McGill University

Qualifying Exam

Mutational Patterns in Founder Populations

Author:
Luke Anderson-Trocmé

Supervisor:
Dr. Simon Gravel

Department of Human Genetics
Faculty of Medecine

November 22, 2018

0.1 Introduction

The human genome is composed of a mosaic of ancestral genetic material passed on for millennia. This pattern of interconnected relatedness is beginning to unveil humanities rich and complex history. Human expansion out of Africa (OOA) is estimated to have occurred between 45,000 and 60,000 years ago [1]. As humans expanded into new habitats, each new colony was founded by a smaller group departing a larger, more diverse source population. The impacts of the OOA bottleneck are still measurable, as distance from Africa is strongly correlated with decreasing genetic diversity [2].

Bottlenecks can drastically change the allele frequencies within a population and lead to rapid divergences [3, 4]. Rare alleles in the source population can increase significantly in frequency or even become fixed in the founder population. Isolation over time can lead to increased diversity between small populations, while the diversity within may remain relatively low. For this reason, migration between populations can increase the genetic diversity of both populations [4].

Recent events of admixture are tractable with genomic analyses like identity by descent (IBD) mapping and coalescence estimations [5]. However, recombination over time will fragment haplotypes eventually making it difficult to determine their ancestry. Despite this, a population will retain a unique set of alleles at frequencies waxing and waning over time. These allele frequencies can then be used to estimate demographic histories of multiple populations [6].

Many disease or phenotype association tests are susceptible to spurious results caused by population structure. In order to properly account for this, a better understanding of human demographic history is needed [6]. Moreover, studying recent founder populations can lead to better understanding of the mechanisms and dynamics of historical founder events.

0.1.1 Thesis proposal

The aim of this PhD project is to study the genomics of founder populations to infer population specific mutational signatures, demographic history and mutation rates. **Our hypotheses are**: **1**. An unusual non reproducible mutational pattern observed in Japan is the result of a technical artifact. **2**. The heterogeneity of Quebec's population structure can be accounted for by ancestral population structure predating the colonization of Canada. **3**. Leveraging phylogenetic records in combination with sequence data can improve the estimates of mutation rates.

1 Mutational signatures in Japan

1.0.1 Batch Effects in Aging Reference Cohort Data

The last 5 years have seen a drastic increase in the amount and quality of human genome sequence data. Reference cohorts such as the International HapMap Project [7], the 1000 Genomes Project [8, 9], and the Simons Diversity project [10], for example, have made thousands of genome sequences publicly available for population and medical genetic analyses. Many more genomes are available indirectly through servers providing imputation services [7] or summary statistics for variant frequency estimation [11].

The first genomes in the 1kGP were sequenced 10 years ago [12]. Since then, sequencing platforms have rapidly improved. Yet, because of the extraordinary value of freely available data, early data from the 1000 Genomes project (1kGP) is still widely used as a reference panel for imputation, allele frequency estimations and to answer a wide range of medical and evolutionary questions. The second phase of the 1kGP implemented multiple technological and analytical improvements over its earlier phases [9], leading to heterogeneous sample preparations and data quality.

Even though such batch effects clearly have the potential to confound analyses, the usefulness and ubiquity of the 1kGP data make such issues particularly insidious. This raises the question of whether and how such legacy data should be included in contemporary analyses alongside more recent cohorts. Here we point out how large and previously unreported batch effects in the early phases of the 1kGP still lead to incorrect genetic conclusions through population genetic analyses and indirect use through prominent imputation servers.

1.0.2 Mutational Signatures

There has been recent interest in population variation in germline mutational signatures which can be revealed in large sequencing panels. In 2015, Harris reported 50% more TCC $\rightarrow$ TTC mutations in European populations compared to African populations, and this was replicated in a different

cohort in 2017 [13, 14, 15]. Strong population enrichments of a mutational signature suggests important genetic or environmental differences in the history of each population [13, 14]. Harris and Pritchard further identified distinct mutational spectra across a range of populations, which were further examined in a recent publication by Aikens et al. [14, 16].

In particular, both studies identified a heterogeneous mutational signature within 1000 Genomes Japanese individuals. This heterogeneity is intriguing because differences in germline signatures accumulate over many generations. A systematic difference within the Japanese population would suggest sustained environmental or genetic differences across sub-populations within Japan with little to no gene flow. We therefore decided to follow up on this observation, by using a newly sequenced dataset of Japanese individuals.

1.1 Results

1.1.1 A peculiar mutational signature in Japan

Harris and Pritchard reported an excess of a 3-mer substitution patterns *AC→*CC in a portion of the Japanese individuals in the 1kGP [14]. While trying to follow up on this observation in a larger and more recent Japanese cohort, we did not find this particular signature. When comparing the allele frequencies between the Japanese individuals from the 1kGP and this larger dataset, we observed a number of single nucleotide polymorphisms (SNPs) private to one of the two groups (Figure 1). Given the similarity of the two populations, this strongly suggests a technical difference rather than a population structure effect. These mismatches were maintained after filtering for low-quality regions of the human genome and sites failing Hardy-Weinberg equilibrium tests.

1.1.2 Reverse GWAS

To identify SNPs that are likely to reproduce poorly across cohorts without having access to a second cohort, we performed an association study in the Japanese from the 1kGP (JPT) for variants that associate strongly with low average quality of mapped bases (Q) (Figure 1). Traditionally, genome wide association studies use genotypes as the independent variable. Here we perform a "reverse GWAS", in the sense that genotypes are now the dependent variable that we attempt to predict using the continuous variable Q as the independent variable. We use logistic regression of the genotypes on Q and identify 587 SNPs with $p < 10^{-8}$ and 1034 SNPs with $p < 10^{-6}$.

1.1.3 Identifying Suspicious SNPs in the 1kGP

The distribution of Q across 1kGP populations shows that many populations have distributions of Q scores comparable to that of the JPT, especially populations sequenced in the phase 1 of the project: sequencing done in the early phases of the 1kGP was more variable and overall tended to include lower quality sequencing data (Figure 2). This variability could result from evolving sequence platform and protocols or variation between sequencing centres. By 2011, older sequencing technologies were phased out, and methods became more consistent, resulting in higher and more uniform quality. We performed the reverse GWAS approach in all populations and identified 3826 variants that were independently associated to low Q in at least two populations with $p < 10^{-6}$ (Figure 3).

1.1.4 Suspicious variants impact modern genomics analyses

We searched the literature for any GWAS that might have reported these dubious variants as being significantly associated with some biological trait, even though there is no particular reason for these SNPS to be associated with phenotypes. The NHGRI-EBI Catalog of published genome-wide association studies identified three recent publications that had reported these variants as close to or above the genome-wide significant threshold. Two of these studies genotyped individuals and imputed the data using the HapMap II as a reference database for imputation [17, 18]. The third study used the 1kGP sequence data and cell lines directly [19]. Despite using state-of-the-art quality controls, these erroneous variants managed not only to be imputed onto real genotype data, but they also nearly reached genome wide significance for biological traits.

The HapMap project and 1000 Genomes project shared some cell lines but used different sequencing platforms to produce the data. It is possible that that a cell line artifact may have produced the same suspicious SNPs in the two sequencing projects.

1.1.5 Recommendations

A conservative approach would be to remove all individuals that do not meet the quality threshold as well as all the variants associated to low Q. In this case, we used a cut-off of Q over 30. Another approach is to remove all the sites that reach a significance below $p < 10^{-6}$ in at least two populations.

2 Population structure of Quebec

2.1 Brief history of Quebec

The Quebec population was founded over four centuries ago from around 8500 settlers from France [20, 21]. This small number of settlers are the direct ancestors of the great majority of Quebec's population of 7 million. Moreover, ancestral contributions of these settlers across the Saint-Lawrence valley is uneven, whereby two thirds of the modern gene pool is attributed to the first 2600 founders [22, 21]. While the rates of consanguineous unions in Quebec were not higher than in Europe, small population sizes and isolation over time led to increased distant consanguinity [21].

The North-East of Quebec is characterized historically as being more isolated with higher emigration rates compared to Western Quebec. Saguenay-Lac-Saint-Jean and Charlevoix also display higher rates of some rare genetic diseases as well increased homozygosity compared to Western Quebec [22]. Variation of disease prevalence across the Saint-Lawrence valley indicate that population structure fluctuates across geographic regions [22]. Indeed, this pattern of genetic diversity is attributed to a number of factors like concentration of founders, native ancestry and geographic isolation [22, 21, 20]. The founders of Quebec are not all from the same regions in France. In the seventeenth-century, while France was evenly populated, there were not many marriages across great distances [21]. French settlers came from regions around Paris and Atlantic port towns. Each group of settlers carried a distinct gene pool and founded a region of 'Nouvelle-France'. The population structure from each of these regions in France provides initial diversity despite small population sizes [21]. Moreover, Native American ancestry has been estimated to be around 1% among these settlements [23].

2.1.1 A case study of founder populations

Many populations experienced founder effects as humans expanded out of Africa and populated the continents. Studying examples of founder effects in human populations are a great way to improve our understanding of human demographic histories. The Quebec population is an good case

study for population genetics for multiple reasons. The source population of this bottleneck is still present, which may not be the case for more ancient migration events. Comparing allele frequencies between France and Quebec can determine more accurately the nature of the colonization. The French-Canadian population also has deep genealogical records that span four centuries and includes over 3 million entries. Quebec is not the only recent founder population, but it is among the largest and in combination with its rich genealogical records, it makes it an excellent cohort to study the impact of founder events [**Bherer2010**, **Gagnon2011**, 24, 25].

There have been several studies in the past few decades that have examined the population structure of Quebec. By analyzing the genealogical data they described distinct sub-populations and inferred the relative contribution of founders [**Bherer2010**, 22]. Genotype data used to analyze linkage disequilibrium and homozygosity has also revealed significant structure measurable today [25]. Quebec is not a homogeneous population, but rather made up of a combination of distinct sub-populations. This was confirmed with Fst statistics of these groups, indicating substantial genetic structure [25]. Roy-Gagnon (2011) also reported that there was greater differentiation between sub-populations of Quebec than between European populations. While this study included 140 genotyped individuals - a small sample by today's standards - they postulate that this difference is accounted for by genetic drift. Alternatively, the heterogeneity of the founder effect in the Quebec population could be explained by a series of separate founder events influenced by variation in the relative isolation, and migration between settlements [22, 21].

2.2 Materials and Methods

While the deep-genealogical data has been studied for over a decade, linking the genealogy to genetic data has only been done with relatively small sample sizes [22, 25]. The aim of the second chapter of this thesis will be to examine the genetic structure of the Quebec population using more recent and larger genomic datasets in combination with the deep-genealogical and geographical data. Recently developed methodologies will also be of use to further characterize this genomic landscape.

We plan on using two main cohorts for this chapter. The Genizon cohort has 9961 samples from Quebec, genotyped on a variety of chips. While this data has yet to be published, it has required a fair amount of cleaning and processing to account for the different genotyping chips used. One key advantage of using this data is that we already have consent to link 516 individuals to genealogical

records. We will also include data from the Cartagene project which has over 12,000 sequenced Quebecers.

The BALSAC database is a genealogical tree of the French-Canadian Quebec population. It was compiled from birth, marriage and death certificates from the Catholic Church. Because of the major role the Church played in French-Canadian society, these records are remarkably detailed and comprehensive. As mentioned above, 516 individuals from the Genizon cohort have been genotyped and are linked to this genealogy.

2.2.1 Founder effect or Ancient population structure?

Using Fst statistics we can infer the relative differences between sub-populations of Quebec (see [25]). Dominic Nelson has written software that can produce single locus forward simulations of the Quebec population using the genealogical data from BALSAC. These simulations can be used to produce estimated differences between sub-populations resulting purely from a founder effect. Comparing the Fst statistics between the real data and simulations will determine if the structure observed in Quebec is the result of preserved genetic structure from France or the result of drift from founder events.

2.3 Results : the geography of genetics

Preliminary analyses of the Genizon dataset are in agreement with previous work. In collaboration with Alex Diaz-Papkovich, we performed a UMAP projection of the Genizon genotype data [26, 27]. This method takes the first 20 principle components of the data as input and reduces this high dimensional data preserving both local and global structure (see [27] and [26]). This projection shows significant population structure within the Quebec population.

We colored these plots based on the three dimensional coordinates of the UMAP projection space (X,Y,Z) converted to three dimensional color space (R,G,B). Therefore, Individuals with similar color are closer to each other in the projection and presumably share more recent ancestry. To link the genetic data to geographic coordinates, we used the marriage locations of the ancestors of those linked to the BALSAC database. Combining the ancestral locations of these individuals with the three dimensional UMAP coloring scheme, geographic clustering of the sub-populations becomes apparent. Five genetic clusters are visible in the UMAP projection, these clusters also appear to be geographically linked. This is in agreement with what Roy-Gagnon observed [25]. These preliminary

results are very promising as they appear to capture the geographic and genetic clustering of the Quebec population.

2.4 Future directions

While Fst gives a good estimate of genome wide past coalescent rates, we plan on analyzing IBD segments as they can provide us with more detailed insight of demographic history [4, 28, 5]. Because Fst takes in all polymorphisms as input for the statistic, it does not distinguish between recent and more ancient common ancestry [28]. Analysis of IBD segments shared among individuals from Quebec can be useful for reconstructing the recent structure of relatedness. The length of the shared segments can be used to estimate the number of generations leading back to a common ancestor [29, 30, 5]. Combining this with geographical data can infer patterns of shared ancestry across the Saint-Lawrence valley. Ralph (2013) identified a general trend of decreasing recent common ancestry with geographic distance as well as distinct regional signals of IBD in different European populations [28]. Europe has a rich history of complex migratory patterns and demographic histories that is difficult to trace back with certainty. Quebec on the other hand has a much more recent history that may be easier to trace with IBD analysis.

While studying the French-Canadian populations on their own can provide a detailed historical inference of demographics, including samples from other European populations can provide more context. We plan on expanding the current analyses to include data from French individuals which could be used to confirm the results from the Fst analysis as well as provide insight into which regions in France are source populations to regions in Quebec. Dr. Christian Dina in the University of Nantes in France has expressed interest in sharing French genomics data for this project.

3 Estimating human mutation rates

Mutation rate estimations are necessary for understanding the timing of evolutionary processes and events. Neutral mutations are expected to arise at a constant rate in the genome, and can therefore be used as an evolutionary clock that can estimate divergence times [31]. The basic equation of mutation rate is $\mu = N/LM$ where N is the number of mutations, L is the length of the region and M is the number of generations [32]. Estimating mutation rates in humans has proven to be difficult for a number of reasons. Mutation rates vary across the genome as some regions are more conserved than others. Mutation rates are also known to vary over time as shorter branches within the Ape family compared to other primates suggest a "hominoid rate slowdown" whereby mutation rates decreased in apes [31, 33]. In addition, more recent studies have found that mutation rates can even vary among human populations. [13, 14].

3.1 Review of current methods and limitations

The first human mutation rates were estimated by Haldane (1935) using the rates of children born with rare diseases from unaffected parents [34]. However, once sequence data became available, estimates were based on phylogenetic data and fossil records. By using fossil record dating as an estimate for coalescence time t and the sequence divergence d between two species as the number of mutation that have occurred over that time span, mutation rates μ can be estimated as $\mu = 2d/t$ [33]. Using this method with the time of divergence of Chimps and Humans of 6 to 10 million years, the mutation rates are estimated to be around 1×10^{-9} up to 2.5×10^{-8} [35, 33, 31, 32]. The limitations of this approach are numerous, ranging from accuracy of fossil record dating, to estimating the average generation time for chimps and humans over millions of years.

Instead of using distant ancestry and fossil records, another more direct approach to estimating mutation rates is to sequence the genomes of parents and offspring. This approach tends to give estimates that are less than half of the rate compared to the phylogenetic approach ($1 - 1.25 \times 10^{-8}$) [36, 32]. One issue with trio sequencing for mutation rates is that false-positive rates of sequencing

technologies are difficult to account for considering how low the true mutation rate is. Another issue is that the samples used for sequencing are not zygotes. This leaves room for germline mosaicisms to artificially increase the number of mutations observed in a generation.

Methods for estimating mutation rates using population genetics are increasingly being used. These approaches expand on the theory behind trio sequencing by including more distantly related individuals. One approach involves autozygous segments observed in a founder population with a known pedigree [30]. These segments shared by multiple individuals are compared and mutations in these segments are then used in combination with the number of generations separating two individuals [30]. Using autozygous regions from 70 trio sequenced Hutterite families, they obtained an estimated mutation rate of 1.2×10^{-8}. Palamara et al. (2015) describe an even more generalized method by identifying IBD segments from unrelated individuals and inferring the number of generations using coalescent theory [29, 5]. This method then regresses the number of mutations observed per base against the time of most recent common ancestor. The slope of the regression is the mutation rate per generation and the intercept is the inferred genotyping false-positive rate. Applying this method to 498 individuals from 250 families, they measured a mutation rate of (2.08×10^{-8}).

3.1.1 Applying new methods to new data

The aim of the third chapter of this thesis will be to explore mutation rate estimations using a combination of sequence data and genealogical data from the French-Canadian population. One preliminary objective might be to replicate the population genetics methods described above by Palamara et al. and Campbell et al. [29, 30]. Knowing that mutation rates can vary across populations, applying these methods to a new dataset may in itself be worthy of publication.

Considering the deep genealogical records available for the French-Canadian population, we may be able to improve on the Palamara method by acquiring more accurate estimates of generation time separating two IBD segments. Indeed, by linking individuals to the phylogenetic tree of the Quebec population, we may be able to better estimate the most recent common ancestor for a given segment of the genome. Furthermore, by grouping IBD segments according to genomic regions (e.g. coding, CpG, etc.) and running the analysis separately would provide us with a more detailed view of the genome wide mutation rate in humans.

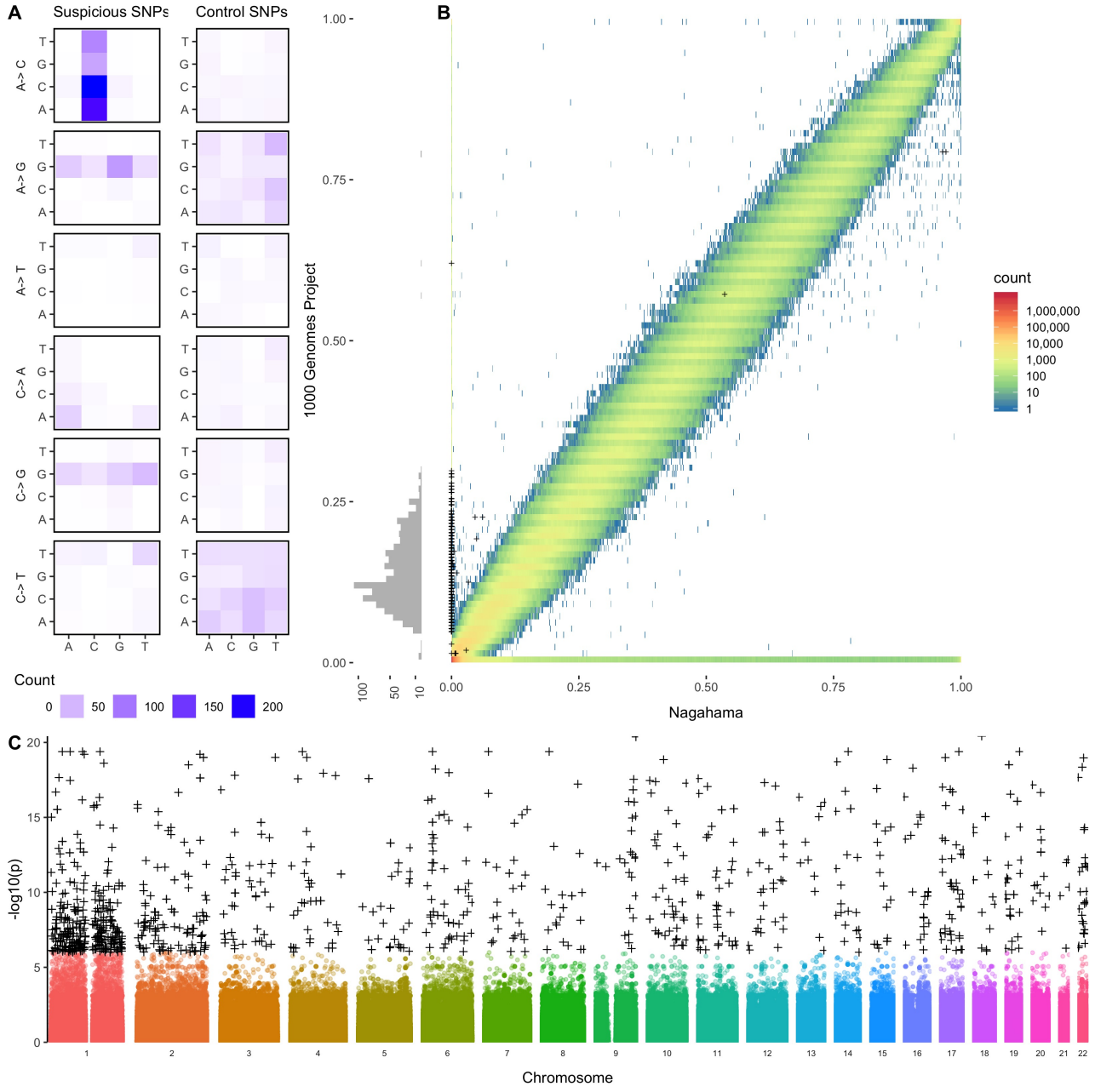

FIGURE 1: **A** Mutation spectrum of the 1034 variants that reached a genome wide significance with a p-value less than $p < 10^{-6}$ in a GWAS of sequencing quality. The majority of the variants with significant associations to Q have the *AC→*CC mutational pattern. There is also a slight enrichment in GA*→GG* and GC*→GG* mutations. These three enrichments can be summarized as G**→GG*. (note: the reverse complement of *AC→*CC is GT*→GG*) **B** Joint frequency spectrum plot of the Japanese from the 1kGP and a more recent Nagahama dataset. Crosses (+) are variants that reached genome wide significance in a GWAS of sequencing quality. The histogram on the left of the plot is the distribution of significant variants. **C** Genome wide association of the average quality of mapped bases Q for the 104 Japanese individuals included in the 1kGP. This GWAS identified 587 $p < 10^{-8}$ and 1034 $p < 10^{-6}$ SNPs that were associated to the average Q of SNPs mapped for an individual The same analysis was performed independently for each of the populations in the 1kGP.

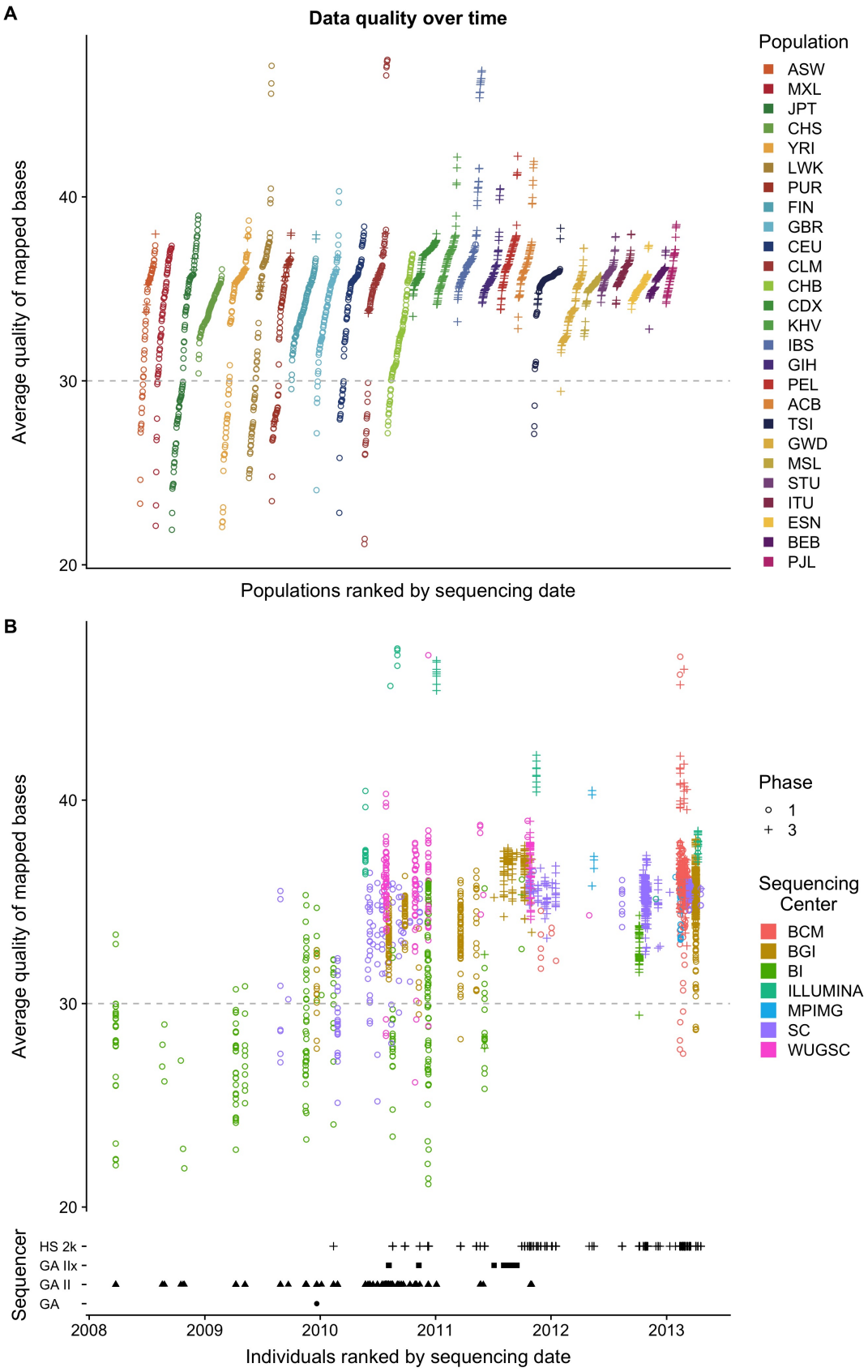

FIGURE 2: **A** The average quality of mapped bases Q for each individual per population included in the 1000 Genomes sequencing project. Individuals are ranked by the date of the earliest sequencing data is used for individuals sequenced more than once. The x-axis is ranked by the mean sequencing date per population. **B** The x-axis is sorted by the sequencing date per individual. The colors indicate the sequencing centres that produced the data for each individual and the shape indicates whether the individual belongs to Phase 1 or Phase 3 of the 1000 Genomes project. The bottom plot indicates the sequencing technologies used over time.

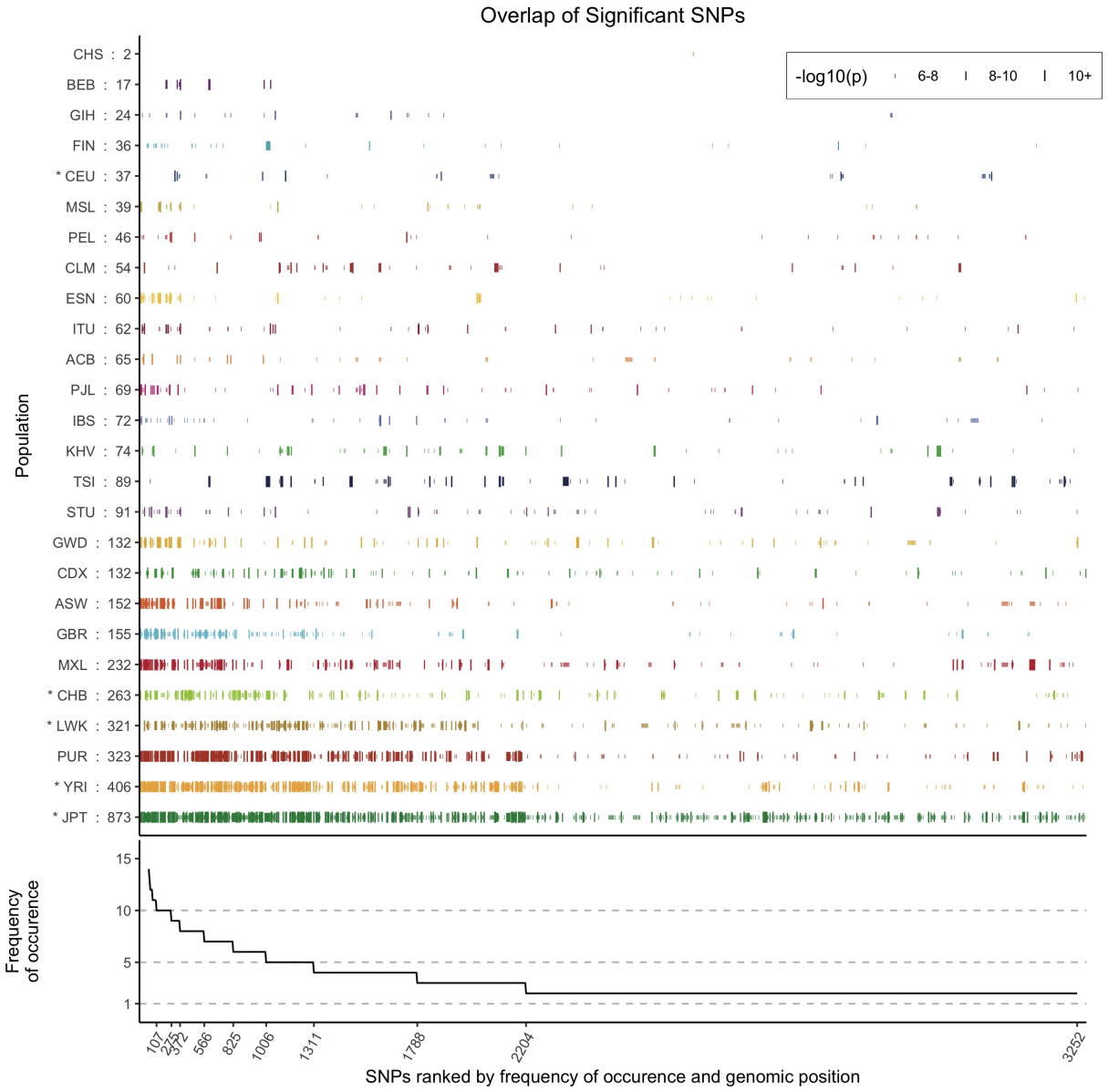

FIGURE 3: Overlap of SNPs identified independently to be associated with average quality of mapped bases Q. The size of the crosses (+) are proportional to the -log10(p) value of that SNP. The x axis is ranked by the frequency of occurrence of a SNP, then by genomic position. Phase 1 populations are marked by a star (*). The line plot underneath shows the number of populations for which a variant has reached significance. The populations that tend to have the most individuals with low Q also tend to have the most variants associated to Q. The same variants identified as being low quality independently in each population are found in other populations.

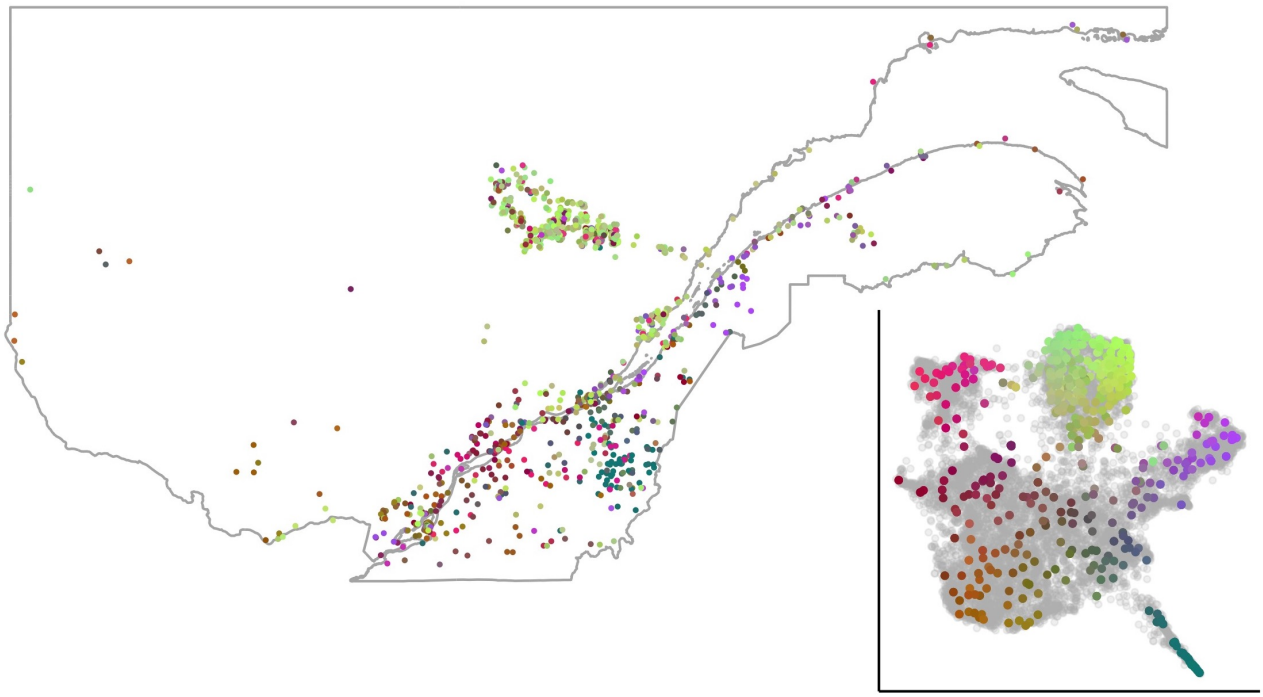

FIGURE 4: A map of Quebec colored by a 3D UMAP projection. The coloring of each individual is based on the three dimensional (X,Y,Z) coordinates of the UMAP projection space converted to three dimensional color space (R,G,B). Therefore, Individuals with similar color are closer to each other in the projection. Five genetic clusters are visible in the UMAP projection, these clusters also appear to be geographically linked. A flattened 2D view of of the 3D projection used for coloring is presented in the bottom right.

Bibliography

[1] Brenna M Henn, L L Cavalli-Sforza, and Marcus W Feldman. "The great human expansion." In: *Proceedings of the National Academy of Sciences of the United States of America* 109.44 (2012), pp. 17758–64. ISSN: 1091-6490. DOI: 10.1073/pnas.1212380109. URL: http://www.ncbi.nlm.nih.gov/pubmed/23077256http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3497766.

[2] Sohini Ramachandran et al. "Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa." In: *Proceedings of the National Academy of Sciences of the United States of America* 102.44 (2005), pp. 15942–7. ISSN: 0027-8424. DOI: 10.1073/pnas.0507611102. URL: http://www.ncbi.nlm.nih.gov/pubmed/16243969http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1276087.

[3] Noah A Rosenberg et al. "Genetic structure of human populations." In: *Science (New York, N.Y.)* 298.5602 (2002), pp. 2381–5. ISSN: 1095-9203. DOI: 10.1126/science.1078311. URL: http://www.ncbi.nlm.nih.gov/pubmed/11954565http://www.ncbi.nlm.nih.gov/pubmed/12493913.

[4] Brian Charlesworth, Deborah Charlesworth, and Nicholas H. Barton. "The Effects of Genetic and Geographic Structure on Neutral Variation". In: *Annual Review of Ecology, Evolution, and Systematics* 34.1 (2003), pp. 99–125. ISSN: 1543-592X. DOI: 10.1146/annurev.ecolsys.34.011802.132359. URL: http://www.annualreviews.org/doi/10.1146/annurev.ecolsys.34.011802.132359.

[5] Soheil Baharian et al. "The Great Migration and African-American Genomic Diversity". In: *PLOS Genetics* 12.5 (May 2016), pp. 1–27. DOI: 10.1371/journal.pgen.1006059. URL: https://doi.org/10.1371/journal.pgen.1006059.

[6] Simon Gravel et al. "Demographic history and rare allele sharing among human populations." In: *Proceedings of the National Academy of Sciences of the United States of America* 108.29 (2011), pp. 11983–8. ISSN: 1091-6490. DOI: 10.1073/pnas.1019276108. URL: http://www.ncbi.nlm.

nih.gov/pubmed/21730125http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3142009.

[7] International HapMap Consortium. "A haplotype map of the human genome." In: *Nature* 437.7063 (2005), pp. 1299–320. ISSN: 1476-4687. DOI: 10.1038/nature04226. URL: http://www.ncbi.nlm.nih.gov/pubmed/16255080.

[8] 1000 Genomes Project Consortium. "A map of human genome variation from population-scale sequencing." In: *Nature* 467.7319 (2010), pp. 1061–73. ISSN: 1476-4687. DOI: 10.1038/nature09534. arXiv: 1302.2710v1. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3042601{\&}tool=pmcentrez{\&}rendertype=abstract.

[9] 1000 Genomes Project Consortium. "An integrated map of genetic variation". In: *Nature* 135 (2012), pp. 0–9. DOI: 10.1038/nature11632.

[10] Swapan Mallick et al. "The Simons Genome Diversity Project: 300 genomes from 142 diverse populations". In: *Nature* 538.7624 (2016), pp. 201–206. ISSN: 14764687. DOI: 10.1038/nature18964. arXiv: NIHMS150003.

[11] Monkol Lek et al. "Analysis of protein-coding genetic variation in 60,706 humans". In: *Nature* 536.7616 (2016), pp. 285–291. ISSN: 14764687. DOI: 10.1038/nature19057. arXiv: 030338.

[12] Erwin L. van Dijk et al. "Ten years of next-generation sequencing technology". In: *Trends in Genetics* 30.9 (2014), pp. 418–426. ISSN: 01689525. DOI: 10.1016/j.tig.2014.07.001. arXiv: arXiv:1312.0570v2. URL: http://linkinghub.elsevier.com/retrieve/pii/S0168952514001127.

[13] Kelley Harris. "Evidence for recent, population-specific evolution of the human mutation rate". In: *Proceedings of the National Academy of Sciences* 112.11 (2015), pp. 3439–3444. ISSN: 0027-8424. DOI: 10.1073/pnas.1418652112. URL: http://www.pnas.org/lookup/doi/10.1073/pnas.1418652112.

[14] Kelley Harris and Jonathan K. Pritchard. "Rapid evolution of the human mutation spectrum". In: *eLife* 6 (2017). ISSN: 2050084X. DOI: 10.7554/eLife.24284.

[15] Iain Mathieson and David Reich. "Differences in the rare variant spectrum among human populations". In: *PLoS Genetics* 13.2 (2017). ISSN: 15537404. DOI: 10.1371/journal.pgen.1006581.

[16] Rachael C Aikens, Kelsey E Johnson, and Benjamin F Voight. "Signals of variation in human mutation rate at multiple levels of sequence context". In: *bioRxiv* (2018). URL: http://biorxiv.org/content/early/2018/08/04/385096.abstract.

[17] Aldi T. Kraja et al. "A bivariate genome-wide approach to metabolic syndrome: STAMPEED Consortium". In: *Diabetes* (2011). ISSN: 00121797. DOI: 10.2337/db10-1011.

[18] Jane L. Ebejer et al. "Genome-wide association study of inattention and hyperactivity-impulsivity measured as quantitative traits". In: *Twin Research and Human Genetics* (2013). ISSN: 18324274. DOI: 10.1017/thg.2013.12.

[19] Rajendra Mandage et al. "Genetic factors affecting EBV copy number in lymphoblastoid cell lines derived from the 1000 Genome Project samples". In: *PLoS ONE* (2017). ISSN: 19326203. DOI: 10.1371/journal.pone.0179446.

[20] Claude Bherer et al. "Admixed ancestry and stratification of Quebec regional populations". In: *American Journal of Physical Anthropology* 144.3 (2011), pp. 432–441. ISSN: 00029483. DOI: 10.1002/ajpa.21424. URL: http://doi.wiley.com/10.1002/ajpa.21424.

[21] A-M Laberge et al. "Population history and its impact on medical genetics in Quebec". In: *Clinical Genetics* 68.4 (2005), pp. 287–301. ISSN: 00099163. DOI: 10.1111/j.1399-0004.2005.00497.x. URL: http://doi.wiley.com/10.1111/j.1399-0004.2005.00497.x.

[22] Alain Gagnon and Evelyne Heyer. "Fragmentation of the Quebec population genetic pool (Canada): Evidence from the genetic contribution of founders per region in the 17th and 18th centuries". In: *American Journal of Physical Anthropology* 114.1 (2001), pp. 30–41. ISSN: 0002-9483. DOI: 10.1002/1096-8644(200101)114:1<30::AID-AJPA1003>3.0.CO;2-L. URL: http://doi.wiley.com/10.1002/1096-8644{\%}28200101{\%}29114{\%}3A1{\%}3C30{\%}3A{\%}3AAID-AJPA1003{\%}3E3.0.CO{\%}3B2-L.

[23] Claudia Moreau et al. "Native American Admixture in the Quebec Founder Population". In: *PLoS ONE* 8.6 (2013). Ed. by Dennis O'Rourke, e65507. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0065507. URL: https://dx.plos.org/10.1371/journal.pone.0065507.

[24] M.H.D. Larmuseau et al. "Genetic genealogy comes of age: Perspectives on the use of deep-rooted pedigrees in human population genetics". In: *American Journal of Physical Anthropology* 150.4 (2013), pp. 505–511. ISSN: 00029483. DOI: 10.1002/ajpa.22233. URL: http://doi.wiley.com/10.1002/ajpa.22233.

[25] Marie-Hélène Roy-Gagnon et al. "Genomic and genealogical investigation of the French Canadian founder population structure". In: *Human Genetics* 129.5 (2011), pp. 521–531. ISSN: 0340-6717. DOI: 10.1007/s00439-010-0945-x. URL: http://link.springer.com/10.1007/s00439-010-0945-x.

[26] Leland McInnes and John Healy. "Umap: Uniform manifold approximation and projection for dimension reduction". In: *arXiv preprint arXiv:1802.03426* (2018).

[27] Alex Diaz-Papkovich, Luke Anderson-Trocme, and Simon Gravel. "Revealing multi-scale population structure in large cohorts". In: *bioRxiv* (2018), p. 423632.

[28] Peter Ralph and Graham Coop. "The Geography of Recent Genetic Ancestry across Europe". In: *PLoS Biology* 11.5 (2013). Ed. by Chris Tyler-Smith, e1001555. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1001555. URL: https://dx.plos.org/10.1371/journal.pbio.1001555.

[29] Pier Francesco Palamara et al. "Leveraging Distant Relatedness to Quantify Human Mutation and Gene-Conversion Rates". In: *The American Journal of Human Genetics* 97.6 (2015), pp. 775–789. ISSN: 0002-9297. DOI: 10.1016/J.AJHG.2015.10.006. URL: https://www.sciencedirect.com/science/article/pii/S0002929715004085.

[30] Catarina D Campbell et al. "Estimating the human mutation rate using autozygosity in a founder population". In: *Nature Genetics* 44.11 (2012), pp. 1277–1281. ISSN: 1061-4036. DOI: 10.1038/ng.2418. URL: http://www.nature.com/articles/ng.2418.

[31] Priya Moorjani, Ziyue Gao, and Molly Przeworski. "Human Germline Mutation and the Erratic Evolutionary Clock". In: *PLOS Biology* 14.10 (2016), e2000744. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.2000744. URL: http://dx.plos.org/10.1371/journal.pbio.2000744.

[32] Vagheesh M Narasimhan et al. "A direct multi-generational estimate of the human mutation rate from autozygous segments seen in thousands of parentally related individuals". In: *bioRxiv* (2016), p. 059436. DOI: 10.1101/059436. URL: https://www.biorxiv.org/content/early/2016/06/17/059436.

[33] Aylwyn Scally and Richard Durbin. "Revising the human mutation rate: implications for understanding human evolution". In: *Nature Reviews Genetics* 13.10 (2012), pp. 745–753. ISSN: 1471-0056. DOI: 10.1038/nrg3295. URL: http://www.nature.com/articles/nrg3295.

[34] J. B. S. Haldane. "The rate of spontaneous mutation of a human gene". In: *Journal of Genetics* 31.3 (1935), p. 317. ISSN: 0022-1333. DOI: 10.1007/BF02982403. URL: https://doi.org/10.1007/BF02982403.

[35] Naoyuki Takahata and Yoko Satta. "Evolution of the primate lineage leading to modern humans: Phylogenetic and demographic inferences from DNA sequences". In: *Proceedings of the National Academy of Sciences* 94.9 (1997), pp. 4811–4815. ISSN: 0027-8424. DOI: 10.1073/pnas.94.9.4811. eprint: http://www.pnas.org/content/94/9/4811.full.pdf. URL: http://www.pnas.org/content/94/9/4811.

[36] Jay Shendure and Joshua M. Akey. "The origins, determinants, and consequences of human mutations". In: *Science* 349.6255 (2015), pp. 1478–1483. ISSN: 0036-8075. DOI: 10.1126/science.aaa9119. eprint: http://science.sciencemag.org/content/349/6255/1478.full.pdf. URL: http://science.sciencemag.org/content/349/6255/1478.