University of South Bohemia in České Budějovice

Faculty of Science

Phylogeny of human populations in Papua New Guinea, a genetic and linguistic diversity hotspot

Master thesis

Klára Kopicová, Bc.

Supervised by: M.Sc. Pavel Flegontov, C.Sc.

České Budějovice, 2019

Kopicová, K., 2019: Phylogeny of human populations in Papua New Guinea, a genetic and linguistic diversity hotspot. Mgr. Thesis, [in English.] – 55p. Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic.

Annotation:

A detailed phylogeny of human populations in Papua New Guinea was constructed using exhaustive topology exploration, and the fit of the model to the data was improved by adding several admixture events. The analysis relied on published genome-wide SNP genotyping data for hundreds of individuals, and *qpGraph* was a principal method employed in the study for testing the fit of admixture graphs to the data.

Date, Location: Signature:

…………………. …………………………

Acknowledgment

I would like to express my gratitude to my supervisor, M.Sc. Pavel Flegontov, C.Sc., and his co-workers who guided me and helped me all the way long.

Contents

Preface

Anatomically modern humans (AMH) spread out of Africa before 75 thousand years ago (ka) (Pagani et al., 2016) and, as cranial fossils of an approximate age 44 – 63 ka from Tam Pà Ling (Laos) confirmed, they further expanded into Southeast Asia (SEA) (Demeter et al., 2017).

About ~4000 years ago, SEA farming economies were developing, expanding, and driving foraging groups to remote habitats. In the Neolithic period, SEA was occupied by indigenous hunter-gatherers of the Hòabìnhian archaeological culture and East Asian farmer migrants, whose admixture formed present-day groups in the region (Lipson et al., 2018; McColl et al., 2018).

During the last glaciation which lasted roughly from 33 to 19 ka (Clark et al. 2009), many islands of the Indonesian archipelago were joined in two large landmasses exposed by the low sea levels. To the west, the Malay Peninsula, Sumatra, Java, Bali and Borneo were joined to mainland Southeast Asia in a landmass known as Sunda. In the east, New Guinea and a number of current islands formed a single landmass with Australia and Tasmania, known as Sahul (Palmer, 2018).

The Sahul shelf of New Guinea and Australia, separated from the Sunda shelf of Southeast Asia, formed a significant biogeographical boundary first recognized by Alfred Russell Wallace in 1859. Wallace's Line separates the Asian fauna which includes primates, carnivores, elephants, and ungulates from the marsupial fauna of New Guinea and Australia. It is highly probable that the water barrier of Wallace's Line also prevented *Homo erectus* from reaching Sahul one million years ago despite reaching the Java Island at the same time (Allen, 2001).

The country of Papua New Guinea (PNG) occupying the eastern half of the New Guinea Island, the Bismarck Archipelago and the Bougainville Island shows human occupation since ca. 50 ka (O'Connell and Allen, 2015), including some of the earliest archaeological evidence for AMH outside of Africa (Rasmussen et al., 2011; Reyes-Centeno et al., 2015). The earliest currently claimed archaeological site in New Guinea dates back to 48000 years before present (YBP) (Hope and Haberle, 2005), which is still well before the arrival of maritime Austronesian-speaking peoples around 3.500 years ago (Palmer, 2018).

The distinctiveness of the Australian archaeological and fossil record has led to a suggestion that the ancestors of Aboriginal Australians and Papuans left the African continent even earlier than the ancestors of present-day Eurasians (Lahr et Foley, 2005), although this result is not supported by recent genetic studies (Malaspinas et al. 2016, Skoglund et al. 2016, Lipson et al. 2018, Posth et al. 2018). Papuan and Australians experienced two pulses of archaic human introgression: the Neanderthal pulse shared with other non-African human groups and the Denisovan pulse that is unique to them (Jacobs et al., 2019). Papuan and Australian ancestors possibly met Denisovans in Southeast Asia, although that is not known for sure. It was found that Papuans and Australians split between 25-40 ka (Malaspinas et al., 2016).

On the basis of their hunter-gathering lifestyle, and their physical characteristics (including small body size, dark skin pigmentation, cranio-facial morphology, and frizzy hair), Malaysian "negritos" are traditionally grouped with other negrito communities in SEA such as Andaman islanders (Onge), Mani in Thailand, and the Aeta in the Philippines (Aghakhanian et al., 2015). These groups and ancient Hòabìnhians are believed to be the closest relatives of Papuans and Australians on the Asian mainland (Mccoll et al., 2018).

Papua New Guinea is the most linguistically diverse region in the world that still keeps its fascinating cultural traditions alive. Despite its extraordinary human diversity - achieved in the absence of massive technology-driven expansions - New Guinea remains underrepresented in modern genetic surveys. Although Papuans are often involved in genomic comparative analyses as reference populations or 'outgroups', the genetic history within PNG itself remains poorly understood.

2. Studies of human history from a multidisciplinary perspective

Study of the human past requires cooperation of many scientific fields (archaeology, anthropology, linguistics, and genetics). Any scientific investigation of the past should start with consideration of different types of evidence available, since this provides the basis for testing hypotheses and refining models. Genetic history of an ethnic group, history of its material culture and language are rarely fully congruent, but considering all these histories side by side is important since they often put constrains on each other.

2.1. Historical linguistics

Human language is unique amongst animal communication systems not only for its structural complexity but also for its diversity at every level of structure and meaning. Historical linguists explores phylogenetic relationship of languages and their interactions, and thus is on the basic conceptual level similar to historical genetics as both aim to build a graph of divergence and admixture events, however on very different types of data. Related languages are grouped into "language families" and borrowing of words and grammatical structures across languages is studied too (Campbell and Poser, 2008). Deeper language relationships across universally recognized families are also studied actively, but just a handful of such proposals has gained universal acceptance in recent decades (Campbell and Poser, 2008).

Historical linguists traditionally used just qualitative methods for analysing language relationships. The comparative method involves finding regular phonetic correspondences across languages, and those correspondences are used for finding cognates, i.e. the same or closely related meanings expressed by related words. If a certain number of cognates within the most stable core set of 100 or 200 meanings called a Swadesh list is accumulated, a language relationship is considered proven (Campbell and Poser, 2008). Although the comparative method can give a vague idea about time depth of a language relationship, it cannot provide reliable date estimates (Gray and Atkinson, 2003).

In 1831, Samuel Rafinesque was the first who with the help of putting a number on the distance between languages explored the origin of Asiatic negritos. Although Rafinesque's negative finding (showing the languages of disparate negrito peoples to be unrelated) was never published, his method became popular (thanks to Jules Dumont d'Urville) and evolved into a new linguistic discipline: glottochronology (Jobling et al., 2014).

Glottochronology is a lexicostatistical approach that uses the percentage of shared cognates between languages. Based on this, divergence times are calculated by assuming a constant rate of lexical replacement (Bergsland and Vogt, 1962).

Nowadays glottochronology is considered an outdated approach, and current linguistic studies adopt more sophisticated phylogenetic approaches developed in evolutionary biology (Gray and Atkinson, 2003). Bayesian phylogenetic methods became especially popular for construction of language phylogenies based on basic vocabulary cognate sets and for dating tree nodes simultaneously (Gray and Atkinson, 2003; Bouckaert et al., 2012; Chang et al., 2015 Kassian et al. 2019).

Austronesian, Indo-European, Bantu and Uto-Aztecan are four language families which together encompass well over a third of the world's approximately 7,000 languages. To explore the linguistic diversity worldwide, it was found that no pairs of word-order features (verb– object order, adposition–noun order, genitive–noun order, relative clause–noun order) were strongly dependent in all of those four language families. Therefore, Dunn et al. (2011) concluded that linguistic diversity does not seem to be a result of universal cognitive factors specialized for language, but rather a product of cultural evolutionThis statement may serve as explanation why PNG developed into such linguistically diverse hot spot without sharing closer relativeness among its language phylums.

2.2 Papua New Guinea: a hotspot of linguistic diversity

The Papuasphere is a linguistic world that encompasses mainland New Guinea, the Bismarck Archipelago, Bougainville, the Solomon Islands, Halmahera, including the islands of Timor, Alor and Pantar. It is also the world's least documented region, both in terms of absolute numbers of languages, and the proportion of languages documented (Palmer, 2018). Papuashpere contains, at the current state of knowledge, 862 languages comprising 43 distinct language families and 37 "language isolates", i.e. language families composed of one language only (Palmer, 2018). Unlike the Austronesian languages that arrived to New Guinea Area around 3500 YBP (Palmer, 2018), other languages spoken in PNG (called "Papuan" for simplicity) share no generally accepted wider phylogenetic links.

Having no phylogenetic or typological status, Papuan refers to a group of families and isolates that share one crucial characteristic: all are endemic to the New Guinea Area (Palmer, 2018). Papuan languages are spoken by some three to four million people. An average Papuan language family includes about 25 member languages. Because there are many hundreds of Papuan languages, most of them are spoken by relatively few individuals, generally less than 3,000. Although the most commonly spoken Papuan language, Enga, has about 165,000 speakers, many Papuan languages have fewer than 100 speakers and some fewer than 50. Only seven languages appear to have 100,000 speakers or more. With one exception from Timor-Alor-Pantar, all of those probably belong to the Trans New Guinea (TNG) language family (Palmer, 2018).

The TNG family, spoken across all of the PNG highlands and large parts of the lowlands, and hypothesized to have spread alongside plant cultivation, is the largest family of Papuan languages (Pawley et al., 2005). Subdivided into three groups (the Finisterre-Huon group, the Eastern Highlands, and the Papuan Highlands group), about 20% of the total Papuan-speaking population speak TNG languages (Foley, 2003).

The list of potential TNG members probably includes also the Enga language family (spoken by more than 400,000 people in the Enga province) and the large Madang family (of more than 80 languages with some 80.000 speakers, spoken in the Madang province). Furthermore, there are still many Papuan language families (e. g. families near the border between Papua (Indonesia) and PNG, in the Sepik-Ramu basin) for which no evidence of genealogical relation with the TNG family has been found yet (Foley, 2003). If all candidate members of the TNG family are ultimately shown to belong to it, then it will include almost 300 languages and two million speakers, no less than 50 percent of the total Papuan-speaking population (Foley, 2003).

The lowland areas of New Guinea are also particularly complex, with many small language families having low numbers of speakers. This is especially notable in the West Sepik, Western, and Gulf provinces of PNG and adjoining areas of Western New Guinea, which also contain a number of isolated languages that cannot be classified with any larger language family.

2.3. Genomics

Based on variable regions in the genome (shaped by genetic recombination, genetic drift, and natural selection) evolutionary geneticists trace relationships among individuals and populations. Various types of genetic variability can be used in principle for population studies, and here I list them in the order they became widely used.

2.3.1. Blood groups and other classical markers

The ABO blood group system, discovered in 1900, was the first human genetic polymorphism to be discovered (Landsteiner, 1900). Based on the antigens exposed on the surfaces of red blood cells and their specific reactivity with antibodies, four classes of individuals were defined: those carrying only the A antigen, those carrying only B, those carrying both (AB), and those carrying neither (O).

The gene underlying the ABO blood group system encodes a glycosyltransferase enzyme that adds a sugar molecule to a carbohydrate structure known as the H antigen on the surface of red blood cells. The A allele codes for an enzyme which adds the sugar *N*-acetylgalactosamine, whereas the enzyme coded by the B allele has two amino acid differences which alter its specificity so that it instead adds D-galactose. The O allele has an inactivating mutation in the gene and so the H antigen remains unmodified (Yamamoto et al., 1990).

Later, more blood groups (such as MN and Rh) were discovered and also used in early population studies. In 1978, Menozzi et al. used allele frequency data in Europe in order to test the cultural diffusion hypotheses. Based on data from single genes (e.g. the Rh-negative alleles, or some HLA-B alleles), they tried to find correlations between genetic gradients across Europe (Menozzi et al., 1978). ABO and Rh loci are clinically very important and represent a key information in clinical and transfusion medicine (blood transfusion and organ transplantation). Between 50’s and 70’s, Mourant et al. also published an outstanding compendium of data on blood groups and other polymorphisms. From the current point of view (Bodmer, 2015), this was the first major work to give worldwide data on the distribution of the human blood groups in a way that enabled a much better genetic assessment of the relationship between the various human populations.

It was this source of information that enabled Edwards and Cavalli-Sforza (1963) and Cavalli-Sforza and Edwards (1965) to construct the first human evolutionary tree using gene frequency data. This evolutionary tree was an outstandingly original piece of work that has formed the basis of all subsequent phylogenetic analyses, and it introduced the use of Principal Components Analysis (PCA) for the interpretation of gene frequency data.

2.3.1.1. Single nucleotide polymorphisms

The ~7 billion people living today carry ~14 billion genome copies. The differences between genomes encompass a wide range of scales: from single nucleotide polymorphisms (SNPs) to structural variation involving millions of base pairs. These differences arise due to mutations or *de novo* changes within the genome (substitutions, indels [insertions, deletions], duplications, inversions). Interest in the identification of SNPs has been driven largely by their potential use as molecular markers in disease association studies. Base substitutions, occurring at an average rate of 10^{-8} per base per generation, are ~10 times more frequent than indels, although this relative frequency varies substantially across loci. Transitions (pyrimidine-pyrimidine/purine-purine exchanges) are almost three times more frequent than transversions (pyrimidine-purine exchanges) (Jobling et al., 2004).

Initially, SNP discovery studies were based on analysing a few loci in several dozens of individuals, and as a result many of the widely used SNPs have diverse and poorly documented origins. SNPs discovered through many different methods, mostly whole-genome resequencing studies (described by numbers prefixed with the letters ”rs”, for RefSNP) are deposited in the database dbSNP (www.ncbi.nlm.nih.gov/projects/SNP). At the time of writing dbSNP contains 686.6 million human SNPs. Many SNPs have been identified by computational analysis of various sequence data held in other databases. Validated SNPs are those that have been ascertained with a noncomputational method, or have been detected and genotyped in a population sample, these have associated allele frequency data (Jobling et al., 2004).

When a diploid genome is sequenced, SNPs can be identified directly as heterozygous base positions; SNPs that are homozygous in the sequenced genome can be identified by comparison with a reference sequence. SNP-based studies focused on so-called common variation normally consider SNPs with a minor allele frequency (MAF, at which the less common allele is found within a given population) of at least 5% (Jobling et al., 2004).

The other alleles are considered rare, and are sometimes targeted specifically for high-resolution exploration of population history, e. g. in East England (Schiffels et al., 2016) or in Chukotka and North Amerika (Flegontov et al., 2019).

2.3.1.2. Variable number of tandem repeats loci

Another class of genetic variation involves changes in the number of repeated DNA sequences arranged adjacently in tandem arrays, collectively known as variable number of tandem repeats loci (VNTR/satellites). VNTRs are usually classified according to the size of a single copy. Microsatellites/short tandem repetitions (STRs) are tandem arrays of repeat units from 1 to 7 bp in length, and those that have a useful degree of polymorphism have a typical copy number of 10-30. While some exceptions exist (e. g. the CAG repeat expansion responsible for Huntington’s disease), in general, variation at most microsatellites has no influence on the phenotype. Microsatellites composed of some specific repeat units may also show clustering, a special variation of microsatellites in which two or more individual microsatellites are found directly adjacent to each other (Kofler et al., 2008) which might be a useful property for population analysis (Parada-Rojas and Quesada-Ocampo, 2018).

Minisatellites consist of repeat units from 8 to 100 bp in length, with copy numbers from as low as 5 to over 1000. The minisatellites that have been well studied are a small and biased subset with particularly high diversity. A method known as minisatellite variant repeat PCR (MVR-PCR) allows these variant repeats to be mapped within arrays conveniently, and allows access to the details of the mutation process in sperm DNA. Mutation rate at many minisatellites is in fact highly elevated in males compared to females; as an extreme example, the minisatellite *CEB1* has a mutation rate of 15% per sperm, but only 0.2% per oocyte (Vergnaud et al., 1991).

Telomeres are essential structures at the ends of chromosomes, and include tandem arrays of the hexanucleotide repeat TTAGGG, typically 10–15 kb in length. The most proximal parts of telomere arrays contain variant repeats (for example, TCAGGG), and this source of variation has been used to study the dynamics of telomeres (Jobling et al., 2004).

Satellites, sometimes called macrosatellites, are large tandem arrays spanning hundreds of kilobases to megabases, and composed of repeat units of a wide range of sizes that can display a higher-order structure.

The mutation processes at these loci is too complex and therefore cannot be studied directly. Historically, some satellite polymorphisms have been used in human evolutionary studies (Rudd et al., 2006), but nowadays they have been superseded by loci that are easier to study and further analyse (e.g. SNPs).

45% of the human genome is composed of dispersed repeat elements with copy numbers ranging from a few hundred to several hundred thousand (Jobling et al., 2004). Human endogenous retroviruses (HERVs) have retained the ability to replicate themselves throughout the genome but do not have the ability to construct the protein coat that allows them to leave the cell. If these HERVs enter the germ line and replicate, integrated DNA copies of their genomes are inherited by future generations.

Structural variation is a broad term covering changes in chromosome structure. Some of these changes are balanced, involving no alteration of sequence copy number—examples are an inversion of a chromosomal segment or a reciprocal translocation between a pair of chromosomes. Others involve differences in the numbers of particular sequences between alleles—copy-number variation. In practice an arbitrary threshold of >1 kbp has generally been used to define structural variation, so it excludes small indels. In principle, insertions of long interspersed nuclear (LINE1) retroelements could be classified as structural variants, but they are generally not included because they form a coherent class with a well-understood basis.

2.3.2. Mitochondrial and Y-chromosome haplogroups

The great majority of the human genome is inherited from both parents, and undergoes reshuffling each generation through recombination. In 2007, a study published data for a set of 270 individuals (over 3.1 million SNPs including 25–35% of common SNP variation in the populations surveyed) (HapMap Consortium, 2007). The results of this study revealed an extremely non-uniform distribution of recombination in the genome (HapMap Consortium, 2007). There are two segments of our DNA that are atypical, being inherited from one parent only, and escaping recombination: most of the Y chromosome, and mitochondrial DNA.

Mitochondria, double-membrane-bound organelles within cells that are critical for energy metabolism, possess their own genome. The human mitochondrial genome (mtDNA) is composed of 16,569 DNA base pairs, whereas the nuclear genome is made of 3.3 billion base pairs. The mitochondrial genome contains 37 genes that encode 13 proteins, 22 tRNAs, and 2 rRNAs.

The 13 mitochondrial encoded proteins all instruct cells to produce protein subunits of the enzyme complexes of the oxidative phosphorylation system (Parr et Martin, 2012).Repeated comparisons of whole human mtDNA and chimpanzee mtDNA sequences indicated that the base-substitution mutation rate in mtDNA is about 10 times higher than the average rate in nuclear DNA (Jobling et al., 2004). An intermediate rate (5×10^{-7}) is predominantly observed in conserved noncoding regions of the control mtDNA region, and the lowest mutation rate (2×10^{-7} per bp per generation) mainly in mitochondrial RNA genes (Soares et al., 2009). A more than tenfold higher mutation rate (5×10^{-6}) observed in hypervariable segments (HVSI and HVSII) of the control region is due to the presence of a number of evolutionarily unstable hotspot positions that make the use of control region data problematic in interspecies comparisons, but useful for population-level studies.

Oocytes contain around 100,000 mitochondria (each containing a single mtDNA molecule) while sperm contain only about 50–75 (Jobling et al., 2004). Because paternal mitochondria from sperm do not persist after fertilization, mtDNA is maternally inherited in haploid form, and therefore escapes recombination. This fact serves is useful for investigating the maternal ancestral linage.

In humans, the sexual differentiation is controlled by a pair of sex chromosomes (also called gonosomes, heterochromosomes, or heterosomes), the X and Y chromosomes. Homogametic *XX* individuals are females and heterogametic *XY* are males. Since it has no homolog with which to recombine, the Y chromosome avoids recombination for more than 90% of its length. However, in specialized regions (pseudoautosomal region, PAR) where sequence similarity with the X chromosome is preserved, recombination occurs between the Y and the X chromosome (Jobling et al., 2004).

Pseudoautosomal region 2 (PAR2), lying at the tips of the long arms of the X and Y chromosomes is a recent evolutionary acquisition specific to humans, and of little importance in chromosomal segregation. However, PAR1, a 2.6-Mb region at the tips of the short arms, derives from the ancient origin of the mammalian sex chromosomes as a pair of homologous autosomes some 300 MYA. At this site a recombination event occurs in the every single male meiosis (Jobling et al., 2004).

MtDNA (and Y) system can be considered as a system of small, sexually isolated genetic units (clonal lineages), accumulating diversity only through mutation, with an evolutionary rate faster than the nuclear genome, which has made these molecules ideal for haplotype-based evolutionary analyses (Castro et al., 1998).

A haplotype refers to a combination of allelic states of polymorphisms along the same DNA molecule (the same chromosome or mtDNA).

The first human population studies based on mtDNA were performed by restriction enzyme analyses (Denaro et al., 1981; Merriwether et al., 1991), and they e.g. revealed differences between the four major races (Caucasian, Amerindian, African, and Asian). Differences in mtDNA patterns have also been shown in communities with a different geographic origin within the same ethnic group (Bonné-Tamir, et al., 1986; Soodyall and Jenkins, 1993).

Although very popular, after the development of high-throughput genome sequencing during the first decade of 21th century maternal and paternal single-locus studies became outdated and were replaced by more powerful methods based on autosomal genetic variations.

2.3.3. Genotyping methods

To detect the variation in a sample of genomes, the ideal way is to carry out comprehensive resequencing. Early methods for studying genetic differences were indirect, based on immunological reactions, later on electrophoretic analysis of gene products or with the help of Southern blotting analysis. Before the revolutionary invention of the polymerase chain reaction (PCR), the sensitivity for analysing single-copy sequences in the large and complex human genome was problematic; specific sequence detection was accomplished by the use of DNA probes, cloned sequences labelled with the radioisotope phosphorus – 32 (Jobling et al., 2014).

2.3.3.1. PCR

In the early 1980s, Kary Mullis got the original idea for PCR. The synthesis of short single-stranded DNA (oligonucleotides) and the use of these to direct target-specific synthesis of new DNA copies using DNA polymerases were already standard tools, but the novelty was to use the juxtaposition of two oligonucleotides, complementary to opposite strands of DNA, to specifically amplify the region between them and to achieve this in a repetitive manner so that the product of one round of polymerase activity was added to the pool of template for the next round, hence the chain reaction (Bartlett and Stirling, 2003).

Today, using pre-designed short (typically 18–24 nucleotides) oligonucleotide primers, PCR provides a way to amplify individual gene copies directly, cheaply, and quickly and became one of the most widely used tools in molecular biology.

2.3.3.2. Shotgun genome sequencing

Soon after the DNA sequencing methods were invented (Sanger et Coulson, 1975, Maxam et Gilbert, 1977) and after the first human gene was isolated and sequenced (Seeburg et al., 1977), the shotgun sequencing strategy was introduced (Anderson, 1981) and remained a fundamental method for large-scale genome sequencing for the next three decades.

The application of shotgun sequencing (a technique in which large pieces of DNA are sheared into smaller fragments, sequenced randomly, realigned, and ordered into larger contiguous pieces that represents the original) was successively extended by applying it to larger and larger DNA molecules from plasmids (4 kb), bacterial genomes (1-2 Mb) up to the human genome sequence, being 25 times as large as any previously sequenced genome and 8 times as large as the sum of all such genomes (Lander et al., 2001).

Direct sequencing with the shotgun method is based on assumption that a genome containing no repeated sequence could have been uniformly sampled at random. However, practical difficulties for shotgun sequencing arise because of repeated sequences and cloning bias. Bacterial genomes contain about 1.5% of repetitive sequences on average, the fruit fly genome about 3%, and the human genome about 50% (Lander et al., 2001).

2.3.3.3 Applied sequencing

In late 1990s, two independent scientific teams focused on sequencing the first complete human genome. The initial one (HUGO Project, 1990-2003), having used the hierarchical shotgun sequencing strategy, was based on a collaboration of 20 groups (from the US, UK, Japan, France, Germany and China) and generated the draft genome sequence from a physical map (made from anonymous volunteers chosen at random) covering more than 96% of the euchromatic part of the human genome. Together with additional sequences in public databases, HUGO achieved sequencing about 94% of the human genome (having finished it over roughly fifteen months) (Lander et al., 2001).

Meanwhile, having used two more advanced assembly strategies (whole-genome assembly and regional chromosome assembly), a parallel effort sequenced a 2.91-billion bp consensus sequence over 9 months from 27,271,853 sequence reads with 5.11-fold average coverage of the genome (Venter et al., 2001).

Whereas the Sanger sequencing technology (applied for the human genome project, 1990-2005) was able to generate up to 700–1000 bp long reads (International Human Genome Sequencing Consortium, 2004) with a very low error rate [www.genome.gov/human-genome-project]).

Since 2005, newer sequencing technologies (commonly referred to as next generation sequencing or NGS) were designed to produce in general shorter reads (100–500 base pairs) but resulting in a higher error rate (~0.1%) (Dal et Alkan, 2018).

2.3.3.4. Illumina sequencing

The most popular NGS technology is Illumina, which provides high-quality reads of 500 nt DNA fragments (Tan et al., 2019), also with relatively high error rates at $0.1 \sim 0.01\%$ frequency [Ma et al. 2019]). Following DNA fragmentation, ligation of adapters, and immobilization on the surface of a flow cell, bridging amplification generates spatially separated clusters each containing ca. 1000 identical molecules. Based on the adapter-primer compatibility (having applied a suitable buffer, sequencing primers complemented to the template base, terminal fluorescently labelled nucleotide, and a specialized DNA polymerase), DNA polymerase starts to replicate small clusters of DNA with the same sequence. After washing steps, the double-stranded DNA is then broken down into single-stranded DNA using heat. The DNA polymerase then adds the first fluorescently-labelled terminator to the new DNA strand so that once a base has been added no more bases can be added to the strand of DNA until the terminator base is cut from the DNA. Lasers are passed over the flowcell to activate the fluorescent label on the nucleotide base and the fluorescence is detected by camera and recorded. When fluorescence imaging is performed to determine the identity of the incorporated nucleotide, and the cyclical process continues, yielding typical read lengths of > 100 bp.

While the sequencing protocols described above starts from one end of each DNA fragment, the practice of paired-end sequencing was introduced by Hong (1981). The first straightforward paired-end sequencing also cloned DNA inserts into bacteriophage vectors. However by sequencing from both ends, twice as much sequencing data from long inserts was produced (Hong, 1981). While both PCR and NGS offer highly sensitive and reliable detection of genetic variants, PCR can only detect known variants. In contrast, NGS is a hypothesis-free approach that does not require prior knowledge of sequence information.

Whereas PCR is effective for low target numbers, the workflow can be cumbersome for genotyping multiple polymorphisms. A single NGS experiment can identify variants across thousands of target regions at a single-base resolution.

2.3.4. Genotyping arrays

The most common type of genetic polymorphism, a SNP, is a difference between chromosomes in the base present at a particular site in the DNA sequence. Any two copies of the human genome differ from one another by approximately 0.1% of nucleotide sites (Halushka et al., 1999). The specific set of alleles observed on a single chromosome, or part of a chromosome, is called a haplotype. New haplotypes are formed by additional mutations or by recombination when the maternal and paternal chromosomes exchange corresponding segments of DNA, resulting in a chromosome that is a mosaic of the two parental haplotypes (Pääbo, 2003). The coinheritance of SNP alleles on these haplotypes leads to associations between these alleles in the population (known as linkage disequilibrium, LD) (Gibbs et al. 2003). Because the likelihood of recombination between two SNPs increases with the distance between them, on average such associations between SNPs decline with distance.

Genome-wide single-nucleotide polymorphism (SNP) chips (arrays) that allow genotyping hundreds of thousands of known SNPs from the human genome have proven useful for studying important questions in human genetics. In 2000, Dunning et al. examined ~1,600 individuals from four European populations with different demographic histories (Afrikaners, Ashkenazim, Finns, and East Anglian British). Based on the comparison of their allele frequencies or LD, they found only little evidence for major differences among those of studied European populations.

One year later, using a freshly available dense genome-wide map of SNPs, Reich et al. (2001) reported that LD in a United States population of North European descent typically extends 60 kb from common alleles. In contrast to Europeans, LD in a Nigerian population displayed an extension markedly less far. Both results illuminated human history, suggesting that LD in northern Europeans was shaped by a marked demographic event about 27,000-53,000 years ago (Reich et al., 2001).

Current versions of such chips can accomplish the remarkable feat of assaying >1 million SNPs simultaneously in a DNA sample with >99% accuracy and reproducibility, for a few hundred dollars. An example of such a technology is the Infinium™ assay used in Illumina™ bead chips such as the 1.2M-Duo~~,~~ (Jobling et al., 2014). The development of such genotyping platform, which can assay several million SNPs (Jobling et al., 2014) across thousands of samples, is essential for e. g. the investigation of genetic differences between patients for the purpose of improving their diagnosis, prognosis and therapeutic intervention (Gunderson et al., 2006), when ultimately~~,~~ one examines whether the polymorphisms occur more frequently in affected individuals or in an appropriate control population (Tebbutt et al., 2004).

2.4. Computational methods for analysing population history

Human genetics utilizes a wide array of data analysis methods, and most common methods can be subdivided into the following classes: ordination or dimensionality-reduction methods, *f*-statistics and model-based methods.

2.4.1. Ordination methods

2.4.1.1. PCA

Since 1901, principal component analysis (PCA) has been employed as a useful statistical tool to reveal internal structure of datasets (Pearson, 1901). PCA is a type of a dimensionality reducing technique, and those reduce a large number of variables to a smaller number, preserving most of the variance present in the original dataset. When thousands of variables are analysed for each sample, a large fraction of them is usually non-independent (correlated). Thus, principal components (PCs) are linear combinations of original variables – allele frequencies at various sites in the case of autosomal genetic data.

A predefined number of PCs is calculated, and they are ordered by decreasing variance represented by the component: PC1 explains the largest share of the variance, PC2 a smaller share, etc. Usually samples are plotted in the PC1 vs. PC2 coordinates only, i.e. a two – dimensional space is visualized. PC1 can be interpreted as the longest line through a multidimensional cloud of sample points, PC2 as the longest line orthogonal to PC1, PC3 as the longest line orthogonal to PC2, and so on.

PCA on multi-locus SNP genetic data is implemented in a number of software packages, most notably smartPCA (Patterson et al., 2006). PCA "embeddings" of genetic data have a useful property: admixed individuals usually fall between ancestry sources on a plot, thus creating ”admixture clines”. However, distribution of samples in the space of PCs can be affected not only by admixture, but by genetic drift too. Moreover, PCA plots are affected by over-representation of certain populations. November et al. (2008) also showed that isolation-by-distance approach can generate PCA gradients that are similar to those that arise from admixture. Thus, interpretation of PCA plots is not straightforward, and they cannot be used for testing hypothesis about genetic admixture. They are very useful for formulating those hypotheses, to be tested by other methods.

2.4.1.2. UMAP

The multidimensional scaling (MDS) is a dimensionality-reduction method widely used in community ecology, and Uniform Manifold Approximation and Projection (UMAP) is a novel non-linear algorithm for dimensionality reduction applicable to a wide variety of data types (Becht et al., 2019; Diaz-Papkovich et al., 2019; McInnes and Healy, 2018). UMAP takes a matrix of all vs. all distances among samples and represents it in two or more dimensions relying on a nearest neighbour graph (McInnes and Healy, 2018). The most important algorithm setting is the number of nearest neighbours considered. This setting controls how UMAP balances local versus global structure in the data (McInnes and Healy, 2018). An optimal nearest neighbour setting depends on the total dataset size (in samples) and can be expressed in absolute or relative terms. UMAP applied to multi-locus SNP genetic data demonstrated increased resolution as compared to PCA in some cases (Diaz-Papkovich et al., 2019), however admixture clines do not appear, which is a disadvantage of the method.

2.4.2. *f*-statistics

Many questions about human genetic history can be answered by examining the patterns of allele frequencies in sets of populations. A useful methodological framework for this purpose called *f*-statistics measures shared genetic drift in sets of two, three, and four populations. *f*-statistics are usually used to test simple hypotheses about admixture between populations.

Demographic models called population phylogenies (or population trees) are models where populations are related in a tree-like fashion, and it serves as a null model for admixture tests. The branch length in the population phylogeny reflects the magnitude of genetic drift so that a branch that is subtended by two different populations can be interpreted as “shared” genetic drift between these populations.

For any four groups there are three possible simple trees: ((A,B),(C,D)), ((A,C),(B,D)), and ((A,D),(B,C)). If the ((A,B),(C,D)) tree is correct, the allele frequency differences between A and B should be uncorrelated with those between C and D, which can be assessed by averaging the quantity (pA-pB)(pC-pD) across SNPs, and further testing for consistency on resampled datasets (Reich et al., 2009; Peterson et al., 2012).

2.4.2.1. f_3-and f_4-statistics

Under a population phylogeny, the three *f*-statistics proposed by Patterson et al. (2012), labelled f_2, f_3, and f_4, have interpretations as branch lengths between two, three, and four taxa, respectively.

f_3-statistics are defined as the product of allele frequency differences between population C to A and B, respectively. f_3-statistics can be used in two ways: first, as a test whether a target population (C) is a mixture of sources (distantly) related to two source proxy populations (A and B); second, as a measure of drift shared between two test populations (A and B) given an outgroup (C) (Patterson et al., 2012).

f_4-statistic was introduced by Reich et al. (2009) and is a powerful measure to distinguish introgression from incomplete lineage sorting, based on allele frequencies in four populations. With populations A, B, C, and D, and the assumed population topology (A,B),(C,D), the f_4-statistic is calculated as the product of the difference of allele frequencies between A and B, and between C and D. Under incomplete lineage sorting alone, the allele frequency differences between A and B should be independent of those between C and D, and the f_4-statistic should thus be zero. If there was introgression however between a pair of populations (e.g. A introgressed into C), this would lead to non-zero f_4-values. Importantly f_4-statistics can be represented as linear sums of f_3- and f_2-statistics, in accordance with their interpretation as phylogenetic distances.

2.4.2.2. *D*-statistics

The *D*-statistic was originally designed to be applied on a genome-wide or chromosome-wide scale to identify an excess of shared derived polymorphisms (Green et al. 2010). The test, considering ancestral “A” and derived “B” alleles, is based on the prediction that two particular SNP patterns without gene flow (termed “ABBA” and “BABA”) should be equally frequent.

Given three populations and one outgroup with the relationship (((P_1, P_2), P_3), O), ABBAs are sites at which the derived allele B is shared between the non-sister taxa P_2 and P_3, whereas P_1 carries the ancestral allele, as defined by the outgroup. Similarly, BABAs are sites at which the derived allele is shared between P_1 and P_3, whereas P_2 carries the ancestral allele. A significant excess of ABBAs over BABAs is indicative either of gene flow between P_2 and P_3, or some form of non-random mating or structure in the population ancestral to P_1, P_2, and P_3. (Martin et al., 2014).

Under the null hypothesis - no gene flow and random mating in the ancestral population - *D* will approach zero, regardless of differences in effective population sizes. If *D* is significantly greater than zero, it indicates a significant excess of shared derived alleles between P_2 and P_3 (Durand et al., 2011).

The D statistic is insensitive to some demographic assumptions such as ancestral population sizes. Therefore, only one assumption is needed: the ancestral populations are required to be randomly mating (Durand et al., 2011). Otherwise, *D*-statistics is a useful approach, which can be used to detect archaic admixture even when no archaic sample is available.

2.4.3 Model-based approaches

2.4.3.1. Allele frequency spectrum

In population genetics, the allele frequency spectrum (AFS, sometimes also called the site frequency spectrum) is the distribution of allele counts for a given set of loci in a population. Calculating the AFS from observed sequence data requires one to be able to distinguish the ancestral and derived (mutant) alleles, often by comparing to an outgroup sequence. The AFS provides a convenient statistic, and much attention has been paid to predicting theoretical expectations of the AFS under a number of different models.

2.4.3.2 ADMIXTURE

ADMIXTURE is a software tool for maximum likelihood estimation of individual ancestries from the multi-locus-SNP-genotype datasets, not relying on any phylogenetic tree. In other words, AMIXTURE implements a model-based Bayesian approach to compute a matrix of ancestral-population fractions in each individual and infer allele frequencies for each hypothetical ancestral population.

As compared to older clustering methods (FRAPPE, STRUCTURE), ADMIXTURE is faster and is also able to generate the allele-frequency estimates for the ancestral populations (Alexander et al., 2009). Because of working either without an explicit historical model, or with an unrealistic model that assumes that all the ancestral populations have radiated from a single group, STRUCTURE/ADMIXTURE results are difficult to interpret historically (Patterson et al, 2012, Lawson et al. 2018).

ADMIXTURE between populations is a fundamental process that shapes genetic variation and disease risk. There are two main classes of methods studying ancestral sources of gene flow (Patterson et al, 2012): local ancestry-based methods (1) and global ancestry-based methods (2). Although the local ancestry-based method (1) traces ancestry at each locus in the genome, and therefore provides individual-level information about ancestry, it displays only reduced power to detect older events. The most commonly used methods for studying global ancestry (2) are: principal component analysis (PCA) (Patterson *et al.* 2006), model-based clustering methods such as STRUCTURE (Pritchard *et al.* 2000) and ADMIXTURE (Alexander *et al.* 2009). STRUCTURE/ADMIXTURE results are also difficult to interpret historically, because these methods work either without explicitly fitting a historical model or by fitting a model that assumes that all the populations have radiated from a single ancestral group, which is unrealistic.

AdmixTools is a software package that implements five methods: the 3-population test (f_3-statistic), *D*- and f_4-statistics, f_4 ratio estimation, admixture graph fitting and rolloff. Admixture graph allows building complex models of population history: based on a statistical test for admixture which might have occurred even hundreds of generations ago (f_3-statistic); based on the directionality of the gene flow (as assessed by the *D*- and f_4-statistics) and estimated mixture proportions (as estimated using f_4 ratios). Admixture graph fitting allows one to build a model of population relationships for an arbitrarily large number of populations and to assess whether it fits the allele frequency correlation patterns among populations.

Four out of five algorithms in AdmixTool rely on allele frequency data at sites across the genome which might be in linkage equilibrium or in disequilibrium. The fifth method, rollof, is an approach for estimating the date of admixture which models the decay of linkage disequilibrium in the target population (Patterson et al, 2012). Simulations published by Patterson et al. (2012) reported an unbiased dating up to 500 generations in the past. With help of AdmixTools Patterson et al. (2012) found a clear signal of admixture into northern Europe, with one ancestral population related to present-day Basques and Sardinians and the other related to present-day populations of northeast Asia and the Americas. Later these results were beautifully confirmed by pioneering archaeogenetic studies (Raghavan et al. 2014, Lazaridis et al. 2014).

Whereas in an AdmixTools workflow, the user must first generate a set of text configuration files tailored to each individual analysis, Petr et al. (2019) introduced a new simplified process automating all low-level configurations. This new R package *admixr* provides a convenient interface for performing reproducible population genetic analyses.

2.4.3.3. GLOBETROTTER

GLOBETROTTER is an R program that can identify and date admixture events occurring in the ancestral history of a given target population within the last ~4,500 years. In other words, it allows us to attempt to “reverse” the admixture process and infer the haplotypic makeup of admixing source groups as well as admixture date(s) (Hellenthal et al, 2014). Compared to similar programs, this method does not require the source surrogates to be *a priori* specified. So GLOBETROTTER may be provided with DNA information on surrogates that may or may not be related to ancestral sources of the target population. GLOBETROTTER then uses these sampled groups to identify whether the target population descends from any admixture event(s). Confirming that, it determines precisely when the event(s) occurred and the admixing source groups involved. However, it is able to infer and date no more than two admixture events.

To describe each original admixing source, GLOBETROTTER uses two characteristics: first, picks a single sampled surrogate group that genetically best represents the admixing source, but second, also represents the admixing source as a mixture of the DNA of all sampled surrogate groups (often many of which are inferred not to contribute to the mixture at all) (Hellenthal et al, 2014).

2.5.3. Papua New Guinea

Melanesia, as a consequence of its colonial and recent post-colonial past, is divided into two halves: the western part belongs to Indonesia (the West Papua and other provinces), and the eastern part is (since 1975) an independent country Papua New Guinea (PNG) with Port Moresby as its capital.

2.5.3.1. Climate and geography

New Guinea, a tropical island located in the southern part of the Pacific Ocean northwest of Australia, belongs with its area of 785,753 km^2 is the second-largest island in the world. Monsoonal climate of two main seasons: in general, the very hot and humid wet season from December to March, and the dry season, interrupted occasionally by strong rains, from May to October, influence the agricultural strategy in this area (farming at fields cut across the rainforest) (Moffatt, 2012). There are significant differences between the lifestyle in the heart of the island, where mountains dominate the landscape, and on flatlands along the coasts.

The territory of PNG had a complicated colonial history. Although Australians had explored the densely-occupied highlands region (altitudes of ca. 1500 m - 2100 m) just before World War II, they did not start interacting with the region (e.g. through coffee plantations) until the post-war period (Bergström et al., 2017). Colonies were set up in the southern coastal region in 1883, and then spread over 100 years ago to the northern lowlands, and very much later to the highlands till 1975, the year of independence, when PNG also joined the United Nations (Gilliam, 1988). Nowadays, 97% of the PNG country’s land is owned and managed under customary tenure and stewardship. Communities have the final say in all resource management decisions, as guaranteed by the country’s constitution. Clans or individuals usually own customary land, which is usually inherited along paternal or maternal lines (JICA, 2002). Current highlands are still by far the most densely populated rural areas in PNG (mostly >50 persons/km^2 and in some areas >100 persons/km^2). Compared to that, the southern coastal region of mainland PNG has a very low population density, especially in the Gulf and Western provinces. The island PNG provinces show patterns similar to the mainland lowland regions (Bergström et al., 2017).

2.5.3.2. Modern history

In 1526, Portuguese sailor Jorge de Meneses as first European visitor named one of the islands "ilhas dos Papuas" or "land of fuzzy-haired people". Britain established a protectorate, British New Guinea (BNG), over south-east New Guinea in 1884, while Germany annexes the northern part of New Guinea (bbc.com, 2018).

Contacts between Europeans and Papuans were rather rare until the end of the 19th century (Soukup, 2010). One of the reasons may be that trading companies preferred quick profits, which New Guinea from thr geographical as well as climatic viewpoints did not offer at the first sight. That is why European traders, travellers and scientists were initially disinterested to gain a detailed knowledge of the New Guinea peoples.

From the anthropological viewpoint it is possible to distinguish two independent phases in Melanesian history since the start of colonisation. The first phase is the very end of the 19th century, when a German colony was established in New Guinea in 1884. To gather first ethnographic records before the colonization, a scientist Nicholas Miklouho-Maclay (1846–1888) spent a long time among natives in the lowlands, in the present-day Madang district, and undertook a long-term research within New Guinean communities. The second phase in Papuan history began at the turn of the 20th century when religious missions appeared on the Papuan mainland. For the first time, the field anthropological research became established. Additionally, philosopher Gunnar Landtman conducted his fieldwork (having used the method of participant observation) on the Kiwai Island in the delta of New Guinea's longest river Fly (Soukup, 2010).

After the war, in 1921, the League of Nations grants Australia a mandate to run German New Guinea. This new Mandated Territory of New Guinea is governed totally separately from the Territory of Papua. In 1942, Japanese forces occupied parts of both territories because of its strategic point in Pacific war. In July **1949,** Australia established a joint administration over both territories called the Territory of Papua and New Guinea. Almost two decades later, United Nations transferred control of West New Guinea to Indonesia (today this region is called West Papua) (bbc.com, 2018).

On 16th September, Papua New Guinea attained full independence from Australia, having its own new currency, the kina (replacing the Australian dollar in April **1975)/**

April - New currency, the kina, replaces the Australian dollar (bbc.com, 2018).

2.5.3.3. Papuans

Some 80% of Papua New Guinea's people live in rural areas (bbc.com, 2018), focusing mostly on farming. Indigenous settlers developed agriculture in PNG very early because of the island’s fertile soils in the highlands, abundant rainfall and the presence of many plant species suitable for cultivation. Agriculture in New Guinea is characterized by mulching (adding organic material on soil and plant roots to protect them from crusting and erosion), crop rotations and tilling, practices which are used on terraces with complex irrigation systems (Diamond, 1997).

Animal food was hard to find until dogs, introduced by the Austronesians, started being used for hunting. Dogs are believed to be the cause of the extinction of several mammal species in New Guinea (Diamond, 1997).

For more than 100 years, researchers from overseas have travelled to what is now the Independent State of Papua New Guinea to conduct ethnographic research. Their numbers have been large (see Bulmer 1969, Strathern 1973). As Department of Anthropology and Sociology, University of Papua New Guinea announced (Morauta et al., 1979), during the second half of 1977, for example, there were 57 noncitizens engaged at some time or other in field research in anthropology or sociologists focusing on variable topics: culture contact (Kituai, 1974; Sarei, 1974), cargo cults (Narokobi, 1974; Sarei, 1974; Waiko, 1973; Waiko, 1976), traditional marriage (Sarei 1974), the oral tradition (Waiko, 1973), and ritual (Talyaga, 1975).

The low population density in the lowlands compared with the very high population density in the highlands (Bourke et Harwood, 2009) has long been regarded, on detailed medical epidemiological evidence (Riley, 1983), to be largely a result of high malaria endemicity in the coastal regions. This high endemicity in the lowlands (<100 m above sea level) contrasts with hypoendemic to absent malaria transmission throughout the highlands.

Village subsistence centers on horticulture, with men clearing forests and bush so that their wives can plant gardens and tend pigs. Some crops, such as bananas, sugarcane, and cash crops (such as coffee and cocoa) are planted and tended by men. While women often help pick cash crops, most of the income goes to men. Men build houses and fences, while women make grass skirts and net bags (*bilums*). Women do the daily cooking, while men butcher pigs for feasts. Both men and women look after small children, with a father tending his infant while the mother weeds her gardens. In town, most women do domestic chores and child care while their husbands are at work (Bray and Smith, 1985).

The basic village household consists of a husband, a wife, their unmarried children, and perhaps the husband's parents. Extended families live in adjacent houses, gathering frequently for meals, companionship, work parties, and ceremonies. Men's houses are no longer common, although young men may live with other bachelors. Household decisions involve consensus between able-bodied adults, although young wives defer to older members. Residence is usually patrilocal. Less common is matrilocality and avunculocality. Land and property rights generally pass from parents to children or from uncles to nieces and nephews (Bray and Smith, 1985).

Each community has its own taboos surrounding class, status, and custodianship of areas, and this differs between each village. The passing down of cultural artefacts, skills and customs is also very complex and there are intricate rules around taboos and beliefs. There are many taboos in Papua New Guinean cultures around gender and sexuality. Homosexuality is illegal in Papua New Guinea (Moffatt, 2012).

Women marry out, and migrants move far from their ancestral territories to find wage employment and other benefits in town. Land is valuable and a way of life for 85 percent of the population (Bray and Smith, 1985).

2.5.3.3.1. Lowlanders

In the lowlands, two main ethnic clusters arose: one in the north and the second in the south. Both of them displaying a low to very low population density (mostly <10 persons/km^2) (Bergström et al., 2017). One of the limiting factors on population growth in the past may have been the nutritional quality of available foods in a country where cereals did not occur, rather than any deficiency in the agricultural potential of the land (Riley, 1983). Malcolm (1970) reported extremely slow growth rates amongst the inland groups living at an altitude of 600-2000 m. Maximum height in males (159"0 cm) was not reached until the age of 24 years and, in females (150"3 cm), until the age of 21 years. The age of menarche was delayed until 18'0 years and could be expressed as a function of the average height of women.

As Scragg further published (1973), probably for cultural reasons, a woman's reproductive span was only 13"7 years and the total fertility rate 4"8. In contrast to that, women on islands, where there has been considerable social and economic development, displayed their reproductive span of 21"7 years and the total fertility rate 9'5.

Vines (1970), in the course of an epidemiological survey of a sample of the population drawn from the New Guinea highlands, islands and mainland regions found significant differences between all regions in the adult heights of both males and females.

Island adults were taller than those from the mainland who in turn were taller than those from the highlands. The changes in growth which seem to be occurring later are associated with increases in the consumption of imported foods and, thus, food dependency in both rural and urban areas (Heywood, 1983).

2.5.3.3.2.Highlanders

In 1930, an Australian traveller, prospector and adventurer Michael Leahy (1901–1979) with his crew penetrated at that time practically unknown New-Guinean plateau at the altitudes of ~1500 m - 2100 m (Soukup, 2010).

Soon, in the beginning of 1935, colonial officials forbade visiting the highland areas by Europeans. This protection remained in place until the Japanese invasion in December 1941 (Soukup, 2010). During 40's, highlands were not controlled by local authorities and scientific research did not take its course due to the Japanese occupation. A strong wave of anthropological research then began in the highlands in the fifties. In 1965, opening of a major highway ended the long-term isolation of the highlands (Soukup, 2010).

The first genetic sample collection, finished in 1984, was carried out only 19 years after the coast-to-highlands highway was commenced. The greater isolation of the highlands means this region was probably less affected than the lowlands by intra-PNG migrations during the colonial period (Bergström et al., 2017).

Current five highlands provinces (Simbu, Eastern Highlands, Enga, Southern Highlands, and Western Highlands) form a discrete, geographically inter-connected and isolated central region, with very limited chain-trade connections to the coastal lowlands. In the highlands, people split into three distinct clusters (western, eastern and Angan speaers) within the past 10,000 years, soon after they began cultivating plants (Bergström et al. 2017).

Becroft (1967), working in the Western Highlands, reported an average birth interval of five years and an interval of 1-5 years following an early death. From the morphological point of view, the highlanders had significantly larger antero-posterior chest, biacromial diameter and bicondylar humerus measurements, and a higher cormic index (Harvey, 1974). Two detailed comparisons of physical traits have been made of people in highland areas. Littlewood (1972) showed a cline of decreasing adult male stature in the Kainantu district of the Eastern Highlands Province from northeast to southwest--158 cm among Gadsup to 150 cm among Awa.

Littlewood suggested that there could have been sufficient gene flow across linguistic boundaries from the taller people in the Markham valley into Gadsup to produce the cline across the study area to the shorter Anga people to the south of Awa. He admits the possibility also that the observed differences in stature and other "environmentally sensitive" traits may be primarily phenotypic and need not imply genetic changes (Littlewood, RA. 1972).

Today permanent settlements reach a maximum altitude of about 2,300 m. The men utilise the mountain forest almost up to the tops of the cordillera, especially for snaring small masupials, less often for hunting them with bow and arrows (Denham, 2005). All highland groups had cultivated plants, using intensive traditional horticulture for up to 10 kya with continuity from previous foraging practices (Bourke et Harwood, 2009), previously growing taro and more recently (300-400 yr) predominantly sweet potato.

2.5.3.3. Genetic studies of Papuans

Several hypotheses have been proposed to explain why present-day indigenous people of Near Oceania (New Guinea, the Bismarck Islands, and the Solomon Islands area) and Remote Oceania have ancestry both from Papuans and from populations of ultimate East Asian origin. In one set of models that has been favoured by recent genetic studies (Kayser, 2010; Wollstein et al., 2010; , Duggan et al., 2014; Matisoo-Smith, 2015) the mixture occurred at around 3,000 YBP, during the expansion of populations of East Asian origin through the New Guinea region (Kayser et al 2000). In the other set of models, the population of ultimate East Asian origin initially mixed little with Papuans (Blust, 2008), and thus later gene exchanges account for the ubiquitous Papuan ancestry today (Friedlaender, 2008, Posth et al. 2018, Skoglund et al. 2016, Lipson et al. 2018).

In stark contrast to the situation today (all present-day populations in Near and Remote Oceania harbor >25% Papuan ancestry implying the additional eastward migration [Lipson et al., 2018]; early and geographically diverse Remote Oceanian individuals had little or no Papuan ancestry. This scenario contradicts the models in which there was a significant Papuan contribution to the Lapita people (prehistoric Oceanian people, ancestors of historic cultures in Polynesia, Micronesia, and some coastal areas of Melanesia) before their dispersal into Remote Oceania (Skoglund et al., 2016).

As Skoglund further introduces, Papuan ancestry may have become ubiquitous in Remote Oceanians approximately 50-80 generations ago, or 1,500 – 2,300 BP assuming 28.1 years per generation. Using qpGraph to explore admixture graphs, certain Polynesian groups were modelled as resulting from the First Remote Oceanians and Papuan populations related to Highland New Guineans (Skoglund et al., 2016).

As Lipson et al. (2018), further suggests, people of almost entirely Papuan ancestry arrived in Vanuatu (a Melanesian island state in Oceania, approximately 2,350 km distant from Australia) by around 2300 BP. Papuan ancestry was subsequently diluted through admixture but remains at least 80%-90% in most islands. Despite these massive demographical changes, Papuan languages did not replace Austronesian languages (Posth et al, 2018). As Posth et al. admits, the almost complete replacement of a population's genetic ancestry that leaves the original languages in situ is extremely rare—possibly without precedent in human history. An undifferentiated proto-Oceanic operating as a lingua franca for linguistically diverse Papuan migrant groups could explain the continuity of Oceanic languages in the face of Papuan expansion. A pidgin language based on English plays the same role in PNG today.

Having genotyped 381 individuals from 85 language groups across PNG at 1.7 million genome-wide markers, a large study was devoted to the genetic history of PNG (Bergstörm et al., 2017). Papuans diverged genetically from Aboriginal Australians, long before rising sea levels separated New Guinea and Australia ~8 ka. The strongest genetic separation within PNG appears to be that between the mainland and the Bismarck archipelago islands (New Britain and New Ireland). Populations located on the island of New Guinea (both highlanders and lowlanders), share uniform relationship to Aboriginal Australians (Bergstörm et al., 2017).

Two genetic clusters were revealed within the highlanders, western and eastern. Speakers of the Angan language family seem to represent a mixture of these two highlander clusters. All highlanders form a clade within a wider diversity of lowlanders and lack East Asian admixture, with the exception of some outliers that display lowland affinities, suggesting that they reflect very recent migration into the highlands. No mitochondrial genomes or Y chromosomes of non-Sahul origin were also found in any highlander (Bergstörm et al., 2017).

Surprisingly from the linguistic point of view, highlanders as a group display slightly higher affinity to groups from the Sepik River region rather than to other lowlanders (the local Sepik-Ramu languages are unrelated to the Trans–New Guinea languages of the highlands). However, archaeology suggests Holocene cultural contact between these two regions (Swadling et al., 2008). Sepik lowlanders separated from highlanders between 10 and 20 kya, whereas all splits within the highlands seem to have happened within the last 10 kya (Bergstörm et al., 2017).

The lowlanders harbour widespread Southeast Asian ancestry, with substantially higher levels in Austronesian speakers than in non-Austronesian speakers (Bergstörm et al., 2017). Compared to regions of similar size in Eurasia, genetic differentiation in PNG is much stronger, which reflects long-lasting *in situ* diversification without major population replacement sweeps. Y chromosome haplogroups reveal more pronounced population structure, suggesting lower male effective population size and/or more active female movement between groups.

A Philippine non-Kankanaey group shows the first evidence of a complex admixture event with a European source around 1710 CE. This is consistent with the arrival of Western explorers in the archipelago starting from the 16th century (Ooi, 2004). The first European contact in Tonga, Samoa, and Tahiti occurred in the mid to late 18th century (Oliver 1974), which coincides with inferred dates for European admixture in Tahiti (1705 CE) and the Tongan–Samoan cluster (1820 CE) (Hudjashov et al., 2017). However, a strong European admixture has never been detected before in Papuan samples. Bergstörm et al. (2017), having included 503 European and 489 South Asian individuals from the 1000 Genomes Project, found no European admixture in Papuans.

3. Aims of the study

The genetic history of present-day individuals includes episodes of mating between divergent groups, which have led to 'introgressed' genetic material persisting in modern genome sequences. Perhaps the most notable examples of such events in humans are the introgressions from Neanderthals into non-Africans about 50,000 years ago, and from a related archaic group known as Denisovans into the ancestors of indigenous people from New Guinea and Australia. Papuan genomes therefore play an important role in studies of the human past. The geographic isolation of PNG is an important factor that influenced population history. A detailed model of Papuan history would be useful not only for historical and evolutionary studies, but also for medical purposes.

No archaeogenetic study of Papuans was published to date. Recent genome-wide studies of present-day Papuans, especially Bergstrom et al. (2017) focused on the PNG mainland, have contributed a lot to understanding Papuan history, however a detailed phylogeny of lowlanders and highlanders was never constructed.

Relying on the admixture graph algorithm (Patterson et al. 2012) and rigorous model-ranking approaches (Flegontov et al. 2019), we plan to build a detailed admixture graph for several major highlander and lowlander lineages genotyped by Bergstrom et al. 2017.

4. Methodology

4.1. Dataset preparation

All tested individuals analysed in this thesis were sampled before as a part of recently published studies: Meyer et al. (2012), Prüfer et al. (2014), Mallick et al. (2016), Mörseburg et al. (2016), Pagani et al. (2016), and Bergström et al. (2017). No extra sampling or genotyping was performed for the purpose of this thesis.

Papuan samples originate mainly from central PNG (Highlands), southern coast (Lowlands), northern coast (East Sepik), and some nearby islands like Manus or Bougainville (Bergström et al., 2017). Papuans were analysed in the context of various populations across the world like Southeast and South Asians, Africans, Siberians, and Europeans (Mallick et al., 2016; Mörseburg et al., 2016; Pagani et al., 2016). Neanderthal (Prüfer et al., 2014) and Denisovan (Meyer et al., 2012) contribution to Papuan genomes was also accounted for in our study.

The final SNP dataset consists of 1,153 individuals sub-divided into 27 meta-populations: African (AFR), North African (AFR_N), Athabaskan-speaking (ATH), Central Asian (CAS), Caucasian (CAU), Chukotko-Kamchatkan-speaking (C-K), Denisovan (Denisovan), Eskimo-Aleut-speaking (E-A), East Siberian (ESIB), European (EUR), European with Indian ancestry (EUR_SAS), Finno-Urgic-speaking (FU), Middle Eastern (ME), Melanesian including Papuan and Australian (MEL), Northern North American (NAM), Northeast Asian (NEA), Neanderthal (Neanderthal), Andamanese negrito (Negrito_IND), negrito from the Philippines (Negrito_PH), Polynesian (POL), Central and South American (SAM), North Indian (SAS_N), South Indian (SAS_S), South Indian with Southeast Asian admixture (SAS_SE), Southeast Asian (SEA), West Siberian (WSIB), while 397 of them are Papuans (sub-divided into 87 groups acording to sampling locations and languages).

4.2. Bioinformatics analyses

4.2.1.PLINK

Using PLINK v.1.9, six previously described SNP-array and sequencing datasets were merged together (Meyer et al., 2012; Prüfer et al., 2014; Mörseburg et al., 2016; Pagani et al., 2016; Mallick et al., 2016; and Bergström et al., 2017). Accounting for the fact that certain sites are missing in some individuals, the dataset was refined by using the --geno option. It filtered out all variants with missing call rates exceeding 5% (cog-genomics.org/plink/1.9).

4.2.2. Principal Components Analysis

EIGENSTRAT v.6.0.1 was used for performing principal components analysis. Input data were pruned for linkage disequilibrium using PLINK 1.9 (using --geno 0.05 setting). Considering the results of PCA analysis (the PC1 vs. PC2 plot), I removed three detected outliers with non-Papuan ancestry. The final PCA analysis was performed on 1.196 individuals and 244.604 SNPs.

4.2.3. qpGraph

I was building admixture graphs using the qpGraph method, and that constitutes the core of my work. This method relies on allele frequencies in populations, and fits admixture graphs to a matrix of f_2-, f_3-, and f_4-statistics for a set of population included into the graph.

Based on previously published scenarios of New Guinea peopling, a basic backbone topology was created to start the analysis (Bergström et al., 2017). qpGraph was run with the following options: outpop, NULL; blgsize, 0.05; lsqmode, NO; diag, 0.0001; hires, YES; initmix, 1000; precision, 0.0001; zthresh, 2.0; terse, NO. Only sites lacking missing data across all groups were used for calculating *f*-statistics (useallsnps, NO).

For next steps of the analysis, it was decided to keep the geographically pre-determined division of Papuans into Highlanders and Lowlanders. Highlanders and Lowlanders were later separated according to their language families and provincial borders into metapopulations: first, I tested all possible topologies including up to 6 consecutive lowlander branches and one highlander (Enga). A group of best topologies was defined according to the worst model residual (i.e. the Z-score of the worst-fitting *f*-statistic). Then a tree of 5 highlander groups was inferred separately and grafted onto the best lowlander tree. The fit of this complex tree to the data was refined by adding intra-lowlander admixture events and by replacing some provincial meta-populations by their constituent populations. I.e. the tree was transformed into an admixture graph. All possible pairs of admixing lowlanders were explored, and in each case non-admixed, two unidirectional and bidirectional admixed models were compared. Below I describe various metrics that should be considered for ranking admixture graph topologies.

4.3. Model fit and model-ranking statistics

Z-score or the worst residual. The Z-score is a difference between fitted and observed f_4 values divided by the standard error interval. And a Z-score of the worst-fitting *f*-statistic is reported by the software as the Z-score of the model. Absolute Z-scores below 3 are usually considered acceptable (Patterson et al. 2012, Lipson and Reich 2017). Ranking alternative topologies by their Z-scores is not an optimal approach since Z-scores of some models tend to be almost identical, but that is the only viable approach for ranking models including different combinations of populations, like all possible combinations of 6 lowlander groups.

Admixture events. When adding a "gene flow", I have always considered both potential directions of the flow, and the bidirectional model too. Admixed unidirectional models were favoured over the unadmixed model if their likelihood ratio exceeded e^4 (Flegontov et al. 2019), and bidirectional models were favoured if the corresponding likelihood ratio exceeded e^5.

Edges. Trifurcations in the tree correspond to 0 genetic drift distance along internal edges of the tree. By "internal" I mean any edges not adjacent to a gene flow (admixture) edge. Since 0-edges resulting from a mis-specified topology are expected to be much more frequent than those resulting from genuine trifurcations (Flegontov et al. 2019), I monitored the number of "internal"/"backbone" 0-length edges in the Papuan clade.

Model likelihood. Likelihood for an admixture graph is calculated and reported by qpGraph as *-ln(likelihood)*, and depends on the sum of all residuals (Z-scores for all possible *f*-statistics for a set of groups) and the covariance matrix. Unlike in the case of the worst residual, model likelihood values are usually not flat across alternative topologies. Thus, model likelihood is a nice metric for ranking alternative topologies. However, likelihood depends on the overall number of sites in the dataset and on the number of model parameters (edges + admixture events), thus, strictly speaking, graphs including different population sets or graphs having different number of parameters cannot be compared according to likelihood (Lipson and Reich 2017). However, in practice it is possible to find a likelihood ratio threshold for comparing not only models having the same populations and the same number of parameters (like alternative branching orders), but also the same populations and an added admixture ("gene flow") event. In the former case we use a likelihood ratio threshold of e^3, and in the latter case of e^4 (Flegontov et al. 2019, Lipson and Reich 2017). I note that this is work in progress and further testing of the method on simulated genetic data is needed to find optimal model-ranking metrics and thresholds.

5. Results and discussion

The map (**Fig. 1**.) of the eastern part of the New Guinea Island shows PNG provincial boundaries as they are used in this study.

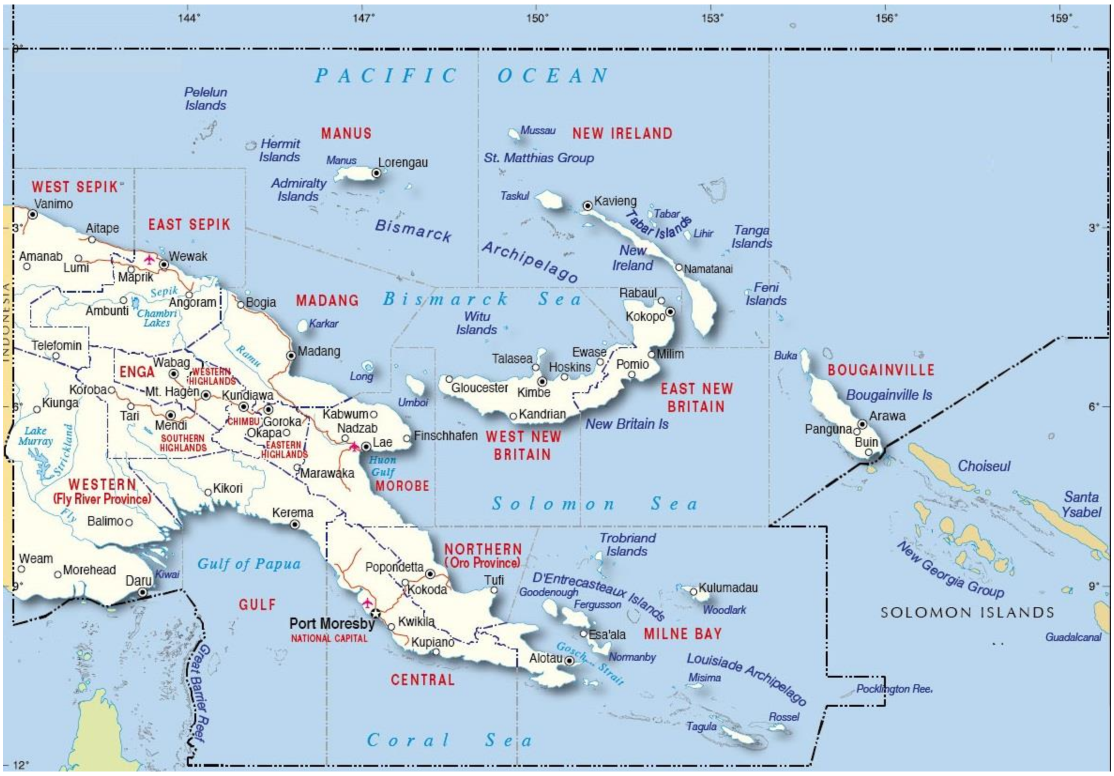

Fig. 1. ***A map of Papua New Guinea***. *Five Highlander groups (West Highlanders, Simbu, East Highlanders, South Highlanders, Enga) and 11 Lowlander groups (Madang, Gulf, Morobe, Milne Bay, East Sepik, East New Britain, Northern, Central, Western, Manus, New Ireland) were originally defined according to provincial borders (available from: nationsonline.org/oneworld/map/papua_map2.htm. ONLINE [11-12-19].*

According to a proclamation by Bergström et al. (2017), every attempt has been made, both during sampling and selection of samples for genotyping, to keep geographical records of samples and ensure that the results reflect the pre-colonial population structure as much as possible. Having assembled the final dataset, two different PCA analyses were run to summarise the variance of studied individuals. The first plot (**Fig. 2**) displays present-day Europeans (EUR), East Asians (EAS), Southeast Asians (SAE), South Asians (SAS), Negritos, Near Oceanians including Papuans and Australians, and Remote Oceanians/Polynesians (POL). The second PCA plot focuses on the variance of present-day Papuan populations and present-day Southeast Asians (SEA) (**Fig. 3**) showing HLs having no SEA admixture, whereas almost all LLs have at least some SEA admixture. This result is in line with previous findings by Bergström et al. (2017).

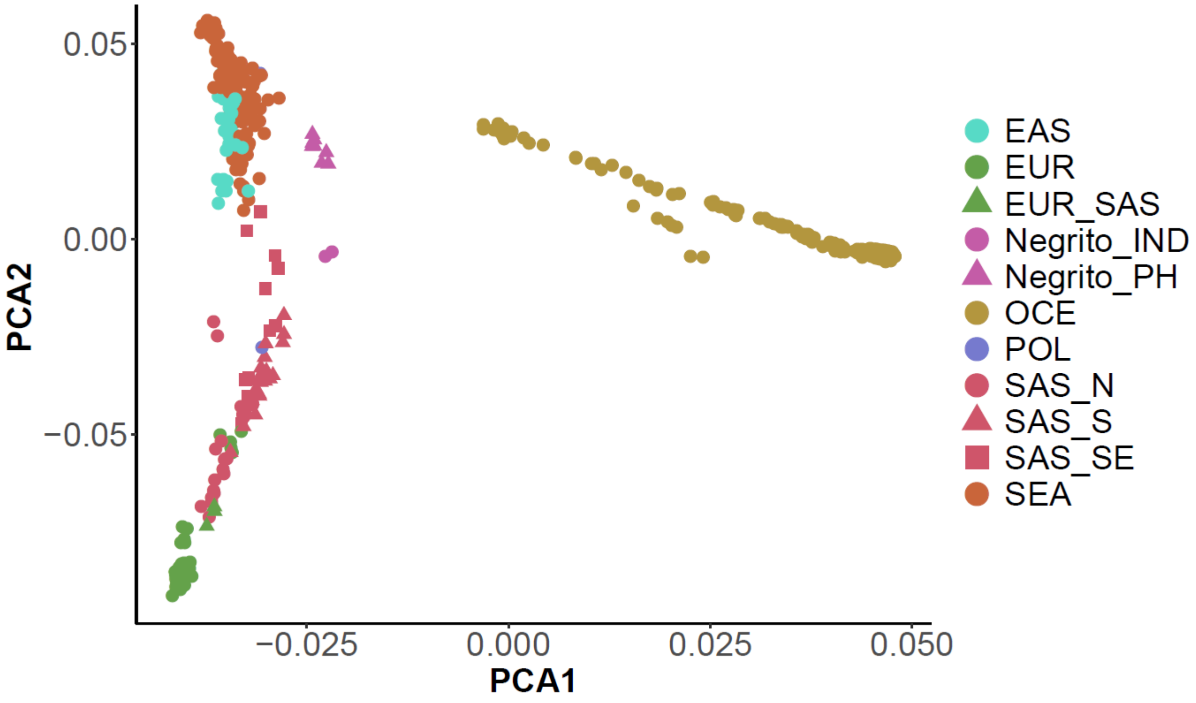

Fig. 2. ***PCA plot of Eurasians and Oceanians.*** *The following meta-populations most relevant for the aim of this study are plotted: present-day East Asians (EAS), Europeans (EUR), Europeans with Indian ancestry (EUR_SAS), Andamanese Negrito (Negrito_IND), Negrito from the Philippines (Negrito_PH), Near Oceanians (OCE), Remote Oceanians or Polynesians (POL), North Indians (SAS_N),), South Indians (SAS_S),), Indians (South Asians) of nuclear affiliation (SAS), Indians with Southeast Asian admixture (SAS_SE), and Southeast Asians (SEA). Plots of two principal components (PC1 vs. PC2) are shown.*

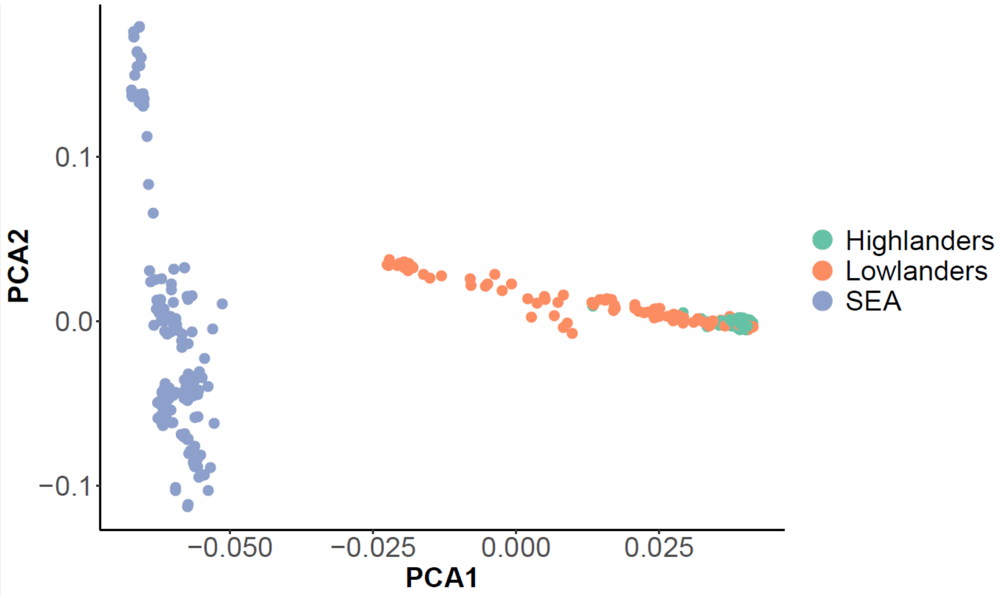

Fig. 3. ***PCA plot of Papuans and Southeast Asians.*** *The following groups most relevant for the aim of this study are plotted: present-day Southeast Asians (SEA), Papuan LLs and HLs. This plot illustrates a cline composed of LLs and formed by widespread SEA admixture among LLs. Plots of two principal components (PC1 vs. PC2) are shown.*

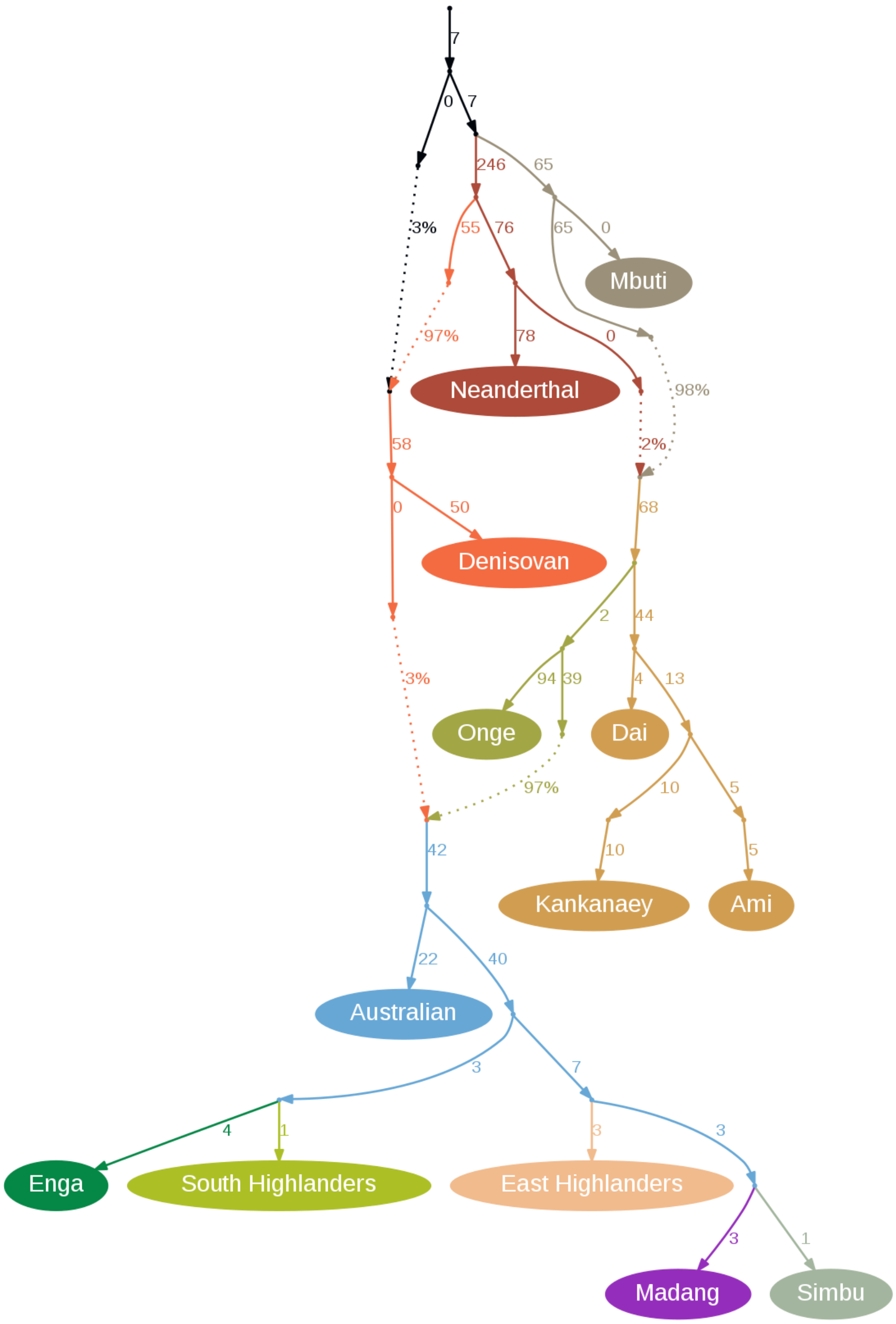

Fig. 4. ***The HL clade***. *The best (with the lowest Z-score), but not fitting (Z-score = 3.64) tree of HLs includes five unadmixed HL meta-populations (Enga, South Highlanders, East Highlanders, Madang and Simbu). No LLs are included here.*

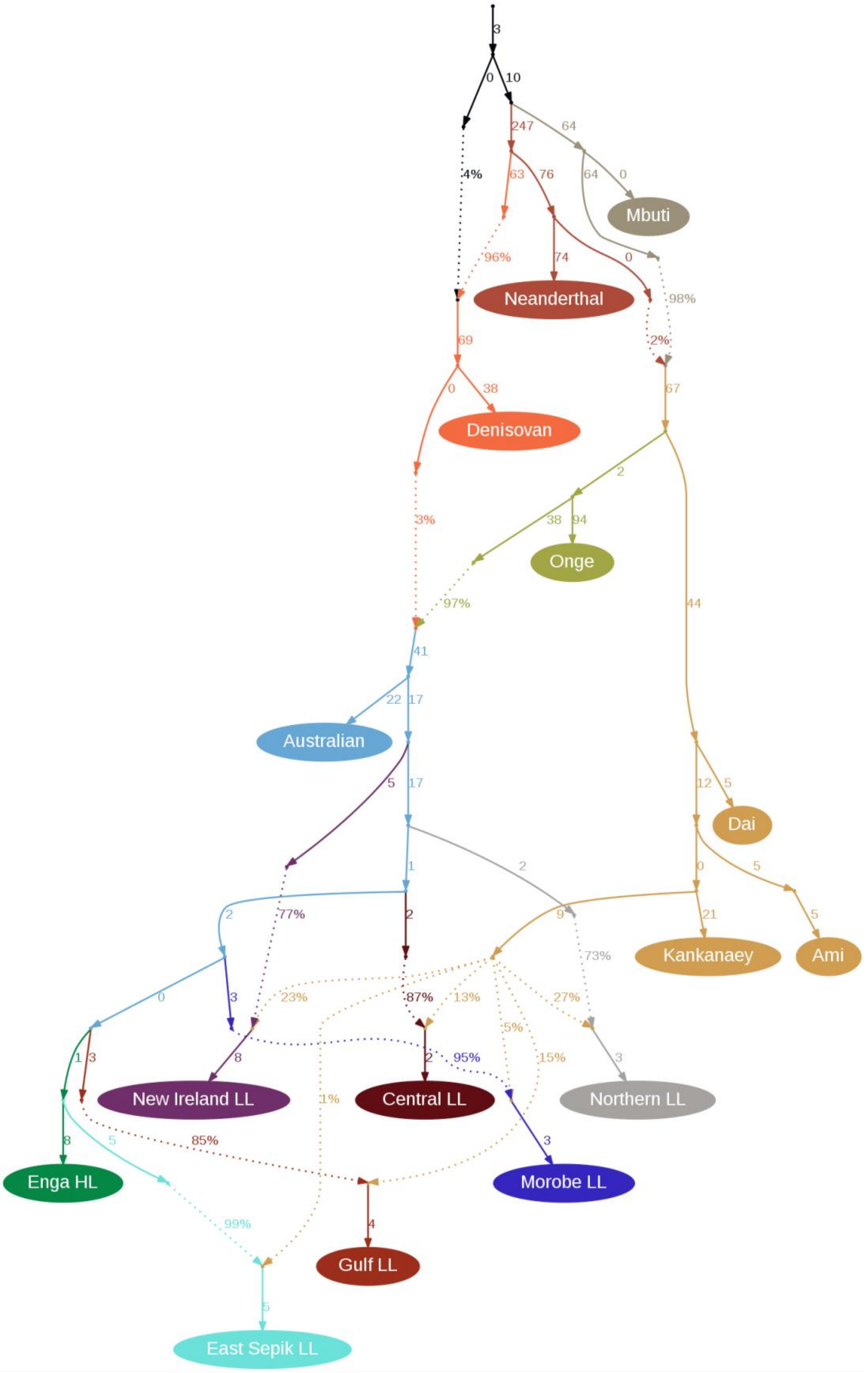

Fig. 5. ***The LL clade.*** *The best (with the lowest Z-score; Table 2) but not fitting (Z-score = 4.71) tree of LLs includes six unadmixed LL meta-populations (Table 1) (East Sepik, New Ireland, Gulf, Central, Morobe, and Northern LLs). One unadmixed HL population (Enga) is included here.*

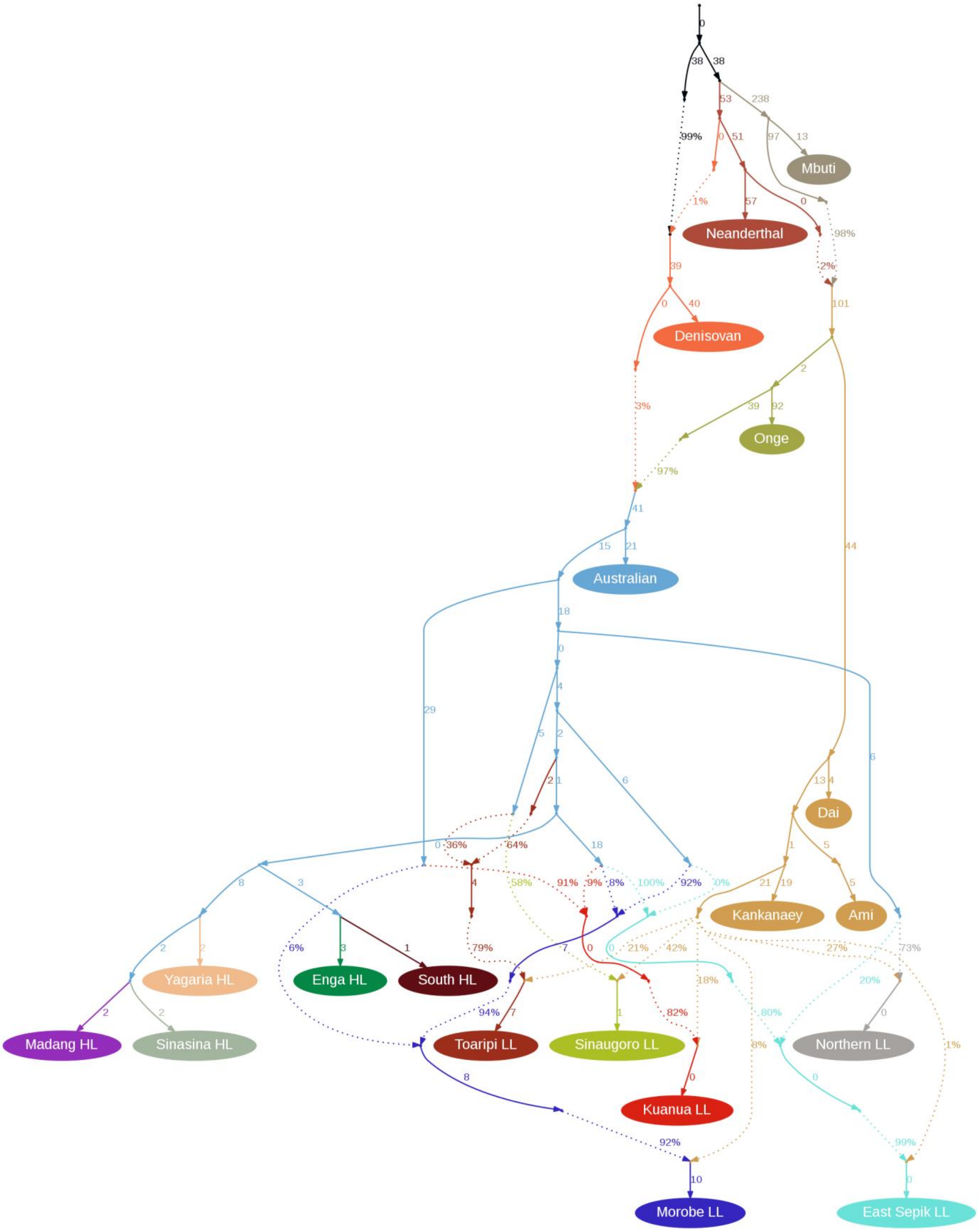

Fig. 6. *The final topology of Papuan clade obtained after optimizing meta-population composition and adding five intra-lowlander admixture events. The worst Z-score = 2.969. The model presented here includes 5 unadmixed HL and 6 admixed LL groups.*

In previous studies (Bergström et al., 2017; Skoglund et al., 2016; Lipson et al., 2018; Posth et al. 2018) just very simple admixture graphs for Papuans were presented I aimed to build a more complex model including several Lowlander and Highlander branches and gene flow events. The process of admixture graph building was for simplicity split into three main phases: 1) searching for the best-fitting tree of Highlanders (HLs), 2) searching for the best-fitting tree of Lowlanders (LLs), 3) a tree incorporating both HLs and LLs, 4) a graph incorporating both HLs and LLs and including several admixture events. This procedure has explored just a small fraction of possible graph topologies, but exhaustively exploring all topologies is unfeasible for a graph of this complexity.

Six TNG-speaking Highlander groups (East Highlanders, South Highlanders, West Highlanders, Enga, Madang, Simbu); and 12 Lowlander groups (Central AN-speaking, Central TNG-speaking, East Sepik Sepik-Ramu-speaking, Gulf TNG, Madang SR, Madang TNG, Morobe AN, Morobe TNG, New Britain AN, New Ireland AN, Northern TNG, Western TNG) consisting of at least two individuals were defined initially according to provincial borders and their language families (Austronesian, AN; Trans-New Guinea, TNG; Sepik-Ramu, SR).

Here is a list of ethnic groups/languages sampled in each province: Madang TNG (Gende, Takia, Amele, Bemal, Nobonob, Wagi, Amaimon, Pondoma, Aruamu, Giri, Kominimung, Rao, Sop); Western Highlands (Maring, Wahgi, Kyaka, Melpa, UmbuUngu, BoUng, Nii), Madang (Biyom), Simbu (Kuman, Sinasina, Chuavve,Golin, Dadibi), Eastern Highlands (Siane, Tokano, Alekano, Yawiuha, Yagaria, Benabena, Kamano, Awiyaana, Yipma, Simbari), Gulf (Akoye, Orokolo KeoruAhia, Toaripi, Tairuma, Ikobi, Purari, other), Southern Highlands (Huli, Foi, Erave, Wiru, Imbongu, East Kewa Aliya, West_Kewa, Heneng, Angal, Samberigi), Enga, Morobe (Yabem, Kate, Nabak, Bugawac), Milne Bay (Umanakaina, Dobu, other), East Sepik (Boikin_Wallis, Boikin, Angoram, Ambulas, Kairiru), East New Britain (Kuanua), Northern (Korafe Yegha), Manus (TuluBohuai), New Ireland (Kara, Patpatar), Central (Motu, Sinaugoro, Keapara, Hula, Grass_Koiari, Mountain_Koiali, Fuyug, Tauade, Waima), Western (Southern Kiwai, other).

The best but not fitting (Z-score = 3.64), tree of Highlanders presented here (**Table 2, Fig. 4**) includes five unadmixed HL meta-populations (Enga, South Highlanders, East Highlanders excluding Angan speakers, Madang, and Simbu). This tree includes no intra-highlander admixture event or any gene flow from Southeast Asia. Adding those admixture events or SEA gene flows did not improve the model metrics (worst residual and likelihood, resluts not shown).

There was only one topology significantly different from the other ones according to the likelihood and worst residual metrics, and that revealed two main genetic clusters within the HLs: the western branch including South HLs and Enga, and its sister eastern branch (East HLs, Simbu and Madang). This subdivision of HLs into two main clusters supports the results previously published Bergström et al. (2017).

I was not able to incorporate West HLs into the model as an unadmixed group (the worst residuals were always higher than 3 SE intervals), and they likely represent a mixture of the western and eastern HL clades in line with the original study (Bergstrom et al. 2017).

More deeply in time, this modelling also confirmed that modern Australians and Papuans descended from an admixture event between an ancestral population related to the Andamanese people (Onge) and ancient Denisovans. According to my model, there is 3% of Denisovan ancestry in present-day Australians and Papuans.

Having solved the basic population structure of HLs, I started building models for Lowlander populations, which resulted in the best (displaying the lowest Z-score) but not fitting topology (Z-score: 4.71) presented here (**Fig. 5**). I tested 840 trees with 4 successive Lowlander branches and Enga as a representative Highlander group, 2,520 trees with 5 successive LL branches, and aligned the branching order in few best trees sorted according to the worst residual. If the grouping by language affiliation (Austronesian vs. TNG) and the distinction between two islands of the Bismarck Archipelago (New Britain and New Ireland) are dropped, all these best trees converge on one topology composed of 6 LL branches. Thus, I am confident that this is the best topology in the class of topologies having successive branches (A, (B, (C, (D, …). The branching order is as follows: Australians, Papuans from the Bismarck Archipelago, Northern LLs, Central LLs, Morobe LLs, Gulf LLs, East Sepik LLs, HLs.

To account for widespread Southeast Asian admixture among LLs (Fig. 3 and Bergstrom et al. 2017), I added gene flows from Southeast Asians to each LL group by default. I found that the best fit is observed when this source is on the Kankanaey branch (an Austronesian-speaking group from the Philippines). For simplicity, all SEA gene flows were derived from the same Kankanaey-related source. According to my best tree (**Fig. 6**), populations from the island of New Ireland split first from the Papuan clade, and have a lot of SEA ancestry (23%). The second and third splits in the Papuan clade are formed by the Northern and Central LLs. Morobe LLs split next, and display the second lowest fraction of SEA ancestry (5%) after East Sepik LLs who have just 1%.

At the next step I attempted to merge the HL and LL trees and improve the fit of the combined tree by reducing selected meta-populations to populations and by adding admixture events in a systematic fashion (Z-score = 2.97 was significantly improved in the end). Two HL and three LL metapopulations were further divided into populations (based on sample description available). The best-fitting model presented here (**Fig. 6**) includes the following replacements: East HLs → Yagaria, Simbu → Sinasina (2 HLs) and New Ireland/New Britain → Kuanua, Gulf → Toaripi, Central LLs → Sinaugoro (3 LLs).

Tab. I. Gene flow testing and final statistic in combined HL/LL tree.

admixture partners	**Gene flow**	**f4 statistic**				**Std. err.**	**Z-Score**
Northern=>E. Sepik	Unidirect.	SHi	ESe	Kua	Toa	0.000456	4.715
Morobe<=>E. Sepik	Bidirect.	Aus	Kua	SHi	ESe	0.000523	3.883
Kuanua=>Morobe	Unidirect.	SHi	ESe	Nor	Toa	0.000364	4.740
Sinaugoro=>Toaripi	Unidirect.	SHi	ESe	Nor	Toa	0.000364	4.764
E. Sepik=>Kuanua	Unidirect.	Den	Mor	Ami	Kan	0.001155	2.969

The first ancestral modern human populations arrived in the Island Southeast Asia more than 40,000 YBP, and contributed to the ancestry of both indigenous Australians and Papuans, and hence to other Pacific islanders (Wollstein et al, 2010). Being the earliest-branching member of the Asian clade (Lipson et al. 2018), Papuans are often used as an important outgroup in studies of genetic history. Occupying islands that virtually connect South Asia and Australia, Papuan genomes have been used while searching signals of admixture from outside the East Asian clade (Lipson et al., 2018). As Lipson et al. (2018) further revealed, the western Indonesians could be modelled well with three (but not two) sources of ancestry: Austronesian-related, Austroasiatic-related, and Papuan-related in respective proportions of ~67%, 29%, and 4%. Thus, Papuan gene flow extends westwards as far as western Indonesia. Being poorly studied in detail, Papuans were often considered as a single entity in genomic studies.

Matisoo-Smith et al. argue that Austronesian (Lapita culture predecessors) came to Papua ca. 3000 YBP. Admixed with Papuans, this population has settled all of the Remote Oceania: Polynesia and Micronesia. However Skoglund et al. (2016), Lipson et al. (2018) and Posth et al. (2018) found, based on few ancient samples sequenced from remote Oceania, that the earliest inhabitants of those islands had nearly 0% Papuan ancestry. Shortly afterwards populations who were nearly 100% Papuan came, the proportions of SEA and Papuan ancestries settled down to modern levels over time (Lipson et al., 2018). However, all Oceanians have at least 25% Papuan ancestry today. Therefore, Lipson suggests that the ultimate source of this Papuan push into remote Oceania is in the N. Britain and N. Ireland.

6. Conclusions

The first aim of this thesis (to generate a detailed phylogeny of Papuans) was fulfilled by testing thousands of topologies. Bergstrom et al. concluded that HLs form a “tight” clade, that LLs are more diverse and differentially related to LLs. These results were based on interpretation of simple *f*-statistics. Our graph topology is in perfect agreement with these results, but much more detailed: (New Britain/New Ireland, (Northern LL, (Central LL, (Morobe LL, (Gulf LL, (Sepik-Ramu-speaking East Sepik and Madang LL, ((Enga HL, South HL), (East HL, (Madang HL, Simbu HL)))))))). However, Lowlanders of the Western and Madang provinces (TNG-speaking) and West Highlanders cannot be fitted onto the graph as unadmixed groups. To keep the graph relatively simple, I have refrained from adding those groups onto the final model which already includes intra-lowlander admixture events

The subdivision of Highlanders into two major clades (Bergstrom et al. 2017) was also confirmed by my analysis. Thus, all of Bergstrom’s results were confirmed, but a much more detailed graph was constructed for both highlanders and lowlanders.

6. List of acronyms

ABO	ABO-antigen-presence-based blood group system
AFS	Fllele Frequency Spectrum
AMH	Anatomically modern humands
BNG	British New Guinea
dbSNP	The Single Nucleotide Polymorphism Database
HapMap	The International HapMap Project
HEVR	Human endogenous retroviruse
HL	Highlanders
HLA-B	Human Leukocyte Antigen B
HUGO	The Human Genome Project
HVS	a hyper variable segments
ka	thousand years ago
LL	Lowlanders
LINE	Long interspersed nuclear elements
MAF	a minor allele frequency
mtDNA	Mitochondrial deoxyribonucleic acid
MVR-PCR	Ministatellite Variants Repeat - Polymerase Chain Reaction
MDS	Multidimensional scaling
MN	MN –antigen-presence-based blood group system
MYA	million years ago
NGS	Next Generation Sequencing
Rh	Rh-factor-presence-based blood group system
PAR	a pseudoautosomal region
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PNG	Papua New Guinea
rRNA	Ribosomal ribonucleic acid
SNP	a single nucleotide polymorphism
STR	a short tandem repetition
TNG	Trans New Guinea
tRNA	Transfer ribonucleic acid
UMAP	Uniform Manifold Approximation and Projection
VNTR	a variable number of tandem repeats

YBP	years before present

World populations:

AFR	African
AFR_N	North-African
ATH	Athabaskan-speaking
CAS	Central-Asian
CAU	Caucasian
C-K	Chukotko-Kamchatkan-speaking
Denisovan	Denisovan
Eskimo-Aleut-speaking	E-A
ESIB	East-Siberian
EUR	European
EUR_SAS	European-with-Indian-ancestry
FU	Finno-Urgic-speaking
ME	Middle-Eastern
MEL	Melanesian-incl.-Papuan-and-Australian
NAM	Northern-North-American
NEA	Northeast-Asian
Neanderthal	Neanderthal
Negrito_IND	Andamanese-Negrito
Negrito_PH	Negrito-from-the-Phillippines
POL	Polynesian
American	Central-and-South
SAS	Indian-loc.-in-South-Asian
SAS_N	North-Indian
SAS_S	South-Indian
SAS_SE	South-Indian-with-Southeast-Asian-admixture
SEA	Southeast-Asian
WSIB.	West-Siberian

7. References

Alexander, DH., Novembre, J., and Lange, K. 2009. *Fast model-based estimation of ancestry in unrelated individuals. Genome Research* (19): 1655–1664.

Allen, J. (2001). *Australia and New Guinea, Archaeology of. International* Encyclopedia of the Social & Behavioral Sciences, 952–956.

Aghakhanian, F., Yunus, Y., Naidu, R., Jinam, T., et al. 2015. *Unravelling the Genetic History of Negritos and Indigenous Populations of Southeast Asia.* Genome Biology and Evolution, 7(5), 1206–1215.

Anderson, S. 1981. *Shotgun DNA sequencing using cloned DNase I-generated fragments.* Nucleic Acids Res. (9): 3015 - 3027.

Bartlett, JMS. and Stirling, D. 2003. *A Short History of the Polymerase Chain Reaction*. PCR Protocols. Methods in Molecular Biology. 226 (2nd ed.): 3–6.

BBC. 2018. *Papua New Guinea Profile – Timeline*. ONLINE[09-12-2019] available from: www.bbc.com/news/world-asia-15593238.

Becht, E., McInnes, L., Healy, J. et al. 2019. *Dimensionality reduction for visualizing single-cell data using UMAP*. Nat Biotechnol (37): 38–44.

Becroft, TC. 1967. *Child-rearing practices in the Highlands of New Guinea*. Medical Journal of Australia 2:810-813.

Bergsland, K. and Vogt, H. 1962. *On the validity of glottochronology*. Curr. Anthropol.(3): 115–153.

Bergström, A., Oppenheimer, SJ., Mentzer, AJ., et al. 2017. *A Neolithic expansion, but strong genetic structure, in the independent history of New Guinea.* Science 357 (6356): 1160-1163.

Blust, R. 2008. *Remote Melanesia: one history or two? An addendum to Donohue and Denham*. Oceanic Linguistics (47) 445–459.

Bodmer, W. 2015. *Genetic Characterization of Human Populations: From ABO to a Genetic* Map of the British People. Genetics, 199(2), 267–279.

Bonné-Tamir, B., Johnson, MJ., Natali, A. et al. 1986. *Human mitochondrial DNA types in two Israeli populations. A comparative study at the DNA level*. Am J Hum Genet (38): 341–351

Bouckaert, R., Lemey, P., Dunn, M. et al. 2012. *Mapping the Origins and Expansion of the Indo-European Language Family.* Science, 337(6097): 957–960.

Bourke, RM. et Harwood, T. 2009. *Food and Agriculture in Papua New Guinea.* Australian National Univ. Press. 638p.

Bray, M., Smith, P. 1985. *Education and Social Stratification in Papua New Guinea.* Melbourne : Longman Cheshire. 225p.

Bulmer, RNH. 1969. *Cultural diversitv and national unitv: Past and future contexts for anthropology and sociology in New Guinea*. Inaugural lecture: University of Papua New Guinea.

Castro, JA., Picornell, A., Ramon, M. 1998. *Mitochondrial DNA: a tool for populational genetics studies.* Internatl Microbiol (1): 327–332.

Campbell, L. and Poser, WJ. 2008 – 1st. *Language classification: History and method*. Camb Uni Press. 548p. ISBN: 978-0521880053.

Chang, W., Cathcart, Ch., Hall, D. et al. 2015. *Ancestry-constrained phylogeneticanalysis supports the indo-european steppe hypothesis*. Language, 91 (1): 194 – 244.

Canizalez-Román, A., Campos-Romero, A., Castro-Sánchez, JA. et al. 2018. *Blood Groups Distribution and Gene Diversity of the ABO and Rh (D) Loci in the Mexican Population.* BioMed Research International: 1–11.

Cavalli-Sforza, LL., and Edwards, AWF. 1965. Analysis of human evolution. *Proceedings of the 11th International Congress of Genetics,* The Hague, 1963. Genetics Today 3: 923–933.

Clark, PU., Dyke, AS., Shakun, JD. et al. 2009. *The Last Glacial Maximum*. Science 325(5941): 710–714.

Dal, E. et Alkan, C. 2018. *Evaluation of genome scaffolding tools using pooled clone sequencing*. Turk J Biol (42): 471-476.

Dawson, E., Abecasis, GR., Bumpstead, S., et al. 2002. *A first-generation linkage disequilibrium map of human chromosome 22.* Nature, 418(6897), 544–548.

Demeter, F. et al. 2017. *Early Modern Humans from Tam Pà Ling, Laos*. Curr. Anthropol. 58 (17) 527 – 538.

Denaro, M., Blanc, H., Johnson, MJ. et al. 1981. *Ethnic variation in HpaI endonuclease patterns of human mitochodrial DNA*. Proc Natl Acad Sci USA (78): 5768–5772.

Denham, T. 2005. *Envisaging early agriculture in the Highlands of New Guinea: landscapes, plants and practices.* World Archaeology, 37(2), 290–306.

Department Japan International Cooperation Agency (JICA). 2002. *Country Profile on Environment – PNG.* Report. Planning and Evaluation Department.

Diamond, J. 1997. *Guns, Germs and Steel: the Fates of Human Societies*. W.W. Norton and Company. 494p.

Diaz-Papkovich, A., Anderson-Trocmé, L., Ben-Eghan, et al. 2019. *UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts.* PLOS Genetics, (15): 11.

Duggan, AT., Evans, B., Friedlaender, FR., et al. 2014. *Maternal History of Oceania from Complete mtDNA Genomes: Contrasting Ancient Diversity with Recent Homogenization Due to the Austronesian Expansion.* The American Journal of Human Genetics, 94(5): 721–733.

Dunn, M., Greenhill, SJ., Levinson, SC. Et al. 2011. *Evolved structure of language shows lineage-specifictrends in word-order universals*. Nature (473): 79-82.

Dunning, AM., Durocher, F., Healey, CS., et al. 2000. *The Extent of Linkage Disequilibrium in Four Populations with Distinct Demographic Histories.* The American Journal of Human Genetics, 67(6), 1544–1554.

Edwards, AWF. and Cavalli-Sforza, LL. 1963. *The reconstruction of evolution*. Heredity (18): 553.

Europe PMC Funders Group. 2007. *A second generation human haplotype map of over 3.1 million SNPs.* Nature 18; 449(7164): 851–861.

Flegontov, P., Altınışık, NE., Changmai, P. et al. 2019. *Palaeo-Eskimo genetic ancestry and the peopling of Chukotka and North America*. Nature (570): 236–240.

Foley, B. 2003. *More one Papuan Languages*.[online].DELP International Encyclopedia of Linguistics.[28-10-2019].Available:
https://sydney.edu.au/arts/research_projects/delp/papuan.php

Friedlaender, JS., Friedlaender, FR., Reed, FA *et al*. 2008. The genetic structure of Pacific Islanders. PLoS Genet. 4(1): e19.

Gibbs, RA., Belmont, JW., Hardenbol, P. et al. 2003 *The International HapMap Project.* Nature, 426(6968), 789–796.

Gilliam, A. 1988. *Anthropology, Geopolitics, and Papua New Guinea.* Central Issues in Anthropology, 8(1): 37–51.

Gray, RD. and Atkinson, QD. 2003. *Language-tree divergence times support the Anatolian theory of Indo-European origin*. Nature (426) 435–439.

Green, RE., Krause, J., Briggs, AW. et al. 2010. *A Draft Sequence of the Neandertal Genome.* Science, 328(5979), 710–722.

Gunderson, KL., Kuhn, KM., Steemers, FJ., et al. 2006. *Whole-genome genotyping of haplotype tag single nucleotide polymorphisms.* Pharmacogenomics, 7(4): 641–648.

Halushka, MK, Fan, JB., Bentley, K., et al. 1999. *Patterns of single-nucleotide polymorphisms in candidate genes for bloodpressure homeostasis*. Nature Genet. (22), 239–247.

Harvey, RG. 1974. *An anthropometric survey of growth and physique of the populations of Karkar Island and Lufa subdistrict, New Guinea*. Philosophical Transactions of the Royal Society of London, Series B. (268): 279-292.

Hellenthal, G., Busby, GBJ., Band, G. et al. 2014. *A genetic Atlas of Human Admixture History*. Since (343): 747-751.

Heywood, PF. 1983. *Growth and nutrition in Papua New Guinea*. Journal of Human Evolution, 12(1), 133–143.

Hong GF. 1981. *A method for sequencing single-stranded cloned DNA in both directions*. Biosci. Rep. (1): 243–252.

Hope, G. and Haberle, S. 2005. *The history of the human landscapes of New Guinea* [in Pawley, A., Attenboroug, R., Golson, J. et al. 2005. *Papuan pasts: cultural, linguistic and biological histories of Papuan-speaking peoples*. Canberra: Pacific Linguistics. 817p.: 541-554.

Hudjashov, G., Karafet, TM., Lawson, DJ., et al. 2017. *Complex Patterns of Admixture across the Indonesian Archipelago.* Molecular Biology and Evolution, 34(10), 2439–2452.

Jacobs, G. S. et al. 2019. *Multiple Deeply Divergent Denisovan Ancestries in Papuans,* Cell (177) 1–12.

Jobling, MA., Hollox, E., Hurles M., Kivisild, T. et Tyler-Smith CH. 2004 – 2nd. *Human evolutionary genetics*. Garland Science, Taylor & Francis Group, NY.

Kassian, AS., Zhivlov, M., Starostin, G. et al. 2019. *Rapid radiation of the Inner Indo-European language approach to Indo-European lexicostatistics* [pre-print accepted in Diachronica 2020].

Kayser, M., Brauer, S., Weiss, G. et al. 2000. *Melanesian origin of Polynesian Y chromosomes*. Curr. Biol. 10 (20): 1237 – 46.

Kayser, M. 2010. *The human genetic history of Oceania: near and remote views of dispersal*. Curr. Biol. (20): 194 – 201.

Kawai, Y., Mimori, T., Kojima, K. et al. 2015. *Japonica array: improved genotype imputation by designing a population-specific SNP array with 1070 Japanese individuals*. J Hum Genet. 60(10): 581–7.

Kituai, A. 1974. *Historical narratives of the Bundi people*. Oral History 2(8): 8-16.

Kofler, R., Schlötterer, C., Luschützky, E. et al. 2008. *Survey of microsatellite clustering in eight fully sequenced species sheds light on the origin of compound microsatellites*. BMC Genomics (9): 612.

Lahr, M. M. et Foley, R. 2005. *Multiple dispersals and modern human origins*. Evol. Anthropol. Issues News Rev. (3) 48–60.

Landsteiner. K. 1900. *Zur Kenntnis der antifermentativen, lytisichen und agglutinierenden Wirkungen des Blutserums und der Lymphe*. Zbl. Bakt. I. Abt. (27) 357–362.

Leppälä, K., Nielsen, SV., Mailund T. 2017. *admixturegraph: an R package for admixture graph manipulation and fitting*. Bioinformatics., 33(11):1738–1740.

Lipson, M., Cheronet, O., Mallick, S., Rohland, N., et al. 2018. *Ancient genomes document multiple waves of migration in Southeast Asian prehistory.* Science, 361(6397): 92–95.

Littlewood, RA. 1972. *Anthropological Studies of the Eastern Highlands of New Guinea.* Physical Anthropology of the Eastern Highlands of New Guinea (H). Seattle: University of Washington Press.

Ma, X., Shao, Y., Tian, L. et al. 2019. *Analysis of error profiles in deep next-generation sequencing data*. Genome Biol (20): 50.

Malaspinas, A.S. et al. 2016. *A genomic history of Aboriginal Australia*. Nature (538) 207 – 214.

Malcolm, L. 1970. *Growth and Development of the Bundi Child of the New Guinea Highlands*. Human Bio 42(2):293-328.

Mallick, S., Li, H., Lipson, M., Mathieson, I., et al. 2016. *The Simons Genome Diversity Project: 300 genomes from 142 diverse populations.* Nature, 538(7624): 201–206.
Matisoo-Smith, E. (2015). *Ancient DNA and the human settlement of the Pacific: A review.* Journal of Human Evolution (79): 93–104.

Martin, SH., Davey, JW., et Jiggins, CD. 2014. *Evaluating the Use of ABBA–BABA Statistics to Locate Introgressed Loci.* Molecular Biology and Evolution, 32(1), 244–257.

Maxam, AM. And Gilbert, W. 1977. *A new method for sequencing DNA*. Proc. Natl Acad. Sci. USA (74): 560 - 564.

Mccoll, H., Racimo, F., Vinner, L. Et al. 2018. The prehistoric peopling of Southeast Asia. Science, 361 (6397): 88-92.

McInnes, L. and Healy, J. 2018. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*.

Menozzi, P., Piazza, A., et Cavalli-Sforza, L. 1978. *Synthetic maps of human gene frequencies in Europeans.* Science, 201(4358), 786–792

Merriwether, DA., Clark, AG., Ballinger, SW. et al. 1991. *The structure of human mitochondrial DNA variation.* J Mol Evol (33): 543–555.

Meyer, M., Kircher, M., Gansauge, MT., Li, H. et al. 2012. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Science, 338(6104): 222–226.

Moffatt, A. 2012. *Papua New Guinea. Cultural Profile*. Review. Diversicare West End.25p.

Morauta, L., Chowning, A. Gilliam, AM., et al. 1979. *Indigenous Anthropology in Papua New Guinea*. Current Anthropology, 20 (3): 561-576.

Mörseburg, A., Pagani, L., Ricaut, FX., et al. 2016. *Multi-layered population structure in Island Southeast Asians.* European Journal of Human Genetics, 24(11), 1605–1611.

Mourant, AE., Kopec, A., Domaniewska-Sobczak, K. 1954, 1976. *The Distribution of the Human Blood Groups and Other Polymorphisms*. Oxford University Press, Oxford.

Narokobi, BM. 1974. *Who shall take up Peli's challenge ? A philosophical contribution to the understanding of cargo cults*. Point (1):93-104.

Novembre, J., Johnson, T., Bryc, K. et al. 2008. *Genes mirror geography within Europe.* Nature, 456(7218): 98–101.

O'Connell, JF. and Allen, J. 2015. *The process, biotic impact, and global implications of the human colonization of Sahul about 47,000 years ago*. Journal of Archaeological Science (56): 73–84.

Oliver, DL. 1974. *Ancient Tahitian society.* Honolulu: University Press of Hawaii

Ooi, KG. 2004. *Southeast Asia: A historical encyclopedia, from Angkor Wat to East Timor.*Santa Barbara, Calif.: ABC-CLIO

Pääbo, S. 2003. *The mosaic that is our genome*. Nature (421), 409–412.

Pagani, L., Lawson, DJ., Jagoda, E., et al. 2016. *Genomic analyses inform on migration events during the peopling of Eurasia.* Nature, 538(7624), 238–242.

Palmer, B. 2018. *The Languages and Linguistics of the New Guinea Area*. Walter de Gruyter GmbH.1020p. ISBN 978-3-11-028642-7

Parada-Rojas, CH. And Quesada-Ocampo, LM. 2018. *Analysis of microsatellites from transcriptome sequences of Phytophthora capsici and applications for population studies*. Sci Rep (8): 5194.

Parr, RL., and Martin, LH. 2012. *Mitochondrial and nuclear genomics and the emergence of personalized medicine*. Human Genomics (6)3.

Patterson, N., Price, A., Reich, D. 2006. *Population Structure and Eigenanalysis*. PLoS Genet.(2): e190.

Patterson, N., Moorjani, P., Luo, Y., et al. 2012. *Ancient Admixture in Human History.* Genetics, 192(3): 1065–1093.

Pawley, A., Attenboroug, R., Golson, J. et al. 2005. *Papuan pasts: cultural, linguistic and biological histories of Papuan-speaking peoples*. Canberra: Pacific Linguistics, Research School of Pacific and Asian Studies, Australian National University. 817p. ISBN 0858835622.

Pearson, K. 1901. *On lines and planes of closest fit to systems of points in space*. Philos Mag A(6), 559–572.

Peterson, BK., Weber, JN., Kay, EH., et al. 2012. *Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species.* PLoS ONE, 7(5): e37135.

Petr, M., Vernot, B., Kelso, J. 2019. *admixr—R package for reproducible analyses using ADMIXTOOLS.* Bioinformatics (1-2).

Pritchard, J., Stephens, M., Donnelly, P. 2000. *Inference of population structure using multilocus genotype data*. Genetics(155): 945–959.

Posth, C., Nägele, K. Colleran, H. et al. 2018. *Language continuity despite population replacement in Remote Oceania*. Cell (175): 1185-1197.

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

Rasmussen, M., Guo, X., Wang, Y., Lohmueller, KE. et al. 2011. *An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia.* Science, 334(6052): 94–98.

Reyes-Centeno, H., Hubbe, M., Hanihara, T., et al. 2015. *Testing modern human out-of-Africa dispersal models and implications for modern human origins.* Journal of Human Evolution (87): 95–106.

Reich, DE., Cargill, M., Bolk, S., et al. 2001. *Linkage disequilibrium in the human genome*. Nature (411), 199–204.

Reich, D., Thangaraj, K., Patterson, N., et al. 2009. *Reconstructing Indian population history.* Nature, 461(7263): 489–494.

Riley, ID. 1983. *Population change and distribution in Papua New Guinea: an epidemiological approach.* Journal of Human Evolution, 12(1), 125–132.

Rudd, MK., Wray, GA., et Willard, HF. 2006. *The evolutionary dynamics of α-satellite*. Genome Res. (16) 88-96.

Sanger, F. and Coulson, A. R. 1975. *A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase*. J. Mol. Biol. (94): 441- 448.

Sarei, AH. 1974. *Traditional marriage and the impact of Christianity on the Solos of Buka Island*. New Guinea Research Bulletin (57).

Schiffels, S., Haak, W., Paajanen, P. et al. 2016. *Iron Age and Anglo-Saxon genomes from East England reveal British migration history.* Nature Communications (7): 10408.

Scragg, RFR. 1973. *Menopause and reproductive span in rural Niugini*. Proceedings of the Ninth Annual Symposium of the Medical Society of Papua New Guinea. Port Moresby.

Seeburg, PH., Shine, J. Martial, JA. et al. 1977. *Nucleotide sequence of a human gene coding for a polypeptide hormone*.Trans. Assoc. Am. Physicians (90): 109.
Skoglund, P., Posth, C., Sirak, K. et al., 2016. *Genomic insights into the peopling of the Southwest Pacific*. Nature, 538(7626): 510–513.

Soares, P., Ermini, L., Thomson, N. et al. 2009. *Correcting for purifying selection: an improved human mitochondrial molecular clock*. Am. J. Hum. Genet (84) 740–759.

Soodyall, H. and Jenkins, T. 1993. *Mitochondrial DNA polymorphisms in Negroid populations from Namibia: New light on the origins of the Dama, Herero, and Ambo*. Ann Hum Biol (20): 477–485.

Strathern, AJ. 1973. *Social science research in Papua New Guinea since 1969*. Man in New Guinea 5(2):2-7.

Swadling, P., Wiessner, P., Tumu, A. 2008. *Prehistoric stone artefacts from Enga and the implication of links between the highlands, lowlands and islands for early agriculture in Papua New Guinea*. J. Soc. Ocean. 126–127, 271–292.

Talyaga, A. and Kundapen, K.1975. *Should we revive initiation rites in Enga society?* Gigibori 2(2):37-41.

Tan, G., Opitz, L., Schlapbach, R. et al. 2019. *Long fragments achieve lower base quality in Illumina paired-end sequencing*. Scientific Reports (9): 2856.

Tebbutt, SJ., He, JQ., Burkett, KM. et al. 2004. *Microarray genotyping resource to determine population stratification in genetic association studies of complex disease.* BioTechniques, 37(6): 977–985.
Vergnaud, G., Mariat, D., Apiou, F. et al. 1991. *The use of synthetic tandem repeats to isolate new VNTR loci—cloning of a human hypermutable sequence*. Genomics (11) 135–144.

Vines, AP. 1970. *An Epidemiological Sample Survey of the Highlands, Mainland and Islan& Regions of the Territory of Papua New Guinea.* Port Moresby: Papua New Guinea Department of Public Health.

Waiko, J. 1973. "European-Melanesian contact in Melanesian tradition and literature," in Priorities in Melanesian development. (edit.: May, R.) Canberra: Research School of Pacific Studies, Australian National University; Port Moresby: University of Papua New Guinea: 417-28.
Waiko, J. 1976. *Komge Oro: Land and culture or nothing*. Gigibori 3(1):16-19.

Wollstein, A., Lao, O., Becker, C., et al. 2010. *Demographic History of Oceania Inferred from Genome-wide Data.* Current Biology, 20(22), 1983–1992.

Yamamoto. F., Clausen. H., White T. et al. 1990. *Molecular genetic basis of the histo-blood group ABO system*. Nature (345) 229–233.