MBE Advance Access published January 15, 2015

Article
Discoveries

The evolution and functional impact of human deletion variants shared with archaic hominin genomes

Yen-Lung Lin[1], Pavlos Pavlidis[2], Emre Karakoc[3], Jerry Ajay[4], Omer Gokcumen[1]

[1] Department of Biological Sciences
State University of New York at Buffalo
Buffalo, New York, U.S.A.

[2] Foundation of Research and Technology – Hellas
Institute of Molecular Biology and Biotechnology (IMBB)
Heraklion, Crete, Greece

[3] Department of Evolutionary Genetics
Max Planck Institute for Evolutionary Biology
Plön, Germany

[4] Department of Computer Science and Engineering
State University of New York at Buffalo
Buffalo, New York, U.S.A.

Correspondence:
Omer Gokcumen, Ph.D.
omergokc@buffalo.edu
Tel: 716-645-4937
Department of Biological Sciences
State University of New York at Buffalo
Buffalo, NY 14260-1300, U.S.A.

Abstract

Allele sharing between modern and archaic hominin genomes has been variously interpreted to have originated from ancestral genetic structure or through non-African introgression from archaic hominins. However, evolution of polymorphic human *deletion*s that are shared with archaic hominin genomes have yet to be studied.

We identified 427 polymorphic human deletions that are shared with archaic hominin genomes, ~87% of which originated before the Human-Neandertal divergence (*ancient*) and only ~9% of which have been introgressed from Neandertals (*introgressed*). Recurrence, incomplete lineage sorting between human and chimp lineages, and hominid-specific insertions constitute the remaining ~4% of allele sharing between humans and archaic hominins.

We observed that ancient deletions correspond to more than 13% of all common (>5% allele frequency) deletion variation among modern humans. Our analyses indicate that the genomic landscapes of both ancient and introgressed deletion variants were primarily shaped by purifying selection, eliminating large and exonic variants.

We found 17 exonic deletions that are shared with archaic hominin genomes, including those leading to 3 fusion transcripts. The affected genes are involved in metabolism of external and internal compounds, growth and sperm formation, as well as susceptibility to psoriasis and Crohn's disease. Our analyses suggest that these *exonic* deletion variants have evolved through different adaptive forces, including balancing and population specific positive selection.

Our findings reveal that genomic structural variants that are shared between humans and archaic hominin genomes are common among modern humans and can influence biomedically and evolutionarily important phenotypes.

Keywords: Neandertal, Denisovan, copy number variation (CNV), DMBT1, LCE3C, GHR, ACOT1, GSTT1

Introduction

The release of ancient Neandertal and Denisovan genomes allowed us to study the relationship between genomes of ancient hominids and modern humans. Neandertal and Denisovan genomes are more closely related to each other than they are to modern human genomes and they diverged from modern human ancestors approximately 500,000 years ago (Prüfer et al. 2014a). Recent studies have shown that archaic hominins, including but not limited to Neandertals and Denisovans, contributed genetic material to modern humans (Veeramah and Hammer 2014). The origin and impact of these introgressions vary geographically and involve different species (Green et al. 2010; Reich et al. 2010; Hammer et al. 2011; Lazaridis et al. 2013). The exact timing and geographical origin of these introgressions have been the focus of several recent studies (Green et al. 2010; Currat and Excoffier 2011; Wall et al. 2013; Hu et al. 2014). In 2014, two papers documented the genome-wide distribution of Neandertal alleles across modern human genomes (Sankararaman et al. 2014; Vernot and Akey 2014). These studies found that regions in modern human genomes that carry Neandertal introgressed sequences overlap with genes less than expected by chance. This implies that purifying (*i.e.*, negative selection against deleterious phenotypes) removed some Neandertal alleles after the introgression event.

Archaic admixture is not the only source of ancient variation in human genome. Previous studies identified highly divergent haplotypes in the human genome, potentially indicating the presence of ancient structure in Africa that has been maintained since before the expected coalescent date for modern human genetic variation (*e.g.*, Barreiro et al. 2005; Cagliani et al. 2008; Teixeira et al. 2014). One hypothesis for preservation of these haplotypes is that they may have been under polymorphism-conserving balancing selection (*e.g.*, heterozygote adaptive fitness advantage).

Genomic structural variants, *i.e.,* deletions, duplications, inversions, and translocations of genomic segments, have recently been recognized as a major part of human genomic variation (Conrad et al. 2009). It has been an ongoing challenge to discover and genotype genomic structural variants (Alkan et al. 2011). However, in the last 5 years, there has been major progress in discovery and genotyping of deletion polymorphisms (1000 Genomes Project Consortium 2012).

We previously described a common deletion polymorphism in modern humans that is shared with Neandertal and Denisovan genomes (Gokcumen et al. 2013). We reasoned that this deletion has evolved before Human-Neandertal/Denisovan divergence in Africa and has been maintained through balancing selection. Herein, we extend our analyses to the entire genome to

identify deletion variants observed among modern humans that are shared with Neandertal and Denisovan genomes.

Results

Identification of polymorphic human deletions that are shared with archaic hominins

To identify the deletion polymorphisms that are also present in the Neandertal and Denisovan genomes, we started with high-confidence deletion polymorphisms documented by the 1000 Genomes Project Phase 1 data release (1000 Genomes Project Consortium 2012). Briefly, this dataset (referred to as *1KG deletions*) includes 14,422 deletion polymorphisms detected among 1,092 human genomes across 14 populations. The deletions were identified by comparing genome resequencing data to the human reference genome (Hg19) and to each other using multiple discovery tools. Furthermore, the breakpoints of these polymorphisms are well characterized, and an extensive validation effort was made to ensure the accuracy of these deletion polymorphisms (See 1000 Genomes Project Consortium 2012). It is important to note that because of the emphasis on accuracy, most complex regions of the genome (*e.g.*, telomeric regions) that could lead to false-positive genotyping may be under-represented in 1KG deletions. As such, this dataset provides an accurate and straightforward starting point for genotyping polymorphic human deletions in Neandertal and Denisovan genomes.

Recent Neandertal (Prüfer et al. 2014a) and Denisovan (Meyer et al. 2012) sequences provide high-depth coverage (~30X) aligned to the human reference genome, the same assembly against which 1KG deletions were compiled. As such, to detect human deletion variants that are shared with archaic hominins, we simply *genotyped* the 1KG deletions using read-depth data from high-coverage Neandertal and Denisovan genomes. Briefly, we calculated the number of Denisovan and Neandertal reads mapping to a given interval in the human reference genome where a deletion polymorphism was previously detected among modern humans. As expected the the number of reads of these regions correlate well with size for both Neandertal and Denisovan sequences (R^2=0.8582 and R^2=0.8713, respectively). However, there are intervals with obviously less than expected read depth as compared to their size (Figure S1AB). These outliers suggest potential deletions in the available archaic genomes

To rigorously identify these outliers, we assumed that the read-depth/size ratio in Neandertal and Denisovan genomes across these intervals follows a normal distribution (Figure 1AB) with the observed mean and standard deviation. We then identify outliers that do not fit into this distribution ($p<0.01$). Using this conservative estimate, we identified 325 and 227 polymorphic *1KG deletions* that are shared with Neandertal and Denisovan genomes, respectively.

To ensure the accuracy of our genotyping pipeline, we conducted multiple checks. First, to avoid any GC bias that may affect mapping, we investigated the GC content of all the human

deletions and those that we found to be shared with archaic hominins. We found no significant difference (Figure S2). Second, we have manually checked all calls in both the Neandertal and Denisovan genomes using Integrative Genome Browser (Nicol et al. 2009) (*e.g.*, Figure S3A). Third, we were also able to take advantage of recently published exome sequences of 3 Neandertal genomes (including the Altai Neandertal used in this study) to verify the presence of 16 exonic deletions in other Neandertal genomes (Figure S3B). These exome sequences also verified the accuracy of our observation for one human deletion that we found to be deleted in Denisovan, but not in Altai Neandertal genome.

Last but not least, we used the software SPLITREAD (http://splitread.sourceforge.net/) to remap Neandertal and Denisovan reads to junctions of the deletion breakpoints as defined in 1KG deletion dataset (Figure S3C, Table S1). We were able to provide strong split-read support for *all* of the 220 non-recurrent deletions we observed in Denisovan genome (please see below for discussion of the recurrent, ancient and introgressed deletions), with at least 20 reads mapping to the breakpoint junctions. Potentially due to differences in sequence lengths and whole genome amplification artifacts, Neandertal sequences performed worse than Denisovan sequences for this analysis, with overall distribution of number of split-reads are an order of magnitude smaller than those observed for Denisovans. Even then, we were able to show that at least 2 split-reads overlap the breakpoint junctions for ~89% (280 out of 315) of the non-recurrent Neandertal deletions. Note that we found strong split-read support from Denisovan reads for all of the 123 non-recurrent deletions that we found in both Neandertal and Denisovans. Based on these observations, we argue that the reduced split-read support from Neandertals is due to lower power, rather than false positives in our deletion dataset. Regardless, we were able to provide split-read support either from Neandertals or Denisovans for ~95% of the non-recurrent deletions.

To further quantify our accuracy, we applied the same procedure for genotyping deletions on a set of random intervals that match the size distribution of 1KG deletion dataset. Based on this analysis, we found 7 Neandertal and 9 Denisovan deletions, corresponding to a false discovery rate of 0.02 and 0.04 for Neandertal and Denisovan deletions, respectively. Overall, we conservatively estimate that at least 427 (~3%) of all polymorphic human deletions are shared with either Neandertals and Denisovans or both (Figure 1C).

Ancient genetic structure, not introgression, explains the majority of polymorphic deletions among humans that are shared with archaic hominins

We considered several scenarios to explain the origins of polymorphic deletions among modern humans with respect to the Neandertal and Denisovan genomes (Figure 2). For the majority of human deletions, we found no evidence of allele sharing with archaic hominins. We will refer to these as *human specific deletions.* It is important to note that our genotyping strategy in archaic hominin genomes is highly conservative and would not be able to pick up deletions that varied among archaic hominins at low frequencies. As such, the *human specific* deletion dataset may

include several variants that are actually shared with other hominin genomes. First, we considered recurrence of the deletion polymorphisms in humans, Neandertals and/or Denisovans. Under this scenario, we expect that the breakpoints of the deletions differ between species. Indeed, we found evidence for 15 recurrent deletions in our manual inspection for different breakpoints (*e.g.*, Figure S3D), explaining ~3.5% of the deletions shared with archaic human genomes. We will refer to these deletions as *recurrent deletions*. With the high-quality Denisovan split-read support, we found no evidence for similar, but not exact breakpoints that we missed in our manual inspection. It is unlikely, but still possible for the deletions to be recurrent even if they share exact breakpoints. The 1KG deletion dataset was compiled with accuracy as a main priority. As such, evolutionarily complex regions of the genome that show high levels of recurrence (*e.g.*, Gokcumen et al. 2011) may have been underrepresented and further studies may uncover important recurrent deletions in these regions. Therefore, our estimate of 15 recurrent events is a lower bound for the recurrence of deletions among Human/Neandertal lineage.

Second, we considered the possibility that these deletion polymorphisms may actually be hominid-specific sequences (*e.g.*, novel insertions or duplications) that evolved in the modern human lineage and remain polymorphic within the species. Under this scenario, these regions then should be observed as deletions in nonhuman outgroups, including chimpanzees and rhesus macaques. We found only two regions that may fit this pattern, as all other deletion regions that we have investigated had an orthologous sequence in chimpanzee or rhesus macaque reference genomes (Table S1). In addition, we found two deletions, for which rhesus macaque but not the chimpanzee reference genome has an orthologous sequence. The most likely explanation of this is incomplete lineage sorting in human-chimpanzee lineage for these variants (Caswell et al. 2008). Albeit interesting, deletions that are explained by these two scenarios constitute less than 1% of the deletions that are shared with Neandertal and Denisovan genomes and will be referred to as the *other deletions.*

Third, we considered previously reported introgression from Neandertals into ancestors of modern Eurasians as a potential source of the observed allele sharing. Under this scenario, the Neandertals contributed genetic material to ancestors of all non-African populations. One challenge was that we could not merely depend on the frequency distribution of deletions in African and non-Eurasian populations to distinguish an introgression scenario from the ancient structure scenario. The frequency distribution of any genetic variant is highly susceptible to drift and recent migrations. As such, we further evaluated the overlap of deletion polymorphisms with Neandertal introgressed regions that were identified based on single nucleotide variation based haplotype construction (Vernot and Akey 2014). Overall, we expect that most polymorphic deletions that introgressed from Neandertals would (i) have breakpoints precisely shared between the Neandertal genome and modern human deletions (*i.e.*, non-recurrent), (ii) the chimpanzee genome will have the sequence that is deleted in the human genome, (iii) these deletions fall into previously reported regions where introgression was detected (Vernot and Akey 2014), (iv) such deletions do not exist or have lower frequency in Africa as compared to Eurasia, and (iv) if they exist in Africa, the haplotypes that carry them have lower nucleotide diversity than Eurasian haplotypes carrying the deletions. Since, Denisovans contributed

genetic materials mostly to the Melanesian populations (Skoglund and Jakobsson 2011; but see for low level Denisovan ancestry in Asia Huerta-Sánchez et al. 2014) we only assumed Neandertal introgression for the present purpose as 1KG deletions do not include Melanesians. In summary, we found that 38 (~9%) of human deletions that are shared with archaic hominin genomes can be explained by Neandertal introgression. We will refer to these as *introgressed deletions* from now on.

To independently estimate potential miscategorization of low-frequency ancient deletions as introgressed, we used the polymorphic human deletions that are shared with the Denisovan genome, but not with the Neandertal genome. Since 1KG deletions do not include sample from Melanesian or any other South East Asian populations, which were reported to have Denisovan introgression, we expect no Denisovan introgression in the *1KG* deletions. As such, the proportion of polymorphic deletions that are shared with only Denisovans and categorized as "introgressed" with our pipeline will give us an indirect estimate of miscategorization. Based on this, we estimate that only 5 out of 102 deletions to be miscategorized as introgressed with our pipeline.

Fourth, we considered the scenario where the age of polymorphic deletion variants we observed in humans actually predates the Human-Neandertal divergence. It has been argued previously that some of the ancient variation that exists in the Human-Neandertal ancestral population has been maintained in humans. As such, the polymorphic deletions among modern humans that also exist in archaic hominins that cannot be explained by introgression should have originated before the Human-Neandertal divergence and been maintained since then in extant humans. Overall, 370 (~87%) of the polymorphic human deletions that are shared with archaic hominins can be traced back to ancient genetic structure predating Human-Neandertal divergence. We will refer to these as *ancient deletions* from here on.

Our observations, taken as a whole, supports the conclusion that the vast majority of allele sharing between humans and archaic hominins affecting deletion variation is due to ancient genetic structure, rather than introgression. In other words, our findings are consistent with the notion that most deletion polymorphisms shared with archaic genomes evolved prior to the Human-Neandertal divergence and have been maintained ever since.

Deletions that are shared with archaic hominins constitute more than 13% of common deletion variation in humans, but are smaller and rarely overlap with functional regions of the genome.

We found that the allele frequencies of the deletion variants that are shared with archaic hominin genomes are significantly higher than human-specific deletion variants ($p<2.2\times10^{-16}$, Wilcoxon rank test, Figure 3B, Figure S4A). We also found that the deletions that we detected both in Neandertal and Denisovan genomes have significantly higher frequency in humans than those we detected only in one of the archaic hominin genomes ($p<0.001$, Wilcoxon rank test).

While ancient deletion variants corresponds to only 2.5% of *all* polymorphic human deletions, they constitute approximately 13% of all *common* (allele frequency>5%) deletion polymorphisms reported among 1KG deletions.

These observations would imply that the ancestral population that gave rise to the Human, Neandertal and Denisova lineages harbored a considerable number of common deletion variants that have been inherited by all three species, and remain polymorphic in extant humans. To investigate whether these deletions are also polymorphic in Neandertals, we manually checked 17 exonic deletions that humans and archaic hominins share among recently released exome sequencing data for 3 Neandertal genomes, including the Altaian individual that we used in this study (Castellano et al. 2014) (Figure S3B). We verified the 16 regions where we previously observed deletions in the Altai Neandertal whole genome sequence. Furthermore, we observed that these regions are homozygously deleted in the other two Neandertal genomes as well. The likelihood of *not* detecting any within species variation across the 16 deleted loci among 2 additional individuals is infinitesimally small, unless the allele frequencies of these deletions are extremely high (Figure S4B). As such, we conclude that the majority of shared deletions described in our study are indeed fixed and not necessarily polymorphic within Neandertals. This is probably due to high inbreeding reported for Neandertals (Castellano et al. 2014; Prüfer et al. 2014b). Using the same dataset, we found no evidence for a deletion in the three Neandertal exomes for one exonic human deletion where we previously observed a deletion in Denisovan, but not for the Altai Neandertal.

Deletion variants have already been shown to be significantly biased away from exonic sequences, indicating the effect of purifying selection (Conrad et al. 2009). We found that the deletions that are shared with archaic genomes have even less overlap with exonic regions in the genome than other deletion variants ($p=0.0004$, Chi-square test, Figure 3A). Taking the heterogeneous nature of genome into consideration, we have simulated genomic intervals using a genome structure correction (Bickel et al. 2010). Our results showed approximately 3.4 ($p=0.0003$) and 5.3 fold ($p=6.6\times10^{-8}$) depletion of the exonic deletions in ancient and introgressed deletions as compared to random expectation, respectively. We also found that ancient deletions are smaller than human specific deletions ($p<2.2\times10^{-16}$, Wilcoxon rank test, Figure S5).

Together, these observations are consistent with recent studies that highlight purifying selection as a major force in shaping the genomic distribution of Neandertal introgressed single nucleotide variation in the human genome (Sankararaman et al. 2014; Vernot and Akey 2014). Our results furthered these observations for ancient and introgressed genomic structural variants.

The majority of ancient deletion polymorphisms have evolved under neutral conditions

The most likely evolutionary scenario to explain the lack of exonic overlap observed among ancient deletions is that purifying selection eliminated most ancient variation. Then, since before the Human-Neandertal divergence, the deletions that were not affected by the initial filtering through purifying selection evolved largely under neutral conditions. According to this hypothesis, we expect that demographic changes and not selective forces operating on ancient deletions are the main processes that would affect neutrality tests.

Ancient deletions provide an especially interesting case. These variants are by definition old, often older than the variation observed in other parts of the genome. We expect that surrounding haplotypes, depending on the recombination rate in that region, are as old as these ancient deletions. Thus, we expect, under neutrality, to observe higher values for Watterson's θ estimator (θ_W) and for Tajima's θ estimator (measured by the average number of pairwise differences, π) as compared to haplotypes harboring non-ancient deletions. Tajima's D measures the normalized difference between θ_W and π. Deviations from 0 indicate either non-neutral evolution or demographic changes. Recent expansions generate negative values for Tajima's D because most of the polymorphisms are recent and are characterized by low allele frequency (Tajima 1989). Human demographic history is characterized by a recent expansion, thus generating negative values for Tajima's D. However, for old regions surrounding ancient deletions, we expect that Tajima's D values will be shifted towards higher values since a proportion of polymorphisms will be as old as the region itself.

To test these expectations, we calculated basic population statistics for ancient and non-ancient deletions including the 10 kb sequence immediately flanking them. We performed these analyses in sub-Saharan African populations, which have higher effective populations sizes than Eurasian populations and consequently less prone to effects of genetic drift. Our results showed that regions harboring ancient deletion variants indeed yield significantly higher θ and π values as compared to regions harboring non-ancient deletions ($p<10^{-4}$ for both measures for both YRI and LWK populations, Student's t test, Figure S6). This observation is consistent with older coalescent times for regions harboring ancient deletions. Our analysis also showed that Tajima's D values for regions harboring ancient deletions do not significantly deviate from zero with means of -0.14 and -0.31 for YRI and LWK, respectively. However, these regions have significantly less negative Tajima's D values, when compared to regions harboring non-ancient deletions ($p<10^{-5}$, one-tail Student's t test for both YRI and LWK, Figure S6). The most plausible explanation for this observation is that regions harboring ancient deletions have been affected to a lesser extent by recent human demography than other genomic regions with a more recent common ancestor. Most of the genomic regions contain polymorphisms affected by the joint effect of bottleneck and recent expansion. These regions in African populations will be characterized by slightly negative Tajima's D as previously shown (*e.g.* Garrigan and Hammer 2006). On the other hand, we observed that polymorphisms on old regions surrounding ancient deletions are characterized by significantly greater values of Tajima's D. These observations are consistent with the scenario that the majority of ancient deletions and their surrounding

haplotypes are indeed older than the genome-wide average and were not subject to major adaptive pressures.

Identifying potentially adaptive ancient deletions among the exonic deletions

As mentioned above, deletions that are shared with archaic hominins are depleted for exonic sequences, indicating the effect of purifying selection. However, we found 17 instances where these deletions overlap with exonic sequences (Table 1, Figure 4ABC). We reasoned that these exonic deletions, which lead to whole gene deletions, fusion transcripts and loss-of-function alleles, are unlikely to evolve under neutrality in contrast to other, non-exonic ancient alleles.

To further investigate the evolution of these exonic deletions, we calculated population differentiation based on the variations in allele frequency within and among populations (F_{ST}) (Hudson et al. 1992) for all 1000 Genomes deletions (Figure 4D). We also used polymorphisms immediately upstream regions of these deletions to calculate Tajima's D (Tajima 1989) (Figure 4E). We then analyzed the human exonic deletions that are shared with Neandertals or Denisovans within the context of all polymorphic human exonic deletions across the genome.

Our analysis highlighted several highly interesting exonic loci (*e.g.*, Figure S7) and below we will discuss several of these genes individually. However, it is important to note that (i) demographic expansions and bottlenecks introduce high levels of noise, increasing the threshold for significance (*e.g.*, false negatives for detecting balancing selection), (ii) recombination that broke the haplotypes between the deletions and flanking regions may have introduced noise into our calculations (*e.g.*, by decreasing otherwise higher Tajima's D), and (iii) that there are multiple (*e.g.*, in contrast to a single external pressure) and complex (*e.g.*, dependent on time, geography, frequency in population, etc.) evolutionary forces that complicate the interpretation of both F_{ST} and Tajima's D.

Ancient and introgressed exonic, loss-of-function deletions are associated with xenobiotic and lipid metabolism, psoriasis, and spermatogenesis

We found that 4 of the non-recurrent exonic deletions that are shared with archaic hominin genomes likely lead to loss-of-function alleles, where either the entire gene or entire coding sequences were deleted (Table 1). One such deletion overlaps with the *LCE3C* gene, which has been strongly associated with psoriasis (de Cid et al. 2009). The allele frequency of *LCE3C* gene deletion is extremely high among Eurasians, reaching to over 70% allele frequency in some European and Asian populations (Figure 4ABC). We found consistently positive Tajima's D values across all 8 non-admixed Eurasian populations analyzed as calculated for the variation in the region harboring the deletion (Figure 4E). When compared to other exonic human deletions, the Tajima's D values reach to 95th percentile in 5 populations. In contrast, differentiation as measured by FST between continental populations show relatively low overall differentiation between continental populations as compared to other exonic deletion variants in

humans (Figure 4D). Positive Tajima's D values and low population differentiation are hallmarks of classical balancing selection (*e.g.*, Cagliani et al. 2008). This observation, in parallel with our understanding that this deletion has been maintained in high allele frequencies since before Human-Neandertal divergence, is consistent with balancing selection acting on the LCE3C deletion variant.

The UGT2B genes comprises of evolutionary dynamic genes that are involved in metabolism of external and internal compounds, including several hormones and steroids. Adaptive deletion variants have already been reported for some members of this family, including deletion of *UGT2B17* gene (Xue et al. 2008). Moreover, UGT2B17 and UGT2B28, both of which are involved in steroid metabolism, have been found to be commonly deleted and the functional impact of these deletions may have cumulative effects (Ménard et al. 2009). Indeed, we found that the deletion that encompasses *UGT2B28* deletion is to be ancient. This deletion is very common in Africa, reaching to almost 40% allele frequency. Similar to what we observed for *LCE3C*, Tajima's D was consistently higher than genome-wide distribution for other exonic deletion variants (Figure 4E), while F_{ST} was consistently low among populations (Figure 4D). These observations are consistent with balancing selection acting on *UGT2B28*.

Another ancient loss-of-function deletion with very high global frequency overlaps with *ACOT1* gene. This gene is involved in lipid biosynthetic pathways and has been argued to play a role in regulation of milk fat synthesis in mammals (Rudolph et al. 2007), which is critical in neonatal development. Interestingly, this whole-gene deletion is shared with the Neandertal genome. The allele frequency of this deletion shows considerable differences between continents, almost reaching fixation among Asian populations, but remains as the minor allele in other continents (Figure 4BC). Unlike the aforementioned *LCE3C and UGT2B28* gene variation, the haplotypic variation around *ACOT1* is characterized by consistently negative Tajima's D values in human populations when compared to values calculated for other exonic human deletions (Figure 4E). However, as expected from high frequency differences, the F_{ST} is consistently higher between Asian and non-Asian populations. This finding may indicate ongoing positive selection pressure on the *ACOT1* deletion, specifically favoring the deletion allele in Asian populations.

We found that only one of the loss-of-function deletions was introgressed from Neandertals. This deletion includes all of the coding sequences of the spermatogenesis-associated gene *SPATA45*. Several evolutionarily important phenotypic trends, such as reproductive efficacy and success, responses to sexual selection pressures or apoptotic pathways that regulate sperm selection, are linked to spermatogenesis. As such, the loss of function of *SPATA45* due to the introgressed deletion mentioned above is a prime candidate for adaptive forces to acting after introgression from Neandertals. Indeed, Vernot and Akey (2014) described a Neandertal-derived haplotype for the functionally similar gene, *SPATA18*. Unlike the *SPATA18* haplotype however, the deletion variant affecting *SPATA45* is relatively rare, found in less than 5% of human genomes.

Incomplete gene deletions lead to novel transcripts, including gene fusions

Nine of the 17 exonic deletions shared with archaic hominins overlap with parts of genes. These deletions potentially lead to alternative protein products, rather than deleting the entire coding sequences. One such deletion overlaps with the exon 3 of growth hormone receptor gene (*GHR*). This well-studied deletion variant leads to a transcript that misses exon 3 (d3). The d3 haplotype was associated with smaller birth size (Sørensen et al. 2010; Padidela et al. 2012). The d3 haplotype was also linked to a 1.7 - 2 times increase in growth acceleration in children that are treated with growth hormone and, consequently is a major target for pharmacogenomics research (Dos Santos et al. 2004). The allele frequency of this deletion is ~44% in Africa, ~31% in Europe, but only 17% in Asia. Not surprisingly, the F_{ST} between Asian and non-Asian populations are consistently higher than genome-wide average (Figure 4D), potentially indicating geography dependent positive selection similar to *ACOT1.*

We identified only one recurrent exonic deletion that is shared with archaic hominins overlapping with *DMBT1*. Unlike the other exonic deletions which are shared among hominin genomes because of common descent or introgression, *DMBT1* deletion sharing is most likely due to recurrent deletions in this region. *DMBT1* also got somatically deleted in malignant brain tumors and contribute to cancer progression (Mollenhauer et al. 1997). Indeed, the region harboring *DMBT1* was discussed within the context of genomic instability (Mollenhauer et al. 1999). Based on our results, we can conclude that it is likely that the genomic instability of this genetic region along chromosome 10q has likely been maintained since Human-Neandertal divergence. Moreover, the deletion variant has been associated with Crohn's disease (Renner et al. 2007). We found that the Tajima's D calculated based on the haplotypic variation upstream of the *DMBT1* deletion is highly negative and in the 5th percentile for two European populations when compared to Tajima's D values calculated similarly for all human exonic deletions in these populations (Figure 4E). Moreover, a recent comprehensive analysis for balancing selection among human genomes identified *DMBT1* as one of the top candidates for balancing selection in both CEU and YRI populations for which the analysis was conducted (DeGiorgio et al. 2014). The low Tajima's D values and the balancing selection reported recently seem to be in conflict. However, as mentioned before, it is plausible that complex mutational and adaptive mechanisms may have shaped the haplotypes carrying some of the deletion variants, leaving complicated signatures of adaptation. It is safe to argue, based on our results and those of previous publications that *DMBT1* deletion variation likely evolved under non-neutral pressures.

One unexpected observation was that three of the ancient deletions led to fusion transcripts, whereby coding sequences of two separate genes are fused. These deletions combine the transcripts of *SNORD115-12* and *SNORD115-13;* as well as *CYP2A6* and *CYP2A7* genes. Another such deletion variant, which is very common (>35%) in all human populations, fuses *GSTT1* and *GSTTP1* genes. *GSTT1* is also involved in metabolizing external compounds. The haplotype surrounding this gene shows one of the highest Tajima's D values measured for exonic deletions in humans (Figure 4E). In addition, there is relatively high population differentiation as measured by F_{ST} between continental populations, especially between African and Eurasian populations (Figure 4D). This observation may be explained by a scenario similar

to that is often put forward for sickle cell trait in malaria-stricken geographies. In essence, the variation that lead to sickle-cell trait has been maintained in the population through geography-specific balancing selection (reviewed in Dean et al. 2002). A similar scenario would explain the extremely high Tajima's D in African populations observed for GSTT1-GSTTP1 fusion deletion, as well as the high F_{ST} values between African and non-African populations observed for this polymorphism.

Conclusion

Deletion variants, the best characterized of all genomic structural variants, have been shown to play an important role in human evolution (McLean et al. 2011). However, the evolutionary role of these variants within species has not been well-established. High-quality sequences and, more importantly, highly improved discovery and genotyping tools primarily developed within the context of the 1000 Genomes project have recently allowed for study of human deletion variation in a population genetics framework. We used these exciting resources to assess deletion variants in humans that are shared with the Neandertal and Denisovan genomes. In so doing, we were able to (i) identify hundreds of ancient and introgressed deletion variants in humans, (ii) investigate deletion variation within their haplotypic backgrounds, and (iii) shed light on the evolution of individual deletion variants that may have phenotypic effects. To the best of our knowledge, this study is the first to document and characterize deletion variation in humans that are shared with Neandertal and Denisovan genomes in a genome-wide context.

Our results suggest that the majority of allele sharing involving deletion variants between modern humans and archaic hominins is due to ancestral structure and, not due to introgression. This observation does not conflict with the recent reports regarding Neandertal and Denisovan introgression to modern humans, but rather highlights a largely unexplored deep ancestry for a considerable portion (~13%) of common deletion variation in humans. Moreover, the genomic distribution of these variants shows signatures of ancient purifying selection, eliminating all but a few exonic variants. This observation complements similar observations made for deletion variants in general (Mills et al. 2011), and further suggests the potential role of recent, rare deletion variants in detrimental phenotypes and disease (Itsara et al. 2009).

Only a very small percentage (~4%) of these maintained ancient and introgressed deletions are exonic, however, the genes involved affect evolutionarily relevant phenotypes, such as growth, immunity and metabolism of external and internal compounds. Some of these deletions were also associated with common human diseases, including Crohn's disease and psoriasis. Exonic deletions are functionally drastic events that are comparable to frameshift, or stop-codon introducing mutations, or if a gene still functions, to multiple non-synonymous single nucleotide variants. As such, we argue that unlike the majority of ancient deletions, those that overlap with exons that have been maintained since Human-Neandertal divergence are unlikely to have evolved under neutral conditions. Instead, these ancient exonic deletions may have been maintained through a combination of (i) geographically different, potentially frequency-

dependent, adaptive forces and (ii) balancing selection. We argue that pathways that involve in these important phenotypes are viable targets for some form of complex adaptive selection that helped maintain the genetic structural variation at these loci for hundreds of thousands of years.

Acknowledgments and funding information

We would like to thank David Radke, Qihui Zhu, Mehmet Somel, and Rebecca Iskow for helpful discussions and insights that lead to conception of this study. We thank Can Alkan for providing helpful advice on SPLITREAD analysis. We also want to thank our colleagues at University at Buffalo, especially to Victor Albert for critical readings of previous versions of this manuscript.

This study is primarily funded by O.G.'s start-up funds from University at Buffalo Research Foundation. P. Pavlidis was funded by the grant FP7 REGPOT-InnovCrete (No. 316223) grant, as well as by the FP7-PEOPLE-2013-IEF EVOGREN (625057).

The authors declare that there is no conflict of interest with organization regarding to the materials discussed in this article.

Tables

Table 1. List of exonic deletions shared with archaic hominin genomes

Chrom	Start	End	Gene	Comment	Ancestral State
chr11	60228164	60229386	*MS4A1*	UTR	Introgressed
chr1	213002368	213013665	*SPATA45*	Loss-of-function	Introgressed
chr13	20077974	20080405	*TPTE2*	UTR	Ancient
chr17	79285360	79286612	*TMEM105*	UTR	Ancient
chr19	41355733	41387636	*CYP2A6, CYP2A7*	Fusion	Ancient
chr10	124369735	124377838	*DMBT1*	Partial-CDS	Recurrent
chr11	3238738	3244086	*MRGPRG*	UTR	Ancient
chr8	144634064	144636239	*GSDMD*	UTR	Ancient
chr15	25436588	25438493	*SNORD115-12, SNORD115-13*	Fusion	Ancient
chr12	27648142	27655163	*SMCO2*	Partial-CDS	Ancient
chr11	128682716	128683410	*FLI1*	UTR	Ancient
chr7	99461389	99463562	*CYP3A43*	UTR	Ancient
chr14	73997051	74024450	*ACOT1*	Loss-of-function	Ancient
chr4	70124301	70230600	*UGT2B28*	Loss-of-function	Ancient
chr1	152555542	152587742	*LCE3C*	Loss-of-function	Ancient
chr5	42628311	42630990	*GHR*	Partial-CDS	Ancient
chr22	24343050	24397301	*GSTTP1, GSTT1*	Fusion	Ancient

Figure legends

Figure 1. Identification of deletions shared with Neandertal and Denisovan genomes
(**A**) We counted the number of Neandertal sequence reads mapping to the genomic intervals where the human deletions were reported by 1000 Genomes project. The histogram shows on the x-axis the distribution of number of reads in each interval divided by the size of the interval and on the y-axis the frequency of the intervals observed. The dotted blue vertical line indicates the mean of this distribution. The solid blue line shows the normal distribution with the mean and the standard deviation of the observed distribution. The dotted red vertical line indicates the read-depth size ratio, where the probability of observing a smaller ratio under normal distribution is 0.01. For the intervals that fall into the red transparent box, the read-depth in Neandertal resequencing is lower than expected by chance/noise, and consequently we assumed that these are deleted in the Neandertal genome. These deletions were indicated by red histogram bars.
(**B**) The read depth/size ratio distribution analysis for the Denisovan genome.
(**C**) Venn diagram showing the number of human deletions that are found only in humans (light brown), shared with Neandertal genome (pink), shared with Denisovan genome (blue), or shared with both (purple).

Figure 2. Possible evolutionary scenarios explaining allele sharing (or lack thereof) between modern and archaic hominins
This figure shows possible evolutionary scenarios, in the form of cartoon phylogenetic trees, explaining the deletion polymorphisms across different lineages. Red color designates branches where the deletion was observed. The number shown under each tree is the number of polymorphic deletions corresponding to the observation. The red headers indicate the likely mechanisms, which were separated from each other by dotted lines, through which the allele sharing have evolved. The *human specific deletion* scenario covers polymorphic deletions which are shared neither with Neandertal nor with Denisovan genomes. The *recurrent* scenario covers Neandertal- or Denisovan-shared human deletions, breakpoints of which vary among different lineages. The *Neandertal introgression* scenario indicates allele sharing due to Neandertal gene flow into non-African human populations. The *ancient genetic structure* scenario indicates deletions that were evolved in the Human-Neandertal ancestral population and have been maintained since then. The *primate incomplete lineage sorting* scenario indicates deletion polymorphisms that have potentially been maintained since before the Human-Chimpanzee divergence. The *hominid-specific insertion* scenario covers polymorphic deletions that are genotyped as deletions in chimpanzee and rhesus monkey, and show polymorphism in hominid genomes. This scenario represents likely novel sequences that evolved in the ancestral population of Neandertals and humans.

Figure 3. Characterization of ancient deletion variants

(**A**) This figure shows the proportions of Neandertal introgressed deletions, maintained ancient deletions, recurrent deletions and human-specific deletions that overlap with exonic regions. H designates human-specific deletions; R designates recurrent deletions. A designates deletions maintained from ancient genetic structure; and N designates Neandertal introgressed deletions. The orange-colored section indicates the fraction of deletions that overlaps with exons. Deletion polymorphisms are known to be depleted for exonic content. However, the ancient deletion variants are even less exonic than human-specific deletion variants (p=0.0011, Chi-square test).
(**B**) Cumulative fraction plot of allele frequency in modern humans for deletion variants shared with Neandertals and/or Denisovans (Shared - Light blue), and human-specific (Human - Red) deletions. The x-axis indicates the frequency of the deletion variants and y-axis indicates the cumulative fraction of the deletions of all the deletions at a given or lower frequency. The steep slope towards the left end of the human-specific deletions curve shows that the majority of these polymorphic deletions have allele frequencies smaller than 0.1. As for deletions common to Neandertals/Denisovans, about 43% of them have allele frequencies over 0.1. Overall, "Shared" deletions are significantly more common than Human specific deletions ($p<2.2\times10^{-16}$, Wilcoxon rank test).

Figure 4. Analysis of exonic variants
(**A-C**) The allele frequency of exonic deletion variants. The x- and y-axes indicates the allele frequencies in a given continent. The heatmap colors represent number of observations. The red to dark blue gradient corresponds to decreased density of observed deletions. Here we plot allele frequencies of all 14,422 1KG deletion variants. As such, red spots designate thousands of observations decreasing to hundreds of observations for yellow pixels and single observations for purple dots. A vast majority of deletions has very low allele frequencies. The exonic variants shared with Neandertal/Denisovan genomes are shown with white colored circles.
(**D**) Heatmap of the percentiles of pairwise F_{ST} values measured for the flanking regions of the ancient exonic deletion variants between 10 non-admixed populations using clustering without *a priori* input. The colors in the heatmap correspond to the percentile of the F_{ST} values as compared to the distribution of all exonic human deletions, with light-yellow/white being the highest values observed (1 indicating the highest percentile) and dark red corresponding to lower values (0 indicating the lowest possible percentile). Exact values to generate this map can be found in Table S2. On the x-axis are the exonic deletion variants, represented by the names of the genes that they affect. The ones highlighted in blue are introgressed deletions, the one highlighted in green is a recurrent deletion, and those that are not highlighted are ancient deletions. On y-axis are population pairs used in the analysis. AFR: African; ASN: Asian; EUR: European; CEU: Utah residents (CEPH) with Northern and Western European ancestry; CHB: Han Chinese in Beijing; CHS: Southern Han Chinese; FIN: Finish in Finland; GBR: British in England and Scotland; IBS: Iberian population in Spain; JPT: Japanese in Tokyo; LWK: Luhya in Webuye, Kenya; TSI: *Toscani* in Italia; YRI: Yoruba in Ibadan, Nigeria.
(**E**) Heatmap of the percentiles of Tajima's D values observed for ancient exonic deletion variants within the Tajima's D distribution of all exonic human deletions. The colors in the heatmap correspond to the percentile of the Tajima's D values as compared to Tajima's D

values calculated for all exonic human deletions, with light-yellow/white being the highest values observed (1 indicating the highest percentile) and dark red corresponding to lower values (0 indicating the lowest possible percentile). Exact values to generate this map can be found in **Table S2.** On the x-axis are the exonic deletion variants, represented by the names of the genes that they affect. The ones highlighted in blue are introgressed deletions, the one highlighted in green is a recurrent deletion, and those that are not highlighted are ancient deletions. On the y-axis are the populations for which Tajima's D values were measured. The population designations can be found above, in the legend of Figure 4D.

Materials and Methods

Genotyping of Neandertal and Denisovan genomes

We used 1000 Genomes dataset Phase 1 dataset (http://www.1000genomes.org/data) (1000 Genomes Project Consortium 2012), as well as the the high-coverage genome-wide sequencing data for Neandertal (Prüfer et al. 2014a) and Denisovan (Meyer et al. 2012) (available at http://cdna.eva.mpg.de) as our main starting point. The sequencing data from both Neandertal and Denisovan genomes, as well as the 1KG deletions are aligned to the same version of the human reference genome (Hg19). As such, we were able to directly measure the number of reads mapping to the intervals where 14,422 polymorphic human deletions were reported. To accomplish this, we used a custom bedtools (http://bedtools.readthedocs.org/) script.

We then constructed a normal distribution of the read-depth (as normalized by size) for Neandertal and Denisovan dataset where the mean and standard deviation were equal to the observed mean and standard deviation. We then established a threshold value that corresponds to 0.01 quantile in the normal distribution, below which we assumed the intervals have lower than expected read-depth/size ratio indicating a deletion in the respective genome.

To determine the false discovery rate, we used the shuffleBed function of bedtools to generate a set of 14,422 random intervals that matches the size distribution of 1KG deletions. We then followed the genotyping workflow described above to identify deletions. We found less than 10 deletions in those random regions in both Neandertal and Denisovan genomes, indicating a false discovery rate is essentially lower than 5% for both genomes.

We used a modified version of SPLITREAD approach (http://splitread.sourceforge.net/) to remap the Neandertal and Denisovan reads to the junctions of the breakpoints of deletions as defined in 1KG deletion dataset, rather than applying it to the whole genome. Specifically, we extended the deletion breakpoints 50,000 bases to the upstream and downstream to generate out mappable reference contigs. The mappings to these reference contigs was performed using BWA-MEM method (Li and Durbin 2009). The PCR duplicates, identified by the reads that are mapping exactly to the same positions, were removed from the downstream analysis. Due to the fact that short reads are single ended, there is an increased rate of support for the regions that are repetitive or in segmental duplication regions. SPLITREAD method also relies on the consistent mapping of the splits at the breakpoints with base pair resolution. The deletions that are recurrent with different breakpoints are expected to have less support overall. We observed significantly reduced number of split-read mapping with the Neandertal genome sequences as compared to what was observed for Denisovan genome sequences. We think this is due to

inherent read-length variation and differential impact of whole-genome-amplification between these two genome sequences.

Categorization of the origin of deletions that are shared with Neandertal and Denisovan genomes

We assessed all of the 14,422 1KG deletions for their occurrence in different lineages. The occurrence of these deletions in Neandertal and Denisovan genomes were determined by genotyping described above. Introgressed deletions in humans were defined based on allele frequency distribution, overlap with introgressed regions in human genome described by Verot and Akey (2014). Specifically, we assumed that those deletions that have lower than 0.01 allele frequency in Africa, accommodating a certain level of possible back migration, is Eurasia specific. To determine whether these deletion are present in chimpanzee lineage, we mapped them onto chimpanzee reference genome panTro4 using Liftover tool (available through USCS Genome Browser (Kent et al. 2002) with minimum ratio of bases remap = 0.95. Then we manually examined those that failed to map using chain and net tracks in USCS genome browser for both chimpanzees and rhesus macaques. The breakpoints of all the Neandertal/Denisovan shared deletions are checked on integrative genome viewer (IGV) (Nicol et al. 2009) to determine whether they are recurrent.

Exon content analysis

We used Galaxy software tools (Goecks et al. 2010) to identify 1KG deletions that overlap with RefSeq exon track (http://www.ncbi.nlm.nih.gov/refseq/). We assessed the significance of the depletion of exonic deletion we observed using Genome Structure Correction software (https://www.encodeproject.org/software/gsc/). Briefly, this software uses a subsampling approach to avoid confounding factors in the localization of genomic elements for which the analysis is being conducted.

Population genetics analysis

We calculated Tajima's D (Tajima 1989) in 10 kb genomic regions located 500 bp upstream each deletion (*i.e.,* the genomic region from 500 bp to 10,500 bp upstream each deletion). We included the phased deletion variant as a single variant to this analysis. Polymorphic data were downloaded from the 1000 Genomes project (www.1000genomes.org; phase1, v3.20101123). Polymorphic data were converted to FASTA alignments using the human reference genome (hg19). Calculations of F_{ST} and Tajima's D were performed by CoMuStats (http://pop-gen.eu/wordpress/software/comus-coalescent-of-multiple-species). F_{ST} (Hudson et al. 1992) was calculated for the total sample as well as for all demes pairs using the allele frequencies of the deletions. We calculated the F_{ST} and Tajima's D values for all exonic sequences. Based on the distribution of these statistics, we calculated the percentile for each of the "shared" exonic deletions and for each "non-admixed" population. All source codes used for the calculations are available from (http://gokcumenlab.org/data-and-codes/).

Statistical tests and graphs

All other statistical tests and graphs were conducted using base statistical and ggplot2 (http://ggplot2.org/) packages available through R statistical environment (http://www.r-project.org/).

References

1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65.

Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. Nat. Rev. Genet. 12:363–376.

Barreiro LB, Patin E, Neyrolles O, Cann HM, Gicquel B, Quintana-Murci L. 2005. The heritage of pathogen pressures and ancient demography in the human innate-immunity CD209/CD209L region. Am. J. Hum. Genet. 77:869–886.

Bickel PJ, Boley N, Brown JB, Huang H, Zhang NR. 2010. SUBSAMPLING METHODS FOR GENOMIC INFERENCE. Ann. Appl. Stat. 4:1660–1697.

Cagliani R, Fumagalli M, Riva S, Pozzoli U, Comi G, Menozzi G, Bresolin N, Sironi M. 2008. The signature of long-standing balancing selection at the human defensin beta-1 promoter. Genome Biol. 9:R143.

Castellano S, Parra G, Sánchez-Quinto FA, et al. 2014. Patterns of coding variation in the complete exomes of three Neandertals. Proc. Natl. Acad. Sci. U. S. A. 111:6666–6671.

Caswell JL, Mallick S, Richter DJ, Neubauer J, Schirmer C, Gnerre S, Reich D. 2008. Analysis of chimpanzee history based on genome sequence alignments. PLoS Genet. 4:e1000057.

De Cid R, Riveira-Munoz E, Zeeuwen PLJM, et al. 2009. Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. Nat. Genet. 41:211–215.

Conrad DF, Pinto D, Redon R, et al. 2009. Origins and functional impact of copy number variation in the human genome. Nature 464:704–712.

Currat M, Excoffier L. 2011. Strong reproductive isolation between humans and Neanderthals inferred from observed patterns of introgression. Proc. Natl. Acad. Sci. U. S. A. 108:15129–15134.

Dean M, Carrington M, O'Brien SJ. 2002. Balanced polymorphism selected by genetic versus infectious human disease. Annu. Rev. Genomics Hum. Genet. 3:263–292.

DeGiorgio M, Lohmueller KE, Nielsen R. 2014. A model-based approach for identifying signatures of ancient balancing selection in genetic data. PLoS Genet. 10:e1004561.

Garrigan D, Hammer MF. 2006. Reconstructing human origins in the genomic era. Nat. Rev. Genet. 7:669–680.

Goecks J, Nekrutenko A, Taylor J, Galaxy Team. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 11:R86.

Gokcumen O, Babb PL, Iskow RC, et al. 2011. Refinement of primate copy number variation hotspots identifies candidate genomic regions evolving under positive selection. Genome Biol. 12:R52.

Gokcumen O, Zhu Q, Mulder LCF, et al. 2013. Balancing Selection on a Regulatory Region Exhibiting Ancient Variation That Predates Human–Neandertal Divergence. PLoS Genet. 9:e1003404.

Green RE, Krause J, Briggs AW, et al. 2010. A draft sequence of the Neandertal genome. Science 328:710–722.

Hammer MF, Woerner AE, Mendez FL, Watkins JC, Wall JD. 2011. Genetic evidence for archaic admixture in Africa. Proceedings of the National Academy of Sciences 108:15123–15128.

Hu Y, Wang Y, Ding Q, He Y, Wang M, Wang J, Xu S, Jin L. 2014. Genome-wide Scan of Archaic Hominin Introgressions in Eurasians Reveals Complex Admixture History. arXiv [q-bio.PE].

Hudson RR, Boos DD, Kaplan NL. 1992. A statistical test for detecting geographic subdivision. Mol. Biol. Evol. 9:138–151.

Huerta-Sánchez E, Jin X, Asan, et al. 2014. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature 512:194-197.

Itsara A, Cooper GM, Baker C, et al. 2009. Population analysis of large copy number variants and hotspots of human genetic disease. Am. J. Hum. Genet. 84:148–161.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. Genome Res. 12:996–1006.

Lazaridis I, Patterson N, Mittnik A, et al. 2013. Ancient human genomes suggest three ancestral populations for present-day Europeans.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.

McLean CY, Reno PL, Pollen AA, et al. 2011. Human-specific loss of regulatory DNA and the evolution of human-specific traits. Nature 471:216–219.

Meyer M, Kircher M, Gansauge M-T, et al. 2012. A high-coverage genome sequence from an archaic Denisovan individual. Science 338:222–226.

Mills RE, Walter K, Stewart C, et al. 2011. Mapping copy number variation by population-scale genome sequencing. Nature 470:59–65.

Mollenhauer J, Holmskov U, Wiemann S, Krebs I, Herbertz S, Madsen J, Kioschis P, Coy JF, Poustka A. 1999. The genomic structure of the DMBT1 gene: evidence for a region with susceptibility to genomic instability. Oncogene 18:6233–6240.

Mollenhauer J, Wiemann S, Scheurlen W, Korn B. 1997. DMBT1, a new member of the SRCR superfamily, on chromosome 10q25. 3–26.1 is deleted in malignant brain tumours. Nature Genetics 17:32-39

Ménard V, Eap O, Harvey M, Guillemette C, Lévesque E. 2009. Copy-number variations (CNVs) of the human sex steroid metabolizing genes UGT2B17 and UGT2B28 and their associations with a UGT2B15 functional polymorphism. Hum. Mutat. 30:1310–1319.

Nicol JW, Helt GA, Blanchard SG Jr, Raja A, Loraine AE. 2009. The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. Bioinformatics 25:2730–2731.

Padidela R, Bryan SM, Abu-Amero S, Hudson-Davies RE, Achermann JC, Moore GE, Hindmarsh PC. 2012. The growth hormone receptor gene deleted for exon three (GHRd3) polymorphism is associated with birth and placental weight. Clin. Endocrinol. 76:236–240.

Prüfer K, Racimo F, Patterson N, et al. 2014a. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505:43–49.

Prüfer K, Racimo F, Patterson N, et al. 2014b. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505:43–49.

Reich D, Green RE, Kircher M, et al. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468:1053–1060.

Renner M, Bergmann G, Krebs I, et al. 2007. DMBT1 confers mucosal protection in vivo and a deletion variant is associated with Crohn's disease. Gastroenterology 133:1499–1509.

Rudolph MC, Neville MC, Anderson SM. 2007. Lipid synthesis in lactation: diet and the fatty acid switch. J. Mammary Gland Biol. Neoplasia 12:269–281.

Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, Pääbo S, Patterson N, Reich D. 2014. The genomic landscape of Neanderthal ancestry in present-day humans. Nature 507:354-357.

Dos Santos C, Essioux L, Teinturier C, Tauber M, Goffin V, Bougnères P. 2004. A common polymorphism of the growth hormone receptor is associated with increased responsiveness to growth hormone. Nat. Genet. 36:720–724.

Skoglund P, Jakobsson M. 2011. Archaic human ancestry in East Asia. Proc. Natl. Acad. Sci. U. S. A. 108:18301–18306.

Sørensen K, Aksglaede L, Petersen JH, Leffers H, Juul A. 2010. The exon 3 deleted growth hormone receptor gene is associated with small birth size and early pubertal onset in healthy boys. J. Clin. Endocrinol. Metab. 95:2819–2826.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595.

Teixeira, de Filippo JC, Weihmann C, et al. 2014. Long-term balancing selection in LAD1 maintains a missense trans-species polymorphism in humans, chimpanzees and bonobos. bioRxiv [Internet]. Available from: http://dx.doi.org/10.1101/006684

Veeramah KR, Hammer MF. 2014. The impact of whole-genome sequencing on the reconstruction of human population history. Nat. Rev. Genet. 15:149–162.

Vernot B, Akey JM. 2014. Resurrecting surviving Neandertal lineages from modern human

genomes. Science 343:1017–1021.

Wall JD, Yang MA, Jay F, et al. 2013. Higher levels of neanderthal ancestry in East Asians than in Europeans. Genetics 194:199–209.

Xue Y, Sun D, Daly A, et al. 2008. Adaptive evolution of UGT2B17 copy-number variation. Am. J. Hum. Genet. 83:337–346.

Figures

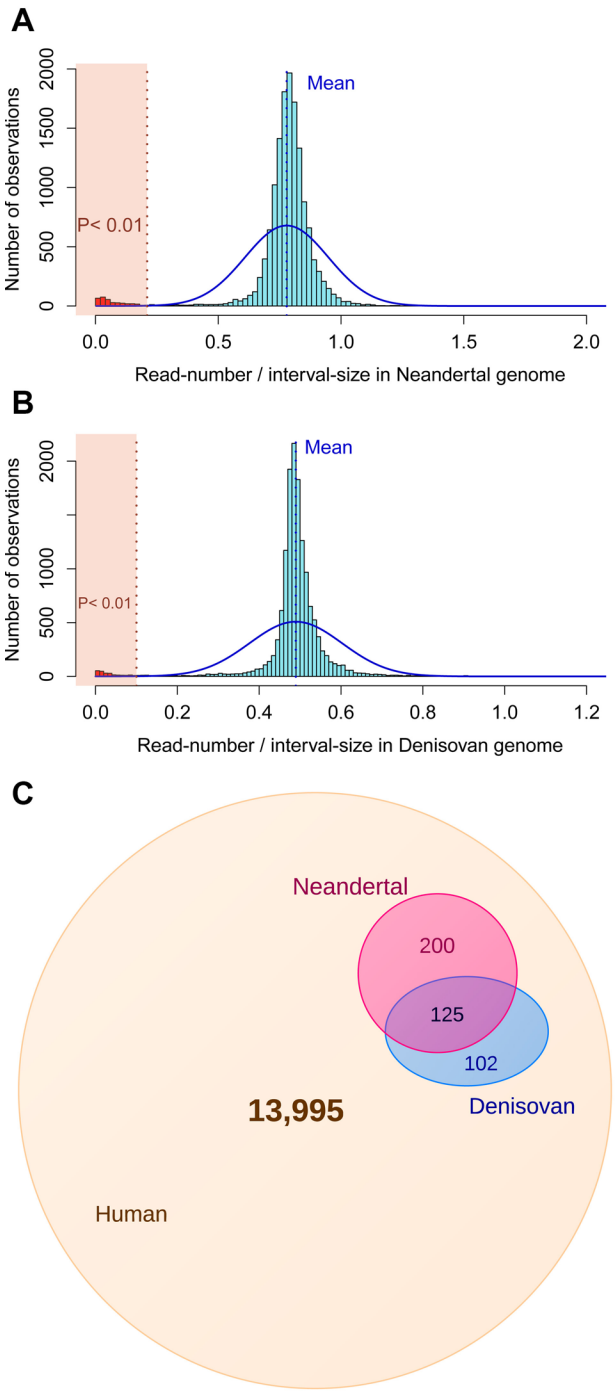
A
Number of observations
0
500
1000
1500
2000
P< 0.01
Mean
0.0
0.5
1.0
1.5
2.0
Read-number / interval-size in Neandertal genome
B
Number of observations
0
500
1000
2000
P< 0.01
Mean
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Read-number / interval-size in Denisovan genome
C
Neandertal
200
125
102
Denisovan
13,995
Human

Figure 1

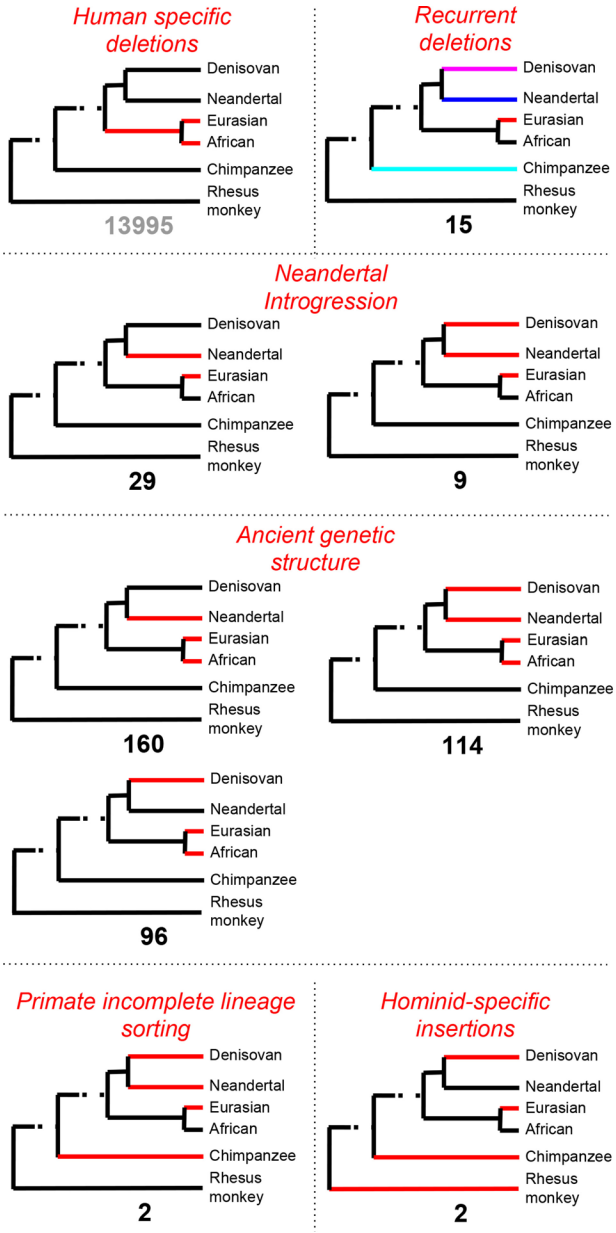
Human specific deletions
Denisovan
Neandertal
Eurasian
African
Chimpanzee
Rhesus monkey
13995
Recurrent deletions
Denisovan
Neandertal
Eurasian
African
Chimpanzee
Rhesus monkey
15
Neandertal Introgression
Denisovan
Neandertal
Eurasian
African
Chimpanzee
Rhesus monkey
29
Denisovan
Neandertal
Eurasian
African
Chimpanzee
Rhesus monkey
9
Ancient genetic structure
Denisovan
Neandertal
Eurasian
African
Chimpanzee
Rhesus monkey
160
Denisovan
Neandertal
Eurasian
African
Chimpanzee
Rhesus monkey
114
Denisovan
Neandertal
Eurasian
African
Chimpanzee
Rhesus monkey
96
Primate incomplete lineage sorting
Denisovan
Neandertal
Eurasian
African
Chimpanzee
Rhesus monkey
2
Hominid-specific insertions
Denisovan
Neandertal
Eurasian
African
Chimpanzee
Rhesus monkey
2

Figure 2

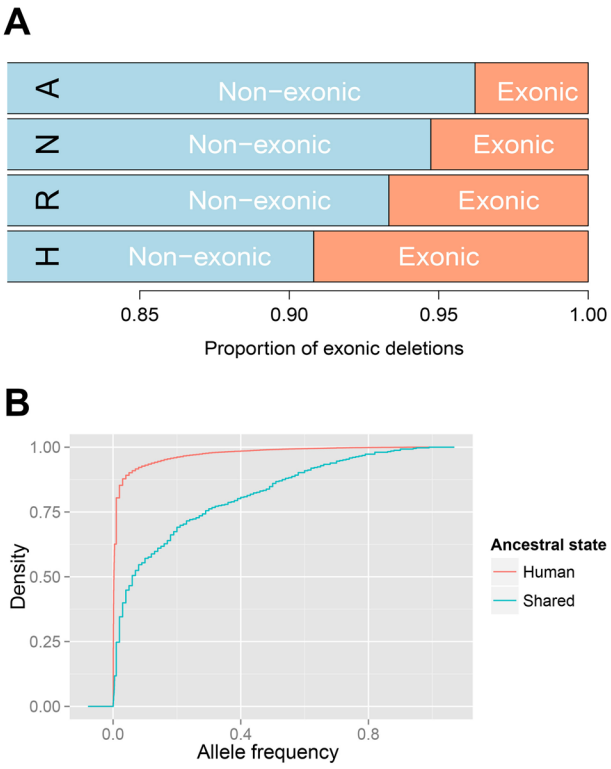
A
A
Non−exonic
Exonic
N
Non−exonic
Exonic
R
Non−exonic
Exonic
H
Non−exonic
Exonic
0.85
0.90
0.95
1.00
Proportion of exonic deletions
B
1.00
0.75
0.50
0.25
0.00
Density
0.0
0.4
0.8
Allele frequency
Ancestral state
Human
Shared

Figure 3

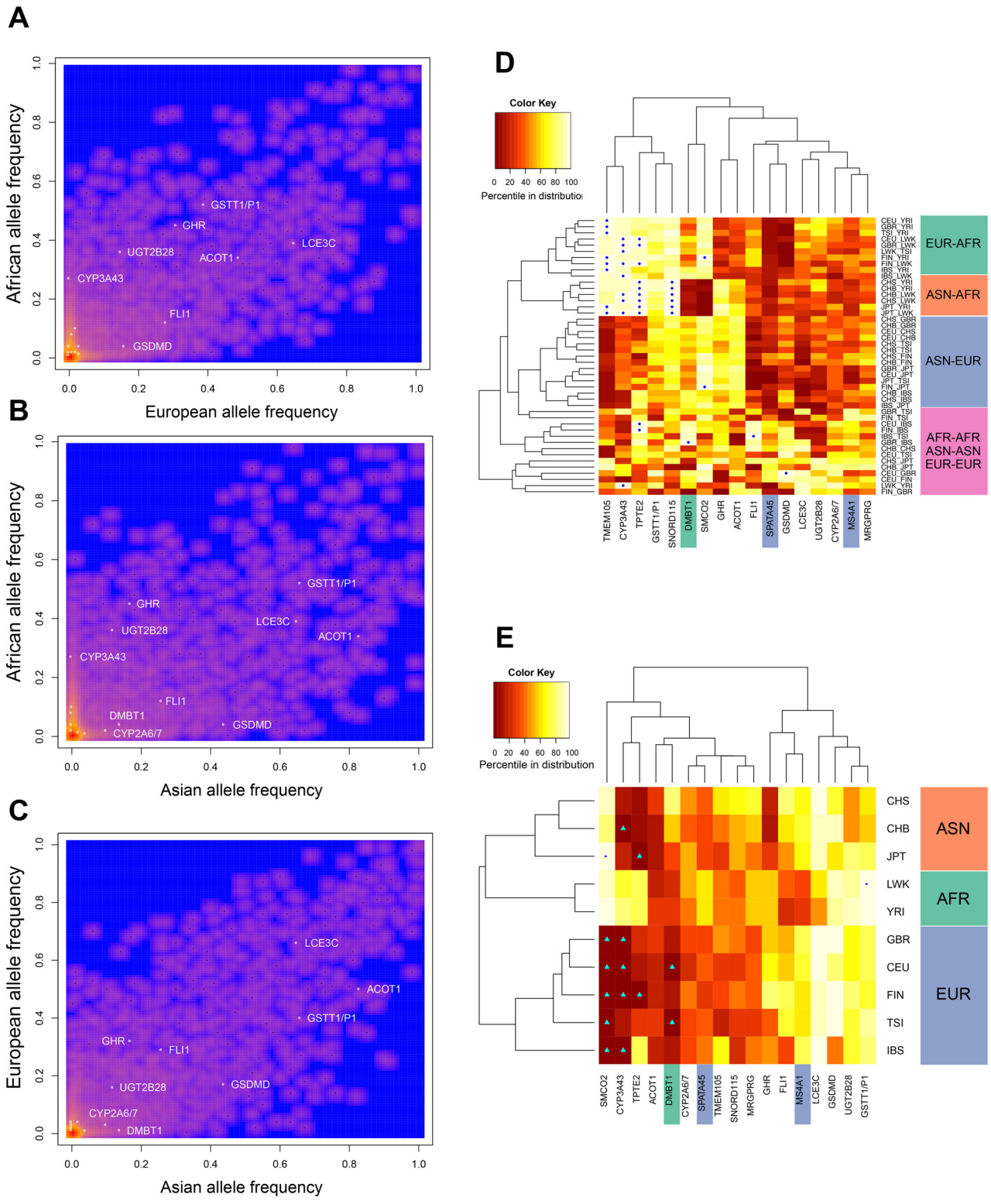
A
African allele frequency
European allele frequency
GSTT1/P1
GHR
LCE3C
UGT2B28
ACOT1
CYP3A43
FLI1
GSDMD
B
African allele frequency
Asian allele frequency
GSTT1/P1
GHR
LCE3C
UGT2B28
ACOT1
CYP3A43
FLI1
DMBT1
GSDMD
CYP2A6/7
C
European allele frequency
Asian allele frequency
LCE3C
ACOT1
GSTT1/P1
GHR
FLI1
UGT2B28
GSDMD
CYP2A6/7
DMBT1
D
Color Key
0 20 40 60 80 100
Percentile in distributio
EUR-AFR
ASN-AFR
ASN-EUR
AFR-AFR
ASN-ASN
EUR-EUR
TMEM105
CYP3A43
TPTE2
GSTT1/P1
SNORD115
DMBT1
SMCO2
GHR
ACOT1
FLI1
SPATA45
GSDMD
LCE3C
UGT2B28
CYP2A6/7
MS4A1
MRGPRG
E
Color Key
0 20 40 60 80 100
Percentile in distribution
CHS
CHB
JPT
LWK
YRI
GBR
CEU
FIN
TSI
IBS
ASN
AFR
EUR
SMCO2
CYP3A43
TPTE2
ACOT1
DMBT1
CYP2A6/7
SPATA45
TMEM105
SNORD115
MRGPRG
GHR
FLI1
MS4A1
LCE3C
GSDMD
UGT2B28
GSTT1/P1

Figure 4