Genetic introgression in chimpanzees and bonobos

Marc de Manuel Montero

TESI DOCTORAL UPF / ANY 2018

Dr. Tomàs Marquès i Bonet,

DEPARTAMENT DE CIÈNCIES EXPERIMENTALS I DE LA VIDA

upf. Universitat Pompeu Fabra Barcelona

A la Júlia,

Agraïments

Tomàs, per acceptar-me al teu grup. Per confiar en mi i donar-me la oportunitat de liderar un projecte important. Per no limitar-te a ser el meu "jefe", sinó estar disposat a donar consell en moments difícils.

Javi, Irene y Tiago por ser mis mentores al principio de la tesis.

Ignasi i Guillem, perquè us trobo a faltar, perquè els jocs de cartes no són el mateix sense vosaltres.

Jéssica, simplement per ser com ets. Perquè ets la definició d'una persona bona.

Raquel, por permanecer siempre risueña. Por ser el mejor ejemplo de determinación y fuerza de voluntad.

Irene, por tu insaciable curiosidad. Por compartir tu sabiduría con los demás.

Aitor, per estar boig en la millor de les maneres. Per saber escoltar a les persones.

Clàudia, per estar sempre oberta a ajudar incondicionalment. Per manternir-te sempre alegre davant les adversitats.

Manolo, por las risas. Por todas las noches de borrachera.

Laura, Luis, Paula i Marina, perquè sou una glopada d'aire fresc. Perquè emaneu alegria i energia a tots els veterans del grup.

Martin, for guiding me through the darkest time in my thesis.

David y Esther. Por ser personas siempre dispuestas a escuchar y ayudar.

Jonas, for being the epitome of kindness.

Tom, for the exciting projects we have started together. For being a good friend.

Fátima, porque te echo de menos. Por tu dulzura y proximidad.

Tom Gilbert, for giving me the opportunity to visit your group in Copenhagen. For your generosity.

A tots aquells del grup "El Merendero". Per fer les hores al laboratori més curtes.

Al grup de música La Social Disfunktion, per reintroduïr-me a la música i sovint esvaïr els maldecaps del doctorat.

A tota la gent de Poble Nou. Per les cerveses, els camions, les festes i les converses.

Aleix, Pablo, Roderic i Frederic. Perquè sou els meus millors amics.

Famílies Domingo i Espinós, per rebre'm de braços oberts. Especialment el Jordi i la Milagros, pels dinars de diumenge i els sobretaules.

Marta i el Gabriel, per ser els millors sogres que algú pot desitjar.

Mercè i Enric, perquè recordo les platges de petit amb vosaltres. Avi, perquè sempre et recordaré per com eres.

Nieves y Paco, por ser la definición de cariño. Porque os quiero mucho y os lo debería decir más a menudo.

Pare. Per transmetre'm el teu amor per la biologia. Perquè sé que sempre podré comptar amb tu. Per posar la meva felicitat pel davant de tot.

Mare. Ets la persona més bonica que conec. Perquè t'estimo i m'estimes.

Júlia. Perquè ets una espurna de joia, l'epicentre de tota la meva força. Perquè tens el superpoder d'irradiar felicitat a tots els que t'envolten. Per fer-me sentir estimat durant les nits més fosques. Perquè se'm fa un nus al pit i ploro escribint aquestes paraules, i això vol dir que t'estimo. Perquè em fas sentir afortunat. A tots aquells que llegeixen, us convido a deslliurar-vos de preocupacions per un instant. Estimeu als que us envolten, digueu-ls'hi! Al cap i a la fi, que és més bonic que estimar i ser estimat?!

Abstract

Our closest living relatives, chimpanzees and bonobos, have a complex demographic history. We analyzed the high-coverage whole genomes of 75 wild-born chimpanzees and bonobos from 10 countries in Africa. We found that chimpanzee population substructure makes genetic information a good predictor of geographic origin at country and regional scales. Multiple lines of evidence suggest that gene flow occurred from bonobos into the ancestors of central and eastern chimpanzees between 200,000 and 550,000 years ago, probably with subsequent spread into Nigeria-Cameroon chimpanzees. Together with another, possibly more recent contact (after 200,000 years ago), bonobos contributed less than 1% to the central chimpanzee genomes. Admixture thus appears to have been widespread during hominid evolution.

Resum

Els nostres parents més propers, els ximpanzés i els bonobos, tenen una història demogràfica complexa. En aquest estudi, hem analitzat els genomes sencers de 75 ximpanzés i bonobos nascuts en 10 països diferents d'Àfrica. Hem descobert que l'estructura poblacional dels ximpanzés fa possible predir l'origen geogràfic a escala nacional i regional mitjançant dades genètiques. Múltiples linies d'evidencia suggereixen que que l'ancestre dels ximpanzés est- i centre-africans va rebre gens de poblacions de bonobo. Juntament amb un altre contacte possiblement més recent (fa més de 200.000 anys), els bonobos han contribuït amb menys del 1% als genomes dels ximpanzé centre-africans. Així doncs, la hibridació entre espècies sembla haver estat generalitzada durant l'evolució dels homínids.

Preface

The origin of the human lineage along with their evolutionary relatives has been a longstanding question in biology. Recently, the advent of molecular data has provided an unprecedented resource to answer this question. In this thesis, I present results that deepen the understanding of the population history of chimpanzees and bonobos, our closest living relatives.

Table of contents

Abbreviations

A: Adenine
C: Cytosine
DNA: Deoxyribonucleic acid
G: guanine
T: thymine

1. INTRODUCTION

In my thesis, I have used the genetic variation of populations to learn about their past. In order to link genetic variation to population history, one must know how genetics behaves under different demographic scenarios, an enterprise that as we shall see is far from obvious. Population genetics is the field in biology concerned to ask such questions, and it has been under thorough development for nearly a century. Here I will review some population genetics principles essential to understand my research.

1.1. A broad overview: evolution, DNA and genetic variation

Evolution by natural selection is the process by which organisms change over time as a result of changes in heritable physical or behavioral traits (Darwin 1869). Notice that this definition carefully states "Evolution by *natural selection* is the process...", a remark that will be relevant later in this section. Indeed, in the mind of Charles Darwin, evolution was driven by natural selection on those individuals with the highest capability to generate offspring (fitness). If we analyse the previous sentence, we can easily see that there must exist some kind of *variation* in a population for evolution to happen, otherwise the population would be composed by clones with identical fitness and it would remain stall. Another important feature of evolutionary change is that the selected traits must be *heritable,* so that they are transmissible from parent to offspring. By the time of Darwin and Wallace, the ultimate recipient of this *variation* and *heritability* was yet to be discovered. Phenotypic variation was thought to be inherited by blending the attributes of each parent into a new individual. However, around the same time Darwin was travelling the world aboard the Beagle, Gregor Mendel observed that inheritance works in discrete units (genes). Although Mendel and Darwin never had the chance to discuss their ideas, scientists in the early 20th century combined Mendelian genetics and Darwinian evolution into what later became the *Modern synthesis of evolution*. Furthermore, between the 1940s and 1950s, DNA was shown to be the molecular repository of genetic information (Avery, MacLeod, and McCarty 1944). The discovery of the double helix (Watson and Crick 1953) cemented DNA as the ultimate recipient of the aforementioned *variation* and *heritability*. DNA has certainly been the main protagonist of my thesis, as I have spent most of my research comparing stretches of DNA between different individuals of a population. For that reason I shall describe its nature and how it is studied in modern times.

DNA is a molecule made by two chains of nucleotides (A, G, T and C) which coil around each other to form a double helix. In eukaryotes, genomes are composed by multiple long stretches of DNA (chromosomes) carrying the genetic instructions used in the growth, development, functioning and reproduction of an organism. Because DNA is formed by 4 discrete "letters" or construction units, it can be read as a simple sequence of characters, very much like a text in a human book. Many times during my thesis I have found myself wondering if alternative systems of storing information would support life. Our computers work on an electronic storage medium based on only two states (1 or 0), and although DNA has more power to store data due to its four-moduled nature, maybe living forms based on a binary chemical version of our electronic bits could exist somewhere. In any case, all known organisms in our planet except some viruses use DNA to store genetic information.

Our ability to sequence and interpret DNA has improved dramatically over the last decades. In the 1960s, the first biochemical methods for detecting differences in DNA sequence were developed. Protein electrophoresis was used to characterize alleles according to the speed a stained protein moved on a gel under standard conditions. In the 1980s, the polymerase chain reaction (PCR), together with Sanger sequencing, provided access to direct DNA sequencing. In humans, the technology was often used to sequence mitochondrial DNA, which was easily retrieved and sequenced due to its high abundance in the cell and short length. However, the greatest revolution in DNA sequencing methods occurred from the 2000s onwards, and one could say we are still living in it. The Human Genome Project took 10 years to finish and cost $3 billion US dollars (Davies 2010). Nowadays, resequencing a whole human genome takes a few hours and routinely costs less than $1.000.

The advent of high-throughput methods has enabled the sequencing of hundreds of thousands of human genomes, and currently ongoing projects aim for 1-2 million sequenced genomes (Ledford 2016). Non-human studies have also greatly benefited from accessible sequencing. Currently there are thousands of sequenced genomes across the whole tree of life, which arguably has transformed every field of biological research. Notably, DNA sequencing has not only strived for cheap and efficient resequencing technologies, but has also invested in a next generation of long-read technologies. However, what advantage would long-read sequencing technologies have over the short-read ones? As we discussed earlier, genomes are extremely long sequences of nucleotides. In primates, the number of nucleotide base-pairs in a genome add up to $3 \cdot 10^9$. This unimaginable large number varies across species, with contrasts as striking as the 10^6 in *E. coli* and the 10^{11} in the marbled lungfish. Organisms with large genome sizes tend to have higher fractions of their genome composed by short sequences repeated thousands of times in a row (terminal repeats, tandem repeats, transposable elements, etc.). In the case of humans, around two thirds of the genome are estimated to be repetitive or repeat-derived (de Koning et al. 2011). Now imagine you are given the task of identifying which part of a zebra is pictured in a photograph. A picture of the head or the tail would be a rather easy task, but we would surely have trouble with a picture showing the 16th and 17th stripe. DNA mappers are software facing similar challenges when mapping short reads to complex reference genomes. If the read sequences are not long enough to span the whole repetitive region, finding a best match turns out to be impossible. In that regard, long-read sequencing technologies have provided great advantages when mapping or assembling repetitive regions, enabling the production of high quality reference genomes for non-human species (Kronenberg et al. 2018).

As you can imagine, processing billions of letters is an unfeasible task without the power offered by modern computers. The exponential growth in DNA sequencing technologies required a similar growth of algorithms capable of dealing with large amounts of data. Bioinformatics is the interdisciplinary field that develops methods and software tools to understand biological data, and has become an essential skill in genomics. In my experience, genomic projects in 2018 are more often hampered by CPU availability and storage than lack of DNA sequences, which suggests that efforts should be invested in improving bioinformatic pipelines to match up with DNA sequencing technologies.

One of the most important functions of bioinformatics is to extract genetic variation out of a collection of DNA sequences. When studying a population of genomes, we should be interested in those positions where individuals carry different genetic information. Technically, this procedure is called single-nucleotide polymorphism (SNP) calling, and it consists of an intricate number of pre- and post-steps that I will not go into. Briefly, the identification of SNPs with short-read data goes as follows. Consider a set of 100 sparkling violetear short-read genomes we have sequenced for our study. Since the violetear genome is roughly 1 billion nucleotides long, our life would be much easier if we focus in those positions where individuals carry different genetic information. In order to spot such positions, we must align our short-pieced genomes to a global coordinate system. We are lucky enough to have an assembly for the violetear, so we proceed to find the best matching place for each of our short reads in the 100 genomes (mapping). Once we have all the genomes mapped to the same reference, we are ready to identify all the positions where at least one individual carries a different nucleotide than the rest (SNP calling). Notice that we have gone from a gibberish of unordered DNA sequences to a well-structured list of positions that contain genetic variation. As we will see, this information can then be exploited to learn about the history, population structure and demography of populations.

The discovery of DNA, the development of technologies to sequence it and the use of computers has allowed us to uncover the genomic variation of populations. This gives us the chance to study the ultimate repository of *variation* and *heritability* that fuels evolution. In the following section, we will see how genetic variation appears and changes through time.

1.2. Allele and genotype frequencies, the Hardy-Weinberg equilibrium

By sequencing the whole-genome of 100 sparkling violetear we detected 3 million variable positions or *polymorphisms*. Each of these positions generally contains two alleles segregating in the population (ie. nucleotides A and T). We can readily compute *allele frequencies* by counting allele copies and dividing by the total number of chromosomes (200 in our case since violetears are diploid). Although multiallelic loci can exist, these are often discarded because are characteristic of regions in the genome hard to align or copy-number variable sites. Additionally, we can also study how the alleles are distributed among the individuals in the population. We can compute *genotype frequencies* by counting homozygous (AA or TT) and heterozygous (AT) individuals and dividing by the total number of individuals (100 in our case). Estimating allele frequencies from genotype frequencies is then an obvious process (ie. in our violetear population; 4 AA, 32 AT and 64 TT would yield allele frequencies of A = 0.2 and T = 0.8). However, can we reverse the process and estimate genotypes frequencies from allele frequencies? For example, knowing that the frequency of G in the 100th polymorphism in violetears is 0.35, what proportion of individuals would we expect to have the genotype GG?

We can answer this question, but only if we make some assumptions. One particularly useful simplifying assumption is that mating is random, ie. that violetears have no preference or taste when choosing mates (panmixis). We will also assume that the population is infinite or very large. Given these assumptions, the chance that an individual of the offspring is of genotype GG is given by the probability of receiving a G allele from both the mother and the father. If the G frequency is equal in females and males (another assumption), the frequency of GG in the next generation is simply the product of the G frequency ($0.35^2 = \sim 0.12$). This property can be extended to the rest of genotypes by the commonly used formulae (if p and q are respectively the G and C allele frequencies; GG = p^2, GC = $2pq$, CC = q^2). This observation was the first milestone in theoretical population genetics, the celebrated Hardy-Weinberg law. The Hardy-Weinberg law describes the equilibrium state of a single locus in a randomly mating diploid population that is free of other evolutionary forces, such as mutation, migration and genetic drift.

One might think that such assumptions are never met in reality. It would be reasonable to think that violetears do not choose mates at random. If the population is composed by several groups isolated by patches of terrain without forest (population structure), panmixis would be difficult to achieve simply due to logistic reasons, as flying time between regions would be too large. Natural selection would also violate the conditions under Hardy-Weinberg equilibrium, as well as populations with reduced census sizes. Because of all these reasons, we would perhaps be surprised to see that the Hardy-Weinberg principle reflects quite well real populations. In figure 1, I plot the allele frequency against the frequency of the 3 genotypes for 20.000 SNPs in 504 humans from Europe (data extracted from the 1000 Genomes Project). The black solid lines show the mean genotype distribution calculated using a loess smoothing, while the dashed gray lines draw the expectation under Hardy-Weinberg equilibrium. The theoretical and empirical trajectories are almost identical. This particular example illustrates one of the most important features of population genetics.

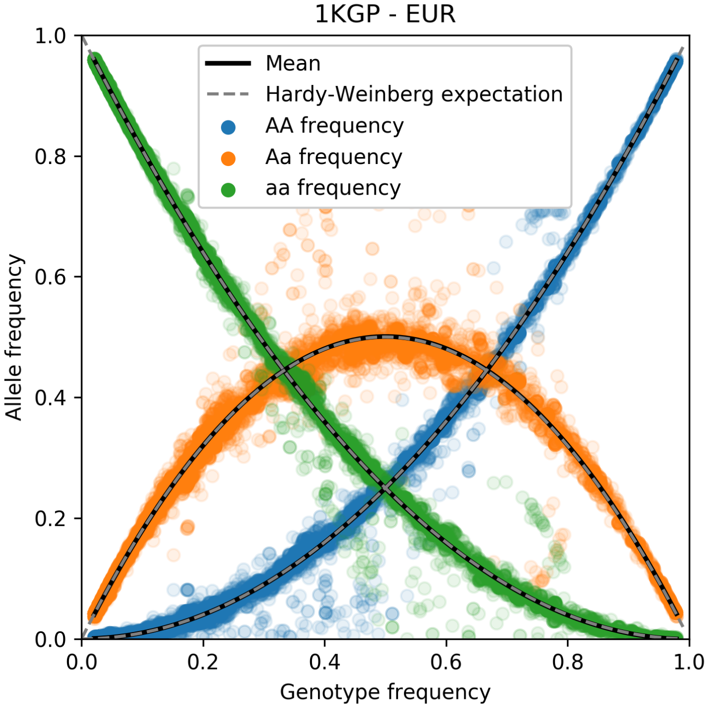

Figure 1. Demonstration of the Hardy–Weinberg proportions using 20.000 SNPs from the CEU European population in the 1000 Genomes Project. The allele frequency of each SNP is plotted against the frequency of the three possible genotypes in a diallelic locus. The solid lines show the mean genotype frequency calculated using a loess smoothing. The dashed line shows the expected genotype frequency under Hardy–Weinberg equilibrium.

It is impossible to fully understand the genetic structure of a population, such knowledge would require a complete description of the genome, spatial location of every individual and environmental conditions at one instant in time. In the next instant, the description would change as most individuals move, some are born and some die, while their genes mutate and recombine. Population genetics have achieved remarkable success by choosing to ignore all these complexities and focus on simple models that seem to reflect reality well. The following quote by the statistician George Box evokes quite well the population genetics mindset: '*All models are wrong, but some are useful*'.

1.3. Genetic drift

As we saw in the previous section, the Hardy-Weinberg law is premised on the population size being infinite. If we check the census size of the sparkling violetear, we can see that the species is under the Least Concern status in the IUCN scale. While this suggests their census size is quite large, it is certainly not infinite, thus the Hardy-Weinberg law may seem hard to accept. Indeed, in populations with finite population size, the allele frequencies in the next generation can be heavily influenced by stochastic events. If a subpopulation of 50 violetears lives isolated in a plateau in a Tepui, the allele frequencies in the next generation will be influenced by variation of offspring in each individual. Perhaps some individuals leave more descendants than others, not because of natural selection, but because of extrinsic factors not related with genetics. Some individuals might leave no offspring because the heavy rains destroyed their nests. On the other hand, randomness can also come from the process of Mendelian segregation. In our imaginary Tepui population, consider a locus with two alleles represented by *A* and *a*. The average number of *A*s and *a*s transmitted to the offspring may not be exactly equal to the expected number (ie. if *a* is only present in 6 heterozygous individuals, the number of *a* alleles in the next generation could very well be 0 due to random chance). *Genetic drift* is the process behind all these scenarios, and it describes the random change of allele frequencies in populations of finite size.

As we have discussed earlier, population genetics concentrates on understanding how allele frequencies change through time. Since real populations are finite, if we manage to incorporate genetic drift in our models, we have learned a great deal about evolution. The founders of population genetics theory, Sewall Wright and Ronald Fisher, provided a model which represents well the inherent stochasticity of genetic drift. The *Wright-Fisher model* assumes a haploid population with discrete generations, where offspring is generated by randomly sampling alleles. This process can easily be simulated by a computer program following an algorithm that iterates over N haploid individuals in the population and; (I) chooses an allele at random from the parent generation, (II) makes a copy of the selected allele and (III) places the copy in the new generation.

The results of such simulations can be found in figure 2. The first obvious observation is that allele frequencies do not remain constant through time, albeit some simulations show more diversity than others. Particularly, genetic drift seems to work much faster in small populations than in large populations. This should not surprise us, since intuitively we can understand that the power hold by randomness is much higher in processes involving a small number of *tests*. The probability of not getting any head in 4 fair coin tosses is relatively high (~6%), but getting zero heads out of 500 tosses sounds like a miracle (<0.0001%). In the Wright-Fisher model, genetic drift operates at a rate inversely proportional to N (1/2N in diploids). This rate can be easily understood if we think of it as a *decay of heterozygosity*. What is the probability that two alleles are the copy of the same allele in the previous generation, thus generating a homozygous individual? The probability is 1/2N, as all alleles are equally likely to be chosen and we have to choose the same allele twice. The probability 1/2N is the backbone of many mathematical expressions in population genetics, so I recommend the reader to keep pondering about this.

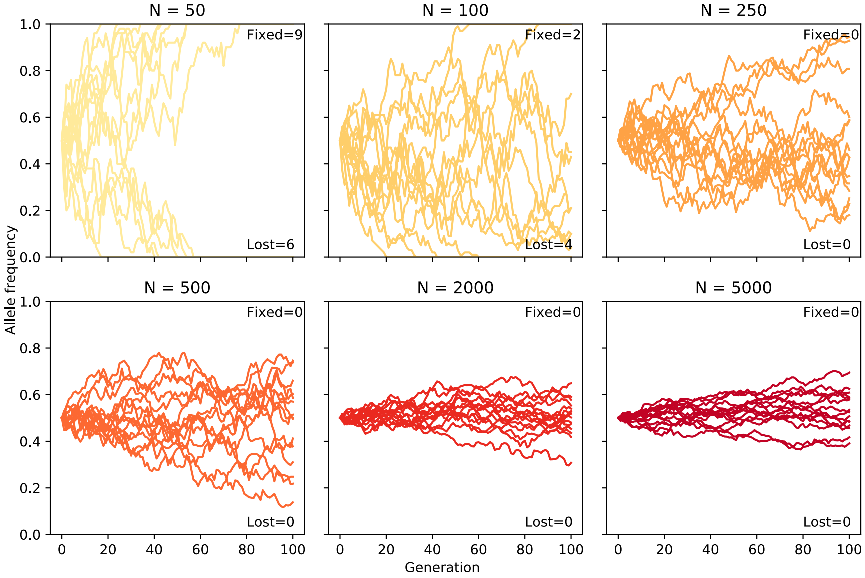

Figure 2. Computer simulations of the Wright-Fisher model of random genetic drift. Each line represents a population of the size N simulated for 100 generations.

A second feature of genetic drift is that alleles are lost from the population, and as we have seen, much more so in populations with small N. From this we might reasonably conclude that genetic drift removes genetic variation from populations. The third feature is more subtle; the direction of the random changes is neutral, there is no systematic tendency for the frequency of alleles to go up or down. I like to think of genetic drift as a drunk person walking in a bridge without rails. They cannot help but to randomly take steps left and right, ultimately falling off the bridge, being swallowed by the river (do not worry, the bridge is very close to the water and they can safely swim back to the shore afterwards). Funnily enough, genetic drift is often modelled as a "drunk" particle. The diffusion approximation approach (Motoo Kimura 1957) assumes that drift disperses allele frequencies in a manner analogous to heat diffusing through a metal rod or particles moving in a Brownian fashion (Motoo Kimura 1957; Kolmogoroff 1931).

Here I must do a little digression from genetic drift to comment on two common features in population genetic analyses. First, the simple computer program described above to perform Wright-Fisher simulations has an interesting characteristic: the probability of each event depends only on the state attained in the previous event. In other words, the frequency in generation *t* only depends on the frequency in the generation *t*-1. This is very good news for any programmer, as the algorithm only needs to keep track of the previous step rather than the whole trajectory. *Markov chains* are stochastic models describing a sequence of events like the Wright-Fisher model, and are widely used in bioinformatics. Additionally, the algorithm works *forward in time*, unlike other population genetics methods that propagate *backward in time* (see coalescent theory in section 1.5). This difference in the orientation of the flow of time is often source of misunderstandings in discussions among population geneticists, as words like *later* or *after* take an opposite meaning.

The Wright-Fisher model, as all models in population genetics, is a simplification of how populations evolve. For example, most populations do not show a constant size, many do not have discrete generations, and almost all have some degree of population structure. However, due to its mathematical simplicity, we tend to want to interpret genetic data in the context of the Wright-Fisher model, or similar models. For this purpose, population geneticists came up with the sometimes enigmatic concept of *effective population size* (*Ne*). The *Ne* of a population is equal to the number of individuals in a Wright-Fisher model that would generate an equivalent amount of genetic drift as in the real population. There are many estimators of genetic drift in real populations (ie. the expected number of average pairwise differences between individuals per site, or π), but one must remember that the *census size* and the effective size can be vastly different. If we go back to our hypothetical violetear population, we may see that the *Ne* is implausibly small given our knowledge of the species. This may be in part because the census size can suddenly drop in some generations (population bottleneck), and in part because typically only a fraction of the total population leaves offspring (remember the heavy rains in Tepui and the destroyed nests).

So far, genetic drift seems to contradict the stability promised by the Hardy-Weinberg law. Are these two fundamental observations incompatible? The answer is that not necessarily. Hardy-Weinberg equilibrium can be attained in only one or two generations. Genetic drift operates in the order of 2N generations, a number vastly larger than two in natural populations. A model population will appear to be close to Hardy-Weinberg equilibrium at any particular generation in time, since the deviation from the expectation due to drift will be in the order of 1/2N. One could say that genetic drift works across millennia, while the Hardy-Weinberg law operates on a yearly basis. However, we have left an important question unanswered. If genetic drift gradually removes genetic diversity from populations, why are not all the populations devoid of variation?

1.4. Mutation, the ultimate source of change

The answer to the last question, of course, is that *mutation* introduces new diversity to populations. Mutations are heritable changes in the genetic material that occur in DNA replication, but also sporadically in non-replicative DNA (ie. spontaneous cytosine deamination results in a nucleotide change to thymine). Organisms have evolved accurate molecular machineries of DNA replication and proofreading to reduce the rate of errors to very low numbers. The rhythm at which mutations appear, the *mutation rate*, can be estimated by comparing the genomes of individuals in a pedigree (ie. how many differences are found between the genomes of parents and their offspring?) (Kong et al. 2012). Scientists have also come up with alternative methods to estimate the mutation rate, such as spotting heterozygous positions in long stretches of homozygous genetic material (Narasimhan et al. 2017), or subtracting the number of derived mutations carried by modern and ancient individuals (Fu et al. 2014) (ie. an individual that died 45.000 years ago *lacks* 45 millennia of evolution, therefore it has accumulated fewer mutations since the ancestor). All these approaches have consistently obtained a mutation rate in humans of ~$0.5\cdot10^{-9}$ per base-pair per year. Assuming that the average age of parents at conception is 25 years old, and knowing that a human genome has ~$3\cdot10^{9}$ base-pairs, we can easily compute that the expected number of novel mutations per child is around 75 (= $0.5\cdot10^{-9} * 3\cdot10^{9} * 25 * 2$). Indeed, in figure 3, we can see that the number of mutations in newborns tends to be around 67 in 1.548 trios from Iceland (Jónsson et al. 2017). Since generations are the time unit of evolution, the mutation rate is often expressed *per generation* instead of *per year*. The rates are simply converted by assuming an average age of parents at reproduction, also called the *generation time* ($0.5\cdot10^{-9}$ becomes $1.25\cdot10^{-8}$ if the generation time is 25 years).

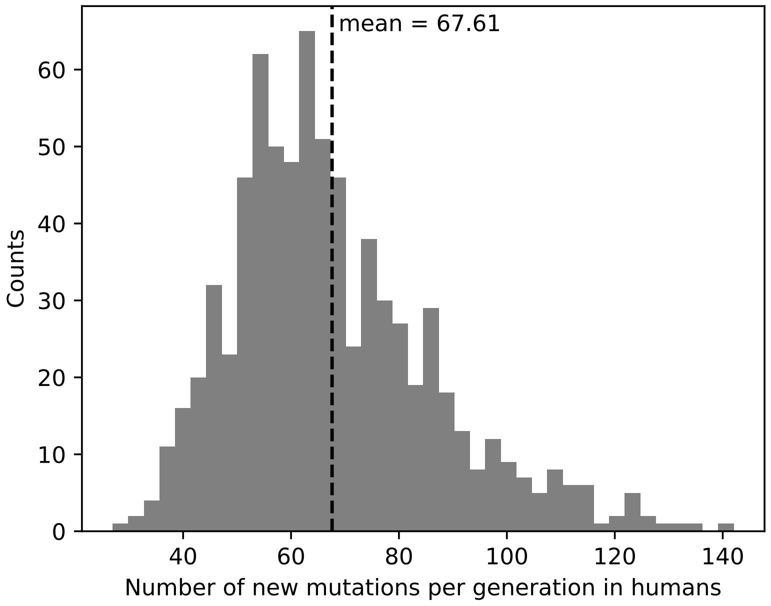

Figure 3. Number of novel mutations in one generation of humans. Data downloaded from Jónsson et al. 2017.

If we look carefully at the previous paragraph, we will see that the yearly mutation rate is expressed in a weird way. Why would anyone write $0.5 \cdot 10^{-9}$ instead of $5 \cdot 10^{-10}$? In fact, I do not know why the scientific community has taken this convention to express the human mutation rate, but it consistently does so. By taking a look at earlier estimates of the human mutation rate, we may be able to do an educated guess on the reasons behind this. Mutation rates can also be estimated by counting genetic differences between the genomes of two different species and dividing by their split time. For example, we know that the genomes of chimpanzees and humans carry different genetic information in 1.27% of the genome (Kronenberg et al. 2018). The mismatch rate between DNA sequences is commonly referred to as *divergence*. We also know that according to the fossil record, chimpanzees and humans shared a common ancestor around 6 million years ago (Benton and Donoghue 2007). From these data we can conclude that the mutation rate in the human and chimpanzee lineages must be around $1 \cdot 10^{-9}$ per base-pair per year ($0.0127 / 6 \cdot 10^{6} = \sim 2 \cdot 10^{-9}$, since the two branches accumulate mutations, we have to divide by two to get to the final $1\cdot10^{-9}$). Scientists have been puzzled for years by the incoungrence between the *slow* pedigree-based mutation rate estimates ($0.5\cdot10^{-9}$) and the *fast* estimates based on the fossil record and sequence divergence ($1\cdot10^{-9}$). Since the fast mutation rate was estimated decades earlier, my hypothesis is that $0.5\cdot10^{-9}$ is used to stress the fact that it is half the original estimate ($5\cdot10^{-10}$ does arguably a worse job conveying this message). Later in this section we will discuss the biological reasons behind the difference in the *slow* and *fast* human mutation rates.

But, how good are human cells at keeping their DNA intact? How low are all these values? Well, they are very low. The probability of me being struck by a lightning in my lifetime is around 1 out of 300.000 ($3.3\cdot10^{-6}$). As we have seen, the probability of a base-pair mutating in a generation of modern humans is around $1.25\cdot10^{-8}$. From that we can conclude that a single nucleotide in the genome has as many chances to mutate in one generation as me being struck by lightning 266 times in my lifetime ($3.3\cdot10^{-6}/1.25\cdot10^{-8} = 266.6$). However, we must remember that populations can exist for thousands of generations and that the genome is very long. Even if the probability of me being struck by a lightning is almost negligible, people across history have died from it. Similarly, the cellular machinery may have extreme fidelity replicating DNA, but a cumulative number of errors still occur and have led to the origin of new species across millennia. In fact, in a population of N diploid organisms, each chromosome in each individual can mutate, thus mutations enter the population at 2N times the mutation rate ($2Nu$ if u is the mutation rate). That means that the $7\cdot10^9$ humans in the planet carry around $252\cdot10^9$ new mutations from the previous generation ($7\cdot10^9 * 3\cdot10^9 * 1.25\cdot10^{-8}$). This number is far larger than the length of the genome, so every mutation that can exist, exists as of today. I always have found this a staggering thought. I wonder if this fact will soon be exploitable as the number of sequenced human genomes keeps increasing. With a cohort large enough, we should be able to spot the positions in the genome where mutations are never tolerated. Such knowledge should allow us to identify which regions in the genome play important functions, as well as boost our general understanding on how genomes operate and are regulated.

Some may have finished reading the last paragraph with confusion. Personally, the fact that the number of mutations entering a population scales with its size troubled me for a long time. Does this mean that populations with more individuals *evolve* faster? Do they accumulate more mutations per unit of time? We may find some answers to these questions in the previous section (see section 1.3). As we have seen, in the absence of mutation and selection, any allele will eventually be lost or fixed (figure 2). From this observation, it follows that one allele must be the ancestor of all the other alleles in the population if we wait long enough. As there are 2N alleles in a diploid population, the chance of any particular one becoming the ancestor is simply 1/2N. If the number of copies of an allele is higher than one (*i* copies), the chance of it being the ancestor is multiplied by the number of copies (1/2N * *i* = *i*/2N). An analogy illustrating this principle might be; if each number in a lottery has the same chance to win the prize, the more tickets I buy, the higher my chance to get the pot. Similarly, under a model without selection and mutation, the *probability of fixation* of an allele is its current frequency (*i*/2N). Since all novel mutations start with a single copy in a population, the probability of them surviving random drift and going to fixation is 1/2N.

You may have noticed that the number of mutations entering a population (2N*u*, *u* being the mutation rate) and the fixation probability of a new mutation (1/2N) both depend on the population size (N). In our hypothetical dataset of violeatears, we sampled an isolated population in Tepui (small N) and some other common population of the Venezuelan rainforest in Imataca (large N). We know that these populations have remained isolated for over 50.000 years, which has lead to the accumulation of genetic differences between them. While fewer mutations have entered the Tepui population per generation (2N*u*), their probability of going to fixation and contributing to genetic differentiation is much higher (1/2N). Indeed, when we analyse the violetear genomes, we see that both populations have become equally differentiated from the ancestor over the 50.000 years. This is explained by the fact that mutations not only have to *appear* in a lineage, but they have to become *fixed* to contribute to genetic differentiation. Therefore, the *rate of substitution* in diverging populations is independent of the effective population size ($2Nu * 1/2N = u$), and is simply equal to the mutation rate. According to the previous chain of reasoning, all populations should diverge from each other at exactly the same pace. This means that the number of mutations separating two lineages accumulate in a *clockwise* manner, and that this information can be used to infer the time since their separation. Such observation was coined as the *molecular clock* (Zuckerkandl and Pauling 1962), and has been widely used since its inception to estimate divergence times between species.

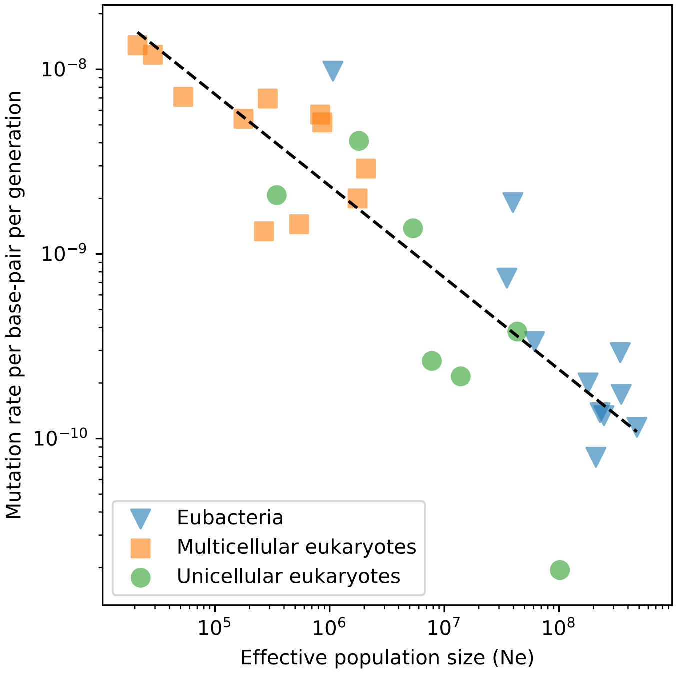

Figure 4. Linear regression of mutation rates on the estimated effective population size. Data extracted downloaded from Lynch et al. 2016.

Is our conclusion true? Do mutations appear at a similar rate in violetears and humans? Figure 4 shows the mutation rate per base-pair per generation against the effective population size in a diverse array of species. Mutations appear at vastly different rates in some lineages (minimum and maximum rates are $1.9 \cdot 10^{-11}$ in *Paramecium tetraurelia* and $1.25 \cdot 10^{-8}$ in humans respectively). Even more surprisingly, there seems to be a strong inverse correlation between effective population size and the mutation rate. These observations contradict our previous argument on how population size and the rate of substitution should be independent. As it often happens in science, we are probably approaching the problem obviating an important variable. Which could be the factor casting diversity in this theoretically constant space? So far, we have assumed simple models with random mating, constant effective population size, etc., always disregarding the role of natural selection. The *neutral theory of molecular evolution* claims that *most* allelic variation and substitutions accumulate due to genetic drift rather than natural selection (Motoo Kimura and Ohta 1971). It is important to stress the *most* in the definition. The neutral theory has been called non-Darwinian evolution, perhaps an unfortunate description given the fact that the theory does not reject natural selection *per se*, but rather undermines its importance in the divergence of species. Indeed, the neutral theory was controversial when first proposed and remains controversial today (Kern and Hahn 2018). To resolve such controversies, one should know how much of the genome is influenced by natural selection. This is a hard problem, as demographic events can often mimic events of natural selection (ie. population bottlenecks and hard selective sweeps). In my opinion, the neutral theory has had an undeniable impact in the development of population genetics, especially when inferring population structure and demographic history. However, as we shall see next, many observations in biology are incompatible with neutrality, and we must invoke natural selection to explain them.

As we have discussed earlier, figure 4 shows that the mutation rate is not constant across species. Moreover, it shows that mutation rate covariates with the effective population size, something that violates the expectation under neutrality. One explanation for such observation might be related to the tight relationship between genetic drift and the effective population size. The *drift-barrier* hypothesis postulates that owing to the stochastic nature of genetic drift, there is an upper bound to the level of refinement achievable by natural selection (Lynch et al. 2016). This hypothesis builds upon that idea that most novel mutations are detrimental to the fitness (deleterious); if we randomly modify a well designed machine, we are probably worsening its performance rather than improving it. Under this premise, most populations would be happy to have a lower mutation rate. However, natural selection can only work on traits that provide an increment in fitness (*s*) higher than the strength of genetic drift (1/2N). Therefore, DNA replication fidelity can only improve until its selective advantage is overwhelmed by the power of random genetic drift (ie. sculpists can only shape a figure to the point their hand steadiness allows it). Since unicellular organisms have higher effective population sizes, their mutation rate tends to be lower than multicellular organisms. The drift-barrier hypothesis illustrates that natural selection is much more efficient in populations of large size, an important principle of evolutionary biology.

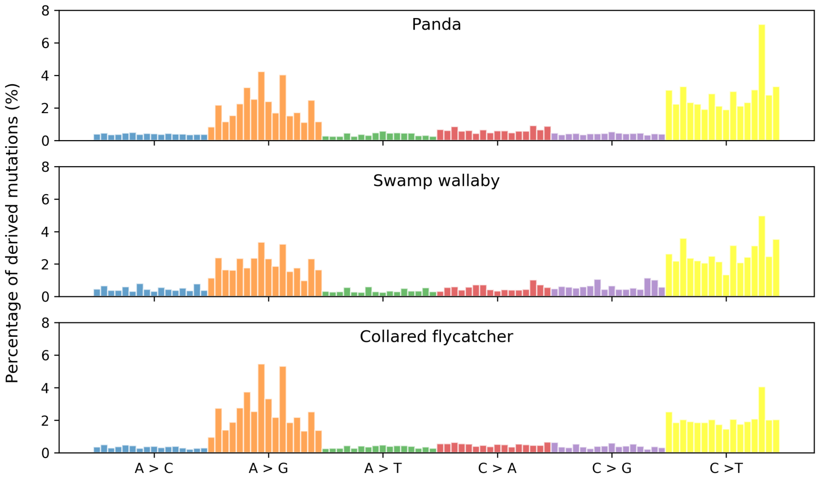

Figure 5. Differences in mutational spectrum across species. Mutations in each species are classified into 96 classes defined by the substitution type and sequence context immediately 3' and 5' to the mutated base (ie. first column

represents transitions from AAA to ACA). The six types of substitutions are displayed in different colours. The mutation types are on the horizontal axes, whereas vertical axes depict the percentage of mutations attributed to a specific mutation type.

The last paragraph has put into question the validity of the molecular clock. To make things even more complicated, there are other known variables influencing the mutation rate. The *generation-time effect* states that organisms with shorter generation time evolve faster, as they copy their genomes more frequently and therefore have more DNA replication errors per unit of time (ie. the human and fruit fly generation times are 25 years and 50 days respectively, so a human generation is equivalent to 180 fruit fly generations). The theory has often been invoked to reconcile the *fast* human mutation rate based on calibration with the fossil-record ($1 \cdot 10^{-9}$ mutations per year) with the *slow* pedigree-based estimates ($0.5 \cdot 10^{-9}$ mutations per year). An increase of the generation time toward the present, could have led to a recent slow down in the yearly mutation rate (Scally and Durbin 2012). Furthermore, mutations are known to appear at different rate in males and females due to the difference in germ cell divisions (Kong et al. 2012), which makes the average age of each parent and the onset of puberty important variables (Gao et al. 2016). Additionally, mutations occur with different probabilities depending on the DNA sequence context (Moorjani et al. 2016). We can go even further and look at the rate of each mutation type and classify them by the flanking nucleotides (ie. the DNA sequence AGC mutating to ATC is classified as AGC->T). Even at this level we see that different mutation types appear at different rates, and that these values change between species (figure 5). It must now be clear to us that mutations are not random. As we have seen, they are determined by a vast amount of variables (some of them overlooked here; methylation, replication time, compaction status in the nucleus, etc.). They might appear to be *effectively* random, in the sense that they do not have *intention* or *purpose* to appear in a given gene, but they ultimately occur in a deterministic manner.

My intention was not to make the reader fall into despair. If the mutation process is so complex, how are we supposed to ever fully understand it? We are far from a comprehensive understanding of how mutations occur, but our knowledge is steadily increasing, often by revising previous studies with overlooked results (Gao et al. 2018). Scientists have also come up with new creative ways to exploit DNA resources to learn about mutation. Some of these fruitful efforts involve studying mutational shifts in cancer cells (Alexandrov et al. 2013) or contrasting the level of mosaicism in somatic mutations (Ju et al. 2017). Additionally, we can be pragmatic and *use* mutations without fully understanding the underlying process generating them (ie. I take the metro every day, but I have no clue how trains work). In figure 6 we can see how the effective population size of orang-utans has changed through time (Prado-Martinez et al. 2013). In order to obtain the time estimates, the authors used the molecular clock and converted divergence units to years using the human mutation rate. By the time of the publication, there were two recognised orang-utan species, one living in Sumatra and the other in Borneo (maybe there are three species? (Nater et al. 2017)). The Toba supervolcanic eruption, one of the largest known eruptions on Earth, occurred 75.000 years ago at the present-day site Lake Toba in Sumatra. It affected the climate of the whole globe, albeit the effects were particularly strong in Sumatra. Figure 6 clearly depicts a sudden population decline in Sumatran orang-utans around 70.000 years ago, coinciding with the time of the Toba eruption. It could all be coincidence. Nonetheless, the stepiness of the population decay and the almost perfect time overlap suggests otherwise. Thus, despite the fact that the molecular clock can deviate due to multiple reasons (mainly because of natural selection and/or changes in life history traits such as the generation time), it seems to be very useful, and can provide a good reflection of reality. As we have seen in multiple occasions, a simple model can sometimes capture well the complexities underlying reality. However, the robustness of the molecular clock should not prompt us to give up on the exploration of mutation. Because mutation is the ultimate source of all variation, our understanding of evolution will always be incomplete without a deep understanding on how mutations occur and prevail.

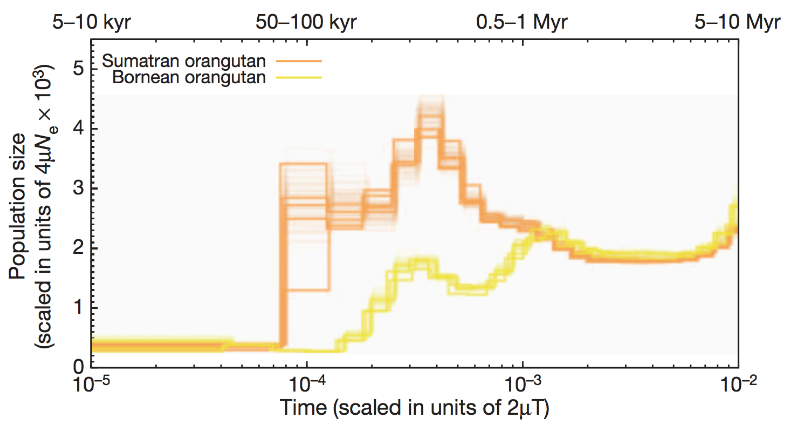

Figure 6. Inferred historical population sizes in orang-utans using a pairwise sequential Markovian coalescent analysis. Figure 3 of Prado-Martinez et al. 2013.

1.5. The coalescent theory

In section 1.3 we developed a theory of genetic drift based on the Wright-Fisher model. Ultimately, we would like to relate theory to genetic data, so that we can learn about populations from their DNA sequences. For instance, we know that the violetears in Tepui differ, on average, in 0.05% sites of the genome. We also know that this value is higher in the population in the Imataca rainforest (0.09% sites of the genome). What do these numbers tell us about the two populations? We can use the Wright-Fisher model to answer this question, although it is often difficult and mathematically awkward to do so. However, in the early 1980s a new theory of population genetics was developed by mathematicians such as Kingman (Kingman 1982), and biologists such as Hudson (Hudson 1983). This theory, called the *coalescent theory,* thinks of genealogies proceeding *backward* in time, in contrast to the *forward* in time mindset of the Wright-Fisher model. This manner of thinking was shown to be very powerful, and established the basis for many modern population genetics methods.

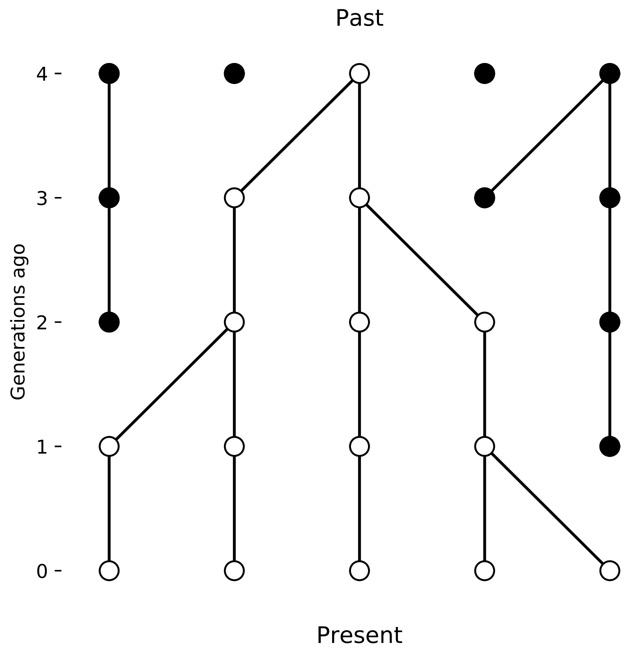

Figure 7. Diagram showing paths of ancestry of a set of alleles sampled in the present (generation 0). Circles in black represent alleles with no descendants in the present.

To understand how the coalescence process works, first consider what happens in figure 7 as time moves forward in time. As we saw in section 1.3, new generations in a Wright-Fisher model are formed by randomly sampling alleles from the current generation (generation 4 if we go forward in time). After choosing the alleles, we place allele copies in the next generation and connect each allele with its parent. This stochastic process inevitably leads to some alleles not leaving descendants in the next generation, thus becoming extinct in the population (colored in black in figure 7). Indeed, if we wait enough generations, all the alleles in the population must be copies of a single *ancestral* parent. In the case of figure 7, the Wright-Fisher model reaches this point in 4 generations.

Consider now the reverse process. In generation zero, each allele must choose a parent from the previous generation. Sometimes two alleles will choose the same parent (see the two alleles in rightmost side of the figure). In such situations we will say that the two alleles have *coalesced* into a single ancestral allele. As we go further back in time, the number of ancestral alleles has to either remain the same or decrease, and each reduction in the number of alleles is called a *coalescent event*. Similarly to the Wright-Fisher model, if this process is repeated enough times, all the alleles coalesce into a single *most-recent common ancestor* (MRCA). Figure 7 illustrates one reason why coalescent thinking is so powerful. If we want to study the process in figure 7 forward in time by means of a computer simulation, our program must generate the ancestry of all the individuals. Because we do not have *a priori* information on which alleles will leave descendants, we must *remember* them all. The 10 alleles colored in black in figure 7 appear to be irrelevant for the population in the present, but that remains unknown until the point they leave no descendants. On the other hand, if we were to study the same process backward in time, none of the alleles tracked by our computer would be wasted. Any allele in any generation must ultimately trace back to an allele in the present. Such alleles are colored in white in figure 7, and they add up to 15 alleles. In other words, the forward simulation wastes nearly half its time generating alleles (10 out of 25) that are of no interest because they do not contribute to the ancestry of the contemporary population. This might not seem a great price to pay in this example, but in samples of thousands of alleles, the vast majority of lineages simulated in a forward direction end up not being used.

The time needed to reach the most recent common ancestor (tMRCA) is of central interest in population genetics, as it allows to link DNA sequences and time. In order to introduce time into the picture, we will first consider a sample of two alleles in a population of size N. How long should we wait until the two alleles coalesce, or in other words, share a parent? Since parent alleles are chosen randomly, the probability of both alleles selecting the same parent in a generation is simply 1/2N. Imagine rolling a die two times. The chance that the second throw results in the same outcome as the first is 1/6, and is independent of the outcome of the first throw (ie. given that we roll a 5 in the first throw, the chance of we getting a 5 a second time is 1/6). Similarly, because in a population there are 2N potential parents equally likely to be chosen, the probability of two individuals having the same parent in the previous generation is 1/2N. From this we can reasonably conclude that the average time we have to wait for the two alleles to *meet* and coalesce is 2N generations.

You may have noticed that I did not clearly answer the last question. What kind of time measure is 2N generations? As we briefly discussed in section 1.4, generations are the time-unit of evolution, something that should make good intuitive sense. Consider the following analogy. You have a coffee machine that makes mediocre coffee, but you keep it because from time to time it produces the *best* coffee. After years of experience, you realise that the perfect coffee only comes out once every 100 preparations in a completely random fashion. How many times on average would you have to use the machine to get the perfect coffee? Since the perfect preparation only happens with a probability 1/100, you would have to wait 100 preparations on average. Now, how long would that be in days? Well, that very much depends on how often you prepare coffee. If you do it every two days, you would wait 200 days on average. Equivalently, generations can be converted to conventional time-units by estimating how often organisms reproduce (ie. 25 years in humans). Therefore, two alleles in a population of 500 individuals with a generation time of 2 years will coalesce, on average, in 2.000 years (= 1.000 generations * 2 years per generation).

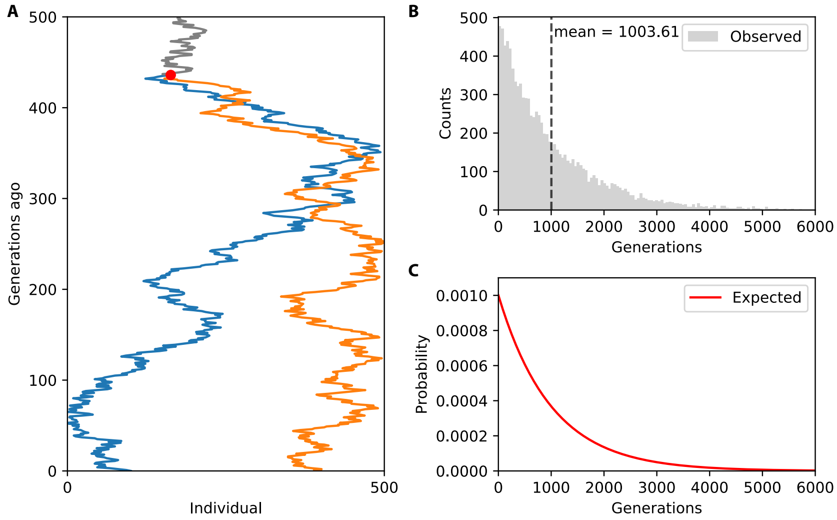

Figure 8. Coalescence in a sample of two alleles. (**A**). The ancestry of two alleles in a population of 500 diploid individuals. The MRCA is found in roughly 400 generations. The coalescent event is highlighted in red. (**B**) Distribution of the number of generations until two alleles find the MRCA in 10.000 simulations of a population size of 500 diploid individuals. (**C)** Continuous approximation of the expected probability of coalescence in a population of 500 diploid individuals.

In figure 8, we can see the distribution of the number of generations until two lineages find a MRCA in a population of 500 diploid individuals. The shape of the distribution might be surprising, as it does not show the typical bell shape of a normal distribution. The coalescence process follows a decreasing exponential distribution, where recent coalescence times are more common than ancient ones. This means that coalescent events in the first generation are the ones most likely to happen, even when the size of the population is very large. How is this possible? Consider two alleles in a model diploid population of size 500. As we have seen, the probability of coalescence in the first generation is 1/2N (0.001 = 1/[2*500]). The probability of coalescence in the second generation is the same as before (1/2N), but now we must also consider the probability that the alleles *have not coalesced* in the first generation (1 - [1/2N]). Therefore, the probability of coalescence in the second generation is [1/2N] * [1 - 1/2N], which is slightly lower than the probability in the first generation. The probability of every successive generation is lowered by the fact that coalescence must not have happened in any previous generation (if *r* is the current generation, then this probability equals $[1 - 1/2N]^{r-1}$). Thus, the probability of coalescence in any given generation can be easily computed with Pr(MRCA in *r* generations) = $[1 - 1/2N]^{r-1} * [1/2N]$. If we come back to the computer simulations shown in figure 8, we can see that indeed the probability of coalescence gradually decreases and asymptotically approaches zero. Nonetheless, the *mean* number of generations until the MRCA is in good accordance with the expectation, and is equal to 2N generations.

As we saw in section 1.3, the *Ne* of a population equals to the number of individuals in a Wright-Fisher model with an equivalent amount of genetic drift. It must now be clear that genetic drift and coalescence are very similar processes, only differentiated by a switch in the lense in which we see time (forward and backward respectively). Keeping this in mind, the *Ne* of a population should also reflect the average tMRCA between two genes in the population (2*Ne* generations). In large populations, coalescence and genetic drift work slowly, and we must wait many generations until two individuals share a parent. In contrast, coalescent events and drift act more rapidly in small populations, so we tend to find an ancestor much faster. The fact that the size of a population directly gives information about *time* is abstract and difficult to grasp at first. However, one of the most important ideas in population genetics particularly exploits this idea. In section 1.4, we saw that the expected number of genetic differences between two diverging lineages is equal to 2*tu* (*t* being the split time in generations and *u* the mutation rate per generation). Since under a neutral model the expected coalescent time is equal to 2N generations, we can substitute *t* with 2N. From this we can conclude that the number of mutations separating two gene copies is simply 4N*u* (= 2 * 2N * *u*). Population geneticists are so excited about this result that they have devoted a greek letter entirely to this, and commonly write θ = 4N*u*. Sooner in this section we saw that the average number of genetic differences within the Tepui and Imataca violetears are 0.02% and 0.09% respectively. Armed with this new result, we can estimate the effective population size of each population. Assuming a mutation rate of $4 \cdot 10^{-9}$ per base pair per generation in violetears, the Tepui population must have an effective population size of 12.500 ($4N * 4 \cdot 10^{-9} = 0.0002$; $N = [0.0002 / 4 \cdot 10^{-9}] / 4$; $N = 12.500$). Applying the same calculation to the Imataca population we obtain an effective population size of 56.000. These numbers might have little to do with actual population sizes, so we must be cautious and should not jump into concluding that there are more individuals in Imataca than in Tepui. It could also be true that the Imataca population has maintained more genetic variability for other reasons (ie. it has existed for longer without population bottlenecks).

Is the expected tMRCA influenced by the number of alleles? Would we have to wait more generations if we consider 5 alleles? So far we have only focused in two gene copies in a model population with N diploid individuals. As we can see in figure 9, adding more alleles will always increase the number of generations needed to reach the MRCA. Imagine 5 alleles in a population of 500 diploid individuals. Since now there are more than 2 alleles under consideration, the time to reach a coalescent event is much shorter. More specifically, there are 10 possible different combinations between the 5 alleles (the number of pairwise comparisons in a set of *k* elements is given by *k**[*k*-1]/2), so on average we have to wait 10 times less time (2N/10, in our toy example 2*500/10, thus 100 generations). This first coalescent event has reduced the number of alleles to 4, so the time needed for a second coalescence event is simply 2N/(4*(4-1)/2), which equals to 266 generations. This process is recursively repeated until there are 2 alleles in the pool, situation in which we go back to square one and the expected time to the MRCA is 2N generations. Once there is a single lineage left, we have found the MRCA of all the alleles. Following this chain of thought, we can derive the expected time to the MRCA for a sample of n alleles as:

$$\text{Expected time to the MRCA} = \sum_{k=1}^{n} \frac{2N}{\frac{k(k-1)}{2}}$$

Although we have already discussed the coalescent in a sample of multiple alleles, I will briefly describe a famous analogy by Joseph

Felsenstein that wonderfully illustrates the coalescent process (Felsenstein 2004), paraphrasing: "imagine a box containing hyperactive, indiscriminate and voracious bugs. We put *k* bugs into the box. They run in the box without paying any attention to where they are going. Occasionally two bugs collide. When they do, one instantly eats the other. Being insatiable, the winning bug then resumes running as quickly as before. It is obvious that the number of bugs falls from *k* to *k*-1, to *k*-2, as the bugs coalesce, until finally only one bug is left. The number of pairs that can collide is *k*(*k*-1)/2, so collisions get rarer when the number of bugs is reduced. If there are 2N *places* in the box that can be occupied, the probability of collision will be proportional to the size of the box. A box with twice as many *places* will slow the coalescence process down by a factor of two. So a simpleminded physical analyses of the bugs-in-a-box process will have the coalescent as the probability distribution of its outcomes".

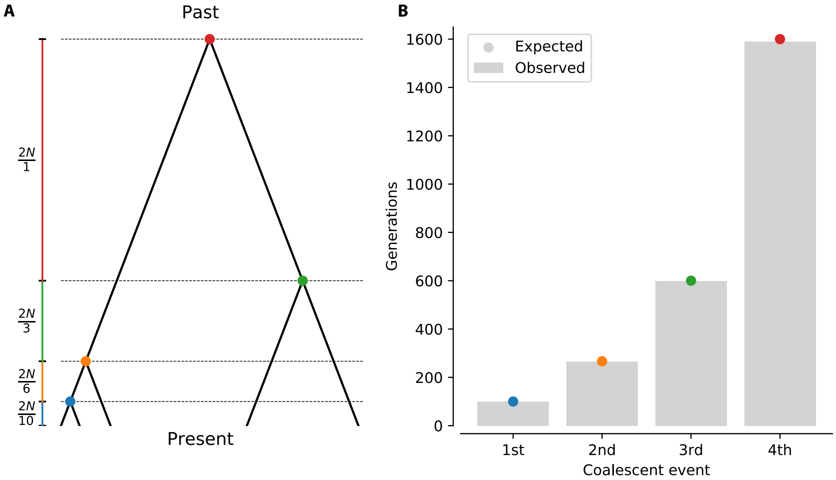

Figure 9. Coalescence in a sample of multiple alleles. (**A**) Gene tree representing the coalescent process in a sample of 5 alleles. In any each generation, if there are *k* alleles present, the expected time back to the next coalescence is given by 4N/[*k*(*k*-1)]. (**B**) Bars show the mean number of generations until each coalescence is produced in 10.000 simulations of a sample of 5 alleles in a population of 500 diploid individuals. Dots represent the expected number of generations for each coalescent event.

Behind the process described in the last paragraph there are several interesting features of population genetics. First of all, the characteristic tree-like shape of phylogenies can be explained by the coalescence of multiple alleles. We have seen that the time required for a *collision* or a coalescent event to happen is proportional to the number of alleles under consideration (ie. the more bugs in the box, the faster a first encounter will happen). For this reason, gene trees tend to show branches that are short near the leaves and get longer when approaching the root. This is explained by the fact that most time to the MRCA is spend waiting for the very last alleles to coalesce (figure 9). This observation is true when the population size remains constant, but can be distorted when the size fluctuates heavily. A population with an expanding population size will display really long branches near the leaves which will get increasingly shorter near the root (star-like shape). The distribution of coalescence times along the genome is a great resource to learn about the genetic history of a population, and we shall discuss it more in-depth in section 1.7. Another important characteristic is how the time to the MRCA changes if we increase the sample size. By performing 10.000 simulations of 5 alleles in a population of 500 diploid individuals (figure 9), we can empirically demonstrate that the time to the MRCA averages at 1600 generations. What would be the expected time to the MRCA in a sample of 10 alleles? We can answer this question without the computer simulations using the equation discussed previously. The number of generations to the MRCA would be around 1770. You may think this is a discrete increase of waiting time given that we have doubled the number of alleles. If we increase it even more to 1000 alleles, the number of generations would be 1995. The number of generations would asymptotically approach 2*2N (2000) as we increase the number of alleles. So a large sample will, on average, have just a slightly older MRCA than a small sample. This illustrates an important principle about the coalescent process; there is less and less additional information as more and more sequences are sampled. Much of the genetic variation can be captured by studying a handful of individuals. These are very good news for any researcher, as sequencing large cohorts of individuals can be unfeasible in most organisms.

1.6. Population subdivision

We have so far considered models for a single population with random mating. In reality, random mating rarely happens, as individuals often choose their mates in a deterministic manner. Generally speaking, pollinisation between plants living 100 meters away from each other is more likely to happen than between plants living 100 kilometers away from each other. There can be other factors than geography deviating a population from random mating. *Assortative mating* is a mating pattern in which individuals choose their mates according to their phenotype. In humans, assortative mating has been reported for traits such as height and educational attainment (Stulp et al. 2017; Yengo et al. 2018). Other assortative traits, such as religious affiliation and ethnicity, have no genetic component but are reflections of cultural preference. Interestingly, cultural assortative mating is not particular to humans, but has also been found in orcas (Foote et al. 2016) and dolphins (Kopps et al. 2014). When any of these factors occur in a population and there is not random mating, we say there is *population subdivision* or *population structure*. Population subdivision is important for understanding evolution and the effects of genetic drift and natural selection, and it is also of direct importance in *conservation biology*. Researchers often use genetic markers to determine which groups of individuals should be considered separate genetic units, considerations that are later taken into account when deciding policies on the management and conservation of species. This section will focus on the genetic effects of population structure and on methods for detecting and quantifying it.

An important genetic consequence of population structure is a reduction in the fraction of heterozygous individuals relative to the expectation under random mating (the so called *Wahlund effect*). Consider all the individuals in a population as a giant family tree. Under random mating, most individuals in the pedigree are very distant cousins, while a few individuals are more closely related (ie first cousins, siblings). The degree of relatedness between pairs of individuals is called *kinship*. When related individuals share alleles that descend from a common ancestor, we say that the alleles are *identical by descent*. This definition does not speak for itself, as all existent DNA sequences must coalesce into a common ancestor at some point, even if we have to go back to *LUCA (Last Universal Common Ancestor).* In order to avoid this vacuous meaning, it is often chosen an arbitrary time in the past at which every allele is assumed to be distinct to every other allele. Therefore, for two individuals to share an identical by descent allele, the alleles must have *survived* the process of Mendelian segregation (*meiosis*). From this it follows that individuals fewer meiosis away from each other will share more alleles. When two individuals mate and are more closely related to each other than two random individuals drawn from the population, we say that there is *inbreeding*. Note that we are describing all over again the process through which population structure arises (deviation of random mating). How do we get a decay of heterozygosity from inbreeding? Inbred offspring will more often carry two alleles identical by descent than offspring from two random individuals in the population. An illustrating way to see it is to think of it as a loop. Inbreeding is promoting the reunion of two copies of the same allele through different paths in the pedigree. Organisms in the same *subpopulation* will often share recent common ancestors, thus their mating will increase the number of inbreeding loops. This inevitably leads to an increased likelihood of being homozygous, thus to a reduction of heterozygous individuals.

The most commonly used statistic for quantifying population subdivision, Wright's F_{ST}, measures the extent of inbreeding in a population. Consider a biallelic locus with alleles *A* and *a* in two populations of the same size. If the frequency of *A* is f_{A1} and f_{A2} in each population respectively, the frequency of *A* when pooling both populations is simply the average frequency ($f_A = [f_{A1} + f_{A2}] / 2$, same applies for f_a). Note that in populations with unequal size we must get f_A by weighting the size of each population. As we have seen in section 1.2, the Hardy-Weinberg law allows to link allele frequencies to genotype frequencies under random mating. Therefore, the expected number of heterozygotes if there is no structure between the two populations would be $H_T = 2f_Af_a$. H_T represents the heterozygosity we would expect if the pooled population is in Hardy-Weinberg equilibrium. However, we could also estimate this parameter by averaging the heterozygosity between the two populations: H_S = $((2f_{A1}f_{a1}) + (2f_{A2}f_{a2})) / 2$. Armed with this new result, we could quantify how different H_T and H_S are. F_{ST} is defined as the difference between H_T and H_S standardized by H_T; $F_{ST} = (H_T - H_S) / H_T$.

To understand what F_{ST} measures, I find enlightening to think of a highly differentiated locus in two populations (ie. $f_{A1} = 0.8$ and $f_{A2} = 0$). The average pooled frequency ($f_A = 0.4$) would suggest that almost half of the individuals should be heterozygous (2*0.4*0.6 = 0.48). However, by averaging the fraction of heterozygous individuals in each population we get a much lower value ([2*0.8*0.2 + 2*0*1]/2 = 0.16). The reason behind this difference is that all the *A* alleles are *trapped* in population 1. Similarly, most of the *a* alleles are found in population 2, so both alleles hardly ever coexist to form heterozygotes. Indeed, by computing F_{ST} we get a value of 0.66 ([0.48 - 0.16]/0.48). This means that the average heterozygosity is a 66% of the total heterozygosity, a very substantial reduction. Wright provided *rule of thumb* guidelines to interpret F_{ST}, with values ranging from 0 - 0.05, 0.05 - 0.15, 0.15 - 0.25 and >0.25 indicating little, moderate, great and very great genetic differentiation respectively.

Beyond Wright's original description of F_{ST}, the measure has been conceptually redefined in many ways (Nei 1973; Weir and Cockerham 1984; R. R. Hudson, Slatkin, and Maddison 1992). The coexistence of multiple expressions to compute F_{ST} has often lead to confusion and mixed results (ie. see differences between F_{ST} values in 1000 Genomes Project Consortium 2012 and HapMap3 discussed in Bhatia et al. 2013). Another important aspect to remember about F_{ST} is that it is a *relative* measure of divergence. As discussed earlier, F_{ST} highly depends on the heterozygosity in each subpopulation (H_S in our nomenclature). From this it follows that F_{ST} will tend to be inflated when diversity within populations is low. One can easily appreciate this effect when computing F_{ST} between populations with different effective population size (which is a reflection of the population diversity as seen in section 1.3). Consider two populations A and B which share a common ancestor with population C at exactly the same time, meaning that A and B are genetically equidistant to C. We also know that diversity within population A is much higher than within population B. After computing F_{ST} between A - C and B - C, we would see that F_{ST} is higher in the later, even though both A and B have equally evolved from the ancestor. This effect can also lead to misinterpretations of genetic data (Cruickshank and Hahn 2014), and must be remembered when interpreting F_{ST} values. In order to circumvent this limitation of F_{ST}, one can use measures of divergence that are independent of the levels of diversity within populations. D_{XY} (Nei and Li 1979), often simply referred as *divergence,* is a measure of *absolute* divergence. It is easily computed and very much related to π (see section 1.3). D_{XY} equals to the average number of pairwise differences between sequences from two populations, excluding all comparisons between sequences within populations. I have used this measure of genetic differentiation on a daily basis during my thesis, and for that reason I will further comment on it in section 1.7.

What other tools are there to explore population structure? Principal component analysis (PCA) is one of the most widely used methods. Technical descriptions of PCA can be found elsewhere, however, its key feature is that it can be used to project samples onto a series of *components*. Each component is an orthogonal axis constructed by a linear combination of several genotyped loci. How can one make linear combinations out of categorical values such as genotypes? Diallelic genotypes can be converted to numerical variables simply by choosing one allele and counting it in the genotype (if we choose *A* in a diallelic *Aa* locus; *AA, Aa* and *aa* are converted to 2, 1 and 0, respectively). The axes are then build such that the projection of samples along the first component explains the greatest possible variance in the data among all possible components. Likewise, projection of samples onto the second axis maximizes the variance for all possible axes perpendicular to the first and so on for the subsequent components. Intuitively, PCA is extracting the underlying structure of a high dimensional dataset, compressing it and stripping away any *unnecessary parts* (at least for our current purposes). The output components are the directions in the data where there is most variance, the directions where the data is most spread out. Thus, the distance between samples in the PCA is a reflection of their genetic distance. Typically, the samples are plotted along the first two components, which often mirror their geographic distribution (Novembre et al. 2008; Lao et al. 2008). This information is commonly used in large scale association studies to correct by genetic stratification (ie. south Europeans are on average shorter than northern Europeans. If we sequence a bunch of Europeans and we see that an allele is clearly associated with height, is the association real or just a product of population structure within Europe?).

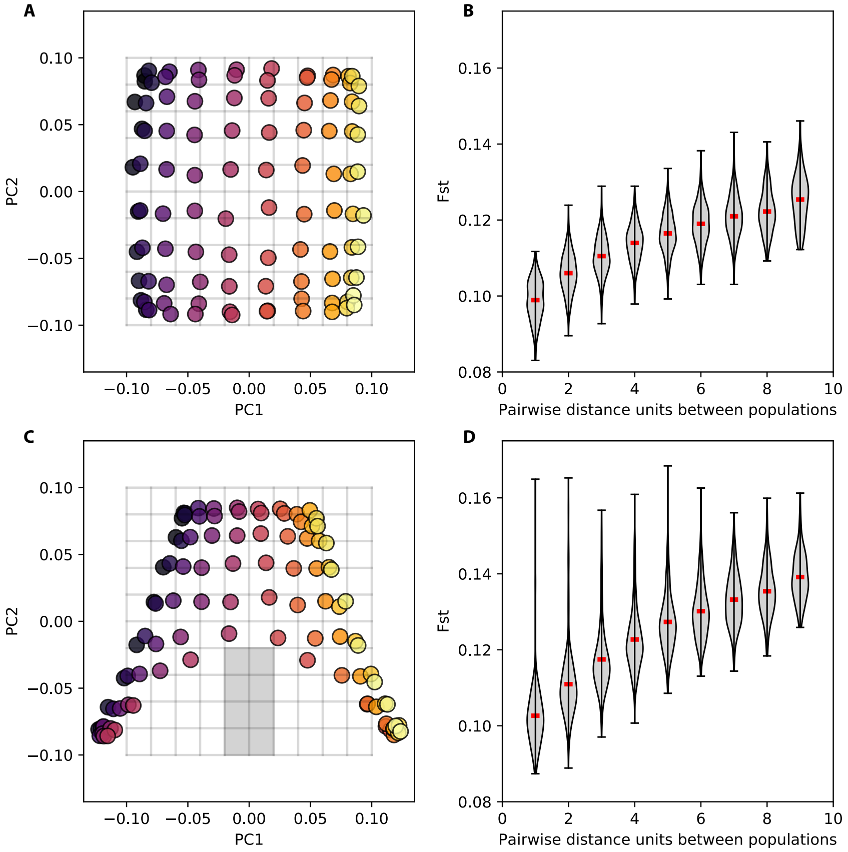

Figure 10. Stepping stone model of migration. (**A**) and (**C**) PCAs of a two-dimensional stepping-stone model of migration where 100 subpopulations are distributed in a 10x10 habitat. Adjacent populations are connected by a unique constant migration rate. Migration edges between populations colored in gray in C are set to zero. 100.000 independent polymorphisms were simulated using scrm

(Staab et al. 2015) and a single haploid individual was sampled from each population. (**B**) and (**D**) Distributions of Wright's F_{ST} for populations separated by differents amount of migration *jumps*. B and C correspond to the models depicted in A and D respectively.

PCA is very appealing due to multiple reasons. First, it is an easily implemented method, fast and not computationally intensive. Its ability to group or separate samples in a striking visual manner is especially useful to get a first impression of genetic data. Additionally, it is a *model-free* method, it does not rely in any underlying demographic model. In order to put PCA into test, consider a population of insects living in a vast archipelago. This hypothetic archipelago is composed by a grid of islands (10x10 as shown in figure 10). As you can imagine, insects most commonly mate with other insects within their island. Since each island has its own endemic subpopulation, this scenario pictures a structured population without random mating. However, from time to time an insect migrates to a neighboring population. This event strictly happens between neighboring islands. The model described above is known as a *stepping stone model* (M. Kimura and Weiss 1964). If we apply PCA to the stepping stone model we observe a good correlation between genetic and geographic distance (figure 10A). Actually, geography is not explicitly incorporated in the model, although a good proxy for it is the necessary migration *jumps* to connect two islands. Since mating between individuals from different islands is rare, allele frequencies in the subpopulations gradually diverge from each other due to random genetic drift. However, mating between neighboring islands still happens from time to time, thus allele frequencies in islands close in space are more similar to each other than distant ones. This explains why the 10x10 archipelago is clearly recognisable when plotting the first two components. Note that distortion near the extremes of the habitat is expected. Islands at the extremes of the archipelago have fewer migration edges than islands in the center, so they appear to be more similar to each other. This effect is especially obvious in islands close to the vertices. Similarly to PCA, F_{ST} measures also reflect well genetic distance between populations. F_{ST} is lowest in comparisons between neighbor islands, and values gradually increase with geographic distance (figure 10B). This tight intercorrelation between geography and genetics is called *isolation by distance* (Wright 1943; Malécot 1948). Patterns of isolation by distance are not only found in theoretical models. In humans, there is a clear pattern of isolation by distance, most likely explained by the sequential colonization of the world after the Out-of-Africa (Handley et al. 2007).

However, real populations often encounter *barriers* that impede their reproduction. These barriers can be physical, but can also be constituted by differences in behavior, morphology, culture etc. For instance, in a mountainous area of a terrestrial species' range, a pair of individuals may be more divergent from each other than a pair of individuals separated by the same distance in a flat and open area of the habitat. In such situations, genetic distance does not bear such a clear resemblance with geography. In figure 10C, I plot the first two components in a PCA of a stepping stone model with a barrier between the islands colored in gray. As expected, distances separating islands along the barrier are much greater than in the rest of the habitat. F_{ST} shows a larger variance than before (figure 10D), with more values at the high-end of the distribution, most likely the result of comparisons between islands at each side of the barrier. The populations in the model no longer show a simple pattern of isolation by distance. Isolation by distance can be statistically tested by means of a Mantel test, which evaluates the association between two matrices (ie. spatial coordinates and genetic distances). Other methods have been developed to explore how genetic divergence covaries with a spatial continuum (Bradburd, Ralph, and Coop 2016; Petkova, Novembre, and Stephens 2014; Al-Asadi et al. 2018).

Another family of tools try to describe population structure by clustering individuals into different groups. The representation of such clustering has become a distinctive feature in many projects in genomics (see barplot in figure 11 for an example). In order to understand how individuals in a population can be sorted into groups, we will expand on our hypothetical study in violetears. Imagine we managed to sequence 90 violetear genomes from regions within the Imataca forest and the neighboring Tepui (see map in figure 11B). Now let us consider we have *a priori* information about allele frequencies in each of these populations. In other words, we know that the *A* allele in locus *l* has a frequency of 0.9 in Imataca ($f_{Ima,l}$) and a frequency of 0.1 in the Tepui ($f_{Tep,l}$). With this knowledge, we can compute the probability of an individual belonging to each population under Hardy-Weinberg equilibrium. For example, if one of the genomes in our dataset carries two *A* at locus *l*, the matching probability for Imacata and Tepui would be 0.81 and 0.01, respectively ($f_{Ima,l}^2$ and $f_{Tep,l}^2$). This simply follows from Hardy-Weinberg equilibrium, and can be extended to other genotypes using the formulae explained in section 1.2. Notice that we are computing the probability of drawing a genotype given the allele frequencies in each population (ie. P(g=*AA*|pop=Imataca) = $f_{Ima,l}^2$). Ideally we would like to know what is the probability of belonging to a population given the genotype P(pop=Imataca|g=*AA*). Here Bayes' rule comes in handy. We can compute P(pop=Imataca|g=*AA*) by assuming that the prior probability that an individual comes from Imataca or Tepui is the same (P(pop=Imataca) = P(pop=Tepui) = 1/2). We then have

$$P(pop = Imataca|g = AA) = \frac{P(g=AA|pop=Imataca)P(pop=Imataca)}{P(g=AA|pop=Imataca)P(pop=Imataca) + P(g=AA|pop=Tepui)P(pop=Tepui)}$$

Probabilities for multiple loci can then be combined by multiplying the match probabilities for each locus. This type of multiplication is only valid if the alleles under consideration are not linked together. That is the reason why a pruning of linkage disequilibrium must be performed before applying this kind of clustering methods (ie. only keeping a single variant in a block of polymorphisms segregating together).

While it is great to be able to assign individuals to particular populations, our current method is dependent on *a priori* information of the allele frequencies. Most probably, allele frequencies in violetears are unknown at the time of our study. Thus, we could greatly benefit from a method that does not rely on *a priori* information and is able to generate clusters only using the genotype data. Particularly, we wish to assign our individuals into *K* unknown populations. This process can be achieved by randomly distributing the individuals into *K* populations and then; (I) allele frequencies in each population *k* are estimated given the current assignments and (II) each individual is reassigned to a population *k* with a probability given by the bayesian scheme described above. Steps I and II are iterated many times (*Gibbs sampling*), which will lead to gradual sorting of individuals according to their genetic make-ups. The output of this process gives us the most likely distribution of individuals into *K* population clusters. Further improvements to the method can be implemented, for instance allowing the assignment of individuals to multiple populations with different mixture proportions. The clustering algorithm is typically run for multiple values of K which allows to get a better picture of hierarchical structure in a population. STRUCTURE (Pritchard, Stephens, and Donnelly 2000) was the pioneer method implementing this idea, and since then multiple extensions and improvements have been developed (Alexander, Novembre, and Lange 2009; Tang et al. 2006).

In figure 11, I show the results of ADMIXTURE and a PCA in 90 simulated violetear genomes. Figure 11A depicts the true phylogeny, only known because the genes are product of a computer simulation. The ADMIXTURE results hint at the genetic distinctiveness between Imataca and Tepui, since it correctly assigns all the individuals to different clusters at *K*=2 (figure 11C). Similarly, the first component in the PCA clearly separates the Tepui from the rest of populations in Imataca (figure 11D). By gradually increasing the value of *K*, we start to detect substructure within Imataca; first the separation between north and south becomes apparent, and differences between east and west start to appear at higher values of *K*. Similar observations can be appreciated in the PCA. As expected, the methods work well at detecting the underlying structure in the population. Furthermore, given that we have geographic coordinates for the populations (figure 11B), we can also hypothesize about how populations interact with each other. Individuals in the Tepui seem to be isolated from the violetears in Imataca, even though they live fairly close to the SE Imataca. That means that they rarely, if ever, mate with each other. Maybe differences in altitude are playing a role, although one cannot tell without knowledge of the terrain topology, etiology of populations, etc. In any case, while the Imataca subpopulations might be in isolation by distance, there seems to be a clear barrier of reproduction between Tepui and the rest.

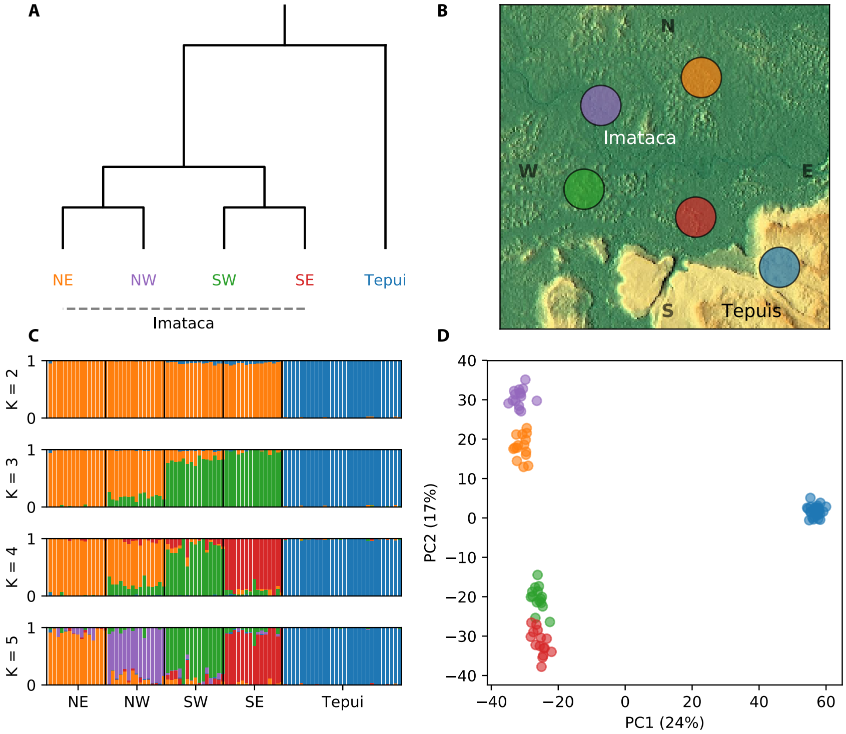

Figure 11. Population structure in violetears. Note that genomes are not real, they have been simulated using scrm (Staab et al. 2015). All populations have a constant effective population size of 20.000 diploid individuals. Split times between Tepui and the Imataca ancestor, South and North Imataca and populations within either Imataca's populations happened 80.000, 40.000 and 20.000 generations ago respectively. Constant migration rate within North and South Imataca and between NW and SW Imataca happens at rates 1.000 individuals and 250 individuals per generation respectively. A single pulse of migration from Tepui into the South Imataca ancestor happens 30.000 generations ago (800 individuals are instantly replaced, which represents a 4% of the population). These parameters are fixed and will also be used in section 1.7. (**A**) Schematic tree showing the true phylogeny. (**B**) Hypothetical geographic distribution of the violetear populations. (**C**) ADMIXTURE analysis of 90 simulated violetear genomes. Each individual is represented with a vertical line

and colors represent the proportion of genetic ancestry that is due to any of *K* populations. (**D**) PCA of the 90 simulated violetears.

Here we must be careful. As I said a few lines ago, it is tricky to draw conclusions from the current results. For instance, consider that our hypothesis is that the genetic relationship between populations is explained by a migration model (ie. Tepui and SW Imataca are more genetically different than SE Imataca and SW Imataca because there are less migrants in the former). We might be in the wrong, as there are other models that can explain the data. Imagine a series of divergence events, each one generating two daughter populations from the original population. Also assume that each new population occupies a different location in space. Following this chain of thought, the ancestral population of all violetears may have originally split into the ancestors of Tepui and all Imataca. Subsequent divergence events derived into the northern and southern Imataca lineages, from which all the present-day population finally appeared. In this divergence model, populations close to each other are more genetically similar, albeit the reason behind it is not migration. The underlying phylogeny we have described makes some populations share more genetic drift than others, but ongoing migration is not needed to explain the data. There are probably endless combinations of parameters that would yield our results, thus we ought to be careful when interpreting them in genomics, especially in analyses such as STRUCTURE and ADMIXTURE (Lawson, van Dorp, and Falush 2018).

Altogether, we have seen that the so called first law of geography *'Everything is related to everything else, but near things are more related than distant things'* (Tobler 1970), has long been obvious to population geneticists. This principle offers far reaching possibilities not described here. For instance, one could try to guess the geographical origin of an individual solely using genetic information. The *spatial assignment* of individuals can be made with models that assign allele frequencies as continuum in a geographical space (Samuel K. Wasser et al. 2004; Yang et al. 2012; Rañola, Novembre, and Lange 2014). This approach has been used to identify local hotspots of ivory trade in Africa (S. K. Wasser et al. 2015), and could be used in other fronts to empower the conservation of endangered species. Alternatively, one can disregard non-sampled locations (thus not assuming allele frequency clines in space) and focus on well characterised populations. Recently, an ever-increasing number of companies offer *ancestry tests*, which exploit the population structure in humans to infer *where you are from genetically*. Note that this statement has very little meaning. Rather, the tests identify to which *present-day* population/s your genetic make-up has more affinity to. Even more so, the increasing number of human DNA sequences makes possible to discover the identities of people who participate in genetic research studies by cross-referencing their data with other publicly available information (Gymrek et al. 2013). While such ability might be useful in some cases (see the Golden State Killer case), it represents routes for breaching genetic privacy, a serious challenge that genetic projects will need to address in the near future.

1.7. Gene flow

In the previous section we have seen that population structure can lead to the genetic differentiation of populations. In a short time scale, the differentiation is mediated by the independent random fluctuations of the allele frequencies within each subpopulation. However, if the isolation lasts long enough, genetic differentiation can be dominated by the appearance of new mutations in each subpopulation. If the reproductive isolation and genetic divergence between the two lineages is large, their genomes may even become *incompatible*. What does this mean? How does genetic incompatibility between lineages evolve without simultaneously causing defects in the recipient lineages? Consider two loci in an ancestral population with genotypes *AA* and *BB* respectively. When the population is split into two, *A* evolves into *a* in one population and *B* evolves into *b* in the other. Imagine that *a* and *b* cause non-synonymous changes in two interacting proteins, impeding the correct association between the molecules. Even though *a* and *b* are mutually incompatible, the *a–b* interaction is not present in the subpopulations, thus the evolution of incompatibility is possible. This model of hybrid incompatibility is often referred to as the Bateson-Dobzhansky-Muller model. Altogether, this means that population structure can be a precursor of *speciation.* If reproductive isolation allows for many genetic incompatibilities to arise, the offspring of subpopulations might become sterile, at which point we would call them species rather than subpopulations. We are walking a thin line now, as the concept of species is controversial. Ernst Mayr provided the classical definition of species (Mayr 1963), which has shown to be impractical in some cases (ie. most would say tigers and lions are different species, although there are multiple reported cases of fertile *tigons*). The so called *Species problem* has been subject of heated debate among biologists. In my opinion, the classical concept of species is not very meaningful. Divergence between two lineages accumulates gradually, so picturing their relationship as a dichotomy seems unrealistic (as it often is said, a continuous scale of gray is more suitable than black or white). Haldane stated this idea very well, describing species as human construct and "*a concession to our linguistic habits and neurological mechanisms*" (Haldane 1941).

Conversely to population structure, migration is an evolutionary force acting against the genetic differentiation due to reproductive isolation. The exchange of genes between diverging lineages or *gene flow* has an homogenising effect, it holds the gene pools of subpopulations together and limits how much genetic divergence can take place. *Introgression* is a special case of gene flow where one *donor* population contributes alleles to the genome of a second *recipient* population. Delineating introgression from other types of gene flow can be complicated. As with the *Species problem*, it can all end up becoming a semantic discussion rather than a biological one. I find enlightening to think of introgression as a relative term; alleles at one locus introgress with respect to alleles at other loci. That is, for the above definition to be applicable, some portion of the gene pool of each of the hybridizing lineages must remain constant such that we can recognize that two distinct gene pools exist. Notice that there is another concept attached to this definition. For us to be able to recognise the two genetic make-ups, introgression must occur rarely, otherwise the genetic differentiation between lineages would be too diluted and indiscernible under constant gene flow.

The most obvious genomic footprint of introgression is very intuitive; two interbreeding lineages should share more alleles than two lineages under reproductive isolation. Going back to the hypothetical population of violetears, in section 1.6 we saw that Tepui is an outgroup to all the populations in Imataca (figure 11). Our intuition may tell us that if neither of the Imataca populations have had any gene flow with the Tepui after their split, then each of the Imataca populations should share approximately the same number of derived alleles with it. In order to test this hypothesis, one can count in how many positions of the genome North Imataca and Tepui share an allele, and then compare it to the number of positions where South Imataca and Tepui share an allele. To increase the sensitivity of this test, we can use the genome sequence of an outgroup (ie. Green hermit is a hummingbird outgroup to all violetears). For instance, we may see that (I) the South Imataca and Tepui populations share an allele *A* and (II) North Imataca and Green hermits share a second different allele *B*. Such pattern of allele sharing can be simplified as BABA (imagine each letter as a leaf in the tree under test; ((North Imataca, South Imataca), Tepui), Green hermit). As discussed earlier, we are interested in checking whether either Imataca population shares more alleles with Tepui, so we also want to pay special attention to the ABBA allele configuration. The *D*-statistic is a widely used measure in population genetics that estimates asymmetries between ABBA and BABA counts (Patterson et al. 2012). In figure 12 I show the results of the *D*-statistic in violetears. Negative and positive values are expected when there is an excess of ABBA or BABA respectively, while values close to zero denote an equal amount of each allele configuration. Thus, the data shows that South Imataca shares more alleles with Tepui than North Imataca, hinting at gene flow between South Imataca and Tepui after their split.

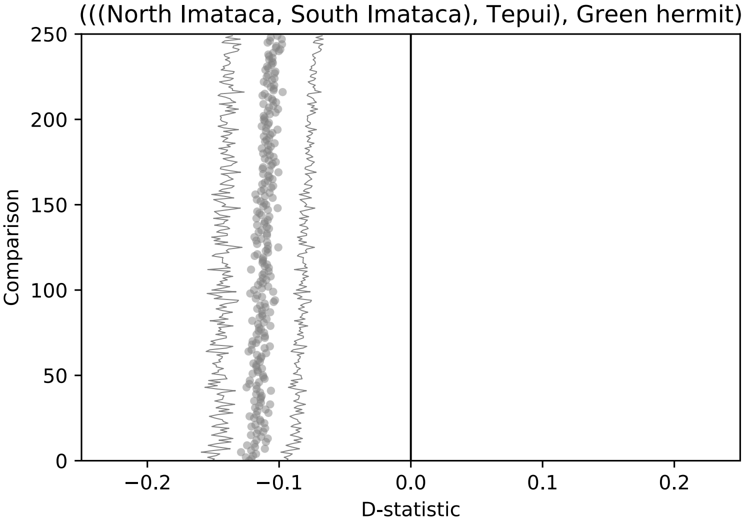

Figure 12. All possible tests of the *D*-statistic in the tree (((North Imataca, South Imataca), Tepui), Green hermit). Comparisons were generated by iterating over 5 individuals in each Imataca population, 10 individuals from Tepui and a single Green hermit. Lines represent three times the standard error.

But, is our interpretation well-grounded? A first step towards answering this question would be to check the validity of the null-hypothesis. In other words, are the ABBA and BABA counts expected to be equal without admixture? Let us first consider under which evolutionary scenarios *without gene flow* we expect to find an ABBA or BABA pattern. Assuming the tree under test captures the underlying phylogeny well, an ABBA/BABA can be found when there is (I) a *back mutation* or (II) *incomplete lineage sorting (ILS)*. Back mutations are mutations occurring independently in multiple lineages. In section 1.4 we saw that mutation rates in the genome are generally low, thus the probability of back mutations in short time scales is also low. Moreover, assuming that mutation rates are very similar in all the populations in the tree, recurrent mutations should happen at the same rate in all pairs of lineages. Thus ABBA and BABA should appear at the same rate and no asymmetries in their counts are expected. This brings us to the other possible explanation, incomplete lineage sorting (ILS). ILS is the process through which single gene trees differ to whole-genome trees.

Notice that if we were to build a phylogenetic tree out of a BABA pattern we would get the false impression that the North Imataca population is more closely related to Green hermits than to other violetears. This pattern can result when A and B have existed in the population since the MRCA of all violetears. Because the polymorphism was present in the MRCA, either allele can be retained by the daughter populations when a branching process occurs (figure 13A). Most importantly, the rate of ILS between lineages should be equal in all the daughter populations irrespectively of their population history. That is, North Imataca and Tepui should share the same number of alleles than South Imataca and Tepui (ABBA = BABA). These features of back mutations and ILS are great for our current purpose. They suggest that the null hypothesis is valid; models without gene flow should yield the same number of ABBA and BABA counts.

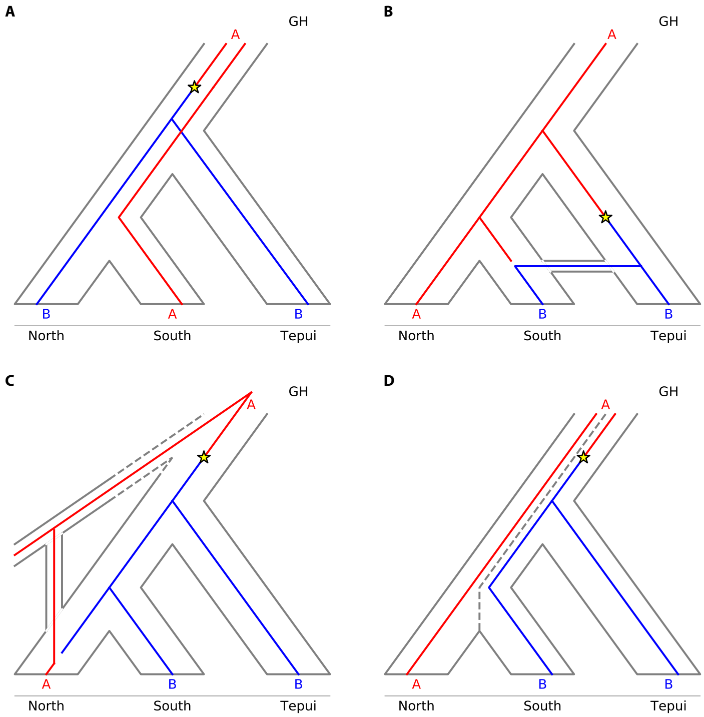

Figure 13. Alleles *A* and *B* are colored in red and blue respectively throughout the figure. Yellow stars represent a mutation event from *A* to *B*. GH stands for

Green hermit. (**A**) Model of incomplete lineage sorting. (**B**) Model of gene flow between South Imataca and Tepui. (**C**) Model of gene flow between North Imataca and an unsampled outgroup. (**D**) Model of ancient substructure in the ancestor of violetears.

Does this mean that values different from zero undoubtedly point to gene flow? Not necessarily. Patterson's D statistic is not an explicit test for admixture, but rather a test of *treeness*. The statistic is equal to zero when the configuration of populations we provide fits well the underlying phylogeny. By placing the populations in configurations that violate the *true* phylogeny, we obtain large deviations from zero. If we test the tree (((North Imataca, Tepui), South Imataca), Green hermit) by means of the *D*-statistic, we will get a vastly larger number of BABA than ABBA counts. This simply comes from the fact that the tree under consideration is *wrong*. However, let us now consider we have a good knowledge of the underlying phylogeny. Similarly to the case above, if we feed the statistic with a seemingly *correct* tree ((North Imataca,South Imataca), Tepui), Green hermit), then values different from zero will also arise when the data does not fit well with the provided tree. However, this time around we are aiming to detect more subtle effects, slight modifications in the ABBA/BABA pattern such as those expected under discrete genetic introgression events. In figure 12, we have seen that the South Imataca population shares more alleles with Tepui than the North Imataca population. As discussed earlier, gene flow between South Imataca and Tepui would disproportionately increase ABBA counts by providing alternative paths through which alleles can move across both populations (figure 13B). However, we must be cautious in invoking gene flow between South Imataca and Tepui, as it is not the only evolutionary scenario by which treeness can be distorted in a similar way. It could also be explained by an influx of genes from an outgroup lineage into the North Imataca population (figure 13C). Such event of admixture would convert fixed mutations in all violetears from BBBA to ABBA, thus generating an unbalance in the ABBA/BABA counts that mimics gene flow between South Imataca and Tepui. The input of alleles could come from any outgroup, for instance Green hermits or other unsampled populations. Additionally, an increased sharing of alleles between South Imataca and Tepui could also be the result of ancient substructure in the ancestor of violetears. This process is best understood when looking at figure 13D. Two diverging lineages could have existed in the ancestor of violetears, one of which evolved into North Imataca and the other into South Imataca and Tepui. Such scenario would generate an excess of ABBA counts with respect to BABA. Notice that gene flow within Imataca should be strong to explain the outgroup position of Tepui in the tree.

So far we have learned that the *D*-statistic is useful to detect asymmetries in allele sharing between populations, albeit these can be generated by multiple scenarios. If we want to narrow down the cause of our observations, we must dig deeper into the data. Some interesting follow-up questions given the pattern in allele sharing in violetears could be; did South Imataca and Tepui populations truly hybridize after their split? If so, was the gene flow directional? Can we find chunks of DNA in South Imataca coming from Tepui, or the other way around? In order to explore these questions, let us further discuss what are the genetic consequences of gene flow between populations.

If there ever was admixture between South Imataca and Tepui, we should be able to detect segments of their genome with an exceptionally short tMRCA. Here, the use of *exceptionally* is justified. Under a model without gene flow and a clean split between the two populations, DNA sequences in South Imataca and Tepui should only start coalescing after their split (*after* backwards in time). The reason behind this is that alleles segregating in separated lineages cannot coalesce or *meet* until the two lineages fuse. When the two populations merge into one, coalescent events will start to occur at a rate determined by the Ne of the ancestral population (section 1.5). From this it follows that the coalescence time between two populations must be older than their split time. However, there are scenarios in which this may not be true. Discrete events of gene flow before the split of populations provide extra routes for alleles to move across populations and meet. Precisely, the subset of genes taking the paths generated by gene flow should show an *exceptionally* short tMRCA compared to the rest of the genome. One could say that most alleles are forced to travel through the species *highway* and must wait until the two lineages fuse, while some *lucky* ones take the shortcut generated by gene flow. In order to test these ideas with genomic data, we must transition from *single-site* methods (ie. the *D*-statistic) to *haplotype* based methods. Such methods are typically applied by screening the genome with sliding windows and computing sequence divergence between populations in each window. In section 1.6, we saw that D_{XY} is an absolute measure of sequence divergence that should reflect well the tMRCA between a set of sequences. Altogether, we can conclude that in a model with gene flow we should expect to find a fraction of windows in the genome with unusually low divergence between the admixing populations.

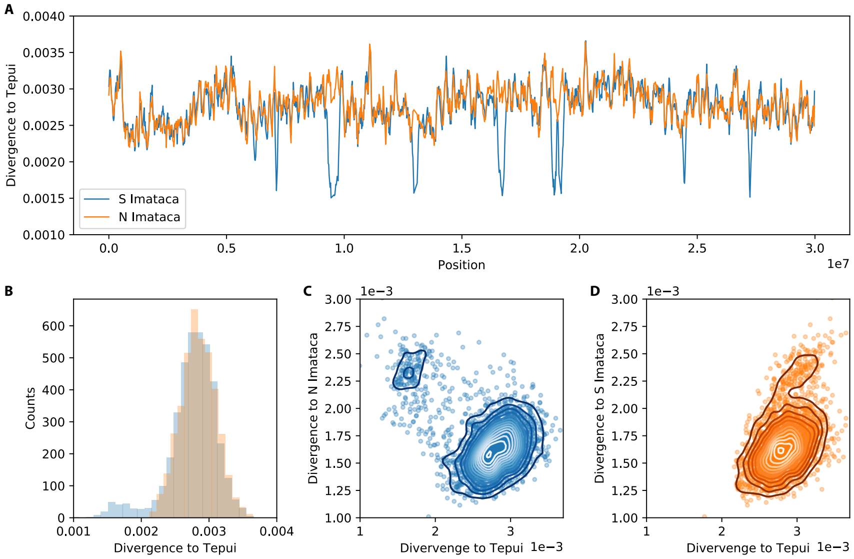

Figure 14. Divergence patterns in simulated sequences of violetears. (**A**) D_{XY} between South Imataca and Tepui (blue) and North Imataca and Tepui (orange) in windows of 100 kilobase-pairs with a slide of 20 kilobase-pairs. The color scheme remains constant throughout the other panels in the figure. **(B)** D_{XY} genome-wide distribution between South Imataca and Tepui, and North Imataca and Tepui. (**C**) D_{XY} between South Imataca and Tepui (x axis) against D_{XY} between South Imataca and North Imataca (y axis). **(D)** D_{XY} between North Imataca and Tepui (x axis) against D_{XY} between North Imataca and South Imataca (y axis).

In figure 14A, I plot D_{XY} between Tepui and a South and North Imataca individuals in windows of 100 kilobase-pairs with a 20 kilobase-pairs slide. The first things that might catch your eye are the sporadic *valleys* of low divergence between Tepui and South Imataca, which grouped together generate the little bump in the low end of the divergence distribution in figure 14B. These observations fit very well with our expectations under gene flow between South Imataca and Tepui. The results also agree well with the *D*-statistics shown in figure 12. Actually, most of the alleles generating the ABBA/BABA asymmetry are probably found within these low divergent segments of the genome. However, the haplotype based approach provides stronger support for gene flow than the single-site test because now we can start to discard the alternative models previously discussed and summarized in figure 13. It is also worth noting that the pattern in figure 14A seems to gather all the properties of genetic introgression; we can easily spot which regions in the genome are Tepui- or South Imataca-like. Nevertheless, we are one step away from inferring the directionality of the genetic introgression. Who are the donor and recipient populations? We might be tempted to say it was DNA from Tepui entering into South Imataca, but at the moment we do not have arguments to refute the opposite. Since D_{XY} is a reciprocal measure (ie. the divergence in the 27th 100 kilobase-pairs genomic window between South Imataca and Tepui has only one value, it does not depend on the order of the populations), the valleys of low divergence could be the reflection of DNA from Tepui into South Imataca, or the other way around.

In order to infer the directionality of the genetic introgression we have to use an extra variable. Let us first consider what are our expectations under a model of genetic introgression from Tepui into South Imataca. When screening the genome of a South Imataca violetear, any introgressed segments from Tepui should show (I) an unusually low divergence to Tepui (we already know this, figure 14B) and (II) an unusually high divergence to North Imataca. The second point comes from the fact that the tMRCA between the Tepui and North Imataca is far larger than that between South and North Imataca. Imagine we have 3 bowls filled with pieces shaped as cubes, spheres and ovoids. The shape of each bowl represents the genetic profiles in Tepui, South Imataca and North Imataca respectively. At some point we decide to move some pieces from the Tepui bowl into the South Imataca bowl. If we now compare the content of each bowl, we will see that a subset of South Imataca pieces are both (I) very similar to the cubes in the Tepui bowl and (II) very different to the rounded shapes peculiar from the North and South Imataca bowls. Of course, this subset of pieces will be the ones we transferred earlier from Tepui into South Imataca. In figure 14C-D, I plot the relationship between the divergence to Tepui and to either Imataca populations. As expected, Figure 14C shows that some sequence windows in South Imataca have really low divergence to the Tepui. However, this time around we can get a hint at the directionality of the introgression by looking at the divergence to North Imataca. The cloud of sequence windows with low divergence to Tepui also displays an unusually high divergence to North Imataca, which strongly supports genetic introgression from Tepui into South Imataca. If we check the divergence patterns in North Imataca in figure 14D, we can appreciate that there is not a subset of windows bearing Tepui-like DNA, or at least not to the extend found in South Imataca. Because D_{XY} is a reciprocal measure, we are also able to appreciate the South Imataca introgressed segments in this panel (look at the subset of data points with unusually large divergence to South Imataca which mirror the introgressed DNA in figure 14C). Altogether, we now have multiple lines of evidence suggesting that South Imataca received genes from Tepui at some point in the past.

Any readers that know about biology may have felt an ever-increasing discomfort reading this thesis. How is it possible that *recombination* has yet to be mentioned?! The truth is that ignoring recombination have made section 1.5 much easier for me. Treating each nucleotide in the genome as an independent unit provides a solid basis for studying population genetics, but it does not take into account that loci are arranged on chromosomes and hence are not inherited independently. I will take this opportunity to digress from the current topic and talk about recombination, although as we shall see, recombination plays an important role in genetic introgression. In fact, it is the reason why introgressed alleles in South Imataca are found in contiguous segments rather than randomly spread in the genome.

In eukaryotes, genetic recombination is the process by which parental chromosomes are combined in gametes to generate unique sets of genetic material. Consider each of your parental genomes as a book with the same amount of pages, one coming from your mother and the other from your father. Haploid gametes are then generated through meiosis by cutting and pasting your parental books into a new book with exactly the same size. That is, assuming that books are 500 pages long, we might take from page 1 to 150 from your father's book and from 150 to 500 from your mother's book. This unique new book can then be combined with other books (your partner's) to generate diploid offspring. Following this chain of reasoning, the gametes of your offspring would undergo a similar process when they are formed. Each of your offspring's unique gamete book will be created by combining the information that your partner and you passed to them. For instance, one of their gametes can be formed by concatenating pages 1 to 214 from your partner's book, pages 215 to 470 from your book, and 471 to 500 from your partner's book again. In this particular example there have been three cut-and-paste or *crossovers*. If we add up the amount of information a gamete from your offspring is carrying from your partner and you, we would see that it approximates 50% from each (in the example above 245 pages are from your partner and 255 are from you). This value would get lower as generations go on. The probability that one of your pages has survived dozens of cut-and-paste events becomes increasingly small with every generation. If we were to trace the amount pages that lives in your grand-grand-grand-grand-grand daughter, we would see that the amount is possibly zero. This implies that you would be her genealogical ancestor, but not necessarily her genetic ancestor. This may sound a bit depressing at first, but one must take into account that is not unusual to have dozens of grand-grand-grand-grand-grand daughters and sons, so your genetic legacy would probably live in some of them.

The human genome is composed by 23 *books* or chromosomes, and recombination occurs independently in each of them (we will ignore for the moment parts in the genome which are inherited in a different manner, X and Y chromosomes, mitochondria). The probability at which chromosomal crossovers occur is called the *recombination rate.* Similarly to mutation rate (section 1.4), the recombination rate is not evenly distributed across the genome. It often shows peaks of activity which generate *hotspots* of recombination. Despite this uneven distribution, approximately two to four crossovers occur in every chromosome. There are important consequences of genetic recombination. In any recombining genome, there will be thousands or even millions of different coalescent trees, each one showing a particular genealogical history. Notice that in the sliding window approach shown in figure 14, we are aiming to identify segments in the genome with a specific genealogy. That is, chunks of DNA that have been unbroken by recombination and share an ancestor with the Tepui population earlier than with the North Imataca population.

In the last two paragraphs, we developed the idea that segments from a genome are destined to become increasingly shorter in its descendants. From this it is reasonable to conclude that time of introgression should be directly related to the length of introgressed tracts. Indeed, we should be able to infer the time of admixture by analyzing the length of Tepui-like haplotypes in the South Imataca genomes. Nonetheless, we are currently assuming to know which are the introgressed segments. By using a sliding window approach we detected regions in the South Imataca genomes that harbour low divergence to Tepui. That is not enough for our purpose; a given 100 kilobase-pairs window may encompass a 63 kilobase-pairs of introgressed haplotype, but our current method is not able to pinpoint the start and end of such stretch of DNA. What could we do to refine the method and enable a fine detection of introgressed haplotypes? Introgressed haplotypes from Tepui should carry a large amount of genetic drift particular to the Tepui population. This signal can be detected by screening for alleles segregating together, in *linkage disequilibrium,* that are in high frequency within Tepui. One possibility could be to compute the average frequency across groups of contiguous polymorphisms in the genomes of South Imataca. In other words, we can screen the genome with windows of 20 SNPs, assigning to each window a value equal to the average frequency of the derived alleles in Tepui. The results of such analysis can be found in figure 15A. Notice that most of the windows have a value of zero. This is expected because many derived mutations in South Imataca have appeared in its own lineage, and thus have a frequency of zero in the Tepui branch. That does not mean there are no shared polymorphisms between all violetears; mutations that occurred in the ancestor can be shared between South Imataca and Tepui (ILS). Nevertheless, such mutations are not expected to segregate in long contiguous tracts of DNA due to the effect of recombination over many generations. In fact, the length of shared haplotypes due to ILS is expected to be orders of magnitude shorter than haplotypes shared due to recent gene flow. A caveat of our method is that it loses power with older gene flow events. The loss of power comes from the fact that genes coming from Tepui into South Imataca will keep accumulating mutations after the introgression. The appearance of novel mutations in the introgressed haplotypes will reduce the average frequency of the derived alleles in Tepui, as those are no longer shared unless back mutations occur. Additionally, the *real* length of an introgressed haplotype is not necessarily delimited by a SNP, it could extend for many base-pairs up- and down-stream in regions where there is no variation. Altogether, these limitations will underestimate the length of introgressed tracts, a far more desirable feature in science than overestimation.

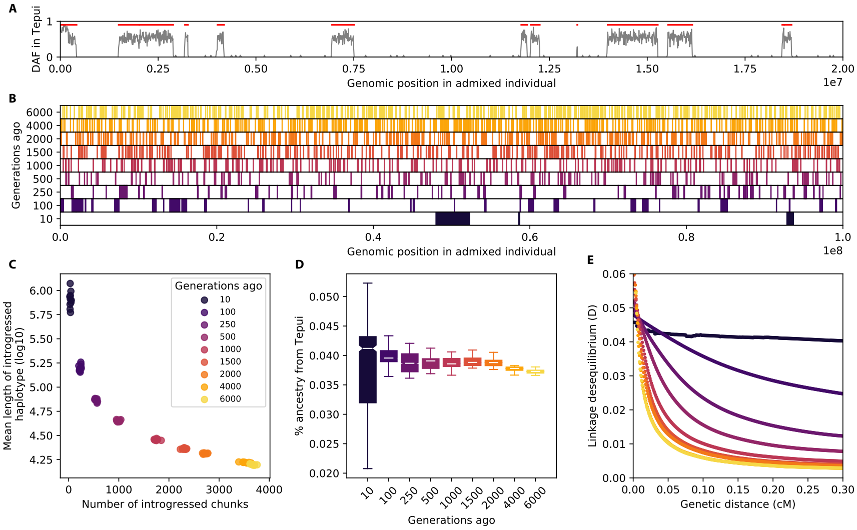

Figure 15. Recombination progressively breaks down introgressed tracts. (**A**) Randomly selected sequence of 2·10^7 base-pairs showing the method to detect introgressed segments from Tepui into a single South Imataca individual. The derived allele frequency in Tepui is plotted in the y axis. (**B**) Introgressed segments are detected using the method depicted in A. Sequences inferred to be introgressed are colored according to the time of gene flow (generations ago). (**C**) Mean length of introgressed segments (log10) against the number of introgressed segments colored by time of gene flow. Ten replicates under the same simulation conditions are shown. (**D**) Distribution of ancestry from Tepui in South Imataca individuals. (**E**) Extent of linkage disequilibrium between sites separated by bins of genetic distance (increment of 0.001 cM).

Now that we have a fairly good method to detect introgressed DNA, we can explore how tract length shrinks with time after a single pulse of gene flow. In figure 15B, I plot a simulated chromosome from an admixed individual in South Imataca after *n* generations since the gene flow episode. The work of recombination is easily appreciated as the Tepui-like haplotypes get shorter as generations pass by. Nevertheless, while the mean length of introgressed haplotypes decays with time, their numbers increase (figure 15C). One could say that introgressed segments 10 generations after the gene flow event look like large bands of DNA expanding through megabases of sequence. In contrast, introgressed segments 6.000 generations after gene flow look like confetti thrown on top of the genome. Do the length and number of introgressed tracts cancel each other out? Is the percentage of DNA coming from genetic introgression constant through time? In figure 15D we can see the proportion of the South Imataca genomes can be assigned to the Tepui population per generation. The percentage of introgressed DNA can be inferred through f4 ratios, a method based on Patterson's D statistic. The variance of Tepui-like DNA is larger in recent events of gene flow. This should make good intuitive sense. Consider a genetic introgression event that happened 2 generations back in time and replaced 4% of the population (that is, if South Imataca has 100 individuals, 4 of them were substituted by migrants from Tepui). Some individuals in the population will carry a large amount of Tepui DNA. In fact, some violetears may have a grandparent from Tepui, which means that roughly 25% of their genome is introgressed. Conversely, most of the individuals will have 0% of DNA from Tepui, as none of their ancestors was one of the migrants; there has not been enough time for the migrant DNA to spread in the population. If we wait for more generations, all the individuals in the population will ultimately have an ancestor with Tepui DNA, and the genetic contribution will converge to the initial percentage of migrants (4%).

However, if we look closely at figure 15D, we can appreciate that the percentage of Tepui ancestry seems to decay with time to values lower than 4%. This could be explained by natural selection acting against the introgressed segments. Earlier in this section we discussed how genetic incompatibilities can arise between diverging lineages. Following this chain of thought, introgressed tracts may be under negative selection due to harmful *epistatic* interactions between South Imataca and Tepui alleles. Another non-exclusive possibility is that the historical *Ne* in Tepui has been very low. Under this scenario, genetic drift in Tepui has been strong, therefore weakly deleterious alleles have had a tendency to persist in the population as if they were neutral (see section 1.4 for more details on the relationship between *Ne* and efficiency of selection). The amount of deleterious alleles in a genome is often referred to as the *mutational load.* If the mutational load in Tepui is higher than in South Imataca, introgressed tracts will be preferentially purged by natural selection in the South Imataca genomic background. Both of the aforementioned effects could be causing the progressive decay of introgressed ancestry. Nevertheless, that is not possible in our case. The genetic data has been simulated under a strictly neutral model, so natural selection is out of the picture. This brings us to the question; what neutral models can explain the data? Tepui-like ancestry may be slowly decaying due to gene flow from an unadmixed population. This *dilution* model is perfectly possible in the simulated data, since the North imataca is modelled and has low levels of constant gene flow with its counterpart in South Imataca.

Earlier in this section we posited a yet unanswered question: can we provide a specific time of gene flow from the length of introgressed tracts? We can, although methods not directly relying on sequence length can be more useful (see limitations of inferring start and end coordinates of introgressed segments above). To do this, one can exploit the rate of linkage disequilibrium between shared alleles between Tepui and South Imataca. As we have seen, ancient gene flow introduces blocks of Tepui ancestry into the South Imataca background that break down at an approximately constant rate per generation as crossovers occur. Therefore, recent events of gene flow should result in higher levels of linkage disequilibrium between introgressed alleles.

In order to quantify the extent of linkage disequilibrium consider two loci with alleles A/a and B/b respectively. The four alleles can be combined into four different haplotypes (AB, Ab, aB and ab). In any sample of chromosomes, we can count the number of each haplotype and compute its frequency in the population (f_{AB},f_{Ab},f_{aB} and f_{ab}). If alleles A and B are segregating in complete independence, then the haplotype frequency should be equal to the product of the allele frequencies ($f_{AB} = f_A f_B$). However, alleles can be segregating in a block (ie. A may be more often paired with B than with b). We can calculate the deviation from linkage equilibrium by subtracting the expected frequency to the observed frequency ($D_{AB} = f_{AB} - f_A f_B$). D, *the coefficient of linkage disequilibrium*, can then be computed across loci separated by n base-pairs. DNA physical distance can be converted to genetic distance (*centimorgans*) by calculating the expected number of crossovers between both positions. Notice that to perform this calculation we must have knowledge of the underlying recombination rate across the sequence (in real populations we need a *recombination map*, but in our toy example this is not a problem because the sequences are simulated under a constant recombination rate). In figure 15E we can see how D in shared polymorphisms between South Imataca and Tepui decays with genetic distance. The steepness of the decay is directly related to the time of admixture. How can we get the number of generations out of these trajectories? The time of admixture (λ) is typically inferred by means of a least squares fit to an exponential distribution ($D(d) = Ae^{-d\lambda}$ with ranging genetic distance values) (Sankararaman et al. 2012).

1.8. Chimpanzees and bonobos

Kuhlwilm M, de Manuel M, Nater A, Greminger MP, Krützen M, Marques-Bonet T. Evolution and demography of the great apes. Curr Opin Genet Dev. 2016 Dec;41:124–9. DOI: 10.1016/j.gde.2016.09.005

2. OBJECTIVES

- Characterise the spatial distribution of genetic diversity within chimpanzee subspecies.
- Explore the use of statistical methods to infer geographic coordinates from genetic data.
- Refine our understanding of chimpanzee population history; split times, historical effective population sizes and gene flow.
- Explore ancient genetic introgression between chimpanzees and bonobos.

3. RESULTS

de Manuel M, Kuhlwilm M, Frandsen P, Sousa VC, Desai T, Prado-Martinez J, et al. Chimpanzee genomic diversity reveals ancient admixture with bonobos. Science (80-). 2016 Oct 28;354(6311):477–81. DOI: 10.1126/science.aag2602

4. DISCUSSION

In the following section, I will discuss the relevance of the results presented herein for the understanding of the population history of chimpanzees and bonobos. As discussed in section 3, my research has focused on (I) unravelling fine-scale population structure within chimpanzee subspecies and (II) exploring ancient admixture between chimpanzees and bonobos.

We analyzed 75 complete genomes from the *Pan* genus, of which 40 were sequenced for this project to a mean sequence coverage of 25-fold. Our samples span 10 African countries, which constitutes the most exhaustive sampling scheme in non-human great ape genomics to date. This dataset brought up the possibility to interrogate the extent to which genetic information can predict geographic origin. Indeed, we found a tight correspondence between genetic diversity and geography in central and eastern chimpanzees. In order to further confirm our results, we analysed a set of georeferenced individuals sequenced from non-invasive samples. By analysing these GPS-labeled individuals, we found that genotyping a few thousand polymorphisms is enough to assign chimpanzees to their country of origin. Although we could not include enough geolocalized samples to assess fine-scale population structure in Nigeria-Cameroon and western chimpanzees, we expect that similar stratification would be found with broader sampling.

The data is largely consistent with patterns of isolation by distance within central and eastern chimpanzees. Nonetheless, some populations show hallmarks of long standing reproductive isolation (ie. eastern chimpanzees from Gombe, Tanzania). Patterns of isolation by distance would imply that inferring the origin of chimpanzees from yet unsampled locations should be possible through the use of allele frequency maps. However, cases like the eastern chimpanzees in Gombe highlight the importance of a denser sampling scheme. The ability to predict geographic origin from genetic data can have a valuable impact on the conservation of declining populations. It can help to localize hotspots of illegal trafficking, which in turn can help governments to increase law enforcement in these areas and empower the conservation of endangered species (S. K. Wasser et al. 2015). In the future, the origins of confiscated chimpanzees will probably be discernible with sufficient data from reference populations, with implications for the *in situ* and *ex situ* management of this species.

Genetic introgression is common in nature. It occurs frequently in plants, insects, fish, birds and mammals (Rieseberg 2009; Nadeau et al. 2012). Thus, perhaps not surprisingly, multiple gene flow events between archaic and modern humans have been suggested in the last decade. To illustrate how convoluted the tree of humans may be, I list all the putative gene flow episodes described as of today: (I) from Neanderthals into non-African modern humans (Green et al. 2010), (II) from Denisovans into Oceanian modern humans (Reich et al. 2010), (III) from early modern humans into Neanderthals (Kuhlwilm et al. 2016), (IV) from Eastern Neanderthals into Denisovans (Prüfer et al. 2014), (V) from an unknown hominin lineage into Southeast Asian modern humans (Mondal et al. 2016) and (VI) from an unknown hominin archaic lineage into Denisovans (Prüfer et al. 2014). Given this intricate graph of connections between modern and archaic humans, more of such events are likely to be detected as more ancient genomes become available. This scenario also puts forward the following question; is genetic introgression also common in chimpanzees and bonobos, our closest living relatives? In section 3 we suggest that the answer is yes, ancient admixture did happen in the *Pan* branch. This finding draws a parallelism with humans and highlights the pervasiveness of gene flow in nature.

However, it is worth noting that gene flow between chimpanzees and bonobos was explicitly explored in the past using whole-genome sequences, yet admixture was found to not be supported by the data (Prüfer et al. 2012). Conversely, several studies have reported allele frequencies in bonobos to be more similar to central chimpanzees than to western chimpanzees (Becquet et al. 2007; Won and -J. Won 2004; Becquet and Przeworski 2007), and gene flow between the chimpanzee ancestor and bonobos has also been suggested (Cahill et al. 2016; Hey et al. 2018). In order to elucidate the conflict between studies, we reanalysed the dataset in Prüfer et al. The main difference between sequence data in Prüfer et al. and our study is the depth of coverage. While our genomes average to 25-fold, genomes in Prüfer et al range between 1-2-fold. Given this difference, we first explored the influence of coverage in the power of *D*-statistics to detect gene flow. We found that low coverages in Prüfer et al. greatly undermines the power to detect reliable polymorphisms, which in turn has an effect in the resolution of *D*-statistics. By detecting polymorphisms with our high-coverage genomes, we found that genomes in Prüfer et al. also harbour the signal of admixture between bonobos and central chimpanzees.

Single-site statistics such as *D*-statistics are known to be especially susceptible to bias. Thus, even though the data in Prüfer et al. seems to agree well with our conclusions, the most suggestive observation in our research is the existence of introgressed haplotypes. Since haplotype-based analyses require high coverage and the calling of diploid genotypes, we could not include data from Prüfer et al. into such analyses. The identification of an excess of long stretches of DNA with unusually low divergence to Neanderthal in non-African modern humans has been interpreted as one of the strongest evidence for genetic introgression between archaic and modern humans (Green et al. 2010). Similarly, finding clusters of bonobo alleles in strong linkage disequilibrium in the genomes of central and eastern chimpanzees is the most convincing line of evidence in favor of gene flow. Such observation is only expected under a model with gene flow from bonobo into chimpanzees, and should be robust to confounding factors known to affect single-site *D*-statistics, such as contamination or technical difficulties in the calling of polymorphisms.

Another evidence for gene flow is the presence of introgression deserts, regions in the chimpanzee genome devoid of introgressed DNA from bonobo. As discussed in section 1.7, this is expected because introgressed bonobo alleles might be disadvantageous in the chimpanzee genetic background. This effect is especially accentuated in the X chromosome, where introgression segments are almost non-existent. Similar observations have been done in humans with respect to Neanderthal introgression (Sankararaman et al. 2014). It has also been found that levels of recombination in the genome can determine the amount of genetic introgression (Schumer et al. 2018). More particularly introgression seems to be more common in regions of high recombination, where there is linkage to fewer putative targets of selection. These results support models in which ancestry from the donor species is more likely to persist when it is rapidly uncoupled from alleles that are deleterious in hybrids. Given that the recombination landscape has a rapid turnover among chimpanzee subspecies (Stevison et al. 2016), it would be interesting to see if the introgression landscape has been shaped differently in central and eastern chimpanzees. Quite interestingly, the opposite situation is also found in nature; evidence is increasing in support of the existence of introgressed variants that may have been selected by natural selection. Recently, it has been suggested that introgressed tracts in chimpanzees have been targeted by selection (Nye et al. 2018).

Nonetheless, genetic introgression is best confirmed when finding admixed individuals few generations after the gene flow event. As discussed in section 1.7, the length of introgressed tracts progressively decays with time due to recombination. Indeed, the sequencing of an ancient individual 4-6 generations after gene flow from Neanderthals into modern humans (Fu et al. 2015), and recently a first generation hybrid of a Neanderthal and Denisovan (Slon et al. 2018), has provided irrefutable evidence of the encounter between human lineages. However, such result would be hard to attain in chimpanzees and bonobos given the age of the putative gene flow event (200-550 thousand years ago), as well as due to a paucity of fossil records (McBrearty and Jablonski 2005). The lack of fossil remains is not only particular to chimpanzees and bonobos, but also found in gorillas and orang-utans, although perhaps less so in the later (Chaimanee et al. 2003). The sequencing of ancient individuals from the *Pan* lineage would enable the exploration of many interesting questions. For instance, it is known that human population history has been characterised by the movement of cultures and common population replacement, at least in Europe (Haak et al. 2015). It would be enlightening to check if the population history of our closest living relatives have experienced a similar amount of migration waves and population replacement across time. The sequencing of ancient remains also opens the possibility to study the genomes of now extinct lineages. Unfortunately, this is a line of research hard to pursue given the almost non-existent fossil record. However, there is room for optimism, as the field of ancient DNA is under constant rapid development. Last year, DNA from multiple ancient organisms was sequenced from Pleistocene sediments (Slon et al. 2017), so it is difficult to imagine what technological advancements the future might bring.

There are other possibilities to explore the past that do not rely on bones or sediments; the excavation of DNA segments from extinct lineages in the genomes of contemporary populations. In our study, we did not explicitly test for gene flow events from unsampled unknown *Pan* lineages into modern chimpanzees and bonobos. However, finding traces of such events would open a window to the past, as it would enable the study of stretches of DNA that directly descend from populations from which there might be little to no fossil remains. In order to detect introgression from unsampled populations, one must use measures that do not rely on the sequence of the donor population (Plagnol and Wall 2006). This strategy has been applied with success in modern humans to recover fragments of the Neanderthal and Denisovan genomes (Vernot et al. 2016). There is tentative evidence that bonobos carry chunks of DNA from an extinct *Pan* population (Brahic 2018), so inferring extinct *Pan* genomic sequences from contemporanean populations may be feasible in the near future.

Finally, I would like to advocate for the study of bonobos. Chimpanzees have historically had a privileged position in comparison to their sister species. This may be due to their larger population size, or quite simply because they live in regions more accessible to researchers. In any case, bonobos are the most understudied great ape population by means of genetic data (Prado-Martinez et al. 2013). Population structure within bonobos has been assessed by fragments of mitochondrial DNA, and found to be determined by the numerous tributaries of the Congo River (Eriksson et al. 2004; Kawamoto et al. 2013). It would be interesting to expand these studies to whole-genome data, with denser sampling schemes that would allow a fine-scale assessment of the geographic distribution of genetic diversity. Increasing the number of bonobo genomic data could also enable an alternative way to contrast our findings of genetic introgression from bonobos into chimpanzees. Maybe there are highly differentiated lineages within bonobos, some of them more closely related to the original source of the genetic introgression. However, for this scenario to be plausible, there should be populations in bonobos sharing an ancestor earlier than the gene flow event (200-550 thousand years ago). That seems unlikely given our current knowledge of bonobos. Nevertheless, the presence of such lineages would suggest that there are multiple bonobo subspecies. In any case, such hypothesis will remain unknown until more sequences of bonobo become available.

BIBLIOGRAPHY

Al-Asadi, Hussein, Desislava Petkova, Matthew Stephens, and John Novembre. 2018. "Estimating Recent Migration and Population Size Surfaces." https://doi.org/10.1101/365536.

Alexander, David H., John Novembre, and Kenneth Lange. 2009. "Fast Model-Based Estimation of Ancestry in Unrelated Individuals." *Genome Research* 19 (9): 1655–64.

Alexandrov, Ludmil B., Serena Nik-Zainal, David C. Wedge, Samuel A J, Sam Behjati, Andrew V. Biankin, Graham R. Bignell, et al. 2013. "Signatures of Mutational Processes in Human Cancer." *Nature* 500 (7463): 415.

Avery, O. T., C. M. MacLeod, and M. McCarty. 1944. "Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types. Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type III. 1944." *J. Exp. Med. 79:137–158.*

Becquet, Celine, Nick Patterson, Anne C. Stone, Molly Przeworski, and David Reich. 2007. "Genetic Structure of Chimpanzee Populations." *PLoS Genetics* 3 (4): e66.

Becquet, Celine, and Molly Przeworski. 2007. "A New Approach to Estimate Parameters of Speciation Models with Application to Apes." *Genome Research* 17 (10): 1505–19.

Benton, Michael J., and Philip C. J. Donoghue. 2007. "Paleontological Evidence to Date the Tree of Life." *Molecular Biology and Evolution* 24 (1): 26–53.

Bhatia, Gaurav, Nick Patterson, Sriram Sankararaman, and Alkes L. Price. 2013. "Estimating and Interpreting FST: The Impact of Rare Variants." *Genome Research* 23 (9): 1514–21.

Bradburd, Gideon S., Peter L. Ralph, and Graham M. Coop. 2016. "A Spatial Framework for Understanding Population Structure and Admixture." *PLoS Genetics* 12 (1): e1005703.

Brahic, Catherine. 2018. "Mystery Ghost Ape Discovered." *New Scientist* 238 (3180): 4.

Cahill, James A., André E. R. Soares, Richard E. Green, and Beth Shapiro. 2016. "Inferring Species Divergence Times Using Pairwise Sequential Markovian Coalescent Modelling and Low-Coverage Genomic Data." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 371 (1699). https://doi.org/10.1098/rstb.2015.0138.

Chaimanee, Yaowalak, Dominique Jolly, Mouloud Benammi, Paul Tafforeau, Danielle Duzer, Issam Moussa, and Jean-Jacques Jaeger.

2003. "A Middle Miocene Hominoid from Thailand and Orangutan Origins." *Nature* 422 (6927): 61–65.

Cruickshank, Tami E., and Matthew W. Hahn. 2014. "Reanalysis Suggests That Genomic Islands of Speciation Are due to Reduced Diversity, Not Reduced Gene Flow." *Molecular Ecology* 23 (13): 3133–57.

Darwin, Charles. 1869. *On the Origin of Species by Means of Natural Selection: Or the Preservation of Favoured Races in the Struggle for Life.*

Davies, Kevin. 2010. *The $1,000 Genome: The Revolution in DNA Sequencing and the New Era of Personalized Medicine.* Simon and Schuster.

Eriksson, J., G. Hohmann, C. Boesch, and L. Vigilant. 2004. "Rivers Influence the Population Genetic Structure of Bonobos (Pan Paniscus)." *Molecular Ecology* 13 (11): 3425–35.

Felsenstein, Joseph. 2004. *Inferring Phylogenies.* Sinauer Associates Incorporated.

Foote, Andrew D., Nagarjun Vijay, María C. Ávila-Arcos, Robin W. Baird, John W. Durban, Matteo Fumagalli, Richard A. Gibbs, et al. 2016. "Genome-Culture Coevolution Promotes Rapid Divergence of Killer Whale Ecotypes." *Nature Communications* 7 (May): 11693.

Fu, Qiaomei, Mateja Hajdinjak, Oana Teodora Moldovan, Silviu Constantin, Swapan Mallick, Pontus Skoglund, Nick Patterson, et al. 2015. "An Early Modern Human from Romania with a Recent Neanderthal Ancestor." *Nature* 524 (7564): 216–19.

Fu, Qiaomei, Heng Li, Priya Moorjani, Flora Jay, Sergey M. Slepchenko, Aleksei A. Bondarev, Philip L. F. Johnson, et al. 2014. "Genome Sequence of a 45,000-Year-Old Modern Human from Western Siberia." *Nature* 514 (7523): 445–49.

Gao, Ziyue, Priya Moorjani, Guy Amster, and Molly Przeworski. 2018. "Overlooked Roles of DNA Damage and Maternal Age in Generating Human Germline Mutations." https://doi.org/10.1101/327098.

Gao, Ziyue, Minyoung J. Wyman, Guy Sella, and Molly Przeworski. 2016. "Interpreting the Dependence of Mutation Rates on Age and Time." *PLoS Biology* 14 (1): e1002355.

Green, Richard E., Johannes Krause, Adrian W. Briggs, Tomislav Maricic, Udo Stenzel, Martin Kircher, Nick Patterson, et al. 2010. "A Draft Sequence of the Neandertal Genome." *Science* 328 (5979): 710–22.

Gymrek, Melissa, Amy L. McGuire, David Golan, Eran Halperin, and Yaniv Erlich. 2013. "Identifying Personal Genomes by Surname Inference." *Science* 339 (6117): 321–24.

Haak, Wolfgang, Iosif Lazaridis, Nick Patterson, Nadin Rohland, Swapan Mallick, Bastien Llamas, Guido Brandt, et al. 2015. "Massive Migration from the Steppe Was a Source for Indo-European Languages in Europe." *Nature* 522 (7555): 207–11.

Haldane, J. B. S. 1941. "Can Science Be Independent?" *Nature* 147 (3727): 416–416.

Handley, Lori J. Lawson, Lori J. Lawson Handley, Andrea Manica, Jérôme Goudet, and François Balloux. 2007. "Going the Distance: Human Population Genetics in a Clinal World." *Trends in Genetics: TIG* 23 (9): 432–39.

Hey, Jody, Yujin Chung, Arun Sethuraman, Joseph Lachance, Sarah Tishkoff, Vitor C. Sousa, and Yong Wang. 2018. "Phylogeny Estimation by Integration over Isolation with Migration Models." *Molecular Biology and Evolution*, August. https://doi.org/10.1093/molbev/msy162.

Hudson, Richard R. 1983. "TESTING THE CONSTANT-RATE NEUTRAL ALLELE MODEL WITH PROTEIN SEQUENCE DATA." *Evolution; International Journal of Organic Evolution* 37 (1): 203–17.

Hudson, R. R., M. Slatkin, and W. P. Maddison. 1992. "Estimation of Levels of Gene Flow from DNA Sequence Data." *Genetics* 132 (2): 583–89.

Jónsson, Hákon, Patrick Sulem, Birte Kehr, Snaedis Kristmundsdottir, Florian Zink, Eirikur Hjartarson, Marteinn T. Hardarson, et al. 2017. "Parental Influence on Human Germline de Novo Mutations in 1,548 Trios from Iceland." *Nature* 549 (7673): 519–22.

Ju, Young Seok, Inigo Martincorena, Moritz Gerstung, Mia Petljak, Ludmil B. Alexandrov, Raheleh Rahbari, David C. Wedge, et al. 2017. "Somatic Mutations Reveal Asymmetric Cellular Dynamics in the Early Human Embryo." *Nature* 543 (7647): 714–18.

Kawamoto, Yoshi, Hiroyuki Takemoto, Shoko Higuchi, Tetsuya Sakamaki, John A. Hart, Terese B. Hart, Nahoko Tokuyama, et al. 2013. "Genetic Structure of Wild Bonobo Populations: Diversity of Mitochondrial DNA and Geographical Distribution." *PloS One* 8 (3): e59660.

Kern, Andrew D., and Matthew W. Hahn. 2018. "The Neutral Theory in Light of Natural Selection." *Molecular Biology and Evolution* 35 (6): 1366–71.

Kimura, Motoo. 1957. "Some Problems of Stochastic Processes in Genetics." *Annals of Mathematical Statistics* 28 (4): 882–901.

Kimura, Motoo, and Tomoko Ohta. 1971. "Protein Polymorphism as a Phase of Molecular Evolution." *Nature* 229 (5285): 467–69.

Kimura, M., and G. H. Weiss. 1964. "The Stepping Stone Model of Population Structure and the Decrease of Genetic Correlation with Distance." *Genetics* 49 (4): 561–76.

Kingman, J. F. C. 1982. "On the Genealogy of Large Populations." *Journal of Applied Probability* 19 (A): 27–43.

Kolmogoroff, A. 1931. "Über Die Analytischen Methoden in Der Wahrscheinlichkeitsrechnung." *Mathematische Annalen* 104 (1): 415–58.

Kong, Augustine, Michael L. Frigge, Gisli Masson, Soren Besenbacher, Patrick Sulem, Gisli Magnusson, Sigurjon A. Gudjonsson, et al. 2012. "Rate of de Novo Mutations and the Importance of Father's Age to Disease Risk." *Nature* 488 (7412): 471–75.

Koning, A. P. Jason de, Wanjun Gu, Todd A. Castoe, Mark A. Batzer, and David D. Pollock. 2011. "Repetitive Elements May Comprise over Two-Thirds of the Human Genome." *PLoS Genetics* 7 (12): e1002384.

Kopps, A. M., C. Y. Ackermann, W. B. Sherwin, S. J. Allen, L. Bejder, and M. Krutzen. 2014. "Cultural Transmission of Tool Use Combined with Habitat Specializations Leads to Fine-Scale Genetic Structure in Bottlenose Dolphins." *Proceedings of the Royal Society B: Biological Sciences* 281 (1782): 20133245–20133245.

Kronenberg, Zev N., Ian T. Fiddes, David Gordon, Shwetha Murali, Stuart Cantsilieris, Olivia S. Meyerson, Jason G. Underwood, et al. 2018. "High-Resolution Comparative Analysis of Great Ape Genomes." *Science* 360 (6393). https://doi.org/10.1126/science.aar6343.

Kuhlwilm, Martin, Ilan Gronau, Melissa J. Hubisz, Cesare de Filippo, Javier Prado-Martinez, Martin Kircher, Qiaomei Fu, et al. 2016. "Ancient Gene Flow from Early Modern Humans into Eastern Neanderthals." *Nature* 530 (7591): 429–33.

Lao, Oscar, Timothy T. Lu, Michael Nothnagel, Olaf Junge, Sandra Freitag-Wolf, Amke Caliebe, Miroslava Balascakova, et al. 2008. "Correlation between Genetic and Geographic Structure in Europe." *Current Biology: CB* 18 (16): 1241–48.

Lawson, Daniel J., Lucy van Dorp, and Daniel Falush. 2018. "A Tutorial on How Not to over-Interpret STRUCTURE and ADMIXTURE Bar Plots." *Nature Communications* 9 (1). https://doi.org/10.1038/s41467-018-05257-7.

Ledford, Heidi. 2016. "AstraZeneca Launches Project to Sequence 2 Million Genomes." *Nature* 532 (7600): 427.

Lynch, Michael, Matthew S. Ackerman, Jean-Francois Gout, Hongan Long, Way Sung, W. Kelley Thomas, and Patricia L. Foster. 2016. "Genetic Drift, Selection and the Evolution of the Mutation Rate." *Nature Reviews. Genetics* 17 (11): 704.

Malécot, G. 1948. *Les Mathématiques de l'hérédité*. Masson.

Mayr, Ernst. 1963. *Animal Species and Evolution*.

McBrearty, Sally, and Nina G. Jablonski. 2005. "First Fossil Chimpanzee." *Nature* 437 (7055): 105–8.

Mondal, Mayukh, Ferran Casals, Tina Xu, Giovanni M. Dall'Olio, Marc Pybus, Mihai G. Netea, David Comas, et al. 2016. "Genomic Analysis of Andamanese Provides Insights into Ancient Human Migration into Asia and Adaptation." *Nature Genetics* 48 (9): 1066–70.

Moorjani, Priya, Carlos Eduardo G. Amorim, Peter F. Arndt, and Molly Przeworski. 2016. "Variation in the Molecular Clock of Primates." https://doi.org/10.1101/036434.

Nadeau, Nicola J., Annabel Whibley, Robert T. Jones, John W. Davey, Kanchon K. Dasmahapatra, Simon W. Baxter, Michael A. Quail, et al. 2012. "Genomic Islands of Divergence in Hybridizing Heliconius Butterflies Identified by Large-Scale Targeted Sequencing." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 367 (1587): 343–53.

Narasimhan, Vagheesh M., Raheleh Rahbari, Aylwyn Scally, Arthur Wuster, Dan Mason, Yali Xue, John Wright, et al. 2017. "Estimating the Human Mutation Rate from Autozygous Segments Reveals Population Differences in Human Mutational Processes." *Nature Communications* 8 (1): 303.

Nater, Alexander, Maja P. Mattle-Greminger, Anton Nurcahyo, Matthew G. Nowak, Marc de Manuel, Tariq Desai, Colin Groves, et al. 2017. "Morphometric, Behavioral, and Genomic Evidence for a New Orangutan Species." *Current Biology: CB* 27 (22): 3487–98.e10.

Nei, M. 1973. "Analysis of Gene Diversity in Subdivided Populations." *Proceedings of the National Academy of Sciences* 70 (12): 3321–23.

Nei, M., and W. H. Li. 1979. "Mathematical Model for Studying Genetic Variation in Terms of Restriction Endonucleases." *Proceedings of the National Academy of Sciences of the United States of America* 76 (10): 5269–73.

Novembre, John, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R. Boyko, Adam Auton, Amit Indap, et al. 2008. "Genes Mirror Geography within Europe." *Nature* 456 (7218): 98–101.

Nye, Jessica, Hafid Laayouni, Martin Kuhlwilm, Mayukh Mondal, Tomas Marques-Bonet, and Jaume Bertranpetit. 2018. "Selection in the Introgressed Regions of the Chimpanzee Genome." *Genome Biology and Evolution* 10 (4): 1132–38.

Patterson, Nick, Priya Moorjani, Yontao Luo, Swapan Mallick, Nadin Rohland, Yiping Zhan, Teri Genschoreck, Teresa Webster, and David Reich. 2012. "Ancient Admixture in Human History." *Genetics* 192 (3): 1065–93.

Petkova, Desislava, John Novembre, and Matthew Stephens. 2014. "Visualizing Spatial Population Structure with Estimated Effective Migration Surfaces." https://doi.org/10.1101/011809.

Plagnol, Vincent, and Jeffrey D. Wall. 2006. "Possible Ancestral Structure in Human Populations." *PLoS Genetics* 2 (7): e105.

Prado-Martinez, Javier, Peter H. Sudmant, Jeffrey M. Kidd, Heng Li, Joanna L. Kelley, Belen Lorente-Galdos, Krishna R. Veeramah, et al. 2013. "Great Ape Genetic Diversity and Population History." *Nature* 499 (7459): 471.

Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. "Inference of Population Structure Using Multilocus Genotype Data." *Genetics* 155 (2): 945–59.

Prüfer, Kay, Kasper Munch, Ines Hellmann, Keiko Akagi, Jason R. Miller, Brian Walenz, Sergey Koren, et al. 2012. "The Bonobo Genome Compared with the Chimpanzee and Human Genomes." *Nature* 486 (7404): 527–31.

Prüfer, Kay, Fernando Racimo, Nick Patterson, Flora Jay, Sriram Sankararaman, Susanna Sawyer, Anja Heinze, et al. 2014. "The Complete Genome Sequence of a Neanderthal from the Altai Mountains." *Nature* 505 (7481): 43–49.

Rañola, John Michael, John Novembre, and Kenneth Lange. 2014. "Fast Spatial Ancestry via Flexible Allele Frequency Surfaces." *Bioinformatics* 30 (20): 2915–22.

Reich, David, Richard E. Green, Martin Kircher, Johannes Krause, Nick Patterson, Eric Y. Durand, Bence Viola, et al. 2010. "Genetic History of an Archaic Hominin Group from Denisova Cave in Siberia." *Nature* 468 (7327): 1053–60.

Rieseberg, Loren H. 2009. "Evolution: Replacing Genes and Traits through Hybridization." *Current Biology: CB* 19 (3): R119–22.

Sankararaman, Sriram, Swapan Mallick, Michael Dannemann, Kay Prüfer, Janet Kelso, Svante Pääbo, Nick Patterson, and David Reich. 2014. "The Genomic Landscape of Neanderthal Ancestry in Present-Day Humans." *Nature* 507 (7492): 354–57.

Sankararaman, Sriram, Nick Patterson, Heng Li, Svante Pääbo, and David Reich. 2012. "The Date of Interbreeding between Neandertals and Modern Humans." *PLoS Genetics* 8 (10): e1002947.

Scally, Aylwyn, and Richard Durbin. 2012. "Revising the Human Mutation Rate: Implications for Understanding Human Evolution." *Nature Reviews. Genetics* 13 (10): 745–53.

Schumer, Molly, Chenling Xu, Daniel L. Powell, Arun Durvasula, Laurits Skov, Chris Holland, John C. Blazier, et al. 2018. "Natural Selection

Interacts with Recombination to Shape the Evolution of Hybrid Genomes." *Science* 360 (6389): 656–60.

Slon, Viviane, Charlotte Hopfe, Clemens L. Weiß, Fabrizio Mafessoni, Marco de la Rasilla, Carles Lalueza-Fox, Antonio Rosas, et al. 2017. "Neandertal and Denisovan DNA from Pleistocene Sediments." *Science* 356 (6338): 605–8.

Slon, Viviane, Fabrizio Mafessoni, Benjamin Vernot, Cesare de Filippo, Steffi Grote, Bence Viola, Mateja Hajdinjak, et al. 2018. "The Genome of the Offspring of a Neanderthal Mother and a Denisovan Father." *Nature* 561 (7721): 113–16.

Staab, Paul R., Sha Zhu, Dirk Metzler, and Gerton Lunter. 2015. "Scrm: Efficiently Simulating Long Sequences Using the Approximated Coalescent with Recombination." *Bioinformatics* 31 (10): 1680–82.

Stevison, Laurie S., August E. Woerner, Jeffrey M. Kidd, Joanna L. Kelley, Krishna R. Veeramah, Kimberly F. McManus, Great Ape Genome Project, Carlos D. Bustamante, Michael F. Hammer, and Jeffrey D. Wall. 2016. "The Time Scale of Recombination Rate Evolution in Great Apes." *Molecular Biology and Evolution* 33 (4): 928–45.

Stulp, Gert, Mirre J. P. Simons, Sara Grasman, and Thomas V. Pollet. 2017. "Assortative Mating for Human Height: A Meta-Analysis." *American Journal of Human Biology: The Official Journal of the Human Biology Council* 29 (1). https://doi.org/10.1002/ajhb.22917.

Tang, Hua, Marc Coram, Pei Wang, Xiaofeng Zhu, and Neil Risch. 2006. "Reconstructing Genetic Ancestry Blocks in Admixed Individuals." *American Journal of Human Genetics* 79 (1): 1–12.

Tobler, W. R. 1970. "A Computer Movie Simulating Urban Growth in the Detroit Region." *Economic Geography* 46: 234.

Vernot, Benjamin, Serena Tucci, Janet Kelso, Joshua G. Schraiber, Aaron B. Wolf, Rachel M. Gittelman, Michael Dannemann, et al. 2016. "Excavating Neandertal and Denisovan DNA from the Genomes of Melanesian Individuals." *Science* 352 (6282): 235–39.

Wasser, Samuel K., Andrew M. Shedlock, Kenine Comstock, Elaine A. Ostrander, Benezeth Mutayoba, and Matthew Stephens. 2004. "Assigning African Elephant DNA to Geographic Region of Origin: Applications to the Ivory Trade." *Proceedings of the National Academy of Sciences of the United States of America* 101 (41): 14847–52.

Wasser, S. K., L. Brown, C. Mailand, S. Mondol, W. Clark, C. Laurie, and B. S. Weir. 2015. "CONSERVATION. Genetic Assignment of Large Seizures of Elephant Ivory Reveals Africa's Major Poaching Hotspots." *Science* 349 (6243): 84–87.

Watson, J. D., and F. H. C. Crick. 1953. “Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid.” *Nature* 248 (5451): 765–765.

Weir, B. S., and C. Clark Cockerham. 1984. “ESTIMATING F-STATISTICS FOR THE ANALYSIS OF POPULATION STRUCTURE.” *Evolution; International Journal of Organic Evolution* 38 (6): 1358–70.

Won, Y-J, and Y. -J. Won. 2004. “Divergence Population Genetics of Chimpanzees.” *Molecular Biology and Evolution* 22 (2): 297–307.

Wright, S. 1943. “Isolation by Distance.” *Genetics* 28 (2): 114–38.

Yang, Wen-Yun, John Novembre, Eleazar Eskin, and Eran Halperin. 2012. “A Model-Based Approach for Analysis of Spatial Structure in Genetic Data.” *Nature Genetics* 44 (6): 725–31.

Yengo, Loic, Matthew R. Robinson, Matthew C. Keller, Kathryn E. Kemper, Yuanhao Yang, Maciej Trzaskowski, Jacob Gratten, et al. 2018. “Imprint of Assortative Mating on the Human Genome.” https://doi.org/10.1101/300020.

Zuckerkandl, Emile, and Linus Pauling. 1962. *Molecular Disease, Evolution, and Genic Heterogeneity*.