12-2-2014

Geographic Distribution And Adaptive Significance Of Genomic Structural Variants: An Anthropological Genetics Perspective

Muthukrishnan Eaaswarkhanth
Department of Biological Sciences, University at Buffalo, The State University of New York

Pavlos Pavlidis
Foundation for Research and Technology–Hellas, Institute of Molecular Biology and Biotechnology (IMBB), Heraklion, Crete, Greece

Omer Gokcumen
Department of Biological Sciences, University at Buffalo, The State University of New York, omergokc@buffalo.edu

Geographic Distribution And Adaptive Significance Of Genomic Structural Variants: An Anthropological Genetics Perspective

Muthukrishnan Eaaswarkhanth[1], Pavlos Pavlidis[2], Omer Gokcumen*[1]

[1]Department of Biological Sciences, University at Buffalo, The State University of New York

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

*Correspondence to: Omer Gokcumen, Ph.D. Department of Biological Sciences, State University of New York at Buffalo, 641 Cooke Hall Buffalo, NY 14260-1300, U.S.A omergokc@buffalo.edu.

Key words: copy number variants, anthropological genomics, local adaptation, keratin-associated proteins.

Abstract

Anthropological geneticists have successfully used single nucleotide and short tandem repeat variations across human genomes to reconstruct human history. These markers have also been used extensively to identify adaptive and

phenotypic variation. The recent advent of high-throughput genomic technologies revealed an overlooked type of genomic variation, namely structural variants (SVs). In fact, some SVs may contribute to human adaptation in substantial and previously unexplored ways. SVs include deletions, insertions, duplications, inversions and translocations of genomic segments that vary among individuals from the same species. SVs are much less numerous than single nucleotide variants, but account for at least seven times more variable base pairs than do single nucleotide variants when two human genomes are compared. Moreover, recent studies have shown that SVs have higher mutation rates than single nucleotide variants when the affected base pairs are considered, especially in certain parts of the genome. The null hypothesis for the evolution of SVs, like single nucleotide variants, is neutrality. Hence, drift is the primary force that shapes the current allelic distribution of most SVs. However, due to their size, a larger proportion of SVs appear to evolve under non-neutral forces (mostly purifying selection) than single nucleotide variants. In fact, as exemplified by several groundbreaking studies, SVs contribute to anthropologically relevant phenotypic variation and local adaption among humans.

In this review, we argue that with the advent of affordable genomic technologies, anthropological scrutiny of genomic structural variation emerges as a fertile area of inquiry to better understand human phenotypic variation. To motivate potential studies, we discuss scenarios through which SVs affect

phenotypic variation among humans within an anthropological context. We further provide a methodological workflow in which we analyzed 1000 Genomes deletion variants and identified 16 exonic deletions that are specific to the African continent. We analyzed two of these deletion variants affecting the keratin-associated protein (KAP) cluster in a locus-specific manner. Our analysis revealed that these deletions may indeed affect phenotype and likely evolved under geography-specific positive selection. We outline all the major software and datasets for these analyses and also provide the basic R and perl codes we used for this example workflow analysis. Overall, we hope that this review will encourage and facilitate incorporation of genomic structural variation in anthropological research programs.

Beyond Neutral Markers in Anthropological Genetics

Anthropological geneticists have been utilizing genetic markers since the early 20th Century. One of the first major studies that incorporated modern genetic methodologies into anthropological questions explored the spread of early farming in Europe from the Near East (Menozzi et al. 1978). Later, more refined attempts of reconstructing the origin and dispersal of humans have frequently focused on analyzing maternally inherited mitochondrial and paternally inherited Y chromosome variations (reviewed in Underhill and Kivisild 2007). These studies have established the African origin of the genetic variation of modern humans (*e.g.*, Hammer 1995; Cann et al. 1987), connected major historical events to contemporary distributions of genetic variations (*e.g.*, Zerjal et al. 2003), scrutinized the impact of social dynamics on genetic structure (*e.g.*, Chaix et al. 2007) and revealed gender-specific human migration patterns (*e.g.*, Seielstad et al. 1998).

The increased use of unlinked autosomal biallelic markers, such as short tandem repeats and single nucleotide variants, provided a more accurate and fine-tuned understanding of the patterns of human genetic structure. For instance, microsatellite diversity studies revealed within and among population genetic variation in worldwide human populations in unprecedented resolution (*e.g.*, Rosenberg 2011; Tishkoff et al. 2009). Similarly, investigations based on single nucleotide variants demonstrated a clear association of genetic variation with

contemporary geographical location (Lao et al. 2008; Novembre et al. 2008; Ralph and Coop 2013). Today, the advent of whole genome sequencing provides even more powerful ways in which to investigate genetic variation among human groups (reviewed in detail in Veeramah and Hammer 2014).

The increased amount of genomic data and availability of more sophisticated computational methods allow anthropological and medical geneticists to move beyond neutral markers to study phenotypic variation among humans. Some of the studies in this line identified genetic variants that are adaptively associated with taste perception (*e.g.*, Campbell et al. 2012), smell perception (reviewed in Hasin-Brumshtein et al. 2009), skin color (*e.g.*, Norton et al. 2007), eye and hair color (*e.g.*, Sulem et al. 2007), various aspects of diet (*e.g.*, Tishkoff et al. 2007), high-altitude lifestyle (*e.g.*, Bigham et al. 2010), and immune response (reviewed in Karlsson et al. 2014). These connections between genetic and phenotypic variation among healthy individuals are often considered within an adaptive framework, and more recently in the light of gene–culture co-evolution (reviewed in Laland et al. 2010). Another major development that has happened in anthropological genetics with the advent of genome-wide resequencing techniques is the scrutinization of whole genomes from now-extinct hominins, such as Neandertals (Green et al. 2010) and Denisovans (Reich et al. 2010). These studies demonstrated that the archaic humans exchanged genetic material with modern humans on multiple occasions (Hu et al. 2014), and these

introgressions may have affected modern human phenotypes (*e.g.*, Huerta-Sánchez et al. 2014). These adaptive genetic variants are starting to give us a glimpse of the complex ecological, cultural and social pressures that shape the contemporary human genetic diversity.

Overall, new sequencing technologies allow affordable, genome-wide investigation of human genetic variation at the population level. At present, there are thousands of human genomes available for data mining and for contextualization of local genetic variation. Moreover, ancient genomes now provide a previously unavailable source of "outgroup" information to better frame contemporary genetic variation. We also have a better understanding of the functional impact of different regions in the genome through myriad locus-specific functional studies (*e.g.*, Kamberov et al. 2013), genome-wide assessment of functional relevance (*e.g.*, ENCODE Project Consortium 2012) and RNA-seq datasets (*e.g.*, Lappalainen et al. 2013).

Most of the genomic studies have so far focused on global trends that shape the genetic variation among major continental populations. Notably, their sampling schemes do not necessarily reflect anthropological concerns (*e.g.*, Gokcumen et al. 2011b). It is apparent that anthropologists are now in a unique position to combine these recently affordable genomic technologies with their thorough understanding of local cultural, ecological and historical diversity to ask powerful questions regarding genetic bases of human phenotypic variation.

Structural Variants: A Hidden, Fertile Ground for Genetic Anthropology

One of the many breakthrough developments in human genetics in the last 10 years is the appreciation of the extent and impact of genomic structural variants (SVs) (Iafrate et al. 2004; Sebat et al. 2004). SVs involve blocks of sequences that vary in copy number (copy number variants), chromosomal locations (translocations) or directionality (inversions) (reviewed in detail in Weischenfeldt et al. 2013). The reason why SVs made such a major splash in the field of human genetics is the realization in the community of their sheer impact to the overall human genetic variation. In fact, SVs constitute at least seven times more variable base pairs than do single nucleotide variants when two human genomes are compared to each other (Conrad et al. 2010). Like other types of genetic variation in humans, it would be safe to argue that majority of common SVs evolve under neutral conditions and may have minimal impact on phenotype. However, some SVs may affect phenotype. Indeed, several studies have further scrutinized the phenotypic impact of SVs through genome-wide association studies linking SVs to diseases such as autism (Sebat et al. 2007), schizophrenia (Stefansson et al. 2008), psoriasis (de Cid et al. 2009), obesity (Falchi et al. 2014; Jacquemont et al. 2011) and Crohn's disease (McCarroll et al. 2008), among others. Other studies have unearthed important evolutionary trends that are shaped by genomic

structural differences between primate species (McLean et al. 2011; Charrier et al. 2012; Iskow et al. 2012a; Gokcumen et al. 2013a). Moreover, especially by the efforts of the 1000 Genomes Project Consortium, several complementary computational and experimental approaches were integrated to identify and analyze different types of SVs (Mills et al. 2011; 1000 Genomes Project Consortium 2012). Therefore, we argue that the time is ripe for studying SVs in an anthropologically contextualized hypothesis-driven manner.

The human genome is packed into 23 pairs of chromosomes and these pairs are inherited with one chromosome coming from each parent. As such, it is expected that each sequence exists in 2 copies across the genome. However, recent studies have shown that a large portion of the genome actually deviates from 2 copies, caused by different genetic mechanisms, such as transposition, simple repeat expansions and segmental duplications (Conrad et al. 2010). For instance, segmental duplications, which are large (>1kb) duplications of otherwise non-repetitive sequences, constitute approximately 5% of the human genome (Bailey et al. 2002). Segmental duplications are also hotspots for new variation in the copy number of duplicated segment to emerge through homology-based recombination errors (reviewed in Weischenfeldt et al. 2013). It is estimated that these homology-based mechanisms alone generate about 2.4×10^{-2} *de novo* large (>100 kbp) copy number variants (deletions or duplications) each generation in humans, affecting on average 100 times more base pairs than single nucleotide

variants do at the same generational time (Campbell and Eichler 2013). This estimate does not include smaller events, such as variable transpositions, insertion–deletion polymorphisms or balanced rearrangements, such as inversions and translocations. Therefore, SVs in every generation provide more variation for non-neutral forces to act upon than do single nucleotide variants in humans.

SVs affect gene function in multifarious ways. Briefly, whole gene duplications, for instance, were shown to lead to an increase in expression (dosage), gain-of-new-function (neofunctionalization), or expression in new tissues (Gokcumen et al. 2013a). Incomplete gene duplications, on the other hand, may create new transcripts that interfere with the ancestral gene function (Dennis et al. 2012). Variation in the number of tandem repeats within exons, albeit under-studied, can affect protein sizes and function with important phenotypic consequences (Barreiro et al. 2005). Exonic or whole gene deletions can lead to loss-of-function alleles (Kidd et al. 2007) or to formation of novel transcripts (Dos Santos et al. 2004). In addition, SVs can affect the functioning of regulatory regions (Stranger et al. 2007; McCarroll et al. 2008).

It is not surprising that recent studies have shown evidence for strong purifying (*i.e.*, negative) selection acting on SVs (*e.g.*, Derti et al. 2006), eliminating SVs overlapping with coding sequences in the genome (Conrad et al. 2010). However, a considerable number of SVs still overlap with known functional parts of the genome, affecting primarily those genes involved in

environmental interactions, *e.g.*, diet, immunity and olfactory reception (Conrad et al. 2010). A comprehensive review of the adaptive significance of these genes can be found elsewhere (Iskow et al. 2012b). Below, we highlight a few important examples within the context of anthropological genetics.

One such example involves the copy number variation of the salivary amylase gene (*AMY1)*. This gene is involved in starch digestion in the mouth. Recent studies have shown that the human genomes harbor more than 2 copies of amylase gene (*AMY1*), in contrast to chimpanzees, which have only 2 copies (Perry et al. 2007). In addition to this cross-species difference, the copy number of *AMY1* varies from 4 to 17 copies among humans. This considerable genetic variation corresponds positively with the expression of this gene at the protein level. In a broader anthropological context, the copy number of *AMY1* varies among human populations and shows positive correlation with starch content in diets of these populations, irrespective of their geographical or ethnic origin (Perry et al. 2007). On top of these important observations, the low copy number *AMY1* was recently associated with susceptibility to obesity (Falchi et al. 2014), linking recent adaptive evolution of SVs to contemporary human disease. It is not clear, however, whether if the evolution of *AMY1* copy number is merely due to increase in the ability of physical breakdown of starch. It is plausible that increased copy number of *AMY1* may have other, perhaps behavioral, consequences. For instance, a recent study puts forward evidence for association

of *AMY1* copy number with perception of taste (de Wijk et al. 2004). Such associations can be studied within a cultural and linguistic context and raises interesting possibilities for further anthropological research.

SVs affect acquired and innate immune system genes more often than they do the genes involved in other functional categories. In fact, a growing body of work shows the importance of copy number variants (one of the major classes of SVs) in the susceptibility to infectious diseases (reviewed in Hollox and Hoh 2014) and autoimmune disorders (*e.g.*, McCarroll et al. 2008). Immune system gene families, which have evolved through gene duplication events (reviewed in Zhang 2003) are prone to further accumulation of structural variation through non-allelic homologous recombination (Hollox et al. 2008; Traherne et al. 2010). Essentially, gene duplication creates homologous sequences that can lead to an erroneous synapsis, which in turn facilitates non-allelic homologous recombination referred above. This process would essentially create additional copies of the homologous sequence in one chromosome, while leading to the loss of some copies of the same sequence in the other chromosome.

For example, the copy number of β-defensin genes, which are highly homologous innate immunity genes, is highly variable among humans. Certain deletions affecting β-defensin gene cluster are associated with the breakdown of the antibacterial barrier in the intestinal wall, potentially leading to inflammation, causing Crohn's disease (Fellermann et al. 2006). In contrast, duplications in β-

defensin gene cluster are associated with more activity in the cytokine, EGF-R and STAT signaling pathways in response to minor skin injury, potentially leading to psoriasis (Hollox et al. 2008).

Three evolutionary scenarios that are not necessarily mutually exclusive can be put forward to explain the ubiquity of SVs affecting the immune system. First, it is possible that the immune system SVs are a side product of recent duplication events in the human lineage and are on their way to elimination from the population through purifying selection. This scenario, however, is unlikely to explain the high allele frequency observed for several immune system SVs in contemporary human populations. Second, it has been put forward that highly variable gene families have evolved in response to changing pathogens and environments (*e.g.*, Gokcumen et al. 2011a). Third, it is possible that a balance exists between protection against pathogens and susceptibility to autoimmune disorders (*e.g.*, Machado et al. 2012). Therefore, it is likely that the SVs affecting immune system are maintained in human populations as a response to ever-changing pathogenic pressures, under a balancing/diversifying selection scenario. Overall, attempts to understand the immune system SVs have resulted in a fascinating set of questions that are yet to be answered. The dearth of phenotypic, environmental and cultural information regarding pathogenic pressure hinders the current understanding of the evolution of immune system SVs. Anthropologists are prime candidates to step in and explore.

These examples highlight three important issues for anthropological genetics regarding SVs. First, a strong case can be made for ongoing adaptive forces acting on SVs in the genomes of contemporary human populations (reviewed in Iskow et al. 2012b). Second, SVs, due to their higher mutation rate and larger phenotypic impact, are likely involved in rapid adaptations and, thus, more likely to be shaped by relatively recent, cultural trends. For instance, the agricultural transition, which is primarily a cultural change, brought with it both a completely different diet and a completely new level of population density, which leads to increased and diversified pathogenic pressures. As exemplified above, both immunity- and diet-related phenotypes are shaped partly by SVs. We think that SVs provide the most likely evolutionary fodder for adaptation to recent cultural transitions. Third, the recent association between the low copy number of *AMY1* and obesity is a clear example in which recent human evolution coincides with contemporary disease. Therefore, SVs may be the prime candidates to explore the impact of recent human evolution on contemporary disease. Overall, genomic structural variation embodies several interesting anthropological angles at the intersection of human evolution, cultural transitions and human health.

A Methodological Case Example for Genome-wide Analysis of SVs for Anthropological Hypothesis Generation

In this section, we present an example study that provides a workflow to identify genes of interest and generate hypotheses for anthropological research (Figure 1). The first issue we tackle is what we call the *"Library of Babel"* problem. That is, since the genomic variation datasets cover almost the entirety of the human reference genome, it is not clear which of the genomic variants are relevant to study to answer a particular question. In other words, the ubiquity of data may hinder the ability to form and test anthropological hypotheses. Our aim here is not necessarily to suggest a particular way of conducting such an analysis, but rather introduce the commonly used tools and datasets, as well as to provide some basic tips and warnings. We provide all the codes for conducting these analyses and generating the graphs at Dryad Digital Repository and our website (http://gokcumenlab.org/data-and-codes/). This basic workflow should familiarize anthropologists with the datasets and approaches in SV studies, which will then help narrow the search space for anthropologically meaningful genetic variation.

We arbitrarily ask the question: "What are the common deletion variants that may affect phenotypes in a continent-specific manner?" Deletion variants are currently the best studied and characterized type of structural variants, and so here we limited ourselves to this type of variation. The most accurate dataset for deletion variants that has substantial sampling is the 1000 Genomes Project (1000 Genomes Project Consortium 2012). This dataset is rapidly being updated, but for this study, we will use the consensus SV dataset of Phase 1 release (referred to as

1KG deletions hereafter), for which the results can be found at ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/consensus_call_sets/sv/.

One of the challenges for the anthropologists is that the dataset is presented in a *vcf* format (defined in Table 1), from which extraction of data requires some basic level of coding know-how. One useful tool is the vcftools (http://vcftools.sourceforge.net/). However, we provide an "Excel-friendly" version that includes the genomic intervals and continental allele frequencies for accessibility on our website (http://gokcumenlab.org/data-and-codes/). Other accessible and highly relevant SV datasets were published by Conrad et al. (2010) and Sudmant et al. (2010).

1KG deletions comprise 14,422 deletion variants compiled across 1,092 individuals from 14 populations. This is a prohibitively large dataset. Thus, our first goal was to narrow down potentially interesting SVs for our purposes. To do this, we first eliminated the SVs on sex chromosomes because the interpretation of sex chromosome SVs may be complicated due to the unique recombination landscape and changes in effective population size of these chromosomes as compared to autosomes. We then determined which of the deletions overlap with exons, as these exonic deletions are the most likely candidates to affect phenotypes. For this analysis, we used the user-friendly, web-based software *Galaxy* (Goecks et al. 2010), (https://usegalaxy.org/). Briefly, this software along

with many other useful tools includes applications to manipulate genomic intervals. In addition, *Galaxy* has the ability to directly download interval data from the *Table Browser* hosted by *UCSC Genome Browser* (Kent et al. 2002). We uploaded our interval dataset to *Galaxy* using "Upload data" function, under "Get data" tab. Please note that the upload file needs to be in the *.bed* format (defined in Table 1), which requires the first 3 columns to be "chromosome", "start position" and "end position", respectively. No headers are allowed (Please see Figure 1 for a snapshot). We also provide an example of the *.bed* file on our website (http://gokcumenlab.org/data-and-codes/) for the *1KG deletions*. Then, we downloaded "Refseq" exon intervals using *Table Browser* for the human reference genome directly from Galaxy using "UCSC Main" function under the "Get data" tab. Note that since *1KG* deletions are annotated using Hg19, it is important to use exon information from the same genome build. Once these two datasets were uploaded, we used "Intersect" function under "Operate on genomic intervals" tab to identify 1,445 autosomal deletion variants that overlap with exons. There are very helpful video tutorials available to lead through these processes at this website: http://vimeo.com/galaxyproject/videos.

Because our initial goal was to identify population-specific and potentially functional variants, we conducted one additional filtering based on continental allele frequencies. Specifically, we compared allele frequencies between African and European continents (Figure 2). We used 1000 Genomes continental

annotations (see for more detailed list http://bit.ly/1wKXe2l), which includes Luhya in Wbuye-Kenya (LWK), Yoruba in Ibadan-Nigeria (YRI) and African-American community in Southwestern United States (ASW) for Africa and Iberian populations in Spain (IBS), British in England and Scotland (GBR), Finnish in Finland (FIN), Toscani in Italy (TSI), Utah-USA residents with Northern and Western European ancestry (CEU) for Europe. The inclusion of ASW in African continent can be problematic given the considerable levels of European admixture in this population (Seldin et al. 2011). We did not include this population in downstream, locus-specific analysis below. But, in this initial filtering we used continental frequencies as defined by 1000 Genome Project.

To find variants that are common in Africa, we identified variants that are rare (<1%) or absent Europe, but had at least 10% frequency in Africa. This is a very simplistic separation. For a more sophisticated analysis a regression residue method may be used. Regardless, since the SV studies are very young, simple allele frequency differences are still unexplored indicators for identifying potentially functionally important, population-specific variants. Overall, we found 16 variable exonic deletions that had high continental differentiation (Table 2).

There are several caveats that should be noted when further analysis is designed. First, drift is a major confounding effect to take into consideration when interpreting the adaptive significance of these candidate exonic variants. Note that for this particular analysis, inclusion of ASW to the African continental group as

defined by 1000 genomes leads to slight reduction in power. Also note that some of these deletion variants are overlapping or proximate to each other, indicating either strong linkage or errors in the breakpoint assignment. These issues need to be considered when further, locus-specific analyses are conducted.

The genes that were affected by the 16 African-specific deletion variants were previously associated with important phenotypes, including sensitivity to pain (SCN9A) (Cox et al. 2006), and drug metabolism (CYP3A43) (Bigos et al. 2011), as well as olfaction (OR52E8). It is remarkable that we found through our simple workflow, two overlapping deletions that are specific to Africa, which overlap with alpha hemoglobins, HBA1 and HBA2. The deletions of these genes have been explored previously within the context of susceptibility to Thalassemia (reviewed in Weatherall 2001). Last, but not least, we found two overlapping deletions that overlap with the keratin-associated protein (KAP) gene cluster, including the genes *KRTAP9-2*, *KRTAP9-3*, *KRTAP9-8.* This gene cluster is associated with the evolution of hair in mammals (Wu et al. 2008). In the next section, we analyzed these deletion variants affecting the KAP gene cluster to give some insights into locus-specific analyses of SVs.

A Methodological Case Example for Locus-specific Analyses of SVs

Among the candidate genes that we highlighted in the previous section, we focused on two overlapping deletion variants on the long arm of chromosome 17. The larger of the two deletions (KAP-Del1) encompasses the smaller one (KAP-Del2) and overlaps with three members of KAP gene cluster (*KRTAP9-2, KRTAP9-3, KRTAP9-8*), while the smaller deletion overlaps with only *KRTAP9-8*. These genes interact with hair keratins to form rigid and resistant hair shafts (reviewed in Rogers et al. 2006). This extensive gene family has been argued to evolve under rapid, diversifying selection among mammals as a response to ever-changing ecological pressures regarding hair formation (Wu et al. 2008).

This dynamic evolutionary history made KAP-Del1 and KAP-Del2 prime candidates for further inquiry. First, hair formation is an important trait in primates involving retention of heat, as well as sexual selection (Schwartz and Rosenblum 1981). In fact, it has been documented that the hair phenotype in humans varies significantly and is occasionally positively selected (*e.g.*, Fujimoto et al. 2008, Kamberov et al. 2013). Moreover, hair patterns can be important in cultural constructs, such as perception of beauty or signifiers. Second, the large tandem repeats that are formed by the KAP gene clusters increase the occurrence of non-allelic homologous recombination-based SV formation, potentially leading to further gene losses and gains. Therefore, deletion variants affecting the KAP gene cluster, including KAP-Del1 and KAP-Del2, remain viable candidates to understand phenotypic variation in hair patterns among modern humans.

We first investigated the geographic distribution of KAP-Del1 and KAP-Del2 among human populations using *1KG* deletion genotypes. For such an analysis, the individual genotypes for specific deletion variants can be found either through *UCSC genome browser*, by simply clicking over the variant of interest using "1000G Ph1 Vars" track. This information can also be reached using the 1000 Genomes Browser (http://browser.1000genomes.org/index.html). Our results showed that individuals who have either of the deletion variants are common in the African continent, but absent in non-African populations (Figure 3). The small fraction of individuals that carry the deletion variants in the Americas likely represent recent gene flow from Africa. Given their potential functional relevance and their geographic confinement to Africa, these deletions may explain some hair phenotypes that are observed only among people with ancestry in Africa (*e.g.*, Loussouarn 2001).

One immediate methodological observation was that there are samples that show homozygous deletions for both KAP-Del1 and KAP-Del2. Because these deletion variants are overlapping, this observation indicates potential genotyping errors in the 1000 Genomes Project. This led us to the next crucial step in studying such gene deletions, which is to develop a locus-specific validation/genotyping method. This is key because the false-positive rates involving structural variation discovery is much higher than those for single nucleotide variants (Mills et al. 2011). Moreover, even if the SV discovery is

accurate, the breakpoints and genotyping may suffer from inconsistencies, especially in complex regions with high repeat content. As such, validation and correct annotation of the breakpoints of SVs are important steps to ensure the healthy progress of the project in later stages.

For gene deletions, such as those overlapping KAP genes, the most straightforward method is to design a polymerase chain reaction (PCR)–based experiment with forward and reverse primers targeting upstream and downstream of the deletions (see Figure 1). For gene duplications, the validation and breakpoint localization is more complicated. However, quantitative and more recently digital PCR-based methods have been shown to be effective in analyzing the copy number variation of multi-copy genes. Inversions and translocations can be validated and genotyped by simple PCR-based methods. However, since the discovery tools for these types of variants are mired with false-positives and other accuracy issues, the validation can be more complicated. Therefore, deletion variants are currently the best-studied and easiest targets for anthropological genetics studies.

The deletion validation method can be used for genotyping as well. Essentially, the PCR product of the sequences carrying the deletion will be smaller than the product of the sequences that do not carry the deletion. As such, for a small deletion variant, it is possible to detect homozygous and heterozygous deletion in a single PCR-based experiment. If the deletion event is large (*i.e.*,

>5kb), the non-deleted sequences will be too large to amplify in a traditional PCR reaction. Thus, under these circumstances PCR would work only when the sequence carries the deletion. To avoid false-negatives, it would be important to add a parallel primer set for the larger events with the primers targeting the sequence within the deletion polymorphism. Regardless, this simple PCR-based genotyping method provides an affordable and straightforward way to study specific SVs of interest in a new population.

Deletion variants are mostly single mutational events. Still, major trends can be observed by comparing allele frequencies across populations using a simple version of F_{ST} (Hudson et al. 1992). This is a relatively simple measure of among population variation as normalized by within population variation and can be calculated using the allele frequencies of a single variant (see Table 1 for a more thorough explanation). Indeed, when we observed the F_{ST} between YRI and CEU populations across chromosome 17, we found that the two deletions that affect the KAP cluster indeed differentiated from other deletions and have the highest F_{ST} among other exonic deletions on this chromosome (Figure 4).

It is also possible to leverage 1000 Genomes sequencing data to extract single nucleotide variation information flanking these deletion variants and use population genetic analysis on the haplotypes surrounding the SVs (*e.g.*, Gokcumen et al. 2013b; Xue et al. 2008). For such an analysis, the manipulation of the *vcf* files is necessary, and we provide a simple code for creating FASTA

files for use with the programs, such as MEGA or dbSNP on our website (http://gokcumenlab.org/data-and-codes/). Once the single nucleotide variation flanking information is extracted (either upstream or downstream or both), it is possible to look into several neutrality tests and population differentiation. For instance, in a quick look, we found that the African populations consistently have lower Tajima's D than those in Europe and Asia in the upstream of KAP-Del2 (Table 3). Negative Tajima's D mean that there are less than expected average pairwise differences, potentially indicating a selective sweep removing most of the variation. However, Tajima's D is highly susceptible to demographic trends, such as population bottlenecks. Additional tests need to be conducted, especially by separating haplotypes for those that carry the deletion variation and those that do not. Overall, the neutrality analysis for KAP-Del2 remains inconclusive. However, higher than expected population differentiation and potential phenotypic impact distinguish KAP-Del1 and KAP-Del2 as two interesting structural variants for further analysis.

To further understand whether the adaptive potential and phenotypic impact of these gene deletions exists, anthropologists can take two approaches. First, we can focus on the allele frequency distribution of this gene among human populations and compare it to a particular ecological, phenotypic or other adaptive trend. This approach was used effectively for the copy number of salivary amylase (Perry et al. 2007), showing that high starch–consuming populations

have higher copy numbers of the *AMY1* gene regardless of their ancestral origins. Another powerful but costly approach would be a within-population association study, showing a particular phenotypic variation is associated with the loss of one of the KAP genes. A relevant example can be found in the association of the losses of the *LCE3B* and *LCE3C* genes with psoriasis prevalence in European population (de Cid et al. 2009). Overall, our aim here is not to resolve the evolution of KAP-Del1 and KAP-Del2, which may be of great interest to anthropologists that work on evolution of hair patterns within and across human populations. Rather, we hope that in this section we were able to introduce some of the basic tenets, methodologies and pitfalls in studying SVs in a locus-specific manner.

A Brief Note on Discovery of New Structural Variants and Working With Duplications

Before concluding, we also want to touch shortly on methodological issues involving SV discovery studies, as well as genotyping of duplications. In addition to bioinformatics analysis of next-generation sequencing data, other major techniques available for genome-wide SV discovery are array-based platforms, such as array comparative genomic hybridization and single channel array platforms that were originally designed for single nucleotide variant genotyping. All these platforms and approaches suffer from major shortcomings. For instance,

array based platforms are highly dependent on the performance of the probes that are used to discover SVs and they are not able to detect balanced variants, such as inversions and translocations. They also have limitations in detecting below a certain size of SV depending on the number of probes presented on the array. Pinto et al. (2011) provides a detailed platform comparison of array-based methods for discovery of copy number variants.

Short reads from next generation sequencing technologies have been successfully leveraged to discover all types of structural variants. However, the approach (*e.g.*, deviations in read-depth, deviations from expected paired-end mapping, split-read mapping, etc.) and the specific algorithms that are used for leveraging these approaches give strikingly different outcomes in the size, types and genomic location of the variants detected. Mills et al. (2011) and Alkan et al. (2011) provide good summaries of these approaches and algorithms. Overall, genome-wide discovery of SVs remain expensive, time-consuming, requires substantial computational know-how and still remain highly prone to sensitivity and accuracy issues (*i.e.*, high levels of false-positives and false-negatives). As such, we believe it is relatively unfeasible to design studies involving genome-wide discovery of structural variation for genetic anthropologists in the near future.

Duplications may be single most important genetic variation category for studies of human adaptive evolution (see Iskow et al. 2012b for a detailed

review). Moreover, initial attempts discovered hundreds of variable gene duplications in humans (*e.g.*, Sudmant et al. 2010). In fact, recently released Phase 3 dataset of 1000 genomes comprises a much-improved set of variable duplications among human populations (http://bit.ly/1tltt7j). However, the approaches for discovery, localization and genotyping of duplications remain less developed than those designed for deletions. Still, with the advent of the next generation of locus-specific quantification methods, such as droplet digital PCR, it is plausible to design genetic anthropology studies, similar in design to experiments suggested for deletions above, to tackle questions involving multi-copy gene duplications. Boettger et al. (2012) provides an excellent recent example of such an approach, resolving the duplication events accompanying a major inversion polymorphism that is common in Europeans.

An Integrative Anthropological Perspective to Study SVs

One important side note about SVs and especially about variation involving multi-copy gene families is that SVs affect traits that are related to environmental interactions. There may be a common evolutionary process through which this variation has evolved. Mechanistically, a gene duplication event, by creating homologous sequences, facilitates formation of further structural variations through homology-based recombination or replication errors (reviewed in Feuk et al. 2006). Functionally, environmental-interaction genes are under ever-changing

adaptive pressures. For instance, olfactory receptor genes, which are affected by SVs significantly more than the genome average (Hasin et al. 2008; Young et al. 2008), are under reduced purifying selection in the primate lineage, allowing the variation within these genes to be tolerated in primate species (Gilad et al. 2004). As a consequence, humans may be randomly losing olfactory receptor gene functions due to deletions (or other types of variants) without major adaptive backlash, increasing the phenotypic variation in the population along the way. Overall, there is ample evidence suggesting that mechanistic and adaptive forces are affecting multi-gene families that are involved in environmental interactions (*e.g.,* sense reception, immunity, xenobiotic metabolism, etc.), leading to an increased rate of structural variation among humans.

SVs affecting the KAP gene family can also be considered within the framework explained above. It is plausible that hair patterns in humans, a highly variable trait, evolved under reduced purifying selection, akin to olfactory receptors in humans. Under this scenario, the hair patterns, controlled by hundreds of genes including the members of KAP gene family evolved as a response to particular adaptive stress, which may not be as strong for humans anymore due to ecological or cultural (*e.g.*, technological) change. As such, the loss-of-function variants, such as the KAP-Del1 and KAP-Del2, may be tolerated for some members of the KAP gene family and in some human groups, but not in others. Under this scenario, random phenotypic variation would be created depending on

which of these gene functions are lost. The deletions KAP-del1 and KAP-del2 may exactly be such loss-of-function variants without major adaptive consequences, but with potential phenotypic consequences. Thus, it will be important to conduct studies on the local context along with phenotypic data to associate these deletions to certain hair features. In addition, it will be interesting further investigate the distribution of these deletion variants in southeastern Asia, India, Central Asia and Australia (See Figure 3).

This is also a good point to mention that hair patterns may have symbolic prominence from one culture to another. It is tempting to argue that a phenotypic difference that may not be crucial in human adaptation from an ecological point of view may be important in shaping the physical attributes of people in a population-specific manner. It is, therefore, tempting to consider the implications of such phenotypic variation for formation of complex cultural concepts, *e.g.*, beauty, attractiveness, gender, group-identity. Such a scenario would explain the observed population differentiation affecting the KAP-cluster. Regardless, SVs affecting multi-copy gene families are of major interest to anthropological research because of the mechanistic ease of creating new variation involving homologous sequences and because of the ontological relevance to "external" traits.

It is an exciting time for both anthropological genetics and genomics. New technologies are paving the way to understand the functional consequences of

genetic variation at an unprecedented pace. Furthermore, it is now better appreciated that "common disease-common variant" framework is not adequate for elucidating the relationship between the genotype and phenotype. Instead, variation at the local scale (Williams et al. 2014), complex interactions of multiple variants (Purcell et al. 2014) and, as mentioned in this review, non-traditional genetic markers, such as SVs, are considered to explain important phenotypic variations. The locus-specific, anthropologically contextualized studies of SVs are fertile grounds for addressing novel and extremely high-impact questions about the genetic bases of human phenotypic variation and its consequences.

Acknowledgements

We thank Austin Reynolds and Kristina Wasson-Blader for insightful comments and corrections in previous versions of this manuscript.

Literature Cited

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

Alkan, C., B.P. Coe, and E.E. Eichler. 2011. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* 12:363–376.

Bailey, J.A., Z. Gu, R.A. Clark et al. 2002. Recent segmental duplications in the human genome. *Science* 297:1003–1007.

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

Bigham, A., M. Bauchet, D. Pinto et al. 2010. Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. *PLoS Genet.* 6:e1001116.

Bigos, K.L., R.R. Bies, B.G. Pollock et al. 2011. Genetic variation in CYP3A43 explains racial difference in olanzapine clearance. *Mol. Psychiatry* 16:620–625.

Boettger, L.M., R.E. Handsaker, M.C. Zody et al. 2012. Structural haplotypes and recent evolution of the human 17q21.31 region. *Nat. Genet.* 44:881–885.

Campbell, C.D., and E.E. Eichler. 2013. Properties and rates of germline mutations in humans. *Trends Genet.* 29:575–584.

Campbell, M.C., A. Ranciaro, A. Froment et al. 2012. Evolution of functionally diverse alleles associated with PTC bitter taste sensitivity in Africa. *Mol. Biol. Evol.* 29:1141–1153.

Cann, R.L., M. Stoneking, and A.C. Wilson. 1987. Mitochondrial DNA and human evolution. *Nature* 325:31–36.

Chaix, R., L. Quintana-Murci, T. Hegay et al. 2007. From social to genetic structures in central Asia. *Curr. Biol.* 17:43–48.

Charrier, C., K. Joshi, J. Coutinho-Budd et al. 2012. Inhibition of SRGAP2

function by its human-specific paralogs induces neoteny during spine maturation. *Cell* 149:923–935.

Conrad, D., D. Pinto, R. Redon et al. 2010. Origins and functional impact of copy number variation in the human genome. *Nature* 464:704–712.

Cox, J.J., F. Reimann, A.K. Nicholas et al. 2006. An SCN9A channelopathy causes congenital inability to experience pain. *Nature* 444:894–898.

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

Dennis, M.Y., X. Nuttle, P.H. Sudmant et al. 2012. Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. *Cell* 149:912–922.

Derti, A., F.P. Roth, G.M. Church et al. 2006. Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. *Nat. Genet.* 38:1216–1220.

de Wijk, R.A., J.F. Prinz, L. Engelen et al. 2004. The role of alpha-amylase in the perception of oral texture and flavour in custards. *Physiol. Behav.* 83:81–91.

Dos Santos, C., L. Essioux, C. Teinturier et al. 2004. A common polymorphism of the growth hormone receptor is associated with increased responsiveness to growth hormone. *Nat. Genet.* 36:720–724.

ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74.

Falchi, M., J.S. El-Sayed Moustafa, P. Takousis et al. 2014. Low copy number of the salivary amylase gene predisposes to obesity. *Nat. Genet.* 46:492–497.

Fellermann, K., D.E. Stange, E. Schaeffeler et al. 2006. A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. *Am. J. Hum. Genet.* 79:439–448.

Feuk, L., A.R. Carson, and S.W. Scherer. 2006. Structural variation in the human genome. *Nat. Rev. Genet.* 7:85–97.

Fujimoto, A., R. Kimura, J. Ohashi et al. 2008. A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. *Hum. Mol. Genet.* 17:835–843.

Gilad, Y., M. Przeworski, and D. Lancet. 2004. Loss of olfactory receptor genes coincides with the acquisition of full trichromatic vision in primates. *PLoS Biol.* 2:E5.

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

Gokcumen, O., P.L. Babb, R.C. Iskow et al. 2011a. Refinement of primate copy number variation hotspots identifies candidate genomic regions evolving

under positive selection. *Genome Biol.* 12:R52.

Gokcumen, O., T. Gultekin, Y.D. Alakoc et al. 2011b. Biological ancestries, kinship connections, and projected identities in four central Anatolian settlements: insights from culturally contextualized genetic anthropology. *Am. Anthropol.* 113:116–131.

Gokcumen, O., V. Tischler, J. Tica et al. 2013a. Primate genome architecture influences structural variation mechanisms and functional consequences. *Proc. Natl. Acad. Sci. USA* 110:15764–15769.

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

Green, R.E., J. Krause, A.W. Briggs et al. 2010. A draft sequence of the Neandertal genome. *Science* 328:710–722.

Hammer, M.F. 1995. A recent common ancestry for human Y chromosomes. *Nature* 378:376–378.

Hasin, Y., T. Olender, M. Khen et al. 2008. High-resolution copy-number variation map reflects human olfactory receptor diversity and evolution. *PLoS Genet.* 4:e1000249.

Hasin-Brumshtein, Y., D. Lancet, and T. Olender. 2009. Human olfaction: from genomic variation to phenotypic diversity. *Trends Genet.* 25:178–184.

Hollox, E., J. Barber, A. Brookes et al. 2008. Defensins and the dynamic genome:

what we can learn from structural variation at human chromosome band 8p23.1. *Genome Res.* 18:1686–1697.

Hollox, E.J., and B-P. Hoh. 2014. Human gene copy number variation and infectious disease. *Hum. Genet.* 133:1217-1233.

Holsinger, K., and S.W. Weir. 2009. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nat. Rev. Genet.* 10:639-650.

Hu, Y., Y. Wang, Q. Ding et al. 2014. Genome-wide Scan of Archaic Hominin Introgressions in Eurasians Reveals Complex Admixture History. *arXiv [q-bioPE] [Internet]*. Available from: http://arxiv.org/abs/1404.7766

Hudson, R.R., D.D. Boos, and N.L. Kaplan. 1992. A statistical test for detecting geographic subdivision. *Mol. Biol. Evol.* 9:138–151.

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

Iafrate, A.J., L. Feuk, M.N. Rivera et al. 2004. Detection of large-scale variation in the human genome. *Nat. Genet.* 36:949–951.

Iskow, R.C., O. Gokcumen, A. Abyzov et al. 2012a. Regulatory element copy number differences shape primate expression profiles. *Proc. Natl. Acad. Sci. USA* 109:12656–12661.

Iskow, R.C., O. Gokcumen, and C. Lee. 2012b. Exploring the role of copy number variants in human adaptation. *Trends Genet.* 28:245–257.

Jacquemont, S., A. Reymond, F. Zufferey et al. 2011. Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. *Nature* 478:97–102.

Kamberov, Y.G., S. Wang, J. Tan et al. 2013. Modeling recent human evolution in mice by expression of a selected EDAR variant. *Cell* 152:691–702.

Karlsson, E.K., D.P. Kwiatkowski, and P.C. Sabeti. 2014. Natural selection and infectious disease in human populations. *Nat. Rev. Genet.* 15:379–393.

Kent, W.J., C.W. Sugnet, T.S. Furey et al. 2002. The human genome browser at UCSC. *Genome Res.* 12:996–1006.

Kidd, J.M., T.L. Newman, E. Tuzun et al. 2007. Population stratification of a common APOBEC gene deletion polymorphism. *PLoS Genet.* 3:e63.

Laland, K.N., J. Odling-Smee, and S. Myles. 2010. How culture shaped the human genome: bringing genetics and the human sciences together. *Nat. Rev. Genet.* 11:137–148.

Lao, O., T.T. Lu, M. Nothnagel et al. 2008. Correlation between genetic and geographic structure in Europe. *Curr. Biol.* 18:1241–1248.

Lappalainen, T., M. Sammeth, M.R. Friedländer et al. 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501:506–511.

Loussouarn, G. 2001. African hair growth parameters. *Br. J. Dermatol.* 145:294–297.

Machado, L.R., R.J. Hardwick, J. Bowdrey et al. 2012. Evolutionary history of copy-number-variable locus for the low-affinity Fcγ receptor: mutation rate, autoimmune disease, and the legacy of helminth infection. *Am. J. Hum. Genet.* 90:973–985.

McCarroll, S.A., A. Huett, P. Kuballa et al. 2008. Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat. Genet.* 40:1107–1112.

McLean, C.Y., P.L. Reno, A.A. Pollen et al. 2011. Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* 471:216–219.

Menozzi, P., A. Piazza, and L. Cavalli-Sforza. 1978. Synthetic Maps of Human Gene Frequencies in Europeans. *Science* 201:786–792.

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

Norton, H.L., R.A. Kittles, E. Parra et al. 2007. Genetic evidence for the convergent evolution of light skin in Europeans and East Asians. *Mol. Biol. Evol.* 24:710–722.

Novembre, J., T. Johnson, K. Bryc et al. 2008. Genes mirror geography within Europe. *Nature* 456:98–101.

Perry, G., N. Dominy, K. Claw et al. 2007. Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.* 39:1256–1260.

Pinto, D., K. Darvishi, X. Shi et al. 2011. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat. Biotechnol.* 29:512–520.

Pruitt, K.D., T. Tatusova, and D.R. Maglott. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucl. Acids Res.* 33:D501-D504.

Purcell, S.M., J.L. Moran, M. Fromer et al. 2014. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* 506:185–190.

Ralph, P., and G. Coop. 2013. The geography of recent genetic ancestry across Europe. *PLoS Biol.* 11:e1001555.

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

Rogers, M.A., L. Langbein, S. Praetzel-Wunder et al. 2006. Human hair keratin-associated proteins (KAPs). *Int. Rev. Cytol.* 251:209–263.

Rosenberg, N.A. 2011. A population-genetic perspective on the similarities and differences among worldwide human populations. *Hum. Biol.* 83:659–684.

Schwartz, G.G., and L.A. Rosenblum. 1981. Allometry of primate hair density and the evolution of human hairlessness. *Am. J. Phys. Anthropol.* 55:9–12.

Sebat, J., B. Lakshmi, D. Malhotra et al. 2007. Strong association of de novo copy number mutations with autism. *Science* 316:445–449.

Sebat, J., B. Lakshmi, J. Troge et al. 2004. Large-scale copy number

polymorphism in the human genome. *Science* 305:525–528.

Seielstad, M.T., E. Minch, and L.L. Cavalli-Sforza. 1998. Genetic evidence for a higher female migration rate in humans. *Nat. Genet.* 20:278–280.

Seldin, M.F., B. Pasaniuc, and A.L. Price. 2011. New approaches to disease mapping in admixed populations. *Nat. Rev. Genet.* 12:523–528.

Stefansson, H., D. Rujescu, S. Cichon et al. 2008. Large recurrent microdeletions associated with schizophrenia. *Nature* 455:232–236.

Stranger, B.E., M.S. Forrest, M. Dunning et al. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 315:848–853.

Sudmant, P.H., J.O. Kitzman, F. Antonacci et al. 2010. Diversity of human copy number variation and multicopy genes. *Science* 330:641–646.

Sulem, P., D.F. Gudbjartsson, S.N. Stacey et al. 2007. Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat. Genet.* 39:1443–1452.

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

Tishkoff, S.A., F.A. Reed, F.R. Friedlaender et al. 2009. The genetic structure and history of Africans and African Americans. *Science* 324:1035–1044.

Tishkoff, S.A., F.A. Reed, A. Ranciaro et al. 2007. Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* 39:31–40.

Traherne, J.A., M. Martin, R. Ward et al. 2010. Mechanisms of copy number

variation and hybrid gene formation in the KIR immune gene complex. *Hum. Mol. Genet.* 19:737–751.

Underhill, P.A., and T. Kivisild. 2007. Use of y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annu. Rev. Genet.* 41:539–564.

Veeramah, K.R., and M.F. Hammer. 2014. The impact of whole-genome sequencing on the reconstruction of human population history. *Nat. Rev. Genet.* 15:149–162.

Weatherall, D.J. 2001. Phenotype-genotype relationships in monogenic disease: lessons from the thalassaemias. *Nat. Rev. Genet.* 2:245–255.

Weischenfeldt, J., O. Symmons, F. Spitz et al. 2013. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* 14:125–138.

Williams, A.L., S.B.R. Jacobs, H. Moreno-Macías et al. 2014. Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. *Nature* 506:97–101.

Wu, D-D., D.M. Irwin, and Y-P. Zhang. 2008. Molecular evolution of the keratin associated protein gene family in mammals, role in the evolution of mammalian hair. *BMC Evol. Biol.* 8:241.

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

Young, J.M., R.M. Endicott, S.S. Parghi et al. 2008. Extensive copy-number variation of the human olfactory receptor gene family. *Am. J. Hum. Genet.* 83:228–242.

Zerjal, T., Y. Xue, G. Bertorelle et al. 2003. The genetic legacy of the Mongols. *Am. J. Hum. Genet.* 72:717–721.

Zhang, J. 2003. Evolution by gene duplication: an update. *Trends Ecol. Evol.* 18:292–298.

Table 1. Glossary of important terms

Term	Description	More In Depth Source
vcf	Variant Call Format (*vcf*) is a tab-delimited file format used by the 1000 Genomes project for storing all variant calls from single nucleotide variants to large-scale structural variants like insertions and deletions, and individual genotypes.	http://bit.ly/1xEp4vn
bed	A BED file (*.bed*) is a tab-delimited text file that defines a feature track. It is essentially a coordinate system depicting the chromosome, start position and end position of a particular feature in the genome (e.g., genes, exons, structural variants, etc.). See Figure 1 for an example.	http://bit.ly/1thlsQP
RefSeq Genes	RefSeq is a publicly available curated reference sequence database, which include, among others, the coding sequences of known genes. This database	Pruitt et al. 2005 describes the database in detail. For more detailed description - http://1.usa.gov/1wFOI4x

	can be accessed as an annotation track in UCSC genome browser.	
Digital PCR	Digital droplet PCR is an accurate, micro-fluidic alternative to conventional quantitative PCR methods. This method involves partitioning a fluorescently labeled PCR reaction into thousands (or millions depending on the technology) of nano-droplets, with the limiting factor being the target DNA (or RNA). The droplets with the target DNA will undergo a PCR, which in turn release a fluorescent signal. The numbers of droplets that give the fluorescent signal directly correlate with the number of target DNA molecules in the sample. It is essentially more accurate and reliable than quantitative PCR for copy number estimation.	For a great example for the application of digital PCR to complicated locus with multiple structural variants, please see Boettger et al. 2012
F_{ST}	Fst is a measure of population divergence. In its basic form it can be summarized as	Please see Holsinger and Weir 2009 for a thorough

	the proportion of among population variation to within population variation. It provides a measure of population differentiation taking into account the variation that exists within populations.	discussion of F-statistics applied to population genetics
Tajima's *D*	Tajima's D is a population level statistical test of neutrality of a genetic region. It compares the estimated nucleotide diversity as expected from the observed number of segregating sites to the observed nucleotide diversity. From another perspective, it measures the deviations from the allele frequency spectrum. For instance, purifying selection or a complete sweep (also recent population expansion) would predict an excess of low frequency alleles in the population. This excess would be reflected in a Tajima's D measure as negative values.	Tajima 1989 For a great video by Dr. Mohammed Noor from Duke University explaining the basics of Tajima's D, please see: http://bit.ly/1thyZrE

Table 2. Exonic deletion variants that show high continental allele frequency differentiation

Chromosome	Start	End	Gene
chr1	16809863	16811594	CROCCP3
chr2	167155679	167158793	LOC101929680, BC051759, SCN9A
chr7	99461389	99463562	CYP3A43
chr8	66091801	66094658	LINC00251
chr10	48664001	48870900	PTPN20B, FRMPD2P1, FRMPD2B
chr11	5873984	5883493	OR52E8
chr15	25464532	25466678	SNORD115 family, SNURF-SNRPN
chr15	25472291	25477569	SNORD115 family, SNURF-SNRPN
chr16	220501	227650	HBA1, HBA2
chr16	90131010	90137150	PRDM7
chr17	39383401	39395500	KRTAP9-2, KRTAP9-3, KRTAP9-8*** (KAP)**
chr17	39391001	39395500	KRTAP9-8*** (KAP)**
chr19	50553601	50560900	FLJ26850
chr21	47609742	47610876	LSS
chr22	44564501	44566100	PARVB
chr6	3195393	3195954	LOC100507194

***** All the three KRTAP9-2, KRTAP9-3, KRTAP9-8 genes belong to the KAP gene cluster

Table 3. Tajima's D values for the upstream of KAP-Del2 breakpoints in 1000 Genomes populations

Population	Population Description	Tajima's D
LWK[a]	Luhya	-1.06788
YRI[a]	Yoruba	-1.2103
CLM[b]	Colombian	-1.44763
PUR[b]	Puerto Rican	-1.63895
MXL[b]	Mexican (from LA)	-1.46258
JPT[c]	Japanese	-0.209768
CHS[c]	Chinese (South)	-0.303256
CHB[c]	Chinese (Beijing)	0.197736
IBS[d]	Iberian	-0.195583
GBR[d]	British	-0.628362
FIN[d]	Finnish	-0.589009
TSI[d]	Toscan	-0.306945
CEU[d]	European (Utah, USA)	-0.258947

[a] African; [b] Americas; [c] Asian; [d] European

Figure 1. Suggested workflow for anthropological studies of genomic structural variation.

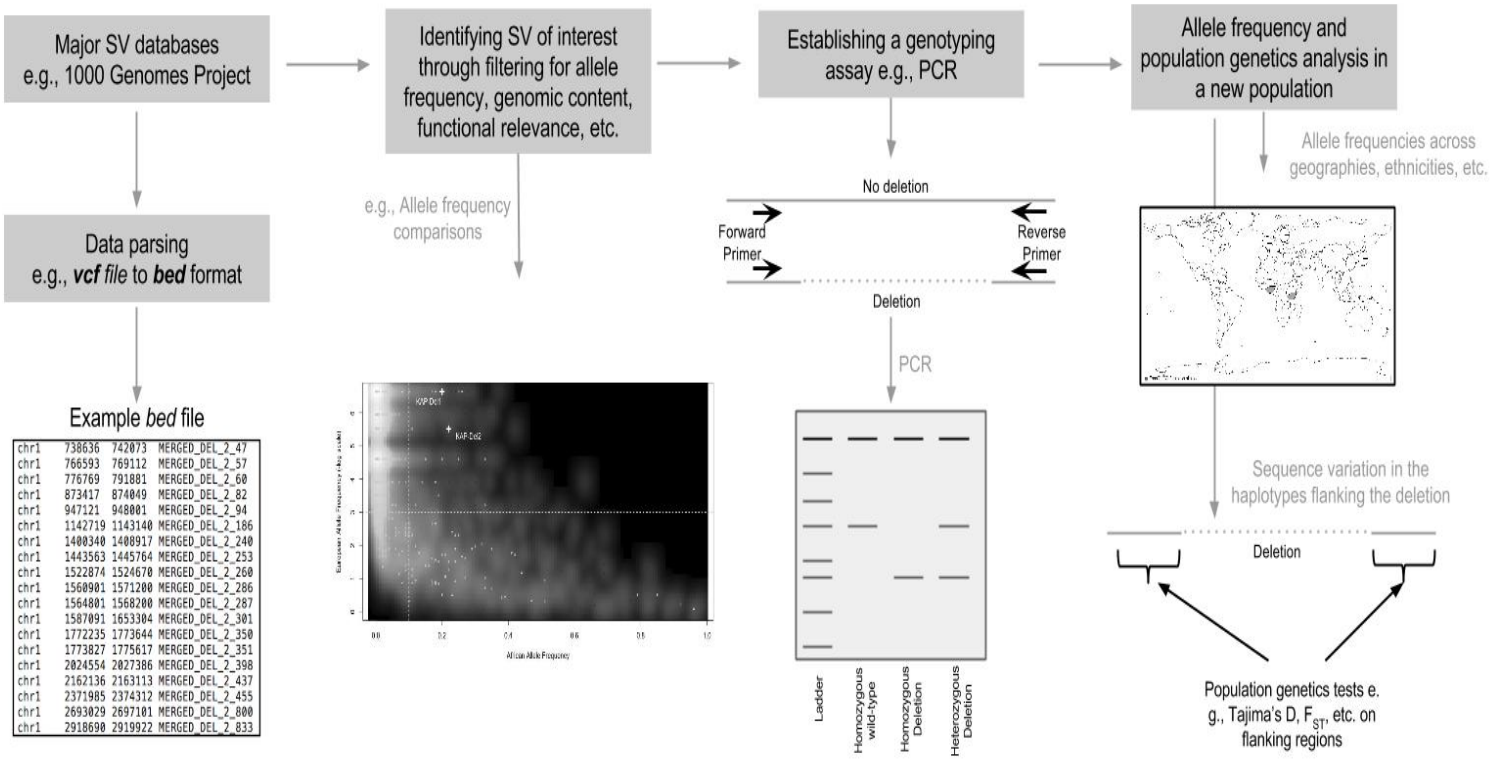

Figure 2. The allele frequency of all *1KG* deletions in African and European populations, highlighting common African alleles that are absent or rare in European populations. The x-axis shows the allele frequency of deletion variants in Africa. The y-axis shows European allele frequency in –log scale so that the lowest frequency alleles separate from the rest. The background "gray" clouds represent all deletion variants. The white cloud dots in lower frequency areas (upper left) indicate thousands of variants; whereas, light gray bubbles indicate single variants. The gray circles indicate allele frequencies of exonic deletion variants. The white stars are the KAP-Del1 and KAP-Del2.

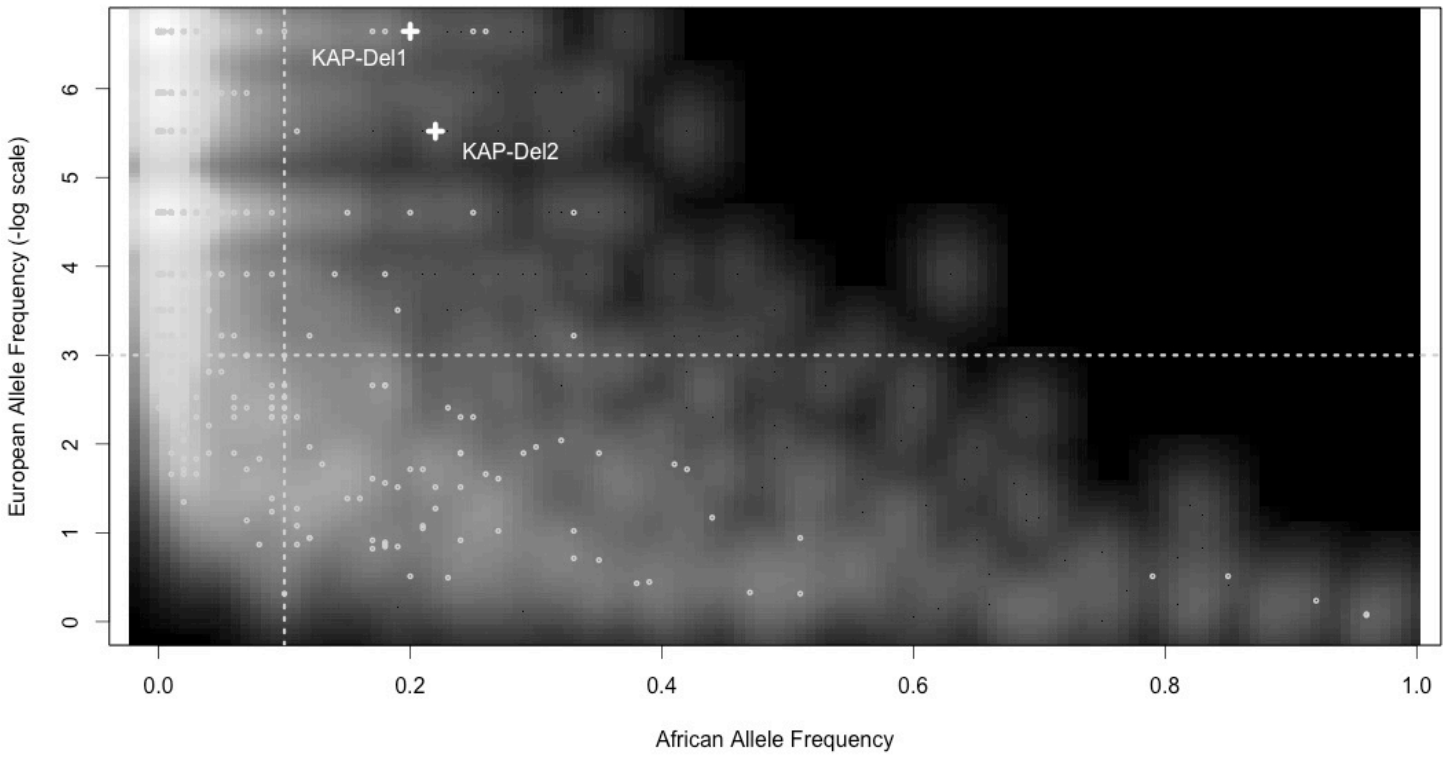

Figure 3. The frequency of individuals carrying either the KAP-Del1 or KAP-Del2 deletion variants. This is not an allele-frequency map, but identifies individuals that carry variants with potential impact on the KAP gene family function.

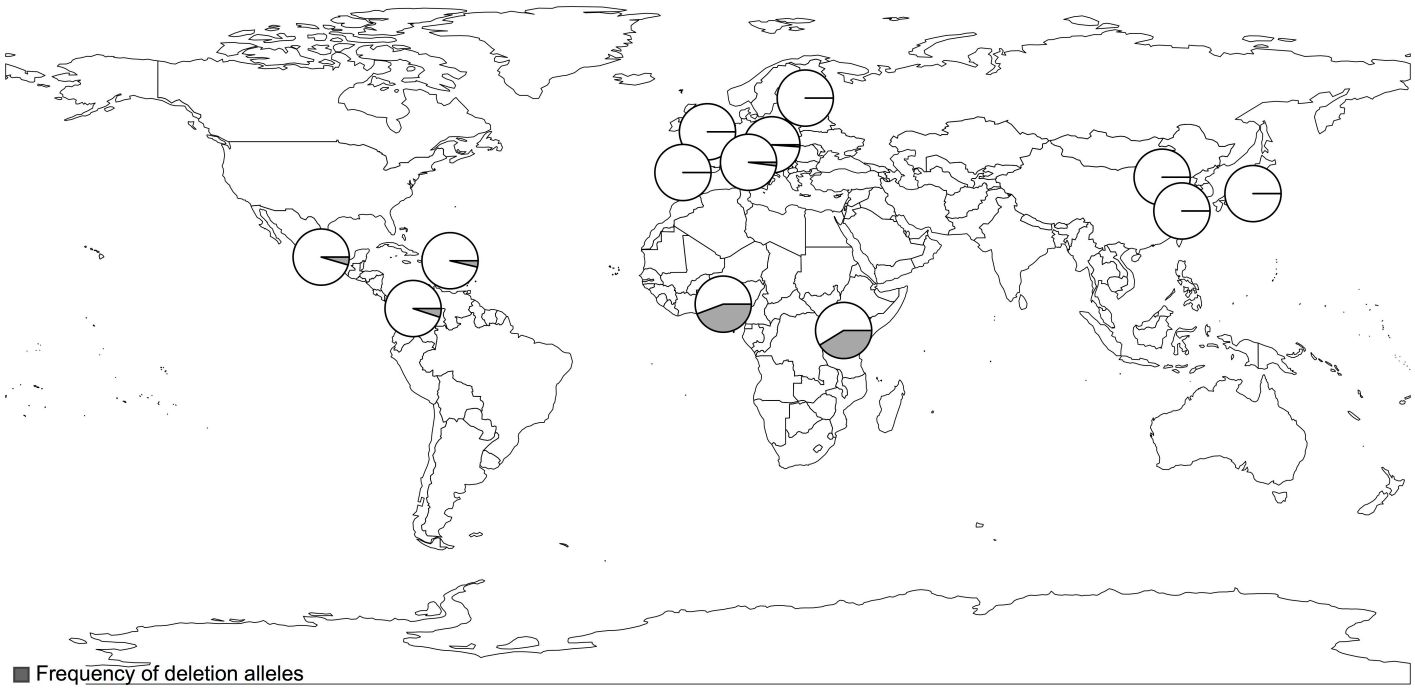

Figure 4. The graph shows the population differentiation between YRI and CEU populations (F_{ST}) across chromosome 17 coordinates. Y-axis shows the F_{ST} values calculated based on allele frequencies. X-axis are the start positions of the deletions based on HG19 version of the human reference genome on chromosome 17. The open gray circles indicate non-exonic deletions, filled gray circles are exonic deletions and black triangles are the two deletions overlapping KAP family of genes. The horizontal dotted line indicates the 95th percentile point when the distribution of all variants is considered.