17 Molecular Genetic Methods

Carolien G. F. de Kovel and Simon E. Fisher

Abstract

Finding the genetic variation that underlies inter-individual variability in language skills is an important approach for deciphering the biological bases of this fascinating human phenomenon. Recent years have seen dramatic advances in the techniques available for identifying DNA variants that influence human traits, not only for disorders but also for variability in the normal range. The method of choice depends on the genetic architecture of the trait being studied. If the difference between people is due to a single alteration in the DNA with a large effect, an effective strategy is to investigate linkage in multigenerational families. Alternatively, if the variability in the trait depends on the accumulation of small effects of many DNA variants, it is optimal to carry out a genome-wide association study with thousands of participants. This chapter describes the principles behind these complementary methods, and how they can be used to study language-related traits, discussing both the pitfalls and the opportunities.

Introduction

Molecular genetics is a subdivision of genetic research concerned with the structure and functions of genes at the molecular (i.e., DNA/RNA) level. An important part of this type of research involves identifying variations in DNA that are associated with

Research Methods in Psycholinguistics and the Neurobiology of Language: A Practical Guide, First Edition. Edited by Annette M. B. de Groot and Peter Hagoort.

variations in the development of a particular trait. In this chapter, we will explain the practical side of searching for genetic variations that influence a person's language skills. In another subdivision of genetic research, which we do not cover here, researchers aim to decipher the biological pathways by which genetic variations have their effects, tracing out intermediate steps between molecules, cells, tissues, and organisms.

Background

Our abilities to understand and use language are undoubtedly influenced by environment and experience. Yet when such effects are accounted for, people still differ in their language skills. At least some of these inter-individual differences are due to variability in genetic make-up. Decades of behavioral research in families and twin cohorts have provided solid evidence that genetic factors can significantly impact on speech, language, and reading proficiency (see Bishop, 2001; Kovas *et al*., 2005). This chapter will discuss the background and general approaches for identifying the genes and genetic alterations involved. Once critical genes have been identified they provide molecular windows for understanding biological processes involved in the trait (Fisher & Scharff, 2009). For example, we could determine which parts of the brain are affected by the relevant genetic alterations, and at what stages of development. We focus here on the (molecular genetics) techniques for first finding connections between genes and language traits.

Before explaining the key methods, it is worth briefly recapitulating the basics of genetics. Every human cell contains strings of DNA. DNA is a huge molecule built by putting together smaller units, usually referred to as nucleotide bases. These bases come in four types: A (adenine), C (cytosine), G (guanine), and T (thymine). DNA is therefore usually represented as a sentence composed of sequences of these four letters. The long string of DNA that makes up our *genome* is organized in 23 different pieces, the *chromosomes*. Our cells contain two copies of each chromosome: one inherited from the mother (maternal) and one inherited from the father (paternal). Sequences of DNA letters (As, Cs, Gs, and Ts) provide the instructions for assembling strings of *amino-acids* into *proteins*, which in turn form the molecular machinery that make our bodies function; enzymes that catalyze reactions, molecules that define the structure of a cell, signaling factors and receptors, to highlight just a few examples. A stretch of DNA that encodes a particular protein is called a *gene*. However, only a small proportion of our genome (<1.5%) codes for proteins. The remainder includes features that regulate when and where proteins are constructed from the DNA code, and how much of each protein should be made. Nevertheless, the potential functional significance of much of the genome's *non-coding DNA* (i.e., the DNA that does not code for protein) remains to be determined.

Gametes (eggs or sperm) carry only one copy from each pair of chromosomes, selected at random during the production of the eggs and sperm. When egg and sperm from two parents fuse, the resulting embryo again has a double set of each chromosome. Each pair of chromosomes is known by a number (1-22), except for the sex chromosomes X and Y. A crucial point for understanding genetic mapping is that during the formation of the gametes, the two chromosomes of a pair line up with each other and may exchange material in a process called *crossing-over*

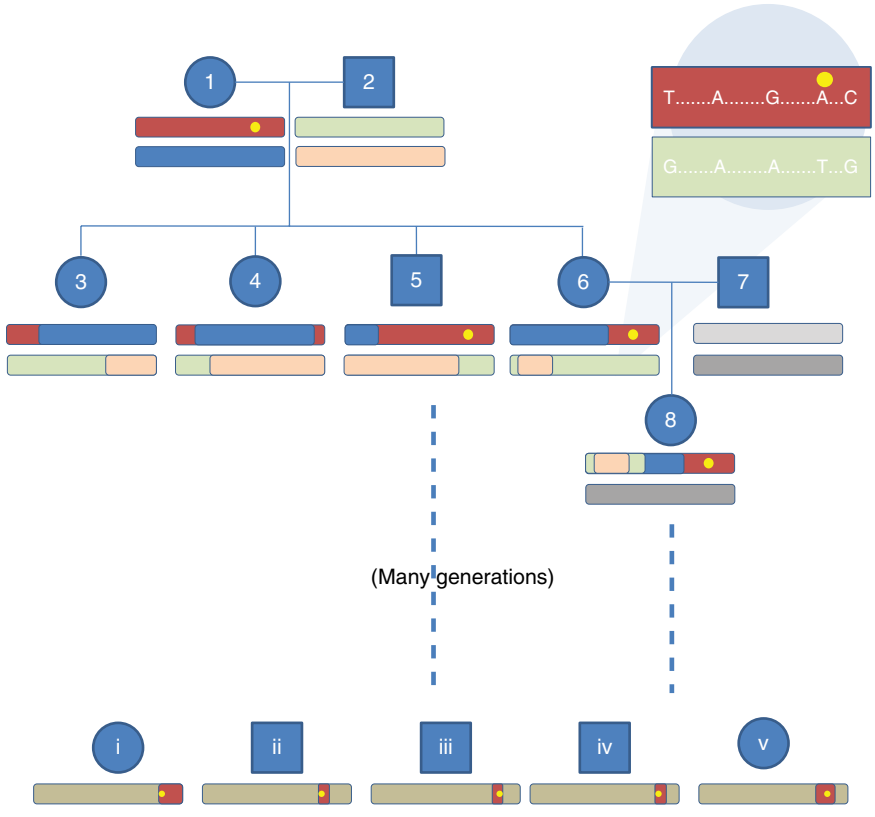

Figure 17.1 Transmission of DNA between generations.
Top: Males are represented by squares, females by circles. In this pedigree one pair of chromosomes (out of 23 pairs) is shown below each individual. Grandmother *1* carries a yellow DNA-variant on her red chromosome, which influences trait X. She transmits her chromosomes to her offspring (individuals *3*, *4*, *5*, *6*), and because of crossing-over between the red and the blue chromosome during egg production, each child gets a different combination. Half of her children inherit the variant that influences trait X.
Top (right): Zoomed in, each chromosome can be represented as a string of letters. At most positions (the dots) the chromosomes are identical to the Human Reference Genome. At some positions (the letters) at least one of the chromosomes differs from the Reference. Such differences are on average a few hundred letters apart. The A with the yellow dot influences trait X.
Bottom: Many generations later some of the descendants of *1* still carry the yellow DNA-variant. The stretch of red chromosome surrounding it has shrunk, but in a different way in each descendant. However, individuals *i*, *ii*, *iv* and *v* still carry the C to the right of the yellow A, while *iii* and *v* still carry the G to the left of it. In a GWAS these two variants may show association with trait X. (*See insert for color representation of the figure.*)

(Figure 17.1). As a consequence, each maternal chromosome in the resulting egg cell is effectively a patchwork of stretches of DNA originating from both maternal grandparents. Similarly, every sperm cell carries a combination of DNA stretches from the different paternal grandparents. Thus, there is a shuffling of genetic information at each generation.

Understanding DNA Variation

The genomes of two unrelated people are typically identical for more than 99% of their length. However, since a human genome is 3.1×10^9 DNA-letters (times two copies), even a ~1% difference means that, on average, one person differs from the next in $\sim 3.6 \times 10^6$ DNA letters (The 1000 Genomes Project Consortium, 2015). The vast majority of the DNA variations that a person carries were inherited from her/his parents. In addition, during the production of gametes, a few errors are made in copying the DNA. As a consequence, each individual also carries about 50 new (*de novo*) variants that were not present in the genomes of either parent. Since most of our genome does not code for proteins, a lot of the variants (whether inherited or new) have little consequence. Even when a variant is located within the coding sequence of a gene, it does not always lead to a change in the encoded protein. This is because the coding system whereby sequence information in DNA is read off for building proteins contains some redundancy: Different three-letter DNA-codes are translated into the same amino acid (e.g., GCA, GCG, GCT, and GCC all correspond to the amino acid *alanine*).

On average, when compared to a standardized reference genome (see https://genome.ucsc.edu/ or http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/), each person carries ~10,000 changes that yield differences in protein sequences. As mentioned above, almost all of these are inherited from one or other parent; typically just one or two of those protein-coding differences are new (Veltman & Brunner, 2012). Most protein-coding variations are relatively harmless. They may contribute to differences in appearance and behavior between us and our neighbors. And because we share inherited variants to some extent with our relatives, they cause similarities within families. However, some DNA variations cause or contribute to susceptibility for disease. In addition to simple changes of a DNA letter, one person's genome differs from that of the next person in missing pieces (*deletions*), having extra pieces (*duplications*), having stretches inverted (*inversions*), or having multiple copies in tandem of a given stretch, and more. For an overview of all sorts of variation and their consequences, see The 1000 Genomes Project Consortium (2015). For simplicity, we will mainly focus here on single-letter alterations.

Genomic diversity has been intensively studied in recent years, and we now know a great deal about different types of changes. Many DNA variants are fairly common in general populations. To take an arbitrary example, at a given chromosomal position, 80% of the genomes in a human population might carry an A, while the remaining 20% carry a C. The alternative letters at the same position are referred to as *alleles*; in this case allele A and allele C. Variations that are common—that is, more than 1% of chromosomes in a given population have the rarer allele—are known as *polymorphisms*. Because each individual has two copies of every piece of DNA (one paternal, one maternal), for our arbitrary example a person may either have two A alleles (*homozygous* A), A plus C (*heterozygous*), or two C alleles (*homozygous* C). The combination of the two alleles at a position is called the *genotype*. Most polymorphisms are close to neutral with respect to health. If they are harmful, fewer children with the damaging allele survive in every generation, and eventually the variant disappears. If, on the other hand, one allele confers an advantage as compared to the alternative allele at that position, it becomes more frequent every generation until it is fixed, meaning that the alternative allele is lost. These processes are called

selection. Without selection, the allele frequencies of a polymorphism remain roughly constant over time in a population.

Genetic Architecture

Different traits or disorders can differ in the nature of their *genetic architecture*. Here, we will consider two well-studied extremes. Certain differences in traits between people can be caused by a single genetic change of large effect, a type of genetic architecture that we call *monogenic*. Many severe diseases are monogenic: An important gene gets disrupted and results in, for example, deafness, blindness, intellectual disability, or some other major disorder affecting one or more tissues of the body. Deleterious DNA variants with large effects are typically rare in the general population, because they are purged by selection. Monogenic traits often show strong clustering within families, and can be identified by their inheritance patterns. Beyond disorders, a frequently cited example of apparent monogenic inheritance in the general population is the ability to taste the bitter compound phenylthiocarbamide (PTC), which is largely determined by variation in the *TAS2R38* receptor, involving two alternative common alleles (a "G" instead of an "A" at one point in this gene). However, in recent years it has become clear that PTC tasting varies along a continuum, and is not purely monogenic (Bufe *et al*., 2005); it is suspected that other genes, as well as environmental factors, modulate a person's abilities in this regard.

This leads us toward the other extreme of genetic architecture. Some traits are far from monogenic, being influenced by the joint action of a large number of DNA variations, occurring in many different genes, that each have a small effect on the trait. Height is a good example of such a *multifactorial* trait. Many of the relevant DNA variations have such small effects by themselves on survival or fecundity that they are not filtered out by selection even if the trait they contribute to is detrimental, thus remaining polymorphic in the general population. While height is a quantitative trait with a continuous distribution, the multifactorial model can also apply to dichotomous traits and diseases. The seesaw provides a useful analogy. A small weight placed on the higher seat will not cause it to topple, but if you keep piling up additional weight, at some point the higher seat will suddenly come down. In a similar way, once a person has a dangerously large number of deleterious DNA variations he or she may develop a particular disease, whereas people with a lower number of those variations are fine. Dichotomous traits with a multifactorial basis cluster less strongly in families than monogenic traits, since there is a low probability that someone will transmit the total package of deleterious DNA variations to a child, given that they may be located at many different sites of the genome. Common experience with multifactorial quantitative traits teaches us that being tall or short clusters in families to a certain degree, but that extreme parents often have less extreme children, while average parents occasionally have an exceptionally tall or short child. This is how it is with most multifactorial traits.

In both types of genetic architecture—monogenic and multifactorial—environmental factors may also contribute. Moreover, it is possible for a trait to lie between these extremes of monogenic and multifactorial architecture, for example by involving interactions of variants of medium effect size in a relatively small number of genes. Such intermediate models are poorly understood at this time.

Introducing the General Approach

If we want to identify genes involved in variation in language skills, the strategy used depends on assumptions about the genetic architecture. Nonetheless, most approaches posit that a DNA variant influencing the trait originated at some point in time, and was transmitted to the next generation, together with surrounding stretches of DNA (Figure 17.1). Because of successive events of crossing-over, the surrounding section of co-transmitted DNA (linked to the variant of interest, and hence to the trait) gets smaller and smaller, the more generations pass (Figure 17.1). People showing the same trait are therefore likely to share the putative DNA variant that contributes to it along with a surrounding stretch of DNA. The more distantly related these people are, the shorter will be this section of shared DNA. Within a family, the regions of shared DNA around a causal variant can be as large as a quarter of a chromosome. If we collect from the general population seemingly unrelated people who share a particular trait, the stretch of shared DNA around a causal variant can be as small as one or two genes. These people may seem unrelated, but they have all inherited that particular stretch of DNA from the same distant ancestor. People who share the same stretch of DNA not only share the variant of interest, but also variants at a number of presumably neutral neighboring polymorphisms (see Figure 17.1, bottom). By determining the genotypes of these common polymorphisms in people, we can map out where shared sections of DNA lie. This is an important step toward pinpointing the locations of the causative variants themselves.

Monogenic traits are usually studied in families with multiple relatives affected by the trait. The aim is to locate a stretch of DNA that the affected relatives share with each other, but not with the unaffected family members, that is, a chromosomal region where all the variants show *linkage* with the trait.

Multifactorial studies, in contrast, involve analyzing a set of unrelated people affected by a trait or disorder (cases) and comparing their genotypes to those who are unaffected (controls). In such a case-control design, we expect variants that contribute to the trait, and additional variants in the surrounding DNA, to be more common among cases than controls. However, not all cases will carry the same set of trait-related DNA variants. Also, within a multifactorial framework, a trait-related variant found in cases can also be carried by controls, since it is ultimately the overall load of risk variants in multiple genes that contributes to whether or not the trait develops (as in the seesaw analogy described earlier). Thus, in multifactorial studies, rather than testing for presence/absence of a particular allele at a polymorphic marker, we compare whether the frequency of the allele differs between cases and controls. Sometimes the trait of interest can be indexed by a quantitative measure that shows continuous variation in a population (standard examples from biomedical fields include height and blood pressure) rather than presence/absence. For these traits we can either compare people who lie at the trait extremes, or collect a random sample of people from throughout the normal distribution showing a range of different values. The choice of traits or trait combinations (*phenotypes*) to study, along with optimal ways to approach quantitative traits, is discussed later, with particular reference to speech, language, and reading skills.

Because of technical limitations in laboratory techniques or in computing possibilities, we may first try to identify the broad location of the suspected causal DNA

variant in the genome and only later search for the particular DNA variant itself. In the past, this was the normal approach for family studies, but recent technological advances have made it possible to start looking for the causal DNA variant directly by means of *next-generation sequencing* (NGS) (Metzker, 2010). The traditional method is still in use, though, for practical reasons or because of the costs. We will go into the details later. In studies of multifactorial traits, unless there is a clear-cut prior hypothesis concerning a specific gene, it is necessary to start with a systematic search of hundreds of thousands of polymorphisms across the genome. This is called a *genome-wide association scan* (*GWAS*) (McCarthy *et al.*, 2008). As discussed later, a GWAS requires thousands of individuals to yield adequate statistical power. Obtaining and analyzing sequence data of the whole genome rather than just a set of polymorphisms in cohorts of this size is not yet feasible for most laboratories, so DNA-chip technology is used to read each individual's genotype (DNA letters) at a great many common polymorphisms.

Whichever approach we choose, statistics are crucial. When we observe a genetic difference between affected and unaffected individuals, rigorous statistical analyses are required to determine whether or not this can be explained by random sampling error. Indeed, the proper statistical methodologies for genetic analyses are an intensely studied field. In addition to robust statistical support for a finding, along with replication in independent cohorts, we often want to collect evidence that the functions of a particular gene are relevant to the trait of interest and that they might be altered by the genetic changes we observe. This can be done by a variety of experiments, for example using cells grown in the laboratory or animal models, which we will not discuss in this chapter. However, a large amount of knowledge about genes, what they do, where and when they are switched on in the body, and other aspects has already been collected. As such, geneticists spend a lot of time mining the available information from public (online) databases for information on the various candidate genes highlighted by their genetic mapping studies.

Techniques for Characterizing Genetic Variation

To give an idea of the lab-work involved in performing a genetic study, we will here describe a few common techniques.

In order to analyze the genes of a study participant we must first isolate the person's DNA. Because the genome sequence is virtually the same in every cell of the body, we preferably use a tissue that is easily sampled and processed. Traditionally, blood has been the tissue of choice, especially since it gives particularly large yields of high quality DNA. In situations where drawing blood is difficult (e.g., participants have fear of needles), we can collect DNA non-invasively from other tissue such as the inside of the cheek, sampled by buccal swabs or saliva sampling. Saliva sampling can even be done by mailing participants a prepared container to spit in, and having it returned to the lab. Once blood or saliva samples are in the lab, extracting and purifying the DNA can be done with commercial kits.

Several alternative techniques are used to read the nucleotide letters from the DNA sample of a participant. We may read out the individual letters in consecutive sequence from an entire stretch of DNA (the size of which might vary depending on

the technique). This type of approach is called DNA *sequencing*. For reasons of speed, costs, and computational ease, we may in some study designs choose to only assess which DNA-letter(s) a person carries at a predefined set of known polymorphisms. In that case, we do not read complete "sentences," but only a single letter here and there. With currently available methods, the number of polymorphisms that are investigated could range from just a single variant to hundreds of thousands of known polymorphisms. This is generically known as *genotyping*. We will describe both sequencing and genotyping in more detail.

Sequencing

The aim of sequencing is to read the code of a given stretch of DNA letter by letter. There are currently two prominent types of technology for doing so: traditional "Sanger"-techniques, and more novel high-throughput *massive parallel sequencing* techniques, also known as *next-generation sequencing* (NGS), which have emerged during the past decade. The output from these techniques is typically not the sequence of just a single DNA molecule from a single cell, but the average of the sequence of many molecules from many cells. If at a given position in the DNA the nucleotide letter you inherited from your father is different from the one that you inherited from your mother (i.e., you are heterozygous at this position), half of the sequenced molecules will have the paternal letter at that position, while the other half will have the maternal letter at that same position. For example, for a particular stretch of DNA sequence a person's code might be read as "GTGCAAGA(**C/T**)GAGACAGGTAAA," indicating that half the molecules are "GTGCAAGA**C**GAGACAGGTAAA" while the other half are "GTGCAAGA**T**GAGACAGGTAAA" (Figure 17.2). Unless the corresponding sequences of the mother and father have also been determined, the result does not tell you which letter (in this case, C or T) was inherited from which parent.

Traditional Sanger techniques are still considered to be of better quality than the available NGS techniques, with higher sensitivity and specificity, but NGS is rapidly catching up. To perform Sanger sequencing, one must first isolate a specific stretch of interest from the long molecules of DNA. This is done with a technique called *polymerase chain reaction* (PCR), in which a particular region of the genome is selectively and exponentially amplified from the original DNA sample, generating large numbers of copies of this target region (https://youtu.be/iQsu3Kz9NYo). The amplified material is then used as a template in a sequencing reaction. A single such reaction typically reads stretches of up to ~800 letters. Most protein-coding genes are substantially larger than this, so it is almost always necessary to carry out multiple reactions to cover the full length of a gene. This technique is relatively low-throughput and preferred if sequencing only a few stretches of DNA per individual, in which case it is faster and cheaper than NGS. The material costs including PCR are around $2 per reaction per individual (as estimated in 2016). During Sanger sequencing, DNA molecules resulting from the PCR-procedure, are read by a sequencing machine. Each "letter" that is encountered generates a fluorescent signal, with a different color for each letter (usually A = green, C = blue, G = black, T = red). A series of differently colored fluorescent signals reveals the DNA sentence that was offered to the machine. More detailed explanations can be found on YouTube, for example, https://youtu.be/e2G5zx-OJIw. Figure 17.2 shows a visualization of Sanger-sequencing results.

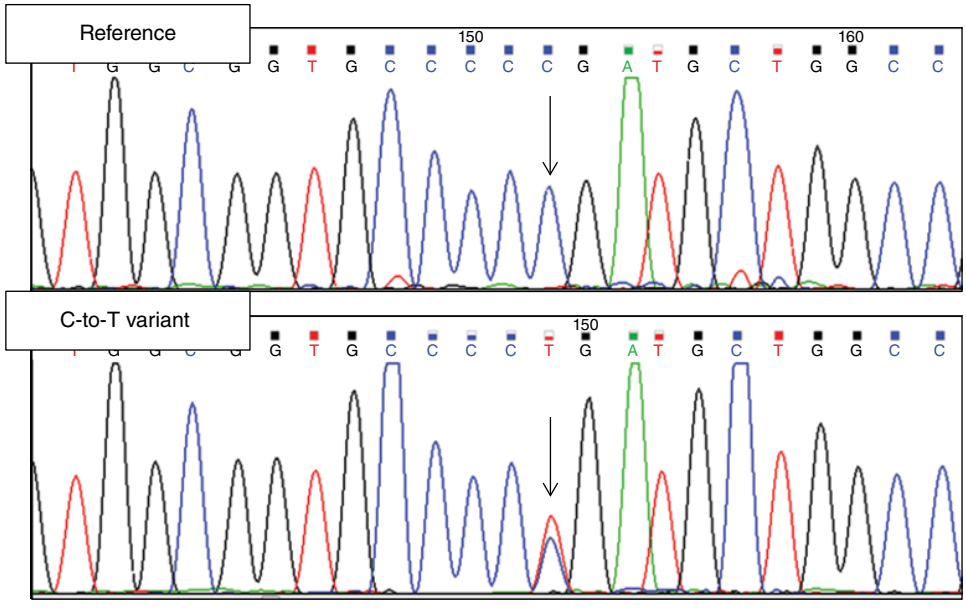

Figure 17.2 Visualization of Sanger sequencing results.
Sanger sequencing results for two individuals for the same stretch of DNA. On the X-axis, the position along the sequenced fragment of DNA; on the Y-axis, the fluorescence intensity for four different colors. A different color lights up for each of the bases that is read: A=green, C=blue, G=black, T=red. In the middle of the lower image, two different colors light up on the same position (arrow), because the individual has inherited different letters from his father and mother: the reference letter C and the variant letter T. Some background coloring can be seen near the bottom of each image. This is an artefact. (*See insert for color representation of the figure.*)

Next Generation Sequencing (NGS) is a technology that is preferred when you need to read a large number of letters of DNA per individual. With NGS it is possible to read all 3.1×10^9 letters of a person's genome in a single experiment (*whole genome sequencing*, or WGS). Alternatively, a method called *enrichment* can be used to initially isolate all known protein-coding parts (~5.5×10^7 letters, known as the *exome*), before sequencing only these sections (*whole exome sequencing*, or WES). NGS reads 50 to 300 letters at a time, depending on the platform and equipment being used (Goodwin, McPherson, & McCombie, 2016). At the end of the experiment, the database in the sequencing machine contains millions of short DNA "sentences," along with information on their reliability. Intensive computer analyses are required to make sense of all these data. Usually, this involves aligning each DNA-sentence to the matching part of the full "text" of the Human Reference Genome (http://www.ensembl.org/Homo_sapiens/Info/Index), a little like assembling the pieces of a huge jigsaw (albeit one that is linear). Multiple sentences will overlap at every position in the text, meaning that each letter has been read several times, increasing the confidence in the accuracy of the sequence information (Figure 17.3). Then, positions in the data that deviate from the reference genome can be identified and listed. Processing of NGS data is highly demanding in terms of computer time, power, and storage capacity.

Depending on the quality required, WGS costs around $1200 per sample (as estimated in 2016). WES is currently cheaper (~$500) including the cost of the enrichment. The investment costs for equipment and for computer infrastructure are

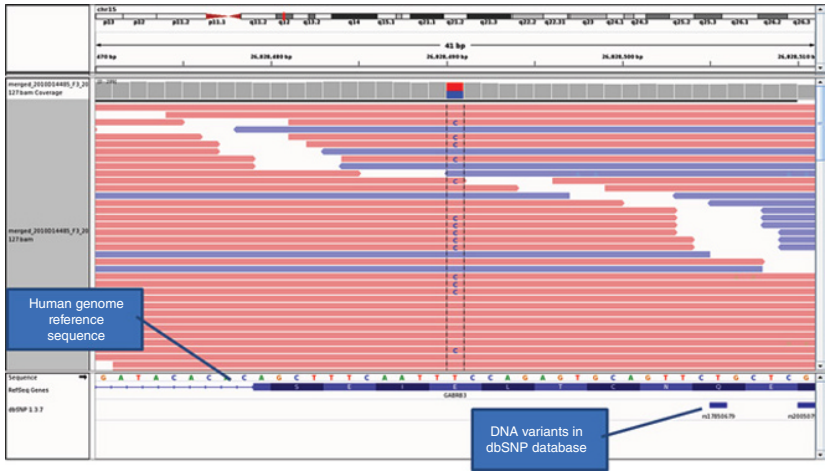

Figure 17.3 Next generation sequencing.
Next Generation Sequencing data for one individual for a short stretch of the DNA of gene *GABRB3*. Each horizontal bar (blue or pink) represents a single sequenced molecule. The sequences are aligned to the Human Reference Genome (bottom). If the sequence differs from the Reference, this is indicated—see blue C in the middle. About half of the molecules carry the C, the other molecules carry the T (as indicated in the Reference). This individual is heterozygous at this point. Either a C was present in the DNA from one of his parents and a T in the other parent, or the C originated *de novo* during egg or sperm production. (*See insert for color representation of the figure.*)

considerable, which makes outsourcing a common solution for most laboratories. While any student can carry out Sanger sequencing, NGS typically needs dedicated technicians and experts in bioinformatics.

Genotyping

For a large number of positions in the human genome, previous experiments have shown that people carry different letters, called polymorphisms, as explained earlier in the chapter. For most known polymorphisms only two of the possible four letters are common in the general population (i.e., there are two alternative alleles). Publically available online databases have collated information about these polymorphisms, including their allele frequencies in various ethnic populations across the world. One of the most well-used databases, dbSNP (http://www.ncbi.nlm.nih.gov/SNP/) catalogues over 150 million different single nucleotide variants (July 2016). In the early days of molecular genetics, it was necessary to genotype variants one by one in DNA-samples of interest. Genotyping practices were revolutionized at the end of the 1990s by the development of glass slides on which assays for hundreds of thousands of known *single nucleotide polymorphisms* (SNPs) can be attached: a SNP-chip. Once DNA from the studied person is added, each individual assay detects the presence of one of the known alleles for the polymorphism of interest, flagging this with a fluorescent label, such that two assays are needed per polymorphism. After computer processing of the signals, the genotypes of the person whose DNA was added, at each of those hundred thousand or more polymorphisms, are known (Figure 17.4, Table 17.1). Nowadays there are several companies (such as Affymetrix and Illumina)

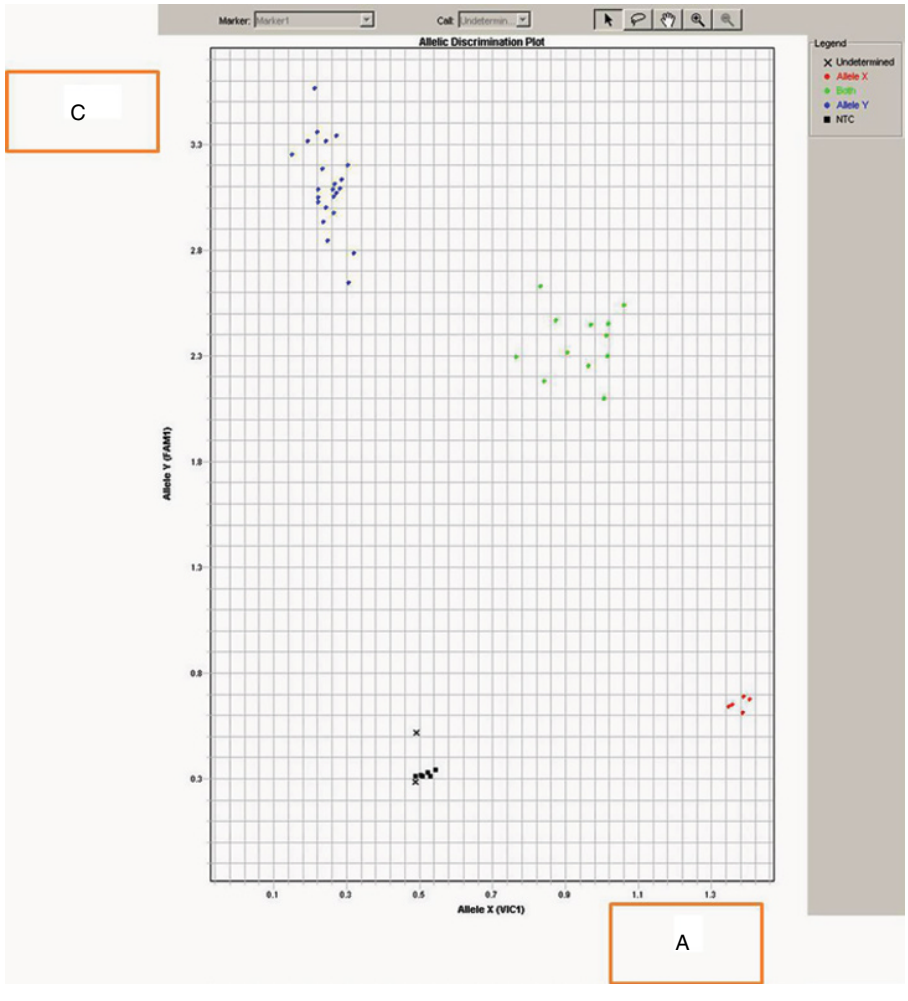

Figure 17.4 Visualization of SNP-chip results.
SNP-chip assay result for a single polymorphism. Each dot represents an individual. X-axis: intensity of the fluorescent label attached to one allele at the polymorphism (e.g., A). Y-axis: intensity of the fluorescent label attached to the other allele (e.g., C). The software recognizes three clusters and assigns a genotype to each individual (e.g., A/A (red), A/C (green), or C/C (blue)). Black dots show samples without signal: controls that contained water instead of DNA. (*See insert for color representation of the figure.*)

Table 17.1 Example of genotyping chip results for four individuals and five polymorphisms.

Polymorphism	*Chromosome*	*Position*	*Ind 1*	*Ind 2*	*Ind 3*	*Ind 4*
rs6051856	20	41499	A/A	A/A	A/G	A/G
rs6038013	20	56187	A/A	A/A	A/G	A/G
rs5038037	20	57272	G/G	G/G	C/G	C/G
rs2298108	20	82476	C/C	C/C	C/T	C/T
rs2298109	20	86125	G/T	T/T	G/T	G/T

that produce standardized commercial SNP-chips allowing genome-wide genotyping at low cost (e.g., $100-$200 per sample).

As can be seen in Table 17.1, each individual has two letters per polymorphism: one inherited from each parent. At this stage of genotyping, it is not known which letter came from which parent, so they are often presented in alphabetical order (C/T, A/G, etc.). Many labs carry out such experiments themselves, but they are also offered as a commercial service.

SNP-chips are a quick and easy way to genotype very large numbers of polymorphisms in one experiment for large cohorts of people. However for other experiments, assays similar to those attached to the SNP-chips are available individually to genotype just a single polymorphism (or perhaps a handful) in a few dozen to a few thousand people. A number of companies sell such assays as kits, to be carried out in your own laboratory, and there are also some that will provide genotyping as a service.

Collecting Phenotypes and Defining Cohorts

The methods described in this chapter essentially consist of uncovering correlations between variability at the level of genotype and that at the level of the trait or trait combination (*phenotype)* in a cohort of interest. As outlined above, there are well-standardized techniques available for obtaining reliable information about the DNA-letters from any study cohort. For a successful outcome it is just as critical to obtain a robust characterization of the traits and characteristics of the study participants. Language-related skills offer considerable challenges when it comes to this side of things.

One strategy that has proved valuable has been to target developmental disorders in which there are unexplained problems with speech, language, or reading occurring against a background of normal intelligence and sensory acuity, along with adequate exposure to spoken/written language in the environment (Bishop, 2001; Fisher & DeFries, 2002; Fisher, Lai, & Monaco, 2003). Research in this area may use performance on a number of different tests, along with clinical reports and case history where available (e.g., from speech/language therapists) to make a formal diagnosis in the individuals taking part in the study. Participants are designated as either being affected or unaffected with the disorder of interest and then the geneticist searches for correlations between genotypic data and this dichotomous affection status in the study cohort. Examples where a "qualitative" approach to trait definition has been particularly effective include studies of *childhood apraxia of speech* (CAS, also known as developmental verbal dyspraxia), a rare disorder where individuals have problems in mastering the rapid coordinated sequences of orofacial movements that underlie fluent speech, leading to inconsistent errors that worsen as the complexity and length of the utterance increases (see later section on Exemplary Studies). Similarly, several investigations of developmental dyslexia (specific reading/spelling disability) have employed qualitative definitions of the disorder to pinpoint suspected candidate genes, such as *DYX1C1* and *ROBO1* (reviewed by Carrion-Castillo *et al*., 2013).

Qualitative all-or-none approaches to defining language-related disorders have certain limitations (discussed in more detail by Fisher & DeFries, 2002; Fisher

et al., 2003). A positive diagnosis might be made based on a child scoring significantly below the level expected for his/her age on one or more measures of language or reading performance. The tasks used to assess performance usually show continuous variation in the general population and the exact choice of threshold can be somewhat arbitrary. Sometimes, formal definitions of these disorders also require a discrepancy between language/reading and general cognition, as assessed by tests of non-verbal IQ, and again the most appropriate degree of discrepancy to apply remains a matter of debate (Fisher & DeFries, 2002). Another difficulty concerns the fact that different types of language-related disorders can co-occur in the same individual, which could reflect biological pathways that impact on multiple skills simultaneously. Traditional diagnostic schemes depend on exclusionary criteria and do not deal well with instances of comorbidity. For instance, the term *specific language impairment* is used to describe a child who has problems with receptive and/or expressive language, without any deficits in speech motor functions, leading to the misleading conclusion that no child could have both *specific language impairment* and *childhood apraxia of speech* together. Moreover, speech, language, and reading skills are developmental traits, such that a child's diagnosis might change at different ages (even though their genetic material remains the same). For instance, a child diagnosed with *specific language impairment* before reading instruction has started, may eventually acquire adequate language skills, but be considered dyslexic later when she or he has problems with learning to read. Overall, a single qualitative diagnosis of a language-related disorder may potentially encompass a heterogeneous mixture of different causes, which could impede the discovery of reliable correlations between traits or disorders and genotypes.

Therefore, an alternative way of studying genetics of language-related disorders is to move away from categorical diagnoses and directly employ the quantitative scores from relevant measures for the genetic analyses. Moreover, such an approach allows researchers to investigate different aspects of our speech, language, and reading faculties, using tasks that are hypothesized to tap into distinct components. Commonly studied traits include the ability to identify and manipulate the sounds in spoken words (*phoneme awareness*), the ability to retain new phonological information without rehearsal (*phonological short-term memory*), the understanding of rules marking tense, number, gender (*grammatical morphology*), the ability to recognize written word forms (*orthographic processing*), and the rapid naming of highly familiar visual symbols (*rapid automatized naming*) (reviewed by Carrion-Castillo *et al.*, 2013; Fisher & DeFries, 2002; Fisher *et al.*, 2003). By focusing on quantitative traits it becomes possible to not only study the molecular basis of disorders (the extremes), but to also investigate the genetic underpinnings of normal variation in language skills in the general population (e.g., Gialluisi *et al.*, 2014; Luciano *et al.*, 2013), which is likely to be highly multifactorial. We later discuss an illustration of this approach under Exemplary Studies.

Another key factor in study design concerns the types of cohorts that are collected. For a monogenic trait, such as a rare language-related disorder, the usual approach is to try to identify multigenerational families in which there are multiple affected individuals, showing an apparently simple inheritance pattern. As discussed in the section "Analyzing the Data," the structure and size of the family, particularly the numbers of affected people in the different generations, is important. Not only does this

indicate whether there is likely to be a monogenic explanation, but it also gives an idea of how much statistical power there is to track down the responsible genetic alteration. In addition, a robust assessment of affection status is crucial, because misdiagnosis of an individual could derail attempts to uncover the relevant gene.

For studies that focus on multifactorial traits, and hence assume the involvement of common DNA-variants with small effect sizes, it is typically necessary to collect large cohorts of thousands of people. As mentioned above, if the trait of interest shows continuous variation in the general population and can be indexed by a reliable quantitative measure, it can be studied in a cohort of people collected randomly from the general population. Examples include birth cohorts like the Avon Longitudinal Study of Parents and Children (ALSPAC) in the UK and Generation Rotterdam (GenR) in the Netherlands.

Analyzing the Data

Once the genetic data (genotypes or sequences) have been collected, the information needs to be analyzed to find out what DNA alterations may be involved in the traits under study. It is necessary to establish whether there is a statistically significant relationship between a variant and a trait. Here we briefly describe the approaches typically used for studying monogenic and multifactorial traits.

Monogenic Traits—Linkage in Large Families

Suppose we have a suspected monogenic disorder clustering in a family, such that each affected person has apparently inherited the disorder from one of his/her parents. We would like to find out if a single genetic variant can explain the disorder in this family. The traditional method, which is still quite often used, involves two steps. In the first step, we search for any sections of the genome in which DNA variants are shared (i.e., the same letters at a polymorphism) between the affected (but not the unaffected) relatives in the family; that is, we want to identify genomic regions that are *linked* to the disorder. At present, this step can be run in a cost-effective way by genotyping DNA variants across the genome in all available family members using SNP-chips (DNA arrays for genotyping, as described above). Since the shared stretches of DNA that we are looking for are expected to be rather large, genotype data from ~10,000 well-distributed common polymorphisms will suffice (Figure 17.1). By using software that systematically considers the inheritance pattern of each polymorphism, we can detect any parts of the genome that show significant linkage to (i.e., are inherited together with) the disorder. Rigorous statistical methods are carried out to determine whether a linkage that we observe is a significant finding. If there is significant linkage, this means the linked polymorphism and the DNA variant that causes the trait are very likely to be located relatively close to each other on the DNA molecule. The section of DNA around this polymorphism is now the place to look for the causal DNA variation. Investigating a number of adjacent polymorphisms and checking their linkage to the trait, will tell us something about the size of the section we need to investigate.

On finding a section of the genome that shows linkage, the assumption is that somewhere within this section there lies a rare variant (perhaps even unique to that family) that is responsible for causing the disorder. The aim of the second step is to then identify that causative variant, usually by reading (sequencing) the DNA of all the genes from the linked DNA-section in the family. Since the regions implicated by the first stage often contain tens to hundreds of different genes, this second step can be very time-consuming, unless an obvious candidate gene (such as one previously implicated in a related disorder) resides in the region. Most monogenic disorders are caused by DNA variants that change proteins, so the search would usually focus primarily on sequencing the protein-coding parts of the linked interval. Sometimes, tracking down the responsible gene can be aided by finding other families who show linkage of a similar disorder to the same region, or independent cases of people in which the region is disrupted by a large-scale *chromosomal rearrangement*.

These days, as an alternative to the above two-step search, we might instead take advantage of new possibilities offered by whole genome/exome sequencing. Ideally, we would like to have sequence data for all members of a family of interest, but because next-generation sequencing (NGS) is still quite expensive, we may only be able to afford this for two or three relatives. In that case, a popular approach is to select two affected people from the family who are not too closely related, such as two cousins. If we have more money, we might choose an unaffected brother or sister of one of them as a control. In these three people, we can sequence all the protein-coding parts of the DNA or even the entire genome (WES or WGS, as described earlier). We can then look for any protein-changing variants that are shared by the two affected people, but are absent in the unaffected sibling control. If the variant is causing a rare and easily recognizable disorder it is unlikely that the causative DNA variant is present in healthy individuals. On the internet, databases are available with sequence data for thousands of apparently healthy individuals. We can discard all variants that are seen in these databases. Usually, these steps leave only a handful of candidate variants. Using Sanger sequencing (see above), we can then inspect these variants in the whole family. The remaining suspects are those variants that are seen in all affected relatives but in none of the unaffected family members, that alter protein sequence in a way that is predicted to alter protein function in a substantive way, and that have never been found in healthy individuals in any other studies. If using a WGS/WES strategy, we can statistically test for linkage at the end of the search, rather than using linkage mapping as a starting point, by performing statistical analyses for just the candidate causal DNA-variants. Even when there is convincing statistical evidence that the causal variant has been found, additional investigations are needed to increase the confidence in this result, such as identification of other causative variants in the same gene in unrelated families/cases or proof of functional effects from, for example, studying the result of manipulating the genes in cultured cells or animal models.

Described above is the ideal situation. In real life, some data from sequencing or SNP-chips may be of low quality, one or more family members may have died or is unwilling to cooperate, the disorder may show variability which makes it hard to be certain about its presence or absence in some people, and so on. For the traditional two-step method to be viable, we need DNA and matching trait data for at least three generations in a family, and the third generation should have at least two people who are affected. Depending on a number of factors, we need about ten to

twelve affected relatives to be able to find a significant result for a *dominant* disorder. A single large family would be perfect for such a study, but a combination of a number of families could be used, assuming that the same gene is disrupted (difficult to establish a priori unless the *phenotype* is particularly distinctive). For the modern WGS/WES-based method different statistics would be possible, in which some of the criteria could be relaxed. You would not necessarily need all family members to be able to trace the inheritance patterns, because a really rare DNA variant is unlikely to occur more than once in a family without being inherited from one family member to the next. So far we discussed so-called *dominant* monogenic inheritance, in which a disorder may result from a DNA-change in one of the two copies you have of each gene. Some disorders only occur when both your maternal and paternal copy of the relevant gene are disrupted. These disorders we call *recessive*. For disorders that show *recessive* inheritance, similar methods can be used with certain adaptations, which we will not go into here.

Multifactorial Traits—Identifying Common Effects with GWAS

When a trait is suspected to have a multifactorial genetic architecture, involving combined effects of a number of different common polymorphisms each with a small effect size, a typical study would collect a large cohort of individuals and test for *association* between gene variants and the trait (e.g., a score on a test or diagnosis of a disorder). If we know of a gene that we already think is very likely to be involved in the trait (a *candidate gene*), we can select polymorphisms in and around that gene to focus on, as described below. However, for traits where we know little about the biology, as in the case of language-related phenotypes, it is difficult to pick out appropriate candidate genes and come up with reasonable hypotheses to test. Technological developments have allowed geneticists to overcome this issue by carrying out a hypothesis-free (with respect to gene choice) search in which polymorphisms across the whole genome are systematically interrogated for association.

While for a monogenic disorder a single causal variant is often sufficient to fully account for the risk of developing the disorder in a family, for multifactorial traits a risk variant might increase a person's chance of having a disorder by <1%. To have enough statistical power to detect the subtle roles of such variants it is necessary to collect DNA and matching phenotypic trait information from large cohorts. For some traits a cohort with thousands of unrelated people could be enough to support a GWAS, but it is becoming common for studies to analyze cohorts comprising tens of thousands of participants (sometimes only possible through meta-analyses). Using SNP-chips, hundreds of thousands of polymorphisms are genotyped in each individual of the cohort. The statistical analyses involved in testing for association are conceptually simple. If we are studying a dichotomous trait, in which we can divide participants into two groups (e.g., cases versus controls) we can use for example a chi-square test to test for each polymorphism whether one of the alleles has a significantly higher frequency in one group than in the other. If we are studying a quantitative trait, we can use an approach like linear regression to test for each polymorphism whether there is a relationship between the alleles that they carry and the trait. For example, for a C/T polymorphism we can ask whether the number of C alleles at that polymorphism (0, 1, or 2) is correlated with the quantitative

score. Because DNA variants that lie close together on a chromosome tend to be transmitted together for many generations, they tend to co-occur on the same stretch of DNA even in people who are seemingly unrelated. Therefore we often see evidence of association for a number of neighboring polymorphisms on a chromosome. A single GWAS requires performance of hundreds of thousands of different statistical tests. Under the null hypothesis of no association between a polymorphism and a trait, 5% of those tests will yield a p-value that is below 0.05. So, clearly, the standard threshold is not suitable since it will deliver an unacceptably high number of false-positive findings. The consensus of the field is that in a GWAS an association between a polymorphism and a trait is only considered significant if the p-value is less than 5×10^{-8}, and even then independent replication in another cohort is usually necessary to be convincing.

Since the people being studied have only very distant relatedness with each other, the DNA-sections implicated by this kind of association testing are much smaller than those identified by family linkage analyses (Figure 17.1). This is because many more generations have passed since the patients shared a common ancestor than in a family. A polymorphism that is significantly associated to the trait or disorder in a genome-wide association study (GWAS) might point to a single gene or, at most, to five or six: a region around it of on average ~300,000 letters. Since genes take up only a small fraction of our genome, it may also happen that these significant polymorphisms are in regions with no gene at all in the neighborhood. Despite the relatively small size of the associated sections, it has turned out to be remarkably difficult to identify which variant is truly responsible for the effect on the trait. In most monogenic disorders, the causal variant obviously disturbs the working of an encoded protein. In multifactorial traits, it is more likely that the relevant variant alters the levels or timing of production of an encoded protein in a subtle manner. We are not so skilled yet at recognizing and characterizing such variants, although there are many genomic initiatives underway that seek to improve the situation. Therefore, GWAS studies usually end with identifying the genes that are likely involved in the disease or trait of interest, without necessarily being able to zero in on the exact variants or mechanisms responsible. Because each sufficiently powerful study will identify multiple genes, subsequent analysis can assess whether these genes act together in a shared biological process, are translated into protein in similar tissues, and so on. Even when a GWAS does not identify individual polymorphisms that meet criteria for genome-wide significance, it can provide useful information on whether association signals are enriched for certain types of genes or biological processes. It has also become popular to study whether variants that were implicated for one trait are also seen for another related trait, so we gain more understanding about the similarities and differences between them (Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013).

Exemplary Studies

To give concrete illustrations of the methods we have introduced, we discuss two exemplary language-related studies from the literature, one concerning a monogenic disorder, the other investigating a multifactorial quantitative trait.

Linkage Analysis Implicates FOXP2 in Speech and Language Deficits

In 1998, a linkage analysis was reported for a three-generation family (known in the literature as the KE family) in which a rare severe speech and language disorder was transmitted in a pattern that could be easily recognized as dominant inheritance (Fisher, Vargha-Khadem, Watkins, Monaco, & Pembrey, 1998). The disorder, which affected fifteen people (about half of the family members) involved *childhood apraxia of speech*, accompanied by wide-ranging impairments in spoken and written language skills, while other aspects of cognition were less affected (http://www.omim.org/entry/602081). The researchers used the traditional design of first genotyping polymorphisms to detect stretches of DNA shared by the affected people. A region on chromosome 7 was found to show highly significant linkage to the disorder. With our current knowledge of the genome, we can see that the size of the linked region covers ~13×10^6 DNA letters and contains ~40 protein-coding genes. At that time, however, the first entire human genome sequence was not yet determined, and there was very limited knowledge about which genes lay in the interval of interest, and what their sequences were.

The researchers pieced together as much information as they could from fragments of data that were known at the time, and began sifting through the available genes using Sanger sequencing to search for causative variants in the affected KE members (Lai *et al*., 2000). Fortunately, they came across a child who was not related to the KE-family and who had a very similar type of speech and language disorder. He had a chromosomal rearrangement disturbing the same genomic region that had been implicated by the KE family linkage analysis. In this rearrangement (a *translocation*) part of chromosome 7 had been swapped with part of chromosome 5, without any loss of genetic material. (Note: the rearrangement affected one of the two copies of each chromosome; the other copies of chromosome 7 and 5 were normal.) The child's parents did not have any speech/language problems and they did not have this genomic rearrangement: It had happened during the formation of the egg or sperm. When pieces of chromosome get swapped in this way, they must have been broken somewhere. If such a breakpoint passes directly through a gene, this gene's function is disturbed. Because a similar section of chromosome 7 was implicated in both the KE family and the unrelated child, it was possible that their language problems were caused by malfunctioning of the same gene. Knowledge and technology was much less advanced than nowadays, so a complicated set of experiments was necessary to find the gene that was broken in the unrelated child and to study it subsequently in the KE family. In the end, a clearly causative DNA variant was found in the affected members of the family, disrupting a novel gene that is now called *FOXP2* (Lai, Fisher, Hurst, Vargha-Khadem, & Monaco, 2001). Unaffected KE family members all had G/G at this position, whereas all those with the language problems had G/A, that is, they were heterozygous for an unusual A allele. The A results in alteration of the protein encoded by *FOXP2*; at one critical point of that protein the amino-acid arginine is replaced by a different one, histidine. Experiments in cultured cells and animal models have shown that this change prevents the protein from working properly (see Fisher & Scharff, 2009). Subsequent screening studies have since identified different rare disruptive variants of the *FOXP2* gene in other unrelated families and cases (reviewed by Graham & Fisher,

2015). Though disruptions of *FOXP2* are rare, explaining only a small proportion of cases of speech and language disorder, the discovery of the gene led to a highly informative series of investigations concerning its roles in cells, neurons, brains, and behavior, as well as providing novel insights into important evolutionary questions. Such work is outside the scope of the current chapter; the interested reader is referred to, for example, Fisher and Vernes (2015) for extensive descriptions.

GWAS Uncovers Effects of ROBO2 on Early Expressive Vocabulary

The expression and understanding of language in children shows considerable inter-individual variation. Twin studies suggest that this is the result of both environmental and genetic factors (e.g., see Kovas *et al.*, 2005). At young ages, the (as yet undetermined) environmental differences seem to explain the larger part of the variation, but clearly not all of it. St Pourcain and colleagues performed a genome-wide association scan (GWAS) of expressive vocabulary scores in unrelated children of European descent from the general population, analyzing early (15–18 months; "one-word stage") and later (24-30 months; "two-word stage") phases of language acquisition (St Pourcain *et al.*, 2014). The study took advantage of large cohorts from the general population that had been followed longitudinally since birth, and had also been genotyped using genome-wide SNP-chips. The phenotypes (traits/trait combinations) of interest for this study were derived from communicative development inventories—parent report instruments, which capture information about children's developing abilities in multiple domains of early language (see Chapter 3 of the present volume). The GWAS was carried out first in a discovery cohort analyzing trait-polymorphism association for $>2 \times 10^6$ polymorphisms. This involved 6,851 infants of mean age 15 months for the "one-word stage" and 6,299 toddlers of mean age 24 months for the "two-word stage." For the trait measured at the younger age, the top association was seen for a polymorphism near the gene *ROBO2*, with a p-value of 9.5×10^{-7}, while analysis of the trait measured at the older age pointed to a polymorphism within the gene *CAMK4* that gave a p-value of 3.5×10^{-7}. As noted above, the accepted threshold for significance in a GWAS is $p < 5 \times 10^{-8}$, meaning that these polymorphisms were only suggestively associated with the traits being studied.

The researchers went on to evaluate their findings further in three independent cohorts from the UK, Netherlands, and Australia, again making use of available measures and genome-wide genotyping data already collected for these cohorts. In this follow-up, which included an additional 2,038 children for the early stage, and 4,520 children for the late stage, the researchers did not perform a full GWAS, but focused on only the most interesting polymorphisms from their discovery cohort. In the combined data from the discovery and the replication cohort, the polymorphism near *ROBO2* was found to be significantly associated with the trait measured at the younger age ($p = 1.3 \times 10^{-8}$). Around 35% of all the children in the cohort had at least one G allele; the others had only A alleles at this position. Having a G allele decreased expressive vocabulary scores at the "one-word stage" by 0.098 standard deviations, illustrating the very small effects typically found for multifactorial traits. Be reminded that this DNA variant is not necessarily responsible itself for the change in the vocabulary scores, but that it lies on a section of DNA that carries the putative

contributing variant (see Figure 17.1, bottom). Curiously the *ROBO2* findings in this study seemed to be specific to the infant sample—no association was found in toddlers at the later stage, and there was no impact on later outcomes for speech, language, or reading skills. *ROBO2* is a convincing candidate for involvement in language, since the gene is known to be important for brain development (particularly in relation to guidance of axons) and prior studies found association of language/reading-related phenotypes with a very similar gene, known as *ROBO1* (Mascheretti *et al*., 2014). Overall, this study shows that even with access to cohorts totaling more than 10,000 participants, identifying common genetic factors that influence a multifactorial trait can be challenging.

Problems and Pitfalls

Research in genetics, like in any field, can encounter difficulties in collecting, analyzing, or interpreting the data. Here we discuss some commonly encountered problems.

No one method is optimal for all research questions. Laboratory-based genetic data can be prone to artefacts. For studies like GWAS, but also linkage, which involve datasets of thousands to millions of data-points, manual inspection of each and every data-point is not an option. Thus, rigorous quality control steps and sanity checks are necessary at every stage of the analyses. Results that generate significant evidence of linkage or association should at least be checked by visualization of the overall data patterns (Figures 17.2-17.4), and/or by testing them again with a second independent technique. However, this cannot guard against false negative results. Note also that the statistical analyses used in gene mapping can only tell us about probabilities in relation to the null-hypothesis of a chance finding, rather than giving absolute proof of the involvement of a variant in the trait under study. The steps we describe in this chapter should be seen as starting points for generating new hypotheses and novel questions, leading to experiments that can further evaluate the contributions of specific genes and genetic variants to language phenotypes.

Studies of monogenic disorders depend on tracking down appropriate families in which developmental language deficits affect large numbers of relatives and are inherited in a simple manner. Suitable families tend to be rare and difficult to find, so there is some serendipity involved. Even the most carefully carried out linkage screens may fail to find any significant results, either because the family is too small to yield adequate power, because the underlying genetic architecture is more complex (i.e., not actually monogenic after all), or due to misdiagnosis of some key family member(s). Even when significant linkage has been found in a family, it might not lead to successful identification of a causal variant, despite extensive searching. Often, the identification of independent families/cases implicating the same gene is crucial for pinpointing causal variants, as already shown by our description of the discovery of *FOXP2*. Moreover, the value of experimental evidence in cell-cultures, animal models, and with other approaches supporting a functional effect cannot be overestimated.

We discussed how Next Generation Sequencing (NGS) offers an alternative to traditional methods in family studies. The advantage of going immediately to whole genome or exome sequencing is that it is more direct and may lead to faster answers,

quickly highlighting potential candidate causal variants. Also, missing a few individuals in the family tree is probably less problematic than for traditional linkage. On the other hand, WGS/WES approaches may miss out on a true causal variant that does not alter a protein, but is instead in some regulatory region of the DNA. Large deletions (i.e., missing stretches of DNA) are hard to detect in NGS results. Notably, the traditional method can pinpoint a linked region of the genome regardless of the type of variant that is causing the disorder.

When it comes to genome-wide association studies (GWAS) and multifactorial traits, one of the main limitations is that the effects of variants in multifactorial traits are so small that you may need very large numbers of participants to ensure adequate statistical power (typically 10,000->100,000 people). This in turn might necessitate the establishment of multi-center consortia, involving collaborations between multiple different groups or even countries. Clearly, consensus must be reached on how the trait is measured, and care must be taken that all centers use the same definitions and inclusion criteria. For the language sciences there is the added complication that data may have to be pooled across diverse languages with distinct properties. Because of the large cohort sizes involved, traits that can be measured reliably without spending too much extra time and money per participant are most suitable for GWAS studies. In the coming years, the field could be transformed by the development of suitable web- or app-based batteries of tests for reliably capturing inter-individual variation in language skills. Thus existing study cohorts from the general population who have already been genotyped with genome-wide chips could be targeted for "phenotyping from a distance," making very large-scale GWAS studies of language traits feasible.

If the GWAS study design involves comparing cases of language disorder with healthy controls, the collection of controls may require extra care. As in all such studies, to adequately compare groups of individuals, it is necessary to match age, gender, and so on. It is in this case also important that the case and control groups are genetically matched, since different ethnic groups sometimes have different allele frequencies for a subset of polymorphisms. Also when we are studying a quantitative trait, the whole group must be as genetically homogeneous as possible; mixing people from different ethnicities will invalidate the standard study design.

An alternative to GWAS that can be run with smaller cohorts is to focus on testing fewer polymorphisms. We could choose only a subset of genes that we are particularly interested in, and test only the polymorphisms that are in and around these genes. This would keep the statistical problem of multiple testing within bounds.

Finally, we note that since systematic genome-wide screens avoid choosing candidate genes, they may seem less elegant than formulating a prior hypothesis based on available biological knowledge. However, for many genes we still know little about their precise functions, and the underlying biology of language-related skills remains very poorly understood. Time and again in human biology, it has turned out that our original ideas about mechanistic underpinnings of a trait or disorder were off the mark, and genetic findings have radically changed our view of these underpinnings, providing important new entry points into the key processes. The rapid major advances in molecular technologies of recent years make it possible to apply systematic screening approaches to more and more questions, even for unraveling the ultimate mysteries of our unique capacities for speech and language.

Key Terms

Allele Alternative DNA-form at a particular position in the genome. For example, a polymorphism may consist of the alleles C and G.

Amino acid Organic compounds that make up proteins. In human biology, 20 different amino acids are used to build proteins.

Chromosome A structure found in living cells, consisting of a single molecule of DNA— encodes genes and much more. Humans have 23 pairs of them.

Crossing-over The exchange of stretches of DNA between two chromosomes from the same pair during egg/sperm formation.

Dominant Inheritance pattern in which alteration of one copy of a gene is enough to change the trait (see *recessive*).

Exome Subset of the genome, encompassing all DNA that codes for protein. In total ~1% of the human genome.

Gene Segment of DNA that codes for a particular protein.

Genome All the genetic material contained in your 23 pairs of chromosomes, including a total of more than 20,000 protein-coding genes.

Genotype Combination of two DNA letters at a particular position in the genome (exceptions on the sex chromosomes). Occasionally, also used to denote all DNA-variants in the whole genome.

GWAS Genome-wide association scan (for example case-control study).

Linkage analysis Family based analysis to identify a DNA-variant whose inheritance pattern matches that of the trait of interest.

Molecular Genetics A subdivision of genetics research concerned with the structure and function of genes at the molecular (i.e., DNA/RNA) level.

Monogenic A disorder or a major trait difference that is caused by disruption of a single gene.

Multifactorial Variations in many genes (plus environmental factors) affect a trait. Both quantitative traits (e.g., blood pressure) and dichotomous traits (e.g., rheumatoid arthritis, presence/absence of a disease) can have such a genetic background.

Next-generation sequencing (NGS) Also known as massive parallel sequencing. A group of relatively new methods (developed in the late 1990s) that enable sequencing of large amounts of DNA per person.

Phenotype An individual's trait or combination of traits, for example, eye color or height. Sometimes used to refer to only the trait(s) under study, sometimes for all of an individual's traits.

Polymorphism Position in the genome where a significant proportion of people in a population carry different DNA letters.

Protein Large molecules composed of one or more chains of *amino acids* in a specific order determined by the base sequence of nucleotides in the DNA coding for the protein.

Recessive Inheritance pattern in which DNA alterations in both the paternal and the maternal copies of the same gene are needed to change the trait (see *dominant*).

Translocation A chromosome abnormality caused by rearrangement of parts of chromosomes, for example, a piece of one chromosome breaks off and gets attached to a different chromosome.

Whole-exome sequencing (WES) Approach to perform NGS only on protein-coding parts of the genome.

Whole-genome sequencing (WGS) Approach to perform NGS on the whole genome (~3 billion DNA letters per person).

References

Bishop, D. V. M. (2001). Genetic and environmental risks for specific language impairment in children. *Philosophical Transactions of the Royal Society B Biological Sciences*, *356*, 369–380.

Bufe, B., Breslin, P. A., Kuhn, C., Reed, D. R., Tharp, C. D., Slack, J. P., … Meyerhof, W. (2005). The molecular basis of individual differences in phenylthiocarbamide and propylthiouracil bitterness perception. *Current Biology*, *15*, 322–327.

Carrion-Castillo, A., Franke, B., & Fisher, S. E. (2013). Molecular genetics of dyslexia: An overview. *Dyslexia*, *19*, 214–240. doi: 10.1002/dys.1464.

Cross-Disorder Group of the Psychiatric Genomics Consortium. (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genetics*, *45*, 984–994.

Fisher, S. E., & DeFries, J. C. (2002). Developmental dyslexia: Genetic dissection of a complex cognitive trait. *Nature Reviews Neuroscience*, *3*, 767–780. doi: 10.1038/nrn936.

Fisher, S. E., Lai, C. S., & Monaco, A. P. (2003). Deciphering the genetic basis of speech and language disorders. *Annual Review of Neuroscience*, *26*, 57–80. doi: 10.1146/annurev.neuro.26.041002.131144

Fisher, S. E., & Scharff, C. (2009). FOXP2 as a molecular window into speech and language. *Trends in Genetics*, *25*, 166–177. doi: 10.1016/j.tig.2009.03.002.

Fisher, S. E., Vargha-Khadem, F., Watkins, K. E., Monaco, A. P., & Pembrey, M. E. (1998). Localisation of a gene implicated in a severe speech and language disorder. *Nature Genetics*, *18*, 168 -170. doi: 10.1038/ng0298-168.

Fisher, S. E., & Vernes, S. C. (2015). Genetics and the Language Sciences. *Annual Review of Linguistics*, *1*, 289–310. doi: 10.1146/annurev-linguist-030514-125024.

Gialluisi, A., Newbury, D. F., Wilcutt, E. G., Olson, R. K., DeFries, J. C., Brandler, W. M., … Fisher, S. E. (2014). Genome-wide screening for DNA variants associated with reading and language traits. *Genes, Brain and Behavior*, *13*, 686–701. doi: 10.1111/gbb.12158.

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*, 333–351.

Graham, S. A., & Fisher, S. E. (2015). Understanding language from a genomic perspective. *Annual Review of Genetics*, *49*, 131–160. doi: 10.1146/annurev-genet-120213-092236.

Kovas, Y., Hayiou-Thomas, M. E., Oliver, B., Dale, P. S., Bishop, D. V., & Plomin, R. (2005). Genetic influences in different aspects of language development: The etiology of language skills in 4.5-year-old twins. *Child Development*, *76*, 632–651.

Lai, C. S. L., Fisher, S. E., Hurst, J. A., Levy, E. R., Hodgson, S., Fox, M., … Monaco, A. P. (2000). The SPCH1 region on human 7q31: Genomic characterization of the critical interval and localization of translocations associated with speech and language disorder. *American Journal of Human Genetics*, *67*, 357–368. doi: 10.1086/303011.

Lai, C. S. L., Fisher, S. E., Hurst, J. A., Vargha-Khadem, F., & Monaco, A. P. (2001). A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature*, *413*, 519–523. doi: 10.1038/35097076.

Luciano, M., Evans, D. M., Hansell, N. K., Medland, S. E., Montgomery, G. W., Martin, N. G., … Bates, T. C. (2013). A genome-wide association study for reading and language abilities in two population cohorts. *Genes, Brain and Behavior*, *12*, 645–652. doi: 10.1111/gbb.12053.

Mascheretti, S., Riva, V., Giorda, R., Beri, S., Lanzoni, L. F., Cellino, M. R., & Marino, C. (2014). KIAA0319 and ROBO1: Evidence on association with reading and pleiotropic effects on language and mathematics abilities in developmental dyslexia. *Journal of Human Genetics*, *59*, 189–197.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P., & Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nature Reviews Genetics*, *9*, 356–369. doi: 10.1038/nrg2344.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews Genetics*, *11*, 31–46. doi: 10.1038/nrg2626.

St Pourcain, B., Cents, R. A. M., Whitehouse, A. J. O., Haworth, C. M. A., Davis, O. S. P., O'Reilly, P. F., … Davey Smith, G. (2014). Common variation near ROBO2 is associated with expressive vocabulary in infancy. *Nature Communications*, *5*, 4831, doi: 10.1038/ncomms5831.

The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation, *Nature*, *526*, 68–74. doi: 10.1038/nature15393.

Veltman, J. A., & Brunner, H. G. (2012). *De novo* mutations in human genetic disease. *Nature Reviews Genetics*, *13(8)*, 565–575. doi: 10.1038/nrg3241.

Further Reading and Resources

Fisher, S. E. (2016). A molecular genetic perspective on speech and language. In G. Hickok, & S. Small (Eds.), *Neurobiology of Language* (pp. 13–24). Amsterdam: Elsevier. doi: 10.1016/B978-0-12-407794-2.00002-X.

Fisher, S. E., & Vernes, S. C. (2015). Genetics and the Language Sciences. *Annual Review of Linguistics*, *1*, 289–310. doi: 10.1146/annurev-linguist-030514-125024.

Neale, B. M., Ferreira, M. A. R., & Medland, S. E. (Eds.). (2008). *Statistical genetics: Gene mapping through linkage and association*. Abingdon: Taylor & Francis.

Strachan, T. & Read A. (2010). *Human Molecular Genetics* (4th ed.). New York, NY: Garland Science.