Addition and Subtraction in the Out of Africa Bottleneck

Delaney Keener
Department of Anthropology
University of New Mexico

Abstract

The Out-of-Africa (OOA) migration marks a pivotal event in human evolution, shaping the genetic landscape of modern populations. This research employs R Programming to analyze DNA sequences from the Thousand Genomes Project, focusing on the genetic consequences of the OOA migration. Our investigation reveals a distinct pattern of genetic variation, with African populations harboring a multitude of common alleles absent or rare outside Africa, and vice versa. Using Group Specific Polymorphisms (GSPs), we identify alleles with high frequency within one geographic region and low frequency in another, shedding light on genetic losses and gains during the OOA migration. Using a sliding window approach, we analyzed chromosome 1, identifying 6,337 GSPs within African populations and 1,851 outside Africa. The observed clustering pattern supports the notion of substantial losses during the migration, and less frequent gains in bottleneck populations. The prevalence of GSPs in African populations signifies an abundance of common variants in ancestral populations, while the smaller and less frequent GSP clusters in non-African populations align with previously proposed founder effects. Our results highlight a higher frequency of losses than gains, pointing towards genetic drift as a significant force during the migration resulting in genetic losses. However, the smaller gains we observe are indicative of natural selection acting over thousands of years. Our findings provide novel insights into the genomic architecture shaped by the OOA migration, emphasizing the impact of both drift and selection on the genetic diversity of modern human populations.

Introduction

When we analyze the structure of genetic variation in human populations by geographic region, we see that an allele that is common in one region is usually common in other regions (Long, Li, and Healy 2009). However, it is well known that people living outside of Africa harbor less variation than African people. Moreover, the common alleles carried by non-Africans tend to be a subset of the common alleles carried by Africans (Long, Li, and Healy 2009; Yu et al. 2002). Researchers believe that a series of ancient founder effects shaped this pattern of genetic variation (Ramachandran et al. 2005; Auton et al. 2015). These founder effects occurred as modern humans migrated out of Africa and filled the remaining habitable continents. The earliest population of modern Homo sapiens lived about 300,000 years ago in Sub-Saharan Africa. The first major founder effect occurred when a small population left Africa sometime between 50,000 and 100,000 years ago. All modern non-Africans carry a major component of ancestry from this population. Non-Africans also carry minor components of archaic hominin ancestry. Population geneticists refer to this migration event as the Out-of-Africa migration (OOA). Analyses of microsatellite loci indicate that the OOA migration created a population bottleneck is consistent with an effective population size of 1,300 people persisting for 5,000 years before expanding and diversifying into modern Eurasians (Niedbalski and Long 2022).

Recent investigations of whole genome sequences reveal a large number of common alleles that are found in African people and are rare or absent outside of Africa. Although these alleles number in the thousands, they are a small percentage of the common alleles in African people. Similarly, there are large number of common alleles that are found in non-African people, but rare or absent in African people. Once again, these alleles number in the thousands, but they are a small percentage of the common alleles in non-African people. A feature of the non-African alleles is that they are common throughout Eurasia. They are not restricted to continental populations or major regions within Eurasia, *e.g.,* Europe, East Asia, etc. (Auton et al. 2015; Niedbalski and Long 2022).

Enhanced genetic drift caused by bottlenecks typically reduces variation in a population and provides a descendant with a subset of the variation that was present in their ancestors. However, genetic drift can also elevate the frequencies of rare alleles and new mutations (Nei 1987). This gain of alleles is less common than the loss of alleles but can be important in the overall diversity in a population. In this light, we propose that the OOA bottleneck is responsible for the appearance of both the African specific alleles and the non-African specific alleles. The African specific alleles appear because the OOA founders lost them in the early phase of their evolution outside of Africa. By contrast, the non-African specific alleles are the ones that rose in frequency in the early phase of the migrant population out of Africa.

Materials and Methods

In this research, we used R Programming to analyze DNA sequences from the Thousand Genomes Project. The Thousand Genomes Project (TGP) includes whole genome analysis of 2,504 individuals from 26 populations around the world (See Figure 1 and Table 1). Of these 26 populations, we can organize them into 5 geographic regions: Americas (N= 6 populations), Africa (N=5 populations), Europe (N=5 populations), East Asia (N=5 populations), and South Asia (N= 5 populations). For the purposes of our research, we chose to not include the 6 populations of the Americas region. This data was not included because all the American samples reflect genetic diversity that is due to intercontinental migrating. The aim of our research instead is regarding genetic diversity due to factors like founder effects and bottlenecks.

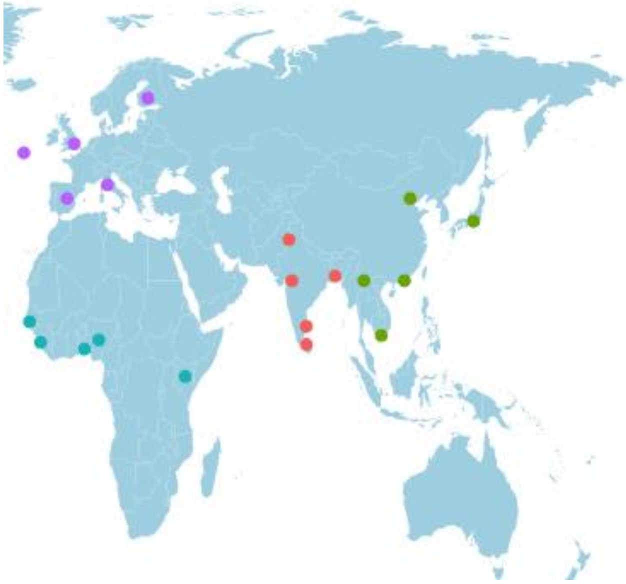

Figure 1: Map of Africa and Eurasia with Thousand Genome populations plotted. Blue- African, Purple- European, Red- South Asian, Green- East Asian. The purple, red, and green points make up the non-African or Out of African migrant populations.

Table 1. Thousand Genome populations categorized into geographic regions with sample sizes per population.

Africa	Europe	East Asia	South Asia
Mende in Sierra Leone N=85	Toscani in Italia N=107	Kinh in Ho Chi Minh City, Vietnam N=99	Gujarati Indians in Houston, Texas N=103
Esan in Nigeria N=99	Iberian Populations in Spain N=107	Chinese in Bejing N=103	Punjabi in Lahore, Pakistan N=96
Yoruba in Ibadan, Nigeria N=108	British from England and Scotland N=91	Han Chinese South N=105	Indian Telugu in the UK N=102
Gambian in Western Division N=113	Ceph European N=99	Chinese Dai in Xishuangbanna, China N=93	Sri Lankan Tamil in the UK N=102
Luhya in Webuye, Kenya N=99	Finnish in Finland N=99	Japanese in Tokyo, Japan N=104	Bengali in Bangladesh N=86

The TGP is completely open to the public for download, but we needed a programming software in order to be able to sift through the amount data. R is a programming language and environment specifically designed for statistical computing and data analysis. It is an open-source software that, like the TGP, can be downloaded and used by anyone in the public. It was the best choice for this project because of its beginner ease and ability to create graphics from the data being analyzed.

We used RStudio to filter the data to include only bi-allelic single nucleotide polymorphisms (SNPs). For our research, we excluded any loci that contained one allele, or greater than two alleles. We wrote an original script in R that allowed us to find what we will further refer to as Group Specific Polymorphisms (GSP) (Niedbalski and Long 2022). A GSP is an allele at a high frequency within one geographic region and at low frequencies in the region(s) outside of the group. We refer to these groups as focal and comparison, respectively. In the research for this paper, we only compared the populations within Africa to the populations outside of Africa collectively. We chose to use the parameters of a frequency of greater than 30% in a focal group and less than 1% in the comparison group (Niedbalski and Long 2022). Our justification for these parameters is such that we apply the Hardy-Weinberg principle to determine expected probability of an individual carrying the GSP. There the expected probability of an individual in the focal group carrying the GSP is a minimum of $(0.3)^2 + 2(0.3)(0.7) = 51\%$, and the expected probability of an individual in the comparison group carrying the GSP is a maximum of $(0.01)^2 + 2(0.01)(0.99) = 2\%$. For example, an African GSP will be present in all African populations at a high frequency, and not found or found at a very low frequency in all populations outside of Africa. In our trials we tested many other thresholds, and these chosen ones allowed us to work with enough GSPs to make a confident assessment, but not so many that we lost any visible patterns.

We then looked at clustering of GSPs on whole chromosomes. We created a sliding "window" of 50,000 base pairs on either side of each locus, and then quantified how many GSPs were present within that window. Graphing the number of GSPs present in each window gave us a visualization of where the GSPs cluster in a chromosome. We intuitively assigned the GSPs within Africa as what has been 'lost' be the migrant populations, and the GSPs in Eurasia as what has been 'gained' by the migrant populations. By plotting these against each other, we can see the frequency and size of loss and gain clusters on whole chromosomes. Although we analyzed multiple chromosomes, we focused only on chromosome 1 in this paper. The other chromosomes we analyzed showed an almost identical pattern of gain and loss, so we chose to condense our results on only one. In the future, we hope to look more closely at each chromosomes.

Results

In our analysis of the entire first chromosome, we identified a total of 6,337 Group Specific Polymorphisms (GSPs) within the African populations. These GSPs represent alleles with a high frequency within the African geographic region and a low frequency in populations outside of Africa. Conversely, outside of the African region, we detected 1,851 GSPs. These GSPs exhibit the same criteria, with high frequency in their respective geographic regions and low frequency in populations from the Africa region. This pattern was created by the out of Africa migration that occurred about 100,000 years ago (Nei 1995). In this migration, a small number of people from a single population would have left Africa and entered Asia, leaving more diversity within populations that stayed in Africa. This small percentage of alleles are what went on to populate the rest of the world. The GSPs found only outside of Africa are derived alleles that have been gained due to the founder effect.

Across all chromosomes, a consistent trend emerges—more losses than gains, with notably less significant clusters in the populations outside Africa. As stated before, we narrowed our focus for this paper onto only chromosome 1. As seen in figure 2, there are more clusters, as well as larger clusters of GSPs found in the African populations. The Out of Africa populations contain less and smaller clusters. There is a lack of clusters between the nucleotide positions 121,485,483 and 142,535,439 due to the fact that there are no SNPs analyzed by the Thousand Genomes project in this region. We can also assume there are no SNPs analyzed by the Thousand Genomes project on the end of the chromosome because of the missing SNPs after 200,000,000 nucleotides.

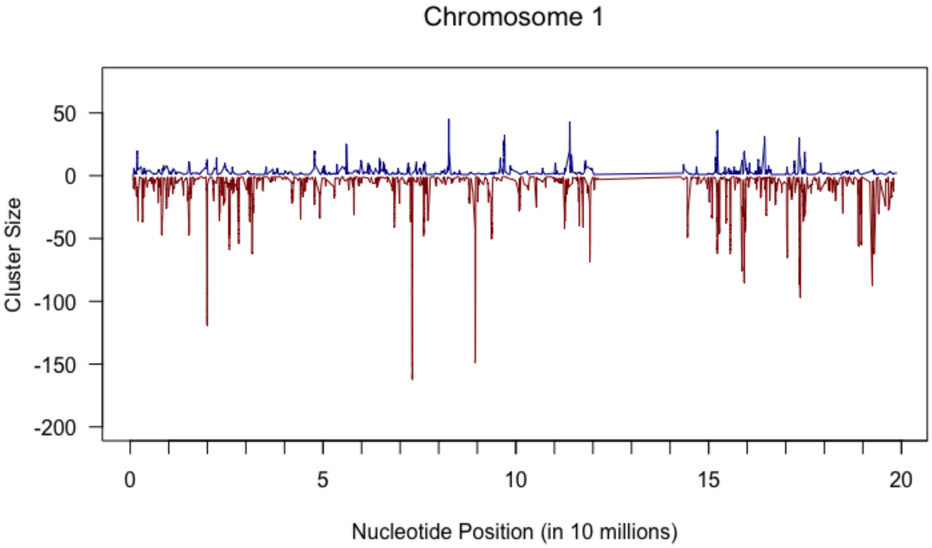

Figure 2: Graphed cluster sizes on chromosome 1. The negative portion of the y-axis represents the common alleles found within African populations, which are alleles that were lost during the OOA bottleneck. The positve portion of the y-axis represents the new mutations and rare variants that rose to high frequency in the OOA bottleneck populations.

Figure 3 shows difference in cluster frequency and size between the African and migrant populations. We see that the populations within Africa show much more frequent and larger clusters with a mean cluster size of 26.12256. The populations outside of Africa have less frequent and smaller clusters with a mean cluster size of 7.061224.

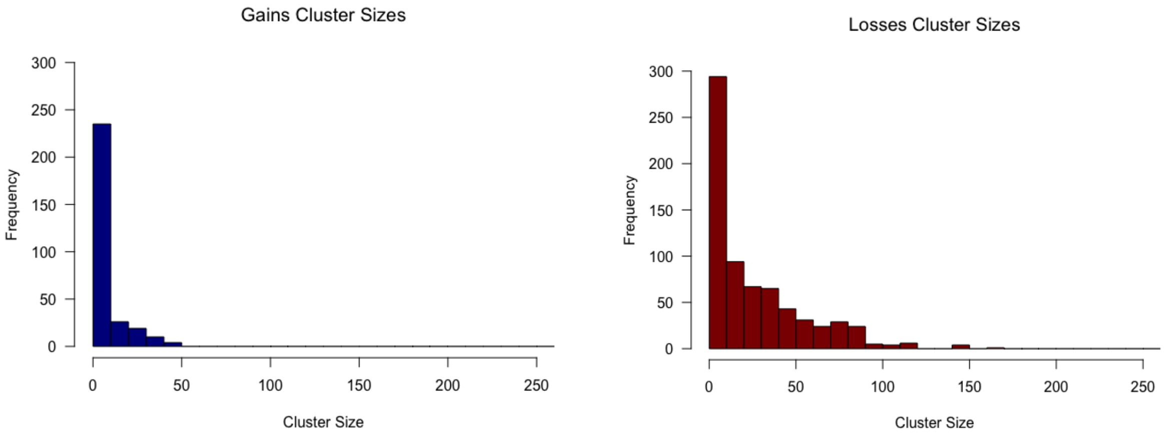

Figure 3: Histograms of the size and frequency of clusters on chromosome 1.

Discussion

The large amount of GSPs in the ancestral African populations in comparison to the bottleneck populations is indicative of already agreed upon findings of more common variants present in the ancestral populations. When we take into account the clustering pattern of the GSPs, we see these variants through a new lens. We know that inbreeding reduces genetic variation in small populations, and there are two relevant modes that it can occur via. One is if the population size has been small but stable for a long period of time, which results in an even distribution of low levels of heterozygosity across the genome. If there has been a recent founder effect with a rebound to a large population size, then we see a jagged pattern of heterozygosity (Figure 4). This is exemplified in recent studies of wolves that have experienced a history of inbreeding (Robinson et al. 2019). The two populations of interest to us are the Ethiopian wolves and the Isle Royale wolves. The Ethiopian wolves represent a small but stable population that has a long history of inbreeding, and therefore results in a very uniform distribution of heterozygosity. The Isle Royale wolves represent a reduction down to a small number of individuals, and then a rebound (with high levels of inbreeding in this case) which results in a more jagged "saw-tooth" distribution of heterozygosity. We can use these results as a model for the out of Africa migration, during which a bottlenecked population has rebounded into the Eurasian populations we are using in our comparison. When we compare these two outcomes to our own data, we see clearly that our data is most like that of the Isle Royale wolves.

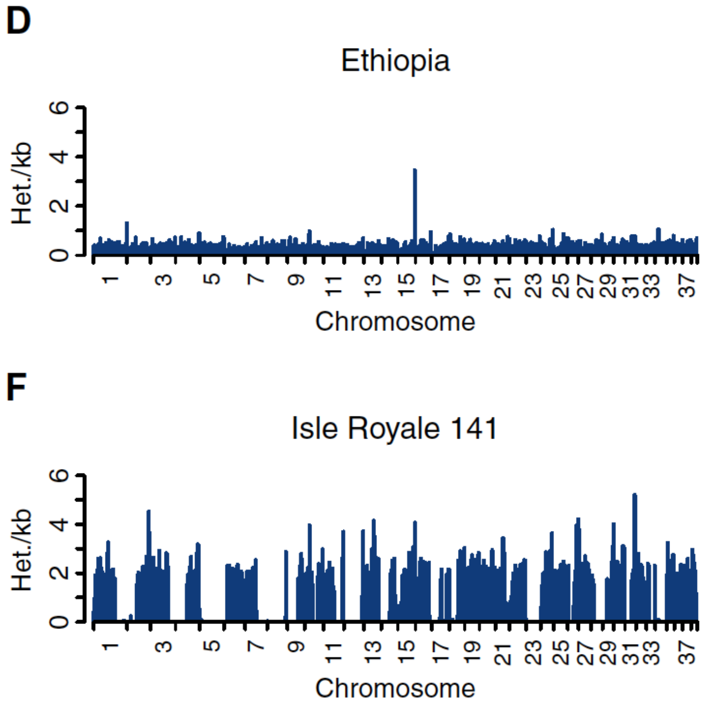

Figure 4: Distribution of heterozygosity across Ethiopian and Isle Royale wolves (Robinson et al. 2019).

Evolution and genetic separation of populations can occur in two ways. It occurs because populations have either lost traits from their common ancestor or because they have gained traits.

The alleles that only present in African populations are considered losses, and these losses occur very fast, almost immediately as populations migrated out of Africa. In the rebound that occurred in the OOA populations, we see gains of alleles. Gains can arise either through mutation and genetic drift or due to natural selection. It is agreed that gains occurring due to natural selection will take place over a long period of time and be slow to increase in frequency. If we compare the out of Africa migrant populations and then non-migrant populations, we can ask to what extent are their differences about addition of new alleles versus loss of old alleles. Do the gains look like they are due to natural selection or do they look like they are due to genetic drift?

From our results, we can dissect implications for natural selection versus genetic drift on the gains in the founder populations. It is clear in the histograms (Figure 3) that there are significantly more losses than gains over the entire genome. As stated in our results, we found about three times as many GSPs within Africa than outside of Africa. There are smaller clusters of GSPs outside Africa with a mean of around 7 GSPs per cluster in comparison to a mean of around 26 GSPS per cluster in African populations, as well as overall fewer clusters in the out of Africa populations. This is explained by our previous claims that losses happened quickly and in large amounts when founder populations migrated out of Africa. The smaller gains we see have taken thousands of years to amass through mutations. It is our inference that these gains were heavily influenced by natural selection. The alignment of some loss locations with gains is representing mutations increasing fitness in the new population in place of the ancestral population's alleles. This increase in frequency could be due to natural selection. We can justify this by assuming that the losses we see are due to genetic drift. Loss of an allele due to drift can occur in one generation, while a gain due to natural selection, one large enough to be seen in our parameters, needs more time to occur. We can corroborate our findings with the difference in mean cluster size between the populations. This lines up with the results of many losses and few gains we see in the plotted data.

There has been a lot of interest in the history of human populations, but here we found it important to look at the impact on genome architecture of this history and were able to make novel observations. In future research, we aim to extend our analysis to all chromosomes, explore patterns of natural selection and drift in the OOA Bottleneck more comprehensively, and compare different geographic groups to gain a deeper understanding of the evolutionary forces shaping global genomic diversity.

References

Auton, Adam, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, et al. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526 (7571): 68–74. https://doi.org/10.1038/nature15393.

Long, Jeffrey C., Jie Li, and Meghan E. Healy. 2009. “Human DNA Sequences: More Variation and Less Race.” *American Journal of Physical Anthropology* 139 (1): 23–34. https://doi.org/10.1002/ajpa.21011.

Nei, M. 1987. *Molecular Evolutionary Genetics*. New York: Columbia University Press.

———. 1995. “Genetic Support for the Out-of-Africa Theory of Human Evolution.” *Proceedings of the National Academy of Sciences* 92 (15): 6720–22. https://doi.org/10.1073/pnas.92.15.6720.

Niedbalski, Sara D., and Jeffrey C. Long. 2022. “Novel Alleles Gained during the Beringian Isolation Period.” *Scientific Reports* 12 (1): 4289. https://doi.org/10.1038/s41598-022-08212-1.

Ramachandran, Sohini, Omkar Deshpande, Charles C. Roseman, Noah A. Rosenberg, Marcus W. Feldman, and L. Luca Cavalli-Sforza. 2005. “Support from the Relationship of Genetic and Geographic Distance in Human Populations for a Serial Founder Effect Originating in Africa.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (44): 15942–47. https://doi.org/10.1073/pnas.0507611102.

Robinson, Jacqueline A., Jannikke Räikkönen, Leah M. Vucetich, John A. Vucetich, Rolf O. Peterson, Kirk E. Lohmueller, and Robert K. Wayne. 2019. “Genomic Signatures of Extensive Inbreeding in Isle Royale Wolves, a Population on the Threshold of Extinction.” *Science Advances* 5 (5): eaau0757. https://doi.org/10.1126/sciadv.aau0757.

Yu, N., F. C. Chen, S. Ota, L. B. Jorde, P. Pamilo, L. Patthy, M. Ramsay, T. Jenkins, S. K. Shyue, and W. H. Li. 2002. “Larger Genetic Differences within Africans than between Africans and Eurasians.” *Genetics* 161 (1): 269–74.