D*, where a value of *D* closer to zero indicates a strong phylogenetic signal for the trait evolution in question, and a value of *D* closer to one indicates a random distribution of the trait on the phylogenetic tree. The value of *D* was estimated for the sexual division of labour traits, as it fit the assumptions of a categorical and binary trait, with an underlying continuous distribution. Essentially, the value of *D* for "Matrilocal and matridominant division of labour" and "Patrilocal and matridominant division of labour" coding schemes should give similar results, as we are estimating the phylogenetic signal only for the matridominant trait in both these schemes. The results meet expectations, where both these schemes have similar *D* values.

Table 3-6: Table depicting the estimation of the strength of phylogenetic signal for the sexual division of labour trait under different coding schemes, through the estimator D. D_MCC is the value of D estimated on the maximum clade credibility tree built from the Bayesian posterior of 1000 trees. Pval0mcc is the p-value indicating whether the estimated D value is significantly different from 0 (therefore not evolving under Brownian motion) or not. Pval1mcc is the p-value indicating whether the estimated D value is significantly different from 1 (therefore not evolving randomly) or not. D_med, D_range, is the median and range respectively of the estimated D values from the Bayesian posterior of 1000 trees. % of D value<0.5 & %of D value>0.5 is the percentage of D values calculated for the Bayesian posterior of 1000 trees, where the value of D is less than 0.5 and greater than 0.5 respectively. %of p-value0>0.05 and %of p-value 1>0.05, indicate the percentage of D values which are not significantly different from zero or 1, respectively (at 95% significance level)

Model	D_MCC	Pval0 mcc	Pval1 mcc	D_med	D range	%of *D* value <0.5	%of *D* value >0.5	%of p-value0 >0.05	% of p-value1 >0.05
Matrilocal & matridominant division of labour	0.53	0.18	0.13	0.66	0.21-1.24	11.8	88.2	94.2	98.9
Ambilocal & equal division of labour	0.71	0.06	0.11	0.76	0.46-0.98	3	99.7	44.2	96.5
Patrilocal & matridominant division of labour	0.53	0.19	0.13	0.65	0.25-1.24	11.5	88.5	94	98.9
Patrilocal & matridominant division of labour (Polygynous societies)	0.36	0.11	0.32	0.35	-1.96	72.8	27.2	99.8	89.6
Patrilocal & equal division of labour (Polygynous societies)	0.41	0.27	0.04	0.36	0.02 - 0.71	91.4	8.6	100	21.1

The value of *D* lies between 0 and 1 for most coding schemes and the p-value is significantly different from both 0 and 1, indicating that neither a phylogenetic clustering nor a random distribution of the trait could be dismissed. Only for equidominant division of labour and matridominant division of labour for the sample including all societies and when only polygynous societies are

included, are the *D* values marginally significant. Interestingly, when controlled for marriage, the value of *D* changes from being significantly different from a phylogenetic distribution to being significantly different from a random distribution trait, in the equidominant division of labour coding scheme. The *D* values when tested over the Bayesian posterior of 1000 trees for the matridominant coding scheme of division of labour in polygynous societies gives a range between -0.35 to 1.61 (extreme clustering to over dispersion), revealing an uncertainty in the estimation of D itself, when accounting for the uncertainty in the phylogeny of Austronesian languages. The value of *D* for other coding schemes in the entire sample and in polygynous societies reveals that we cannot discount the effect of phylogeny on the inheritance of these traits.

3.6.2.2 Testing for co-evolutionary relationship between residence norms and sexual division of labour traits

The dependent and independent runs of each for each of the models were compared using the likelihood ratio statistic (LR). The p-value denotes whether the dependent model is significantly better than the independent model (Table 3-7). The probability of the ancestral state (root) is also estimated for the combination of traits.

Table 3-7 Table depicts the results of the likelihood ratio statistic for all the models based on the maximum likelihood results. LR is the likelihood ratio, p-value denotes the significance of the LR test. The columns labelled "Root" depict the probability of that particular state being root, with the suffixes following the root denoting the combination of trait states for which the probability value is calculated.

S.no	Model	LR	p-value
1	Matrilocal & matridominant division of labour	4.85	0.30
2	Ambilocal & equal division of labour	1.45	0.84
3	Patrilocal & matridominant division of labour	4.08	0.40
4	Patrilocal & matridominant division of labour (Polygynous societies)	3.91	0.42
5	Patrilocal & equal contribution (Polygynous societies)	4.43	0.35

The maximum likelihood analysis of five of the models yielded no significant Likelihood ratio statistic in any of the models. In majority of the coding schemes,

the dependent model of trait evolution was as likely as an independent model of evolution for the residence and SDL traits.

3.7 Discussion

3.7.1 Co-evolution of division of labour and residence norms

Several debates exist regarding the presence of a co-evolutionary relationship between sexual division of labour and residence (Driver and Massey, 1957; Ember and Ember, 1971). We investigated whether SDL and residence norms evolve in a dependent and co-evolutionary manner in Austronesian societies. To test the "Classic" hypothesis, i.e., "the sex that contributes the most is the sex that stays put", the behaviour of the traits in societies with (a) matridominant SDL and matrilocal residence norms vs societies with other norms, (b) equal division of labour and ambilocal residence norms vs societies with other norms (Driver and Massey, 1957) and (c) matridominant division of labour and patrilocal residence norms vs societies with other norms was observed. To test the "Marriage mediated" hypothesis (Korotayev, 2003b), tests of whether there exists a co-dependent and predictable relationship between equidominant division of labour and patrilocal residence norms, and matridominant division of labour and patrilocal residence norms, in polygynous societies alone was investigated. For both the tests, there was no significant evidence of a dependent evolution of traits over an independent evolution, thus indicating no direct support for either hypothesis.

While testing for the phylogenetic signal (*D*), some interesting patterns emerged. As only test for the sexual division of labour traits could be conducted, due to the limitations of the index (only valid for a trait with an underlying continuous distribution), I was not able to test whether the traits of post-marital residence were phylogenetically clustered or not. The results (Table 3-6) indicated that a phylogenetic clustering of traits could not be dismissed under any of the coding schemes in our study. This reinforces the need to control for phylogeny while testing for correlation between traits.

The *D* values indicated that trait evolution in a phylogenetic manner could not be discounted, but a random evolution could not be discounted either. Also, the range of the *D* values across the 1000 Bayesian posterior trees indicated an inherent uncertainty of the D value, with wide ranges. There is no clear pattern emerging for most coding schemes, across the sample distribution, indicating that probably certain clades of the tree have these traits in a random. Interestingly, when the phylogenetic signal was tested against the background of polygyny, the equidominant trait dispersion trend reversed, and the phylogenetic signal was strong and significant. Whereas, in the absence of this control, the trait dispersion was marginally significant to indicate random trait dispersion. This indicates that

the behaviour of the sexual division of labour trait is complex, and needs further investigation on the variables that influence its evolution. We know that by controlling for marriage ecology, we are able to discover the underlying phylogenetic signal in the data, i.e., by controlling for a factor affecting sexual division of labour in a particular setting (equidominant), we are able to clearly observe the phylogenetic evolution of the trait. This gives us reason to believe that perhaps more forces are at play than what we accounted for.

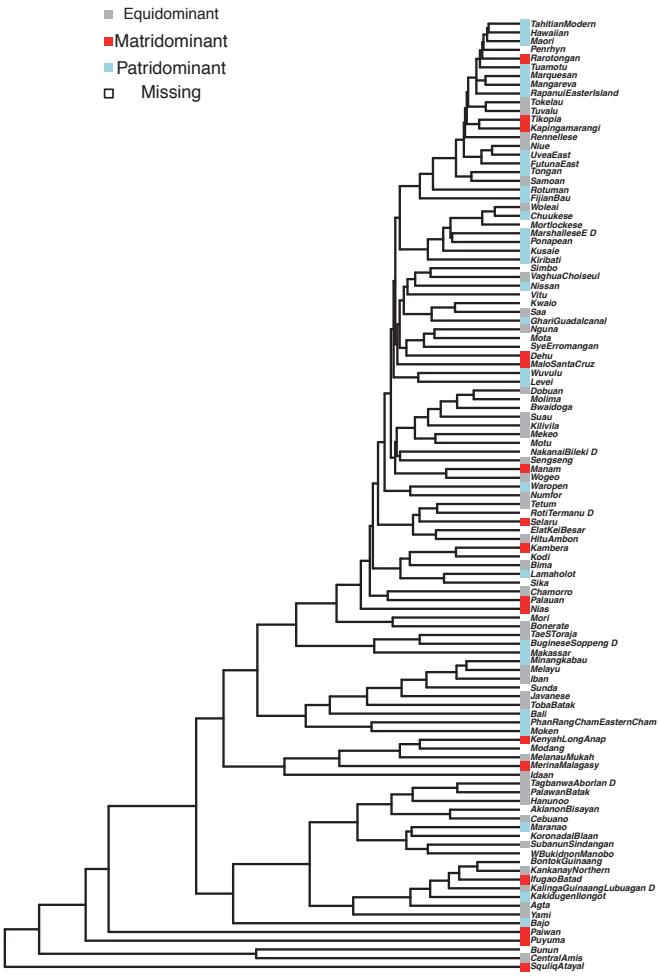


Figure 3–10: Sexual division of labour traits mapped onto the maximum clade credibility tree of the Austronesian societies in the sample. The maximum clade credibility tree is built from a Bayesian posterior of 1000 trees. distribution, while in other sub-clades, the trait is evolving in a phylogenetic manner.

We see that the majority of the traits are patridominant, followed by equidominant and matridominant traits. However, even though the traits are dispersed across the tree, their occurrence is clustered. For e.g.: Palauan and Nias, Paiwan and Puyuma, while these are societies are far apart on the three, the matridominant traits are however clustered within the subclades. This is an example of the prevalent pattern, and might account for the value of D not being significantly different from a phylogenetic distribution or a random inheritance of traits.

3.7.2 Revisiting the basis of sexual division of labour

The fundamental basis for sexual division of labour to exist is the different role social norms have in maximizing the energy invested and the way these norms interact with each other to achieve this (Codding *et al*, 2011). The difference in division of labour is hypothesised to emerge on the differential foraging goals of men and women, in an effort to increase fitness (Elston *et al*, 2014). While women are thought to target risk free provisioning, men target specialised, high risk – high reward resources which are likely to increase their social status. However, it is understood that the energetic constraints or contribution of women to division of labour in subsistence depends on the extent of parental investment that is available, i.e., the amount of kin-cooperation that is available, and this is where residence norms comes into play (White *et al*, 1981). The energy investment is also a function of the type of subsistence, for example, it is thought that even if men and women participate equally in terms of time invested in the subsistence activity, they might perform distinct tasks that enable high efficiency. The prevalent ecological conditions also play an important role in sexual division of labour economics, for e.g.: it is physically more intensive to clear a forest than a savannah to cultivate crops and in this case, it would require the physically stronger sex to invest more energy for a successful subsistence strategy (White *et al*, 1981). Thus, the evolution of the sexual division of labour trait is also inherently dependent on the adaptive strategies a society has to follow to be successful. This is particularly relevant for Austronesian societies, where, as they colonised new areas, encountered varied ecological conditions, had to exist in previously uninhabited land, they had to develop new strategies to adapt. It is undoubted that the effect of environment is substantial on subsistence and the energetics involved to survive.

When the Austronesians reached islands where the environment and climate was possibly unsuitable for them or unfamiliar, they had to adapt different strategies to survive. We see that when faced with adverse conditions, some islands (e.g.: Chatham Islands, Punan hunter gatherers) have reverted back to hunter-gatherer life-styles (Bellwood, 1987; Diamond and Bellwood, 2003). In some islands, they had to abandon traditional shifting cultivations and move to

more intensive agriculture (e.g. Easter Island), this required upland forested regions to be tilled and used for cultivation, while in some islands, they adapted to more environment friendly way of sustenance (Kirch, 2002). Generally, a fallow period of 1 year is followed after 4 years of cropping, but in Futuna, to adapt to the small island size and limited resources, the rain fed gardens were cropped for only 3 years and followed with a fallow period of 10 years (Kirch, 1994). So the subsistence strategies have in each society have completely depended on the environment and the strategy of how to adapt to the environment. A change in subsistence strategy was an eventuality for the expanding Austronesian societies. As discussed in Chapter 2, archaeological evidence in the region pointed to a substantial resource depletion just after the start of colonization of Remote Oceania (AD 1300 – AD 1800) (Anderson, 2002; Field and Lape, 2010), combined with less competition probably fuelled a shift in the principal mode of subsistence in the region away from a dependence on marine resources. This hypothesis is coupled with evidence of people moving from the coast to inland during this time (Anderson and Clark, 1999) and change of subsistence to intensive agriculture (Enright and Gosden, 1992).

A change in subsistence strategy means that there is change in energy investment required. For example, while women are known to participate more in extensive agriculture, men do more work in intensive agriculture (White *et al*, 1981). Another dimension is the changing techniques involved in subsistence, especially agriculture. For example, Bismarcks and Solomons remained forested until modern methods of clear-cutting were adapted. Other factors such as high rainfall, large island size and volcanic ash, also determined how subsistence strategies were adapted (Rolett, 2008). These differences in conditions also meant that each society had to adapt a different strategy to survive, as the use of resources would differ vastly, and the prediction of how a society reacts is a function of the ecological constraint it faces rather than a culturally inherent trait (Atkinson *et al*, 2016). So it would be important to revisit these hypotheses or test evolution of division of labour traits alone, by placing constraints on environmental conditions.

3.7.2.1 Post-marital residence trait evolution

We know that even post-marital trait evolution was hypothesised to be driven by environmental and demographic conditions during the Austronesian expansion, either directly or indirectly. Jordan *et al* (2009) showed that though probably ancestral, the matrilocal states are unstable and a change towards patrilocality is observed in Oceanic societies. Prolonged male absences were posited as one of the main reasons for matrilineality and matrilocality to thrive. The resource depletion (mentioned before), combined with less competition, probably fuelled a shift in

the principal mode of subsistence in the region away from a dependence on marine resources and to the development of intensive agriculture. This change in subsistence resulted in furthering the reduction of male-absences from home, one of the main proponents for the stability of matrilocality (Divale, 1974b; Hage and Marck, 2002; Hage and Marck, 2003; Lévi-Strauss, 1969) and thereby dislodging the main driving force between matrilineality and matrilocality (Hage and Marck, 2003). Demographic events, like depopulation events during migration and adaptation to fluctuating environments have been known to encourage the emergence of ambilocal societies (Lane, 1961). As discussed in Chapter 4, the Austronesian expansion was punctuated by pauses, and a severe bottleneck and depopulation event (loss of life due to environmental risks), before a rapid spread through eastern Polynesia. So the evolution of post-marital residence traits probably does not have a straight forward relationship with division of labour as we thought previously, and is most likely influenced by several other factors.

3.7.3 Summary

I might have approached this test too simplistically, by looking at sexual division of labour and post-marital residence directly, and not controlling for other factors apart from phylogeny. We now know that both these traits are influenced by environmental variables. By not controlling and looking at a more intrinsic model that incorporates all of these variables (ecological conditions, social conditions like marriage ecology, phylogeny), the conclusion is that there is no strong basis to either reject or support either of the proposed hypotheses on sexual division of labour and post-marital residence correlation. Given that subsistence strategies and energetics are clearly linked to environmental conditions, it is imperative to explore the mechanisms of how environmental conditions can affect division of labour traits and residence traits, and then test for a co-evolutionary hypothesis by controlling for phylogeny.

3.8 Conclusion

While the importance of controlling of phylogeny is highlighted, I could not find support for a straight-forward co-evolutionary behaviour of sexual division of labour and post-marital residence traits. It was clear that the dynamics of sexual division of labour traits is much more complex than previously thought. I propose that after accounting/controlling for ecological change, phylogeny and relevant marriage dynamics, the hypothesis of co-evolution of sexual division of labour and post-marital residence needs to be revisited. Also, by testing it in other language families would further provide a conclusive test. The ability to infer the dynamics of cultural traits is prized and gives an insight into the social organization of the past societies, which is crucial in deciphering past history.

4 Testing sex-biased dispersal hypothesis in Remote Oceania: Drift or ancient matrilocality?

"History employs evolution to structure biological events in time."
- Stephen Jay Gould

4.1 Abstract

The patterns of diversity in genetic markers aid in clarifying hypotheses gathered from different lines of evidence (anthropology, archaeology and linguistics). Taking advantage of this "gene-culture correlation", a trend has emerged to use evidence from either the genetic or the cultural state of a society to infer the state of the other traits (Destro-Bisol *et al*, 2004; Oota *et al*, 2001). In Remote Oceania, competing hypotheses on the origin of these societies have emerged (Kayser *et al*, 2006). While Y-chromosome evidence (male history) points to a Near Oceanic origin, mitochondrial DNA evidence (female history) points to an Asian origin. Scholars have tried to reconcile this differential origin of SSM markers by attributing this pattern to be a result of post-marital residence norms. However, differential diversity of sex-specific markers could also be due to stochastic evolutionary processes like drift, or demographic processes like bottlenecks, and population size reductions that may have affected males and females differently. Also, previous studies have not classified the populations into culturally and historically meaningful groups and this could also affect inferences made in the past. In this Chapter, I have tried to address the pitfalls of previous studies and tested different plausible hypotheses against the hypothesis of sex-specific migration due to post-marital residence in a robust coalescent driven framework. Results based on different admixture estimates and a coalescent based framework suggest the following: a) during colonisation of Remote Oceania, a bottleneck affected males and females differently, and this hypothesis was more likely than a conscious sex-biased dispersal process, b) for both males and females, the contribution from Near Oceanic Austronesians was the maximum, and we did not find evidence for a differential origin, c) the timing and quantum of admixture supported the VC Triple I model and, d) segregation of populations into historically meaningful categories plays a very vital role in interpretations of human history.

4.1 Introduction

The phenomenon of sex-biased dispersal (SBD) is known to occur in many animal societies, including humans (Handley and Perrin, 2007). The mechanism underlying this phenomenon in human societies has been discussed in detail in Chapter 2. The main evidence supporting SBD has come from the analysis of sex-specific genetic markers (SSM) (Wilkins and Marlowe, 2006). Often, the mechanistic process driving SBD is attributed to differential origin of men and women in a population, driven by cultural processes like marriage migration (Chaix *et al*, 2007) and post-marital residence (Oota *et al*, 2001). There have been several studies where a better understanding of the dispersal history and the cultural aspects of a society was attained by using genetic data from SSM markers (Destro-Bisol *et al*, 2004; Mesa *et al*, 2000). Remote Oceania is one such part of the world where the origin, history of colonisation and information on social aspects of this region was inferred by heavily relying on SSM patterns (Hage and Marck, 2003; Kayser *et al*, 2006; Kayser *et al*, 2008).

The occupation of Remote Oceania by humans represents the last significant colonisation event in human history (Kayser, 2010). The origins of the ancestors of the present day Remote Oceanians is a matter of constant debate (Kayser *et al*, 2006; Kayser *et al*, 2000; Kayser *et al*, 2008; Lum *et al*, 1994; Oppenheimer and Richards, 2001a). Many theories have been proposed to explain the history of this region and by drawing inferences from different lines of evidence: linguistics, archaeology, anthropology and genetics. The Oceanic region is divided into Near and Remote Oceania based on the arrival of the first humans in the region (Green, 1991b). The first out of Africa migration carried hunter-gatherers into Near Oceania and at least 40,000 YBP and the colonisation of the entire Sahul-Near Oceania (uphill New Ireland) was accomplished by 35,000 – 36,000 YBP (Roberts, 1991). This initial colonisation did not extend beyond the main Solomon Islands (see Chapter 1). The extent of human colonisation in the east did not progress beyond this point, until about at least 20,000 YBP. Linguistic, archaeological and some genetic evidence pointed out that Southeast Asia and the West Pacific encountered new populations, with close links to South Central China and Taiwan about 5,000 YBP. This second wave of dispersal is associated with Austronesian speaking, Lapita bearing, farming communities with developed seafaring technology, which took people beyond the boundary of Near Oceania and into Remote Oceania (Diamond and Bellwood, 2003; Kirch and Green, 2001). There is strong evidence that this second wave of migration is related to the spread of Austronesian speaking people into Near and Remote Oceania (Kirch and Green, 2001). However, the origin, timing and spread of the Austronesians have been highly contentious (See Chapter 1).

Genetics has contributed significantly to understand the origins of Polynesians. The initial studies from genetics came from the investigations of

mtDNA (Melton *et al*, 1995; Redd *et al*, 1995), which indicated an Asian origin for ancestral Polynesians. Later studies with mtDNA (Richards *et al*, 1998) and mainly Y chromosome (Kayser *et al*, 2000) supported a more indigenous, Melanesian origin for the Polynesians. The analysis of the SSM in a comparative framework revealed different male and female origins (Kayser *et al*, 2006). These differential origins of males and females led to the hypothesis of a sex-biased Austronesian dispersal (Hage and Marck, 2003; Kayser, 2010; Kayser *et al*, 2006). Studies revealed that while 94% of Polynesian mtDNA's had an East Asian origin, more than half (66%) Y-chromosomes had a Near Oceanic origin (Kayser *et al*, 2006; Kayser *et al*, 2000). With a strong possibility of the proto-Austronesian societies being matrilocal (Hage and Marck, 2003; Jordan *et al*, 2009), it was reasoned that because of this practice of post-marital residence, the Austronesian women married non-Austronesian men while colonizing Remote Oceania and there was relatively less significant movement of women between the two populations. This sex-biased admixture resulted in the genetic pool of males in Polynesia largely stemming from Melanesia, while the female gene pool could be traced back to Asian origins. In other parts of the world too, differing SSM patterns were attributed to post-marital residence (Destro-Bisol *et al*, 2004; Oota *et al*, 2001 for more detailed review, see Chapter 2).

While there are several drawbacks to the assumptions of sex-biased dispersal without accounting for stochastic evolutionary processes or demographic processes (see Chapter 2), there are also several concerns regarding the methodology and assumptions of previous relevant studies in the Pacific, which might have led to erroneous conclusions regarding the origins and timing of events. The majority of the studies of Pacific populations have concentrated on haplotype inference to assign origins for sample populations. Haplotypic data represents a complex and unique genetic history of an individual, which is not tractable without understanding the process that led to its current state. Any number of random events of drift, bottlenecks, population size changes, admixture and mutation could have led to its current state, but traditional population genetic analyses using haplotypic information do not consider this. A stochastic process called the "coalescent" (Kingman, 1981) provides an apt statistical framework for the analysis of genetic polymorphisms. Just as two haplotypes could have evolved to be similar by descent, they could also have evolved to the same state due to random genetic variation (through mutations) and it is important to incorporate this randomness and test for different possible models of evolution of the two individuals. In case of population demography, the genetic structure of a population is not only determined by population affinities and origins but also by population size fluctuations due to random genetic drift, bottlenecks and population decline. Apparent relationships by different genealogical samples need not necessarily reflect the actual relationship between them. Importantly, the coalescent provides us with a robust statistical framework to test the different

possible models of evolution of a population (Rosenberg and Nordborg, 2002). For the studies in the Pacific, particularly regarding sex-biased dispersal towards Polynesia, there is need to test SBD along with models of drift and bottlenecks to determine whether the affinities noticed today are a result of sex-biased admixture or are due to some other demographic process. For example, one scenario could be a bottleneck during colonisation of Remote Oceania, where there is a loss of genetic variation (bottleneck or drift) and this bottleneck could have affected males and females differently, resulting in sex-biased affinities, i.e., the genetic similarities of males will be to a different group from that of females. Similarly, there are several other plausible scenarios that could result in the current genetic pattern we see in Polynesia, and we need to test these scenarios in a rigorous statistical framework to understand what the most likely scenario is, before arriving at a conclusion on the admixture history of the Austronesians.

Another problem in previous studies is the grouping of populations under the broad category of "Melanesia(n)" (Hage and Marck, 2003; Kayser *et al*, 2006; Kayser *et al*, 2000; Lum *et al*, 2002). It has long been argued that Melanesia is not a cohesive cultural or historical unit (Green, 1991b). The area defined under Melanesia consists of people originating from the first "Out of Africa" migration into the then "Sahul" continent, who are non-Austronesian speakers, and also some Austronesian people from the second wave of migration which led to the colonization of Polynesia. The category conflates geography, language ancestry, phenotype/genotype, and cultural practice, and grouping people as "Melanesians" is not culturally, genetically or historically correct. The distinction between Near and Remote Oceania, coupled with grouping populations under linguistic affiliations into Austronesians and Non-Austronesians proves to be more useful in deliberating the settlement of the Pacific (Green, 1991b). The clubbing of Near Oceanic Austronesians (NO-AN) & Near Oceanic non-Austronesians (NO-NAN), blurs the line between two different migrations and two different time-lines of history and is not useful specifically when trying to disentangle the affiliations and origins of Remote Oceanians (RO). By grouping the two populations together under Melanesians, arguing that Polynesians are closely affiliated to Melanesians proves to be circular, as the affiliation is a historical artefact and has significant implications on understanding the pre-history of a population.

In the current study, by using coalescent simulations under an approximate Bayesian framework, different plausible scenarios were tested against a statistically rigorous backdrop. This provided a solution to incorporate the uncertainty caused by evolutionary and demographic processes in influencing genetic variation, not possible using haplotypic analyses. Modelling past history also provided us with an opportunity to quantify different demographic variables such as: male and female effective population sizes during the colonization of RO; admixture contributions from different populations; population bottlenecks or splitting events; and other variables. Testing for the effective population size of

males and females during the colonization of Remote Oceania gave us an indication of the extent of effect of drift on these populations and also a framework to test whether drift and bottlenecks could have led to the differing genetic patterns of variation we see in SSM.

To address the problem of clustering of populations, the populations were segregated based on the timing of migration and linguistic affiliations, into Asia, Island Southeast Asia (ISEA), West New Guinea (WNG), Near Oceanic non-Austronesians (NO-NAN), Near Oceanic Austronesians (NO-AN) and Remote Oceanians. This was to ensure a robust estimation of history by incorporating known knowledge of affiliations of populations that made biological, cultural and historical sense. Even though Remote Oceania is comprised of populations in Polynesia and Micronesia, populations from Micronesia were not included due to lack of adequate genetic sampling for both mtDNA and MSY in the region. Therefore, for the purposes of this study, Remote Oceania (RO) henceforth is interchangeably used with Polynesia.

4.2 Objectives

In this chapter, I set out to understand the cultural and genetic pre-history of Polynesians and more importantly to test the validity of a sex-biased dispersal hypothesis of Remote Oceanian populations. I collated all relevant genetic information from previous genetic studies on the Pacific and tested different models of plausible demographic history in a statistically rigorous framework. The possibility of a bottleneck or drift leading to the current different genetic variation of SSM versus a sex-biased admixture hypothesis was also tested. This would add to the history of the region and also add to our knowledge of understanding and inferring sex-biased dispersal from SSM. The objectives I set out to test in this chapter are:

a) In Remote Oceania, what colonisation scenarios derived from the study of pre-history most plausibly account for the observed patterning in human genetic data?
b) Is a sex-biased dispersal scenario for Remote Oceanian SSM genetic variation more plausible than a scenario involving bottlenecks or drift?

4.3 Data

To address questions regarding a metapopulation, it is essential to adequately sample across the region, with representation from all possible populations, to make reliable inferences about pre-history. Therefore, all possible genetic data published from the above-mentioned five regions was collated to deduce the most likely model of admixture history of Polynesians stemming from a concrete and

reliable genetic basis. Published data was collected from different sources (Table 4-1 and Figure 4-1) representative of populations in Near and Remote Oceania, Island Southeast Asia (ISEA), Aboriginal Taiwanese and Coastal/Mainland China. Effort was made to collect all the available data from the region.

Populations were grouped based on their historical and linguistic affiliations into the following groups:

a. Asia (Chinese & Aboriginal Taiwanese)
b. ISEA (only Austronesian speakers)
c. West New Guinea (WNG, both highland and lowland – the linguistic origin of the sampled groups is unknown)
d. Near Oceanic non-Austronesians (NAN)
e. Near Oceanic Austronesians (NO-AN)
f. Remote Oceania / Polynesians (RO)

The details of the different populations under each group and their sources are given in Table 4-1. While all possible efforts were made to gather information that would be a good representation of Oceania, there is a distinctive gap in our dataset in the New Caledonian region and as mentioned earlier, in Micronesia. All inferences made from here on are cautious of these facts.

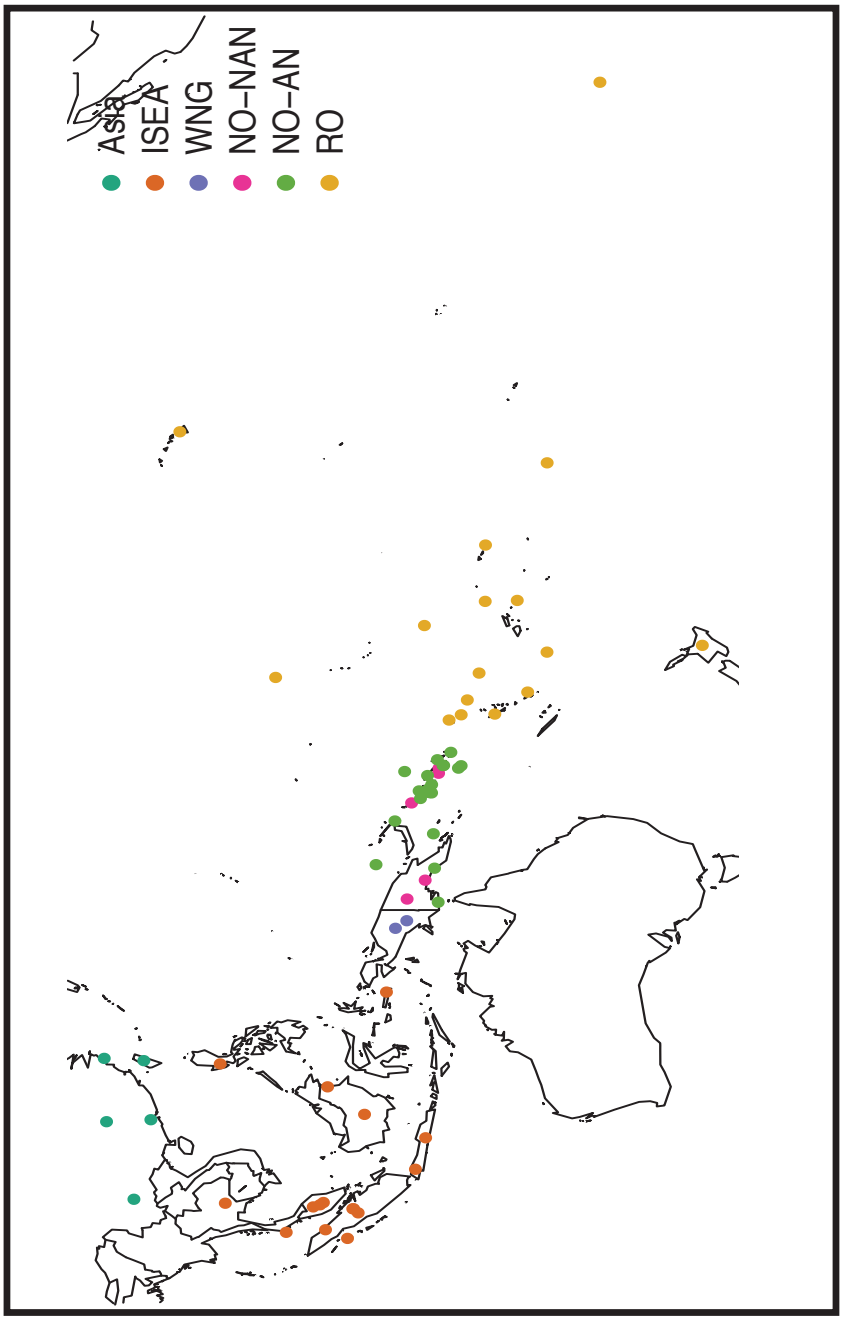


Figure 4–1: Map depicting sampled population groups for both mtDNA and MSY. Asia = Asia & Taiwan, ISEA = Island Southeast Asia, WNG = West New Guinea, NO-NAN = Near Oceanic non-Austronesians, NO-AN = Near Oceanic Austronesians, RO = Remote Oceanians.

Table 4-1: Detailed information regarding mtDNA (HVS-I) samples used in the study, the corresponding sample size for mtDNA and MSY and their sources

Group	Sample Size		Populations	Source/Reference
	mtDNA	MSY		
Asia	1110	140	Bai, Dai, Lisu, Lahu,Oroqen Miao, Zhuang, Han-Chinese, Aboriginal Taiwanese	(Hill *et al*, 2006; Kayser *et al*, 2006; Kayser *et al*, 2000; Lum *et al*, 1998; Tajima *et al*, 2003; Trejaut *et al*, 2005)
ISEA	234	952	Borneo (Barito River), Java, Moluccas (Hiri, Ternate), Adonara, Nusa Tenggara (Flores, Rote, Timor), Philippines, Sumatra (Karo Batak, Riau)	(Hill *et al*, 2007; Kayser *et al*, 2000; Lum and Cann, 1998; Lum *et al*, 1998; Redd and Stoneking, 1999; Van Oven *et al*, 2011)
WNG	227	116	WNG Highlands, WNG lowlands	(Kayser *et al*, 2003; Mona *et al*, 2007; Tommaseo-Ponzetta *et al*, 2002)
NO-NAN	207	243	Kapuna, PNG highlands, Solomons (Savo, Vella Lavella, Bougainville)	(Delfin *et al*, 2012; Friedlaender *et al*, 2007; Karafet *et al*, 2010; Kayser *et al*, 2000; Redd and Stoneking, 1999)
NO-AN	695	768	Admiralties (Andra-Hus, Kurti,Ere Kele,Lele,Mokerang, Nali, Nyindrou, Seimat-Wuvulu, Titan), Bereina, PNG Coast, Solomons (Bellona, Choiseul, Gela, Gizo, GuadalCanal, Isabel, Kolombangara, Makira, Malaita, New Georgia, Ontong Java, Ranongga, Rennell, Shortland, Simbo), Tolai, Trobriand	(Delfin *et al*, 2012; Karafet *et al*, 2010; Kayser *et al*, 2006; Kayser *et al*, 2000; Kayser *et al*, 2008)
RO	543	446	Cook Islands, Fiji, Futuna, Hawaii, Maori, Niue, Easter Island, Samoa, Solomons (Reef Islands, Tikopia, Santa Cruz), Tokelau, Tahiti, Tonga, Tuvalu, Vanuatu, Kapingamarangi	(Delfin *et al*, 2012; Karafet *et al*, 2010; Kayser *et al*, 2006; Lum and Cann, 1998; Lum and Cann, 2000; Whyte *et al*, 2005)
Total	3016	2665		

4.4 Methods

A number of approaches were employed to estimate admixture amongst populations and test various demographic scenarios. These included a) standard measures of genetic differentiation such as F_{ST}, and R_{ST} b) mean pairwise difference to measure genetic variation c) admixture estimates using haplotypic data, with estimator *my* d) gene-flow and genetic structure estimates using the Bayesian partitioning method employed by BAPS and e) approximate Bayesian computation method. This uses coalescent modelling to understand the probability of different demographic histories for Remote Oceanic males and females.

4.4.1 Genetic distance

Genetic distance was calculated using F_{ST} for mtDNA and R_{ST} for MSY. F_{ST} (Weir and Cockerham, 1984) and R_{ST} (Equivalent to F_{ST} for Microsatellites, for more details see (Slatkin, 1995). Indicators were used to determine the genetic distance between the different groups of populations. This determined whether there was significant genetic difference between the populations and if our grouping of populations based on their linguistic and cultural history made sense. Genetic distance is also an indicator of the gene-flow between populations. As F_{ST} is inversely related to migration i.e. $F_{ST} = (1/1+4Nm)$, where Nm is the number of migrants. Calculating genetic distances also gave us an idea regarding the strength of gene-flow and if the mtDNA and MSY markers showed similar relationships amongst the different populations in our dataset.

4.4.2 Genetic variation

The genetic variation in the population was calculated using mean pairwise differences of mtDNA and MSY in each of the populations. The mean number of pairwise difference was calculated using Arlequin (Excoffier and Lischer, 2010). When there are k haplotypes in a population, and if P_I is the frequency of population I and P_J is the frequency of population J and D_{ij} is the number of mutations that have occurred since I and J have diverged (i.e., the number of nucleotides that are different between I and J in case of sequence data), and n is the sample size,

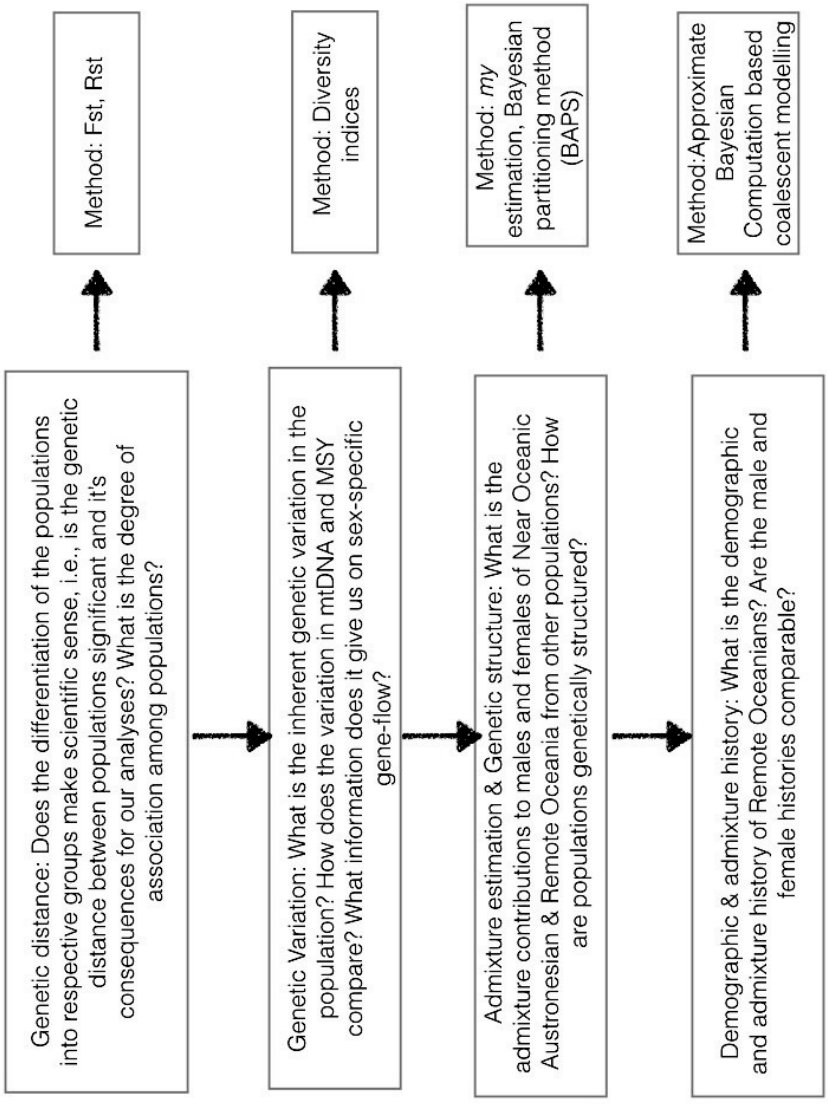


Figure 4–2: Schematic representation of the analyses followed

The mean number of pairwise difference (π) is given by

$$\pi = \frac{n}{n-1}\sum_{i=1}^{k}\sum_{j=1}^{k} p_i p_j d_{ij}$$

Equation-4-1: Equation to determine mean number of pairwise differences (π)

The amount of variation gave us information regarding the sex-specific gene flow, along with its direction, i.e., for example, whether mtDNA harboured more variation than MSY and if so, in which populations.

4.4.3 Admixture estimates

Genetic structure and admixture ratios were quantified by using the admixture estimator (*my*) (Dupanloup and Bertorelle, 2001) and Bayesian analysis of population structure (BAPS) (Corander *et al*, 2008). These analyses were conducted in a two-step manner. In the first step, Remote Oceania was considered to be a population resulting from the admixture between the remaining populations in the data set (Asia, ISEA, WNG, NO-NAN & NO-AN). In the second step, the admixture history of Near Oceanic Austronesians was determined in the same manner (assigning it as a population of admixed descent) after excluding Remote Oceania from the analyses. Remote Oceanians were excluded from the analyses as potential parents to the Near Oceanic Austronesians as they are a much younger population when compared to NO-AN. Also, Remote Oceania, the last and most recent region on the earth to be colonized by humans, was found to have the origin of its immediate ancestors in Near Oceanic Austronesians.

4.4.3.1 Admixture estimation using my

Admixture proportions (*my),* were calculated using the program ADMIX 2.0 (Dupanloup and Bertorelle, 2001). Using this estimator, the relative contribution of parental populations to a daughter population can be quantified using comparison of shared haplotypes and the degree of molecular divergence by specifying the mutation rate at our locus of interest. It was first assumed that the current population of Remote Oceania was an admixture of the remaining group of populations in our sample set (Asia, ISEA, WNG, NO-NAN & NO-AN). In each step, the non-significant contributors to the gene pool of Remote Oceanians were removed till there were only significant contributors remaining. After calculating the admixture estimates of Remote Oceania, we proceeded to quantify the admixture history of Near Oceanic Austronesians in the same manner. Each calculation was bootstrapped 1000 times.

4.4.3.2 Admixture and genetic structure estimation using a Bayesian analysis of population structure approach

Genetic structure of each of the population groups was estimated by classifying the inherent variation of a population into its own variation and variation introduced from other populations through gene flow. The source of variation was identified using a Bayesian approach through the software BAPS (Corander and Marttinen, 2006; Corander *et al*, 2008). In recent years there has been a tremendous increase in the variety of tools that have been developed to assess genetic structure, especially those utilizing Bayesian approaches (Corander and Marttinen, 2006; Corander *et al*, 2004; Corander *et al*, 2003; Dawson and Belkhir, 2001; Falush *et al*, 2003; Pritchard *et al*, 2000). Bayesian model based approaches provided us with a reliable and fast method of assessing the number of underlying putative source populations and therefore were relevant to our analyses wherein we needed to identify the source of variation. This method was used (Corander and Marttinen, 2006; Tang *et al*, 2009) to estimate the amount of admixture in each population. A stochastic partitioning approach was used in this method to assign individuals to genetically distinct clusters and helped in ascertaining the ancestral population in the presence of admixture. By establishing the admixture and gene flow patterns in the population, it was possible to determine if the source of variation in a population is different for males and females. The likelihood of the individual belonging to the cluster assigned is calculated and compared to the likelihood when the individual is assigned to a different cluster. By way of repetitive comparison and iteration, the structure of different clusters in a population is defined. The same approach is applied while evaluating the admixture history of individuals in a cluster and determining the proportion of admixture from each population. Using this Bayesian partitioning method, Remote Oceania was assigned as an admixed population in the first step, and admixture proportions determined. In the second step, Near Oceania was assigned as an admixed population and the genetic contributions from the remaining populations were determined.

4.4.4 Demographic history

To obtain a detailed inference of the evolutionary history of the Remote Oceanians, several demographic scenarios were constructed and analysed using the coalescent (see below) with an approximate Bayesian computation procedure (Beaumont *et al*, 2002) in DIYABC v1.0.4.46 (Cornuet and Ravigné, 2010; Cornuet *et al*, 2008). Most inferences on demographic and evolutionary history are based on genealogical tree building and coalescent theory (Kingman, 1981). The gene pools of the current populations are a representation of only those ancestral

individuals who were successful in passing on their genes to the next generation and by sampling current variation, we are only studying the history of the lineages that have survived. If we envision the genealogical tree represented by our sampled population as a small tree within the tree of the entire population, the larger tree represents the lineages that were present in the ancestral gene pool but did not survive. Understanding the processes that might have led to loss of genetic diversity, and also determining possible genetic affiliations and variations in a gene pool that did not survive or that are not represented by the current genetic variation has a deep impact on our understanding of the history of a population. If a population has undergone a genetic bottleneck or drift and it has resulted in the loss of genetic diversity and therefore loss of a particular line of genealogy, then we do not have any means of tracing this particular line of genetic history. It is a huge effort to understand and predict the number of possible demographic scenarios with the available genetic variation and if we add the onus of predicting it for genealogies that are lost, the task seems insurmountable. However, the main interest here was to understand the possible admixture contribution from different populations to the ancestral gene pool of the Remote Oceanians along with understanding the demographic history. To test whether there was a possible common ancestor between two populations, whose progeny might not have survived to the current generation, is a much more achievable task than to model for all possible demographic scenarios for genealogies that have not survived.

The genealogical process that underlies the history of a population is called the coalescent process (Kingman, 1981). Based on the work of Hudson (1983) and Tajima (1983), Kingman (1981) coined the term "coalescence". The advantage of coalescent simulations is the power to simulate a number of different genealogies with different parameters and history, by randomly dropping mutations across the genealogies and the one critical assumption is that all mutations are neutral i.e., mutations do not influence the probability of reproduction. In the coalescent framework, the probabilities of different events of demography can be understood by estimating the probabilities of parameters like population size change, admixture, migration rate, etc. However, we need a likelihood framework for testing the probabilities of the different events on the genealogies to have resulted in the genetic variation existing currently in our sample populations. These likelihood analyses help determine the set of parameter values and model under which the probability of observing the data sampled is maximum. Studies employing coalescent to understand demographic history have mostly been restricted to full likelihood methods (for example: IM, by Hey (2004)). But in a coalescent framework, the parameter space is large and to reach convergence, even while using the MCMC algorithm, is computationally intensive and if and when convergence is reached, the extent of uncertainty of the posterior is quite large. With an increase in sample size and complexity of scenarios, the probability

of making an inference from the full likelihood method becomes increasingly difficult and therefore, for large datasets (like the present one), a full likelihood method is not an efficient method to understand the demographic history. Approximate Bayesian Computation (ABC) provides a robust alternative, where the likelihood criterion is replaced by a similarity criterion.

In the ABC approach (Figure 4-3),

1. A range of the priors (Φ) is defined for the scenarios we are interested in modelling (Beaumont *et al*, 2002). These priors are population parameters like size, admixture, and mutation rate. Since we used a hierarchical Bayesian approach, the value for each vector of the prior, Φi, (where I =1,2,3... N, N is the number of simulated datasets), is also drawn from a distribution (hyperprior), such that $\Phi_i \sim p(\Phi)$.
2. For the next step, genetic data *D* is simulated by drawing values for parameters from the defined prior range and is given by $\Phi_i : D \sim p(D|\Phi)$.
3. In the third step, D is summarized with a set of chosen summary statistics S_i and our dataset (the actual data) is also summarized, *S'*, by the same summary statistics. The choice of the summary statistics depends on the model defined and the parameters we are trying to extract. In population genetics, most statistics characterize genetic variation and distance and our aim is from these to extract as much information as we can that would represent the characteristics of a population. Care has to be taken to not choose summary statistics that are auto-correlated. For a detailed discussion, please refer to Joyce and Marjoram (2008).
4. In the fourth step, there are two methods that can be employed
    a. Rejection (Pritchard *et al*, 1999): In this step, the summary statistics of the simulated data set S_i, is compared with the actual data , *S'*. Points whose S_i lies within δ from *S'* is then accepted i.e., $|S_i - S'| < \delta$. The points that lie beyond the acceptable δ are rejected. This step provides an estimate of the conditional density of the posterior distribution $p(\Phi| S' = S_i)$. i.e., we chose the parameter vectors Φ_i for which $|S_i - S'|$ is small.
    b. Regression (Beaumont *et al*, 2002): In the second method, a smooth weighting regression is used in the fourth step to evaluate the posterior distribution, i.e., the Φ_i is weighted according to the distance from the actual data by measuring δ and is adjusted by performing a local linear regression.
5. In the final step, the posterior distribution is attained from the retained simulated data after either rejection or regression. i.e., $p(\Phi| S' = S_i| < \delta)$.

Although there has been a debate about using the approximate Bayesian method instead of the full likelihood method to test complex demographic scenarios

(Beaumont *et al*, 2010; Templeton, 2010), we proceed with the ABC method. ABC has proven to be a robust tool to understand and test inferences on demographic and phylogeographic scenarios (Bertorelle *et al*, 2010; Csillery *et al*, 2010). Given the large amount of data in our sample, and sufficient representation of different population groups, it makes for an appropriate tool both theoretically and practically. The results from genetic distance measures and admixture estimations provided us with robust information on population affiliations and an insight into probable admixture histories. Taking into account this information and the known demographic history of Near and Remote Oceania facilitated in constructing appropriate models of demographic history to test with the hierarchical Bayesian method in DIYABC (Cornuet *et al*, 2008).

Two main models are defined, of seven different scenarios each, to understand the demographic history of Remote Oceania. In the first model, West New Guinean population (WNG) is grouped with Austronesian ISEANs, whereas in the second model the WNG is grouped with NO-NANs. As mentioned earlier, the linguistic and cultural affiliations of the societies in WNG, whose genetic data was collected, is not clear. But this genetic information from WNG was indispensable to the model as WNG lies in the crossroads of Oceanic migration and forms an intricate part of the puzzle revealing the history of the Austronesian expansion. Without the information from WNG, our inferences on the demographic history of Remote Oceanians are potentially flawed. Since WNG shows gene flow both to Austronesians in ISEA and NO-NAN, we decided to test the grouping of this population, given the current genetic variation.

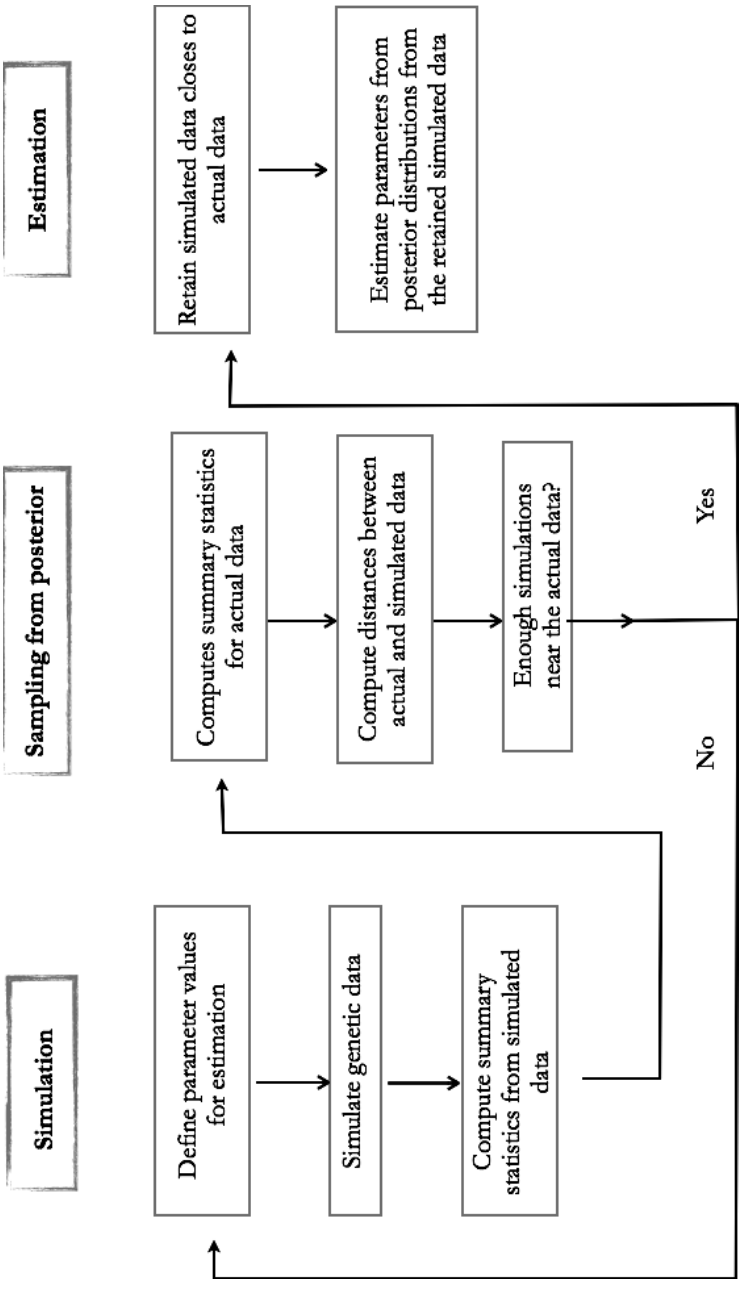


Figure 4–3: Representation of the Approximate Bayesian Computation process. Adapted from Lopes and Beaumont (2010)

For each "scenario" of population history modelled, there were two models, i) WNG is paired with ISEA ii) WNG is paired with NO-NAN. Population parameters like effective population size, timing of each event, population splitting, population admixture events and rate of admixture, population size changes through time (bottleneck etc.,) were defined to describe different scenarios of each model. The range of these parameters were defined for each of these scenarios. The mutation rates per generation. Generation times were defined from Fenner (2005), mutation rates with generation times for mtDNA from Soares *et al* (2009) and for MSY from Zhivotovsky *et al* (2004). Below are the descriptions of the scenarios under each model.

Scenario 1

Scenario 1 is the second most complex of the modelled scenarios in terms of the number of parameters. Here, the admixture history of RO, NO-AN & ISEA was tested. All three populations were modelled to arise from an admixture event. This scenario mainly tests admixture compositions of Remote Oceania, NO-NAN and NO-AN. Here, the hypothesis tests admixture of NO-AN and RO, with ISEA+WNG and NO-NAN, indicating extensive admixture during the dispersal of Austronesian speaking people i.e., in line with the "slow boat to Polynesia" hypothesis (Oppenheimer et *al* 2001) and the VC Triple-I hypothesis (Green, 2003) and competing against the Out of Taiwan hypothesis (Diamond, 1988).

Model 1

ISEA & WNG are combined together as one population. The populations in our dataset are sampled at time *0 (t0)*. Going backwards in time, at *t1,* an admixture event between Near-Oceanic Non-Austronesians & Near Oceanic Austronesians is modelled, at proportions r1 & 1-r1 respectively, which gave rise to the current Oceanic population. Preceding this event, at *t2*, another admixture event involved NO-NAN & ISEA+WNG contributed r2 & 1-r2 respectively to form NO-AN population. At *t3*, Asian & NO-NAN population admixed at the rates of r3 & 1-r3 and formed ISEA+WNG population. At *t4 & t5* population split from the ancestral population (we assumed to be out of Africa), and established ISEA+WNG & NO-NAN populations. At *t6*, from the ancestral population, Asia (including Taiwan) was colonized.

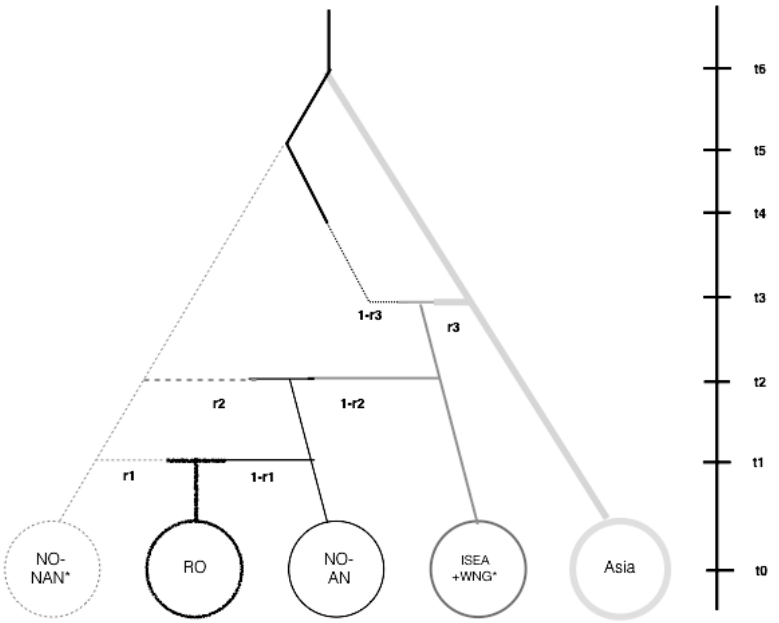


*Figure 4–4: Scenario 1 under model 1. The different colours/patterns indicate that different population sizes were modelled for each lineage and coalescent/splitting event. The timing of each event is defined by parameter t (given on the side scale). In model 1, ISEA & WNG are combined together as one population. *In Model 2, of Scenario 1, WNG is grouped with NO-NAN, instead of ISEA.*

Model 2

In the first scenario under model 2, the only difference from model 1 is the grouping of WNG with NO-NAN instead of ISEA. At *t1*, *r1* represents the admixture contribution of NO-NAN & WNG to RO. At *t2*, *r2* represents the admixture contribution of NO-NAN & WNG to NO-AN.

Scenario 2

In the second scenario, we test the possibility of Remote Oceania arising out of a splitting event from NO-AN, rather than from an admixture event as modelled in scenario 1. Genetic evidence from previous studies point to admixture contributions to Remote Oceania stemming mainly from NO-AN & Asia and hardly from other populations, i.e., with the recent genetic studies on admixture contributions from mtDNA and MSY (Kayser et al, 2006). If this is the case, then scenario 2 should be preferred over scenario 1 for mtDNA. This is also in line with

the Out of Taiwan hypothesis, where the dispersal is said to have happened with very limited gene-flow during the course of colonization (Diamond, 1988).

Model 1

In the second scenario, ISEA & WNG are combined together as one population. The populations in our dataset are sampled at time *t0*. Going backwards in time, at *t1,* the split of Remote Oceania from Near-Oceanic Austronesians was modelled. Preceding this event, at *t2*, another admixture event involved NO-NAN & ISEA+WNG contributed r2 & 1-r2 respectively to form NO-AN population. At *t3*, Asian & NO-NAN population admixed at the rates of r3 & 1-r3 and formed ISEA+WNG population. At *t4 & t5* population split from the ancestral population (we assumed to be out of Africa), and established ISEA+WNG & NO-NAN populations. At *t6*, from the ancestral population, Asia (including Taiwan) was colonized.

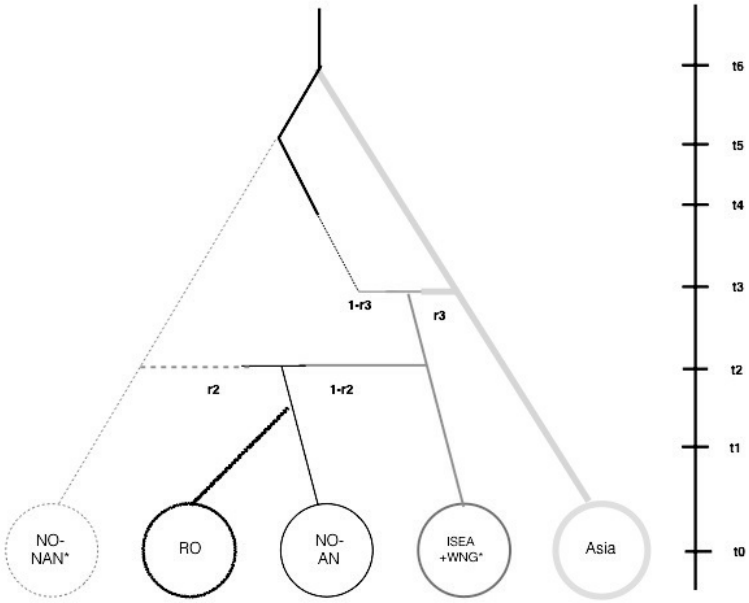


*Figure 4–5: Scenario 2 under model 1. The different colours indicate that different population sizes were modelled for each lineage and coalescent/splitting event. The timing of each event is defined by parameter t (given on the side scale). In model 1, ISEA & WNG are combined together as one population. * In Model 2, of Scenario 2, WNG is grouped with NO-NAN, instead of ISEA.*

Model 2

In the second scenario under model 2, the only difference from model 1 is the grouping of WNG with NO-NAN instead of ISEA. Therefore, at *t2*, the admixture rate of *r2* represents admixture contribution from NO-NAN & WNG, and *1-r2* is the admixture contribution from ISEA. At *t4 & t5*, the split from ancestral population divides NO-NAN+WNG & ISEA.

Scenario 3

In Scenario 3, the admixture history of Near Oceanic Austronesians is tested. The establishment of NO-AN population composition is crucial in determining the admixture contribution to Remote Oceanic populations. It is also important to understand whether the proposed sex-biased admixture occurred during the settlement of Near Oceania by Austronesians or during the Remote Oceanic settlement. If it was only the former, then scenario 2 should be chosen over scenario 3. Here, we are testing if the mtDNA and MSY patterns we see could be i) a result of admixture during the colonization of Remote Oceania vs. ancient admixture and ii) post-colonisation admixture, where during or after the colonization of Remote Oceania, Near Oceanic Non-Austroneisans contributed to the gene-pool of RO, with limited contribution to NO-AN gene pool. The basis of constructing this model is to test whether admixture was during the colonization of RO, or whether during the Austronesian dispersal, admixture was a constant feature. This model also tests the hypothesis for MSY patterns we see, where during colonization of RO, due to the predicted post-marital residence practice, there was indeed high admixture of men into Austronesian speaking RO. If this was true, then the model should be selected for MSY and not mtDNA.

Model 1

In the third scenario, ISEA & WNG are combined together as one population. The populations in the dataset are sampled at time *t0*. Going backwards in time, at *t1*, we modelled an admixture event between NO-NAN & NO-AN at proportions *r1*&*1-r1* respectively, which then gave rise to the current Oceanic population. And unlike scenario 1 & 2, where there is an admixture event, instead at *t2* there is a splitting event of a population from ISEA+WNG leading to the establishment of the NO population. At *t3*, Asian & NO-NAN population admixed at the rates of *r3* &*1-r3* and formed ISEA+WNG population. At *t4 & t5* population split from the ancestral population (assumed to be out of Africa), and established ISEA+WNG & NO-NAN populations. At *t6*, from the ancestral population, Asia (including Taiwan) was colonized.

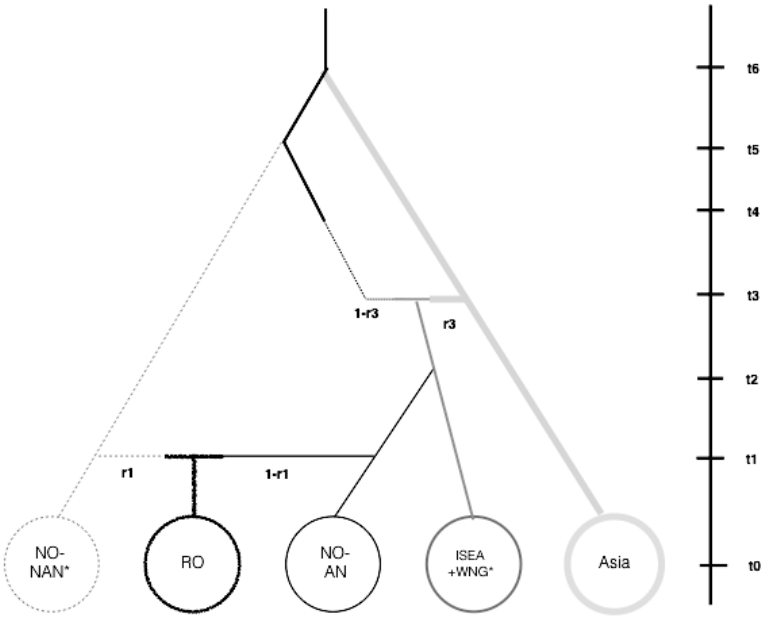


Figure 4–6 Scenario 3 under model 1. The different colours indicate that different population sizes were modelled for each lineage and coalescent/splitting event. The timing of each event is defined by parameter t (given on the side scale). In model 1, ISEA & WNG are combined together as one population. In Model 2, of Scenario 3, WNG is grouped with NO-NAN, instead of ISEA.*

Model 2

In the third scenario under model 2, the only difference from model 1 is the grouping of WNG with NO-NAN instead of ISEA. Therefore, at *t1*, the admixture rate of r1 represents admixture contribution from NO-NAN & WNG, and 1-r1 is the admixture contribution from NO-AN. At *t4 & t5*, the split from ancestral population divides NO-NAN+WNG & ISEA.

Scenario 4

Scenario 4 allowed us to test further the admixture history of NO-AN & RO. In comparison with scenario 2 & 3, if admixture did not occur, but if populations were established by a splitting event during the settlements of both NO-AN & RO, then scenario 4 would be preferred. The hypothesis states that there was a sex-biased gene-flow, and hardly any admixture happened in the female gene-pool. If this was the case, then scenario 4 is expected to be the most likely scenario for

mtDNA data and a different scenario chosen for males. This scenario also is constructed on the basis of the Out of Taiwan hypothesis model (Diamond, 1988).

Model 1

In the fourth scenario, ISEA & WNG are combined together as one population. The populations in the dataset are sampled at time *t0*. The difference between scenario 4 and the previous scenarios is the lack of admixture events during the colonization of Remote Oceanic and Near Oceanic Austronesian populations. At *t1 & t2,* a population splitting event is modelled. At *t3*, Asian & NO-NAN population admixed at the rates of r3 & 1-r3 and formed ISEA+WNG population. The events at *t4, t5 & t6* represented events modelled similar to the previous scenarios.

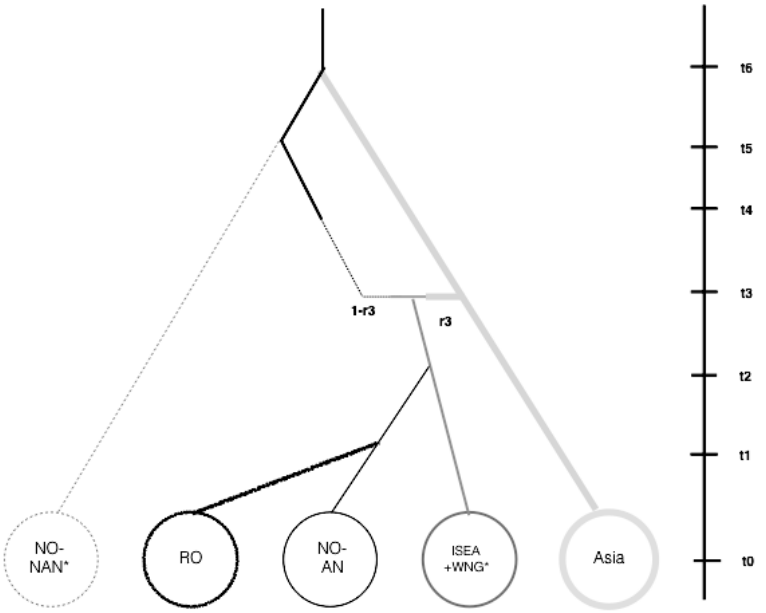


*Figure 4–7: Scenario 4 under model 1. The different colours indicate that different population sizes were modelled for each lineage and coalescent/splitting event. The timing of each event is defined by parameter t (given on the side scale). In model 1, ISEA & WNG are combined together as one population. * In Model 2, of Scenario 4, WNG is grouped with NO-NAN, instead of ISEA.*

Model 2

In the fourth scenario under model 2, the only difference from model 2 is the grouping of WNG with NO-NAN instead of ISEA. Therefore, at *t3*, the admixture rate of r3 represents admixture contribution from Asia, and 1-r3 is the admixture

contribution from Asia to ISEA. At *t5* NO-NAN & WNG split from the ancestral population.

Scenario 5

In scenario 5, the presence of a bottleneck during the colonization of RO is tested against the backdrop of scenario 1. The effects of a bottleneck, a decline in population size, and drift, when there is a loss of genetic variation, could have also contributed to the difference in male and female histories perceived in the Polynesians today. If the bottleneck scenario was true for only either male or female, then it would mean that there is a significant loss of genetic information in one of the sexes. This loss of information could also lead to the differential admixture contribution for males and females that we see today.

Model 1

In the fifth scenario under model 1, ISEA & WNG are combined together as one population. The populations in the dataset are sampled at time *t0*. Scenario 5 is similar to scenario 1, except in the presence of a bottleneck at *t1*. Remote Oceanic population that is formed by the admixture of NO-NAN & NO-AN at *t1,* undergoes a bottleneck through *t*. The remaining events modelled at *t2, t3, t4, t5 & t6* represented events modelled similar to the previous scenarios. *Nb* is defined as the population size change during bottleneck.

Model 2

In the fifth scenario under model 2, the other difference from model 1 apart from the grouping of WNG with NO-NAN instead of ISEA, there is a bottleneck modelled during the colonization of Remote Oceania. At *t1, r1* represents the admixture contribution from NO-NAN & WNG to RO after which, at *t*, there was a bottleneck event leading into the colonization of RO. At *t2, 1-r2* represents the admixture contribution of ISEA to NO-AN. *r3* and *1-r3* represents the admixture contributions to ISEA.

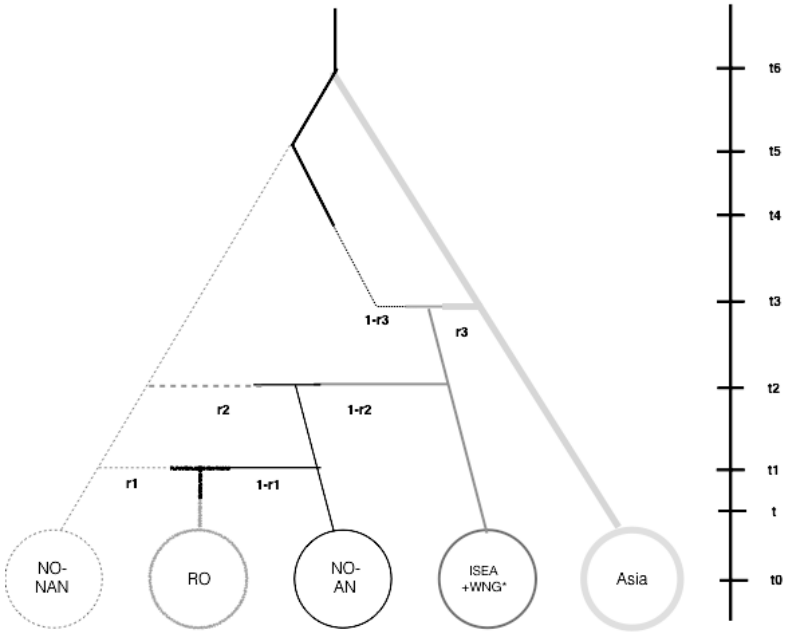


*Figure 4–8: Scenario 5 under model 1. The different colours indicate that different population sizes were modelled for each lineage and coalescent/splitting event. The timing of each event is defined by parameter t (given on the side scale). In model 1, ISEA & WNG are combined together as one population. * In Model 2, of Scenario 5, WNG is grouped with NO-NAN, instead of ISEA.*

Scenario 6

In scenario 6, like scenario 5, the possibility of a bottleneck during the colonization of Remote Oceania is tested. But scenario 6 is similar to scenario 2, in that the test is whether RO was established by a splitting event from NO against scenario 5, where RO is a result of an admixture event between NO-NAN & NO-AN.

Model 1

In the sixth scenario under model 1, ISEA & WNG are combined together as one population. The populations in our dataset are sampled at time *t0*. The scenario 6 is similar to scenario 2, except in the presence of a bottleneck at *t1*. Remote Oceanic population that is split from NO-NAN at *t1* undergoes a bottleneck. The remaining events modelled at *t2, t3, t4, t5 & t6* represented events modelled similar to the scenario 2. *Nb* is defined as the population size change during bottleneck.

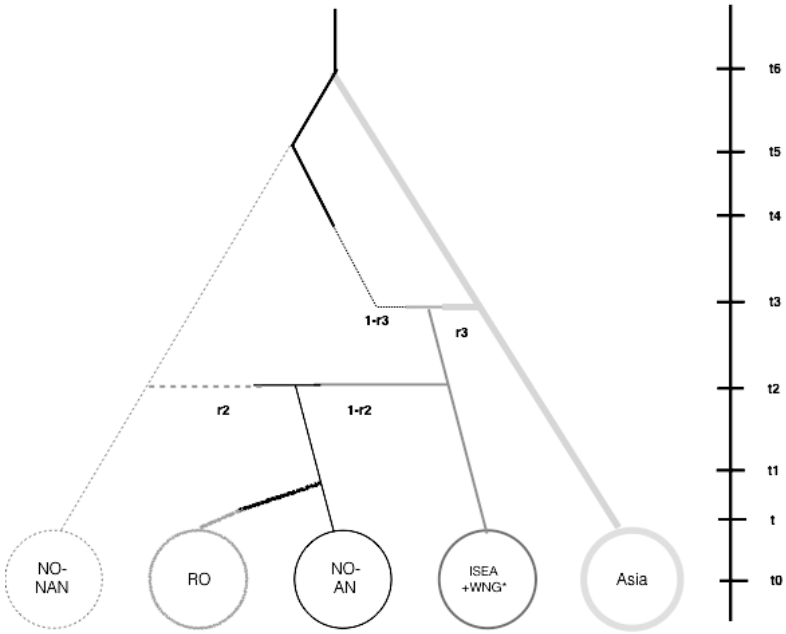


*Figure 4–9: Scenario 6 under model 1. The different colours indicate that different population sizes were modelled for each lineage and coalescent/splitting event. The timing of each event is defined by parameter t (given on the side scale). In model 1, ISEA & WNG are combined together as one population. * In Model 2, of Scenario 6, WNG is grouped with NO-NAN, instead of ISEA.*

Model 2

In the sixth scenario under model 2, the other difference from model 1 is the grouping of WNG with NO-NAN instead of ISEA. There is a bottleneck modelled during the colonization of Remote Oceania. At *t2*, *r2* represents the admixture contribution of NO-NAN & WNG to NO-AN.

Scenario 7

Scenario 7 was to test an extreme model of sex-biased admixture. In this scenario, there are no admixture events with NO-NAN or NO-NAN+WNG, for any of the other populations during to the colonization of Remote Oceania. This was to test the hypothesis of Austronesians colonizing RO, without any admixture with non-Austronesians. But previous studies have shown that in both mtDNA and MSY, there is at least some extent of admixture with non-Austronesians, so this scenario

was not expected to emerge as the most likely scenario for either the male or female history.

Model 1

In the seventh scenario under model 1, ISEA & WNG are combined together as one population. The populations in our dataset are sampled at time *t0*. This scenario was a no-admixture scenario, where populations were established only by splitting events. All the populations except NO-NAN, have split sequentially i.e., ISEA+WNG arose from Asia, while NO-AN split from ISEA+WNG and finally RO was established by a splitting event from NO-AN at *t1*.

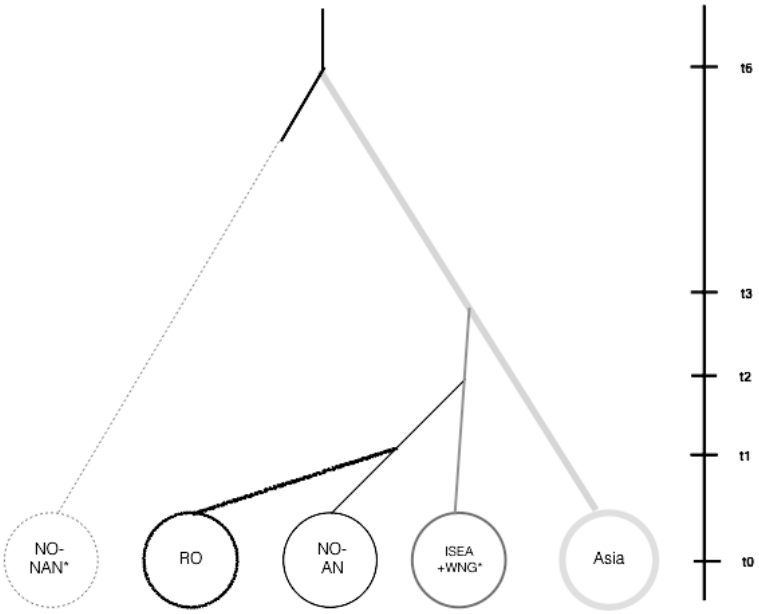


*Figure 4–10: Scenario 7 under model 1 The different colours indicate that different population sizes were modelled for each lineage and coalescent/splitting event. The timing of each event is defined by parameter t (given on the side scale). In model 1, ISEA & WNG are combined together as one population. * In Model 2, of Scenario 7, WNG is grouped with NO-NAN, instead of ISEA.*

Model 2

In model 2 of the seventh scenario, NO-NAN & WNG are grouped together. Here too, similar to model 1, populations are established by splitting events. At *t1*, RO splits from NO-AN and at *t2* NO-AN splits from ISEA and at *t3* ISEA splits from Asia.

Table 4-2: Parameter range values defined for the prior set for simulating 7 scenarios of 2 models. The hyperprior values are all defined under a uniform distribution; Ne is in number of individuals, while time t is given above in number of generations.

Parameter	Details	Range min	Range max	Step
N1	Ne of Asia	10	2000000	1
N2	Ne of NO-NAN+WNG	10	2000000	1
N3	Ne of NO-AN	10	2000000	1
N4	Ne of RO	10	2000000	1
N5	Ne of ISEA	10	2000000	1
N6	Ne of NAN pop in ISEA involed in admixture to form ISEA	10	2000000	1
N7	Ancestral Ne (out of Africa)	10	2000000	1
t1	Time of split /bottleneck in generations	1	500	1
r1	Admixture proportion contribution of NAN to RO	0.001	0.999	0.001
t2	Time of admixture event leading to formation of NO-AN in generations	5	1000	1
r2	Admixture proportion contribution of NO-NAN to NO-AN	0.001	0.999	0.001
t3	Time of admixture event leading to formation of ISEA in generations	10	1200	1
r3	Admixture proportion contribution of Asia, leading to the formation of ISEA	0.001	0.999	0.001
t4	Time of settlement of ISEA by ancestral population (NO-NAN) in generations	10	1500	1
t5	Colonization event where NO-NAN popualtion split from ancestral population in generations	10	2000	1
t6	Time of Out of Africa event in generations	10	3000	1
t0	Colonisation of Remote Oceania in generations	10	1000	1
Nb	Population bottleneck during colonization of RO	10	2000000	1

For all the above scenarios, parameters were drawn from prior vectors, which were defined as a range of values (see Table 4-2). The summary statistics that were used to summarize the simulated and actual data sets were then defined. First, an initial control check was performed by simulating 100,000 runs for each scenario, to see if the datasets simulated were close to the observed dataset and if not, the parameter ranges were re-defined. Using these parameter ranges, 500,000 simulated datasets were produced for all the scenarios. The closest 0.1% (500) and 1% (5000) of simulated data was determined after performing a weighted logistic regression under the same DIYABC program framework, by comparing the summary statistics of simulated and actual data (Cornuet *et al*, 2008). This information was then used for estimating posterior probabilities of each scenario.

Using the posterior probabilities of each scenario, the most likely scenario was quantified (from among the 7 scenarios). The posterior distribution of the parameters under the most likely scenario was estimated by performing a logistic regression using a logit transformation for parameter values, using the same simulated data chosen for estimating the most likely scenario (0.1% of the simulated data closest to the actual data).

4.5 Results

The results for the mtDNA and MSY are described following the schematic in Figure 4-2.

4.5.1 Estimation of genetic distance

Summary:

When comparing mtDNA and MSY genetic distances for RO with other populations, it was found that for both RO-MSY and RO-mtDNA, the maximum affinity was with NO-AN, followed by NO-NAN. For NO-AN as well, for both mtDNA and MSY, NO-NAN was closest genetically. While an increase in genetic distance is seen as geographic distance increases, the only outlier to this pattern for mtDNA is WNG, which shows relatively large genetic distances with all populations. Analysing MSY, an interesting pattern is observed. The MSY of Asians have a very high affinity with ISEAns, the geographically closest population, and then the distance increases till NO-NAN, but then the genetic distance suddenly drops with NO-AN, and thereafter increases with RO. The genetic distance between our populations are significant, reinforcing the classification of populations

based on linguistic history rather than just geographical or political boundaries.

mtDNA:

F_{ST} measures revealed that the maximum genetic distance is between WNG & RO (0.54). Interestingly, the least distance is found between NO-NAN & NO-AN (0.036), followed by NO-AN & RO (0.038). The mtDNA patterns of Near Oceanic Austronesians and non-Austronesians are significantly closer than the remaining population pairs.

Table 4-3: Genetic distance (F_{ST}), calculated for mtDNA between all population groups. All values are significant. (p<0.001)

Population	Asia	ISEA	WNG	NO-NAN	NO-AN	RO
Asia	0					
ISEA	0.055	0				
WNG	0.312	0.306	0			
NO-NAN	0.115	0.103	0.257	0		
NO-AN	0.162	0.157	0.413	0.036	0	
RO	0.265	0.269	0.541	0.129	0.038	0

MSY:

Table 4-4 Pairwise R_{ST} distance between populations based on MSY, depicting the associations between males of different societies. All distances are significant (p<0.01)

Population	Asia	ISEA	WNG	NO-NAN	NO-AN	RO
Asia	0					
ISEA	0.056	0				
WNG	0.153	0.251	0			
NO-NAN	0.056	0.116	0.11	0		
NO-AN	0.034	0.07	0.104	0.007	0	
RO	0.084	0.086	0.234	0.038	0.027	0

Distance measures show that MSY of WNG and ISEAns are genetically most distant from each other, while, NO-AN and NO-NAN MSY are the closest (Table 4-4). It is observed that populations of RO, are genetically most similar to NO-AN, followed by NO-NAN and most distant from WNG.

4.5.2 Estimation of genetic diversity: mean pairwise difference (pi)

mtDNA: The estimation of mtDNA genetic diversity revealed RO & NO-AN to be the least diverse (Table 4-5), and WNG to be the most diverse of all populations in the study. The estimates of diversity from WNG and NO-NAN were comparable.

MSY: WNG MSY were found to be the least diverse, and NO-AN to be genetically the most diverse of all male populations in our study. However, the estimates of diversity from RO and NO-NAN were comparable.

Table 4-5: Genetic diversity (pi), calculated for mtDNA & MSY of all population groups.

Population	mtDNA	MSY
Asia	7.151	10.484
ISEA	6.589	11.736
WNG	7.732	9.705
NO-NAN	7.11	14.574
NO-AN	5.213	17.282
RO	3.373	15.943

4.5.3 Admixture estimates

4.5.3.1 Estimating admixture using my estimator

While NO-AN seems to be the single most significant contributor to the gene-pool of the RO mtDNA, evidence for contributions from ISEA (mainly Nias) and NO-NAN is apparent. For MSY, ISEAn (Nias, 0.24) and NO-AN contributions to the gene-pool of RO outweighed the contributions from the other populations.

Admixture was estimated by using the *my* estimator (Dupanloup and Bertorelle, 2001). In the first step, RO was designated to be of an admixed origin and then in the next step the admixed origins of NO-AN was quantified.

It was found that NO-AN was, in itself, of a highly admixed origin, with NO-NAN contributing significantly for both the MSY and mtDNA gene-pool. A strong genetic signal from Asia (mainly aboriginal Taiwanese) contributing to the mtDNA gene-pool of NO-AN is apparent, with a relatively smaller contribution from ISEA. For NO-ANs MSY, it was observed that after NO-NAN, the highest and comparable contributor was ISEA (with a small portion of it stemming from Nias).

4.5.3.1.1 Step 1: Remote Oceanic admixture

In the first step, where all possible contributors to the mtDNA gene-pool of Remote Oceania were included, it was observed that the ISEANs (0.28), Near Oceanic Non-

Austronesians (0.38) & Near Oceanic Austronesians (0.41), contributed the most. In ISEA, the population from Nias (0.27) contributed the most compared to the rest of the ISEANS. When the non-significant contributions were removed, only contributions by Near Oceanic Austronesians (1.45) to Remote Oceania mtDNA gene-pool remained significant.

Table 4-6: Step 1 of my estimator of admixture contributions for mtDNA and MSY of all populations to Remote Oceania.

Remote Oceania	mtDNA			MSY		
	Step1-a	Step1-b		Step1-a	Step1-b	Step1-c
Asia	-0.23 ±0.02			-0.059 ±0.28		
ISEA	0.28±0.06 (Nias:0.27)	-0.4±0.09		0.45 ±0.28	0.62 ±0.11	0.67±0.05
WNG	-0.15 ±0.02			0.20 ±0.18	-0.003 ± 0.02	
NO-NAN	0.38±0.16	-0.43±0.02		-0.34±0.24		
NO-AN	0.41±0.01	1.45 ±0.17		1.02 ±0.67	0.40 ±0.26	0.33±0.04

Admixture results for the MSY (Table 4-6) revealed in the first step, when all possible contributors to the gene pool of Remote Oceanians were included, that the Island Southeast Asians (0.45), NO-AN (1.02) & WNG (0.20), contributed the most. When the non-significant contributors were removed, and the analyses was run again iteratively till no non-significant contributors were left in the sample, ISEA (0.666) & NO-AN (0.333) emerged to be the most significant contributors to the MSY gene pool of Remote Oceanians. Given the significant contribution of NO-AN to RO, we tested if the NO-AN was in itself an admixed population.

4.5.3.1.2 Step 2: Near Oceanic admixture

The main contributor to the mtDNA gene pool of NO-AN was the NO-NANs. There was also a strong signal of admixture from Asians (mainly Aboriginal Taiwanese) and Island Southeast Asians (Significantly from NIAS). When the Near Oceanic Austronesian MSY was tested for admixture (Table 4-7), there was a strong signal of admixture from Island Southeast Asians (0.48), Nias (0.03) & Near Oceanic Non-Austronesians (0.49). An almost equal contribution of the ISEAn and NAN MSY to Near Oceanic Austronesian MSY was detected.

Table 4-7: Step 2 of my estimator of admixture contributions for mtDNA and MSY of all populations to Near Oceanian Austronesians

NO-AN	mtDNA			MSY	
	Step2-a	Step2-b		Step2-a	Step2-b
Asia	0.203±0.13 (AT: 0.11)	0.07 ± 0.09 (AT: 0.04)		-0.09 ±0.06	
ISEA	0.13 ±0.24 (NIAS: 0.07)	0.28 ±0.09 (NIAS: 0.12)		0.46±0.10 (Nias:0.32)	0.48 ±0.14 (Nias: 0.03)
WNG	-0.29 ± 0.06			-0.13 ±0.055	
NO-NAN	1.02 ±0.09	0.64±0.11		0.65± 0.01	0.49±0.19

4.5.4 *Admixture and genetic structure estimation using a Bayesian analysis of population structure approach*

We found a strong genetic signature of NO-AN and Asian mtDNA in the RO mtDNA gene pool. For MSY it was mainly NO-AN followed by ISEA and NO-NAN. Results from the admixture coefficients of NO-AN, were comparable to admixture estimates by the my estimator, with NO-NAN contributing the most for both mtDNA and MSY. Evidence of gene flow from Asia in both the SSM of NO-AN was observed.

Table 4-8: Admixture estimations made by the Bayesian analysis of partition structure for mtDNA and MSY of RO. The numbers indicate the proportion of admixture from each population (columns), in individuals that we have assigned to populations (in rows). The admixture scenario is where we have assumed that RO is an admixed population, derived from the remaining populations sampled in the dataset. All values are significant ($p<0.05$)

	mtDNA	MSY
Asia	36.3	4.26
NO-NAN	8.2	15.38
WNG	6.5	9.64
ISEA	3.2	16
NO-AN	45.8	54.7

The same procedure of first assigning RO to be of admixed origin and then testing if NO-AN was an admixed population was followed. By assigning RO to be of admixed origin, it is fixed that RO in itself cannot contribute to the gene-pool of the other populations and it could only be the recipient of genes. The admixture contributions by grouping WNG with ISEA in model 1 and with NO-NAN in model 2 was also estimated. This was to maintain consistency and test the

admixture estimates from the approximate Bayesian computation method employed in the next step, as described in 4.4.3.2 and 4.4.4.

Bayesian estimation of population structure results showed that mtDNA of RO is mainly composed of individuals whose putative origin lies in Asia, NO-AN, ISEA and NAN (Figure 4-18, Table 4-8).

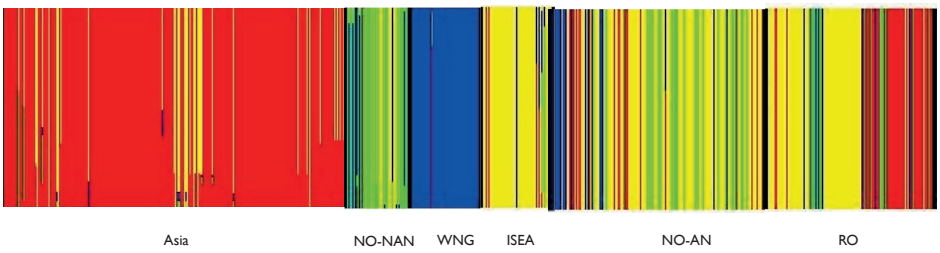


Figure 4–11: : Figure showing the genetic structure of mtDNA of the 6 populations, when NO-AN & RO are considered to be admixed populations, derived from the other populations in the data set. Each line represents an individual, and the amount of admixture is determined by amount of colour filled in each line. Where the p value of admixture estimated is less than 0.05, it is represented with a single colour.

Similar to the admixture results, the main contributor of mtDNA to Remote Oceania is Near Oceanic Austronesians (45%) & Asians (36%) (Table 4-8). A strong genetic signal of Asian ancestry for the mtDNA history of the population is observed. A significant but relatively low exchange of gene flow from NO-NAN (8%), WNG (6%) and ISEA (3%) is also evident.

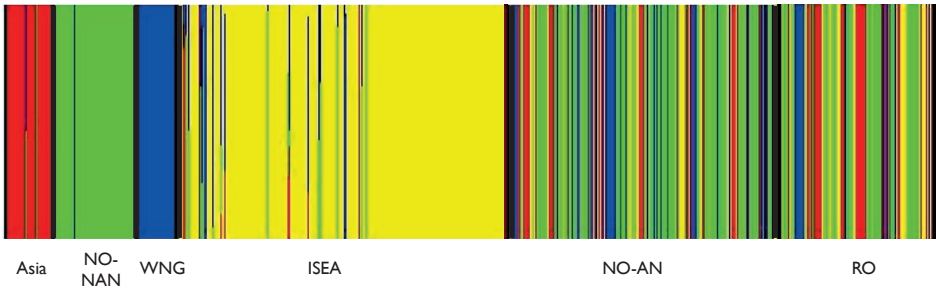


Figure 4–12: Figure showing the genetic structure of MSY of the 6 populations, when NO-AN & RO are considered to be admixed populations, derived from the other populations in the data set. Each line represents an individual, and the amount of admixture is determined by amount of colour filled in each line. Where the p value of admixture estimated is less than 0.05, it is represented with a single colour.

The estimation of admixture proportions in Remote Oceanic MSY (Figure 4–12**,** Table 4-8) revealed that MSY of RO is of a highly admixed origin, with

significant amount of gene flow from Near Oceanic Austronesians, ISEAns, NANs, WNG and even signatures of admixture from mainland Asia. The MSY gene flow from Island Southeast Asia (16%) and Near Oceanic non-Austronesians (15.38%) is almost equal with that from Near Oceanic Austronesians (54.7%) into RO.

Table 4-9: Admixture estimations made by the Bayesian analysis of partition structure for mtDNA and MSY of NO-AN. The numbers indicate the proportion of admixture from each population (columns), in individuals that we have assigned to populations (in rows). The admixture scenario is where we have assumed that NO-AN is an admixed population, derived from the remaining populations sampled in the dataset (except RO). All values are significant (p<0.05)

Population	mtDNA	MSY
Asia	12.2	19.5
NO-NAN	63.5	45.7
WNG	9.8	17.3
ISEA	14.5	17.5

In the case of Near Oceanic Austronesian mtDNA, in line with the results from the *my* estimator, there is a significant admixture of mtDNA from Near Oceanic Non-Austronesians (63.5%) to the Near Oceanic Austronesian populations. There is also a strong genetic signature of Asian contribution (12.2%) to the mtDNA gene pool of Near Oceanic Austronesians.

When the calculations for MSY of the Near Oceanic Austronesian were repeated, and gene flow into NO-AN from the remaining 4 populations (Asia, NAN, WNG & ISEA) calculated, we found a significant amount of genetic exchange with NAN+WNG (45.7%), followed by Asia (19.5%), WNG and ISEA (~17%) (Figure 4–12 and Table 4-9).

4.5.5 Demographic history

As described in 4.4.4, demographic history was modelled in an ABC framework. Seven scenarios under two models were tested, with Model 1 and Model 2 differing only in their grouping of WNG. In Model 1, WNG was grouped with ISEA, and in model 2, WNG was grouped with NO-NAN.

For the 7 scenarios in model 1, only parameters from MSY data could be estimated and not from mtDNA. This was because, the given parameters and grouping of populations and the subsequently simulated datasets under the model were not close enough to the actual dataset and therefore it was not possible to quantify the parameters of the model. Even for MSY, only 60% of the summary statistics of the simulated data lay well within the summary statistics of the observed dataset. However, for both mtDNA and MSY, Scenario 6 (Figure 4–9) was chosen to be the most likely scenario representing the demographic history. The

scenario predicted a bottleneck event during or just preceding the colonization event of RO, which affected MSY (males) and mtDNA (females) differently. The effective population size of MSY emerging from the bottleneck, which lasted from 9980 YBP to 3908 YBP was estimated to be around 3920, under Scenario 6.

In model 2 for both mtDNA and MSY, Scenario 6 was chosen to be the most plausible model of demographic history. An admixture event in this scenario took place during the establishment of the NO-AN population following which, the RO population split off NO-AN following a population bottleneck. In this model, more than 90% of the simulated data fell within the observed data distribution, thereby lending confidence to the inferences drawn from this analysis.

By comparing the results from model 1 and model 2, it is observed that when we group WNG with NO-NAN, we get substantially better results for both mtDNA and MSY. In general, the timeline established from mtDNA is almost twice as old as the MSY timeline. The effective population size of MSY through the bottleneck during the colonization of RO was around 5470 (CI: 3720 – 13000), which was much lesser than the effective population size of females colonizing Remote Oceania where at least 13400 women (CI: 6640 – 52500) were estimated (Table 4-13). This clearly shows that gene pool from which the genetic diversity of the males of Remote Oceania was derived from was much smaller than the female gene pool.

4.5.5.1 Model1

4.5.5.1.1 mtDNA

The coalescent modelling of population demography, using the approximate Bayesian computation method with seven different scenarios under model 1, revealed that Scenario 6 was the most likely. Scenario 6 refers to the model where during the colonization of Remote Oceania there was a bottleneck, and the Remote Oceanic population was a split from the Near Oceanic population, without an admixture event (Figure 4–9). However, the NO-AN population itself was a result of the admixture of NO-NAN and ISEA+WNG. The comparison of the observed summary statistics to the summary statistics of the simulated data revealed that 70% of the summary statistics were in the 0.01 tail of the distribution. Therefore, parameter estimations and model choice performed could not be relied upon to reflect the true demographic history and are not reported here.

4.5.5.1.2 MSY

Under the approximate Bayesian computation framework, possible scenarios of population demographic history for the Remote Oceanians was tested.

Table 4-10: Results from logistic regression using 5000 of the closest points to the actual dataset in estimating the probability of scenarios under model 1 for MSY.

Scenario	Confidence
Scenario 1	0.021 [0.000,0.042]
Scenario 2	0.105 [0.005,0.205]
Scenario 3	0.000 [0.000,0.000]
Scenario 4	0.000 [0.000,0.001]
Scenario 5	0.125 [0.018,0.231]
Scenario 6	0.748 [0.583,0.914]
Scenario 7	0.000 [0.000,0.001]

Following results from logistic regression (Table 4-10), Scenario 2 (Figure 4–5), Scenario 5 (Figure 4–8) and Scenario 6 (Figure 4–9) emerged to be the most likely, amongst the 7 scenarios under model 1. Scenario 2, 5 and 6 are essentially comparable, except in how the population of RO was established. In these scenarios, NO-AN is the result of an admixture event between NO-NAN and ISEA+WNG. Following this, in Scenario 2 and 6, RO splits off from the parental NO-AN population, whereas in scenario 5, it arises from an admixture event from NO-NAN and NO-AN. The difference between scenarios 2 and 6, was the presence of a bottleneck during the settlement of RO in scenario 6. From the results of model choice, Scenario 6 emerged to be 60 times more likely than the rest (Figure 4–13).

While testing the confidence in our model choice, scenario 6 was observed to be chosen 60% of the times over scenario 2 (24.4%) and scenario 5 (15.6%). The estimated summary statistics from the simulated dataset under Scenario 6 showed about 60% of the summary statistics lying well within the distribution of the observed summary statistics, and the remaining 30 % lying in the tails of the distribution.

Since there was a significant amount of simulated data points near the observed data set, it was possible to estimate the demographic parameters under model 1, Scenario 6 (parameters of interest are discussed below). Parameter estimation with 2.5% quantile and 95% quantile & mode under model 1, Scenario 6 are reported (Table 4-11).

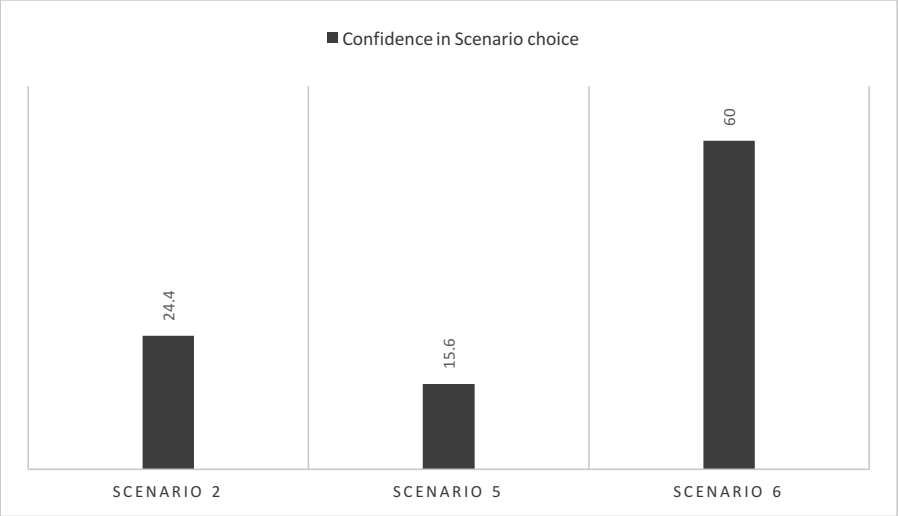


Figure 4–13: Figure showing the confidence in scenario choice estimation under model 1 of testing population demography for MSY. The percentage likelihood of each scenario is given.

Table 4-11: Parameter estimation of variables in model 1, based on the most likely scenario (2), for the demographic history of Remote Oceanian males, using Approximate Bayesian Computation based coalescent modelling.

Parameter	Details	Mode	2.5% quantile	95% quantile
N2	Ne of NAN	961000	252000	1910000
N3	Ne of NO-AN	1940000	431000	1960000
Nb	Pop size during bottleneck	3920	718	428000
N4	Ne of RO	1980000	105000	1970000
N5	Ne of ISEA+WNG	259000	122000	1770000
r2	NO-NAN contribution to the genetic pool of NO-AN	0.727	0.19	0.934
t1	Time of split/bottleneck	9980	8300	10000
t	Time of Colonization of RO, after bottleneck	3980	510	12640

Bayesian modelling showed Remote Oceanic colonisation to be a genetic population subdivision of the Near Oceanic Austronesian population. The NO-AN population was modelled as a genetic admixture of the NO-NAN and a population comprising of WNG & ISEA. The admixture proportion from the parameter estimation shows 49% of the genetic contribution to the male gene pool of NO-AN

derived from NO-NAN gene pool, while the remaining 51 % was from ISEA+WNG. This is similar to our results from admixture estimates (Table 4-7 and Table 4-9), where contribution from ISEA+WNG to NO-AN is more than NO-NAN. The estimation of admixture contribution to ISEAns revealed an almost equal contribution of NO-NAN (43%) and Asians (57%). This showed gene flow across linguistic boundaries during the initial expansion of Austronesians out of Taiwan. The time line estimated from the model parameters establishes a time of about 9980 YBP (CI: 8300 – 10,000 YBP), when there was a bottleneck following a split from NO-AN population. The number of males during this splitting bottleneck event was around 3920. This bottleneck continued till around around 3908 YBP (CI: 510 – 12640 YBP).

4.5.5.2 Model2

4.5.5.2.1 mtDNA

In model 2, the WNG population was coupled with NO-NAN populations. Here too, Scenario 6 emerged to be the most likely scenario (Figure 4–9).

Table 4-12 Table: Results from logistic regression using 5000 of the closest points to the actual dataset in estimating the probability of scenarios under model 2 for mtDNA.

Scenario	Confidence
Scenario 1	0.006 [0.000,0.007]
Scenario 2	0.101 [0.000,0.144]
Scenario 3	0.000 [0.000,0.001]
Scenario 4	0.000 [0.0000,0.001]
Scenario 5	0.121 [0.092,0.284]
Scenario 6	0.771 [0.481,0.973]
Scenario 7	0.000 [0.000,0.001]

When confidence in Scenario choice was tested by logistic regression, Scenarios 5 (Figure 4–8) and 6 (Figure 4–9) had a significantly better likelihood than the other scenarios. Scenario 6 emerged to be ~70 times more likely than Scenario 5. A close contender to Scenario 5 was Scenario 2 (Figure 4–5). The Scenarios 2 and 5 were essentially same, except for the presence of a bottleneck event in scenario 5 during the colonization of Remote Oceania. The demographic model in Scenarios 2, 5 and 6 followed the same events till time *t2*, where an admixture event between NO-NAN+WNG and ISEA gave rise to NO-AN. In the next step, in Scenario 2 and 5, an admixture event between NO-AN and NO-NAN+WNG

gave rise to RO, whereas in scenario 6, RO population split NO-AN without any admixture.

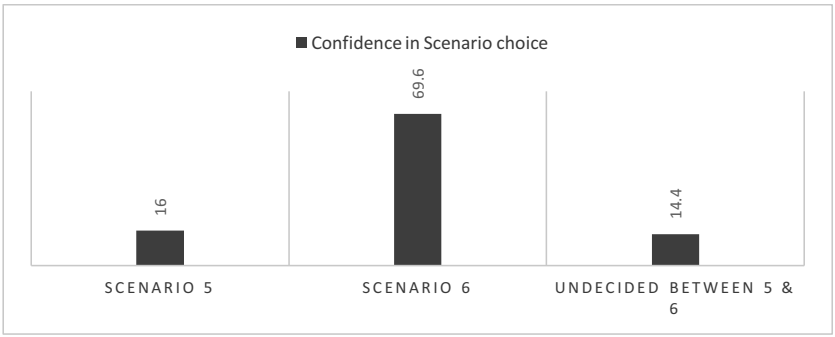


Figure 4–14: Figure showing the confidence in scenario choice estimation under model 2 of testing population demography for mtDNA. The percentage likelihood of each scenario is given

When tested between Scenario 5 and Scenario 6, 70% of the times scenario 6 was more likely than Scenario 5 (Figure 4-14). The estimated summary statistics from the simulated dataset under Scenario 6 showed around 90% of the summary statistics lying well within the distribution of the observed summary statistics. Hence, the simulation of demographic parameters under model 2, Scenario 6 was considered to be robust for parameter estimation (

Table 4-13).

A strong signature of bottleneck appears in the mtDNA history of Remote Oceania. The parameter estimation with 2.5% quantile and 95% quantile and mode are reported below (

Table 4-13). The admixture estimations show that colonization of Remote Oceania was a population subdivision of the Near Oceanic Austronesian population. However, the NO-AN population was most likely a result of genetic admixture of the Asians, NO-NAN & WNG and ISEA. The admixture proportion estimation shows that 76% of the genetic contribution to the mtDNA gene pool of NO-AN was derived from NO-NAN gene pool (& WNG), while the remaining 24% was from the Island Southeast Asians. The results show a significant amount of gene flow between non-Austronesian and Austronesian populations, across linguistic boundaries. This is congruent with the results from the analysis of population structure (Table 4-9), where the NO-NAN (& WNG) contribution to Near Oceanic Austronesians was estimated at 76.4%. The results from *my* estimator, also establish the Near Oceanic Non-Austronesians to be the main contributor to the gene pool of Near Oceanic Austronesians (Table 4-7).

Table 4-13: Parameter estimation of variables in model 2 for mtDNA, based on the most likely scenario (6), for the demographic history of Remote Oceania, using Approximate Bayesian Computation based coalescent modelling.

Parameter	Details	Mode	2.5% quantile	95% quantile
N2	Ne of NO-NAN+WNG	31500	19700	388000
N3	Ne of NO-AN	18600	13300	362000
Nb	Population bottleneck during colonization of RO	13400	6640	52500
N4	Ne of RO	8.73E+05	7.64E+05	9.98E+05
r2	Admixture contribution of NO-NAN+WNG to NO-AN	0.766	0.326	0.911
t1	Time of split/ bottleneck	11200 YBP	6000YBP	12500 YBP
t	Colonisation of Remote Oceania, after bottleneck	9350 YBP	5850 YBP	11600 YBP

The timeline of events established from the coalescent modelling were based on the mutation rate and generation time. The results from mtDNA established that there was a bottleneck before the colonization of Remote Oceania around 11200 YBP (CI: 6000 – 12500 YBP), where it split from NO-AN. The effective population size during the bottleneck was estimated as 13400 (CI: 6640 – 52500). RO recovered from the bottleneck around 9350 YBP (CI: 5850 – 11600 YBP).

4.5.5.2.2 MSY

In model 2 for MSY, we tested the confidence in each scenario and observed scenario 6, followed by Scenario 2 to be contenders for the plausible models for RO's male demographic history (Table 4-14**)**.

As described in the sections above, scenario 6 and 2 differ in the presence of a bottleneck during the colonization of RO, after the splitting event from NO-AN.

Scenario 6 emerged to be 70 times more likely than any other scenario. Between Scenario 2 and Scenario 6, the results of the logistic regression clearly showed Scenario 6 to be the most plausible amongst the two (Figure 4–15**).** From the results of model 1 and model 2 for MSY, it was apparent that the grouping of WNG with ISEA or NAN did not make a difference, and both models chose the same scenario to be the most plausible Scenario for RO males. But with mtDNA, the grouping of populations made a significant difference in the confidence of parameter estimation.

Table 4-14: Results from logistic regression using 5000 of the closest points to the actual dataset in estimating the probability of scenarios under model 2 for MSY.

Scenario	Confidence
Scenario 1	0.0180 [0.0000,0.0374]
Scenario 2	0.2090 [0.0418,0.3762]
Scenario 3	0.0000 [0.0000,0.0001]
Scenario 4	0.0001 [0.0000,0.0002]
Scenario 5	0.0412 [0.0000,0.0828]
Scenario 6	0.7314 [0.5487,0.9141]
Scenario 7	0.0002 [0.0000,0.0007]

The estimated summary statistics from the simulated dataset under Scenario 6 showed approximately 95% of the summary statistics lying well within the distribution of the observed summary statistics, and the remaining 5% lying in the 0.1 tail of the distribution. Since there was a significant amount of simulated data points near the observed data set, it is assumed that the estimation of parameters of the variables in the model are a robust estimation and most plausible representation of the demographic history of the MSY of Remote Oceanians.

The time line estimated from the parameters revealed that around 4720 YBP (CI: 750 – 9340 YBP), there was a population bottleneck of the RO population that split from NO-AN. RO recovered from the bottleneck around 2000 YBP. The admixture estimations showed that Remote Oceania colonisation was like a genetic population subdivision of the Near Oceanic population, the same model chosen to be most likely for mtDNA evolutionary history. The effective population size (Ne) of MSY during the bottleneck was 5470 (CI: 3720 – 13000).

Figure 4–15: Figure showing the confidence in scenario choice estimation under model 2 of testing population demography for MSY. The percentage likelihood of each scenario is given.

The parameter estimation with 2.5% quantile, 95% quantile and mode under Scenario 6 are reported (Table 4-15).

Table 4-15: Parameter estimation of variables in model 2 for MSY, based on the most likely scenario (6), for the demographic history of Remote Oceanian males, using Approximate Bayesian Computation based coalescent modelling

Parameter	Details	Mode	2.5% quantile	95% quantile
N2	Ne of NAN+WNG	224000	97700	1780000
N3	Ne of NO-AN	2000000	598000	1980000
Nb	Pop size during bottleneck	5470	3720	13000
N4	Ne of RO	7880	3770	1110000
r2	NAN+WNG contribution to the genetic pool of NO-AN	0.928	0.381	0.977
t1	Time of split/ bottleneck	4720	750	9340
t	Time of colonization of RO, after bottleneck	2000	508	11920

The admixture proportion from the parameter estimation shows that 72% of the genetic contribution to the male gene pool of Near Oceanic Austronesians was derived from Non-Austronesian & WNG gene pool, while the remaining 37 % was from the Island Southeast Asian. This is similar to our results from admixture estimates (Table 4-7) and from analysis of the population structure (Table 4-9). In

comparison with other populations, in all three models (*my* estimator, BAP and ABC), the amount of admixture contribution to NO-AN MSY from NO-NAN is significantly more than the ISEAns. The admixture contribution to the current population of ISEAns seems to be majorly from the Asians (84%), with 16% from the Non-Austronesian & WNGs.

4.6 Discussion

4.6.1 Contrary to findings of previous studies, coherent clustering of populations revealed similar mtDNA and MSY histories

The aim of this study was to understand the admixture history of Remote Oceanians and their immediate ancestors during the colonization of Polynesia. Data was collected from all possible published studies, where information on geographic and ethnic origin was available. In previous studies where the demographic history and origin of RO populations was inferred from mtDNA and MSY, lack of representation from crucial regions in the Pacific such as the Solomon islands, Vanuatu and New Caledonia, and its implications for inferring demographic history were stressed (Kayser *et al*, 2006). In this study, by filling in most of these gaps, it was discovered, surprisingly, that the MSY (male) and mtDNA (female) histories of ancestral Remote Oceanians did not differ as previously thought (Hage and Marck, 2003; Kayser *et al*, 2006; Kayser *et al*, 2000). The clustering of population groups based on language and culture had a critical impact on the results for modelling demographic histories and estimation of admixture contributions. As mentioned in section 4.1, the clustering of populations as Melanesia and Polynesia, led us to believe that Austronesian expansion was sex-biased. While Polynesia is a sound cultural-historical category (Kirch, 2002), the term “Melanesian” does not make any biological, cultural or historical sense (Green, 1991a). The partitioning of known populations of Austronesian and non-Austronesian origin in Near Oceania, and grouping them into separate categories, revealed intricate processes of the Austronesian expansion and details regarding the colonization process that would have otherwise been unknown. The delineation of these two lines of history, 1) the initial out of Africa expansion and 2) the Austronesian expansion, proved vital in understanding the extent of genetic admixture, the direction of gene flow of males and females and in testing the hypotheses of Austronesian dispersal.

The hypothesis was that since ancestral Oceanians were predicted to be ancestrally matrilocal, during the Austronesian expansion which eventually colonized Remote Oceania, significant male gene flow occurred across linguistic boundaries (between Austronesians and Non-Austronesians) but the female gene flow remained restricted. As discussed earlier, most previous studies clubbed the

NO-AN and NO-NAN populations as "Melanesians," and inferred that the male affinities of Polynesians lie with Melanesia, while the female affinities lie with Asia. Contrary to this, we did not find any evidence for sex-biased affinities for the Polynesians. For both males and females of the Polynesians, their closest affinity is with NO-AN. Also, for both the NO-AN females and males the principal admixture stemmed from NO-NAN. Therefore, one of the important findings of this study was that grouping of populations without prior biological or cultural knowledge affects our understanding of history and it is essential to make a sound decision on how we group populations to have a robust insight into population history.

4.6.2 Differential process affecting mtDNA and MSY: pulses and pauses

From the results of genetic distance and admixture estimates, a substantial male and female gene flow between NO-AN and NO-NAN was observed. Interestingly, when the MSY genetic distance of MSY of Asia was investigated, NO-AN was genetically the least distant (0.034), but the distance suddenly increased for the next geographically proximal population, RO. Even though NO-AN were the immediate ancestors for the RO populations, it seemed that there was only a subset of genetic variation from the NO-AN male population that was passed down to RO males. The diversity estimates also point to a similar effect, probably a founder effect with reduced population size. But when we looked at Asian mtDNA (females), there was only an increasing genetic distance with geographical distance. This points to a different process affecting males and females of NO-AN while colonising RO.

Coalescent modelling predicted that the colonization of RO did not proceed until a NO-AN population was established, i.e., there was a pause in NO before humans voyaged to Polynesia. The splitting event in lieu of an admixture event during colonization of RO points to a long pause by Austronesians in NO, enabling significant genetic exchange with local populations of ISEA & NO-NAN. Even though molecular dating is still not a very reliable technique, in relative terms, the timelines established for colonization of NO by Austronesians and the start of colonization of RO revealed evidence for a long pause in Near Oceania, consistent with studies from linguistics and archaeology (Kirch and Green, 2001).

With mtDNA analyses, the time of colonization of NO by Austronesians was established to be 19,000 YBP (CI: 12,000 – 25,000) and the time of bottleneck from which the colonization of RO proceeded was dated around 11,200 YBP (CI: 6,000 – 12,500), there is clearly a long pause before the colonization to RO proceeded. The dates for the colonization of Remote Oceania from mtDNA, do not fit with the archaeological dates or with dates established from MSY. The timeline established by MSY was much younger, with the colonization event of NO by Austronesians

estimated around 6500 YBP (CI: 2,720 – 17,760) and the time of bottleneck to colonize RO was estimated around 4720 YBP (CI: 750 – 9340) and this fits in well with the archaeological dates. The difference in male and female timelines could be attributed to the difference in the rate of mutations of the two SSM (Scally and Durbin, 2012). The MSY mutates almost 100 times faster than mtDNA, and it was found that mtDNA has a two-fold deeper time coalescence and two-fold larger effective population size (Ne) than MSY (Wilder *et al*, 2004). Therefore, the bottleneck process that we see for mtDNA could well be the process preceding the colonisation of RO, but the effect of which is still observable in mtDNA of RO. If we accounted for the two-fold deeper coalescence of the mtDNA time-line, then both the mtDNA and MSY time lines were remarkably similar and seemed to align with the time-line established by the Out of Taiwan model.

4.6.3 Drift followed by extensive gene-flow from NO is a more plausible history for RO than sex-biased admixture

The next step was to look at admixture contributions for males and females of RO, to ascertain if there was sex-biased gene flow during the Austronesian expansion to Polynesia. In the current analyses from all three methods, a strong genetic signature of Asian ancestry for mtDNA was observed. The genetic admixture from NO-NAN females was significantly higher than the contribution from Asia, for both NO-AN and RO females. When looked at haplotypic analyses alone for mtDNA, using the *my* estimator, a clear genetic signal from the Taiwanese population in Asia was evident (the proposed homeland for Austronesians) to RO. However, the same analyses for males(MSY) revealed no indication of admixture contribution from either Asia or populations representative of the Austronesian expansion, like Nias (Van Oven *et al*, 2011).

When the NO-AN populations were tested further for admixture, substantial contribution from populations of Austronesian affiliation (Nias and Asia) for both the SSM was discovered. The lack of a strong Asian genetic signal in RO MSY could be a result of bottleneck during the colonization of RO. The results from our coalescent analyses, genetic distance and variation statistics and admixture estimations, strongly supported the presence of a bottleneck and subsequent drift, during the colonization of RO, which might have affected males and females differently. For example, there might have been a selective loss of male individuals carrying Asian ancestry. Drift is indicated, because the effective population sizes of males and females involved in the colonization of RO were different from each other.

Effective size of a population is indicative of the effect of drift on a population, the smaller the effective population size, higher will be the effect of drift (Charlesworth, 2009). The effective population size of MSY and mtDNA

estimated through the ABC analysis was 5470 and 13400 respectively. Even if the two-fold difference between males and females (Wilder *et al*, 2004) was accounted for, the effective population size of females was larger than that of males. Lower effective size of males would have meant that there was a higher reduction in the genetic diversity of males during the bottleneck and ensuing colonization of RO. The random process of drift could have eliminated males of largely Asian ancestry and this is where, males having a genetic signature from Taiwan could have been lost. NO-AN MSY clearly showed Asian (Taiwanese) genetic signature and somewhere between NO and RO, the individuals of this particular descent have been lost. This was also seen in the genetic distance measures, while the distance between NO-AN and Asia was relatively smaller, but the genetic distance between RO and Asia underwent a sudden increase. Though essentially Near Oceanic Austronesians colonized RO, the sudden increase in genetic distance indicated that due to selective mechanisms (bottleneck and effect of drift), only certain lineages survived, which skewed the ancestry of the individuals colonizing RO towards NO-NAN. Therefore, no evidence of a sex-biased dispersal to RO was found and both males and females showed the maximum affinity to NO-AN, with traces of NO-NAN and Asia.

4.6.4 Forces driving drift to affect differently on MSY and mtDNA

The results from previous studies (Hage and Marck, 2003; Kayser *et al*, 2006; Kayser *et al*, 2000) propose a largely "local" origin for males of RO, we propose that males and females of RO have originated from the same ancestral population, and it was more likely that the bottleneck affected males and females differently, resulting in the diverging genetic affinities of MSY and mtDNA of RO established from haplotypic analyses. There are several mechanisms that would have led to this phenomenon.

4.6.4.1 Austronesian custom of primogeniture

Firstly, the Austronesian voyaging was probably propelled by the Austronesian custom of primogeniture, i.e., only the first born could inherit the chieftainship and this would have resulted in the younger sons venturing out and seafaring in search of new lands to settle in, as a more constructive outcome of their ambition than agitation or fratricide (Fenney, 1996). They would have gathered a voyaging crew and several followers and set out to explore new islands to establish their territory. Most of these voyages first had an exploratory phase, wherein new lands to inhabit were sought out and once suitable dwelling was found, they returned back to their original inhabitation to carry back with them a larger troupe. During

the exploratory phase, the sailing would have continued for many days without sighting land or food, but if they searched eastward using the westerly wind shifts, they would have eventually been able to reach home before their supplies ran out. This made searching eastwards a better tactic than searching west, thereby shaping the direction of the voyaging and colonization move eastwards. Once the explorers found a suitable uninhabited island for settlement, they would have been able to come back and organize a larger colonizing expedition to settle the newly discovered island (Irwin, 1994). These exploratory voyages probably consisted mostly of men, and there was a high chance of perishing at the high seas before they could actually navigate and settle to an island. Archaeological, linguistic and genetic evidence points to a long pause before/during the colonization of RO. This was evident even in our analyses, where there was a strong bottleneck during the initial phase, which lasted for about 2000 years.

4.6.4.2 Lack of seafaring techniques – perilous to the voyaging sex

Linguistic and archaeological evidence points to an initial expansion into Remote Oceania, where the Eastern Lapita societies colonized the Western Polynesian complex (WPC: Fiji, Tonga, Samoa and the nearby islands), followed by a long pause before voyaging further east (Kirch, 2002; Kirch and Green, 2001). This pause could have been due to the lack of sea faring technologies to navigate the rough waters which were much more dangerous than previously traversed. Beyond Fiji, the trade winds were sporadic and also the direction of the mean wind-flow only deflected to the Northeast, instead of reversing the westerlies (not making a complete circle). This meant that the waters were much more treacherous to traverse than during the initial phase of colonization and there would have been a considerable loss of life. For example, there is a recollection in the oral traditions in the Cook Islands, that around 100 canoes left in search of new lands, and only one canoe survived (Allen and Johnson, 1997). These recollections arose from actual events, and we can only assume that there were repeated voyaging attempts to colonize new islands of which only a few of them turned successful. It was long before the two-rudder canoes developed which would have helped navigate the waters. Therefore, there is a clear evidence for a pause after the start of colonization of RO, most probably in the WPC.

Evidence from genetics (Wollstein *et al*, 2010), commensal studies (Matisoo-Smith, 1994; Matisoo-Smith and Robins, 2004) and bacterial studies (Falush, 2003; Moodley *et al*, 2009), also show that eastern Polynesians were founded from the WPC. Archaeological evidence pointed to the WPC (Futuna to be more precise) as the homeland for the transition of eastern Lapita culture into Polynesian (Kirch, 2002). Recent genomic studies also point towards a large degree of admixture post the start of colonization process from Near Oceania

(Skoglund *et al*, 2016), and this could have well happened in WPC. Evidence from linguistic phylogenies also pointed to a long pause in WPC, before a rapid expansion east (Gray *et al*, 2009), to the boundaries of the Polynesian triangle. If all this were true, then there was an even more severe founder effect than initially expected. The repeated voyages attempting to colonize the rest of RO, resulting in loss of human life (mainly male), compounded by the fact that most of eastern Polynesia was colonized from the western Polynesian complex, which itself had a small gene-pool, probably resulted in the survival of a selective male lineages. Therefore, ample support is seen for the results from this study, that a bottleneck during the colonization of RO probably affected males and females differently and this effect was compounded by subsequent founder effects. Given that the greatest affinity of RO was observed to be with NO-AN, it is strongly believed that the key to answering the probability of sex-biased Austronesian dispersal probably lay with Near Oceanic Austronesians, rather than with Remote Oceanians.

4.6.5 Admixture history of mtDNA and MSY of Near Oceanic Austronesians revealed high levels of gene-flow with Near Oceanic Non-Austronesians of both sexes

To trace the ancestors and origin of Austronesians, it was important to look at the history of NO-AN rather than RO, as RO contained only a subset of genetic diversity of NO-AN following a bottleneck during the colonization event.

The NO-AN population admixture estimations revealed significant genetic exchange with ISEA, WNG, Asia and NO-NAN. NO-NAN and Asia emerged to be the largest contributors to the gene pool of male and female NO-ANs. For both the sexes we observed that the contribution of NO-NAN was significantly greater than the Asian contribution. The MSY genetic pool was more diverse than the mtDNA genetic pool, as evidenced by admixture results and genetic diversity measures (Table 4-5, Table 4-6 and Table 4-7). Also, from coalescent-based estimators, the admixture contribution of NO-NAN+WNG to the NO-AN gene pool was 0.76 to mtDNA and 0.92 to MSY. A greater amount of gene flow of men between the two societies was observed and this could probably be attributed to the matrilocal post-marital residence form of ancestral Oceanians (Jordan *et al*, 2009).

However, when we look at it from the perspective of Austronesian dispersal, the affinity of both men and women lay closer to NO-NAN than the proposed sex-biased affinities of males with NO-NAN and females with Asia (Table 4-6 and Table 4-8). If the Austronesian expansion was in fact strongly sex-biased, then we should see, as detected by previous studies, a very strong Asian contribution to the female gene pool over-shadowing the contribution from NO-NAN. Contrary to expectations, we observed significant female gene flow between the two linguistic groups. The coalescent models also tested for restricted female

gene flow by modelling scenario 7 (Figure 4–10), which would have been selected as the most likely scenario, if there was restricted female gene flow, but we found no support for this model (Table 4-10), and it was in fact estimated as one of the least likely models. If there was a sex-biased admixture history, two different models should have been selected for male and female demographic history (with and without admixture of NO-NAN to NO-AN/RO), but we found the most likely demographic history chosen, with a high degree of confidence, for both the sexes was the same.

The hypothesis put forward by Kayser *et al* (2006) and Hage and Marck (2003) equate the disparities in the origin of the SSM to post-marital residence practices. The findings of the previous studies can be attributed to two main reasons: 1) the previous studies were based on an erroneous classification of the NO groups into Melanesia and not into cultural-historically sensible groups of NO-Austronesians and NO-non Austronesians; and 2) the Austronesian expansion was fairly rapid and the longest pause was probably in Near Oceania. This would explain why diversity increased with genetic distance, when compared to females, from the proposed Austronesian homeland (Asia/Taiwan), whereas the NO-AN group in itself did not show such a pattern. As the Austronesian expansion progressed rapidly, there was very little time for the voyaging group which eventually reached NO to integrate to a great extent with the locals. Instead, the Austronesian groups which settled and now inhabit these regions (Asia and ISEA), had ample time to integrate and probably in a sex-biased manner, as would have been expected, given their ancestral matrilocality, supporting the VC-Triple I hypothesis. However, in Near Oceania, there were probably other mechanisms at play which drove sex-independent gene flow between Austronesians and non-Austronesians.

4.6.6 Sex-independent gene-flow in Near Oceania

Currently, the post-marital residence practices in Oceania are quite diverse (Fortunato and Jordan, 2010; Murdock, 1940). From cultural phylogenetics, it was determined that Proto-Austronesian populations, were most likely to be matrilocal than patrilocal (Jordan *et al*, 2009). This seemingly related well to the hypothesis on sex-biased dispersal to RO. If the ancestors were indeed matrilocal, why did we not find evidence of a sex-biased dispersal? As discussed in the earlier section, the expansion from Taiwan to NO probably proceeded quite rapidly. When they reached Near Oceania, they possibly adapted by integration with the already existing non-Austronesian inhabitants to survive successfully. The results from analysis of cultural data also indicates that while the Proto-Austronesian was reconstructed as matrilocal, the reconstruction of Proto-Oceanic node had a higher probability of being patrilocal than matrilocal . Proto-Oceanic is the ancestor for all the Austronesian languages spoken in the Oceanic region

(consisting of Near Oceania, Micronesia and Polynesia). This shows that there was probably a shift from ancestral matrilocality to patrilocality in Oceania and hence the extensive gene flow of both males and females between AN and NAN during this stage. This also supports the "integration" aspect of the VC-Triple I hypothesis.

Some of the reasons for a shift from matrilocality to either patrilocality or ambilocality was proposed due to a depopulation event (Ember, 1974; Service, 1962) like migration. Migration events are essentially depopulation events, as a small group of individuals have to establish themselves in a new territory and survive. Ambilocality provides the perfect setting for such flexibility. In these societies, post-marital residence was probably decided based on primogeniture, rather than sex, i.e., the first born irrespective of the sex would remain with the natal family (Murdock, 1949b). When this is the case, there is no restriction on gene flow of either sex. We see evidence for this in Polynesia, where there is a prevalence of ambilocality (Jordan, 2007)

Apart from residence, the inheritance of wealth and descent systems also play a major role in determining the movement pattern of males and females in a society. Austronesian and mainly Oceanic societies largely followed an ambilineal form of descent, where individuals could choose to affiliate with either the matriline or patriline (Firth, 1957). The choice maybe based on the advantages and obligations that are associated with membership to that particular group. Another form of descent is the bilateral descent system, where members of both matrilines and patrilines could be present. This system is widespread in ISEA and in some societies in RO, like the Maori (Firth, 1957; Scheffler, 1964) and these societies are referred to as having a social organisation called the "House society" (Lévi-Strauss, 1969). The transmission of wealth and resources is strictly organised based on these house societies (Kirch and Green, 2001), and the rules vary depending on the type of behaviour specified, for example, inheritance of land, inheritance of property, residence etc., and these also varied among the different house societies. Though bilateral descent encompasses different modes of transmission (based on the house society), it is not as flexible as ambilineality. One proposition could be that during the Austronesian expansion, when the largely bilateral Proto Malayo-Polynesian societies came across the strongly patrilineal and patrilocal non-Austronesians (Brown, 1978), they found it more beneficial to turn to a descent system like ambilineality or for that matter matrilineality or patrilineality, that allowed flexibility to choose the kin group, and to enjoy the associated benefits (Bellwood, 1997). Jordan (2007) found the ancestral node for Proto-Oceanic to be Patrilineal, with some weak matrilineal aspects originating and being lost multiple times. Coupled with a shifting residence practise to ambilocality, and a change from bilateral descent to unilineal descent, provided an opportunity for individuals in the NO-AN society to adapt and maximize benefits through association, and thereby provided a favourable atmosphere for sex-independent gene flow. These results were not surprising, as

flexible cultural norms that allow for adaptation of different features was essential for the survival and perpetuation of a migrating population like Austronesians into NO & subsequently to RO (Vayda and Rappaport, 1963).

4.6.7 Triangulation

The proposed sex-biased dispersal theory is largely based on the assumption that ancestral RO were matrilocal and matrilineal (Hage and Marck, 2003) and this tied in with the genetic evidence from previous studies (Kayser *et al*, 2006) of contrasting origin and admixture histories for mtDNA and MSY. But studies since then using cultural phylogenetics (Jordan, 2007; Jordan *et al*, 2009) have shown that while the ancestral Austronesian societies were most likely matrilocal, there was no support for ancestral Oceanic societies being matrilocal and in fact, there is evidence to suggest that there might have been a shift from the ancestral state of matrilocality to ambilocality or patrilocality and evidence also points to a shift from a more rigid bilateral system to a flexible unilineal descent system. Both these shifts provide the flexibility for sex-independent gene flow across societies. This tied in well with the results from the current study, where a clear indication of significant male and female gene flow across linguistic boundaries was found, between AN and NAN societies in Near Oceania, before the colonization of RO. Using historical and cultural knowledge to group population made a significant difference in the interpretation of genetic results. Another important finding of the study is that appropriate groupings of societies, based on foundations of sound knowledge on the cultural and social past of a society are critical to making robust and historically accurate interpretations of pre-history.

Coalescent modelling with genetic data revealed that both mtDNA and MSY underwent similar demographic histories, and did not find any support for sex-biased origin of Remote Oceanians. Further, it also showed the presence of a bottleneck during the colonization of Polynesia, which affected males and females differently. The cultural practice of primogeniture coupled with the risks involved in voyaging supports the hypothesis of a differential risk for males and females. These differential effects could have led to the pattern of sex-biased affinities we see today.

Genetic signal from Asia was still strong in NO-AN and RO populations, indicating a support for the Austronesian homeland in Taiwan, in conjunction with hypotheses emerging from linguistic data analysis (Gray *et al*, 2009; Gray and Jordan, 2000). While the timeline established from molecular clocks is still not very reliable, support for the previously proposed pauses during the Austronesian expansion, using linguistic data, (Gray *et al*, 2009; Kirch, 2002; Kirch and Green, 2001), after an initial rapid expansion phase from the ancestral homeland before reaching RO, is observed based on the admixture results from our study. This

supports a theory where during the pauses, there was integration with local societies. Recent genomic results from ISEA (Lipson *et al*, 2014) also provide support for an extensive admixture of Austronesians in ISEA. Similarly, in the case of NO, where we see extensive gene-flow between NO-NAN and NO-AN, and NO-AN and ISEA, supporting the VC-Triple I hypothesis.

4.7 Conclusion

By collating genetic data from across the entire range of Austronesian societies, and following a culturally and historically sound grouping of societies, I was able to address most of the drawbacks of the previous studies in elucidating the demographic history of the Remote Oceanians. By employing a statistically rigorous coalescent framework, aided by other genetic analyses, it was possible to deduce that during the colonisation of Remote Oceania, a bottleneck that affected males and females differently was more likely than a conscious sex-biased dispersal process. Using cultural and historical accounts, and thereby, drawing on triangulating evidence, ample support for a sex-independent gene-flow in Near Oceania was observed. By finding support to the VC Triple I model, the results from this study integrated findings from comparative anthropology and linguistics and brought together the different lines of evidence, to give a unique perspective on the Austronesian expansion.

5 Tracing the evolution of the Dravidian language family

5.1 Abstract

The Indian sub-continent lies at the cross-roads of human migration, with several waves of human migrations traversing it. In such a setting, understanding human history using molecular tools becomes extremely challenging, as the admixture history muddles any robust inference garnering. In such a scenario, linguistics can play an important role in understanding events that have led to the current state of human diversity. In societies where there is no clear signal of human history, and where archaeology and genetics has failed to delineate zones of contact and clearly differentiate different waves of human expansion, linguistics can throw light on past processes. In this Chapter, I take advantage of the evolutionary nature of language to understand the evolution of the Dravidian language family, and test different hypotheses regarding its origin and subsequent spread. Several hypotheses put-forth regarding the sub-grouping of the language family using phylogenetic methods are also tested. This study is the first attempt at inferring Dravidian language phylogeny by using contemporary linguistic data in a quantitative evolutionary framework. The position of certain languages within the language family and a relative chronology to the spread and diversification of this language family were also determined. Results from this study suggest that Brahui is not an ancient lineage as was previously thought, and is most likely to be the result of a recent divergence, probably during the spread of Indo-European into the sub-continent.

5.2 Introduction

The Dravidian language family is a small language family centred on Southern India and consisting of about 80 languages (Hammarström *et al*, 2016; Lewis *et al*, 2009). Out of these languages, only four are major literary languages with a substantial history of written literature and a distinct written script. While these four languages are geographically restricted to southern India, the remaining minor literary and non-literary languages have a wider spread, from the Nilgiri hills in the south, to the central Indian plains, northern India, with Brahui a geographical outlier in the Balochistan region of Pakistan (Figure 5-1). These languages range from being spoken by small language communities (Vishavan, 150 speakers), by far larger communities (Kodava, 200,000 speakers), to global

languages with literary histories that go back hundreds of years: Malayalam, 33 million speakers, Kannada, 38 million speakers, Tamil, 61 million speakers, and Telugu, 74 million speakers.

The influence of the Dravidian speakers is visible through literature, art, and culture throughout the world. For example, the Pallavas, a Tamil dynasty, carried and spread their writing system on their voyages to the east, where it forms the basis for several writing systems of Southeast Asia and beyond: Myanmar, Thailand, Laos, Cambodia, and as far as Sulawesi and the Philippines (Steever, 1997). Several other Dravidian speakers have contributed to literature through other languages and their own. For example, India's best known philosophers, Sankaracharya, Ramanuja and Madhvacharya, who developed the *advaita*, *visista advaita* and *dvaita* philosophies, were from South India and spoke Dravidian languages, but spread their philosophy by writing in Sanskrit (an Indo-Aryan language).

5.2.1 Study, Origins and Spread

The term "Dravida" was first used by Caldwell, the Irish missionary/linguist, to describe the people of Southern India (Caldwell, 1956) and remains generally accepted for the description of this language family, as it is mostly restricted to the southern part of India. Only about 30 years after the concept of language family was popularised by Jones (1786), Dravidian was recognized as a distinct language family, separate from that of the sympatric Indo-European language family by Ellis (1816). Until Ellis' discovery, these languages were considered descendants from Sanskrit; this was possibly why these languages did not receive any special interest or consideration at that time. Ellis not only recognised the presence of the Dravidian language family, although he acknowledged borrowing from Sanskrit, he also presented evidence regarding the relationships between the different languages within the language family (Krishnamurti, 1969). Tuttle (1927) worked on the reconstruction of proto-Dravidian and held that links between the Dravidian languages and Sanskrit have existed since Rig Vedic times. This work on the links to Sanskrit ignited interest in the language family, especially in North America. Following this reconstruction, and based on Caldwell's (1956) comparative grammar, the work on minor literary and non-literary languages, which had till then received minimal attention, picked up pace. Emeneau, an eminent scholar from Yale, worked and published grammars of Toda (Emeneau, 1938) and Kota (Emeneau, 1944), non-literary minor languages which till then did not have any reliable grammar.

The origin and spread of the Dravidian language family has been linked to several other societies and language families such as Uralic (in particular Finno-Ugric), Elamite, Japanese, Indo-European, and languages of Indus valley and Harappa civilisations. Caldwell (1956) linked the Dravidian language to the Scythians, and

in particular the Finno-Ugric branch of the Uralic language family. Burrow (1969) investigated the hypothesized link between these two language families by comparing 72 etymologies referring to body parts. Since then several linguists (Bouda, 1956; Krishnamurti, 1969; Menges, 1964; Menges, 1969; Zvelebil, 1970) have reviewed this theory, and proposed a counter explanation, in terms of prolonged contact between Dravidian and Finno-Ugric which could have led to diffusion; i.e. their similarity is not necessarily due to shared ancestry (Zvelebil, 1990). The foundation of this link was laid by showing parallels in some selected features between languages of Dravidian and Finno-Ugric, rather than Figure 5–1equating the proto-systems of the different languages which is common practice (Krishnamurti, 1969). Although some form of contact relationship between

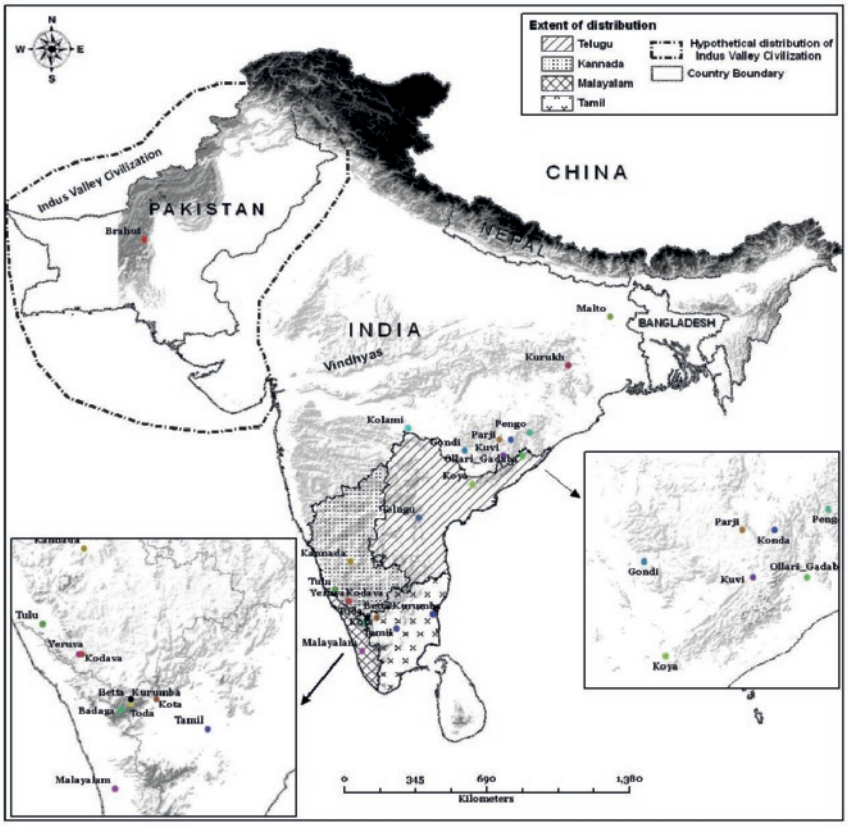


Figure 5–1: Distribution of Dravidian languages. The shaded regions show the distribution of the four extant major literary languages. The other langauges are represented as point locations, as their current distribution is not concretely known. Also represented are the Indus valley civilization and the Vindhya mountains, places of interest in the history of Dravidian languages.

Dravidian and Uralic remains a promising lead, proof is still needed regarding the nature and historical context of this relationship.

Another well-known hypothesis was proposed by Campbell (1998), who suggested the South Caspian area as a contender for the centre of origin for the Dravidian family, and proposed links between the Elamites and the Dravidians. McAlpin (1974) put forward several lines of comparison between Dravidian and Elamite, among which was the agglutinative structure of the noun and the use of plural markers. However, most of the evidence presented in support of this relationship was *ad hoc,* and the comparisons between the two language families were based on a small dataset of Elamite originating in the sixth century BCE, by which time major Dravidian languages were already thought to have split. Krishnamurti (1997) found several important cognates between Elamite and proto-Dravidian, like the verb 'to die' in proto-Dravidian and Middle Elamite.
Krishnamurti (2003) later noted that while McAlpin (1974) compares Dravidian to Elamite, and found cognates for 81 items out of a corpus of 5000 lexical items, the similarities in these 81 word forms could be attributed to chance. While links to Japanese (Ōno, 1980) and Nostratic (Pedersen, 1903; Krishnamurti 2003) have been proposed, these hypotheses have never gained momentum.

Campbell (1998; 1999) suggested that the Dravidian languages were native to the Indus valley civilization (Figure 5-1) and proposed the centre of origin to be somewhere in Northern India, as we can still find Dravidian languages in scattered and isolated pockets of northern India. There is no particular archaeological date or horizon associated with a putative Dravidian "entry" of language and/or speakers into India. Krishnamurti (2003: 5) states that it is best to consider the Dravidians "natives of the Indian subcontinent who were scattered throughout the country by the time the Aryans entered India around 1500 BCE". There is linguistic evidence for a much wider dispersal of the Dravidian languages than found today, with most famously Marathi, an Indo-Aryan language, spoken in West-Central India, having a Dravidian substrate (Southworth, 2005: 288, but see Kulkarni-Joshi, 2012).

The Indo-Aryan language family is a sub-division of the Indo-Iranian language family, which is a branch of the Indo-European language family. The nature of evidence of the relationship between Indo-Aryan and Dravidian is speculative, as the archaeological evidence is uninformative and most of the crucial period of this interaction was during pre-history. While there is no evidence to show where and when exactly the Dravidians entered India, the interaction between Indo-Aryan and Dravidian is crucial for understanding the

distribution and evolution of the Dravidian language family. Krishnamurti (1978; 1985; 2003) supported the principle of the Dravidians being widespread throughout north India and Aryans entering India at a later stage. Zvelebil (1970; 1972) linked the history of the Harappan (Indus Valley) culture to Proto Dravidian and thus supporting arguments for a widespread Dravidian language in the Northern part of India. The Indus civilization ended before the Aryans entered India, and the seals recovered from sites such as Harappa showed that the language contemporaneous with Indus was a non-prefixing language like Dravidian. There was no linguistic evidence that the speakers had knowledge of fauna like lions, camel or rhino (based on lack of any expression for these fauna in proto-Dravidian), unlike Aryans. Knowledge of other native Indian fauna like tigers and elephants is commonly recorded in the Indus valley seals as well as being reconstructable in proto-Dravidian, but which is however conspicuously absent in early Aryan languages (Thapar, 2001). This indicates some support for Indus valley civilizations speaking a Dravidian language. Sjoberg (1992), examined the nature of structural and lexical changes in Dravidian and Indo-Aryan and concluded that, given the interaction between the two language families for over three millennia, the amount of structural change that Dravidian has undergone is relatively small when compared to the Indo-Aryan languages. This relative difference in the nature and amount of structural or lexical change has been attributed to the difference in distribution ranges and timing of entry of these language families in India, as explained below.

If the Dravidian languages had a distribution that was spread throughout northern India, then why are they limited largely to the southern part of the subcontinent in the present day? Krishnamurti (1985; 2003) proposed that the prevalence of Indo-European languages in the North and the restrictive distribution of the Dravidian to the south have largely to do with the political scenario during the Indo-Aryan expansion. On his account, the current substrate of Indo-Aryan languages in India is largely Dravidian. There were many instances of lexical borrowing from Indo-European to Dravidian and Munda language families. Clear evidence of Dravidian loanwords in Indo-Aryan is found in the middle Rigvedic period (c. 1200 BCE) in a source area that might have been Sindh, contemporary Southwest Pakistan (Southworth, 2004; Witzel, 1999). Southworth (2004) posits Sindh, Gujarat, and eastern Maharashtra as areas where Dravidian would have been spoken at earlier stages. Krishnamurti (2003: 35-42) goes further and suggests a Dravidian substratum for Middle Indo-Aryan and all modern Indo-Aryan languages, suggesting that these languages arose when Dravidian speakers merged with the Aryan society and learned their language. However, Dravidian speakers were clearly not the first inhabitants of the subcontinent. Southworth (2005) discusses foreign words and features including retroflex consonants that are not native to Old-Indo-Aryan, and neither are they borrowed from Dravidian or Munda (Afro-Asiatic). They originate in the South

Asian linguistic interaction zone that predates the entry of the Indo-Aryan languages into the subcontinent and that includes languages that have left no descendants we know about.

Krishnamurti (1985; 2003) proposes that since the incoming Aryans politically controlled the native Dravidians and Munda tribes, the Indo-European languages were probably instilled as a second language (L2). He proposed that, if the speakers of a particular language (first language speakers, L1) are compelled to speak another language as their "lingua franca", then a third language (L3) would develop with the structural features from L1 and lexical features of L2. It was evident that middle Indic had a Dravidian substratum, with lexical features of Indo-European. He further substantiated this claim by giving the example of English spoken by Indians in the 20th century, where the English (L2) has a large number of structural features of the Indian languages (L1) and we see a significant amount of borrowing of the English lexicon into Indian languages, but with no evidence of structural borrowing. With this, he concluded that the ancestors of the current Indo-Aryan speakers were probably originally Dravidian or Munda speakers (Caldwell, 1956). The Dravidian speakers did not migrate south, but rather adapted the Indo-Aryan languages in a language shift. This scenario explains the current distribution of the Dravidian languages as well as the existence of a Dravidian substratum in middle Indic.

In line with the hypothesis of the existence of a historically large spread of Dravidian language, is the claim put forth by Fuller (2003), based on archaeo-botanical and linguistic data. This hypothesis argues that the hypotheses proposed till date assumed that Dravidians were agriculturists (McAlpin, 1974; Southworth, 1976), but Fuller (2003; 2007) argues that reconstructed vocabulary points towards some practices that must have preceded agriculture amongst hunter-gatherer groups with traditions of wild-seed and tuber use. This theory posits that Dravidians were a "Mesolithic" society with technology for de-husking, grinding and storage. With several terms for millets and tubers native to central India, reconstructed to proto-Dravidian, central India was proposed as the likely homeland for proto-Dravidian, from where division of the language family progressed towards the current distribution of South Dravidian and North Dravidian in two waves, prior to the spread of agriculture. According to Fuller, the arrival of Indo-European and agriculture drove the already diverging language families further. This proposal suggests that the status of Brahui in the west is the result of this westward expansion, rather than it being the first branch diverging from proto-Dravidian based on the hypothesis of a westward entry.

Given that there is strong evidence supporting a larger presence of Dravidian language family in the north of India in the past, Zvelebil (1970) claimed that Brahui, which is spoken in the north-western border of the Indus plain, represented the most ancient lineage and the first split of Proto-Dravidian. However, Krishnamurti (1978; 2003) argued that this was highly unlikely given

that Brahui does not have most of the features of Proto-Dravidian, like the dental-alveolar-retroflex contrast in the stop series or the lack of voice contrast among the stops, which are some of the extant features in the Dravidian languages. Zvelebil (1990) himself later backed away from this proposal of a northwest entry and concluded "All this is still in the nature of speculation. A truly convincing hypothesis has not even been formulated yet".

The momentum of work on Dravidian languages gained pace after the independence of India in 1947. American linguists trained several Indian scholars in modern and historical linguistics, which led to an increase in the descriptive work on the Dravidian languages. Emeneau (1938; 1944), worked on the languages of the Nilgiri region (Kota and Toda) which were until then not described, while Burrow and Battacharya (1953; 1960; 1961; 1963; 1970), worked on several Dravidian languages in central India (Gondi, Manda, Pengo, Parki, Kui and Kuvi). Burrow and Emeneau (1961) published the first Dravidian etymological dictionary in 1961, which has since then proved to be indispensible for comparative studies and for work on understanding the evolution of the Dravidian language family itself. As a result, the first comparative analysis of the language family was described in Krishnamurti (1961), where he also reconstructed many proto-Dravidian roots and formatives. This work validated the subgroupings of the Dravidian language family and discussed many important features of the language family. These include the major sound changes that occur within the Dravidian language family as well as more focused details on some languages, such as the two step sound changes in Gondi, where there is a change from $s > h > \phi$ amongst the Gondi dialects (Krishnamurti, 1998), and a clear description of the structural similarities that Indo-European has with Dravidian. He further described the pattern of lexical diffusion within the Dravidian language family, and innovations that led to establishment of subgroups.

5.2.2 Distribution and subgrouping

The initial subgrouping of the Dravidian language family was based on glottochronological studies by Andronov (1964). Andronov used cognate data from Swadesh's 100-word list (Swadesh, 1952) of the known Dravidian languages and grouped them into **Southern Dravidian**, consisting of Tamil, Malayalam, Kannada and Telugu; **Central Dravidian**, containing Kolami, Parji and Gondi; and **Northern Dravidian**; consisting of Malto, Kurukh and Brahui. These three groups are geographically structured, with South Dravidian restricted to Southern India, Central Dravidian spread across the central Indian plateau and Northern Dravidian having a spotty and isolated distribution across Northern India, including the language isolate Brahui in Balochistan. There were several drawbacks of the methodology used for this work, which were compounded by

errors in collection of data and thereby leading to questionable inferences. One example: for the meaning "all" in the Swadesh list, the lexical difference of this word for humans and non-humans was not taken into account while collecting data and this choice affected the cognate coding with other languages. Another example: the Telugu word given in the list for the entry "foot" is "*padamu*", a known borrowing from Sanskrit, whereas the commonly used word is "*adugu*" is cognate with Tamil, Malayalam and Kannada (Krishnamurti, 2003). Such erroneous data coding led to incorrectly inferring relationships and historical time depths. Krishnamurti (1961), revised the three sub-groups of the Dravidian language family based on comparative phonology and morphosyntax and included Telugu in Central Dravidian, instead of South Dravidian as previously proposed by Andronov. The details of this classification are given below:

- ***South Dravidian:*** Tamil, Malayalam, Irula, Kurumba, Kodava, Toda, Kota, Badaga, Kannada, Koraga, Tulu
- ***Central Dravidian:*** Telugu, Gondi, Konda, Kui, Kuvi, Pengo, Manda, Kolami, Naikri, Parji, Ollari, (Kondekor) Gadaba
- ***North Dravidian:*** Kurukh, Malto, Brahui

This proposal was widely accepted, until Krishnamurti (1961) proposed new evidence for a Telugu-Manda subgroup separate from Central Dravidian. For this he lists the following shared innovations exclusive to this subgroup: a set of atypical sound changes; word-initial apicals, consonant clusters with apicals as second members, the loss of PD *c>s>h>∅, and the distinction between human and non-human plural forms.

He re-designated this branch as South II Dravidian or South-Central Dravidian and rechristened the erstwhile South Dravidian as South I Dravidian. He argued for a common origin for South I Dravidian and South II Dravidian, i.e., Proto-South Dravidian based on evidence from shared innovations. These included sound changes particular to the subgroup, where Proto-Dravidian **i *u* transformed to **e *o* before a low vowel, and morphological innovations such as the development of paired intransitive and transitive stems with –(N)P/-(N)PP alternations in verbs. This version of the subgrouping is currently widely accepted. In this version of the subgrouping, there is a ternary branching proposed from the Proto-Dravidian node. Proto-South further splits into South I Dravidian and South II Dravidian, followed by the second branch which leads to Central Dravidian and the remaining branch delineating the Northern group of languages (Figure 5-2a). However, Krishnamurti notes that a binary branching is more likely to occur than a ternary division, and proposed an alternative scenario, where the first division defines Proto-South-Central from Proto-North Dravidian (Figure 5-2b)

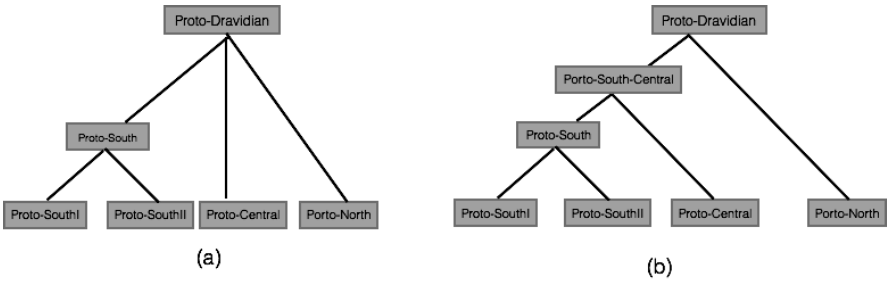


Figure 5–2: Figure indicating the two models of sub-branching of the Dravidian language family proposed by Krishnamurti (1961, 2003)

Apart from the currently accepted four subgroups, Southworth (1976) proposed seven subgroups besides North Dravidian. Drawing on McAlpin (1974) hypothesis that Dravidian and Elamite are related and that the origin of the Dravidian speakers lay somewhere near Mesopotamia, Southworth (1976) proposed that Dravidian speakers entered South Asia sometime around the third millennium BCE. From this, he proposed a new subgrouping, but these seven subgroups largely reflect the different sub-groups within each branch of Krishnamurti's proposed four branches and do not actually depict any new groupings of languages. The origin and subsequent delineation of these different subgroupings continue to be a matter of debate, and are constantly being revisited (Rama *et al*, 2009; Southworth, 1976; Winters, 2007) to understand their manner and mode of evolution, and to substantiate the standings of different sub-groups. There are still several remaining unanswered questions.

5.2.3 Molecular evidence

India is at the cross roads of modern human migrations (Cann, 2001) and is populated by inhabitants who speak four different language families : Austroasiatic, Sino-Tibetan, Dravidian and Indo-European. Molecular evidence from mtDNA supports the theory that the Dravidian speakers represent the first modern human migration into India (Cavalli-Sforza *et al*, 1993; Kivisild *et al*, 1999; Kivisild *et al*, 2000; Kumar *et al*, 2008; Lahr and Foley, 1994; Majumder, 1998; Metspalu *et al*, 2004; Quintana-Murci *et al*, 1999). However, Renfrew (1992) contends that it is the descendants of Austroasiatic speaking people who account for the first settlers of India, and that later agricultural dispersal can be accounted for by the Dravidian and Sino-Tibetan languages. These hypotheses are based on the presence of the haplogroup M of the mtDNA in Dravidian and Austroasiatic speaking groups, which is considered to be a subset of the L3 haplogroup found in Africa (Kivisild *et al*, 1999; Kivisild *et al*, 2000; Quintana-Murci *et al*, 1999). The frequency of this particular haplogroup is high in Dravidian speakers (Kivisild *et*

al, 2003) and is also present at a substantial frequency in Austroasiatic populations (Bamshad *et al*, 2001; Basu *et al*, 2003). However, the coalescence age of this mtDNA haplogroup, the M, is comparable in both the Dravidian and the Austroasiatic (Kumar *et al*, 2008). So, there is no way of delineating if Austroasiatic speakers were already present in India, when Dravidians entered India or vice-versa and if there was substantial gene-flow leading to admixture between the two groups. Genome wide data indicates that Dravidian speaking groups are as distantly related to Austroasiatic/Sino-Tibetan as they are to Indo-Europeans (Reich *et al*, 2009). It was found that the Dravidian speaking populations bear very little admixture signatures with the western Eurasians or central Asians. In contrast, the Indo-European speakers exhibit high admixture signatures with the Western Eurasians, and this cline of admixture increases as one moves from South to North in India (Reich *et al*, 2009).

The Y chromosome data shows a substantial gene flow of Western Asian and European admixture in Dravidians, with a large diversity of haplogroups present in India. This evidence of high rates of admixture is interpreted as support for the Indo-Aryan expansion hypothesis with significantly male-mediated geneflow (Sengupta *et al*, 2006). However, Metspalu *et al* (2004) argue that in India linguistic groups do not cluster into distinct groups, and there is high gene-flow across social and linguistic groups and therefore due to this high admixture, it is not prudent to make inferences regarding the history of these groups through genetic data. So, the spread of Dravidians, and their subsequent admixture history cannot be ascertained with confidence through molecular evidence. The debate of whether proto-Dravidians entered from the North and Brahui is a remnant of that migration, or whether Brahui branched from the rest of the then well distributed indigenous Dravidian at a later stage, cannot not be tested through genetic data.

5.3 Outstanding questions

The two broad areas which deserve further attention are 1) The overall topology and subgrouping of the Dravidian family tree; 2) the status of individual languages. Further, the contribution that information emerging from these two broad areas would make to the history of Dravidian peoples is of interest.

5.3.1 Topology and Subgrouping

The validity and position of different subgroups in the family, and the languages within each subgroup, are still a matter of debate. While Southworth (1976) hypothesized seven distinct subgroups in the language family, (Krishnamurti, 2003) argued for four major branches versus Andronov's three. Thus, the branching topology from Proto-Dravidian through to the major

purported subgroups is worth investigating and in particular, the following questions:

1. What was the **sequence of branching** from Proto-Dravidian?
2. Did Northern Dravidian branch off first? Did the Central group branch out from the Northern group or the Southern group? Krishnamurti (2003), suggested that the Northern group branched first, followed by the split of Proto-South Central into their respective subgroups of South I Dravidian, South II Dravidian and Central Dravidian. He was still unsure of a binary versus a ternary division at this stage.
3. Did South II Dravidian share a common ancestor with South I Dravidian or, as Krishnamurti (2003a) suggests, with the Central Dravidian? While Burrow and others have placed Telugu with the South I Dravidian Krishnamurti (1969), Krishnamurti and Emeneau (2001), argued that the Telugu sub-group was more closely related to its northern neighbours, the Central Dravidian languages and re-defined it as South II Dravidian or South Central Dravidian. Outstanding questions still remain regarding the proximity of South II Dravidian to South I Dravidian and Central Dravidian.
4. In South-I Dravidian languages, there is uncertainty regarding the relationships of the Kota, Badaga, Toda and Kodava branches. Emeneau (1967) placed Kota-Toda as a group splitting from Proto-Toda-Kota node, along with the Kodava, branch from the Proto-Malayalam-Tamil node, separate from Badaga and Kannada. But the position of Kodava in this grouping is a point of controversy. Krishnamurti presents evidence to place Kodava closer to Tamil-Malayalam than Kota-Toda as Emeneau presents it. Therefore, the position of Kodava and also the position of the Nilgiri languages (Kota, Toda and Badaga) are contentious.

5.3.2 The status of individual languages

To understand the history of Dravidian languages, it is also important to validate the position of key individual languages within the language family. As stated earlier, the placement of Kodava within South-I is of interest. Then, the question whether Brahui represents the oldest lineage of the Dravidian language family is of interest. The notion that Proto-Dravidians entered the subcontinent from the northwest is based on the notion that Brahui was the result of the first split of Proto-Dravidian. In Andronov's work based on glottochronology, Brahui was separated from the rest of the languages by at least 5000 years. This time depth could be artificially inflated by heavy borrowing from Balochi (McAlpin, 1974) and therefore it is imperative to answer the question regarding the position and evolution of Brahui.

5.3.3 Current Study

During two visits to India, I sampled 20 different languages, using improved methodology to address errors in some of the previous work on Dravidian languages. The quantitative work on Dravidian languages based on the lists by Andronov (1964) has several shortcomings like the lack of comprehensiveness of data on non-literary languages, and even the accuracy of the data collected on the basic word list has been criticized by scholars (Krishnamurti, 2003). I tried to address the issues of accuracy and decided to start fresh and collect a larger (twenty languages) and higher quality sample of Dravidian languages. In the current study, primary data was gathered from native speakers, while focussing on avoiding the errors made by Andronov during data collection, like the existence of multiple words corresponding to a meaning, and contextusal differences in usage. I took advantage of the robust computational framework of phylogenetic methods in order to address questions regarding: the structure of the Dravidian language family, the mode of evolution of the different subgroups, the robustness of the subgroups and to infer the underlying chronology of the Dravidian language family to shed light on the timeline of Dravidian family evolution.

Previous attempts at understanding questions on the evolution and association between languages have largely focused on using methods like glottochronology (Swadesh, 1952) and lexicostatistics. Model-based hypothesis testing approaches available with computational phylogenetic methods borrowed from evolutionary biology have enhanced our ability to infer trees and to test competing hypotheses about history using linguistic evidence in a statistically rigorous manner. In recent years, investigators have applied these methods to language families in different regions: for example, inference of the Austronesian language phylogeny allowed for the testing of hypotheses regarding the settlement of the Pacific (Gray *et al*, 2009; Gray and Jordan, 2000; Greenhill and Gray, 2005). Similarly, language phylogenies were used to investigate questions regarding the origin and spread of Indo-European (Bouckaert *et al*, 2012; Gray and Atkinson, 2003) and Bantu languages (Holden, 2002; Holden and Gray, 2006).

Given that we now possess new methods of understanding and deciphering pre-history, here we use phylogenetic methods to address some of the questions that remain unanswered along with testing and validating hypotheses pertaining to the Dravidian language family. The following section describes the underlying methodological background to phylogenetic methods and details of the approach taken in this Chapter to answer each question.

5.4 Phylogenetic approaches to language: A methodological overview

5.4.1 Parallels of biological and linguistic evolution: applications for inferring historical relationships

The fundamental characteristics of linguistic and biological evolution are evidently parallel, where both modes of evolution evolve in a Darwinian manner (explained in introduction). While the fundamental unit of biological evolution is DNA sequences, in language this can be either grammatical, phonological structures or lexicons. Just as DNA sequences might differ between individuals,

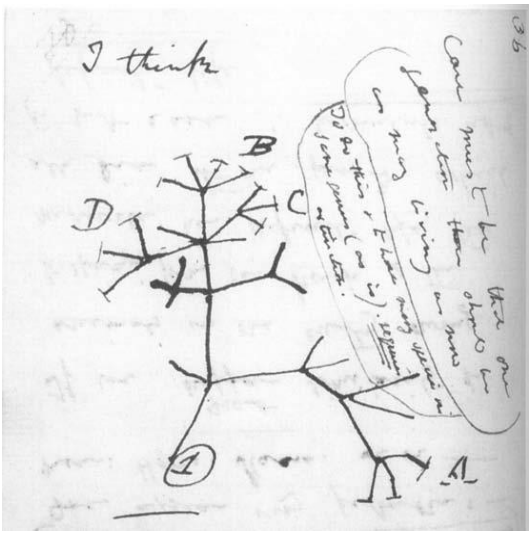


Figure 5–3: Sketch from Darwin's diaries (1837) depicting genealogical relationships in a tree like fashion (Atkinson, 2006)

the units of linguistic evolution can also differ from language to language. Just as homologous features in biological structures indicate common ancestry, homology in linguistic characteristics can be taken as evidence for inheritance from a common ancestor, a proto-language (Atkinson, 2006; Croft, 2000). Darwin, had sketched in his dairies probably one of the earliest, and most well-known depictions of genealogy through evolutionary trees (Figure 5-3).

With the publication of the *Origin of Species* (1859), the connection between similarity by descent and similarity between species characteristics became clearer and their depictions through evolutionary trees became more meaningful (Mayr, 1982). However, evolutionary thinking was also being pursued in other fields, like linguistics, where a language tree was published by Schleicher (1853). He represented the genealogical relationships between the different languages of the Indo-European family in the form a tree in a paper titled "*Die*

Darwinsche Theorie und die Sprachwissenschaft". In his paper, Schleicher (1873) also expounds on the point that, using trees to represent genealogical relationships between languages was a well-established method before Darwin's assertion regarding species and their evolution in a tree-like manner. He used ideas from biology and comparative anatomy to arrive at these tree-like relationships, and introduced the "Stammbaum approach". The idea of phylogenetic understanding had advanced considerably in linguistics, even before Darwin and continued to do so through the 19th century, when Schleicher (1873) published his paper on comparative linguistics much before the publication of the *Origin of Species* and there was also substantial work on evolution of manuscripts by philologists who used tree-based depictions (Atkinson, 2006). However, there has been some interdisciplinary borrowing of ideas for the depiction of languages and species. According to Koerner (1983), the first tree based approach was used by Schlegel (1808), to depict comparative grammar. Another major advancement in the comparative method in linguistics was the distinction between retentions and innovations made by Bruggman and Osthoff (1878). They noticed that there were two types of shared characters present in languages, shared innovations and shared retentions. Shared innovations are those shared characters that are exclusive to one group and the character is not present in the ancestral form, while a shared retention is the shared character that is present in the ancestral form. This differentiation is crucial, as shared retentions tend to exist for a longer duration of time period, and are not useful in classifying closely related organisms or in deciphering the degree of relatedness, whereas shared innovations are likely to differentiate closely related groups and are more important in phylogenetic classification. The same advances in biology came only around 1950, when there was a distinction made between synapomorphic (shared innovations) and sympleisomorphic (shared retentions) characters (Hennig, 1950).

As demonstrated, there were considerable similarities in the comparative methods of linguistics and biological phylogenetics. However, there was never an effort in linguistics to use computer algorithms to find the best tree. Some efforts were made by Gleason (1959) and Hoenigswald (1965) who formalized the criterion implicit in finding the best tree. Thomason and Kaufman (1988) later identified the sequential steps involved in depicting genealogical relationships between languages, where determining the phonological correspondences of words of the same meaning and establishing their phonological systems, followed by establishing and reconstructing grammatical correspondences would lead to identifying a sub-group of languages and further a model of diversification. With the advent of the "modern synthesis", where evolutionary theory was explained in terms of Mendelian characters of inheritance (Dobzhansky, 1937; Fisher, 1930) and the discovery of the structure of DNA, the amount of data that could be analysed increased and so did methodological innovations to infer phylogenies

(see Felsenstein, 2004 for details) from distance based methods (Michener and Sokal, 1957; Sneath, 1957). These methodological innovations could now be applied to linguistic characters that were evolving in a Darwinian manner.

Before these methodological innovations arose, Swadesh introduced lexicostatistics (Swadesh, 1952) and glottochronology (Swadesh, 1955; Swadesh, 1972) to understand genealogical relationships between languages. Lexicostatistics used cognate-coded lexical data to infer language trees from the percentage of shared cognates between languages. For example, homologous words or cognates for the meaning "water" are found in English (*water)*, German (*wasser*), Swedish (*vatten*) and Gothic (*wato*). This reflects that the meaning for water in all these languages descended from a common ancestor of the Proto-Germanic languages and these words are cognates with each other. Cognates are lexical items that are inherited from an ancestor and identified through corresponding sound changes. While the words themselves are not homoplastic, the changes they undergo are homoplastic and hence are suitable for phylogenetic analyses.

Swadesh developed a list of 100 words (initially 200 words) now known as the "Swadesh list" (Swadesh, 1952). Usually, words in each language change or are lost over time, based on the culture, ecology and daily activities of usage. But there was a need to identify a set of concepts/words that would remain in a language and were resistant to change, loss or borrowing over time. Therefore, Swadesh in his list had chosen words like numbers, colours, elements of the nature, words for daily activities like eating, sleeping, and parts of body. Based on these words, the process of analysing relationships between languages following his methodology has the following steps. First, data is collected for these concepts in the languages that need to be compared. Then, from the collected word lists in individual languages, sets of cognates are identified. In this process, words that are borrowed rather than inherited from a common ancestor are removed from the analysis. Cognates are identified on the basis of corresponding sound changes, following which; a distance matrix is constructed based on the number of cognates shared between languages. This distance matrix is then used to depict genealogical relationships between the different languages.

The first step is cognate coding the lexical data. The data is in the form of a matrix. Rows represent the languages and the columns represent the traits of the language. The important ideology of these traits is that they should be homoplastic. The table depicts a sample of the raw form of data collected from the Swadesh word list (Nichols and Warnow, 2008).

Table 5-1: Table depicting data collected from Swadesh word list for four different languages (English, French, Russian, Ingush) for Blood, Bone, Dog, heart and Sun. The numbers in the brackets indicate the cognate identity value. (Nichols and Warnow, 2008)

	blood	bone	dog	heart	sun
English	blood (1)	bone (1)	dog (1)	heart (1)	sun (1)
French	bang (2)	os (2)	chien (2)	coeur (1)	soleil (1)
Russian	krov' (3)	kost' (2?)	sobaka (3)	serdce (1)	solnce (1)
Ingush	c'ii (4)	t'exk (3)	zhwalii (4)	dog (2)	maalx (2)

These words are taken from English, French, Russian and an unrelated Causasian language, Ingush. The state of characters for each feature/variable is coded. Cognates are coded as the same states. For example, for the concept HEART in the word list, *heart* in English, *coeur* in French, *serdce* in Russian are coded as state 1 and *dog* in Ingush is coded as character state 2. From this data set, a distance matrix is created based on the cognate coding (in brackets, in Table 5-1).

In principle, to understand genealogical relationships or infer phylogenies, anything that carries a phylogenetic signal can be used. However, the information that we use constraints the inferences we can make regarding the phylogeny. For example, when we use typological features which are shared retentions, there exists the problem of homoplasy, lateral diffusion, low rate of change and convergent evolution leading to similarly. This in turn will affect the reconstruction of historical relationships, unless appropriate measures are taken to account for these effects. Reesink *et al* (2009) have shown how structural typological data presumes that unrelated languages may have identical typological parameters, and this feature makes it particular useful to detect contact and admixture (Hunley *et al*, 2008). However, certain lexicons items are known to be stable over time and depict relationships between languages accurately (Dunn *et al*, 2005).

As for inferring divergence time, Swadesh extended the approach of lexicostatistics to glottochronology, which is an approach to estimate time depth on the resulting distance based tree (from lexicostatistics) under the assumption of a constant rate of lexical evolution. Using this glottochronological constant, the divergence time for each node can be calculated. If *c* is the percentage of shared cognates and *r* is the glottochronological constant, (the expected proportion of cognates remaining after 1,000 years of separation; Swadesh, 1955), then time of divergence, *t* is calculated by *t=(logc)/(2logr).* Generally, 81% retention/millennium is used as the value of *r,* when Swadesh's 200-word list is used.

Lexicostatistics and glottochronology as investigated through this methodology has since been abandoned due to several problems. By converting to a distance matrix, we lose important information about the data. Glottochronology assumes a constant rate of change, which might not be true as the rate of lexical replacement can different in different parts of the phylogenetic tree (Campbell, 2002; Gray and Atkinson, 2003). Using just distance methods, and assuming constant change for all languages reduces the robustness of the tree. These measures also cannot account for borrowing (within and across a language family), which might lead to the inference of spurious relationships. If there is a high degree of contact between languages, resulting in a large amount of borrowing, using percentage-shared measures that do not account for shared cognates due to contact might result in erroneously inferring relationships between languages that might be similar just due to contact. It would also confound the divergence time estimates due to the overestimation of shared cognates. Andronov's (1964) attempt to understand the chronology of evolution of the Dravidian language family, has drawn criticism for the methods used as well as for the quality of data (Krishnamurti, 2003). Krishnamurti (2003) explained that by choosing the wrong set of lexical representations, Andronov (1964) had erroneously grouped Telugu with South-I instead of South-II. For example, the meaning "*foot"* can be expressed by two lexemes in Telugu, *padamu* and *adugu*. While *padamu* is a borrowing from Sanskrit, *adugu* is cognate with the languages Tamil, Malayalam and Kannada. But by only using *padamu* for Andronov's (1964) lexicostatistic analysis, he did not account for borrowing and instead erroneously inflated the distance measures. Languages are changing through time and as these changes accumulate, akin to biological evolution, languages too evolve and diverge from each other. These changes can be losses, changes in morphology, and changes in sound. With lexicostatistics, measurement of evolution is based solely upon loss and replacement. But as we now know, this is not a complete representation of linguistic evolution.

There has been tremendous development in evolutionary methodology to represent historical language relationships, especially in the field of phylogenetics. This lead to a substantial amount of work using these methods to test already existing hypothesis about languages as well as to seek answers to questions that traditional linguistics have not been able to answer till now (Bouckaert *et al*, 2012; Dunn *et al*, 2011a; Dunn *et al*, 2011b; Dunn *et al*, 2005; Gray and Atkinson, 2003; Gray *et al*, 2009; Gray and Jordan, 2000). Currently, with phylogenetic inference, we can define the model of evolution to estimate the evolutionary history for a set of related languages. Given sets of cognate-coded information on the lexicon of different languages, there are two methods that we can use to build phylogenetic trees: distance-based and character-based methods. Previous attempts at quantitatively understanding the structure of Dravidian

languages were made using employing distance-based methods described in the following section (Andronov, 1964).

5.4.2 Distance based methods:

Tree building techniques to represent relationships have since grown and new methods introduced (Felsenstein, 2004). The tree-building techniques of the distance based methods (UPGMA, NJ), produce accurate trees in the absence of rate heterogeneity but tend to cluster languages evolving at a similar rate (i.e., slowly evolving languages in one cluster), rather than languages that share a common ancestor (Blust, 2000). Recent efforts were made using distance based phylogenetic methods by Rama *et al* (2009) address some of the concerns on the tree topology of the Dravidian languages. In the study by Rama *et al* (2009) tested the grouping of the different branches of the Dravidian language family, using distance based algorithms (UPGMA and Neighbour Joining). However, they were not able to reconstruct any of the proposed subgroupings of the family with confidence and did not address the problems in the primary dataset collected by Andronov.

5.4.3 Character based methods

A more effective approach that has since emerged to do phylogenetic inference are character-based methods, i.e., models where inference is based on the behaviour of individual linguistic features present in the dataset to try and address the shortcomings of the distance based methods (Swofford and Sullivan, 2003). Character based methods, instead of finding the shortest distance between two languages, try to find an evolutionary plausible pathway between a language and its most recent ancestor. Therefore, character based models are much more realistic in depicting the process of actual language evolution than distance based methods (Steel and Penny, 2000). The following sections describe the different character based methods.

5.4.3.1 Maximum Parsimony

Maximum parsimony (MP) has been the most popular choice of method for tree building until recently when other robust methods like maximum likelihood and Bayesian methods have gained popularity (Sullivan and Swofford, 2001). MP works on identifying the least amount of evolutionary change required to produce the observed states of variation in a dataset, i.e., the tree topology where innovations maximize the amount of diversity observed are preferred. However,

this approach has many flaws and experiments with simulated data show that it depicts the true tree only under certain conditions. But recently, likelihood and Bayesian methods have been identified as being more robust at arriving to a much more realistic account of evolution and diversification and have overtaken maximum parsimony methods as a scientifically stronger choice.

5.4.3.2 Likelihood and Bayesian methods

Maximum likelihood methods calculate the likelihood (L) of the observed data (D) being produced by a particular model of evolution (M), i.e., L = P(D|M) (Felsenstein, 1981). Some of the parameters are the tree topology, the branch lengths and the probability that a new cognate set appears or that a reflex of a cognate set is lost. With likelihood methods, the model is actually our hypothesis of language evolution and the validity of the model for our dataset is quantified. Different models of evolution can be compared to arrive at the most likely model of evolution for our data (Nichols and Warnow, 2008; Pagel and Meade, 2004).

To find the parameters (branch lengths, rate of change, etc.,) that maximize the likelihood of the tree is mathematically tractable, but to find the best tree topology, given the number of parameters that influence phylogenetic inference and the vast tree space, is computationally exhaustive and there is no known algorithm to identify the best tree in a reasonable amount of time (Schmidt and von Haeseler, 2009). The different searches might just result in a small proportion of random trees that may or may not be close to the highest possible likelihood in the tree search space, as only a small proportion of the tree space represents good solutions to the evolutionary hypothesis. For example, the search space for an Oceanic language sample of 16 taxa was calculated to be no fewer than 213,458,046,676,875 distinct unrooted tree topologies alone (Felsenstein, 2004). The solution for a productive search emerged in the form of Bayesian Monte Carlo Markov chain (MCMC) sampling (Pagel and Meade, 2004). Bayesian approaches estimate the probability that each possible tree is the true tree and therefore they do not produce a single tree, but a posterior probability distribution on the set of possible trees (Nichols and Warnow, 2008; Pagel and Meade, 2004; Pagel *et al*, 2004; see Dunn, 2008 for an introduction to Bayesian methods, and appendix for further details on Bayesian and ML)

5.4.3.2.1 Search Algorithm and sample of trees

The search algorithm starts at a random point in the likelihood space and searches for the region of trees with the highest likelihood (Felsenstein, 1993; Swofford, 1998). At any point in the likelihood space, the algorithm perturbs the parameter

values at that space, compares the likelihood of the current position to the new set of values and if the new values have a higher likelihood, then it repeats the exploration from the new position. Imagine this, in simplistic terms, as a hill-climbing algorithm. If the likelihood were to be compared to elevation (higher likelihood, higher elevation), then the algorithm measures the height difference between its current location and that of a nearby point. If that point is higher, it takes a step in that direction, otherwise it searches for a new point. Imagine the terrain of the likelihood space to be rugged, then there is a probability of the algorithm reaching local maxima, and getting stuck (at the top of a mound, instead of at the peak of the mountain). The algorithm deals with this by not completely rejecting every lower likelihood moves in the search, and by randomly accepting to move to a lower likelihood position if the difference in likelihood is low (i.e., it would more likely move in the direction of a position with a slightly lower likelihood than to a position with a large difference in likelihood).

The tree parameters and likelihood values are saved to the posterior sample for instance once every thousand steps (as defined by the researcher). This is done in order to maintain statistical independence between data points, and to make sure that the values of the parameters are not too close in the parameter space. The initial phase of the tree search is called "burn-in", where the search algorithm moves erratically and it takes a while for it to reach a stable, maximum state. The values collected during this period are discarded, as these values do not represent highly likely sets of parameters, until it reaches the stable state of search space. The parameters obtained in the posterior, after reaching the stable, maximum state are more-or-less comparable sets of highly likely parameter values (as they are sampled in proportion to their likelihood).

The result of this search is a random set of trees that have the highest probability for a combination of parameters (Nichols and Warnow, 2008). For example, we can consider the proportion of trees containing particular taxa at a particular node as a measure of confidence of its position in the tree topology; similarly, confidence measures of other parameters can be estimated. Therefore, the advantage of a Bayesian framework is that the measure of uncertainty in a predicted model is explicit and quantifiable. In this manner, the result, i.e., the sample of trees obtained by the search algorithm are held to approximate the "true" posterior probability distribution.

There are several models of character evolution that can be used in a phylogenetic inference, and the fit of the model to a dataset is defined as the likelihood score, which is tractable (Pagel and Meade, 2004). Character-based methods allow us to maximize likelihood by inferring the phylogeny most likely to produce the observed variation. Bayesian inference is based on the model of conditional probability, i.e., the prior probability defines our estimate of the performance of the model and the posterior probability helps us define and revise the range of the priors and arrive at a true estimate for the parameters of the

model. To calculate the likelihood of each model, we need to have three variables, the model of the evolutionary process, the tree topology, and our priors. Different models of evolution of the characters can be specified depending on prior, if possible empirical, knowledge of the behavior of cognate changes or retention across the dataset, whereas estimation of the tree topology and parameter values is the goal of the search. For example, a simple model could be that, "each character changes state with probability P per unit time" (Dunn *et al*, 2005). The character history of states of a particular character can be simulated randomly (the character state changes along the branches of the tree from the root to the tips depending on the rate defined by P). The simulations would produce many result distributions, which were different from the observed distribution in the particular dataset under investigation. When the simulations were carried out a large number of times, for a particular probability of change of state, the proportion of result distributions equivalent to the observed distribution would be analogous to the likelihood of that particular value of P (for that model and tree). If we have specified prior ranges that are broad enough, then there is a high probability of simulating data that can be comparable to the observed dataset, and the maximum likelihood of the observed distribution under the model can be identified (Pagel and Meade, 2004). Therefore, the combination of parameters that produces the highest likelihood of explaining the observed data is the best phylogenetic explanation of the data for that model (Nichols and Warnow, 2008).

5.4.3.2.2 Rates of Change

The example above defines a simplistic model, with one parameter. Adapted from biology, realistic models include several other parameters, such as those allowing for different classes of characters to be able to change at a different rate: some characters are conserved and change at a slower rate, whereas other characters tend to produce variations through innovations, and change at a faster rate (Atkinson, 2006; Gray and Atkinson, 2003; Pagel *et al*, 2004). In likelihood models, rate of change is defined under the parameter, clock models. A strict clock model constrains the rate of change to be constant and allows for no variation across the lineages of a tree. This has proven to be inappropriate, in many cases for biological models of evolution (Felsenstein, 2004), and in linguistics too (Greenhill and Gray, 2009; Posada and Buckley, 2004; Posada and Crandall, 2001). A more appropriate and popular clock model used is the relaxed clock model, where the rates are allowed to vary across the tree, and the value of this variation can be drawn from a probability distribution, whose mean is determined by the rate of the parent branch (correlated clock model), or can be assigned by a parameter value, where there is no relationship between the rate of a branch and that of its ancestor

(uncorrelated clock model), as implemented in BEAST v.1.7 (Drummond *et al*, 2012).

Another parameter of importance is the parameter that specifies how the rate of cognate change differs across different cognate sets, also known as the substitution model. The substitution models for binary coded data are simple one or two rate models, where the rate of change from character state 1 to 0 or 0 to 1 is expressed as a single rate q or as two rates q10 and q01 respectively (where 0 is coded as absence, and 1 is coded as presence). The latter two-rate model is more realistic, as it treats innovation of a cognate separate from the rate of loss. To apply this model of change, it is necessary to specify the root of the tree to determine whether a change is $0 \rightarrow 1$ or $1 \rightarrow 0$ (see appendix for details on rooting). It is possible to determine the position of the root as part of the topology (Drummond and Rambaut, 2007), as the position of the root influences the likelihood of the tree. The gamma substitution model assigns each character to one of the many rate classes, where the rates for each class are drawn from a gamma distribution (Yang *et al*, 1994). The shape of the gamma distribution is controlled by a single shape parameter, and this model of substitution is deemed to be apt for linguistic data from Swadesh lists, as the different terms on the list have different levels of stability (Greenhill and Gray, 2009). The covarion model allows for the rate of each character to vary along the branches of the tree, overwhelming the individual differences in rates controlled by the gamma model. Another substitution model is the stochastic Dollo model, where, according to Dollo's law, the same cognate cannot not arise independently in several languages i.e., it is homoplasy free (Ringe *et al*, 2002). Language change can occur either due to innovation, loss or reproduction (i.e., when languages split, forming two descendant copies of the parent language) (Greenhill and Gray, 2005); (Nicholls and Gray, 2008). As birth, death and reproduction of cognates play a vital role in the evolution of characters, the stochastic Dollo model always produces a rooted tree. The Dollo models are also designed for capturing linguistic realism as they well approximate the known process of cognate substitution (Dunn *et al*, 2013).

5.4.3.2.3 Assessing the best model

The advantage of using the Bayesian approach is the opportunity to incorporate different forms of prior knowledge in a model. One way to incorporate the knowledge is through the extent of distributions of the prior parameters, and knowledge on subgrouping and tree topology can be included as constraints in the model (Dunn, 2014). As Dunn (2014) emphasizes, incorporating knowledge on subgrouping in tree priors, based on our knowledge of phonological and morphological innovations is the most direct way of integrating traditional comparative method and computational phylogenetic methods of tree inference.

The constraints restrict the MCMC search algorithm to the likelihood space where the tree topologies are consistent with the prior knowledge. When strong apriori knowledge on particular parameters exists, reducing the search efforts on such parameters and constraining the search algorithm can increase the speed of the analysis. By incorporating these constraints, different models of subgrouping can also be tested against one another, where uncertainty of grouping exists (Dunn *et al*, 2013).

Likelihood scores for the same data using different models helps determine the model that has a significantly better fit over others. Ideally, the model with the simplest parameters and highest likelihood is sought after (Kelchner and Thomas, 2007). An increase in the complexity of the model, with small increases in likelihood can lead to over-fitting of the model, where the model has the predictive power of just that dataset, and would probably not be suitable to slight changes in the dataset, hence losing its predictive power. The performance of the models can be assessed by the Bayes Factor test (See Chapter 3 and Kass and Raftery, 1995), which determines the best performing model for the dataset. Bayes Factor is a Bayesian analog to the likelihood ratio test. The model likelihoods being compared, however, are derived from integration over all possible parameter values, rather than from maximum likelihood estimates for the model. The likelihood score of each analysis represents the probability that the observed data could have evolved under that particular model. The difference in the performance of the two models can be evaluated using the Likelihood ratio, which is the ratio of $L(M_1)$, the likelihood of model one, to the likelihood of model two $L(M_2)$, i.e., $LR_{12} = L(M_1)/ L(M_2)$. Incidentally, the Bayes Factor statistic is expressed as twice its natural logarithm, $2\log LR_{12}$ rather than the raw value of the ratio. i.e., $2\log BF_{12} = 2(\log L(M_1) - \log L(M_2))$. Bayes factor is not an independent value like a p-value; rather it is only meaningful in testing one hypothesis against the other. A positive Bayes Factor means that it supports model 1 over model 2 (Posada and Buckley, 2004). The strength of this support is directly correlated to the magnitude of the Bayes Factor statistic (Table 5-2; Berger and Pericchi, 1996; Kass and Raftery, 1995).

Table 5-2: Table depicting the use of Bayes Factor statistic and its consequences on hypothesis testing

LR_{12}	$2\log BF_{12}$	Evidence for M_1 Over M_{12}
0 - 2	1 - 2	Negligible
3 - 20	2 - 6	Positive
20 - 50	6 - 10	Strong
>150	>10	Very Strong

The Bayes Factor test is just a comparative test, and does not make judgements on whether the model is correct or not, and this responsibility solely

rests at our discretion to determine the validity of the model being applied. As mentioned earlier, the best model is in most cases, the simplest model (Kelchner and Thomas, 2007). When the results from the likelihood ratios are not able to distinguish between two models, then the model with the fewest parameters should be considered, as more often than not, the complicated model is an extension of the simpler model.

5.4.3.2.4 Visualizing the tree sample

Once the most likely model of evolution is determined and the posterior distribution of tree sample obtained, it is imperative to understand and decipher the values of parameters that define the evolutionary relationship between the languages in our dataset and this is typically achieved by observing the tree sample. As stated above, the posterior distribution of this analysis typically contains thousands of trees, and it is not practicable to inspect it by eye. Visualizing the tree sample is deemed as the best technique to understand and formulate a single coherent narrative regarding the phylogenetic inference (Nichols and Warnow, 2008). Currently, the most widely used method is to calculate the maximum clade credibility tree. The maximum clade credibility (MCC) tree is treated as the best representative of the posterior distribution of the tree sample which is constructed using the *TreeAnnotator* tool in *BEAST* package (Drummond *et al*, 2012). The MCC tree is the tree out of the sample that maximizes the product of the frequency (over the whole sample) of the branches. Branch lengths are usually taken from the median or mean of corresponding branches in the sample. The motivation with this is to find a single 'point estimate' tree that is in some way central to the distribution of trees. This tree is then given (annotated with) summary information for the full set of trees from the sample (Drummond *et al*, 2012). It is important to be noted here that visualisation is just a convenience and analysis regarding the divergence and testing different hypotheses are done using the full sample.

The consensus tree method has also been popular in summarizing sets of trees. This is estimated by transcribing all the non-conflicting branching present in the posterior distribution of tree sets, in order of their frequency i.e., the most commonly occurring bifurcation is first transcribed and then the next most commonly occurring branching that does not contradict the earlier bifurcation is added, and so on, till the tree is completely resolved, or until there are no more branching present in more than 50% of the trees (Pagel and Meade, 2004; Sharkey and Leathers, 2001). For example, if a branch has a credibility of 1, that means that it is present in all of the posterior distribution sample of trees, and any number less than 1 means that this particular branching is represented in a subset of the posterior sample. The drawback of the consensus technique is that, there is no

guarantee that there will actually be a tree with the topology similar to the consensus tree in the posterior sample set and there is also no generally acceptable way of specifying the branch lengths of a consensus tree (Drummond and Rambaut, 2007; Dunn, 2014; Sharkey and Leathers, 2001). Therefore, the MCC tree is widely preferred today to represent the posterior tree sample set. Other tree visualization techniques include *DensiTree* (Bouckaert, 2010), *SplitsTree* (Holland *et al*, 2004), which can be used to detect reticulation and uncertainties in the data (for further description see Dunn, 2014).

5.5 Inferring the Dravidian language phylogeny

5.5.1 Data

5.5.1.1 Primary Data collection

I collected data using Swadesh's 100-word elicitation list (Swadesh, 1972), for twenty languages listed in Table 5-3, during two field seasons : 2010 and 2013. The team (V Kolipakam, M Dunn, F Jordan & A Verkerk), perfomed cognate coding using the Dravidian Etymological dictionary (Burrow and Emeneau, 1984). All cognacy judgements recorded and were compared for quality control. Cognate judgements from languages not in etymological dictionary carried out by comparison to other similar languages. For Malto, data was collected from C Puttuswamy (Indian Institute of Technology, Kanpur, pers. comm.) in the form of voice recordings and for Betta Kurumba, from Gail Coehlo (University of Delhi, Delhi, pers. comm.) in an already IPA transcribed format.

The procedure of eliciting the data was as follows. For all the languages (except Malto and Betta Kurumba), a native language speaker of these languages was given the 100-word Swadesh list on paper in the commonly used literary language for that particular language. During the course of data collection, I observed that informants would confuse between words used under different contexts. This was observed after the initial data collection phase was completed, however to eliminate further error, data collected during 2013 (see table 5-3) for languages Brahui, Kurukh, Ollari-Gadaba, Parji, Kolami, Kurukh, Kuwi and Kota, along with the Swadesh list, the context of each word was provided to the informant as described in Kassian *et al* (2010). The informants were asked to say the same word in their native language, and where possible (i.e., where the script of the language was available), it was also written by the informant.

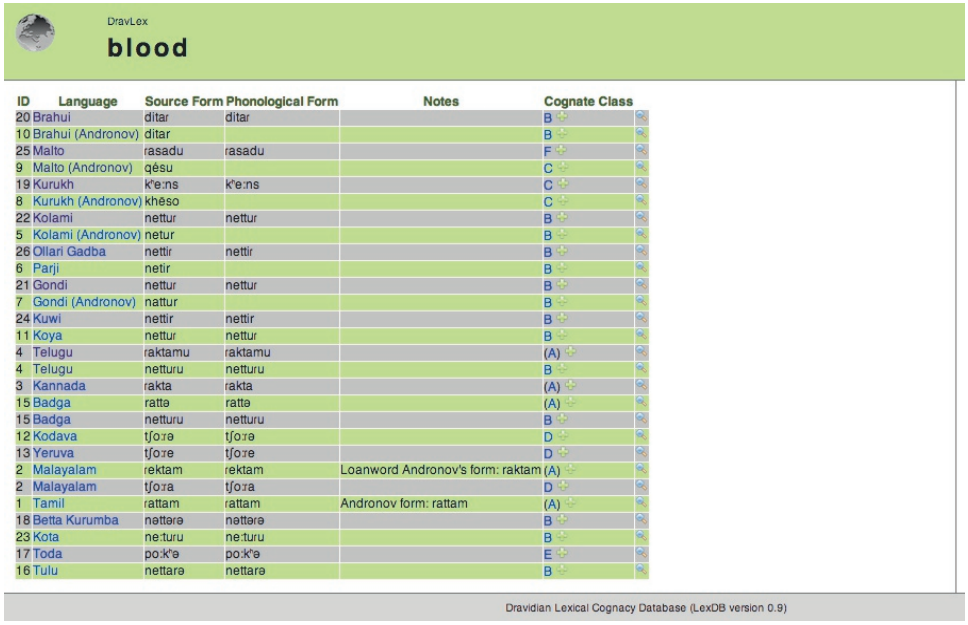
For languages where there was no script, or if the informant was not comfortable using the script of another language, the words pronounced by the informants were recorded using a voice recorder (for details, see Table 5-3). The written transcripts and the voice recordings were crosschecked and transcribed to IPA by me.

Table 5-3: Table depicting the use of Bayes Factor statistic and its consequences on hypothesis testing

Language	ISO code	Classification according to Krishnamurti (2003)	Source Form	Native speaker	Primary Source
Brahui	brh	North	Written (Arabic, Latin Scripts), Audio	Abdul Raziq, Balochistan	V Kolipakam (2013)
Malto	mjt	North	Written (Arabic Script), Audio	-	C Puttaswamy (pers comm; 2010)
Kurukh	kru	North	Written (Arabic Script), Audio	AK Baxla, Jharkhand	V Kolipakam (2013)
Ollari_Gadba	gdb	Central	Audio	Rajesh (elicited by MK Mishra, Orissa)	V Kolipakam (2013)
Parji	pci	Central	Audio	Mohan (elicited by MK Mishra, Orissa)	V Kolipakam (2013)
Kolami	kfb	Central	Audio	Karan, Orissa	V Kolipakam (2013)
Kuwi	kxv	South -II	Audio	Anup (elicited by MK Mishra, Chhattisgarh	V Kolipakam (2013)
Gondi	gno	South -II	Written (Latin Script), Audio	Pawan, Madhya Pradesh	V Kolipakam (2013)
Koya	kff	South -II	Written (Telugu Script), Audio	Peter Daniels, Andhra Pradesh, Khammam district	V Kolipakam (2010)
Telugu	tel	South -II	Written (Telugu Script), Audio	V Kolipakam, Andhra Pradesh	V Kolipakam (2010)
Tamil	tam	South -I	Written (Tamil, Telugu Scripts), Audio	Venkatachalam Chokkalingam, Tamil nadu	V Kolipakam (2010)
Malayalam	mal	South -I	Written (Malayalam, Telugu Scripts), Audio	Anil Nair, Kerala	V Kolipakam (2010)

Language	ISO code	Classification according to Krishnamurti (2003)	Source Form	Native speaker	Primary Source
Kannada	kan	South -I	Written (Kannada, Telugu Scripts), Audio	Ponnappa, Karnataka	V Kolipakam (2010)
Kodava	kfa	South -I	Written (Kannada, Telugu Scripts), Audio	Bopanna, Karnataka	V Kolipakam (2010)
Tulu	tcy	South -I	Written (Kannada, Telugu Scripts), Audio	Sunanda, Karnataka	V Kolipakam (2010)
Yeruva	yea	South -I	Audio	Chubakki, Karnataka	V Kolipakam (2010)
Toda	tcx	South -I	Audio	Kishore, Tamilnadu	V Kolipakam (2010)
Kota	kfe	South -I	Written (Tamil, Telugu Scripts), Audio	Mohanraj, Tamilnadu	V Kolipakam (2013)
Badaga	bfq	South -I	Written (Tamil, Telugu Scripts), Audio	Yellapa, Tamilnadu	V Kolipakam (2010)
Betta Kurumba	xub	South -I	IPA transcription	-	G Coehlo (pers.comm; 2010)

5.5.1.2 Cognate judgments


DravLex

blood

ID	Language	Source Form	Phonological Form	Notes	Cognate Class
20	Brahui	ditar	ditar		B
10	Brahui (Andronov)	ditar			B
25	Malto	rasadu	rasadu		F
9	Malto (Andronov)	qésu			C
19	Kurukh	kʰeːns	kʰeːns		C
8	Kurukh (Andronov)	khēso			C
22	Kolami	nettur	nettur		B
5	Kolami (Andronov)	netur			B
26	Ollari Gadba	nettir	nettir		B
6	Parji	netir			B
21	Gondi	nettur	nettur		B
7	Gondi (Andronov)	nattur			B
24	Kuwi	nettir	nettir		B
11	Koya	nettur	nettur		B
4	Telugu	raktamu	raktamu		(A)
4	Telugu	netturu	netturu		B
3	Kannada	rakta	rakta		(A)
15	Badga	rattə	rattə		(A)
15	Badga	netturu	netturu		B
12	Kodava	ʧoːrə	ʧoːrə		D
13	Yeruva	ʧoːre	ʧoːre		D
2	Malayalam	rektam	rektam	Loanword Andronov's form: raktam	(A)
2	Malayalam	ʧoːra	ʧoːra		D
1	Tamil	rattam	rattam	Andronov form: rattam	(A)
18	Betta Kurumba	nettərə	nettərə		B
23	Kota	neːturu	neːturu		B
17	Toda	poːkʰə	poːkʰə		E
16	Tulu	nettarə	nettarə		B

Dravidian Lexical Cognacy Database (LexDB version 0.9)

Figure 5–4: Example entry from DravLex showing how lexical data is structured in the database under each meaning in the Swadesh list (here "blood"). For each language the source form in IPA and the cognate class is listed

The IPA coded wordlist were collated and assimilated in an online database (http://dravlex.mpi.nl). These wordlists were then cognate-coded with reference to Burrow *et al* (1984). For words where cognate judgements were not available, I and three colleagues[3] made judgements based on sound changes from Proto-Dravidian to the contemporary languages as listed in (Burrow and Emeneau, 1984). Participant or transliteration errors were coded as "low" quality words and were excluded from the analysis. Known loan words which were listed in the DED (Burrow et al., 1984) and words identical in Indo-European languages were checked for possible borrowings, and excluded from the analysis.

For each meaning in the Swadesh list, cognates can be of several different classes. The words that are cognates with each other belong to one cognate class. For example, in Figure 5-4, the Koya meaning for the word "blood", "*nettur"* and the Tulu meaning, "*nettare"* are cognates with each other and are put under the same cognate class, "B". However, the Malayam meaning, "*ʧorə"*, and is categorised as a different cognate class, "D".

[3] Dunn, M., Jordan, F.M and Verkerk, A.

For each language, all the different forms of meaning for the corresponding word in the Swadesh list was listed and cognate coded. Figure 5-4 shows that there are several entries for the meaning "blood" in the database under one language. For example, the forms for the meaning "blood" were entered as "*raktamu"* and "*netturu"* for the language Badga. However, the word "*raktamu"* is a known loan word from Indo-European, and this is indicated with brackets around the cognate class (

Figure 5–5).

While retained in the database, for phylogenetic analyses loan words and words with low reliability i.e. words where the IPA form could not be transcribed with confidence, or where the context was possibly misjudged by informants, were excluded. The data matrix was then assimilated from cognate classes of each word, as described in the previous section (See section 5.2).

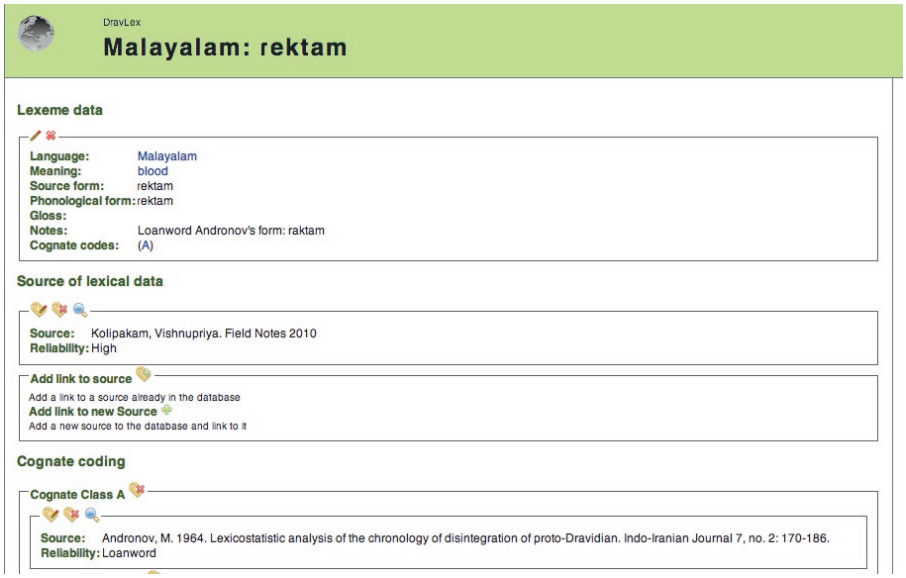


Figure 5–5: Example lexeme entry in DravLex showing the information recorded about each lexeme, including details on the source. Here, the lexeme "rektam" meaning "blood" in the language Malayalam is recognized as a loan word.

5.5.2 Tree building

Beast v.1.8.0 (Drummond *et al*, 2012) was used for all phylogenetic analyses. All the models were run with 2 billion iterations, sampling every 5000 steps and with at least 3 independent runs for each model, to find the model with the best fit for our data. The information for the parameters of this model was largely drawn

from understanding the behaviour of cognate evolution from other studies, like the Aslian dataset (Dunn *et al*, 2011a) and the Austronesian dataset (Gray *et al*, 2009; Greenhill and Gray, 2005). The calibration points for our dataset were obtained from the earliest literary records, as follows: a) Tamil 245 BCE, b) Kannada 450 CE, c) Telugu 620 CE, d) Malayalam 830 CE, e) Tulu 1500 CE, and f) Brahui 800 CE. These calibrations were only used on the lower bound of the prior and no constraints were placed on the upper bound.

The posterior distribution of samples of the above models was obtained, and the likelihoods compared to arrive at the best-fit model. A quality check was performed to see if the distributions of all the independent runs of the chosen model were comparable. Subsequently, independent runs of the chosen dataset were combined, and used to build the MCC tree (Drummond *et al*, 2007; Rambaut and Drummond, 2007) using *LogCombiner* and *TreeAnnotator* (Drummond *et al*, 2012). The parameter estimates were summarized as medians of the posterior tree sample. It was then possible to visualize the tree topology and parameters like branch length, rate of substitution and time stamps on the tree, with confidence intervals.

For each model, it was defined if it was a strict or relaxed clock and the possibility of different substitution models were tested:

a) Strict clock
    a. Simple
    b. Gamma
    c. Covarion
    d. Dollo

b) Relaxedclock
    a. Simple
    b. Gamma
    c. Covarion
    d. Covarion+gamma[4]
    e. Dollo

We tested between two clock models (strict and relaxed) and several substitution models (simple, gamma-distributed, covarion and dollo). The Dollo model with a relaxed clock emerged to be the most likely model for our dataset. It outperformed all other models by a substantial margin (Table 5-4). This is in line with other studies which have used BEAST to infer models of lexical evolution (Bouckaert *et al*, 2012; Gray *et al*, 2011; Gray *et al*, 2009; Lee and Hasegawa, 2013).

[4] Only in the relaxed clock model were we able to model a covarion+gamma model.

Following the relaxed clock Dollo model, the next most likely model was the Dollo model with a strict clock. The difference between the two clock models was around 13 BF units, indicating a clear support for the relaxed clock model. This means that rate variation is needed to adequately account for the division of the languages in the Dravidian language family. The covarion model was the next most likely substitution model, with a difference of around 50 BF units from the relaxed Dollo model and therefore it was unlikely that there was any indication of the covarion model being even slightly preferred over the Dollo model. The Dollo model is apt for linguistic data, as it captures realistically the process of cognate substitution. Under a Dollo model, new cognates can be born only once, meaning there is no chance of homoplasy. The consistency index and retention index calculated by parsimony analysis were 0.69 and 0.4 respectively, which were well within the range of what would be expected of a biological datasets of a similar size (Sanderson and Donoghue, 1989), indicating that the model chosen fits the evolution of characters well.

Even though there are criticisms regarding the draw backs of using Dollo models on data where unidentified borrowing might exist, it has already been shown that the effect of this unidentified borrowing is negligible (Atkinson and Gray, 2005). Given that the Dravidian languages exist geographically proximal to each other, and that Krishnamurti (2003) stressed the uncertainty of borrowing across subgroups, the implications for these unidentified borrowings would be underestimation of divergence estimates, and not the performance of the model. Atkinson *et al* (2005) showed that even with the existence of unidentified borrowings as high as 30%, the Dollo model performed well and captured the precision required to explain language evolution. Therefore, it was assumed that the critiques of the model do not particularly affect the assumptions in the context of the Dravidian language family and the questions that were being addressed.

5.5.3 Dravidian Language Phylogeny

Here I present the first model-based inference of the Dravidian language phylogeny, based on primary data collected from speakers using a common elicitation framework. The result was a successful resolution of a Dravidian tree, with adequate statistical confidence. Most of the branches had a high posterior probability values (above 0.95), indicating good support for the relaxed Dollo model. The model also reflects the branching pattern proposed by Krishnamurti (1961); Krishnamurti (2003) using historical linguistic data and by using the linguistic comparative method. The Northern and South I branches are clearly delineated with strong posterior probabilities. There is some degree of uncertainty within the Central and South II branches. The advantage of using the stochastic Dollo model was that it was possible to estimate the position of the root

without having to specify an out-group. This was advantageous, as in the Indian context of panmixia, the selection of an outgroup is controversial.

The maximum clade credibility tree, summarising the Bayesian posterior distribution sample of trees, largely coincides with the tree structure proposed by Krishnamurti (1961) and Krishnamurti (2003). The two branches that emerge from the root are Proto-Northern Dravidian and Proto-South-Central Dravidian. The Southern and Central Proto node further divides into South II and Central as one lineage and South I as the second clade branching from the split. However, relatively low support (<50%) was observed for the bifurcation of the Central–South II Dravidian and South I Dravidian branches of the tree. This uncertainty was dealt with by examining the performance of constrained models, which are explained later in the section. The internal branching and position of languages within the major subgroups however was reconstructed with high degrees of confidence. The MCC trees are better at summarising the branches with high degree of confidence, i.e., the branches that have a value of 1 (i.e., 100%) are present in all the trees in the sample (Figure 5-6).

In this Chapter, I only focus on topology and do not address the aspect of dating the Dravidian language family tree. In order to get a more robust date estimate of the Dravidian language family, in a forthcoming paper (Kolipakam et al., 2018), my co-authors and I try to narrow down the age estimates reported here through constraining topology, narrowing down age priors, and including several other aspects of analysis. Firstly, we use the newer version of BEAST V.2 (Bouckaert *et al*, 2014) as opposed to V.1.7 used here. In this version of the program, updated methodology, like the stepping stone algorithm that is reported to be more exhaustive and reliable method to calculate marginal likelihood of tested models, is available (Baele *et al*, 2012). Another aspect that we explore for analysis is ascertainment correction to compensate for the fact that only sites included in the cognate alignment are those that have at least one 1 in it, taking missing data in account (Chang *et al*, 2015). Based on our current results, and support from literature, we included several priors on subgrouping. We constrained the languages of North, South I and South II to be monophyletic. We also constrain the root to be a maximum of 10,000 years and try similar constraints on the individual subgroups, for e.g. On Brahui, we place a calibration such that this group cannot be older than 2250 years. Results point to an age of around 4500 years. The reason for the difference with our results and the new analyses could be due to one of several reasons, i.e., calibration constraints, monophyly constraints, ascertainment bias, and a change in the way BEAST handles data and calculates marginal likelihoods of the different models.

Table 5-4: : Table depicting the pairwise comparison of the performance of the different clock and substitution models of the Dravidian lexical dataset as evaluated by BayesFactor units. Relaxed dollo model outperforms all other models by a wide margin

	lm P(model)	S.E.	Strict + Simple	Strict+ Gamma	Strict+ Covarion	Strict + Dollo	Relaxed + Simple	Relaxed+ Gamma	Relaxed+ Covarion	Relaxed+ Covarion+ Gamma	Relaxed+ Dollo
Relaxed+Dollo	-3885.81	+/- 0.36	80.37	82.41	68.66	13.73	68.03	65.44	51.72	54.99	-
Strict+Dollo	-3917.43	+/- 0.27	66.64	68.67	54.93	-	54.3	51.71	37.99	41.26	-13.73
Relaxed+Covarion	-4004.9	+/- 0.27	28.65	30.69	16.94	-37.99	16.32	13.72	-	3.27	-51.72
Relaxed+Covarion+ Gamma	-4012.44	+/- 0.27	25.37	27.41	13.67	-41.26	13.04	10.45	-3.27	-	-54.99
Relaxed+Gamma	-4036.51	+/- 0.30	14.92	16.96	3.21	-51.71	2.59	-	-13.72	-10.45	-65.44
Relaxed+Simple	-4042.46	+/- 0.32	12.34	14.37	0.63	-54.3	-	-2.58	-16.31	-13.04	-68.03
Strict+Covarion	-4043.91	+/- 0.41	11.71	13.75	-	-54.93	-0.63	-3.21	-16.94	-13.67	-68.66
Strict+Simple	-4070.88	+/- 0.34	-	2.04	-11.71	-66.64	-12.34	-14.93	-28.65	-25.38	-80.37
Strick+Gamma	-4075.57	+/- 0.32	-2.04	-	-13.75	-68.68	-14.38	-16.96	-30.69	-27.41	-82.41

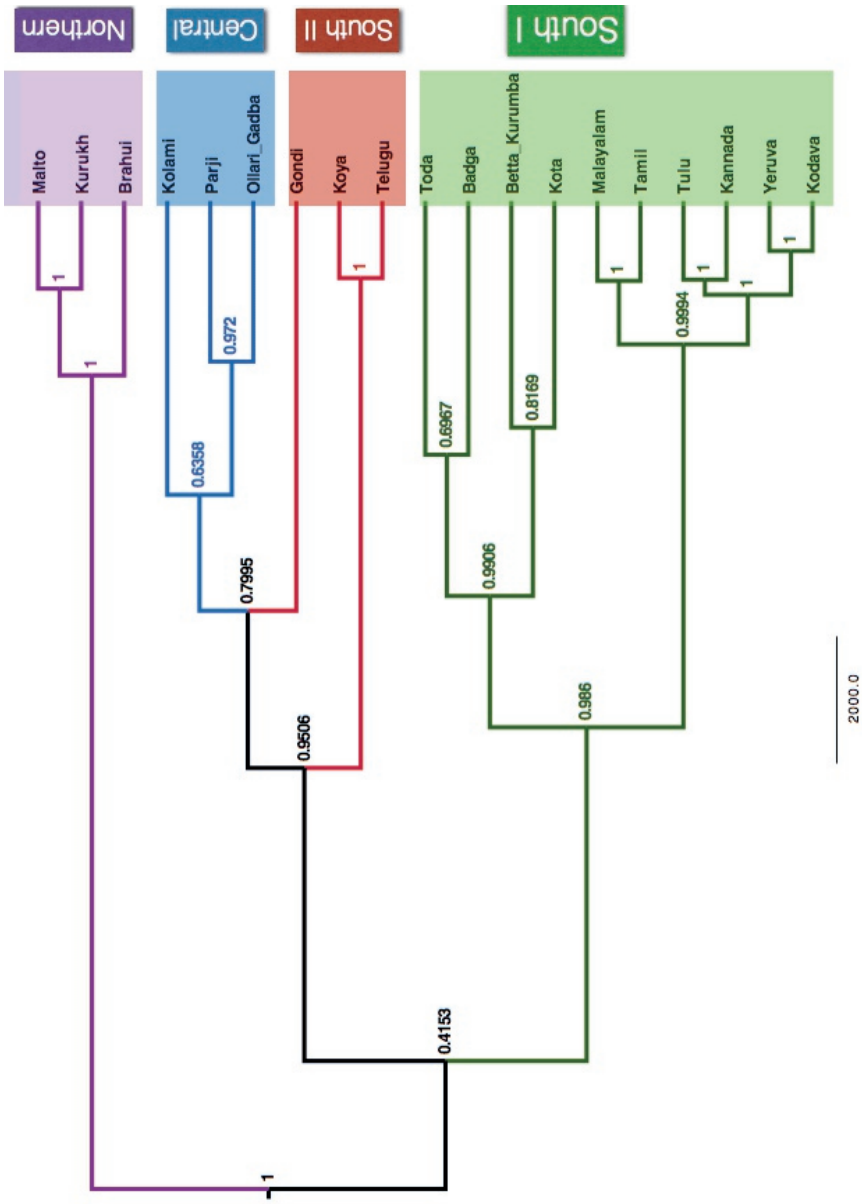


Figure 5–6: Maximum Clade Credibility tree of the Dravidian language family inferred by modeling the lexical data in a Bayesian framework under the relaxed clock with Dollo substitution model. The numbers at the nodes represent confidence of branching (posterior probabilities)

In Figure 5-6, it is apparent that the northern branch of the Dravidian language family is reconstructed with a great degree of confidence. The branching

pattern shows a recent divergence of Brahui from Malto and Kurukh, where the latter two form a monophyletic group. This coincides with the relationship put forth by earlier linguists regarding the relationship between the three northern languages (Emeneau, 1967). Andronov (1964), proposed that Brahui was separated from the rest of the Dravidian languages by 5000 years, based on lexicostatistics. But according to Krishnamurti (2003:491), "The misleading time depth [of Brahui] is caused by loss of many cognates in Brahui because of heavy borrowing from Balochi and Indo-Aryan". Also, Elfenbein (1998:389) says about Brahui, "in terms of shared phonological and morphological innovations, it could not have been separated for more than a thousand years or so from Kurukh–Malto.". In our study too, there is no evidence to support the hypothesis of Brahui being an ancient lineage and/or the first split from the Proto-Dravidian root. Why might Andronov's (1964) claims about Brahui's ancient status not hold? First, the data for his analysis was taken from dictionaries prepared by the then-British rulers, the accuracy of which cannot be accounted for. This might have influenced date estimates, as Brahui when documented had been in contact with Balochi for over 1000 years, and had borrowed a majority of the lexicon from the latter (Krishnamurti, 2003). If the documentation did not differentiate Brahui from Balochi, or did not account for borrowings from other languages and eliminate them from the analysis, it would have inflated the distance between Brahui and the rest of the languages. The data in this study is from a wordlist collected from a contemporary native Brahui speaker, which has been checked for borrowings, and hence is probably more accurate. Current analyses did not reflect Brahui as the earliest language to split from Proto-Dravidian.

Labov (2007) posited that groups often use language to establish identity by differentiating their language from their neighbours and it has also been shown that if this linguistic social change is occurring, then there should be a diversification of languages in punctuational bursts, followed by gradual divergence (Atkinson *et al*, 2008). This process would establish initial short branches, at the nodes, followed by long branches. A very similar phenomenon with the Northern Dravidian subgroup is observed from the current results. While there is need for substantiation for this hypothesis by placing calibrations and constraints on the tree with age estimates, the emergence of the Indo-European languages in India may well have stimulated the diversification of this branch of Dravidian languages. The findings are also in line with Krishnamurti's (2003) argument, that there is no basis for delineating Brahui as an ancient lineage descending from Proto-Dravidian, but that Dravidian languages were probably spread throughout northern India, and their range contracted and diversification ensued following the spread of Indo-European languages throughout the Indian subcontinent. The analyses in our forthcoming paper (Kolipakam et al, in prep) also suggest a similar timing for the split and divergence of the North sub-group, at the root of the tree, at the beginning of the third millennium BCE.

As explained earlier, the split of South I from South II and Central Dravidian was not highly supported, but the maximum clade credibility tree grouped the South II branch with Central Dravidian. Telugu and Koya form a cluster, while the Central Dravidian groups segregate together, with Ollari-Gadaba and Parji sharing a common ancestor, with Kolami forming a paraphyletic group, as expected. The languages that were thought to belong to each group *a priori* segregated accordingly, forming monophyletic groups, except for Gondi. Krishnamurti grouped Gondi with South II, but the inferred MCC tree groups Gondi to share a common ancestor with the Central Dravidian languages. However, Krishnamurti observed several small clusters within South II, and our lack of representation from all of these clusters might have reflected in Gondi assuming an outlier position. The placement of Gondi needs to be inspected more closely, and more languages from this region (Manda, Pengo, Kui) are needed to resolve the uncertainty in the position of Gondi.

In the South I Dravidian clade, there was a clear demarcation between the Nilgiri (Toda, Badga, Kota and Betta Kurumba) languages and other languages. Krishnamurti (Krishnamurti, 2003; Krishnamurti and Emeneau, 2001) clustered Kodava, Kurumba, Toda and Kota closer to the Tamil Malayalam node, while separating out the Kannada, Tulu and Badaga languages (Figure 5-6). However, phylogenetic reconstruction of these subgroups placed Kodava, Yeruva, Kannada and Tulu as sister clades to Malayalam and Tamil, while the remaining languages maintained a unique cluster. Krishnamurti however indicated that the position of these languages within the subgroup was ambiguous. With this result, the amount of uncertainty in the grouping was measured to some extent and it was possible to resolve the relative positions of the languages within the subgroups. The posterior probability of the internal grouping of the South I languages was highly supported, with most of the branches having a probability of greater than 90%, indicated high degree of certainty of the position of the languages within the South I cluster. Kodava, whose position was considered to be ambiguous, clustered with the Tamil–Malayalam-Kannada-Tulu-Yeruva node in 100% of the posterior tree sample. This grouping was also tested by constraining the tree evolution, as explained below (Section 5.6.2).

5.6 Testing the relative positions of the different pre-classified subgroups

To test the evidence in favour of different subgrouping hypotheses, constraints were placed on tree topology and likelihood inferred of these different tested constraints. The most likely model of substitution, inferred while building the Dravidian phylogenetic tree was used in all the subsequent tests/models on the data set. Each model was run in BEAST (Drummond and Rambaut, 2007) with the following parameters: 2 million iterations, sampled every 5000 steps, with a burn-in of 100,000 iterations and for at least three independent replicates. The runs

were compared using *Tracer* (Drummond and Rambaut, 2007) to check for consistency across independent replicates of each model. The effective sample size (ESS) of all parameters and the likelihood traces of the independent replicates were compared. If the traces of the likelihood values were not falling in the same distribution range or if the ESS values were low, then three more independent runs of the model were performed. Once the runs reached convergence, and the ESS were satisfactory, a Bayes Factor test was performed to determine the best performing model.

5.6.1 Position of the different subgroups

Three different constrained models were mainly tested, to understand the broad divisions of the proposed four clusters of the Dravidian language family - North, Central, South I and South II.

a) Geo_model: All four proposed branches of the family were forced to evolve separately from each other, but coherently with the languages within the group
    a. Northern: Brahui, Malto and Kurukh
    b. Central: Kolami, Parji, Ollari_Gadaba
    c. South II: Telugu, Koya and Gondi
    d. South I: Toda, Badga, Betta_Kurumba, Kota, Malayalam, Tamil, Tulu, Kannada, Yeruva and Kodava
b) SouthII_Central: This model constrained only the languages of South II and central Dravidian to be monophyletic, i.e., Kolami, Parji, Ollari_Gadaba, Telugu, Koya and Gondi, while the remaining languages were unconstrained.
c) South_south2: In this model, South-I and South – II branches were constrained to evolve together.

5.6.1.1 Results

The South II_Central model was slightly more preferred than the Geo model (Table 5-5; see Figure 7-2 in appendix 7.2 for MCC tree), where the latter had each of the four designated subgroups constrained to be monophyletic, versus the South II_Central model that had the South II and Central languages constrained to cluster together. The South I_South_II model clustered South I and South II together to the exclusion of the Central and North group.

The grouping of the South II clade with the Central clade implied that Krishnamurti's re-classification of South II, by separating it from South I and indicating that it was probably much more closely related to Central, was indeed sound (Krishnamurti and Emeneau, 2001). As seen in the unconstrained model

presented in section 5.5.3, Gondi of the South II group seems to emerge with the Central Dravidian group, while Koya emerges as a clear sister branch to Telugu. However, Koya is considered a dialect of Gondi by linguists (Tyler, 1969). One reason could be the limitation with regard to the coverage of Gondi languages in our dataset. This is probably because our dataset is limited with regard to coverage of Gondi languages. Gondi is spoken in four Central Indian states (Maharashtra, Madhya Pradesh, Andhra Pradesh and Orissa), and has extensive dialect variation due to its geographical spread and lack of a written tradition (Steever 1998). The Gondi dialects are only mutually intelligible when they are geographically adjacent (Tyler 1969: 3). Our sample of Gondi is from Northern Gondi spoken in Madhya Pradesh, close to the border with Maharashtra – far away from the lands of the Koyas (Figure 5-1). Koya is spoken in the Telugu heartland, and has been influenced by Telugu in various ways (Tyler 1969, Krishnamurti 2003). Our classification of Koya must be a reflection of the shared history Koya has with Telugu. Only a language sample which includes various other Gondi languages can shed further light on the relationships of the Gondi languages and their connection with Telugu. Even though South II_Central model is weakly preferred over Geo model, the true test was between the former and the South I_South II model. This resolves the question of the affinity of South II Dravidian.

Table 5-5: Table depicting the pairwise comparison of the performance of the different hypotheses of Model 1 of the Dravidian lexical dataset as evaluated by Bayes Factor units

Model	ln P(model)	S.E.	Geo	South II_Central	South I_SouthII
South II_Central	-3886.71	+/- 0.161	1.98	-	6.188
Geo	-3891.268	+/- 0.155	-	-1.98	4.209
South I_South II	-3900.959	+/- 0.174	-4.209	-6.188	-

5.6.2 Position of Kodava

We tested the two competing hypothesis for the position of Kodava by constraining the evolution of Kodava with a) Tamil and Malayalam and b) Kota and Toda.

The BF test revealed a strong likelihood for the grouping of Kodava with Tamil and Malayalam. Even in the results from the unconstrained model, in the entire posterior distribution of the trees, Kodava grouped with Tamil and Malayalam rather than with Kota and Toda. In addition, Kodava and Yeruva were reconstructed as a sister branches to Tulu and Kannada.

Table 5-6: Table depicting the pairwise comparison of the performance of two hypotheses, testing the robustness of grouping Kodava with Kota and Toda or Tamil and Malayalam

	lm P(model)	S.E.	Kodava_ Kota & Toda	Kodava_Tamil & Malayalam
Kodava_Tamil & Malayalam	-3924.29	+/- 0.159	16.293	-
Kodava_Kota & Toda	-3961.8	+/- 0.172	-	-16.293

Krishnamurti, and Emeneau (1967) grouped Kodava as sharing the most recent common ancestor with Kota, Toda, Tamil and Malayalam. It has clearly emerged from our results that Kodava is a sister branch to Tamil and Malayalam, with a much closer affinity towards Kannada and Tulu than previously thought (Table 5-6).

5.7 Testing the sequence of evolution of the different branches of the Dravidian language family

To test the divergence of branches and hypotheses pertaining to the evolution of the different branches of the Dravidian language family, we posed the following constraints on the language groups and tested different hypotheses.

a. N_CS2_S1

In this model, the first split from Proto-Dravidian differentiated North Dravidian (which then evolved into the current North Dravidian branch) from the rest of the subgroups. The second split, which included Central and South, further split into one branch containing Proto-Central and South II and the second split contained Proto-South I. While the Proto-South I evolved into the present day South I, the Proto-Central and South II first split into Proto Central and Proto South II, before evolving into the current languages of Central Dravidian and South II Dravidian. This model allows us to test if South II Dravidian is more closely related to Central Dravidian than to South I Dravidian (Figure 5-7).

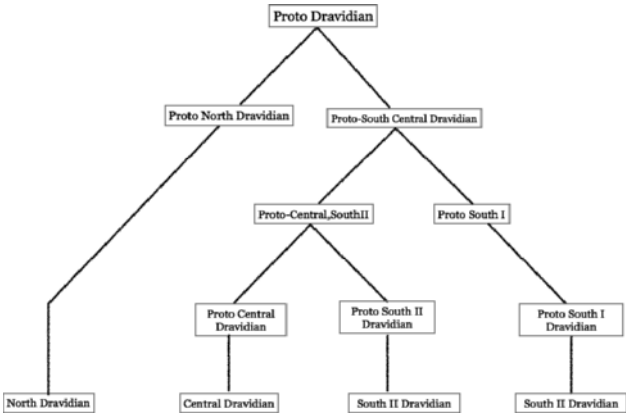


Figure 5–7 : Phylogenetic depcition of N_CS2_S1 model of evolution of Dravidian language family, used as a constrained model for testing against other hypotheses

b. NC_SS

Here we tested the strength of association of South II with South I, instead of Central Dravidian. The split of north Dravidian was modelled first, with the remaining three subgroups having a common ancestor: Proto-South-Central Dravidian. From Proto-South-Central Dravidian, the Central-South II branch evolved and the Proto-South node then split into South I and South II See Figure 5-8.

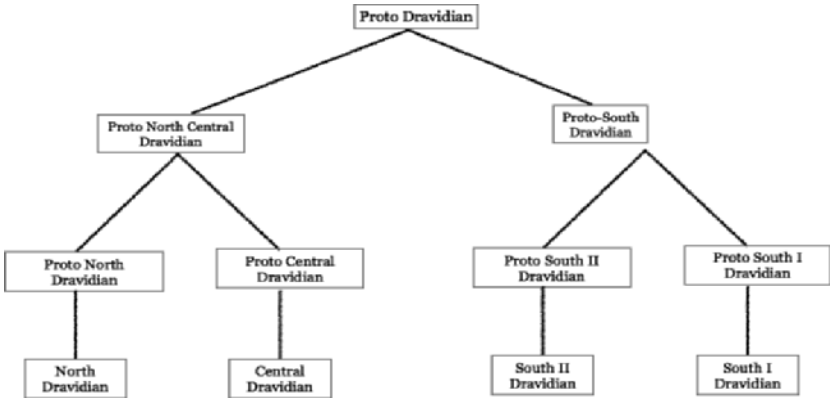


Figure 5–8: Phylogenetic depcition of NC_SS model of evolution of Dravidian language family, used as a constrained model for testing against other hypothese

c. N_CS_SS

The main aim of this model was to test whether the Central branch of the Dravidian language family shared the most recent common ancestor with North Dravidian rather than with the southern branches (as was the case in the previous two models, See Figure 5-9).

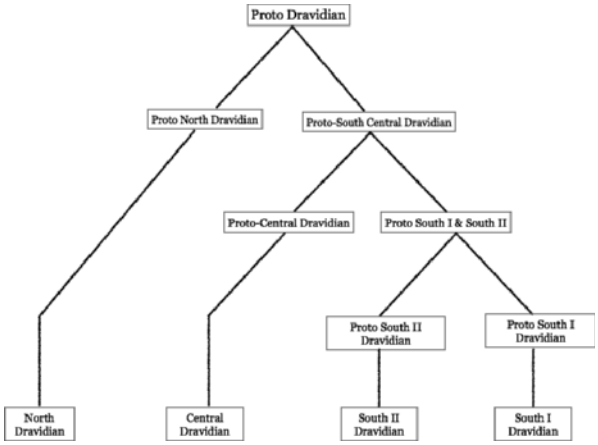


Figure 5–9: Phylogenetic depcition of N_CS_SS model of evolution of Dravidian language family, used as a constrained model for testing against other hypotheses

d. NCS2_S1

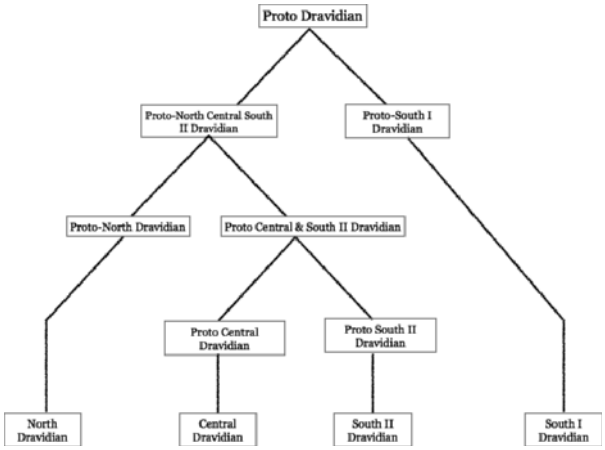


Figure 5–10: Phylogenetic depcition of NCS2_S1 model of evolution of Dravidian language family, used as a constrained model for testing against other hypotheses

Finally, in this model, it was tested if a Central and South II Dravidian grouping evolved from a common ancestor with the Northern group, instead of the South I Dravidian group. A split of Proto Dravidian into Proto-North Central South-II and Proto-South I was modelled, see Figure 5-10.

5.8 Results

The N_Cs2_S1 model emerged to be the most plausible topological model (Table 5-7) of evolution for the Dravidian language family (for MCC tree, see Figure 7-3 in appendix 7.2). The hierarchical models testing the evolution of the subgroups within the Dravidian language family, revealed several critical pieces of information. Firstly, the northern clade split from the rest of the clades first. Secondly, the central and South II branches of the Dravidian language family shared a common ancestor. South I and South II are paraphyletic and not monophyletic as previously suggested. The MCC tree of the best-supported model (See appendix 7.2, Figure 7-3) showed that the split into three branches (Central, South I and South II) happened very rapidly, and thus was rather like a trifurcation, instead of sequential bifurcation. While Krishnamurti argued against a common stage for South-II and central branches, in both constrained and un-constrained models, we found that there was a strong possibility for a common stage for South II and Central branches.

Table 5-7: Table depicting the pairwise comparison of the performance of the different hypotheses of Model 2 of the Dravidian lexical dataset as evaluated by BayesFactor units

Model	ln P(model)	S.E.	N_CS2_S1	NC_SS	N_CS_SS	NCS2_S1
N_CS2_S1	-4073.817	+/- 0.989	-	12.195	10.245	2.821
NCS2_S1	-4080.312	+/- 0.801	-2.821	9.374	7.425	-
N_CS_SS	-4097.408	+/- 0.897	-10.245	1.949	-	-7.425
NC_SS	-4101.896	+/- 0.811	-12.195	-	-1.949	-9.374

Unsurprisingly, the best performing topology model resembles the subgrouping in our MCC tree. The N_CS2_S1 model has Central and South II forming an exclusive subgroup. The next best performing topology model is, surprisingly, NCS2_S1, where the main division in the tree is between South I and North+Central+South II. This suggests that the close affiliation between Central and South II is more important than the higher order branching of North, South I, and Central+South II. This is confirmed by the structure of the worst performing topology model, NC_S2S1, in which Central and South II are placed furthest apart.

5.9 Summary & Conclusions

The current study is the first attempt at inferring Dravidian language phylogeny by using contemporary linguistic data in a quantitative evolutionary framework. By applying phylogenetic methods to linguistic data, I was able to answer some of the important unanswered questions regarding Dravidian language pre-history and test hypotheses regarding the evolution of the Dravidian language family. The position of certain languages within the language family, along with a relative chronology to the spread and diversification of this language family was also ascertained.

From these results, it is established that Brahui is not an ancient lineage as has been sometimes claimed, but has emerged from a relatively recent divergence from the rest of the Northern Dravidian subgroup (Cavalli-Sforza *et al*, 1994; Krishnamurti, 2003; Zvelebil, 1972). If one might speculate, the erstwhile distribution of the Dravidian language family across northern India was intact until a recent change that led to its restricted distribution along with the diversification of the extant languages. The divergence of the different northern clades into their current state seemed to have coincided with that of the diversification of the Indo-Iranian branch of the languages that are prevalent in India (Gray and Atkinson, 2003). This coincides with Krishnamurti's (2003) conjecture regarding the Dravidian languages of northern India being slowly replaced by Indo-European over generations. The surviving Dravidian languages in the north are Malto, Kurukh and Brahui, and are all spoken by geographically isolated communities. Isolation and population size might be a factor for their persistence, as language has been known to be used as group identification, and this might increase the conservation in its form, especially when surrounded by other linguistic groups (Labov, 2007). However, we also see that the diversification between the three North Dravidian languages is quite low, and given their isolation and geographic distance from each other, it could be argued that isolation should lead to faster diversification of languages (Gray *et al*, 2009). Evidence also suggests that smaller communities have denser social networks (Bowern, 2010; Trudgill, 2011), and this would slow down the spread of innovations and as a consequence the rate of diversification from its sister lineages other languages (Granovetter, 1973). This is applicable to the small communities of North Dravidian speakers, for example Malto, where the community is of around 12,500 people and the number of speakers is unknown (Bhaskararao, 2006).

If the situation in Northern India of replacement of languages by Indo-European occurred, why did not a similar situation arise in southern India? Firstly, according to the "Continental Axis theory", an outcome of the Farming Language Dispersal Hypothesis, farming societies tended to expand east to west, rather than north to South (Diamond, 1997; Greenhill, 2015). Given that the expansion of the

Indo-European language family was most likely coupled with expansion of farming (Campbell, 2002; Diamond and Bellwood, 2003), and growing conditions fluctuate much more rapidly while moving latitudinally, the Indo-European speakers probably never attempted to venture beyond the Vindhyas. Laitin *et al* (2012) have shown how language persistence is greater in countries with north-south axes, and given the direction of spread of Indo-European, it is highly plausible that persistence of Dravidian languages in southern India could be explained by the continental axis theory. Also, as is evident today, probably in the past too, the Southern Dravidian languages were literarily and culturally more embedded (writing system, literature, etc.,), and encouraged to develop by the rulers, which made them much more difficult to replace (Goody and Watt, 1963).

More data from the Central Dravidian branch is required to confidently hypothesise the point of origin and diversification of the Dravidian languages, but given the current diversity of language distribution and sequence of evolution, Krishnamurti's (2003) and Fuller's (2003) hypothesis of Dravidian originating in south or central India and spreading across the central Vindhya Mountain range, seems plausible. To test the hypothesis that the Indus valley civilization could have been Dravidian or could have interacted and co-existed with Dravidian language speakers, a more detailed phylogeographic work, and dating of the tree needs to be undertaken.

Another hypothesis that was successfully tested was regarding the steps in the evolution and diversification of the different subgroups in the Dravidian language family. Strong evidence supporting a common ancestor for the South II and Central subgroups of the language family was observed. However, since some of the languages in the Central and South II branches are missing, this needs to be retested with more data. Nevertheless, ample support for Krishnamurti's separation of Telugu from South I, and positioning it as a distinct subgroup was present. The results also point to the need for a re-investigation into the hypothesis of a common stage for South and Central Dravidian, prior to their individual divergence.

Finally, I emphasise the advantage of using linguistic data in an evolutionary framework, to test hypotheses and to understand pre-history. The next step is targetted work to (a) attempt a more robust dating estimate of the Dravidian phylogeny (b) collect more language data) and (c) to see what biological anthropology/genetics and archaeology could contribute to our understanding of Dravidian history. In this Chapter, by bringing together linguistic data and evolutionary methodology, I have been able to make significant contributions to filling the gaps in the knowledge of Dravidian history as well as providing a foundation for further comparative studies.

6 Conclusion

I started the thesis with the intention of testing several hypotheses regarding human pre-history in an evolutionary framework, with information from biological and cultural traits. As discussed in Chapter 1, the evolutionary nature of cultural traits, in comparison to biological traits, has been a topic of debate and interest for over a decade now. The "curious parallels" exhibited by the evolution of these two traits (Atkinson and Gray, 2005) has been robustly established through several studies (Fortunato and Jordan, 2010; Jordan, 2007; Kushnick *et al*, 2014; Mace *et al*, 2003; Murdock, 1940; Tehrani and Collard, 2002; Tehrani and Collard, 2009; Watts *et al*, 2015a). This has resulted in a flurry of studies employing evolutionary methodology, using insights regarding cultural evolution to explain human diversity, and using data from human genetic diversity to explain cultural evolution of human societies (Besaggio *et al*, 2007; Hamilton *et al*, 2005; Kayser *et al*, 2006; Oota *et al*, 2001). Culture has since been considered a "force of history", and in turn, claims have been made to show that it has shaped the genetic diversity of human populations as we see it today. The co-evolutionary and co-dependent nature of genes and culture has contributed to the discovery and hypotheses of several facets of information regarding pre-historic human societies (Gunnarsdottir *et al*, 2011; Hage and Marck, 2003a; Hammer *et al*, 2008). The focus of this thesis was to a) understand the dynamics of interaction between genes and culture and test the robustness of this relationship and b) determine how the triangulation technique, especially drawing information from the evolutionary nature of cultural traits, can be used to understand pre-history.

6.1 Genes and Culture: the extent of their spheres of influence

Cultural traits like post-marital residence have been hypothesised to influence the SSM diversity seen in human populations today (Oota et al., 2001). In the Pacific, a discrepancy in the SSM diversities of mtDNA and MSY was hypothesised to be driven by the social norm of post-marital residence, leading to sex-biased origin of the Polynesians (Kayser *et al*, 2006; Kayser *et al*, 2000). Conclusions regarding the sex-biased origin of present day Pacific islanders and many other societies were hypothesised based on similar patterns of SSM (Besaggio *et al*, 2007; Oota *et al*, 2001; Ségurel *et al*, 2008). But social norms do not fossilize, and conclusions about the influence of social norms on human genetic diversity have been at best speculative or based on very broad comparison with non-human primates (Foley and Lee, 1989). However, humans are outliers in primate cultural diversity, which makes cross-species comparison correspondingly difficult. While parallels are

useful, humans differ from non- humans on many potentially relevant parameters simultaneously. It is imperative to critically assess the robustness of such inferences regarding the nature of interaction between culture and genes. Until recently, we lacked the theoretical and methodological frameworks that were required to test the influence of cultural norms on genes. With the recent advancements in comparative methodology (Holden and Mace, 2003; Jordan and Shennan, 2009; Mace and Pagel, 1994; Pagel, 1999b), this "virtual archaeology" (Mesoudi, 2011) provided us with a framework for rigorous testing. Chapter 2 and Chapter 3 focussed on the aspect of post-marital residence, its evolution and influence on SSM genetic diversity. Chapter 2 explored the effects of residence on genetic diversity, and through Chapter 3, I delved deeper into the evolution of residence norms and the influence of other cultural traits on residence.

Advances in molecular tools helped me simulate the movement patterns of SSMs in correspondence to different post marital residence practices in the Pacific and as a result, I was able to test and comprehend the relationship between cultural traits and genetic diversity (See Chapter 2). To tease out the cause-effect parameters in the gene-culture relationship, forward simulations proved to be a useful tool, allowing all other factors that can influence the patterns or movements of genes to be kept constant or varied according to choice, unlike a real world scenario. Through this experiment it was evident that i) social practices of post-marital residence did influence SSM genetic diversity, in a predictable and anticipated manner, and ii) these predictions do not hold good in the face of stochastic demographic processes like drift. Therefore, the take home message is that caution must be observed in making direct correlations between genes and particular post-marital residence practices without taking cognisance of the evolutionary forces at play.

Studies have also attributed to the lack of congruency in SSM diversity and post-marital residence practise to the presence of a "time lag" (Gunnarsdóttir *et al*, 2011). This "time lag" refers to the duration from when the post-marital residence in a society changes to time when the SSM diversity reflects this change. We were able to quantify the "time-lag" genetic markers take to reflect the corresponding changes in residence to be between five and twenty-five generations. However, this effect also was valid only in populations where the effect of drift was negligible. Thereby showing that if a population underwent a bottleneck or drastic population reduction, this relationship between SSM diversity and prediction of post-marital residence becomes intractable.

When testing this relationship on Pacific populations, it was found that the currently patrilocal populations of Fiji, Tuvalu and Tonga exhibited SSM genetic diversity resembling matrilocal populations. This would mean that a recent change in the post-marital residence practice of these societies (within the time-lag phase of 5-25 generations) could be the most likely explanation for the pattern observed (as none of the populations exhibited signs of decline or drift). In the

period following these societies' colonization of the islands, a change in climatic conditions (Anderson and Clark, 1999) and change in subsistence (Enright and Gosden, 1992), substantiated our hypothesis of a change from matrilocality to patrilocality in recent times (650-750 years ago). A further detailed sampling and analyses of these societies' history (using coalescent simulation models) and a wider survey of this pattern, in societies with good records or documentation of post- marital traditions, would help consolidating the evidence for the time-lag hypothesis. From this effort of quantification and teasing apart forces that drive the relationship between SSM diversity and post-marital residence, it is clear that a great degree of caution and taking cognizance of demographic and evolutionary forces is necessary to interpreting evidence of social influence on the genetic structure of a society. It is crucial to be aware that tractable insights into gene-culture correlations are only valid when the effect of stochastic processes are negligible. An estimation of this effect is an important step before any inference based on this correlation is drawn.

In Chapter 3, the hypothesis of another cultural trait thought to co-evolve and influence post-marital residence practices directly, while indirectly driving SSM diversity was tested. It was proposed by Murdock (Murdock, 1949b) and Driver (Driver and Massey, 1957) that post-marital residence is primarily influenced by elements like sexual division of labour towards subsistence, against the background of marriage ecology. By taking advantage of the evolutionary nature of cultural traits, I was able to explicitly test for this relation, while controlling for historical relatedness amongst societies. The results showed no evidence to either support or reject Murdock (Murdock, 1949b) and Driver's (Driver and Massey, 1957) predictions. Instead, based on the results of the present analysis on presence of phylogenetic signal of traits, my approach to the test of correlation or coevolution between cultural traits in these societies was perhaps at too coarse a level to account for variation. Subsistence and societal organisations are factors that are deeply dependent on environmental factors. Especially when societies are colonising new areas and new environments, and they are adaptation to the environment, can result in a cultural change. I propose that after accounting/controlling for ecological change, phylogeny and relevant marriage dynamics, the hypothesis of co-evolution of sexual division of labour and post-marital residence needs to be revisited. The ability to infer the dynamics of cultural traits is prized and gives an insight into the social organization of the past societies, which is crucial in deciphering past history.

The lack of congruency in SSM pattern diversity and post-marital residence (mentioned above) in Fiji, Tonga and Tuvalu, revealed that a change in climatic conditions drove a change in subsistence and eventually post-marital residence. Archaeological evidence pointed to a substantial resource depletion around 2100 YBP, and combined with a change in climatic conditions (little ice age, Anderson, 2002; Field and Lape, 2010), affected the mode of subsistence and sexual division

of labour to subsistence in these communities (Enright and Gosden, 1992). Rising sea levels, and increasing population size, and loss of resources led people to turn to intensive agriculture from fishing. With male contribution increasing with increase in intensive agriculture, and reasons for male-absence reduced (reduced long-distance sea- voyage), the main hypothesised drivers of matrilocality disappeared and populations evolved to practising patrilocality, a more stable state in these conditions (Jones, 2009). We find evidence for these factors and change in post-marital residence to have coincided through the genetic data in Chapter 2. From the co-evolutionary analyses in Chapter 3, we deduce the need to investigate the effects of sudden environmental change on cultural traits. Once the dynamics of trait evolution is teased apart, we can test the presence of a co-evolutionary relationship between traits. We would then need to confirm that against the backdrop of marriage ecology, and test how sexual division of labour towards subsistence drives post-marital residence norms. This would throw light on our hypothesis that the change in subsistence in recent times due to change in resources, could have well driven Fiji, Tuvalu and Tonga, to change their post-marital residence. Since this change could be recent enough, within the time lag phase, we do not see the SSM diversity matching with the practised post-marital residence.

These findings could also help predict social aspects of pre-historic societies, which are often not known, when mode of subsistence or sexual division of labour could be ascertained. The quantification of the time-lag phase in societies could also add to the information on the process of change in post-marital residence pattern, as well as garner information regarding correlated elements like introduction of agriculture, trade and change in lifestyle (Marlowe, 2004). Further insights from investigating marriage ecology would be invaluable in understanding this equation. Further work could show how far this process is generalizable after incorporating environmental variables and its influence on social organisation and accounting for historical relatedness. The work from Chapter 3 neatly ties into the consilient nature of all aspects of a human society that needs to be accounted for while explaining the observed human genetic diversity.

6.2 Triangulation and rigorous hypothesis testing is crucial for robust inferences

Thomas Kuhn (1970) (Pigliucci, 2007), described paradigms as being a framework involving a set of ideas and principles that allows a scientist to solve puzzles. Also, these paradigms work well for a long period of time, until a new discovery or phenomenon questions the very basis of the paradigm which cannot explain certain processes. This situation can be compared to the discovery of differential male and female origins for the present day Polynesians (Kayser *et al*, 2006), which was latter attributed to their post-marital residence practice. This

rationalisation was initially used to explain the differential diversity in SSMs in Thailand (Oota *et al*, 2001). So, according to this paradigm, an explanation for contradicting SSM diversity was a result of post- marital residence practices. Atkinson (2006) stressed the importance of rigorous hypothesis testing by coalescent based simulations to further validate the models of evolution or dispersal history proposed through different lines of evidence. While we were able to establish that indeed post-marital residence does have a detectable effect on genetic diversity of SSM markers, tests also confirmed that this relationship is not always predictable. Stochastic and demographic effects do throw patterns of SSM out of sync with post-marital residence. Using this knowledge, and taking the advantage of coalescent modelling, in Chapter 4, we were able to further shed light on the quantum of influence post-marital residence practices had on the SSM marker diversity of societies in the Pacific.

At the outset, I tested the hypothesis that male and female admixture histories of the present day Polynesians were different. For this, I first determined the contribution of different "ancestral" populations to present day Polynesians. It was found that the entirety of the RO population diversity was a subset of the NO-AN diversity and that the actual admixture estimations of the NO-AN populations shed more light on the colonization of RO. In essence, the partitioning of populations on the basis of their cultural and linguistic traits than geographical or political boundaries revealed crucial information on the admixture history of the Remote Oceanians. Studies till now focused only on Near and Remote Oceania, but did not consider the diverging histories of Austronesians and Non-Austronesians in Near Oceania.

For both males and females, there were signals of Asian ancestry and gene flow from NO-NAN, and we did not find any evidence for sex-biased origin as previously thought. Coalescent based modelling provided information that the Polynesian populations were only but a subset of NO-AN and based on these results we tested the ancestral contributions to NO-AN. The coalescent-based admixture estimator pegged the NO-NAN contribution to the NO-AN gene pool as 0.92 for male and 0.76 for female. This uneven contribution towards male and female gene pools could be due to the ancestral matrilocality of the ancestral Austronesians, as proposed. While, the presence of sex-biased admixture could not be ruled out, there was no evidence for the highly skewed affinities and origins of men and women as understood before (Hage and Marck, 2003; Kayser *et al*, 2006). There seems to be gene-flow of both males and females between Austronesians and Non-Austronesians, even if differentially restricted. This differential gene-flow could be attributed to the fact that immigration is more tightly regulated in patrilocal than matrilocal societies (Hamilton *et al*, 2005). More importantly, current results show that evidence that the diverging genetic affinities of MSY and mtDNA that other studies have found (Hage and Marck, 2003; Hertzberg *et al*, 1989; Hurles *et al*, 2002; Kayser *et al*, 2006a; Lum *et al*, 1998), was

most likely due to bottleneck events (~2000 YBP) during the colonization of Remote Oceania, which probably affected the two sexes differently. The effect of a bottleneck means smaller effective population sizes, and larger effects of drift. As we saw in Chapter 2, this means, the SSM pattern that we see does not necessarily match the post-marital residence practise of the society of interest. However, I re-iterate that the affinities of the Polynesians for males and females were essentially the same.

Another interesting finding was that coalescent and admixture based estimates revealed support for the long pause of the Austronesian populations at NO, before RO was colonized. Coalescent based modelling showed support for a splitting event of the Polynesian population in lieu of an admixture event with NO-NAN, during the colonization of RO. The findings from this study corroborated the pause-pulse scenario put forward by linguists regarding the colonization of Polynesia. There was a long pause of about 2000 years, the bottleneck, before the colonization of Polynesia (Gray *et al*, 2009). Along with this, evidence for the male and female affinities to date back to Asia and particularly to Taiwan was observed. The signals of admixture also reveal affinities to the first Austronesian settlers of ISEA, NIAS as well as from other populations in ISEA and WNG. This showed that the Austronesian dispersal happened with cultural integration across their route of dispersal, thereby finding support for the proposed "VC triple I" model of dispersal proposed by Green (2003). Coalescent based modelling further provided additional substantiation of this hypothesis, by placing the ancestral homeland of the present day Polynesians to be from Taiwan.

Coupled with the evidence from modelling and by other admixture estimators, and pooling the populations into historically coherent groups, no evidence was found for a strong sex-biased dispersal. The fact that there was a bottleneck with severe founding effect, which affected males and females differently and ample evidence of Taiwan being the ancestral homeland of Polynesians, the most viable theory of dispersal was the "VC Triple I voyaging corridor" hypothesis. The evidence supported the cultural and genetic integration of the Austronesians along the route of dispersal and the ancestral homeland of Taiwan put forth by the VC Triple I hypothesis. There is need for further work to investigate the Near oceanic populations in more depth, and to also collate and include data from New Caledonia and other regions of Polynesia where information is sparse. I hope to have shown that by taking cognisance of history while testing a hypothesis, and including information from all lines of evidence, gives an insight into several previously unknown facets, provides a robust way of testing proposed hypotheses (pause- pulse scenario, VC Triple I hypothesis, sex-biased origin), and also a more comprehensive picture of history. If the cultural and linguistic aspects of Pacific pre-history were not incorporated into model testing, these valuable insights could not have been found.

6.3 Linguistic data – of utmost importance in deciphering human history

In India the signals of admixture history are not clear, genetic history is far too admixed and is only informative at a broad scale (Cordaux *et al*, 2004; Reich *et al*, 2009). Linguistics can play an important role in understanding events that have led to the current state of human diversity in such societies, where archaeology and genetics have failed to delineate zones of contact and clearly differentiate different waves of human expansion.

The patterns of language divergence have shown to be indicative of the forces of history that have moulded them to their current state. The evolution of language is not, contrary to common belief, an intrinsic nature of language itself (semantics, grammar etc,) but rather reflects the demographic, social and cultural and political forces that have acted on the communities that speak these languages (Heggarty and Beresford-Jones, 2010). This cause- effect relationship is therefore an invaluable surviving record of the linguistic history as a function of what and how these different forces have acted and shaped human societies

In the Pacific, linguistics has helped immensely in contributing information regarding the dispersal of the ancestral Polynesians by understanding the evolution of the Austronesian language family (Blust, 1988; Gray *et al*, 2009; Gray and Jordan, 2000; Lum *et al*, 1998). In India, as it is harder to make sense of history using gene admixture, and in such a scenario, linguistic evidence plays an important role in deciphering and understanding pre-history. In Chapter 4, I model population histories and hypotheses emerging from genetic, linguistic, anthropological and archaeological data, using genetic data but taking into cognisance knowledge from all the other lines of evidence. In Chapter 5, I took advantage of the evolutionary nature of language to model population history, where archaeological and genetic data is confusing at best. By using phylogenetic methods, the spread of Dravidian languages of India is investigated and in turn information regarding the spread of their speakers is garnered. In addition, phylogenetic methods allowed us to test several hypotheses that were proposed regarding this language family.

The efforts invested in Chapter 5 resulted in the first phylogenetic tree of the extant Dravidian languages, based on original data from the 100-word Swadesh list collected and coded as part of this project. Using a phylogenetic framework helped test several hypotheses regarding the Dravidian language. Contrary to previous expectations (Zvelebil, 1970), Brahui was not found to be an ancient lineage but rather a recent divergent from the northern clade of the Dravidian languages. The divergence of the different sub-clades of the Dravidian language family tree places the root of the Dravidian tree to be located in south/central India, with a later divergence to the northern sub clades. Given the many different hypotheses regarding the origin and spread of Dravidian language family, it seems likely that the hypothesis of a central Indian origin for Dravidian

language family (Fuller, 2007), with subsequent spread towards the rest of India seems highly plausible. A more rigorous testing with dates from archaeobotanical records, more sampling from languages of central Dravidian language family is needed. However, this test has proven how important language data is in deciphering human pre-history when other lines of evidence are limited.

This may be evidence supporting the continental axis theory (Diamond, 1997; Greenhill, 2015). According to this theory, the spread of humans along a longitudinal gradient was far more probable than along a latitudinal gradient. Evidence for this is observed in the intact distribution of the Dravidian languages of South of India, whereas the Indo-European languages have their range from the western most to the eastern parts of India. This work also fills in the gap on several unanswered questions regarding the sub-branches of the Dravidian language family, and provides and insight into population dynamics. This information is invaluable as prehistoric societies rarely leave evidence regarding the linguistic affinities or social traits.

6.4 Consilience – the key to a holistic picture of history

In the search to explain the origins of human populations, each discipline offers a partial but unique insight into our past. However, when all the relevant threads connecting the different lines of evidence are untangled and integrated into a common framework, the whole picture emerges much more clearly.

The Pacific and Dravidian regions represented two contrasting contexts: in terms of both the time depth of events under investigation, and of the evidential constraints that each system brings with it. But a consilient framework of language and biology has helped integrate the systems of culture (including language) and biology together in understanding and testing hypotheses regarding human pre-history in a robust framework. I also hope to have shown that it is very important to incorporate information from the different lines of evidence present to gather a holistic picture and that by not doing so we are most likely to make erroneous conclusions. The inferences drawn here have substantially contributed to our knowledge of Pacific and Dravidian pre-history.

I hope that I have demonstrated the importance of model testing and hypothesis validation. It must take a larger role in understanding pre-history, especially by making use of the integrative, comparative, phylogenetic and molecular tools available today. The information from the non-biological realm cannot be discounted while studying pre-history and a consilient framework is the key to understanding the diversity of human societies present today.

7 References

Abbott RJ, James JK, Milne RI, Gillies ACM (2003). Plant introductions, hybridization and gene flow. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **358**(1434)**:** 1123-1132.

Alkire WH (1960). Cultural adaptation in the Caroline Islands. *The Journal of the Polynesian Society* **69**(2)**:** 123-150.

Allen MS, Johnson KT (1997). Tracking ancient patterns of interaction: Recent geochemical studies in the southern Cook Islands. *Prehistoric Long-Distance Interaction in Oceania: An Interdisciplinary Approach New Zealand Archaeological Association Monograph* **21:** 111-113.

Anderson A, Clark G (1999). The age of Lapita settlement in Fiji. *Archaeology in Oceania***:** 31-39.

Anderson C (2002). Gender Matters: Implications for climate variability and climate change and for disaster management in the Pacific islands. *Intercoast Network* **41:** 24-05.

Andronov M (1964). Lexicostatistic analysis of the chronology of disintegration of Proto-Dravidian. *Indo-Iranian Journal* **7**(2)**:** 170-186.

Atkinson Q (2006). From species to languages: A phylogenetic approach to human prehistory. PhD thesis, University of Auckland.

Atkinson Q (2010). The prospects for tracing deep language ancestry. *Journal of Anthropological Sciences* **88:** 231-233.

Atkinson Q, Gray R (2005). Curious parallels and curious connections - Phylogenetic thinking in biology and historical linguistics. *Systematic Biology* **54**(4)**:** 513-526.

Atkinson Q, Coomber T, Passmore S, Greenhill SJ, Kushnick G (2016). Cultural and environmental predictors of pre-European deforestation on Pacific islands. *PLoS ONE* **11**(5)**:** e0156340.

Atkinson Q, Meade A, Venditti C, Greenhill SJ, Pagel M (2008). Languages evolve in punctuational bursts. *Science* **319**(5863)**:** 588-588.

Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, Alekseyenko AV (2012). Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Molecular Biology and Evolution* **29**(9)**:** 2157-2167.

Bamshad MJ, Kivisild T, Watkins WS, Dixon ME, Ricker CE, Rao BB *et al* (2001). Genetic evidence on the origins of Indian caste populations. *Genome research* **11**(6)**:** 994-1004.

Bamshad MJ, Watkins WS, Dixon ME, Jorde LB, Rao BB, Naidu JM *et al* (1998). Female gene flow stratifies Hindu castes. *Molecular Biology and Evolution* **395**(6703)**:** 651-652.

Basu A, Mukherjee N, Roy S, Sengupta S, Banerjee S, Chakraborty M *et al* (2003). Ethnic India: A genomic view, with special reference to peopling and structure. *Genome research* **13**(10)**:** 2277-2290.

Beaumont MA, Nielsen R, Robert C, Hey J, Gaggiotti O, Knowles L *et al* (2010). In defence of model-based inference in phylogeography. *Molecular Ecology* **19**(3)**:** 436-446.

Beaumont MA, Zhang W, Balding DJ (2002). Approximate Bayesian computation in population genetics. *Genetics* **162**(4)**:** 2025-2035.

Belle EMS, Ramakrishnan U, Mountain JL, Barbujani G (2006). Serial coalescent simulations suggest a weak genealogical relationship between Etruscans and modern Tuscans. *PNAS* **103**(21)**:** 8012-8017.

Belledi M, Poloni ES, Casalotti R, Conterio F, Mikerezi I, Tagliavini J *et al* (2000). Maternal and paternal lineages in Albania and the genetic structure of Indo-European populations. *European Journal of Human Genetics* **8**(7)**:** 480-486.

Bellwood P (1987). *The Polynesians: Prehistory of an island people*. Thames and Hudson: London.

Bellwood P (1991). The Austronesian dispersal and the origin of languages. *Scientific American* **265**(1)**:** 88-93.

Bellwood P (1997). *The Prehistory of the Indo-Malaysian archipelago*, 2nd edn. University of Hawai'i Press: Honolulu.

Berger JO, Pericchi LR (1996). The Intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association* **91**(433)**:** 109-122.

Bertorelle G, Benazzo A, Mona S (2010). ABC as a flexible framework to estimate demography over space and time: Some cons, many pros. *Molecular Ecology* **19**(13)**:** 2609-2625.

Besaggio D, Fuselli S, Srikummool M, Kampuansai J, Castri L, Tyler-Smith C *et al* (2007). Genetic variation in Northern Thailand Hill Tribes: Origins and relationships with social structure and linguistic differences. *BMC Evolutionary Biology* **7**(Suppl 2)**:** S12.

Bhaskararao P. (2006). *Ethnologue: Languages of the world*. Simons GF and Fenning CD (eds). SIL International: Dallas, Texas.

Blomberg SP, Garland T, Ives AR (2003). Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. *Evolution* **57**(4)**:** 717-745.

Blust R (1988). The Austronesian homeland: A linguistic perspective. *Asian Perspectives* **26:** 45-678.

Blust R. (1999). *Selected papers from the eighth international conference on Austronesian linguistics, Vol. 1*, pp 31-94.

Blust R (2000). Why lexicostatistics doesn't work: the 'universal constant'hypothesis and the Austronesian languages. In: Renfrew C, McMahon A and L.Trask (eds) *Time depth in historical linguistics*. McDonal Institute for Archaeological Research: Cambridge. Vol. 2, pp 311-331.

Boivin N (2007). Anthropological, historical, archaeological and genetic perspectives on the origins of caste in South Asia *The Evolution and History of Human Populations in South Asia*. Springer, pp 341-361.

Bolnick DA, Bolnick DI, Smith DG (2006). Asymmetric male and female genetic histories among Native Americans from Eastern North America. *Molecular Biology and Evolution* **23**(11)**:** 2161.

Borgerhoff Mulder M (1991). Human behavioral ecology: studies in foraging and reproduction. In: Krebs J and Davies N (eds) *Behavioral Ecology: An Evolutionary Approach*. Blackwell Scientific: New Jersey, USA, pp 69-98.

Borgerhoff Mulder M, George-Cramer M, Eshleman J, Ortolani A (2001). A Study of East African kinship and marriage using a phylogenetically based comparative method. *American Anthropologist* **103**(4)**:** 1059-1082.

Borgerhoff Mulder M, Nunn C, Towner M (2006). Cultural macroevolution and the transmission of traits. *Evolutionary Anthropology : Issues News and Reviews* **15**(2)**:** 52-64.

Bouckaert R (2010). DensiTree: Making sense of sets of phylogenetic trees. *Bioinformatics* **26**(10)**:** 1372-1373.

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D *et al* (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Computational Biology* **10**(4)**:** e1003537.

Bouckaert R, Lemey P, Dunn M, Greenhill SJ, Alekseyenko AV, Drummond AJ *et al* (2012). Mapping the origins and expansion of the Indo-European language family. *Science* **337**(6097)**:** 957-960.

Bouda K (1956). Dravidisch und uralaltaisch. *Lingua* **5:** 129-144.

Bowern C (2010). Correlates of language change in hunter-gatherer and other 'small' languages. *Language and Linguistics Compass* **4**(8)**:** 665–679.

Boyd R, Borgerhoff Mulder M, Durham WH, Richerson PJ (1997). Are cultural phylogenies possible. *Human by nature: between biology and the social sciences***:** 355-384.

Boyd R, Richerson PJ (1985). *Culture and the evolutionary process*. University of Chicago Press: Chicago.

Boyd R, Richerson PJ (1992). Punishment allows the evolution Of cooperation (Or anything else) In sizable groups. *Ethology and sociobiology* **13**(3)**:** 171-195.

Boyd R, Richerson PJ (1996). Why culture is common, but cultural evolution is rare. *Proceedings-British Academy* **88:** 77-93.

Brown J (1963). A Cross-cultural study of female initiation rites. *American Anthropologist New Series* **65**(4)**:** 837-853.

Brown P (1978). New Guinea: ecology, society and culture. *Annual Review of Anthropology* **7**(1)**:** 263-291.

Bruford MW, Ancrenaz M, Chikhi L, Lackman-Ancrenaz I, Andau M, Ambu L *et al* (2010). Projecting genetic diversity and population viability for the fragmented orang-utan population in the Kinabatangan floodplain, Sabah, Malaysia. *Endangered Species Research* **12**(3)**:** 249-261.

Bruggman K, Osthoff H. (1878). *A Reader in Nineteenth-Century Historical Indo-European Linguistics., Vol. 1*.

Burrow T (1969). Dravidian and the decipherment of the Indus script. *Antiquity* **43**(172)**:** 274-278.

Burrow T, Bhattacharya S (1953). *The Parji Language*: Hertford.

Burrow T, Bhattacharya S (1960). A comparative vocabulary of the Gondi dialects. *Journal of Asiatic Society* **2:** 73-251.

Burrow T, Bhattacharya S (1961). Some notes on the Kui dialect as spoken by the Kuttia Kandhs of north-east Koraput. *Indo-Iranian Journal* **5**(2)**:** 118-135.

Burrow T, Bhattacharya S (1963). Notes on Kuvi with a short vocabulary. *Indo-Iranian Journal* **6**(3)**:** 231-289.

Burrow T, Bhattacharya S (1970). *The Pengo language; Grammar, texts, and vocabulary.* Clarendon Press: Oxford.

Burrow T, Emeneau MB (1984). *A Dravidian Etymological Dictionary [DEDR]*, Second edn. Clarendon Press: Oxford.

Burton ML, Reitz K (1981). The Plow, female contribution to agricultural subsistence and polygyny: A log linear analysis. *Cross-Cultural Research* **16**(3-4)**:** 275-305.

Caballero A (1994). Developments in the prediction of effective population size. *Heredity* **73**(6)**:** 657-679.

Calafell F, Grigorenko EL, Chikanian AA, Kidd KK (2001). Haplotype evolution and linkage disequilibrium: a simulation study. *Human Heredity* **51**(1-2)**:** 85-96.

Caldwell R (1956). *A comparative grammar of the Dravidian or South-Indian family of languages*. University of Madras: Madras.

Campbell LT (1998). Nostratic: a personal assessment. In: Joseph B and Salmons J (eds) *Amsterdam studies in the theory and history of linguistic science series 4*. John Benjamins: Amsterdam, pp 107-152.

Campbell LT (1999). Nostratic and linguistic palaeontology in methodological perspective. In: Renfrew C and Nettle D (eds) *Nostratic: examining a linguistic macrofamily*. The McDonald Institute for Archaeological Research and Cambridge University Press: Cambridge, UK, pp 179-230.

Campbell LT (2002). What drives linguistic diversification and spread? In: Bellwood P and Renfrew C (eds) *Language-Farming Dispersals*. McDonald Institute for Archaeological Research: Cambridge, pp 49-63.

Cann RL (2001). Genetic clues to dispersal in human populations: Retracing the past from the present. *Science* **291**(5509)**:** 1742-1748.

Cann RL, Stoneking M, Wilson AC (1995). Mitochondrial DNA and human evolution. *Nature* **325**(1)**:** 31-36.

Carvajal-Rodríguez A (2010). Simulation of genes and genomes forward in time. *Current Genomics* **11**(1)**:** 58-61.

Cavalli-Sforza L (1975). Cultural and biological evolution: A theoretical inquiry. *Advances in Applied Probability* **7:** 90-99.

Cavalli-Sforza L, Feldman MW (1981). *Cultural transmission and evolution: A quantitative approach*. Princeton University Press: New Jersey.

Cavalli-Sforza LL, Menozzi P, Piazza A (1993). Demic expansions and human evolution. *Science* **259**(5095)**:** 639-646.

Cavalli-Sforza LL, Menozzi P, Piazza A (1994). *The History and geography of human genes*. Princeton university press: New Jersey.

Chaix R, Quintana-Murci L, Hegay T, Hammer MF, Mobasher Z, Austerlitz F *et al* (2007). From social to genetic structures in central Asia. *Current Biology* **17**(1)**:** 43-48.

Chang W, Cathcart C, Hall D, Garrett A (2015). Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* **91**(194-244).

Charlesworth B (2009). Fundamental concepts in genetics: Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics* **10**(3)**:** 195.

Chaubey G, Metspalu M, Kivisild T, Villems R (2007). Peopling of South Asia: Investigating the caste–tribe continuum in India. *Bioessays* **29**(1)**:** 91-100.

Chikhi L, Sousa VC, Luisi P, Goossens B, Beaumont MA (2010). The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. *Genetics* **186**(3)**:** 983-995.

Cleveland WS, Grosse E, William MS (1992). Local regression models In: Chambers JM and Hastie TJ (eds) *Statistical models in S*. CHapman & Hall: New York, pp 309-376.

Codding B, Bird R, Bird D (2011). Provisioning offspring and others: risk–energy trade-offs and gender differences in hunter–gatherer foraging strategies. *Proceedings of the Royal Society of London B: Biological Sciences* **278**(1717)**:** 2502-2509.

Collard M, Shennan SJ, Tehrani JJ (2006). Branching, blending, and the evolution of cultural similarities and differences among human populations. *Evolution and Human Behavior* **27**(3)**:** 169-184.

Corander J, Marttinen P (2006). Bayesian identification of admixture events using multilocus molecular markers. *Molecular Ecology* **15**(10)**:** 2833-2843.

Corander J, Marttinen P, Siren J, Tang J (2008). Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. *BMC Bioinformatics* **9**(1)**:** 539.

Corander J, Waldmann P, Marttinen P, Sillanpaa MJ (2004). BAPS 2: Enhanced possibilities for the analysis of genetic population structure. *Bioinformatics* **20**(15)**:** 2363-2369.

Corander J, Waldmann P, Sillanpaa MJ (2003). Bayesian analysis of genetic differentiation between populations. *Genetics* **163**(1)**:** 367-374.

Cordaux R, Aunger R, Bentley G, Nasidze I, Sirajuddin SM, Stoneking M (2004). Independent origins of Indian caste and tribal paternal lineages. *Current Biology* **14**(3)**:** 231-235.

Cornuet J-M, Ravigné V (2010). Inference on population history and model checking using DNA sequence and microsatellite data with the software DIYABC (v1. 0). *BMC Bioinformatics*(11)**:** 7.

Cornuet J-M, Santos F, Beaumont MA, Robert CP, Marin J-M, Balding DJ *et al* (2008). Inferring population history with DIY ABC: A user-friendly approach to approximate Bayesian computation. *Bioinformatics* **24**(23)**:** 2713-2719.

Cowlishaw G, Mace R (1996). Cross-cultural patterns of marriage and inheritance: A phylogenetic approach. *Ethology and Sociobiology* **17**(2)**:** 87-97.

Croft W (2000). *Explaining language change: An evolutionary approach*. Pearson Education: London.

Cronk L (1991). Human behavioral ecology. *Annual Review of Anthropology* **20:** 25-53.

Csillery K, Blum MG, Gaggiotti OE, Francois O (2010). Approximate Bayesian computation (ABC) in practice. *Trends in Ecology & Evolution* **25**(7)**:** 410-418.

da Silva S, Tehrani J (2016). Comparative phylogenetic analyses uncover the ancient roots of Indo-European folktales. *Royal Society open science* **3**(1)**:** 150645.

Davis R (1984). Muang matrifocality. *The Australian Journal of Anthropology* **14**(4)**:** 263-271.

Dawson KJ, Belkhir K (2001). A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genetical Research* **78**(01)**:** 59-77.

de Filippo C, Barbieri C, Whitten M, Mpoloka S, Gunnarsdóttir E, Bostoen K *et al* (2011). Y-chromosomal variation in Sub-Saharan Africa: Insights into the history of Niger-Congo groups. *Molecular Biology and Evolution* **28**(3)**:** 1255–1269.

Destro-Bisol G, Donati F, Coia V, Boschi I, Verginelli F, Caglia A *et al* (2004). Variation of female and male lineages in sub-Saharan populations: The importance of sociocultural factors. *Molecular Biology and Evolution* **21**(9)**:** 1673-1682.

Di Piazza A, Di Piazza P, Pearthree E (2007). Sailing virtual canoes across Oceania: revisiting island accessibility. *Journal of Archaeological Science* **34**(8)**:** 1219-1225.

Diamond JM (1988). Express Train to Polynesia. *Nature* **336**(6197)**:** 307-308.

Diamond JM (1997). *Guns, germs, and steel : The fates of human societies*. W.W.Norton: New York.

Diamond JM, Bellwood P (2003). Farmers and their languages: The first expansions. *Science* **300**(5619)**:** 597-603.

Divale WT (1974a). *The Causes of matrilocal residence: A cross-ethnohistorical survey*. University Microfilms: Ann Arbor, MI.

Divale WT (1974b). Migration, external warfare, and matrilocal residence. *Cross-Cultural Research* **9**(2)**:** 75-133.

Divale WT, Chamberis F, Gangloff D (1976). War, peace, and marital residence in pre-industrial societies. *Journal of Conflict Resolution* **20**(1)**:** 57-78.

Dobzhansky T (1937). *Genetics and the Origin of Species*. Columbia University Press: New York.

Doolittle WF (1999). Phylogenetic classification and the universal tree. *Science* **284**(5423)**:** 2124-2129.

Driver HE, Massey WC (1957). Comparative studies of North American Indians. *Transactions of the American Philosophical Society* **47:** 165-456.

Driver HE, Ulvestad BE (1956). *An Integration of functional, evolutionary, and historical theory by means of correlations: an approach to describing usage of language variants*. Waverly Press: Baltimore, MD.

Drummond AJ, Rambaut A (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* **7**(1)**:** 214.

Drummond AJ, Rambaut A, Shapiro B, Pybus OG (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution* **22**(5)**:** 1185.

Drummond AJ, Suchard MA, Xie D, Rambaut A (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* **29**(8)**:** 1969-1973.

Dumont d'Urville, MJ (1832). Notice sur les îles du Grand Océan et sur l'origine des peuples qui le habitent. *Société de Géographie Bulletin* **17:** 1-21.

Duncan RP, Williams PA (2002). Ecology: Darwin's naturalization hypothesis challenged. *Nature* **417**(6889)**:** 608-609.

Dunn M (2008). Contact and phylogeny in Island Melanesia. *Lingua* **119**(11)**:** 1664-1678.

Dunn M (2014). Language Phylogenies. In: Bowern C and Evans B (eds) *The Routledge Handbook of Historical Linguistics*. Routledge: New York.

Dunn M, Burenhult N, Krusped N, Tufvesson S, Beckerb N (2011a). Aslian linguistic prehistory: A case study in computational phylogenetics. *Diachronica* **28**(3)**:** 291–323.

Dunn M, Greenhill SJ, Levinson SC, Gray RD (2011b). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* **473**(7345)**:** 79.

Dunn M, Kruspe N, Burenhult N (2013). Time and Place in the Prehistory of the Aslian Languages. *Human Biology* **85**(1-3)**:** 383-400.

Dunn M, Terrill A, Reesink G, Foley R, Levinson S (2005). Structural phylogenetics and the reconstruction of ancient language history. *Science* **309**(5743)**:** 2072.

Dupanloup I, Bertorelle G (2001). Inferring admixture proportions from molecular data: extension to any number of parental populations. *Molecular Biology and Evolution* **18**(4)**:** 672-675.

Durham WH (1991). *Coevolution: Genes, culture, and human diversity*. Stanford University Press: Stanford, California.

Durham WH (1992). Applications of evolutionary culture theory. *Annual Review of Anthropology***:** 331-355.

Eggan F (1966). *The American Indian: Perspective for the study of social change*. Aldine: Chicago.

Ellis FW (1816). Note to the Introduction. In: Campbell A (ed) *A Grammar of the Teloogoo Language, 2nd ed. AD Campbell, ed*. University of Madras: Madras. Vol. 12, pp 1-35.

Elston R, Zeanah D, Codding B (2014). Living outside the box: An updated perspective on diet breadth and sexual division of labor in the Prearchaic Great Basin. *Quaternary International* **352:** 200-211.

Ember C (1974). An evaluation of alternative theories of matrilocal versus patrilocal residence. *Cross-Cultural Research* **9**(2)**:** 135.

Ember M, Ember CR (1971). The Conditions favoring matrilocal versus patrilocal residence. *American Anthropologist New Series* **73**(3)**:** 571-594.

Emeneau MB (1938). Echo-words in Toda. *New Indian Antiquity* **1:** 109-117.

Emeneau MB (1944). *Kota texts*, Vol 2. University of California Press: Berkley.

Emeneau MB (1967). The South Dravidian languages. *Journal of the American Oriental Society* **87**(4)**:** 365-413.

Endicott P, Mait M, Kivisild T (2007). Genetic evidence on modern human dispersals in South Asia: Y-chromosome and mitochondrial DNA perspectives. In: Petraglia M and Allchin B (eds) *The Evolution and History of Human Populations in South Asia: Inter-disciplinary Studies in Archaeology, Biological Anthropology,Linguistics and Genetics*. Springer/Kluwer Academic Publishers: Dodrecht, Netherlands, p 390.

Enright N, Gosden C (1992). Unstable archipelagos—southwest Pacific environment and prehistory since 30,000 BP. In: Dodson J (ed) *The Naive Lands: prehistory and environmental change in Australia and the southwest Pacific*. Longman Cheshire: Melbourne, pp 160-198.

Excoffier L, Foll M (2011). Fastsimcoal: A continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* **27**(9)**:** 1332-1334.

Excoffier L, Laval G, Schneider S (2005). Arlequin (version 3.0): An integrated software package for population genetics data analysis. *Evolutionary bioinformatics online* **1:** 47.

Excoffier L, Lischer HE (2010). Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Molecular ecology resources* **10**(3)**:** 564-567.

Falush D (2003). Traces of Human Migrations in Helicobacter pylori Populations. *Science* **299**(5612)**:** 1582-1585.

Falush D, Stephens M, Pritchard JK (2003). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* **164**(4)**:** 1567-1587.

Feldman M (1996). Gene-culture coevolutionary theory. *Trends in Ecology & Evolution* **11**(11)**:** 453-457.

Felsenstein J (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* **17**(6)**:** 368-376.

Felsenstein J (1985). Phylogenies and the comparative method. *American Naturalist***:** 1-15.

Felsenstein J (1993). PHYLIP: Phylogenetic inference package (Version 3.5). *Cladistics* **5:** 164-166.

Felsenstein J (2004). *Inferring phylogenies*, Vol 2. Sinaeur Associates: Sunderland, MA.

Fenner J (2005). Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *American Journal of Physical Anthropology* **128**(2)**:** 415-423.

Field JS, Lape PV (2010). Paleoclimates and the emergence of fortifications in the tropical Pacific islands. *Journal of Anthropological Archaeology* **29**(1)**:** 113-124.

Firth R (1957). A Note on Descent groups in Polynesia. *Man* **57:** 4-8.

Fisher RA (1930). *The Genetical theory of natural selection*. Oxford University Press: Oxford.

Foley RA, Lee PC (1989). Finite social space, evolutionary pathways, and reconstructing hominid behavior. *Science* **243**(4893)**:** 901-906.

Forster P, Toth A (2003). Toward a phylogenetic chronology of ancient Gaulish, Celtic, and Indo-European. *PNAS* **100**(15)**:** 9079-9084.

Fortunato L (2011). Reconstructing the history of residence strategies in Indo-European speaking societies: neo-, uxori-, and virilocality. *Human Biology* **83**(1)**:** 107-128.

Fortunato L, Holden C, Mace R (2006). From bridewealth to dowry?: A Bayesian estimation of ancestral states of marriage transfers in Indo-European. *Human Nature* **17**(4)**:** 355-376.

Fortunato L, Jordan FM (2010). Your place or mine? A phylogenetic comparative analysis of marital residence in Indo-European and Austronesian societies. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **365**(1559)**:** 3913-3922.

Freckleton RP, Harvey PH, Pagel M (2002). Phylogenetic analysis and comparative data: A test and review of evidence. *American Naturalist* **160**(6)**:** 712-726.

Friedlaender J, Hunley K, Dunn M, Terrill A, Lindström E, Reesink G *et al* (2009). Linguistics more robust than genetics. *Science* **324**(5926)**:** 464-465.

Friedlaender JS, Friedlaender FR, Hodgson JA, Stoltz M, Koki G, Horvat G *et al* (2007). Melanesian mtDNA complexity. *PLoS ONE* **2**(2)**:** e248.

Friedlaender JS, Friedlaender FR, Reed FA, Kidd KK, Kidd JR, Chambers GK *et al* (2008). The Genetic structure of Pacific Islanders. *PLoS Genetics* **4**(1)**:** e19.

Fritz SA, Purvis A (2010). Selectivity in mammalian extinction risk and threat types: A new measure of phylogenetic signal strength in binary traits. *Conservation Biology* **24**(4)**:** 1042-1051.

Fuller DQ (2003). An agricultural perspective on Dravidian historical linguistics: Archaeological crop packages, livestock and Dravidian crop vocabulary. In: Bellwood P and Renfrew C (eds) *Examining the farming/language dispersal hypothesis*: McDonald Institute for Archaeological Research, Cambridge, United Kingdom, pp 191-213.

Fuller DQ (2006). Agricultural origins and frontiers in South Asia: A working synthesis. *Journal of World Prehistory* **20**(1)**:** 1-86.

Fuller DQ (2007). Non-human genetics, agricultural origins and historical linguistics in South Asia. In: Petraglia M and Allchin B (eds) *The Evolution and history of human populations in South Asia: Inter-disciplinary studies in archaeology, biological anthropology, linguistics and genetics*. Springer: Doetinchem, The Netherlands, pp 393-443.

Fuller DQ, Korisettar R, Venkatasubbaiah PC, Jones MK (2004). Early plant domestications in southern India: Some preliminary archaeobotanical results. *Vegetation History and Archaeobotany* **13**(2)**:** 115-129.

Geraghty PA (1983). The History of the Fijian languages. *Oceanic Linguistics Special Publications*(19)**:** i-483.

Geraghty PA. (1996). *Oceanic studies: Proceedings of the First International Conference on Oceanic Linguistics*, pp 83-91.

Gibson MA, Mace R (2007). Polygyny, reproductive success and child health in rural Ethiopia: Why marry a married man? *Journal of Biosocial Science* **39**(2)**:** 287-300.

Gleason HA (1959). Counting and calculating for historical reconstruction. *Anthropological Linguistics***:** 22-32.

Goody J, Watt I (1963). The consequences of literacy. *Comparative studies in society and history* **5**(03)**:** 304-345.

Granovetter MS (1973). The Strength of weak ties. *American Journal of Sociology***:** 1360-1380.

Gray RD, Atkinson QD (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* **426**(6965)**:** 435-439.

Gray RD, Atkinson QD, Greenhill SJ (2011). Language evolution and human history: What a difference a date makes. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **366**(1567)**:** 1090.

Gray RD, Drummond AJ, Greenhill SJ (2009). Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* **323**(5913)**:** 479-483.

Gray RD, Greenhill SJ, Ross RM (2007). The Pleasures and perils of Darwinizing culture (with phylogenies). *Biological Theory* **2**(4)**:** 360-375.

Gray RD, Jordan FM (2000). Language trees support the express-train sequence of Austronesian expansion. *Nature* **405**(6790)**:** 1052-1055.

Green RC (1991a). The Lapita cultural complex: Current evidence and proposed models. *Bulletin of the Indo-Pacific Prehistory Association* **11:** 295-305.

Green RC (1991b). Near and remote Oceania—disestablishing "Melanesia" in culture history. In: Pawley A (ed) *Man and a half: essays in Pacific anthropology and ethnobiology in honour of Ralph Bulmer*. Polynesian Society: Auckland, New Zealand, pp 491-502.

Green RC (2003). The Lapita horizon and traditions—signature for one set of Oceanic migrations. *Pacific archaeology: assessments and prospects***:** 95-120.

Greenhill SJ (2015). Demographic correlates of language diversity. In: Bowern C and Evans B (eds) *The Routledge Handbook of Historical Linguistics*. Routledge Taylor & Francis Group: Abingdon, UK and New York, USA, pp 557-578.

Greenhill SJ, Blust R, Gray RD (2008). The Austronesian basic vocabulary database: From bioinformatics to lexomics. *Evolutionary Bioinformatics* **4:** EBO-S893.

Greenhill SJ, Gray RD (2005). Testing population dispersal hypotheses: Pacific settlement, phylogenetic trees and Austronesian languages. *The evolution of cultural diversity: A phylogenetic approach***:** 31-52.

Greenhill SJ, Gray RD (2009). Austronesian language phylogenies: Myths and misconceptions about Bayesian computational methods. In: Adelaar A and Pawley A (eds) *Austronesian historical linguistics and culture history: a festschrift for Robert Blust*. Pacific Linguistics, pp 375-397.

Guillot EG, Cox MP (2014). SMARTPOP: inferring the impact of social dynamics on genetic diversity through high speed simulations. *BMC Bioinformatics* **15**(1)**:** 175.

Gunnarsdottir ED, Nandineni MR, Li M, Myles S, Gil D, Pakendorf B *et al* (2011). Larger mitochondrial DNA than Y-chromosome differences between matrilocal and patrilocal groups from Sumatra. *Nature Communications* **2:** 228.

Haak W, Brandt G, de Jong HN, Meyer C, Ganslmeier R, Heyd V *et al* (2008). Ancient DNA, strontium isotopes, and osteological analyses shed light on social and kinship organization of the Later Stone Age. *PNAS* **105**(47)**:** 18226-18231.

Hage P, Marck J (2002). Proto-Micronesian kin terms, descent groups, and interisland voyaging. *Oceanic Linguistics* **41**(1)**:** 159-170.

Hage P, Marck J (2003). Matrilineality and the Melanesian origin of Polynesian Y chromosomes. *Current Anthropology* **44**(S5)**:** S121-S127.

Hagelberg E, Clegg JB (1993). Genetic polymorphisms in prehistoric Pacific islanders determined by analysis of ancient bone DNA. *Proceedings of the Royal Society of London B: Biological Sciences* **252**(1334)**:** 163-170.

Hamilton G, Stoneking M, Excoffier L (2005). Molecular analysis reveals tighter social regulation of immigration in patrilocal populations than in matrilocal populations. *PNAS* **102**(21)**:** 7476.

Hammarström H, Forkel R, Haspelmath M, Bank S. (2016). *Online:* http://glottolog.org/*.* Max Planck Institute for the Science of Human History: Jena.

Handley LJL, Perrin N (2007). Advances in our understanding of mammalian sex-biased dispersal. *Molecular Ecology* **16**(8)**:** 1559-1578.

Harris M (1980). *Cultural materialism*. Random House: New York.

Harris M (1985). *Culture, people, nature: An Introduction to general anthropology*. Harper & Row, New York.

Harvey PH, Pagel M (1991). *The comparative method in evolutionary biology*, Vol 239. Oxford University Press: Oxford.

Heath DB (1958). Sexual division of labor and cross-cultural research. *Social Forces* **37:** 77-79.

Heggarty P, Beresford-Jones D (2010). Agriculture and language dispersals. *Current Anthropology* **51**(2)**:** 163-191.

Hennig W (1950). Grundzuge einer Theorie der phylogenetischen Systematik. *Annual Review of Entomology* **10:** 97-116.

Hey J (2004). Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* **167**(2)**:** 747-760.

Heyer E, Chaix R, Pavard S, Austerlitz F (2011). Sex-specific demographic behaviours that shape human genomic variation. *Molecular Ecology* **21**(3)**:** 597-612.

Heyerdahl T (1950). *The Kon-Tiki expedition*. Allen and Unwin: London.

Hiatt B (1970). Woman the gatherer. *Woman's role in Aboriginal society***:** 2-8.

Hoban S (2014). An overview of the utility of population simulation software in molecular ecology. *Molecular Ecology* **23**(10)**:** 2383-2401.

Hoban S, Bertorelle G, Gaggiotti OE (2012). Computer simulations: Tools for population and evolutionary genetics. *Nature Reviews Genetics* **13**(2)**:** 110-122.

Hoenigswald HM (1965). *Language change and linguistic reconstruction*. University of Chicago Press: Chicago.

Holden CJ (2002). Bantu language trees reflect the spread of farming across sub-Saharan Africa: A maximum-parsimony analysis. *Proceedings of the Royal Society of London B: Biological Sciences* **269**(1493)**:** 793-799.

Holden CJ, Gray RD (2006). Rapid radiation, borrowing, and dialect continua in the Bantu languages. In: Forster P and Renfrew C (eds) *Phylogenetic Methods and the Prehistory of Languages*. McDonald Institute for Archaeological Research: Cambridge, pp 19-31.

Holden CJ, Mace R (2003). Spread of cattle led to the loss of matrilineal descent in Africa: a coevolutionary analysis. *PNAS* **270**(1532)**:** 2425-2433.

Holland BR, Huber KT, Moulton V, Lockhart PJ (2004). Using consensus networks to visualize contradictory evidence for species phylogeny. *Molecular Biology and Evolution* **21**(7)**:** 1459-1461.

Hudson RR (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology* **23**(2)**:** 183-201.

Huelsenbeck JP, Rannala B, Masly JP (2000). Accommodating phylogenetic uncertainty in evolutionary studies. *Science* **288**(5475)**:** 2349-2350.

Hunley K, Dunn M, Lindström E, Reesink G, Terrill A, Healy ME *et al* (2008). Genetic and linguistic coevolution in northern Island Melanesia. *PLoS Genetics* **4**(10)**:** e1000239-1000383.

Hurles ME, Nicholson J, Bosch E, Renfrew C, Sykes BC, Jobling MA (2002). Y chromosomal evidence for the origins of Oceanic-speaking peoples. *Genetics* **160**(1)**:** 289-303.

Ingold T (2000). The poverty of selectionism. *Anthropology Today* **16**(3)**:** 1-2.

Ingold T (2007). The trouble with 'evolutionary biology'. *Anthropology Today* **23**(2)**:** 13-17.

Irwin G (1994). *The prehistoric exploration and colonisation of the Pacific*. Cambridge University Press: Cambridge.

Johnson MTJ, Stinchcombe JR (2007). An emerging synthesis between community ecology and evolutionary biology. *Trends in Ecology & Evolution* **22**(5)**:** 250-257.

Jones D (2003). Kinship and deep history: Exploring connections. *American anthropologist* **105**(3)**:** 501-514.

Jones JH (2009). The Force of selection on the human life cycle. *Evolution and Human Behavior* **30**(5)**:** 305-314.

Jordan FM (2007). A Comparative phylogenetic approach to Austronesian cultural evolution. PhD thesis, University College of London, London.

Jordan FM, Gray RD, Greenhill SJ, Mace R (2009). Matrilocal residence is ancestral in Austronesian societies. *Proceedings of the Royal Society of London B: Biological Sciences* **276**(1664)**:** 1957-1964.

Jordan P, Shennan S (2009). Diversity in hunter–gatherer technological traditions: Mapping trajectories of cultural 'descent with modification' in northeast California. *Journal of Anthropological Archaeology* **28**(3)**:** 342-365.

Joyce P, Marjoram P (2008). Approximately sufficient statistics and Bayesian computation. *Statistical Applications in Genetics and Molecular Biology* **7**(1).

Kass R, Raftery A (1995). Bayes Factors. *Journal of the American Statistical Association* **90**(430)**:** 773-795.

Kassian A, Starostin G, Dybo A, Chernov V (2010). The Swadesh wordlist. An attempt at semantic specification. *Journal of Language Relationship* **4**(16)**:** 46-89.

Kayser M (2010). The Human genetic history of Oceania: Near and Remote views of dispersal. *Current Biology* **20**(4)**:** R194-R201.

Kayser M, Brauer S, Cordaux R, Casto A, Lao O, Zhivotovsky LA *et al* (2006). Melanesian and asian origins of Polynesians: mtDNA and Y chromosome gradients across the Pacific. *Molecular Biology and Evolution* **23**(11)**:** 2234-2244.

Kayser M, Brauer S, Weiss G, Underhill P, Roewer L, Schiefenhovel W *et al* (2000). Melanesian origin of Polynesian Y chromosomes. *Current Biology* **10**(20)**:** 1237-1246.

Kayser M, Lao O, Saar K, Brauer S, Wang X, Nürnberg P *et al* (2008). Genome-wide analysis indicates more Asian than Melanesian ancestry of Polynesians. *American Journal of Human Genetics* **82**(1)**:** 194-198.

Kelchner SA, Thomas MA (2007). Model use in phylogenetics: nine key questions. *Trends in Ecology & Evolution* **22**(2)**:** 87-94.

Kendal J, Tehrani JJ, Odling-Smee J (2011). Human niche construction in interdisciplinary focus. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **366**(1566)**:** 785-792.

Kingman JFC (1981). The coalescent. *Stochaistic Processes and their Applications* **13:** 235-248.

Kirch PV (1994). *The Wet and the dry: Irrigation and agricultural intensification in Polynesia*. University of Chicago Press: Chicago.

Kirch PV (2002). *On the road of the winds: An archaeological history of the Pacific Islands before European contact*. Univ of California Press: Berkley.

Kirch PV (2010). Peopling of the Pacific: A holistic anthropological perspective. *Annual Review of Anthropology* **39:** 131-148.

Kirch PV, Green RC (2001). *Hawaiki, ancestral Polynesia: an essay in historical anthropology*. Cambridge University press: Cambridge.

Kittles R, Bergen A, Urbanek M (1999). Autosomal, mitochondrial, and Y chromosome DNA variation in Finland: Evidence for a male-specific bottleneck. *American Journal of Physical Anthropology* **108**(4)**:** 381-399.

Kivisild T, Bamshad MJ, Kaldma K, Metspalu M, Metspalu E, Reidla M *et al* (1999). Deep common ancestry of Indian and western-Eurasian mitochondrial DNA lineages. *Current Biology* **9**(22)**:** 1331-1334.

Kivisild T, Papiha SS, Rootsi S, Parik J, Kaldma K, Reidla M *et al* (2000). An Indian ancestry: A key for understanding human diversity in Europe and beyond. *Archaeogenetics: DNA and the population prehistory of Europe***:** 267-279.

Kivisild T, Rootsi S, Metspalu M, Metspalu E, Parik J, Kaldma K *et al* (2003). The Genetics of language and farming spread in India. *Examining the farming/language dispersal hypothesis McDonald Institute Monographs Series, McDonald Institute for Archaeological Research, Cambridge, UK***:** 215-222.

Kobben A, Altschuler M, Bailey W, Carstens P, Driver H, Jorgensen J *et al* (1967). Why exceptions? The logic of cross-cultural analysis [and comments and reply]. *Current Anthropology* **8**(1/2)**:** 3-34.

Koerner K (1983). Preface. In: Koerner K (ed) *August Schleicher: Die Sprachen Europaas in systematischer Übersicht*. John Benjamins Publishing Co: Amsterdam.

Korotayev A (2003a). Division of labor by gender and postmarital residence in cross-cultural perspective: A reconsideration. *Cross-Cultural Research* **37**(4)**:** 335-372.

Korotayev A (2003b). Form of marriage, sexual division of labor, and postmarital residence in cross-cultural perspective: A reconsideration. *Journal of Anthropological Research* **59**(1)**:** 69-89.

Krishnamurti B (1961). *Telugu verbal bases: A comparative and descriptive study (UCPL 24)*. Berkeley and Los Angeles: University of California Press.

Krishnamurti B (1969). Comparative Dravidian studies. *Current trends in Linguistics* **5:** 309-333.

Krishnamurti B (1978). Areal and lexical diffusion of sound change: Evidence from Dravidian. *Language*(1)**:** 1-20.

Krishnamurti B (1985). An overview of comparative Dravidian studies since current trends 1969. *Oceanic Linguistics Special Publication* **20:** 212-231.

Krishnamurti B (1997). Proto-Dravidian laryngeal* H revisited. *PILC Journal of Dravidic Studies* **7:** 2.145-165.

Krishnamurti B (1998). Regularity of sound change through lexical diffusion: A study of s> h> in Gondi dialects. *Language Variation and Change* **10**(02)**:** 193-220.

Krishnamurti B (2003). *The Dravidian Languages*. Cambridge University Press: New York.

Krishnamurti B, Emeneau MB (2001). *Comparative Dravidian linguistics: Current perspectives*. Oxford University Press: Oxford.

Kumar S, Padmanabham P, Ravuri RR, Uttaravalli K, Koneru P, Mukherjee PA *et al* (2008). The earliest settlers' antiquity and evolutionary history of Indian populations: Evidence from M2 mtDNA lineage. *BMC Evolutionary Biology* **8**(1)**:** 230.

Kumar V, Langstieh BT, Madhavi KV, Naidu VM, Singh HP, Biswas S *et al* (2006). Global patterns in human mitochondrial DNA and Y-chromosome variation caused by spatial instability of the local cultural processes. *PLoS Genetics* **2**(4)**:** e53.

Kushnick G, Gray RD, Jordan FM (2014). The sequential evolution of land tenure norms. *Evolution and Human Behavior* **35**(4)**:** 309-318.

Labov W (2007). Transmission and diffusion. *Linguistic Society of America* **83**(2)**:** 344.

Lahr MM, Foley R (1994). Multiple dispersals and modern human origins. *Evolutionary Anthropology: Issues, News, and Reviews* **3**(2)**:** 48-60.

Laitin DD, Moortgat J, Robinson AL (2012). Geographic axes and the persistence of cultural diversity. *PNAS* **109**(26)**:** 10263-10268.

Laland KN (1993). The Mathematical modelling of human culture and its implications for psychology and the human sciences. *British Journal of Psychology* **84**(2)**:** 145-169.

Laland KN, Odling-Smee J, Feldman M (2000). Niche construction, biological evolution, and cultural change. *Behavioral and Brain Sciences* **23**(1)**:** 131-146.

Laland KN, Kumm J, Feldman MW (1995). Gene-culture coevolutionary theory: A test case. *Current Anthropology* **36**(1)**:** 131-156.

Laland KN, Odling-Smee J, Myles S (2010). How culture shaped the human genome: Bringing genetics and the human sciences together. *Nature Reviews Genetics* **11**(2)**:** 137.

Laland KN, Sterelny K, Odling-Smee J, Hoppitt W, Uller T (2011). Cause and effect in biology revisited: Is Mayr's proximate-ultimate dichotomy still useful? *Science* **334**(6062)**:** 1512-1516.

Lane RB (1961). A Reconsideration of Malayo-Polynesian social organization. *American Anthropologist* **63**(4)**:** 711–720.

Lansing JS, Watkins JC, Hallmark B, Cox MP, Karafet TM, Sudoyo H *et al* (2008). Male dominance rarely skews the frequency distribution of Y chromosome haplotypes in human populations. *PNAS* **105**(33)**:** 11645-11650.

Leavesley MG, Bird MI, Fifield LK, Hausladen PA, Santos GM, Di Tada ML (2002). Buang Merabak: Early evidence for human occupation in the Bismarck Archipelago, Papua New Guinea. *Australian Archaeology* **54**(1)**:** 55-57.

Lebar FM (1972). *Ethnic groups of insular Southeast Asia*. Human Relations Area Files Press: New Haven.

Lee S, Hasegawa T (2013). Evolution of the Ainu language in space and time. *PLoS ONE* **8**(4)**:** e62243.

Lévi-Strauss C (1969). *Elementary structures of kinship*, Vol 340. Beacon Press: Boston, MA.

Levinson D (1993). *Encyclopedia of World Cultures*. GK Hall & Co: Boston, MA.

Levinson SC, Gray RD (2012). Tools from evolutionary biology shed new light on the diversification of languages. *Trends in cognitive sciences* **16**(3)**:** 167-173.

Lewis MP, Simons GF, Fenning CD (eds) (2009). *Ethnologue: Languages of the world*. SIL International: Dallas, TX, USA.

Li H, Durbin R (2011). Inference of human population history from individual whole-genome sequences. *Nature* **475**(7357)**:** 493-496.

Lindenfors P, Revell LJ, Nunn CL (2010). Sexual dimorphism in primate aerobic capacity: a phylogenetic test. *Journal of Evolutionary Biology* **23**(6)**:** 1183-1194.

Lipo CP (2006). *Mapping our ancestors: Phylogenetic approaches in anthropology and prehistory*. Transaction Publishers: New Brunswick, USA.

Lipson M, Loh P-R, Patterson N, Moorjani P, Ko Y-C, Stoneking M *et al* (2014). Reconstructing Austronesian population history in island Southeast Asia. *Nature Communications* **5:** 4689.

Lipson M, Skoglund P, Spriggs M, Valentin F, Bedford S, Shing R *et al* (2018). Population Turnover in Remote Oceania Shortly after Initial Settlement. *Current Biology*.

Lopes J, Beaumont M (2010). ABC: A useful Bayesian tool for the analysis of population data. *Infection, Genetics and Evolution* **10**(6)**:** 825-832.

Losos JB (2008). Phylogenetic niche conservatism, phylogenetic signal and the relationship between phylogenetic relatedness and ecological similarity among species. *Ecology letters* **11**(10)**:** 995-1003.

Lum JK, Cann RL (1998). mtDNA and language support a common origin of Micronesians and Polynesians in Island Southeast Asia. *American Journal of Physical Anthropology* **105**(2)**:** 109-119.

Lum JK, Cann RL, Martinson JJ, Jorde LB (1998). Mitochondrial and nuclear genetic relationships among Pacific Island and Asian populations. *American Journal of Human Genetics* **63**(2)**:** 613-624.

Lum JK, Jorde LB, Schiefenhovel W (2002). Affinities among Melanesians, Micronesians, and Polynesians: A neutral, biparental genetic perspective. *Human Biology* **74**(3)**:** 413-430.

Lum JK, Rickards O, Ching C, Cann RL (1994). Polynesian mitochondrial DNAs reveal three deep maternal lineage clusters. *Human Biology* **66**(4)**:** 567.

Mace R (2000). Evolutionary ecology of human life history. *Animal behaviour* **59**(1)**:** 1-10.

Mace R, Holden CJ (2005). A Phylogenetic approach to cultural evolution. *Trends in Ecology & Evolution* **20**(3)**:** 116-121.

Mace R, Jordan FM (2011). Macro-evolutionary studies of cultural diversity: A review of empirical studies of cultural transmission and cultural adaptation. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **366**(1563)**:** 402-411.

Mace R, Jordan FM, Holden CJ (2003). Testing evolutionary hypotheses about human biological adaptation using cross-cultural comparison. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology* **136**(1)**:** 85-94.

Mace R, Pagel M (1994). The Comparative Method in Anthropology. *Current Anthropology* **35**(5)**:** 549-564.

Madella M, Fuller DQ (2006). Palaeoecology and the Harappan civilisation of South Asia: A reconsideration. *Quaternary Science Reviews* **25**(11)**:** 1283-1301.

Majumder PP (1998). People of India; biological diversity and affinities. *The Indian human heritage***:** 45-59.

Marck J (2008). Proto Oceanic society was matrilineal. *The Journal of the Polynesian Society* **117**(4)**:** 345.

Marck J, Bostoen K (2010). Proto Oceanic (Austronesian) and Proto East Bantu (Niger-Congo) kin terms ca. 1000 BC. In: Jones D and Milicic B (eds) *Kinship, language, and prehistory: Per Hage and the renaissance in kinship studies*. University of Utah Press: Salt Lake City, pp 83-91.

Marlowe FW (2004). Marital residence among foragers. *Current Anthropology* **45**(2)**:** 277-284.

Marlowe FW (2007). Hunting and gathering: The human sexual division of foraging labor. *Cross-Cultural Research* **41**(2)**:** 170-195.

Matisoo-Smith E (1994). The Human colonisation of Polynesia. A novel approach: genetic analyses of the Polynesian rat (Rattus exulans). *The Journal of the Polynesian Society* **103**(1)**:** 75-87.

Matisoo-Smith E, Robins JH (2004). Origins and dispersals of Pacific peoples: Evidence from mtDNA phylogenies of the Pacific rat. *PNAS* **101**(24)**:** 9167-9172.

Mayr E (1982). *The growth of biological thought: Diversity, evolution, and inheritance*. Harvard University Press: Boston, MA.

McAlpin D (1974). Toward Proto-Elamo-Dravidian. *Language* **50**(1)**:** 89-101.

McCartney AP, Veltre DW (1999). Aleutian Island prehistory: Living in insular extremes. *World Archaeology* **30**(3)**:** 503-515.

Melton T, Peterson R, Redd AJ, Saha N, Sofro AS, Martinson J *et al* (1995). Polynesian genetic affinities with Southeast Asian populations as identified by mtDNA analysis. *American Journal of Human Genetics* **57**(2)**:** 403.

Menges KH (1964). Altaijisch und Dravidisch. *Orbis* **13:** 66-103.

Menges KH (1969). The Dravido-Altaic Relationship. *Journal of Tamil Studies* **1:** 35-39.

Mesa NR, Mondragón MC, Soto ID, Parra MV, Duque C, Ortíz-Barrientos D *et al* (2000). Autosomal, mtDNA, and Y-chromosome diversity in Amerinds: Pre-and post-Columbian patterns of gene flow in South America. *American Journal of Human Genetics* **67**(5)**:** 1277-1286.

Mesoudi A (2007). Biological and cultural evolution: Similar but different. *Biological Theory* **2**(2)**:** 119-123.

Mesoudi A, Whiten A, Laland K (2004). Perspective: Is human cultural evolution Darwinian? Evidence reviewed from the perspective of The Origin of Species. *Evolution* **58**(1)**:** 1-11.

Mesoudi A, Whiten A, Laland K (2006). Towards a unified science of cultural evolution. *Behavioral and Brain Sciences* **29**(4)**:** 329-347; discussion 347-383.

Metspalu M, Kivisild T, Metspalu E, Parik J, Hudjashov G, Kaldma K *et al* (2004). Most of the extant mtDNA boundaries in south and southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. *BMC Genetics* **5**(1)**:** 26.

Michener CD, Sokal RR (1957). A Quantitative approach to a problem in classification. *Evolution* **11**(2)**:** 130-162.

Mirabal S, Herrera KJ, Gayden T, Regueiro M, Underhill PA, Garcia-Bertrand RL *et al* (2012). Increased Y-chromosome resolution of haplogroup O suggests genetic ties between the Ami aborigines of Taiwan and the Polynesian Islands of Samoa and Tonga. *Gene* **492**(2)**:** 339-348.

Moodley Y, Linz B, Yamaoka Y, Windsor HM, Breurec S, Wu J-Y *et al* (2009). The Peopling of the Pacific from a bacterial perspective. *Science* **323**(5913)**:** 527-530.

Murdock GP (1940). The Cross-cultural survey. *American Sociological Review* **5**(3)**:** 361-370.

Murdock GP (1949a). Social Organization. *American Anthropologist* **51**(2)**:** 298-300.

Murdock GP (1949b). *Social structure*. Macmillan: New York.

Murdock GP (1967). Ethnographic Atlas: A summary. *Ethnology* **6**(2)**:** 109-236.

Murdock GP, White DR (1969). Standard cross-cultural sample. *Ethnology* **8**(4)**:** 329-369.

Naroll R (1961). Two solutions to Galton's problem. *Philosophy of Science* **28**(1)**:** 15-39.

Neiman FD (1995). Stylistic variation in evolutionary perspective: Inferences from decorative diversity and interassemblage distance in Illinois Woodland ceramic assemblages. *American Antiquity* **60**(1)**:** 7-36.

Nicholls G, Gray R (2008). Dated ancestral trees from binary trait data and their application to the diversification of languages. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(3)**:** 545-566.

Nichols J, Warnow T (2008). Tutorial on computational linguistic phylogeny. *Language and Linguistics Compass* **2**(5)**:** 760-820.

Nunn C, Mulder M, Langley S (2006). Comparative methods for studying cultural trait evolution: A simulation study. *Cross-Cultural Research* **40**(2)**:** 177-209.

O'Connell JF, Allen J (2004). Dating the colonization of Sahul (Pleistocene Australia–New Guinea): A review of recent research. *Journal of Archaeological Science* **31**(6)**:** 835-853.

O'Meara BC, Ané C, Sanderson MJ, Wainwright PC (2006). Testing for different rates of continuous trait evolution using likelihood. *Evolution* **60**(5)**:** 922-933.

Odling-Smee FJ, Laland KN, Feldman MW (2003). *Niche construction: the Neglected process in evolution*. Princeton University Press: New Jersey.

Ōno S (1980). *Sound correspondences between Tamil and Japanese*. Gakushuin University: Tokyo.

Oota H, Settheetham-Ishida W, Tiwawech D, Ishida T, Stoneking M (2001). Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. *Nature Genetics* **29**(1)**:** 20-21.

Opie C, Shultz S, Atkinson Q, Currie T, Mace R (2014). Phylogenetic reconstruction of Bantu kinship challenges Main Sequence Theory of human social evolution. *PNAS* **111**(49)**:** 17414-17419.

Oppenheimer SJ, Richards M (2001a). Fast trains, slow boats, and the ancestry of the Polynesian islanders. *Science Progress* **84**(3)**:** 157-182.

Oppenheimer SJ, Richards M (2001b). Polynesian origins - Slow boat to melanesia? *Nature* **410**(6825)**:** 166-167.

Orme CDL, Freckleton RP, Thomas GH, Petzoldt T, Fritz SA, Isaac NJB *et al* (2012). Caper: Comparative analyses of phylogenetics and evolution in R. *Methods in Ecology and Evolution* **3:** 145– 151.

Osmond MW (1965). Toward monogamy: A cross-cultural study of correlates of type of marriage. *Social Forces* **44**(1)**:** 8-16.

Otterbein KF, Otterbein CS (1965). An eye for an eye, a tooth for a tooth: A cross-cultural study of feuding. *American Anthropologist New Series* **67**(6)**:** 1470-1482.

Pagel M (1994). Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society of London B: Biological Sciences* **255**(1342)**:** 37-45.

Pagel M (1997). Inferring evolutionary processes from phylogenies. *Zoologica Scripta* **26**(4)**:** 331-348.

Pagel M (1999a). Inferring the historical patterns of biological evolution. *Nature* **401**(6756)**:** 877-884.

Pagel M (1999b). The Maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Systematic Biology* **48**(3)**:** 612-622.

Pagel M (2002). Modelling the evolution of continuously varying characters on phylogenetic trees. In: Macleod N and Forey PL (eds) *Morphology, shape and phylogeny*. Taylor & Francis: New York, pp 269-286.

Pagel M (2009). Human language as a culturally transmitted replicator. *Nature Reviews Genetics* **10**(6)**:** 405-415.

Pagel M, Harvey P (1988). Recent developments in the analysis of comparative data. *The Quarterly review of biology* **63**(4)**:** 413-440.

Pagel M, Meade A (2004). A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Systematic Biology* **53**(4)**:** 571-581.

Pagel M, Meade A (2006). Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *American Naturalist* **167**(6)**:** 808-825.

Pagel M, Meade A, Barker D (2004). Bayesian estimation of ancestral character states on phylogenies. *Systematic Biology* **53**(5)**:** 673-684.

Pakendorf B, Bostoen K, de Filippo C (2011). Molecular perspectives on the Bantu expansion: a synthesis. *Language Dynamics and Change* **1**(1)**:** 50-88.

Panter-Brick C (2002). Sexual division of labor: Energetic and evolutionary scenarios. *American Journal of Human Biology* **14**(5)**:** 627-640.

Pawley A, Ross M (1995). The prehistory of Oceanic languages: A current view. *The Austronesians***:** 39-74.

Peng B, Kimmel M (2005). simuPOP: a forward-time population genetics simulation environment. *Bioinformatics* **21**(18)**:** 3686-3687.

Petchey F, Spriggs M, Bedford S, Valentin F, Buckley H (2014). Radiocarbon dating of burials from the Teouma Lapita cemetery, Efate, Vanuatu. *Journal of Archaeological Science* **50:** 227-242.

Pigliucci M (2007). Do we need an extended evolutionary synthesis? *Evolution* **61**(12)**:** 2743-2749.

Posada D, Buckley TR (2004). Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology* **53**(5)**:** 793-808.

Posada D, Crandall KA (2001). Selecting the best-fit model of nucleotide substitution. *Systematic Biology* **50**(4)**:** 580-601.

Posth C, Nägele K, Colleran H, Valentin F, Bedford S, Kami KW *et al* (2018). Language continuity despite population replacement in Remote Oceania. *Nature ecology & evolution***:** 1.

Pritchard JK, Seielstad M, Perez-Lezaun A (1999). Population growth of human Y chromosomes: A study of Y chromosome microsatellites. *Molecular Biology and Evolution* **16**(2)**:** 1791-1798.

Pritchard JK, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**(2)**:** 945-959.

Quinn GP, Keough MJ (2002). *Experimental Design and Data Analysis for Biologists*. Cambridge University Press: New York.

Quintana-Murci L, Krausz C, Zerjal T, Sayar SH, Hammer MF, Mehdi SQ *et al* (2001). Y-chromosome lineages trace diffusion of people and languages in southwestern Asia. *American Journal of Human Genetics* **68**(2)**:** 537-542.

Quintana-Murci L, Semino O, Bandelt H-J, Passarino G, McElreavey K, Santachiara-Benerecetti AS (1999). Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. *Nature Genetics* **23**(4)**:** 437-441.

Raff JA, Bolnick DA, Tackney J, O'Rourke DH (2011). Ancient DNA perspectives on American colonization and population history. *American Journal of Physical Anthropology* **146**(4)**:** 503-514.

Rama T, Kolachina S, B LB (2009). Quantitative methods for phylogenetic inference in historical linguistics: an Experimental case study of South Central Dravidian. *Indian Linguistics* **70:** 265-282.

Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *PNAS* **102**(44)**:** 15942-15947.

Ramakrishnan U, Hadly E, Mountain J (2005). Detecting past population bottlenecks using temporal genetic data. *Molecular Ecology* **14**(10)**:** 2915-2922.

Rao RPN, Yadav N, Vahia MN, Joglekar H, Adhikari R, Mahadevan I (2009). Entropic evidence for linguistic structure in the Indus script. *Science*(324)**:** 1165.

Ray N, Currat M, Excoffier L (2003). Intra-Deme Molecular Diversity in Spatially Expanding Populations. *Molecular Biology and Evolution* **20**(1)**:** 76-86.

Redd AJ, Takezaki N, Sherry ST, McGarvey ST, Sofro AS, Stoneking M (1995). Evolutionary history of the COII/tRNALys intergenic 9 base pair deletion in human mitochondrial DNAs from the Pacific. *Molecular Biology and Evolution* **12**(4)**:** 604-615.

Reesink G, Singer R, Dunn M (2009). Explaining the linguistic diversity of Sahul using population models. *PLoS Biology* **7**(11)**:** e1000241.

Reich D, Thangaraj K, Patterson N, Price AL, Singh L (2009). Reconstructing Indian population history. *Nature* **461**(7263)**:** 489-494.

Renfrew C (1992). Archaeology, genetics and linguistic diversity. *Man***:** 445-478.

Revell L, Harmon L, Collar D (2008). Phylogenetic signal, evolutionary process, and rate. *Systematic Biology* **57**(4)**:** 591.

Rexova K, Bastin Y, Frynta D (2006). Cladistic analysis of Bantu languages: A new tree based on combined lexical and grammatical data. *Naturwissenschaften* **93**(4)**:** 189-194.

Richards A (1950). *Some types of family structure amongst the central Bantu*. Oxford University Press: London.

Richards M, Oppenheimer S, Sykes B (1998). mtDNA suggests Polynesian origins in eastern Indonesia. *American Journal of Human Genetics* **63**(4)**:** 1234-1236.

Ridley M (1983). *The explanation of organic diversity: The comparative method and adaptations for mating*. Oxford University Press: New York.

Riede F (2008). Maglemosian memes: Technological ontology, craft traditions and the evolution of Northern European barbed points. In: O'Brien M (ed) *Cultural Transmission and Archaeology: Issues and Case Studies* Society for American Archaeology: Washington, D.C.

Ringe D, Warnow T, Taylor A (2002). Indo-European and Computational Cladistics. *Transactions of the philological society* **100**(1)**:** 59-129.

Rivera MC, Lake JA (2004). The Ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature* **431**(7005)**:** 152-155.

Roberts M (1991). Origin, dispersal routes, and geographic distribution of Rattus exulans, with special reference to New Zealand. *Pacific Science* **45**(2)**:** 123-130.

Rogers A, Harpending H (1992). Population growth makes waves in the distribution of pairwise genetic differences. *Molecular Biology and Evolution* **9**(552-569).

Rolett BV (2008). Avoiding collapse: Pre-European sustainability on Pacific islands. *Quaternary international* **184**(1)**:** 4-10.

Ronquist F (2004). Bayesian inference of character evolution. *Trends in Ecology & Evolution* **19**(9)**:** 475-481.

Rosenberg NA, Mahajan S, Gonzalez-Quevedo C, Blum MGB, Nino-Rosales L, Ninis V *et al* (2006). Low levels of genetic divergence across geographically and linguistically diverse populations from India. *PLoS Genetics* **2**(12)**:** e215.

Rosenberg NA, Nordborg M (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics* **3**(5)**:** 380-390.

Ross MD (1988). *Proto Oceanic and the Austronesian languages of western Melanesia*. Australian National University: Canberra.

Ross MD (1996). Contact-induced change and the comparative. In: Durie M and Ross M (eds) *The comparative method reviewed: Regularity and irregularity in language change*. Oxford University Press: New York, pp 180-217.

Sahoo S, Singh A, Himabindu G, Banerjee J, Sitalaximi T, Gaikwad S *et al* (2006). A prehistory of Indian Y chromosomes: evaluating demic diffusion scenarios. *PNAS* **103**(4)**:** 843-848.

Salas A, Richards M, De la Fe T, Lareu M, Sobrino B, Sánchez-Diz P *et al* (2002). The Making of the African mtDNA landscape. *American Journal of Human Genetics* **71**(5)**:** 1082-1111.

Sanday PR (1974). Female status in the public domain. *Woman, culture, and society* **133:** 189.

Sanderson MJ, Donoghue MJ (1989). Patterns of variation in levels of homoplasy. *Evolution* **43**(8)**:** 1781-1795.

Sapir E (1916). *Time perspective in aboriginal American culture: a study in method*. Government Printing Bureau: Ottowa, Canada.

Saraswat KS (2004). Plant economy of early farming communities. *Early farming communities of the Kaimur (excavations at Senuwar) Jaipur: Publication Scheme***:** 416-435.

Saslis-Lagoudakis C, Hawkins J, Greenhill S, Pendry C, Watson M, Tuladhar-Douglas W *et al* (2014). The evolution of traditional knowledge: Environment shapes medicinal plant use in Nepal. *Proceedings of the Royal Society of London B: Biological Sciences* **281**(1780)**:** 20132768.

Scally A, Durbin R (2012). Revising the human mutation rate: implications for understanding human evolution. *Nature Reviews Genetics* **13**(10)**:** 745-753.

Scheffler H (1964). Descent concepts and descent groups: The Maori case. *The Journal of the Polynesian Society* **73**(2)**:** 126-133.

Schlegel A, Barry III, H (1986). The cultural consequences of female contribution to subsistence. *American anthropologist* **88**(1)**:** 142-150.

Schleicher A (1873). *Die Darwinsche theorie und die sprachwissenschaft: Offenes sendschreiben an herrn Ernst Häckel*, Vol 2. Böhlau.

Schmidt HA, von Haeseler A (2009). Phylogenetic inference using maximum likelihood methods. In: Salemi M, Lemey P and Vandamme A-M (eds) *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*. Cambridge University Press: Cambridge. Vol. 2, pp 181-209.

Scozzari R, Cruciani F, Santolamazza P, Malaspina P, Torroni A, Sellitto D *et al* (1999). Combined use of biallelic and microsatellite Y-chromosome polymorphisms to infer affinities among African populations. *American Journal of Human Genetics* **65**(3)**:** 829-846.

Ségurel L, Martínez-Cruz B, Quintana-Murci L, Balaresque P, Georges M, Hegay T *et al* (2008). Sex-specific genetic structure and social organization in Central Asia: insights from a multi-locus study. *PLoS Genetics* **4**(9)**:** e1000200.

Seielstad MT, Minch E, Cavalli-Sforza LL (1998). Genetic evidence for a higher female migration rate in humans. *Nature Genetics* **20**(3)**:** 278-280.

Sellen D, Mace R (1997). Fertility and mode of subsistence: A phylogenetic analysis. *Current Anthropology* **38**(5)**:** 878-889.

Sengupta S, Zhivotovsky LA, King R, Mehdi SQ, Edmonds CA, Chow C-ET *et al* (2006). Polarity and temporality of high-resolution y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. *American Journal of Human Genetics* **78**(2)**:** 202-221.

Service ER (1962). *Primitive social origanization: An Evolutionary perspective*, 2nd edn. Random House: New York.

Shapiro B, Drummond AJ, Rambaut A, Wilson MC, Matheus PE, Sher AV *et al* (2004). Rise and fall of the Beringian steppe bison. *Science* **306**(5701)**:** 1561-1565.

Sharkey M, Leathers J (2001). Majority does not rule: The trouble with majority-rule consensus trees. *Cladistics* **17**(3)**:** 282-284.

Sjoberg AF (1992). *The Impact of Dravidian on Indo-Aryan: an overview*. Mouton de Gruyter: Berlin and New York.

Skoglund P, Posth C, Sirak K, Spriggs M, Valentin F, Bedford S *et al* (2016). Genomic insights into the peopling of the Southwest Pacific. *Nature* **538**(7626)**:** 510-513.

Slatkin M (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**(1)**:** 457-462.

Sneath PHA (1957). The application of computers to taxonomy. *Journal of General Microbiology* **17**(1)**:** 201-226.

Soares P, Ermini L, Thomson N, Mormina M, Rito T, Rohl A *et al* (2009). Correcting for purifying selection: An improved human mitochondrial molecular clock. *American Journal of Human Genetics* **84**(6)**:** 1-20.

Southworth FC (1976). On subgroups in Dravidian. *International Journal of Dravidian Linguistics* **5**(1)**:** 114-137.

Southworth FC (2004). *Linguistic archaeology of South Asia*. Routledge: Abingdon, UK.

Southworth FC (2005). *Linguistic archaeology of South Asia*. Routledge: London.

Sperber D, Claidiere N (2006). Why modeling cultural evolution is still such a challenge. *Biological Theory* **1**(1)**:** 20-22.

Spriggs M (1984). The Lapita cultural complex: Origins, distribution, contemporaries and successors. *The Journal of Pacific History* **19**(4)**:** 202-223.

Steel M, Penny D (2000). Parsimony, likelihood, and the role of models in molecular phylogenetics. *Molecular Biology and Evolution* **17**(6)**:** 839-850.

Steever S (1997). Introduction to the Dravidian languages. In: Steever S (ed) *The Dravidian Languages*. Routledge: New York, pp 1-39.

Stocking GWJ (1968). Edward Burnett Tylor. In: Sills DL (ed) *International Encyclopedia of the Social Sciences*. McMillan Company: New York. Vol. 16, pp 170-177.

Stoneking M (1998). Women on the move. *Nature Genetics* **20**(3)**:** 219-220.

Stoneking M, Krause J (2011). Learning about human population history from ancient and modern genomes. *Nature Reviews Genetics* **12**(9)**:** 603.

Sullivan J, Swofford DL (2001). Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? *Systematic Biology* **50**(5)**:** 723-729.

Summerhayes GR, Leavesley M, Fairbairn A, Mandui H, Field J, Ford A *et al* (2010). Human adaptation and plant use in highland New Guinea 49,000 to 44,000 years ago. *Science* **330**(6000)**:** 78-81.

Swadesh M (1952). Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society* **96**(4)**:** 452-463.

Swadesh M (1955). Towards Greater Accuracy in Lexicostatistic Dating. *International Journal of American Linguistics* **21**(2)**:** 121-137.

Swadesh M (1972). What is glottochronology. *The origin and diversification of languages***:** 271-284.

Swenson NG, Enquist BJ, Thompson J, Zimmerman JK (2007). The Influence of spatial and size scale on phylogenetic relatedness in tropical forest communities. *Ecology* **88**(7)**:** 1770-1780.

Swofford DL (1998). *PAUP 4.0: phylogenetic analysis using parsimony*. Sinaeur Associates: Sunderland, MA.

Swofford DL, Sullivan J (2003). Phylogeny inference based on parsimony and other methods using PAUP*. In: Salemi M and Vandamme A (eds) *The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny*. Cambridge University Press: Cambridge. Vol. 7, pp 160-206.

Tajima F (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**(2)**:** 437-460.

Tamura K, Dudley J, Nei M, Kumar S (2007). MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Molecular Biology and Evolution* **24**(8)**:** 1596-1599.

Tang J, Hanage WP, Fraser C, Corander J (2009). Identifying currents in the gene pool for bacterial populations using an integrative approach. *PLoS Computational Biology* **5**(8)**:** e1000455.

Taraka R, Kolachina S (2013). Distance-based phylogenetic inference algorithms in the subgrouping of Dravidian languages. In: Borin L and Saxena A (eds) *Approaches to Measuring Linguistic Differences*. De Gruyter Mouton: Berlin, Germany. Vol. 265, p 141.

Tarazona-Santos E, Carvalho-Silva DR, Pettener D, Luiselli D, De Stefano GF, Labarga CM *et al* (2001). Genetic differentiation in South Amerindians is related to environmental and cultural diversity: Evidence from the Y chromosome. *American Journal of Human Genetics* **68**(6)**:** 1485-1496.

Tehrani JJ, Collard M (2002). Investigating cultural evolution through biological phylogenetic analyses of Turkmen textiles. *Journal of Anthropological Archaeology* **21**(4)**:** 443-463.

Tehrani JJ, Collard M (2009). On the relationship between interindividual cultural transmission and population-level cultural diversity: a case study of weaving in Iranian tribal populations. *Evolution and Human Behavior* **30**(4)**:** 286-300.e281.

Temkin I, Eldredge N (2007). Phylogenetics and material cultural evolution. *Current Anthropology* **48**(1)**:** 146-153.

Templeton AR (2010). Reply to Berger et al.: Improving ABC. *PNAS* **107**(41)**:** E158.

Terrell JE (1988). *Prehistory in the Pacific islands*. Cambridge University Press: Cambridge.

Thangaraj K, Chaubey G, Kivisild T, Reddy AG, Singh VK, Rasalkar AA *et al* (2005). Reconstructing the origin of Andaman Islanders. *Science* **308**(5724)**:** 996-996.

Thanseem I, Thangaraj K, Chaubey G, Singh VK, Bhaskar LVKS, Reddy BM *et al* (2006). Genetic affinities among the lower castes and tribal groups of India: Inference from Y chromosome and mitochondrial DNA. *BMC Genetics* **7**(1)**:** 1.

Thapar R (2001). The Rgveda: encapsulating social change. In: Panikkar K, Byres T and Patnaik U (eds) *The Making of History: Essays presented to Irfan Habib*. Tulika: New Delhi, pp 11-40.

Thomas PK, Joglekar PP (1994). Holocene faunal studies in India. *Man and Environment* **19**(1-2)**:** 179-202.

Thomason SG, Kaufman T (1988). *Language contact, creolization, and genetic linguistics*. University of California Press: Berkley.

Thomson B, Corney BG, Stewart J (1908). *The Fijians: A study of the decay of custom*. W. Heinemann: London.

Tishkoff SA, Verrelli BC (2003). Patterns of human genetic diversity: Implications for human evolutionary history and disease. *Annual review of Genomics and Human Genetics* **4**(1)**:** 293-340.

Trejaut JA, Kivisild T, Loo JH, Lee CL, He CL, Hsu CJ *et al* (2005). Traces of archaic mitochondrial lineages persist in Austronesian-speaking Formosan populations. *PLoS Biology* **3**(8)**:** e247.

Trudgill P (2011). *Sociolinguistic typology: social determinants of linguistic complexity*. Oxford University Press.: Oxford.

Tuttle EH (1927). Dravidian 1 and 2. *American Journal of Philology* **48:** 267-272.

Tyler SA (1969). *Koya: an outline grammar: Gommu dialect*, Vol 54. University of California Press: Berkeley.

Tylor EB (1889). On a method of investigating the development of institutions; applied to laws of marriage and descent. *The Journal of the Anthropological Institute of Great Britain and Ireland* **18:** 245-272.

Underhill P, Kivisild T (2007). Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annual Review of Genetics* **41**(1)**:** 539-564.

Van Oven M, Hämmerle JM, van Schoor M, Kushnick G, Pennekamp P, Zega I *et al* (2011). Unexpected island effects at an extreme: reduced Y chromosome and mitochondrial DNA diversity in Nias. *Molecular Biology and Evolution* **28**(4)**:** 1349-1361.

Van Wyhe J (2005). The descent of words: Evolutionary thinking 1780–1880. *Endeavour* **29**(3)**:** 94-100.

Vayda AP, Rappaport RA (1963). Island Cultures. In: Fosberg FR (ed) *Man's Place in the Island Ecosystem*. Bishop Museum Press: Honolulu, pp 133-142.

Veeramah KR, Connell BA, Pour NA, Powell A, Plaster CA, Zeitlyn D *et al* (2010). Little genetic differentiation as assessed by uniparental markers in the presence of substantial language variation in peoples of the Cross River region of Nigeria. *BMC Evolutionary Biology* **10**(1)**:** 1.

Verdu P, Becker NS, Froment A, Georges M, Grugni V, Quintana-Murci L *et al* (2013). Sociocultural behavior, sex-biased admixture, and effective population sizes in Central African Pygmies and non-Pygmies. *Molecular Biology and Evolution* **30**(4)**:** 918-937.

Watkins JC (2004). The role of marriage rules in the structure of genetic relatedness. *Theoretical population biology* **66**(1)**:** 13-24.

Watts J, Greenhill SJ, Atkinson Q, Currie TE, Bulbulia J, Gray RD (2015a). Broad supernatural punishment but not moralizing high gods precede the evolution of political complexity in Austronesia. *Proceedings of the Royal Society of London B: Biological Sciences* **282**(1804)**:** 20142556.

Watts J, Sheehan O, Greenhill SJ, Gomes-Ng S, Atkinson QD, Bulbulia J *et al* (2015b). Pulotu: Database of Austronesian supernatural beliefs and practices. *PLoS ONE* **10**(9)**:** e0136783.

Weir BS, Cockerham CC (1984). Estimating F-statistics for the analysis of population structure. *Evolution* **38**(6)**:** 1358-1370.

Wells RS, Yuldasheva N, Ruzibakiev R, Underhill PA, Evseeva I, Blue-Smith J *et al* (2001). The Eurasian heartland: A continental perspective on Y-chromosome diversity. *PNAS* **98**(18)**:** 10244-10249.

White DR, Burton ML, Dow MM (1981). Sexual division of labor in African agriculture: A network autocorrelation analysis. *American anthropologist* **83**(4)**:** 824-849.

Wickler S, Spriggs M (1988). Pleistocene human occupation of the Solomon Islands, Melanesia. *Antiquity* **62**(237)**:** 703-706.

Wiens JJ, Graham CH (2005). Niche conservatism: Integrating evolution, ecology, and conservation biology. *Annual Review of Ecology, Evolution, and Systematics* **36:** 519-539.

Wilder J, Kingan S, Mobasher Z, Pilkington M, Hammer (2004). Global patterns of human mitochondrial DNA and Y-chromosome structure are not influenced by higher migration rates of females versus males. *Nature Genetics* **36**(10)**:** 1122-1125.

Wilkins JF (2006). Unraveling male and female histories from human genetic data. *Current opinion in genetics & development* **16**(6)**:** 611-617.

Wilkins JF, Marlowe F (2006). Sex-biased migration in humans: What should we expect from genetic data? *Bioessays* **28**(3)**:** 290-300.

Winterhalder B, Smith E (2000). Analyzing adaptive strategies: Human behavioral ecology at twenty-five. *Evolutionary Anthropology : Issues News and Reviews* **9**(2)**:** 51-72.

Winters C (2007). Did the Dravidian speakers originate in Africa? *BioEssays* **29**(5)**:** 497-498.

Witzel M (1999). Substrate languages in Old Indo-Aryan (Rgvedic, Middle and Late Vedic). *Electronic Journal of Vedic Studies* **5**(1)**:** 1-67.

Wollstein A, Lao O, Becker C, Brauer S, Trent RJ, Nürnberg P *et al* (2010). Demographic history of Oceania inferred from genome-wide data. *Current Biology* **20**(22)**:** 1983-1992.

Wood B, Marlowe F (2011). Dynamics of postmarital residence among the Hadza. *Human Nature* **22**(1-2)**:** 128-138.

Wood ET, Stover DA, Ehret C, Destro-Bisol G, Spedini G, McLeod H *et al* (2005). Contrasting patterns of Y chromosome and mtDNA variation in Africa: Evidence for sex-biased demographic processes. *European Journal of Human Genetics* **13**(7)**:** 867-876.

Xu S, Pugach I, Stoneking M, Kayser M, Jin L, Consortium HP-AS (2012). Genetic dating indicates that the Asian–Papuan admixture through Eastern Indonesia corresponds to the Austronesian expansion. *PNAS* **109**(12)**:** 4574-4579.

Yang Z, Goldman N, Friday AE (1994). Comparison of models for nucleotide substitution used in maximum likelihood phylogenetic estimation. *Molecular Biology and Evolution* **11:** 316-324.

Yang Z, Rannala B (1997). Bayesian phylogenetic inference using DNA sequences: A Markov Chain Monte Carlo method. *Molecular Biology and Evolution* **14**(7)**:** 717-724.

Yngvadottir B (2007). Insights into modern disease from our distant evolutionary past. *European Journal of Human Genetics* **15**(5)**:** 603-606.

Zhivotovsky LA, Underhill PA, Cinnioğlu C, Kayser M, Morar B, Kivisild T *et al* (2004). The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. *American Journal of Human Genetics* **74**(1)**:** 50-61.

Zvelebil K (1970). *Comparative Dravidian phonology*, Vol 80. Mouton: The Hague.

Zvelebil K (1972). The descent of the Dravidians. *International Journal of Dravidian Linguistics* **1**(2)**:** 57-63.

Zvelebil K (1990). *Vol. 3*. Pondicherry Institute of Linguistics and Culture: Pondicherr

8 Appendix

8.1 Genes to Culture: Correlation of Sex-Specific Markers and Post-marital Residence

Table 8-1: Details of main subsistence type ("Subsistence"), the sexual division of labour corresponding to the main subsistence type ("Sexual Division of Labour"), the main residence type ("Residence Main"),and whether the society is monogamous or polygyny is allowed ("Marriage Composition"). The "Subsistence" is coded based on what activity contributes most to subsistence based on the information from ethnographic atlas. "Agriculture" is casual agriculture, "Agriculture_U" is agriculture where the type type is unknown, "Agriculture_I" is intensive agriculture, "Agriculture_E" is extensive agriculture, "Fishing" is denoted as "Fishing", "Horticulture" is denoted as "Horticulture", "Hunting" is denoted as "Hunting", "Gathering" is denoted as "Gathering", and when there are more than two activities contributing most to subsistence, then both the activities are denoted. The The final codes used for the sexual division of labour ("SDL"), can take the vaues of "E" for equidominantl societies, "F" for matridominant societies and "M" for patridominant societies. The final codes used for the main residence type ("RES"), can take the values of "A" for ambilocal and neolocal societies, "M" for matrilocal societies, "P" for patrilocal societeis. The final codes used for the marriage composition ("MAR") can take on the values "M" for monogamous societies and "P" for polygynous societies.

No	Ethnonym_EA	ABVD	Subsistence	Sexual Division of Labour	Residence Main	Marriage Composition	SDL	RES	MAR
1	Agta	Agta	Agriculture	Equal participation, no marked differentiation	Ambilocal	Monogamy	E	A	M
2	Bissayam	AklanonBisayan	Agriculture_U	Missing Data	Neolocal	Monogamy	-	A	M
3	Badjau	Bajo	Fishing	Males appreciably more	Ambilocal	Monogamy	M	A	M

No	Ethnonym_EA	ABVD	Subsistence	Sexual Division of Labour	Residence Main	Marriage Composition	SDL	RES	MAR
4	Balinese	Bali	Agriculture_I	Males only or almost alone	Patrilocal	Polygyny allowed	M	P	P
5	Sumbawanese	Bima	Agriculture_E	Equal participation, no marked differentiation	Neolocal	Monogamy	E	A	M
6	Bonerate	Bonerate	Agriculture	Equal participation, no marked differentiation	Matrilocal	Polygyny allowed	E	M	P
7	Bontok	BontokGuinaang	Agriculture_I	Missing Data	Patrilocal	Monogamy	-	P	M
8	Buginese	BugineseSoppeng_D	Agriculture	Males appreciably more	Matrilocal	Polygyny allowed	M	M	P
9	Bunun	Bunun	Agriculture_E	Missing Data	Patrilocal	Monogamy	-	P	M
10	Bwaidoga	Bwaidoga	Agriculture_U	Missing Data	Patrilocal	NA	-	P	-
11	Sugbuhanon	Cebuano	Agriculture_I	Equal participation, no marked differentiation	Neolocal	Monogamy	E	A	M
12	Ami	CentralAmis	Agriculture_I	Equal participation, no marked differentiation	Matrilocal	Monogamy	E	M	M

No	Ethnonym_EA	ABVD	Subsistence	Sexual Division of Labour	Residence Main	Marriage Composition	SDL	RES	MAR
13	Chamorro	Chamorro	Agriculture_E	Equal participation, no marked differentiation	Neolocal	Monogamy	E	A	M
14	Trukese*	Chuukese	Agriculture	Males appreciably more	Matrilocal	Monogamy	M	M	M
15	Lifu	Dehu	Agriculture_E	Females only or almost alone	Patrilocal	Polygyny allowed	F	P	P
16	Dobuan	Dobuan	Agriculture_E	Equal participation, no marked differentiation	Matrilocal	Monogamy	E	M	M
17	Kei	ElatKeiBesar	Agriculture_E	Missing Data	Patrilocal	Polygyny allowed	-	P	P
18	MbauFijian	FijianBau	Fishing	Males appreciably more	Patrilocal	Polygyny allowed	M	P	P
19	EastFutunan	FutunaEast	Agriculture_E	Males only or almost alone	Patrilocal	Polygyny allowed	M	P	P
20	Kaoka	GhariGuadalcanal	Agriculture_E	Differentiated but equal participation	Matrilocal	Monogamy	M	M	M

No	Ethnonym_EA	ABVD	Subsistence	Sexual Division of Labour	Residence Main	Marriage Composition	SDL	RES	MAR
21	Hanunoo	Hanunoo	Agriculture_E	Equal participation, no marked differentiation	Matrilocal	Polygyny allowed	E	M	P
22	Hawaiian	Hawaiian	Agriculture_I	Males only or almost alone	Ambilocal	Polygyny allowed	M	A	P
23	Ambonese	HituAmbon	Agriculture_E	Equal participation, no marked differentiation	Patrilocal	Monogamy	E	P	M
24	Iban	Iban	Agriculture_E	Equal participation, no marked differentiation	Ambilocal	Monogamy	E	A	M
25	Dusun	Idaan	Agriculture_E	Differentiated but equal participation	Patrilocal	Polygyny allowed	E	P	P
26	Ifugao	IfugaoBatad	Agriculture_I	Females appreciably more	Ambilocal	Monogamy	F	A	M
27	Javanese	Javanese	Agriculture_I	Differentiated but equal participation	Neolocal	Polygyny allowed	E	A	P

No	Ethnonym_EA	ABVD	Subsistence	Sexual Division of Labour	Residence Main	Marriage Composition	SDL	RES	MAR
28	Ilongot	KakidugenIlongot	Agriculture_I/Hunting/Fishing	Males only or almost alone	Matrilocal	Monogamy	M	M	M
29	Kalinga	KalingaGuinaangLubuagan_D	Agriculture_I	Differentiated but equal participation	Matrilocal	Monogamy	E	M	M
30	Sumbanese	Kambera	Agriculture_I	Females appreciably more	Patrilocal	Polygyny allowed	F	P	P
31	SagadaIgorot	KankanayNorthern	Agriculture_I	Equal participation, no marked differentiation	Ambilocal	Monogamy	E	A	M
32	Kapingamarangi	Kapingamarangi	Agriculture_E	Females only or almost alone	Matrilocal	Polygyny allowed	F	M	P
33	Kenyah	KenyahLongAnap	Fishing	Females appreciably more	Ambilocal	Monogamy	F	A	M
34	Trobriand	Kilivila	Agriculture_E	Differentiated but equal participation	Matrilocal	Polygyny allowed	E	M	P
35	Gilbertese	Kiribati	Fishing	Males appreciably more	Patrilocal	Polygyny allowed	M	P	P
36	Kodi	Kodi	Agriculture_U	Missing Data	Patrilocal	NA	-	P	-

No	Ethnonym_EA	ABVD	Subsistence	Sexual Division of Labour	Residence Main	Marriage Composition	SDL	RES	MAR
37	Bilaan	KoronadalBlaan	Agriculture_U	Missing Data	Neolocal	Polygyny allowed	-	A	P
38	Kosrae	Kusaie	Agriculture_E	Males only or almost alone	Patrilocal	Polygyny allowed	M	P	P
39	Kwaio	Kwaio	Agriculture_U	Missing Data	Patrilocal	Polygyny allowed	-	P	P
40	Ili-Mandiri	Lamaholot	Agriculture_E	Males appreciably more	Patrilocal	Polygyny allowed	M	P	P
41	Manus	Levei	Fishing	Males appreciably more	Patrilocal	Polygyny allowed	M	P	P
42	Macassarese	Makassar	Agriculture_I	Males appreciably more	Ambilocal	Polygyny allowed	M	A	P
43	SantaCruzIslanders	MaloSantaCruz	Agriculture_E	Females appreciably more	Patrilocal	Polygyny allowed	F	P	P
44	Manam	Manam	Agriculture_E	Females appreciably more	Patrilocal	Polygyny allowed	F	P	P
45	Mangarevan	Mangareva	Fishing	Males appreciably more	Patrilocal	Polygyny allowed	M	P	P

No	Ethnonym_EA	ABVD	Subsistence	Sexual Division of Labour	Residence Main	Marriage Composition	SDL	RES	MAR
46	Maori	Maori	Agriculture	Males appreciably more	Patrilocal	Polygyny allowed	M	P	P
47	Maranao	Maranao	Agriculture	Males appreciably more	Matrilocal	Polygyny allowed	M	M	P
48	Marquesan	Marquesan	Agriculture_E	Males appreciably more	Patrilocal	Polygyny allowed	M	P	P
49	Marshallese	MarshalleseE_D	Agriculture_E	Males appreciably more	Ambilocal	Monogamy	M	A	M
50	Mekeo	Mekeo	Agriculture_E	Equal participation, no marked differentiation	Patrilocal	Polygyny allowed	E	P	P
51	Malay	Melayu	Agriculture_I	Equal participation, no marked differentiation	Patrilocal	Polygyny allowed	E	P	P
52	Melenau	MelanauMukah	Horticulture/Fishing	Differentiated but equal participation	Matrilocal	Polygyny allowed	E	M	P

No	Ethnonym_EA	ABVD	Subsistence	Sexual Division of Labour	Residence Main	Marriage Composition	SDL	RES	MAR
53	Malagasy	MerinaMalagasy	Agriculture_I	Females appreciably more	Patrilocal	Polygyny allowed	F	P	P
54	Minangkabau	Minangkabau	Agriculture_I	Males only or almost alone	Missing Data	Polygyny allowed	M	-	P
55	Modang	Modang	Agriculture	Missing Data	Matrilocal	Polygyny allowed	-	M	P
56	SelungMawken	Moken	Fishing	Males appreciably more	Neolocal	Polygyny allowed	M	A	P
57	Molima	Molima	Agriculture_U	Missing Data	Patrilocal	NA	-	P	-
58	Mori	Mori	Agriculture	Missing Data	Matrilocal	Polygyny allowed	-	M	P
59	Nomoian	Mortlockese	Agriculture_U	Missing Data	Matrilocal	Monogamy	-	M	M
60	Mota	Mota	Agriculture_E	Missing Data	Matrilocal	Polygyny allowed	-	M	P
61	Motu	Motu	Agriculture_E	Missing Data	Patrilocal	NA	-	P	-
62	Lakalai	NakanaiBileki_D	Agriculture_E	Missing Data	Patrilocal	Polygyny allowed	-	P	P
63	Nguna	Nguna	Horticulture	Differentiated but equal participation	Patrilocal	Monogamy	E	P	M
64	Nias	Nias	Agriculture_E	Females appreciably more	Patrilocal	Polygyny allowed	F	P	P

No	Ethnonym_EA	ABVD	Subsistence	Sexual Division of Labour	Residence Main	Marriage Composition	SDL	RES	MAR
65	Nissan	Nissan	Agriculture_E	Males appreciably more	Patrilocal	Polygyny allowed	M	P	P
66	Niuean*	Niue	Fishing/Agriculture	Equal participation, no marked differentiation	Patrilocal	Polygyny allowed	E	P	P
67	Numfor	Numfor	Fishing	Differentiated but equal participation	Patrilocal	Polygyny allowed	E	P	P
68	Paiwan	Paiwan	Agriculture_E	Females appreciably more	Patrilocal	Polygyny allowed	F	P	P
69	Palauan	Palauan	Agriculture_E	Females appreciably more	Matrilocal	Polygyny allowed	F	M	P
70	PalawanBatak	PalawanBatak	Gathering	Differentiated but equal participation	Matrilocal	Polygyny allowed	E	M	P
71	Tongarevan	Penrhyn	Fishing	Missing Data	Patrilocal	Polygyny allowed	-	P	P
72	Cham	PhanRangChamEasternCham	Agriculture_I	Males appreciably more	Matrilocal	Polygyny allowed	M	M	P

No	Ethnonym_EA	ABVD	Subsistence	Sexual Division of Labour	Residence Main	Marriage Composition	SDL	RES	MAR
73	Ponapean	Ponapean	Agriculture_E	Males appreciably more	Matrilocal	Polygyny allowed	M	M	P
74	Puyuma	Puyuma	Agriculture_I	Females appreciably more	Matrilocal	Monogamy	F	M	M
75	RapaNui	RapanuiEasterIsland	Agriculture_E	Males appreciably more	Patrilocal	Polygyny allowed	M	P	P
76	Mangaian	Rarotongan	Agriculture_E	Females only or almost alone	Patrilocal	Polygyny allowed	F	P	P
77	Rennellese	Rennellese	Horticulture	Differentiated but equal participation	Patrilocal	Monogamy	E	P	M
78	Rotinese	RotiTermanu_D	Agriculture_U	Missing Data	Patrilocal	Polygyny allowed	-	P	P
79	Rotuman	Rotuman	Agriculture_E	Males only or almost alone	Matrilocal	Monogamy	M	M	M
80	Ulawan	Saa	Agriculture_E	Equal participation, no marked differentiation	Patrilocal	Polygyny allowed	E	P	P
81	Samoan	Samoan	Agriculture_E	Differentiated but equal participation	Ambilocal	Polygyny allowed	E	A	P

No	Ethnonym_EA	ABVD	Subsistence	Sexual Division of Labour	Residence Main	Marriage Composition	SDL	RES	MAR
82	Taninbarese	Selaru	Agriculture_E	Females appreciably more	Patrilocal	Polygyny allowed	F	P	P
83	Sengseng	Sengseng	horticulture	Differentiated but equal participation	Ambilocal	Polygyny allowed	E	A	P
84	Sika	Sika	Agriculture_E	Missing Data	Ambilocal	Polygyny allowed	-	A	P
85	Simboese	Simbo	Agriculture_U	Missing Data	Ambilocal	Polygyny allowed	-	A	P
86	Atayal	SquliqAtayal	Agriculture_E	Females appreciably more	Patrilocal	Monogamy	F	P	M
87	Dahuni	Suau	Agriculture_E	Differentiated but equal participation	Patrilocal	Polygyny allowed	E	P	P
88	Subanun	SubanunSindangan	Agriculture_E	Equal participation, no marked differentiation	Ambilocal	Polygyny allowed	E	A	P
89	Sundanese	Sunda	Agriculture_I	Missing Data	Ambilocal	Polygyny allowed	-	A	P
90	Eromangan	SyeErromangan	Agriculture_U	MIssing Data	Patrilocal	Polygyny allowed	-	P	P

No	Ethnonym_EA	ABVD	Subsistence	Sexual Division of Labour	Residence Main	Marriage Composition	SDL	RES	MAR
91	Toradja	TaeSToraja	Agriculture_E	Equal participation, no marked differentiation	Matrilocal	Monogamy	E	M	M
92	Tagbanua	TagbanwaAborlan_D	Agriculture_E	Equal participation, no marked differentiation	Matrilocal	Polygyny allowed	E	M	P
93	Tahitian	TahitianModern	Agriculture_E	Males appreciably more	Ambilocal	Polygyny allowed	M	A	P
94	Belu	Tetum	Agriculture_E	Differentiated but equal participation	Matrilocal	Polygyny allowed	E	M	P
95	Tikopia*	Tikopia	Agriculture	Females appreciably more	Patrilocal	Polygyny allowed	F	P	P
96	TobaBatak	TobaBatak	Agriculture_I	Equal participation, no marked differentiation	Patrilocal	Polygyny allowed	E	P	P
97	Tokelaun	Tokelau	Fishing	Differentiated but equal participation	Matrilocal	Polygyny allowed	E	M	P

No	Ethnonym_EA	ABVD	Subsistence	Sexual Division of Labour	Residence Main	Marriage Composition	SDL	RES	MAR
98	Tongan	Tongan	Agriculture_E	Males appreciably more	Patrilocal	Polygyny allowed	M	P	P
99	Rareian	Tuamotu	Fishing	Males appreciably more	Ambilocal	Polygyny allowed	M	A	P
100	Tuvaluan	Tuvalu	Fishing	Differentiated but equal participation	Patrilocal	Polygyny allowed	E	P	P
101	EastUvean	UveaEast	Agriculture_E	Males only or almost alone	Patrilocal	Polygyny allowed	M	P	P
102	Choiseulese	VaghuaChoiseul	Agriculture_E	Differentiated but equal participation	Patrilocal	Polygyny allowed	E	P	P
103	Koobe	Vitu	Fishing	Missing Data	Patrilocal	NA	-	P	-
104	Waropen*	Waropen	Gathering, Fishing	Males appreciably more	Patrilocal	Polygyny allowed	M	P	P
105	Manobo	WBukidnonManobo	Agriculture_E	Missing Data	Patrilocal	Polygyny allowed	-	P	P
106	Wogeo	Wogeo	Agriculture_E	Differentiated but equal participation	Patrilocal	Polygyny allowed	E	P	P

No	Ethnonym_EA	ABVD	Subsistence	Sexual Division of Labour	Residence Main	Marriage Composition	SDL	RES	MAR
107	Woleaian	Woleai	Agriculture_E	Differentiated but equal participation	Ambilocal	Monogamy	E	A	M
108	Aua	Wuvulu	Fishing	Males appreciably more	Missing Data	Polygyny allowed	M	-	P
109	Yami	Yami	Agriculture_E	Differentiated but equal participation	Patrilocal	Monogamy	E	P	M

* -Denotes societies where in the Ethnographic Atlas (EA), more than one activity was denoted as the main subsistence type, and the Pulou databse contributed information on the important subsistence type, and that was retained for further analysis

1)Trukese, Tikopia: For both these societies, according to EA, fishing with a differentiated but equal participation of division of labour and agriculture with a patridominant division of labour, equally contributed to the main subsistence type. But according to the Pulou database, agriculture was the “prinicipal”” mode of subsistence, while fishing was classified as “major”. Hence, agriculture with patridominant division of labour was retained

2)Niuean: According to the EA, Niuean societies have fishing with a differentiated but equal participation of division of labour and agriculture with equal and no marked distinction in the divison of labour as contributing to the main subsistence. However, fishing was classified as “medium” importance to subsistence, while agriculture was classified as “major” mode of subsistence, so the latter was retained.

3) Waropen:In this society, fishing with males doing appreciably more division of labour and gathering with males only contributing to division of labour in this activity, were classified as main subsistence categories. The pulou database has classified both as “major” modes of subsistence. As both these activities have a patridominant division of labour, it was retained as “Fishing/Gathering” with a patridominant division of labour.

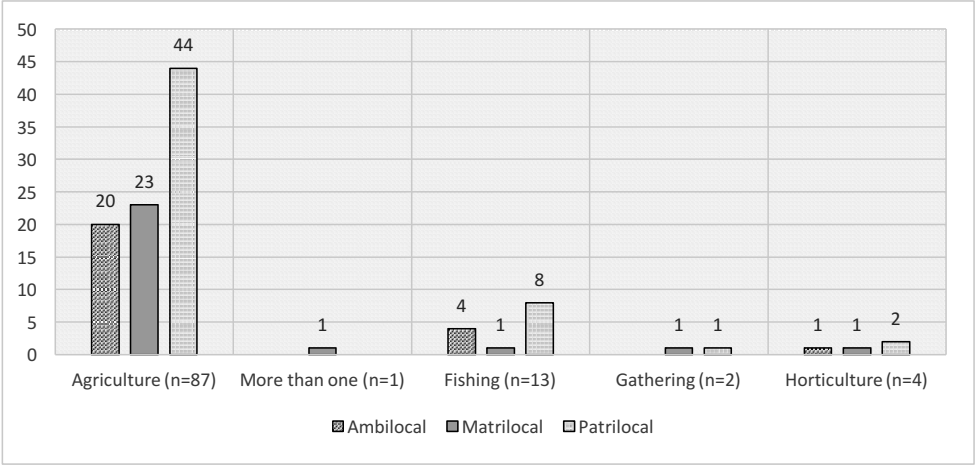


Figure 8–1: Chart depicting the distribution of residence traits amongst the different subsistence types. Y-Axis represent the number of societies.

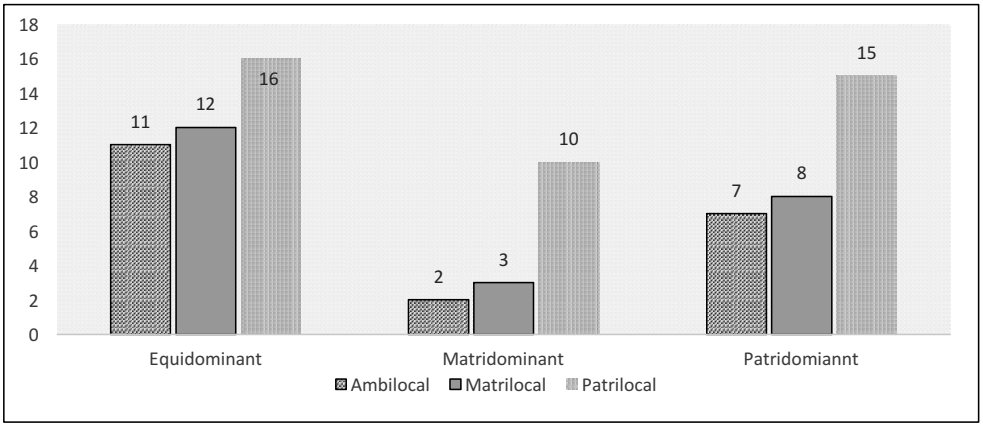


Figure 8–2: Chart depicting the distribution of division of labour traits amongst the different residence types. Y-Axis represents the number of societies.

8.2 Tracing the evolution of the Dravidian language family

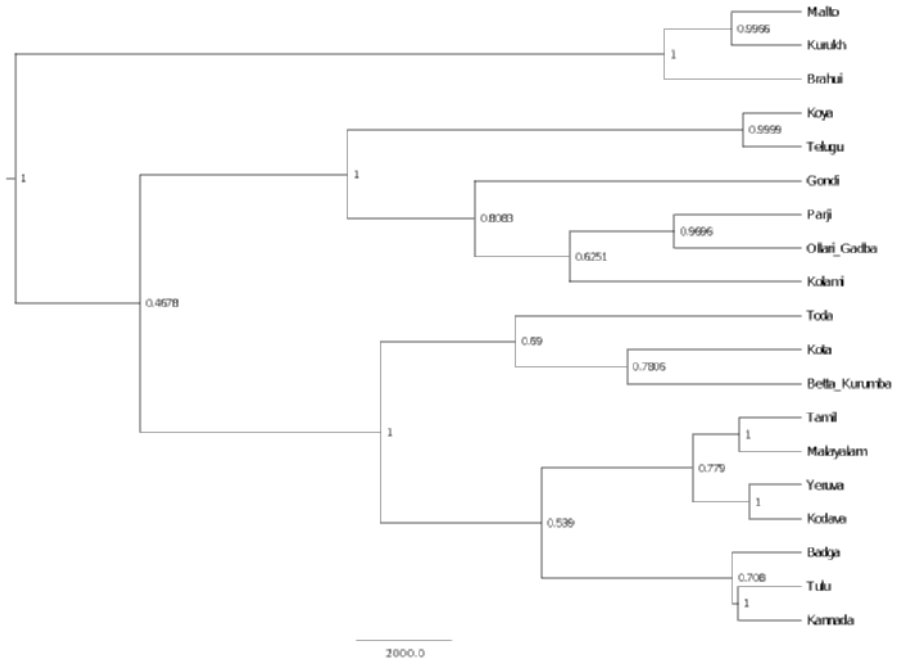


Figure 8–3: : MCC tree from the South II_Central constrained model using relaxed Dollo model of evolution. The numbers on the nodes indicate the confidence of branching. In this model, the South II and Central group of languages were constrained to evolve together

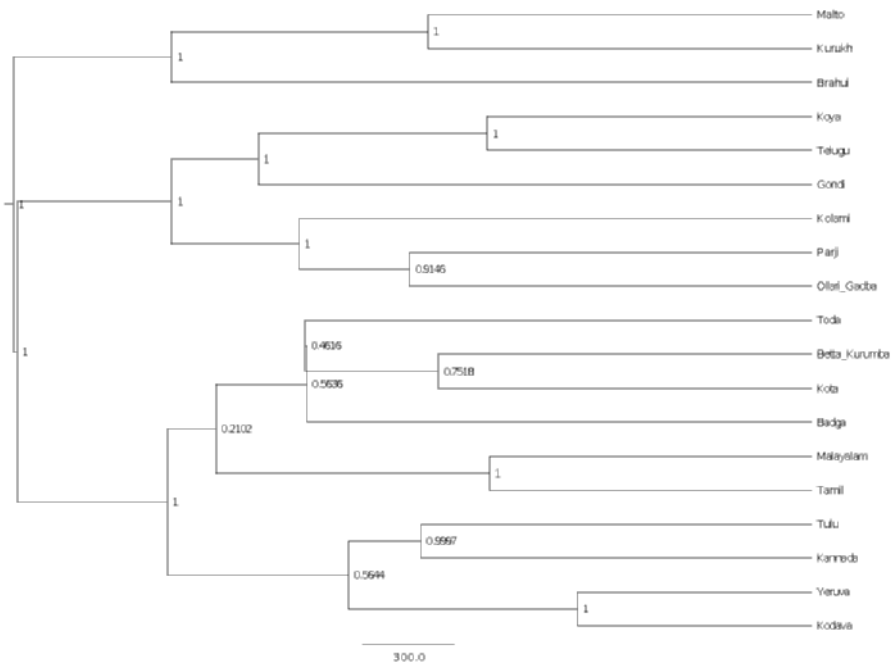


Figure 8–4: MCC tree from N_CS2_S1 constrained model using relaxed Dollo model of evolution. The numbers on the nodes indicate the confidence of branching. In this model, the northern branch was constrained to delineate first, followed by South I from Proto-South II_Central node. The final split separated the South II from the Central group of languages.

Samenvatting

Onderzoek naar de evolutionaire processen die onderliggend zijn aan de patronen van diversiteit in hedendaagse menselijke samenlevingen is cruciaal voor een integraal begrip van onze prehistorie. De oorsprong en de geschiedenis van menselijke samenlevingen zijn complex, en er is geen eenduidige weg om ze te begrijpen. De inzichten van verschillende academische disciplines die licht werpen op deze processen, zoals onder andere de archeologie, antropologie, genetica en taalwetenschap, kunnen alleen op waarde geschat worden als we inzien dat eenieder slechts een deel van het verhaal vertelt. Het begrijpen van de prehistorie vereist een holistische benadering, waarin informatie die gegenereerd is in deze verschillende disciplines geïntegreerd wordt in een gemeenschappelijk kader. In deze dissertatie gebruik ik empirische methoden om interdisciplinaire resultaten te integreren. Technieken uit de populatiegenetica en de evolutiebiologie worden gebruikt om hypotheses over de interactie tussen de evolutie van genen en de evolutie van talen en andere culturele kenmerken te testen. Het uiteindelijke doel is om een beter begrip van de prehistorie mogelijk te maken. Hierbij benader ik talen en culturen als evolutionaire entiteiten. Ik richt me in deze dissertatie vooral op de manieren waarop een multidisciplinaire benadering van de prehistorie a) een aantal van onze aannames met betrekking tot de onderliggende modellen van de evolutie van menselijke samenlevingen kan verhelderen en b) evolutionaire processen onderliggend aan de diversiteit van menselijke samenlevingen kan verklaren.

De genetica of erfelijkheidsleer heeft aanzienlijk bijgedragen aan ons begrip van de prehistorie, vooral sinds de ontwikkeling van moderne moleculaire genetische methoden. Een van de eigenschappen van genen die belangrijk is voor populatiegenetica, is het bestaan van zogenaamde geslachtsgebonden genetische markers. De evolutie van genen op het Y-chromosoom kan ons informatie geven over de herkomst van mannen in een populatie, terwijl genen op het mitochondriaal DNA ons informatie verschaffen over de herkomst van vrouwen. Recent onderzoek in Polynesië heeft de verschillende genetische afkomst van mannen en vrouwen in verband gebracht met culturele processen, zoals de verblijfplaats van een getrouwd stel na het huwelijk: is die bij familie van de vrouw (‘matrilocal’, matrilokaal), bij familie van de man (‘patrilocal’, patrilokaal), of ergens anders (‘neolocal’)? In Polynesië geven genen op het Y-chromosoom aan dat de mannen waarschijnlijk uit Melanesië komen, terwijl de vrouwen (middels onderzoek naar het mitochondriaal DNA) uit Azië komen. Wetenschappers hebben betoogd dat de voorouders van de Polynesiërs hoogstwaarschijnlijk matrilokaal waren, en zodoende trouwden met de plaatselijke mannen die ze tegenkwamen tijdens hun reis (onder andere door Melanesië) om Polynesië te

koloniseren, maar dat zij niet met plaatselijke vrouwen trouwden. Dit is een mogelijke verklaring voor de verschillende herkomst van mannen en vrouwen in Polynesië.

Er is echter een discussie over de vraag of dergelijke culturele praktijken op deze manier de genetische patronen van een populatie daadwerkelijk kunnen beïnvloeden en of dit effect ook waarneembaar is. Een bijkomende vraag is of culturele praktijken de genetische signatuur meteen beïnvloeden, of dat de genen achterlopen op veranderingen in de cultuur. Samenlevingen worden beïnvloed door demografische processen zoals veranderingen in de populatiegrootte door een genetische flessenhals, het kleiner en groter worden de populatie, uitbreiding over nieuwe gebieden, en door stochastische evolutionaire processen zoals genetische drift. Deze processen hebben een bewezen effect op de genetische variatie van een populatie, en daarom is het belangrijk om het effect van culturele praktijken binnen de populatiegenetica te bezien in het licht van demografische en evolutionaire processen. Alleen dan kunnen we bepalen of een culturele verklaring van genetische variatie mogelijk is. In hoofdstuk 2 gebruik ik een nieuwe benadering om dergelijke hypothesen te testen, middels voorwaartse genetische simulatie tests. Ik onderzoek in de Pacific of een verandering in de verblijfplaats na het huwelijk een verandering in genetische variatie teweegbrengt onder verschillende demografische scenario's. Daarnaast kwantificeer ik het aantal generaties dat nodig is om een verandering in genetische patronen te zien volgend op een culturele verandering in de verblijfplaats na het huwelijk. Mijn resultaten geven aan dat er inderdaad een effect is van culturele normen wat betreft verblijfplaats na het huwelijk op veranderingen in de geslachtsgebonden genetische markers, en dat genetische veranderingen achterlopen op culturele, maar dit effect is afhankelijk van demografie, de snelheid van migratie, en genetische drift. Deze laatstgenoemde demografische en evolutionaire processen overschaduwen het effect van verandering in culturele normen, en daarom is het belangrijk om voorzichtig te zijn bij de interpretatie van geschiedenis op basis van alleen genetische informatie. Als genetische informatie wordt gebruikt om de geschiedenis van een populatie te begrijpen, is het de eerste en belangrijkste stap om rekening te houden met demografische en evolutionaire processen zoals populatiegroei, populatiereductie, en genetische drift. In de Pacific moeten we zowel kijken naar culturele normen over de woonplaats na het huwelijk als naar demografische en evolutionaire processen om de verschillende herkomst van mannen en vrouwen te kunnen verklaren.

Culturele normen wat betreft de nieuwe woonplaats van een pasgetrouwd stel zijn dus belangrijk om de genetische variatie in menselijke samenlevingen te begrijpen. Een cultureel kenmerk dat belangrijk is om de evolutie van deze normen te begrijpen, is de arbeidsverdeling tussen man en vrouw. Antropologen beweren dat de verblijfplaats na het huwelijk bepaald wordt door het geslacht wat het meest bijdraagt aan het levensonderhoud. Wereldwijd antropologisch

onderzoek ondersteund deze bewering echter niet. In hoofdstuk 3 test ik de relatie tussen normen wat betreft de verblijfplaats na het huwelijk en de arbeidsverdeling tussen man en vrouw middels comparatieve fylogenetische methoden. Ik focus daarbij op samenlevingen die een Austronesische taal spreken, aangezien zij veel variatie vertonen wat betreft hun verblijfplaats na het huwelijk. Comparatieve fylogenetische methoden zijn statistische methoden die gebruikt kunnen worden om comparatieve data, zoals informatie over verblijfspatronen en arbeidsverdeling, te modelleren op een fylogenetische boom. Een fylogenetische boom is een representatie van de geschiedenis van een groep verwante talen of culturen. Ik maak bij deze analyses dus gebruik van een expliciet evolutionair perspectief op culturele normen.

Om deze analyses te doen verzamel ik culturele informatie over verblijfplaats en arbeidsverdeling in 109 Austronesische samenlevingen, en modelleer ik de co-evolutie van deze kenmerken op een set van Austronesische fylogenetische bomen. Sommige van de kenmerken vertonen clustering naar gelang fylogenetische relatie, anderen niet, en ik vind geen bewijs voor een directe correlatie tussen verblijfplaats en arbeidsverdeling. Vooral de evolutie van de arbeidsverdeling tussen man en vrouw is veel complexer dan voorheen werd gedacht. Mijn voorstel is om deze hypotheses verder te onderzoeken, en om binnen de comparatieve fylogenetische analyse ook ecologische verandering en andere culturele normen over het huwelijk in acht te nemen. Dit hoofdstuk geeft goed aan wat het voordeel is om evolutionaire analyses te gebruiken om culturele normen te begrijpen, en wat voor inzichten een evolutionair perspectief kan geven in de prehistorie.

Informatie over de genetische herkomst en geschiedenis van samenlevingen wordt regelmatig gebruikt om hypotheses uit andere disciplines te verhelderen, waaronder hypotheses uit de antropologie, archeologie, en taalwetenschap. Wetenschappers proberen voordeel te behalen uit de 'relatie' die bestaat tussen genen en cultuur, en zodoende is er een trend ontstaan om informatie over de genetische herkomst van een samenleving te gebruiken om af te leiden wat voor culturele praktijken zij hadden, en vice versa. In hoofdstuk 2 hebben we een voorbeeld hiervan gezien, waar ik de hypothese test dat culturele normen wat betreft verblijfplaats na het huwelijk verschillen in de herkomst van geslachtsgebonden genetische markers kunnen verklaren. Maar zoals ik daar aangetoond heb, verschillen in de geslachtsgebonden genetische markers kunnen ook ontstaan door evolutionaire processen zoals genetische drift en demografische processen zoals populatieflessenhalzen en populatiereductie die een verschillende impact op mannen en vrouwen kunnen hebben.

Om deze lijn van onderzoek nader te beschouwen, test ik in hoofdstuk 4 hypotheses wat betreft de herkomst van de Polynesiërs, en specifiek of demografische of stochastische evolutionaire processen de hedendaagse geslachtsgebonden variatie in genetische patronen kunnen verklaren. Ik probeer

hierbij valkuilen van eerder onderzoek te vermijden en test verschillende hypothesen die geslachtsgebonden genetische variatie kunnen verklaren naast de hypothesen over de migratie van mannen en vrouwen en woonplaats na het huwelijk. De methode die hierbij gebruikt wordt 'coalescent theory' genoemd. Middels dit denkkader en bijbehorende analyses kunnen verschillende hypothesen over de geschiedenis van een set samenlevingen getoetst worden. Als Polynesische mannen en vrouwen daadwerkelijk een verschillende genetische afkomst hebben, komt dit naar voren middels verschillen in de zogenaamde 'admixture estimates' van het best scorende model. De resultaten van deze analyses geven aan dat er geen bewijs was voor een verschillende herkomst van mannen en vrouwen, aangezien beide een sterke genetische bijdrage hebben van de 'Near Oceanic' Austronesische groep. Daarnaast hebben er gedurende de kolonisatie van Polynesië een aantal populatieflessenhalzen plaatsvonden, die een grotere impact op mannen hadden dan op vrouwen, en deze verklaring is waarschijnlijker dan een bewuste strategie tot geslachtsgebonden migratie. Dit geeft aan dat het onderscheiden van populaties in categorieën met een historische betekenis, zoals 'Near Oceanic Austronesians', 'Near Oceanic Non-Austronesians' of 'Remote Oceanians', in tegenstelling tot een generieker label zoals 'Melanesians', een cruciale rol speelt in onze interpretatie van de prehistorie.

De taalwetenschap kan een belangrijke rol spelen om beter te begrijpen welke processen en gebeurtenissen een rol hebben gespeeld in het ontstaan van de hedendaagse diversiteit aan menselijke samenlevingen. Dit is vooral relevant voor gebieden waar studies in archeologie en populatiegenetica niet toereikend zijn geweest om verschillende golven van migratie te onderscheiden of om zones van contact tussen verschillende populaties af te bakenen. Het Indische subcontinent ligt op een kruispunt van migratie, waar van oudsher verschillende populaties doorheen getrokken zijn, en waar zich ook een rijke diversiteit aan samenlevingen gevestigd heeft. In een dergelijke context wordt het begrijpen van de geschiedenis van de mens met behulp van moleculaire genetische methoden een enorme uitdaging, omdat de mix van genen bijna compleet willekeurig wordt en er geen inferenties met zekerheid gemaakt kunnen worden. Hoofdstuk 5 is een fylogenetisch onderzoek naar de oorsprong en evolutie van de Dravidische taalfamilie, die voornamelijk in Zuid-India gesproken wordt, maar ook in Pakistan en Nepal. Ik probeer daarbij specifieke hypothesen over de subgroepen van de taalfamilie die volgen uit eerder onderzoek te testen. Dit onderzoek is een eerste poging om de Dravidische taalfamilie te reconstrueren aan de hand van nieuwverworven data binnen het kader van kwantitatieve, evolutionaire methoden. De resultaten van dit onderzoek geven aan dat Brahui niet een oude tak is binnen de taalfamilie, zoals eerder wel gedacht werd, maar waarschijnlijk een meer recente afstammeling is van de Centrale groep van de Dravidische talen, die ten tijde van de verspreiding van Indo-Europese talen in het subcontinent naar Pakistan gemigreerd is. De topologie van de taalfamilie plaatst de herkomst van

de familie in Centraal India, met een verspreiding van het taalfamilie naar het zuiden, oosten, en noorden.

In de zoektocht naar de herkomst en verspreiding van menselijke samenlevingen, biedt iedere discipline (onder andere de archeologie, antropologie, genetica, en taalwetenschap) een unieke maar gedeeltelijke blik in onze geschiedenis. Als de relevante informatie uit iedere discipline wordt geïntegreerd in een gemeenschappelijk en evolutionair denkkader, komt een compleet beeld veel duidelijker naar voren. De Pacific en de Dravidische taalfamilie zijn heel verschillend, zowel wat betreft de duur van menselijke aanwezigheid, die in India veel langer is als in de eilanden van de Stille Oceaan, als wat betreft de inherente regionale beperkingen die aan historisch onderzoek kleven. Maar ik heb laten zien dat in beide gevallen een gemeenschappelijk denkkader, waarin genetisch onderzoek gepaard wordt met antropologisch of taalkundig onderzoek, kan helpen om hypotheses over de prehistorie van de mens op robuuste wijze te testen. Het is van groot belang om hypothesen en onderzoek uit verschillende disciplines samen te brengen tot een holistisch geheel, want indien dit niet gebeurd is het te gemakkelijk om een verkeerde conclusie te trekken. Naast deze belangrijkste conclusie, hebben de individuele studies in deze dissertatie in belangrijke mate bijgedragen aan onze kennis van de Pacific en de Dravidische taalfamilie. Ik hoop ook het belang van het testen van verschillende modellen en het proactief vergelijken van verschillende hypothesen te hebben gedemonstreerd. Het expliciet testen van modellen en hypothesen moet een grotere rol gaan spelen in het onderzoek naar de prehistorie, vooral middels de comparatieve, phylogenetische, en moleculaire genetische methoden die vandaag de dag beschikbaar zijn. De inzichten die de archeologie, antropologie en taalwetenschap te bieden hebben voor de biologie zijn onmisbaar voor de studie van onze prehistorie en het bouwen aan een gemeenschappelijk denkkader is onmisbaar voor een compleet inzicht in de hedendaagse diversiteit van samenlevingen.

Biographical Note

Vishnupriya Kolipakam was born in Chennai (India) on 4 September, 1985. She did her BSc in Biotechnology at Bharatiar University, Coimbatore, India and graduated in 2006 (*with distinction*). She then proceeded to do her MSc in Biosciences with specialisation in Human genetics, from University of Leeds, UK. She wrote her MSc dissertation on understanding the peopling of Island Southeast Asia, using molecular tools and graduated in 2007. She then worked for 6 months at the National Centre for Biological Sciences (India) on using molecular approaches to understand the history of North-eastern hill tribes in India, and non-invasive genetics to understand status of carnivores in India. She was employed as a researcher at the Centre for Ecological Sciences, Indian Institute of Science till August 2009, to understand the drivers of perception towards human wildlife conflict in semi-arid landscapes of India. In September 2009, Vishnupriya started her Ph.D project *A holistic approach to understanding pre-history* at the Evolutionary processes in language and culture group, Max Planck Institute for Psycholinguistics, Nijmegen. She is currently employed by the National Tiger Conservation Authority, Ministry of Environment Forests and Climate Change and Wildlife Institute of India, Dehradun, as a Project Scientist (Conservation Genetics), of the "Tiger Cell". Her work mainly involves using molecular approaches to understand and preserve India's biodiversity.

MPI Series in Psycholinguistics

1. The electrophysiology of speaking: Investigations on the time course of semantic, syntactic, and phonological processing. *Miranda van Turennout*
2. The role of the syllable in speech production: Evidence from lexical statistics, metalinguistics, masked priming, and electromagnetic midsagittal articulography. *Niels O. Schiller*
3. Lexical access in the production of ellipsis and pronouns. *Bernadette M. Schmitt*
4. The open-/closed-class distinction in spoken-word recognition. *Alette Haveman*
5. The acquisition of phonetic categories in young infants: A self-organising artificial neural network approach. *Kay Behnke*
6. Gesture and speech production. *Jan-Peter de Ruiter*
7. Comparative intonational phonology: English and German. *Esther Grabe*
8. Finiteness in adult and child German. *Ingeborg Lasser*
9. Language input for word discovery. *Joost van de Weijer*
10. Inherent complement verbs revisited: Towards an understanding of argument structure in Ewe. *James Essegbey*
11. Producing past and plural inflections. *Dirk Janssen*
12. Valence and transitivity in Saliba: An Oceanic language of Papua New Guinea. *Anna Margetts*
13. From speech to words. *Arie van der Lugt*
14. Simple and complex verbs in Jaminjung: A study of event categorisation in an Australian language. *Eva Schultze-Berndt*
15. Interpreting indefinites: An experimental study of children's language comprehension. *Irene Krämer*
16. Language-specific listening: The case of phonetic sequences. *Andrea Weber*
17. Moving eyes and naming objects. *Femke van der Meulen*
18. Analogy in morphology: The selection of linking elements in Dutch compounds. *Andrea Krott*
19. Morphology in speech comprehension. *Kerstin Mauth*
20. Morphological families in the mental lexicon. *Nivja H. de Jong*
21. Fixed expressions and the production of idioms. *Simone A. Sprenger*
22. The grammatical coding of postural semantics in Goemai (a West Chadic language of Nigeria). *Birgit Hellwig*
23. Paradigmatic structures in morphological processing: Computational and cross-linguistic experimental studies. *Fermín Moscoso del Prado Martín*
24. Contextual influences on spoken-word processing: An electrophysiological approach. *Daniëlle van den Brink*
25. Perceptual relevance of prevoicing in Dutch. *Petra M. van Alphen*
26. Syllables in speech production: Effects of syllable preparation and syllable frequency. *Joana Cholin*
27. Producing complex spoken numerals for time and space. *Marjolein Meeuwissen*
28. Morphology in auditory lexical processing: Sensitivity to fine phonetic detail and insensitivity to suffix reduction. *Rachèl J. J. K. Kemps*
29. At the same time...: The expression of simultaneity in learner varieties. *Barbara Schmiedtová*
30. A grammar of Jalonke argument structure. *Friederike Lüpke*
31. Agrammatic comprehension: An electrophysiological approach. *Marlies Wassenaar*
32. The structure and use of shape-based noun classes in Miraña (North West Amazon). *Frank Seifart*
33. Prosodically-conditioned detail in the recognition of spoken words. *Anne Pier Salverda*
34. Phonetic and lexical processing in a second language. *Mirjam Broersma*
35. Retrieving semantic and syntactic word properties. *Oliver Müller*
36. Lexically-guided perceptual learning in speech processing. *Frank Eisner*

37. Sensitivity to detailed acoustic information in word recognition. *Keren B. Shatzman*
38. The relationship between spoken word production and comprehension. *Rebecca Özdemir*
39. Disfluency: Interrupting speech and gesture. *Mandana Seyfeddinipur*
40. The acquisition of phonological structure: Distinguishing contrastive from non-contrastive variation. *Christiane Dietrich*
41. Cognitive cladistics and the relativity of spatial cognition. *Daniel B.M. Haun*
42. The acquisition of auditory categories. *Martijn Goudbeek*
43. Affix reduction in spoken Dutch. *Mark Pluymaekers*
44. Continuous-speech segmentation at the beginning of language acquisition: Electrophysiological evidence. *Valesca Kooijman*
45. Space and iconicity in German Sign Language (DGS). *Pamela Perniss*
46. On the production of morphologically complex words with special attention to effects of frequency. *Heidrun Bien*
47. Crosslinguistic influence in first and second languages: Convergence in speech and gesture. *Amanda Brown*
48. The acquisition of verb compounding in Mandarin Chinese. *Jidong Chen*
49. Phoneme inventories and patterns of speech sound perception. *Anita Wagner*
50. Lexical processing of morphologically complex words: An information-theoretical perspective. *Victor Kuperman*
51. A grammar of Savosavo, a Papuan language of the Solomon Islands. *Claudia Wegener*
52. Prosodic structure in speech production and perception. *Claudia Kuzla*
53. The acquisition of finiteness by Turkish learners of German and Turkish learners of French: Investigating knowledge of forms and functions in production and comprehension. *Sarah Schimke*
54. Studies on intonation and information structure in child and adult German. *Laura de Ruiter*
55. Processing the fine temporal structure of spoken words. *Eva Reinisch*
56. Semantics and (ir)regular inflection in morphological processing. *Wieke Tabak*
57. Processing strongly reduced forms in casual speech. *Susanne Brouwer*
58. Ambiguous pronoun resolution in L1 and L2 German and Dutch. *Miriam Ellert*
59. Lexical interactions in non-native speech comprehension: Evidence from electro-encephalography, eye-tracking, and functional magnetic resonance imaging. *Ian FitzPatrick*
60. Processing casual speech in native and non-native language. *Annelie Tuinman*
61. Split intransitivity in Rotokas, a Papuan language of Bougainville. *Stuart Robinson*
62. Evidentiality and intersubjectivity in Yurakaré: An interactional account. *Sonja Gipper*
63. The influence of information structure on language comprehension: A neurocognitive perspective. *Lin Wang*
64. The meaning and use of ideophones in Siwu. *Mark Dingemanse*
65. The role of acoustic detail and context in the comprehension of reduced pronunciation variants. *Marco van de Ven*
66. Speech reduction in spontaneous French and Spanish. *Francisco Torreira*
67. The relevance of early word recognition: Insights from the infant brain. *Caroline Junge*
68. Adjusting to different speakers: Extrinsic normalization in vowel perception. *Matthias J. Sjerps*
69. Structuring language. Contributions to the neurocognition of syntax. *Katrien R. Segaert*
70. Infants' appreciation of others' mental states in prelinguistic communication: A second person approach to mindreading. *Birgit Knudsen*
71. Gaze behavior in face-to-face interaction. *Federico Rossano*
72. Sign-spatiality in Kata Kolok: how a village sign language of Bali inscribes its signing space**.** *Conny de Vos*
73. Who is talking? Behavioural and neural evidence for norm-based coding in voice identity learning. *Attila Andics*
74. Lexical processing of foreign-accented speech: Rapid and flexible adaptation. *Marijt Witteman*
75. The use of deictic versus representational gestures in infancy. *Daniel Puccini*
76. Territories of knowledge in Japanese conversation. *Kaoru Hayano*

77. Family and neighbourhood relations in the mental lexicon: A cross-language perspective. *Kimberley Mulder*
78. Contributions of executive control to individual differences in word production. *Zeshu Shao*
79. Hearing speech and seeing speech: Perceptual adjustments in auditory-visual processing. *Patrick van der Zande*
80. High pitches and thick voices: The role of language in space-pitch associations. *Sarah Dolscheid*
81. Seeing what's next: Processing and anticipating language referring to objects. *Joost Rommers*
82. Mental representation and processing of reduced words in casual speech. *Iris Hanique*
83. The many ways listeners adapt to reductions in casual speech. *Katja Poellmann*
84. Contrasting opposite polarity in Germanic and Romance languages: Verum Focus and affirmative particles in native speakers and advanced L2 learners. *Giuseppina Turco*
85. Morphological processing in younger and older people: Evidence for flexible dual-route access. *Jana Reifegerste*
86. Semantic and syntactic constraints on the production of subject-verb agreement. *Alma Veenstra*
87. The acquisition of morphophonological alternations across languages. *Helen Buckler*
88. The evolutionary dynamics of motion event encoding. *Annemarie Verkerk*
89. Rediscovering a forgotten language. *Jiyoun Choi*
90. The road to native listening: Language-general perception, language-specific input. *Sho Tsuji*
91. Infants'understanding of communication as participants and observers. *Gudmundur Bjarki Thorgrímsson*
92. Information structure in Avatime. *Saskia van Putten*
93. Switch reference in Whitesands. *Jeremy Hammond*
94. Machine learning for gesture recognition from videos. *Binyam Gebrekidan Gebre*
95. Acquisition of spatial language by signing and speaking children: a comparison of Turkish sign language (TID) and Turkish. *Beyza Sümer*
96. An ear for pitch: on the effects of experience and aptitude in processing pitch in language and music. *Salomi Savvatia Asaridou*
97. lncrementality and Flexibility in Sentence Production. *Maartje van de Velde*
98. Social learning dynamics in chimpanzees: Reflections on (nonhuman) animal culture. *Edwin van Leeuwen*
99. The request system in Italian interaction. *Giovanni Rossi*
100. Timing turns in conversation: A temporal preparation account. *Lilla Magyari*
101. Assessing birth language memory in young adoptees**.** Wencui Zhou
102. A social and neurobiological approach to pointing in speech and gesture. *David Peeters*
103. Investigating the genetic basis of reading and language skills. *Alessandro Gialluisi*
104. Conversation Electrified: The Electrophysiology of Spoken Speech Act Recognition**.** *Rósa Signý Gisladottir*
105. Modelling Multimodal Language Processing. *Alastair Smith*
106. Predicting language in different contexts: The nature and limits of mechanisms in anticipatory language processing. *Florian Hintz*
107. Situational variation in non-native communication. *Huib Kouwenhoven*
108. Sustained attention in language production. *Suzanne Jongman*
109. Acoustic reduction in spoken-word processing: Distributional, syntactic, morphosyntatic, and orthographic effects. *Malte Viebahn*
110. Nativeness, dominance, and the flexibility of listening to spoken language. *Laurence Bruggeman*
111. Semantic specificity of perception verbs in Maniq. *Ewelina Wnuk*
112. On the identification of FOXP2 gene enhancers and their role in brain development. *Martin Becker*
113. Events in language and thought: The case of serial verb constructions in Avatime. *Rebecca Defina*
114. Deciphering common and rare genetic effects on reading ability. *Amaia Carrión Castillo*
115. Music and language comprehension in the brain. *Richard Kunert*

116. Comprehending Comprehension: Insights from neuronal oscillations on the neuronal basis of language**.** *Nietzsche H.L. Lam*
117. The biology of variation in anatomical brain asymmetries. *Tulio Guadalupe*
118. Language processing in a conversation context. *Lotte Schoot*
119. Achieving mutual understanding in Argentine Sign Language. *Elizabeth Manrique*
120. Talking Sense: the behavioural and neural correlates of sound symbolism. *Gwilym Lockwood*
121. Getting under your skin: The role of perspective and simulation of experience in narrative comprehension. *Franziska Hartung*
122. Sensorimotor experience in speech perception. *Will Schuerman*
123. Explorations of beta-band neural oscillations during language comprehension: Sentence processing and beyond. *Ashley Lewis*
124. Influences on the magnitude of syntactic priming. *Evelien Heyselaar*
125. Lapse organization in interaction. *Elliott Hoey*
126. The processing of reduced word pronunciation variants by natives and foreign language learners: Evidence from French casual speech. *Sophie Brand*
127. The neighbors will tell you what to expect: Effects of aging and predictability on language processing. *Cornelia Moers*
128. The role of voice and word order in incremental sentence processing. *Sebastian Sauppe*
129. Learning from the (un)expected: Age and individual differences in statistical learning and perceptual learning in speech. *Thordis Neger*
130. Mental representations of Dutch regular morphologically complex neologisms. *Laura de Vaan*
131. Speech production, perception, and input of simultaneous bilingual preschoolers: Evidence from voice onset time. *Antje Stoehr*
132. A holistic approach to understanding pre-history. *Vishnupriya Kolipakam*