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, 2020

Kopicová, K., 2020: Phylogeny of human populations in Papua New Guinea, a genetic and linguistic diversity hotspot. Mgr. Thesis, [in English.] – 64p. 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.

Prohlašuji, že svou diplomovou práci jsem vypracovala samostatně pouze s použitím pramenů a literatury uvedených v seznamu citované literatury.

Prohlašuji, že v souladu s § 47b zákona č. 111/1998 Sb. v platném znění souhlasím se zveřejněním své diplomové práce, a to v nezkrácené podobě elektronickou cestou ve veřejně přístupné části databáze STAG provozované Jihočeskou univerzitou v Českých Budějovicích na jejích internetových stránkách, a to se zachováním mého autorského práva k odevzdanému textu této kvalifikační práce. Souhlasím dále s tím, aby toutéž elektronickou cestou byly v souladu s uvedeným ustanovením zákona č. 111/1998 Sb. zveřejněny posudky školitele a oponentů práce i záznam o průběhu a výsledku obhajoby kvalifikační práce. Rovněž souhlasím s porovnáním textu mé kvalifikační práce s databází kvalifikačních prací Theses.cz provozovanou Národním registrem vysokoškolských kvalifikačních prací a systémem na odhalování plagiátů.

Date, Location: **Signature:**

22-05-2020, České Budějovice ……

Keywords

Admixture graph, blood groups, common ancestry, D-statistic, f-statistic, genomic analyse, haplotypes, Highlanders, historical linguistic, Lowlanders, Papua New Guinea, PCA, phylogeny, sequencing, SNP, variance.

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

1. Preface

Anatomically modern humans (AMH) spread out of Africa earlier than 75 thousand years ago (kya) (Pagani et al., 2017). As cranial fossils of an approximate age 44 – 63 kya from Tam Pà Ling in Laos (Demeter et al., 2017) confirmed, AMHs further expanded into Southeast Asia (SEA), and later towards Indonesia (Oppenheimer, 2009) and America Flegontov et al., 2019; Waters, 2019). This Out-of-Africa theory (Ashraf & Galor, 2013) operates with a concept of migration waves that established human diversity across the world. Migrants were regularly exposed to multiple introgression events with local indigenous populations: at first Neanderthals and Denisovans, and then AMH themselves. Any individual represents a colourful mosaic of past admixture events that can be investigated.

If we are interested in the human history of Island Southeast Asia and Oceania, it is necessary to trace the past back to the time of last glaciation, which lasted roughly from 33 to 19 kya (Clark et al., 2009). Two large landmasses existed at that time: Sunda (consisting of the Malay Peninsula, Sumatra, Java, Bali and Borneo, connected with the Southeast Asian mainland) and Sahul (New Guinea together with Australia and Tasmania) (Riley, 1977). Being separated from the Sunda shelf by sea and strong currents, Sahul was an isolated biogeographical region, which was first recognized by Alfred Russell Wallace in 1859 (Vetter, 2006).

Referred as Wallace's line, this border explains the current geographical distribution of the Asian and Oceanian fauna (including primates, carnivores, elephants, and ungulates) and the marsupial fauna of New Guinea and Australia. Although in the past it was speculated that the water barrier of Wallace's line also prevented *Homo erectus* from reaching Sahul about a million years ago despite reaching the Java Island at the same time (Allen, 2001), later archaeological excavations on the island of Flores (Brown et al., 2004) confirmed the ability of *Homo erectus* to pass Wallace's line. In any case, Wallace's line has shaped the independent human history on the New Guinea Islands.

The country of Papua New Guinea (PNG) occupying the eastern half of the New Guinea Island, the Bismarck Archipelago and the Bougainville Island shows a stable human occupation since ca. 50 kya (O'Connell & Allen, 2015).

The closest relatives of Papuans are Australian Aboriginals (ATSI; Aboriginal and Torres Strait Islander people), with whom they share ancient Denisovan admixture (Rasmussen et al., 2011). Papuans split from ATSI between 25 and 45 kya (Malaspinas et al., 2016).

First PNG settlers started exploring their environment (deforestation) in New Guinea not before 48 kya (Hope & Haberle, 2005). However, the distinctiveness of the Papuan archaeological record raised doubts about this dating. It was hypothesized that the common ancestors of Papuans and ATSI left the African continent even earlier than the ancestors of the other present-day Eurasians (Lahr & Foley, 1994). However, these ideas were refuted by recent genetic studies (Lipson et al., 2018; Malaspinas et al., 2016; Posth et al., 2018; Skoglund et al., 2016). Nowadays it is believed that Papuans and Australians experienced two pulses of archaic human introgression: the Neanderthal pulse shared with the other non-African human groups and the Denisovan pulse that is unique to them (Jacobs et al., 2019).

Papuans share not only their hunter-gathering lifestyle, but also their phenotypic characteristics (small body size, dark skin pigmentation, cranio-facial morphology, and frizzy hair) with so-called negritos from Malaysia, Thailand, India, and the Philippines: the Maniq in Thailand, Andaman islanders (the Onge and Jarawa) in India, and the Aeta in the Philippines (Aghakhanian et al., 2015). Together with indigenous SEA hunter-gatherers of the Hòabìnhian culture (McColl et al., 2018), they all are believed to be the closest relatives of Papuans and Australians on the Asian mainland (McColl et al., 2018).

Papua New Guinea is linguistically the most 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.

1.1. Papua New Guinea: geography and recent history

New Guinea, a tropical island located near the equator, in the southern part of the Pacific Ocean northwest from Australia, is the second-largest (after Greenland) island in the world (Barton, 1965). As a consequence of its colonial and recent post-colonial past, the island 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.

The local monsoonal climate has two seasons (the very hot and humid wet season from December to March, and the dry season, interrupted occasionally by strong rains from May to October) and influences the agricultural strategy in PNG (farming on fields cut in the rainforest) (Moffatt, 2012). PNG has on average 2,000-4,000 mm of rainfall per year across the country (exceptionally even more than 7,000 mm of rainfall), which is ideal for agriculture since tropical crops require annually between 1.500 mm to 3.000 mm of rainfall (Bourke & Harwood, 2009).

As Bourke & Harwood noted (2009), the humidity of land also influences the local human presence: the wetter the area, the lower its population. According a report by the Japan International Cooperation Agency (JICA, 2002), nowadays 97% of the land in PNG is owned and managed by clans, exceptionally by individuals, whose rights are claimed based on heritage along either paternal or maternal lines. The highland regions (altitudes between ca. 1,500 m and 2,100 m) are by far the most densely populated rural areas in the country (ca. 50 persons/km^2, see worldometers.info or population.un.org). In contrast, the southern coastal region of mainland PNG has a very low population density, especially in the Gulf and Western provinces (Bourke & Harwood, 2009). The PNG provinces located on islands are comparable to the mainland lowland regions in population density (Bergström et al., 2017).

PNG experienced a period of complicated colonial settlements: the first European colonies were set up in the southern coastal region in 1883 (Bergström et al., 2017). Since then, the process of colonisation has further continued towards northern lowlands; and very much later also into the highlands; although Europeans did not colonize the highlands before 1930s (Diamond, 1997).

Australian exploring missions revealed the existence of the densely-occupied highlands region just before World War II. Despite that, Australians avoided interactions with locals (e.g. through coffee plantations or trading) until the post-war times (Bergström et al., 2017). The final post-colonial year came in 1975, the year of independence, when PNG also joined the United Nations (Gilliam, 1988).

As mentioned before, properly documented contacts between Europeans and Papuans were lacking until the end of the 19th century. According to Soukup (2010), at that time, the arrival of Europeans was most likely believed as a return of spirits of dead ancestors who has magical powers. From today's perspective, climatic and geographical conditions in PNG were not deemed worthwhile of attention by European travellers or trading companies who were looking for easy and quick profit (Soukup, 2010). This is the reason why serious European colonisation began not earlier than the turn of the 20th century, together with religious missions on the Papuan mainland (Soukup, 2010).

First migrations in PNG have been documented since 1890, when colonial governments used the labour of natives from undeveloped areas for plantations and mines (Bourke & Harwood, 2009). Since 1942, Japanese forces occupied PNG during the Pacific war. As Bourke & Harwood (2009) further mentioned, at that time, many young Papuans joined both the Japanese and Allied forces. The problem of suddenly missing workers was solved by the Highland Labour Scheme that relocated young men from the highlands to the lowlands for two years. Similar attempts continued in the 1970s (for instance in the East Sepik Province) in order to support rubber production (Bourke & Harwood, 2009).

Those written sources reveal just a small fraction of human migrations across Papua, in the very recent past only. Therefore, if we seek for deeper ancestral links we should look at other evidence provided by anthropology and genetics.

1.2. Papua New Guinea's people

In 2019, an annual PNG statistical report announced that at that time, some 87% of Papuans lived in rural areas and their occupation was mostly agriculture; the majority of income came from farming and selling products like coffee, cocoa, palm oil, and copra (*Papua New Guinea Economic Upadte*, 2019). Agriculture originated independently in PNG, more precisely in the highlands (Denham & Haberle, 2008), very early, ~10 kya (Bergström et al., 2017). Until these days, Papuan agricultural fields (although fenced, drained and intensively managed) still display some traditional agricultural approaches like maintaining of tree plantations, crop rotations, and organic mulching that prevents erosion (Diamond, 1997). Concerning animal food production, beside pigs, chickens (introduced by Southeast Asians) and dogs (introduced by Austronesian speakers), PNG people knew no domesticated large mammals, which means for Papuans the need to overcome the protein deficit in a diet consisting of taro and sweet potatoes (Diamond, 1997).

Current anthropological knowledge about Papuans was assembled by a numbers of researchers conducting terrain ethnographic studies in PNG for more than 100 years as reviewed by (Morauta et al., 1979). These studies laid grounds for understanding of the basic social patterns in PNG, for instance the traditional gender specialization: men are responsible for deforestation and planting crops like bananas, sugarcane, coffee, and cocoa, whereas women take care of pigs and smaller gardening activities (Bray & Smith, 1985). Additionally, women make clothing by knitting traditional grass skirts and bags called *bilums*, and run the household including daily cooking. Men take basically absolute control over the family income, ensure the family technical support (repairing and building of houses and fences) and butcher animals (Bray & Smith, 1985).

The family unit is usually patriarchal and patrilocal, and consists of a husband (not exceptionally with his parents), a wife, and children. Larger families often gather for meals, to celebrate different occasions (work, cultural or religion ceremonies) (Bray & Smith, 1985). Papuan daily life is strongly governed by cultural traditions and sexual taboos like e.g. bride prices paid in pigs (brides are paid off by groom's family) or a taboo on homosexuality, which is illegal in PNG (Moffatt, 2012).

The country of Papua New Guinea is naturally divided into two regions: lowlands, <100 m above sea level (Bourke & Harwood, 2009), and highlands. Populations in these regions differ in several aspects like anthropological features (see below in chapters 1.2.1. and 1.2.2.), their lifestyle and population density. The lowlands have a much lower population density in contrast to a very high population density in the highlands (Bourke & Harwood, 2009), which is a phenomenon that has long been investigated in detail from the medical and epidemiological perspectives.

The difference is largely a result of high malaria endemicity in the coastal regions, in contrast to rare/absent malaria throughout the highlands (Riley, 1983). As Bergström et al. (2017) further added, this inverse correlation between population density and malaria also explains the intensity of land use, which is low in the lowlands.

1.2.1. Lowlanders

From geological and geographical point of view, lowlands are characterized by open beaches continuing towards coastal swamps or dry savannah. For island regions, also very hot and humid, rich volcanic soils are typical (JICA, 2002). Malaria is not the only limiting factor that prevents population growth in the lowlands. Another factor might have been the very low nutritional quality of traditional food. As Riley (1983) suggested, despite sufficient amount, the protein-poor diet can result in low fertility and long birth intervals. Second, such a poor nutrition also affects phenotypes like height.

In 1970s, several studies reported extremely slow growth rates amongst the inland groups living at altitudes of 600-2000 m (Malcolm, 1969; 1970). According to Malcolm (1970), maximum height in males (159 cm on average) was not reached until the age of 24 years and, in females (150 cm on average), until the age of 21 years. Woman's reproductive span was only 13.7 years and total fertility rate was 4.8. In contrast, women on the PNG islands, where there has been considerable social and economic development, had an average reproductive span of 21.7 years and total fertility rate of 9.5 (Ring & Scragg, 1973).

Having monitored positive changes in lowlander children growth for 25 years, Heywood, (1983) highlighted a correlation between their improving physical traits and increased food consumption and food import in the lowlands.

1.2.2. Highlanders

Since 1930s, an Australian traveller Michael Leahy has started his investigations of by that time practically unknown highland plateaus (Soukup, 2010). Soon afterwards though (in the beginning of 1935), colonial officials forbade any Europeans to visit highland areas, which remained unexplored until the Japanese invasion in December 1941 (Soukup, 2010). In the 1940s it was still impossible to conduct any scientific research because of the Japanese occupation. Therefore a wave of anthropological research in the highlands has not begun before 1950s. The long-term isolation of the highlands ended in 1960s.

The highlands are nowadays divided into five central provinces: Simbu (or Chimbu), Eastern Highlands, Enga, Southern Highlands, and Western Highlands, which have very limited trade connections to the coastal lowlands (Bergström et al., 2017). Both geographic and cultural isolation suggests that there are major differences between lowland and highland populations. Comparative measurements between coastal children and children from mountain regions revealed that young highlanders have longer trunks and shorter legs or have a significantly larger antero-posterior chest (Harvey, 1974) in contrast to children from lowlands.

From the linguistic point of view, people from the highlands are grouped into three rather distinct clusters (western, eastern ,and Angan speakers) that have likely been formed within the past 10,000 years, soon after the origin of agriculture (Bergström et al., 2017).

Permanent settlements in the highlands reach a maximum altitude of about 2,300 m (Schiefenhövel, 2014). Similarly to lowlander societies, the social roles among highlanders are also assigned according to gender: whereas men are occupied mostly with forest management and less often with hunting (using bow and arrows), highlander women cultivate plants (Schiefenhövel, 2014) using intensive traditional horticulture, mostly focusing on the exportable Arabica coffee and food crops like sweet potato in rotation with peanut (Bourke & Harwood, 2009).

1.2.3. Ancestry of Papuans

Present-day indigenous people from Remote and Near Oceania share their ancestry not only with Southeast Asians (SEA) but also with Papuans (Skoglund et al., 2016). According to some models in recent genetic studies (Duggan et al., 2014; Kayser, 2010; Matisoo-Smith, 2015; Wollstein et al., 2010), the admixture between Papuans and SEA occurred around 3,000 years ago as soon as Austronesian-speaking populations of SEA origin reached the New Guinea region on their way to Remote Oceania.

According to another set of models, populations of SEA origin first settled Oceania without mixing with Papuans, but subsequently Papuans mastered advanced seafaring technologies introduced by the SEA peoples and that drove the second migration wave into Remote Oceania (Friedlaender et al., 2008; Lipson et al., 2018; Posth et al., 2018; Skoglund et al., 2016). Lipson et al. (2018) analysed ancient samples from Vanuatu, an island state in Remote Oceania, and detected the arrival of individuals with almost entirely Papuan ancestry at around 2.3 kya. Later, the proportion of Papuan ancestry fell again. Despite these massive demographic changes, Papuan languages failed to replace Austronesian languages in Oceania (Posth et al., 2018). In contrast, many Austronesian languages were introduced in PNG. This suggests that the second wave of settlement in Oceania could have been Papuan genetically but Austronesian-speaking.

In stark contrast to the latest results (all present-day populations in Near and Remote Oceania have >25% Papuan ancestry) (Lipson et al., 2018), earlier studies suggested that Remote Oceanian individuals had little or no Papuan ancestry (Skoglund et al., 2016).

In 2017, Bergström et al. published a large-scale study that reported SNP genotypes for >350 individuals from 85 language groups across PNG at 1.7 million genome-wide sites. The study was devoted to the genetic history of PNG only, used standard analytical tools (PCA, ADMIXTURE, *f*-statistics) and revealed just a very general picture of the Papuan genetic history. The results of this study are discussed in this thesis (see below in chapter 8).

Jacobs et al. (2019) published a genomic study focused on the legacy of the archaic hominins (Denisovan groups sampled in Altai, Siberia) to present-day populations in Island Southeast Asia and New Guinea. Jacobs et al. (2019) confirmed that modern Papuans got ancestry from two divergent Denisovan lineages (that separated 350 kya) east of Wallace's line, whereas yet another Denisovan lineage contributed to present-day East Asians.

2. Studies of human history from a multidisciplinary perspective

A study of human past requires cooperation between many scientific fields (linguistics, archaeology, anthropology, and genetics). Genetic history of an ethnic group, history of its material culture and language are rarely fully congruent. Only considering all of those histories, side by side, we can achieve a clearer understanding of the past.

2.1. Historical linguistics

Historical linguists often explore phylogenetic relationships and interactions of languages. Such a basic concept is very similar to historical genetics because both are aimed at constructing a graph of divergence and admixture events, on very different types of data though. Related languages are grouped into so-called language families, and deeper relationships among families are actively studied. Additional analyses investigate borrowing of words or grammatical structures (Gray & Atkinson, 2003).

Linguists search for systematic sound correspondences across languages, which are later used for finding cognates, i.e. the same or closely related meanings expressed by related words (Campbell & Poser, 2008). If a certain number of cognates within the most stable core set of 100 or 200 meanings (known as a Swadesh list) is accumulated, a language relationship is considered proven (Campbell & Poser, 2008). Date estimates for language divergence events remained unavailable until the advent of Bayesian phylogenetic methods (Bentz et al., 2018, List et al., 2017).

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 negative findings made by Rafinesque (showing the languages of disparate Negrito peoples to be unrelated) was never published, his method became popular and laid grounds for new linguistic chronology methods (Jobling et al., 2014). This example shows how linguistics might support studies of human population history.

2.1.1. A linguistic characteristic of the Papuasphere

The Papuasphere is a linguistic world consisting of the mainland New Guinea, the Bismarck Archipelago, the Bougainville Island, the Solomon Islands, the islands of Halmahera, Timor, Alor, and Pantar (Palmer, 2018). As a part of his linguistic review, Palmer (2018) describes the current state of the Papuashpere: 862 languages comprising 43 distinct language families and 37 language isolates (language families composed of one language only). Language catalogues are available at glottolog.org or transnewguinea.org. However, being world's least documented region, final estimates of the number of Papuan languages vary.

The Papuan linguistic family lacks any phylogenetic or typological status, and includes any language family endemic to the New Guinea area (Palmer, 2018). Current estimates suggest that approximately three to four million inhabitants use Papuan languages actively. With about 165 thousand speakers, the most commonly spoken Papuan language is Enga (Palmer, 2018).

The total number of Papuan languages was estimated to reach many hundreds, and most of them are spoken by relatively few individuals, generally less than 3,000. Some languages have under 100 or 50 speakers (Palmer, 2018). Only seven Papuan languages have more than 100,000 speakers. Most of those seven languages belong to the Trans New Guinea (TNG) language family, with one exception from the Timor-Alor-Pantar area (Palmer, 2018).

The small number of large language communities means that language barriers were until recently common in Papuan societies. Despite this, the basics of Papuan communication were laid on shared trade jargon and alliances with economically superior neighbours (Foley, 2003). Tracing deep linguistic links among Papuan languages is difficult. Whereas detailed studies comparing individual Papuan languages exist (Vries et al., 2019), proper written evidence that would help to distinguish "indigenous" words and loanwords are missing. If not, they are hardly older than 50 years (Foley, 2003).

A major feature of the Papuan linguistic landscape is pre-historic colonisation by Australian speakers. In contrast to their rapid expansion across Melanesia 3.5 kya, in New Guinea Austronesian speakers had to deal with obstacles such as a high malaria prevalence, dense rain-forests and high mountains (Ross, 2018).

The most solid signs of the first contact between Austronesians and Papuans comes from the Timor area (Ross, 2018).

2.1.2 A linguistic overview of populations in Papua New Guinea

Without knowing exact linguistic genealogy, an overview of language families is still very helpful. The largest language family, Trans-New Guinea (TNG), is spoken across all of the PNG highlands and large parts of the lowlands. 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 (composed of more than 80 languages with some 80,000 speakers in the Madang province). As Foley (2003) noted, if all candidate languages were confirmed as members of the TNG family, then it would include almost 300 languages and two million speakers.

The lowland areas of New Guinea are in general very complex linguistically, occupied by many small language families, among them highly exotic ones: the Sentani family, the Lakes Plain family, the Cenderawasih Bay family, the East Bird's Head family, and the Western Bird's Head family (Foley, 2003). Taken together the above-named families have approximately 90,000 speakers and include 52 languages in this region (Foley, 2003).

Linguistic groupings are often used in genomic studies for defining populations like in (Bergström et al., 2017).

2.2. Archaeology

Archaeology, together with paleoecology, provides direct records of former human activity linked to environmental conditions at that time. Whereas genetic or linguistic research focuses on modelling of human genealogy in the first place, archaeology studies lifestyle and culture.

2.2.1. Archaeological findings in Papua New Guinea

In PNG, the oldest agricultural archaeological findings date back to 8 to 3 kya in the highlands (Swadling et al., 2008), whereas in across large lowland areas and many PNG islands no archaeological sites have been recorded. A PNG chronology (Swadling et al., 2008) was correlated with conclusions of previous studies of a coastal Lapita culture in South Pacific (Skoglund et al., 2017). Using their first long-distance sea boats, people of the Lapita culture were able to reach extremely distant localities such as the southern and eastern Pacific and successfully spread their culture across Remote Oceania (Skoglund et al., 2017).

Based on missing records of human settlement before 3.5 kya, it seems that inhabitants of Papuan islands (especially the West New Britain Island) lived a very mobile lifestyle based on hunting and fishing (Clark et al., 2000). In contrast to that, increasing variation in grass types, palm types, herbaceous and forest types starting around 2.3 kya shows increased environment disturbance by the Lapita culture people (Clark et al., 2000).

Two highland sites in the Upper Waghi Valley (from 1500 m to 2400 m a.s.l.) belong among most well-studied archaeological sites in PNG: Warrawau (excavated since the 1960s) and Kuk Swamp (excavated since the 1970s) (Denham & Haberle, 2008). Mapping the arrival of agriculture into wetlands, archaeologists have so far revealed 12 archaeological sites dated to mid-Holocene, 8 to 4 kya (Haberle & David, 2004; Pawley & Hammarström, 2017).

Archaeological sites are mostly located on lands formerly used as gardens or building sites (Swadling et al., 2008). As most of those sites are dated to the time of local agriculture origin (like ditched field systems dated to ca. 4 kya (Denham & Haberle, 2008)), a majority of archaeological conclusions (discussing dispersal of plants and human populations) refer to that period. Mid-Holocene agricultural tools found in at the Waim site in the Jimi valley witness socio-behavioural changes linked to slowly developing food processing at that time (Shaw et al., 2020).

Waim axe artefacts reveal a technologically highly developed culture that appeared long before the arrival of the Lapita people (by at least 1000 years, Shaw et al., 2020). Successful technological innovations (with centers in the Sepik-Ramu basin and Huon Gulf) did not remain confined to the highlands. Figurative stone carvings from Waim also share their decorative features across northern lowlands, several coastal and island populations, which suggests rich social networks in the region (Shaw et al., 2020). Waim finds of obsidian cores were traced even to a source located 800 km away in New Britain (Shaw et al., 2020).

The distribution of prehistoric stone mortars and pestles in Papua is believed to match the area where taro pudding was produced because it follows areas with taro cultivation (rather than bananas, yams or sago) (Swadling et al., 2008). Taro cultivation in PNG begins probably during the mid-Holocene period and accompanies the early-mid Holocene human occupation in Enga (Highlands), wherefrom it later spread across the highlands (Swadling et al., 2008). Until 4 kya, traces of taro cultivation were regularly found in the Sepik-Ramu region. However, with increasing sea level coastal-highland contacts were supressed to a minimum. Since then, only communities located along rivers mediated the exchange of goods (Swadling et al., 2008).

3. Genetic markers

Evolutionary geneticists trace relationships among individuals and populations based on variable regions in the genome (shaped by genetic recombination, genetic drift, and natural selection). Historically, various types of genetic variation were used for population studies.

3.1. Blood groups

The ABO blood group system, discovered in 1900, was the first human genetic polymorphism to be discovered (Owen, 2000). Based on 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 (0). Between 1950s and 1970s, Mourant et al. assembled an outstanding compendium of data on blood groups and other polymorphisms (Mourant et al., 1976). In addition to that, Cavalli-Sforza and Edwards constructed the first human evolutionary tree (of fifteen populations, three for each continent) using allele frequency data (Cavalli-Sforza & Edwards, 1963). This evolutionary tree was also an outstandingly original piece of work and formed the basis of all subsequent phylogenetic analyses. Cavalli-Sforza also introduced the use of principal component analysis (PCA) for visualization of allele frequency data, which shaped the methodology of human genetic studies for decades to come.

Other blood groups (such as MN or Rh) were also used in early population studies. In 1978, Menozzi et al. used allele frequency data in order to test for ancient demic diffusion across Europe. Based on variation in single genes (e.g. of Rh-negative alleles, or some HLA-B alleles), researchers discovered several centers of migration radiation in the Middle East (Menozzi et al., 1978).

3.1.1. Blood group surveys in Papua New Guinea

In 1963, the Red Cross Transfusion Centre at Port Moresby (the capital of PNG) took a part in a genetic survey in the Papuan Highlands (Vines & Booth, 1963). A final analysis on 192 individuals included blood group data from 20 separate locations in Eastern, Western and Southern Highlands.

All of those blood samples were matched against each other and tested for their heterogeneity. While no significant genetic differences were found between the Southern and Eastern region, the comparison allele frequencies revealed geographic isolation of Western Highlanders (Vines & Booth, 1963).

Shields et al. extended previous research by adding Coastal Papuan and Australian Aboriginal populations into his analysis (Shields et al., 1986). Based on a common belief that the Austronesians settlement of the coastal regions was a formative event in the Papuan history; the study focused on the origin of the Karimui population from Central PNG Highlands and its connections to coastal groups in PNG. Relying on HLA, blood group and serum protein markers, Shields et al. (1986) traced the ancestry of the Karimui population in Northeast and Eastern parts of Highlands, which most probably followed trade routes into those regions; and also confirmed the genetic separation of Highlanders from coastal populations.

Those examples show that blood-group-based analysis can give a glimpse into the human past.

3.2. Mitochondrial and Y-chromosomal haplogroups

Thanks to rapid development of sequencing approaches, researchers could start reading the human genome (Human Genome Project [1990] and Celera [1998]). In contrast to the early era of simple blood markers, geneticists were suddenly given the possibility to examine thousands of loci.

Whereas most of the human genome is inherited in from both parents (although remodelled by recombination), two DNA varieties are atypically inherited from one parent only (escaping recombination): the mitochondrial DNA and most of the Y chromosome (Jobling et al., 2014). Mitochondria are double-membrane organelles found with few exceptions (Shiflett & Johnson, 2010) in most eukaryotic cells. Trying to explain the origin of the mitochondrion, a complex membrane structure harbouring its own genome, led scientists to an idea of endosymbiosis, which probably happened as a reaction to changes in atmospheric oxygen level in the Precambrian (Margulis, 1972). Since then, the mitochondrion has become critical for energy metabolism.

In contrast to the human nuclear genome that is made of 3.3 billion base pairs, the mitochondrial genome (mtDNA) is composed of 16,569 base pairs (Jobling et al., 2014). In total, mtDNA contains 37 genes that encode 13 proteins, 22 tRNAs, and 2 rRNAs (Jobling et al., 2014) that are mostly involved in the process of oxidative phosphorylation, i.e. the ATP/energy production (Shiflett & Johnson, 2010). The 13 mitochondrial-encoded proteins are incorporated into the enzyme complexes of the electron transport system (Parr & Martin, 2012).

Mitochondria in humans are inherited uniparentally, in most cases maternally. In animals, the exclusively maternal inheritance is ensured, firstly, by reducing the mtDNA content in male sperm [oocytes contain around 100,000 mitochondria, each containing at least one mtDNA molecule, while sperm cells contain only about 50–75 mtDNA molecules (Jobling et al., 2014)]; and secondly, by selective elimination of paternal mtDNA within an hour after fertilisation (Wallace, 2007). Since paternal mitochondria do not persist after fertilization, mtDNA is maternally inherited in haploid form, and therefore escapes recombination. Both, the uniparental inheritance and the lack of recombination are useful features for genetic tracing of the maternal ancestral linage.

In humans, the sexual differentiation is controlled by a pair of sex chromosomes (also called gonosomes or heterochromosomes), the X and Y chromosomes: homogametic *XX* individuals are females and heterogametic *XY* are males. Due to the lack of homolog partner required for recombination, no recombination happens along more than 90% of the Y-chromosome (Jobling et al., 2014). Random mutations and rare recombination events import into such a deeply conserved system as mtDNA or Y chromosome a random variability which might be traced by geneticists while studying population history.

After human and chimpanzee mtDNA sequences were compared, the base-substitution mutation rate in mtDNA was found to be about 10 times higher than the average rate in nuclear DNA (Jobling et al., 2014). Moreover, in contrast to the average mutation rate observed in non-coding regions of mtDNA, hypervariable segments (HVS) of the mtDNA "control region" appeared to display more than a tenfold higher mutation rate (5×10^{-6}) (Jobling et al., 2014).

These two regions, HVSI and HVSII, represent two evolutionary very unstable regions that cause problems while comparing them between species, but bring advantages in population studies (Jobling et al., 2014). The rapid mutation rate, the existence of hypervariable regions, and the lack of recombination make mtDNA and Y-chromosome ideal for haplotype-based population studies.

According to a definition by Glusman et al., a haplotype is the longest sequence on one chromosome which was inherited from one parent, but might be affected by random mutagenesis (Glusman et al., 2014). Autosomal haplotypes are often assembled computationally relying on allele correlations in large sets of individuals or in parent-offspring trios (Glusman et al., 2014).

First human population studies based on mtDNA were performed by restriction enzyme analyses (Denaro et al., 1981; Torroni et al., 1992). They e.g. revealed differences between the four major races (Caucasian, Amerindian, African, and Asian) or the migration history of Native Americans.

3.2.1. Uniparental genetic studies in Papua New Guinea

Based on linguistic suggestions, several studies also examined mtDNA and Y-chromosome haplotypes on the island of New Guinea. In Austronesian-speaking groups (living mainly on islands north from New Guinea) were observed haplotypes of East Asian origins, and in non-Austronesian-speaking groups haplotypes unique to Melanesia (Kayser et al., 2003; Stoneking et al., 1990). In 2007, Mona et al. analysed 162 Y-chromosome samples from northwestern New Guinea, which is a region inhabited by non-Austroneasian speakers, to reveal its local history. Published results confirmed the major impact of Melanesians along with minor East Asian admixture, but also noticed an unexpected local impact of TNG speakers from the central highlands of PNG (Mona et al., 2007). These conclusions supported a hypothesis about Papuan spread to the west which probably started about 6 to 7 kya (Mona et al., 2007).

3.3. Currently popular genetic markers

Although uniparental studies became very popular, after the development of high-throughput genome sequencing during the first decade of the 21th century, uniparental historical genetics became outdated and was replaced by more powerful methods based on autosomal genetic variation.

Microsatellites/short tandem repeats (STRs) are tandem arrays with structural units of 1 to 7 bp in length, with a typical copy number of 10-30 (Jobling et al., 2014). While some exceptions exist (e.g. the CAG repeat expansion responsible for Huntington's disease), in general, variation at most microsatellites has most probably no influence on the phenotype (Jobling et al., 2014). In contrast to microsatellites, minisatellites are built of 8 to 100 bp long repeats, with copy numbers from as low as 5 to over 1000 (Jobling et al., 2014). Microsatellites used to be a popular tool for population genetics (Pemberton et al., 2009).

In general, the mutation process and evolution at repetitive loci is too complex. 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 more easily recognizable and which evolution is easier to model. These structural variants are termed single nucleotide polymorphisms.

3.3.1. Single nucleotide polymorphisms

Single nucleotide polymorphisms (SNPs) are one-base variants that can be found at specific positions in the genome (termed SNP-discovered loci, SDL) and can differ at the level of individuals as well as subpopulations (Wakeley et al., 2001). Considering the fact that a single human genome contains about $3x10^9$ base pairs (Olson, 1993), and that the current world human population (~7 billion people) carries ~14 billion genome copies (Jobling et al., 2014), then the worldwide frequency of SNPs might be estimated as one variant per 1,000 bases on average (HapMap, 2003).

SNPs arise via different mechanisms: by substitutions, indels (insertions, deletions), duplications, or inversions. Their frequency has been noticed as unequal though: 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 (Jobling et al., 2014). Transitions (pyrimidine-pyrimidine and purine-purine substitutions) are almost three times more frequent than transversions (pyrimidine-purine exchanges) (Jobling et al., 2014).

Since 2003, members of the International HapMap project aimed at detailed characterising of SNP patterns: their structural differences, various frequencies and frequency correlations in populations from Africa, Asia and Europe (Belmont et al., 2005; HapMap, 2003). In 2007, a final summary reported a set of 270 individuals and over 3.1 million SNPs including 25–35% of common SNP variation in the populations surveyed (Frazer et al., 2007). Novel SNPs discovered mostly through whole-genome resequencing studies are deposited in the dbSNP database (www.ncbi.nlm.nih.gov/projects/SNP). At the time of writing, dbSNP contains 686.6 million human SNPs.

The discovery of SNPs and their later characterization gave scientists a new, advanced tool that is perfectly suitable for comparative genetic studies. SNPs surveys started to be applied within many disciplines such as medicine, forensics, and agriculture (Kumar et al., 2012).

SNP-based analyses usually deal with so-called common variation, i.e. with SNPs having a global minor allele frequency of at least 5% (Jobling et al., 2014). The other alleles are considered to be rare, and are sometimes targeted specifically for high-resolution exploration of population history as for example in an archaeogenetic study of Chukotka and North Amerika (Flegontov et al., 2019).

4. Genotyping and sequencing methods

For studying genetic variation the ideal way is to carry out comprehensive genome resequencing. However, early methods were rather indirect: immunological reactions and electrophoretic analysis of gene products (Jobling et al., 2014). The amount of information supplied by those old methods was inadequate for constructing complex models of genetic history (Hurst & Jiggins, 2005).

Two revolutionary discoveries were turning points for genomics in the 20th century: the first was in the 1970s, when first chemical-based sequencing approaches were introduced and further evolved (Maxam & Gilbert, 1977; Sanger & Coulson, 1975); and the second one in the 1980s, when an original idea of polymerase chain reaction, currently known as PCR, was developed (Mullis et al., 1986).

4.1. Early sequencing approaches

Soon after the DNA sequencing methods were invented and after the first human gene was isolated and sequenced (Seeburg et al., 1977), the first shotgun sequencing strategy was also introduced (Anderson, 1981) and remained a fundamental method for large-scale genome sequencing for the next three decades. 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 represent the original) was further successively applied to larger and larger DNA molecules: from plasmids (4 kb), bacterial genomes (1-2 Mb) up to the human genome sequence (Lander et al., 2001). In practice, assembling repetitive regions is tricky, and cloning bias makes sequencing GC-rcih regions difficult too. For example, an average bacterial genome contains about 1.5% of repetitive sequences, the fruit fly genome about 3%, and the human genome about 50% (Lander et al., 2001).

In late 1990s, two independent scientific teams focused on sequencing the very first complete human genome. The initial project, the Human Genome Organisation (HUGO), was based on collaboration between six nations (United States, United Kingdom, Japan, France, Germany, and China). Having applied the hierarchical shotgun approach on genomic sequences of randomly chosen, anonymous volunteers, HUGO covered more than 96% of the human euchromatin.

Together with additional sequences in public databases, HUGO successfully assembled about 94% of the human genome (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).

4.2. Illumina sequencing

Finishing both projects caused a lively debate, which has led to a detailed comparison of both approaches: Venter's random "shotgun" sequencing approach and hierarchical "chromosome walking" practiced by the HUGO. So, unintentionally, the assembly of the first human genome laid grounds for a new era of the second- and third-generation sequencing methods, collectively termed next-generation sequencing (NGS). In general, NGS is characterized by the shift from long 500-800 nt reads to short reads ranging roughly from 50 to 500 nt, depending on the protocol, by higher error rates in comparison to the Sanger sequencing protocol, and sequencing output increased by many orders of magnitude (Slatko et al., 2018). Such a high-throughput approach reduces the costs dramatically (Slatko et al., 2018). Among second-generation sequencing methods (e.g. 454 pyrosequencing or Ion Torrent), the most popular method is Illumina sequencing developed by the company of the same name.

Third-generation sequencing methods again feature very long reads (up to 30 kb) and relatively modest throughput (Slatko et al., 2018), however Illumina is still the most popular sequencing technology today. Illumina enables reading 500 nt DNA or cDNA fragments from both ends (Tan et al., 2019), with relatively high error rates of 0.1 - 0.01% (Ma et al., 2019). The protocol starts with DNA fragmentation, which, if using e.g. the Infinitum protocol (Mason, 2013), avoids a PCR step, and therefore streamlines the procedure and makes sequencing more bias-free.

After molecule immobilization on chips, clusters of identical molecules on chips are generated by a special PCR protocol. Fluorescently labelled nucleotides are then added one by one, a fluorescent signal is induced, recorded, then a blocking group us removed, and the machine continues reading the next position (Mason, 2013). Several estimates suggest that current NGS protocols are capable of assaying more than one million SNPs simultaneously, and also guarantee 99% accuracy and reproducibility (Jobling et al., 2014).

Obviously, NGS offers highly sensitive and reliable detection of genetic variants at a single-base resolution. Finally, as a hypothesis-free approach, NGS requires no prior knowledge of the sequence information (Lohmann & Klein, 2014).

5. Computational methods for analysing population history

As said before, the most commonly used genetic polymorphisms are SNPs. A SNP is located at a particular site on the DNA sequence. A contiguous set of alleles co-inherited and on a chromosome is called a haplotype. Correlation between allele frequencies at adjacent loci is termed linkage disequilibrium (LD) (Reich et al., 2001). Analytical methods used for investigating genetic history rely on either allele frequencies at unlinked loci or on "semi-stable" autosomal haplotypes, or on uniparental haplotypes. Here we will discuss several most common method classes: ordination or dimensionality-reduction methods, *f*-statistics and model-based methods.

5.1. Ordination methods

When thousands of variables are analysed, a large fraction of them is usually non-independent (correlated). It is often useful to reduce the dimensionality of the data, i.e. to replace thousands of correlated variables with few uncorrelated variables and plot the samples in the space of these new variables. That is a concise description of dimensionality reduction (ordination) methods: multi-dimensional scaling (MDS) is widely used in community ecology; 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 et al., 2018); and finally principal component analysis (PCA) is a tool most popular in archaeogenetic studies.

Uniform Manifold Approximation and Projection (UMAP) takes a matrix of all vs. all distances among samples and pictures them in two- or more dimensions relying on a nearest neighbour graph (McInnes et al., 2018). The most important algorithm setting is a number of nearest neighbours considered, which controls how UMAP balances local versus global structure in the data (McInnes et al., 2018). In comparison to PCA, UMAP applied to multi-locus SNP genetic data demonstrated in some cases an increased resolution (Diaz-Papkovich et al., 2019), however admixture is not manifested as "clines", which is a disadvantage for population studies.

Principal component analysis (PCA) is a dimensionality reduction technique most widely used in genetic studies (Jolliffe & Cadima, 2016). Since 1901, PCA has been employed as a statistical tool for revealing structure in datasets (Pearson, 1901). Principal components (PCs) are linear combinations of original variables (in the case of genetic data, allele frequencies at various sites), and PCs are orthogonal with respect to each other. 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.

Interpreting PCA plots is not straightforward (Novembre and Stephens, 2008; McVean 2009, Patterson et al., 2006). Closely related individuals are clustered together, and admixed individuals lie between the ancestral clusters, forming "admixture clines". However, imbalanced dataset composition (over-representation of certain groups) skews the resulting plots (McVean, 2009), and clines can appear if gene flows are homogeneous in space and time, i.e. in the absence of isolation, migration, and admixture (Novembre et al., 2008). Thus, PCA should be used for formulating hypotheses that should be tested using other methods. PCA on multi-locus SNP genetic data is implemented in a number of software packages, most notably smartPCA (Patterson et al., 2006).

5.2. *f*- and D-statistics

For investigating shared genetic drift in sets of two, three, or four populations, and for testing of simple hypotheses about admixture events, *f*-statistics are used (Patterson et al. 2012, Peter, 2016, Soraggi and Wiuf, 2019). *f*-statistics form a basis for demographic models: non-phylogenetic admixture models (qpAdm) or admixture graphs. For any four groups there are three possible unrooted 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. This can be assessed first by averaging the quantity (pA-pB)(pC-pD) across SNPs, and further by testing for consistency on resampled datasets (Patterson et al., 2012; Reich et al., 2009).

***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 statistic for to distinguishing introgression from incomplete lineage sorting. 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. A closely related *D*-statistic (Green et al., 2010) was developed originally for detecting Neanderthal admixture in AMH. Both f_4- and *D*-statistics deviate significantly from 0 if the topology tested is wrong, or if there was a gene flow between a pair of branches in the unrooted four-population tree. Thus, these statistics can be used as tests for admixture, although their interpretation is not unambiguous (Green et al., 2010)

5.3 Model-based approaches

ADMIXTURE is a software tool for ancestry modelling in unrelated individuals relying on multi-locus SNP genotypes and not relying on any phylogenetic tree (Alexander et al., 2009). ADMIXTURE takes a pre-defined number of hypothetical ancestral populations and computes a matrix of ancestral population fractions ("admixture proportions") for each individual (Alexander et al., 2009). As compared to older methods (FRAPPE, STRUCTURE), ADMIXTURE is much faster (Alexander et al., 2009). On the other hand both STRUCTURE and ADMIXTURE analyses lack an explicit historical model and assume unrealistically that all ancestral populations originate from a single group. Their results can be affected by strong genetic drift in certain groups and thus are tricky to interpret (Lawson et al., 2018).

GLOBETROTTER is a haplotype-based tool that enables detection and dating of one or two admixture events that happened in the history of the target group within the last ~4,500 years (Hellenthal et al., 2014). As Hellenthal et al. (2014) mentioned, to identify an admixture event, the GLOBETROTTER method *a priori* requires no detailed sampling of source populations. On the other hand, the reliability of admixture inference decreases if tested groups are genetically too similar. Similarly challenging is the identification and interpretation of a bidirectional gene flow.

qpAdm and **qpGraph** methods included into the AdmixTools package (Patterson et al., 2012). A pair of closely related tools named qpWave (Reich et al., 2012) and qpAdm (Haak et al., 2015) are now used widely in archaeogenetic studies. The qpAdm approach is based on *f*-statistics and does not require a detailed knowledge of the population phylogeny beyond few simple assumptions (Lazaridis et al., 2014; Haak et al., 2015)T. his method allows testing combinations of proxy ancestral groups ("sources") for a target, given a set of outgroups. Outgroups are usually ancient genomes distant in time from the supposed ancestry sources. Admixture proportions in the target are also estimated by qpAdm. Non-nested combinations of proxy sources cannot be ranked by model likelihood, but can be classified according to a chosen *p*-value threshold. qpAdm relies on a matrix of f_4-statistics f_4(target/source, target/source$_j$; outgroup, outgroup), which can be reduced to a smaller matrix f_4(target, source; outgroup$_1$, outgroup$_j$) if all statistics are calculated on the same set of variable sites (Harney et al., 2020). A conceptual problem of qpAdm is that it can hardly detect ancestry sources distant from any sampled groups, i.e. "ghost" populations.

Answering questions about human history often requires testing complex demographic models, i.e. admixture graphs (phylogenetic trees with admixture events added). Diverse graph-fitting methods exist, and qpGraph (Patterson et al., 2012) is a widely used tool relying on *f*-statistics. It produces admixture graphs with admixture proportions and edge lengths (scaled in units of relative genetic drift) and is very fast as compared to Bayesian methods relying on site frequency spectra: momi2 (Kamm et al., 2019), rarecoal (Flegontov et al., 2019; Schiffels et al., 2016)), fastsimcoal (Excoffier et al., 2013) so that topologies can be explored exhaustively for crucial sub-graphs.

The popularity of the methods listed above is explained by their modest computational requirements and versatility: the ability to analyze biallelic genotype data of various types (pseudohaploid or diploid) and generated using targeted enrichment or shotgun technologies, and the ability to analyze low-coverage ancient genomes with a high proportion of missing sites (Harney et al., 2020). In contrast, other analytical methods, such as identity by descent (IBD) block sharing, ChromoPainter, GlobeTrotter, MSMC, diCal, fastsimcoal, momi, Rarecoal, Tsinfer, and Relate, require phased data (haplotypes) and/or high-confidence allele calls generated from high-coverage data, and hence are not applicable to a great majority of ancient DNA data now analyzed in most studies.

6. Aims of the study

Although recent genome-wide studies of present-day Papuans, especially (Bergström et al., 2017) also focused on the PNG mainland and have contributed a lot to understanding the Papuan history, a a complex admixture graph including several lowlander and highlander groups has never been constructed.

Stated goals:

- To re-check Papuan population structuring as described in (Bergström et al., 2017).
- To build a detailed phylogeny of major mainland Papuan groups relying on the admixture graph algorithm (Patterson et al., 2012) and rigorous model-ranking approaches (Flegontov et al., 2019).
- To reveal intra-Papuan admixture events relying on the admixture graph algorithm (Patterson et al., 2012) and rigorous model-ranking approaches (Flegontov et al., 2019).

Stated hypotheses:

- Papuans are believed to share a widespread Southeast Asian ancestry.
- Papuans form two clades (Highlanders and Lowlanders) respecting the population geographic distribution.
- Papuan Highlanders, falling into two major groups (Western and Eastern) are believed to form a clade within a wider diversity of lowlanders

7. Methodology

7.1. Dataset composition

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

Papuan individuals originate mainly from central PNG (Highlands), the southern coast (Lowlands), the northern coast (the East Sepik province), 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 Australians, 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 consisted 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, 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, 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).

Based on description provided by (Bergström et al., 2017), Papuan tested populations, which means 337 individuals (out of total 397, when 60 remained unidentified) were further classified into 17 geographically-based-on metapopulations, and later into 80 populations according to their sampling locations. Finally, their identity was associated with 26 main Papuan language families (Tab. I). When building the admixture graphs, the population structuring was used as seen in the classification in Tab. I.

Tab. I. *A list of Papuan (meta-) populations including the numbers of analysed individuals and their linguistic classification available from: ethnologue.com (*TNG: Trans-New Guinea).*

Papuan meta-populations	Papuan populations	Individual Counts	Language classification	ISO 639-3
Gulf	Akoye	3	TNG - Angan	miw
Eastern_HL	Alekano	4	TNG – K.Goroka	gah
Madang	Amaimon	1	TNG - Madang	ali
East_Sepik	Ambulas	2	Sepik	abt
Madang	Amele	2	TNG - Madang	aey
Southern_HL	Angal	8	TNG - Engan	age
Madang	Aruamu	1	Ramu – L.Sepik	msy
Eastern_HL	Awiyaana	5	TNG – K.Goroka	auy
Madang	Bemal	1	TNG – Madang	bmh
Eastern_HL	Benabena	1	TNG – K.Goroka	bef
Madang	Biyom	1	TNG - Madang	bpm
East_Sepik	Boikin	4	Sepik	bzf
Bougainville	Bougainville	2	-	-
Western_HL	BoUng	2	TNG - Chimbu	mux
Chimbu	Chuave	3	TNG - Chimbu	cjv
Chimbu	Dadibi	1	TNG - Teberan	mps
Milne_Bay	Dobu	1	Austronesian	dob
Southern_HL	East_Kewa_Aiya	7	-	-
Southern_HL	East_Kewa_Aliya	10	-	-
Enga	Enga	10	TNG - Engan	eng
Southern_HL	Erave	14	TNG - Engan	kjy
Southern_HL	Foi	2	TNG – E.Kutubu	foi
Central	Fuyug	2	TNG – SE.Papuan	fuy
Madang	Gende	44	TNG – K. Goroka	gaf
Madang	Giri	1	Ramu	geb
Chimbu	Golin	6	TNG - Chimbu	gvf

Papuan meta-populations	Papuan populations	Individual Counts	Language families	Languages names
Central	Grass_Koiari	3	TNG – SE.Papuan	kbk
Southern_HL	Heneng	7	TNG - Engan	akh
Central	Hula	6	Austronesian	hul
Southern_HL	Huli	20	TNG - Engan	hui
Central	Humene	1	TNG – SE. Papuan	huf
Gulf	Ikobi	1	TNG – T. Kikorian	meb
Southern_HL	Imbongu	4	TNG - Chimbu	imo
East_Sepik	Kairiru	1	Austronesian	kxa
Eastern_HL	Kamano	2	TNG – K.Goroka	kbq
New_Ireland	Kara	1	Austronesian	leu
Morobe	Kate	2	TNG – F.Huon	kmg
Central	Kepara	4	Austronesian	khz
Madang	Kominimung	1	Ramu	xoi
Northern_HL	KorafeYegha	3	TNG – G.Binaderean	kpr
East_New_Britain	Kuanua	3	Austronesian	ksd
Chimbu	Kuman	13	TNG - Chimbu	kue
Western_HL	K(a)yaka	1	TNG - Engan	kyc
Western_HL	Maring	1	TNG - Chimbu	mbw
Western_HL	Melpa	3	TNG - Chimbu	med
Central	Motu	15	Austronesian	meu
Central	Mountain_Koiali	1	TNG – S. Papuan	kpx
Morobe	Nabak	1	TNG – F. Huon	naf
Western_HL	Nii	4	TNG - Chimbu	nii
Madang	Nobonob	2	TNG - Madang	gaw
Northern_HL	Orokaiva	1	TNG – G.Binaderean	okv
Gulf	Orokolo	5	TNG - Eleman	oro
New_Ireland	Patapatar	1	Austronesian	gfk
Madang	Pondoma	1	TNG - Madang	pda

Papuan meta-populations	Papuan populations	Individual Counts	Language families	Languages names
Gulf	Purari	1	TNG - Eleman	iar
Madang	Rao	2	Ramu	rao
Southern_HL	Samberigi	1	TNG - Engan	ssx
Eastern_HL	Siane	4	TNG – K.Goroka	snp
Eastern_HL	Simbari	1	TNG - Angan	smb
Chimbu	Sinasina	10	TNG - Chimbu	sst
Central	Sinaugoro	3	Austronesian	snc
Madang	Sop	1	TNG - Madang	urw
Western_HL	Southern_Kiwai	5	TNG - Kiwaian	kjd
Gulf	Tairuma	6	TNG - Eleman	uar
Madang	Takia	1	Austronesian	tbc
Central	Tauade	1	TNG – S.Papuan	ttd
Gulf	Toaripi	7	TNG - Eleman	tqo
Eastern_HL	Tokano	1	TNG – K.Goroka	zuh
Manus	TuluBohuai	1	Austronesian	rak
Milne_Bay	Umanakaina	1	TNG – S.Papuan	gdn
Western_HL	UmbuUngu	4	TNG - Chimbu	ubu
Madang	Wagi	6	TNG - Madang	fad
Western_HL	Wahgi	7	TNG - Chimbu	wgi
Central	Waima	5	Austronesian	rro
Southern_HL	West_Kewa	1	TNG - Engan	kew
Southern_HL	Wiru	12	TNG - Wiru	wiu
Morobe	Yabem	2	Austronesian	jae
Eastern_HL	Yagaria	2	TNG – K.Goroka	ygr
Eastern_HL	Yawiyuha	2	TNG – K.Goroka	yby
Eastern_HL	Yipma	4	TNG - Angan	byr

In total:				
17	**80**	**337**	**26**	**-**

7.2. Bioinformatic analyses

7.2.1. Dataset preparation

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

7.2.2. Principal Components Analysis

EIGENSTRAT v.6.0.1 was used for performing principal components analysis (PCA). Input data were pruned for linkage disequilibrium using PLINK v.1.9. The PCA analysis was performed on 1,196 individuals and 244,604 SNP sites. Considering the results of the PCA analysis (the PC1 vs. PC2 plot), I removed from the Papuan groups three outliers with non-Papuan ancestry.

7.2.3. Admixture graphs

I built 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 further steps of the analysis, it was decided to keep the geographically and genetically pre-determined division of Papuans into Highlanders and Lowlanders. Highlanders and Lowlanders were 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 group (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 according to model likelihood. Below I describe various metrics that should be considered for ranking admixture graph topologies.

7.2.4. Model-ranking metrics for admixture graphs

Z-score or the worst residual. The Z-score is a difference between the observed f_4 statistic and the statistic expected under the model, 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 admixture graph model. Absolute Z-scores below 3 are usually considered acceptable (Patterson et al., 2012;Lipson et al., 2017). Ranking alternative topologies by their Z-scores is not an optimal approach since Z-scores of multiple 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.e an admixture event, I 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 the same threshold was used when comparing bidirectional and unidirectional models.

Edges. Trifurcations in the tree manifest themselves as "internal" graph edges having a 0 length. 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 squared 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 more suitable 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 et al., 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 et al., 2017). I note that this is a work in progress and further testing of the method on simulated genetic data is needed to find optimal model-ranking metrics and thresholds.

8. Results and discussion

The sub-division of Papuan populations analysed here is based on the geographical division of the country into provinces as showed in the maps below (Fig. 1 and Fig. 2). In the first figure (Fig. 1), boundaries of both highland and lowland regions are visible; nevertheless, to get a better resolution, only lowland and island provinces are marked in the first map.

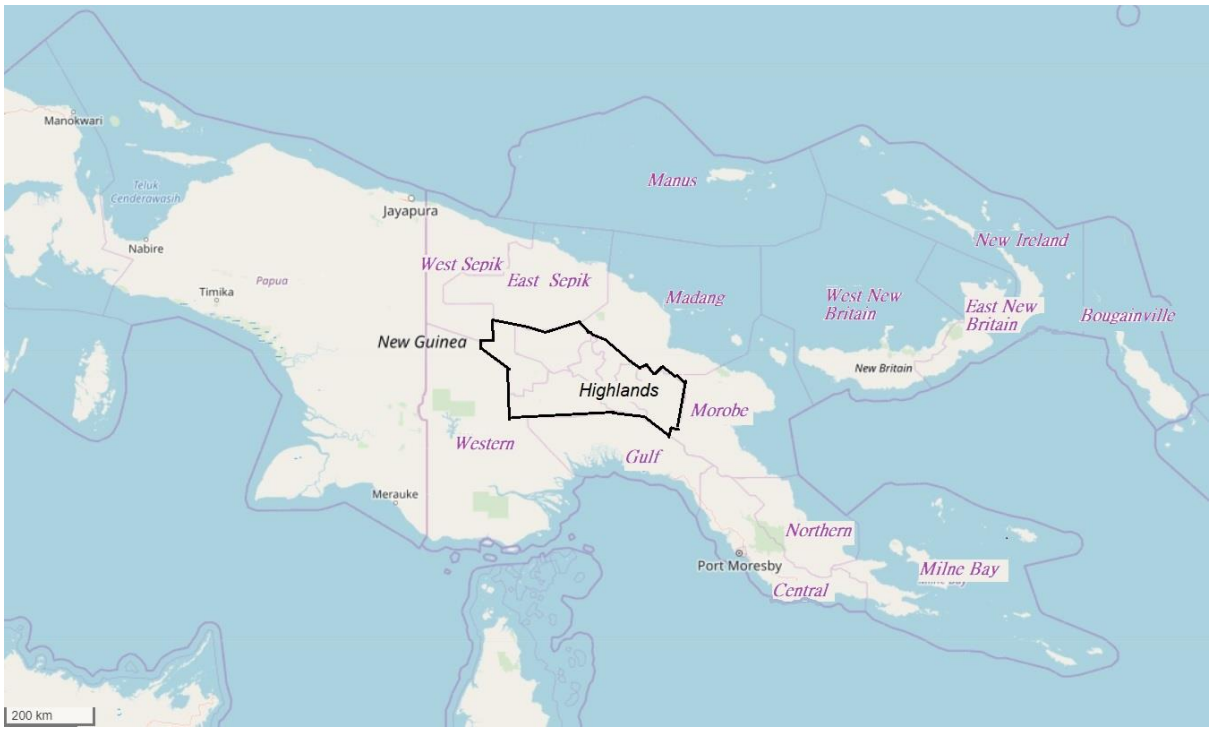

Fig. 1. ***A map of PNG focused on the lowland parts****. The figure shows geographic boundaries of 14 Lowlander groups (Madang, Gulf, Morobe, Milne Bay, East and West Sepik, East and West New Britain, Northern, Central, Western, Manus, New Ireland) that were used for initial admixture graph analyses. All of them match the provincial borders (available from: transnewguinea.org. ONLINE [07-05-2020]).*

The second figure (Fig. 2.) shows the central region of highland provinces and their boundaries only.

Fig.2. ***A map of PNG focused on the highland parts***. *The figure shows geographic boundaries of five Highlander groups (West Highlanders, Simbu, East Highlanders, South Highlanders, Enga) that were used for initial admixture graph analyses. All of them match the provincial borders (available from: transnewguinea.org. ONLINE [07-05-2020]).*

Two different PCA analyses were performed to visualize genetic distances among the studied individuals. The first plot (Fig. 3) displays present-day Europeans (EUR), East Asians (EAS), Southeast Asians (SEA), South Asians (SAS), Negritos, Remote and Near Oceanians including Polynesians (POL), Papuans and Australians The second PCA plot (Fig. 4) shows the present-day Papuans and Southeast Asians only.

PCA is a low-dimensional representation ("embedding") of the genetic diversity, in this case, visualizing genetic distances in just two dimensions: PC2 here reflects internal genetic structure in SEA and partially the genetic structure in Papuans; PC1 reflects just SEA admixture in Papuans. The clustering of Papuans does not reflect any intra-Papuan structure, but rather varying levels of SEA admixture in Papuans.

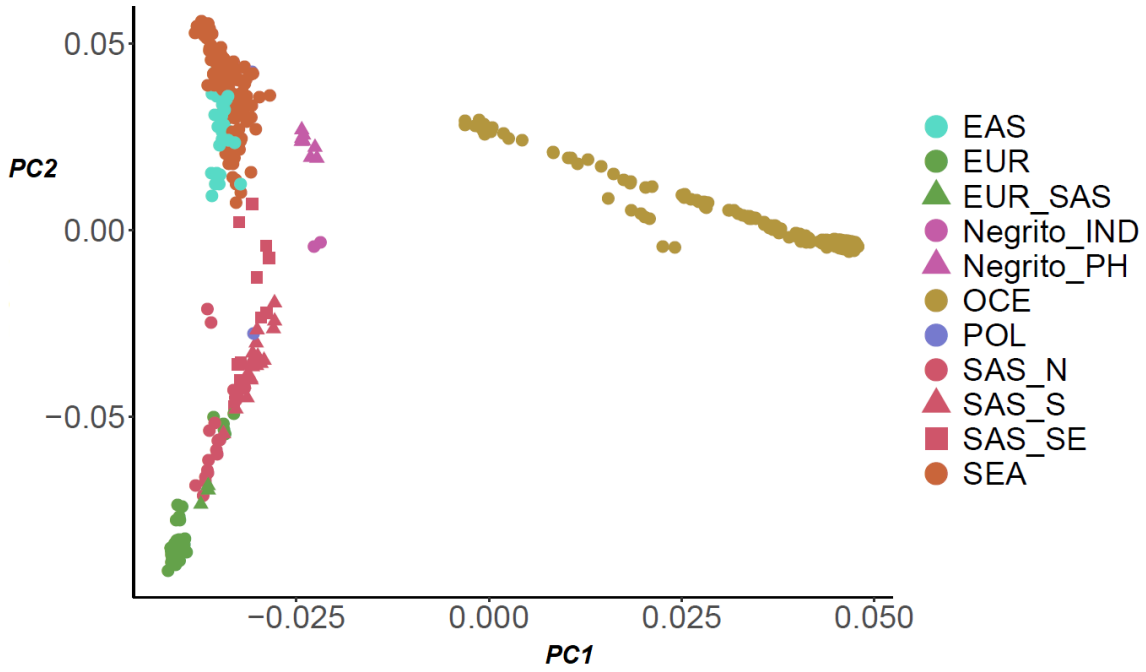

Fig.3. ***A PCA plot of Eurasians and Oceanians.*** *The following meta-populations are plotted: present-day East Asians (EAS), Europeans (EUR), Europeans with Indian ancestry (EUR_SAS), Andamanese Negrito (Negrito_IND), Negrito from the Philippines (Negrito_PH), Near Oceanians (OCE), Remote Oceanians or Polynesians (POL), North Indians (SAS_N),), South Indians (SAS_S),), Indians (South Asians) of unclear affiliation (SAS), Indians with Southeast Asian admixture (SAS_SE), and Southeast Asians (SEA). A plot of two principal components (PC1 vs. PC2) is shown.*

Papuans form a linear cline directed towards SEA. Clines on PCA plots do not necessarily reflect admixture of previously isolated groups (Novembre et al., 2008), but that interpretation was supported by other analyses (ADMIXTURE and *f*-statistics) in (Bergström et al., 2017).

Taking a closer look at the Papuan cluster itself, it is noticeable that a great majority of HLs have the lowest proportion of SEA admixture among Papuans (probably no SEA admixture), whereas almost all LLs have at least some SEA admixture.

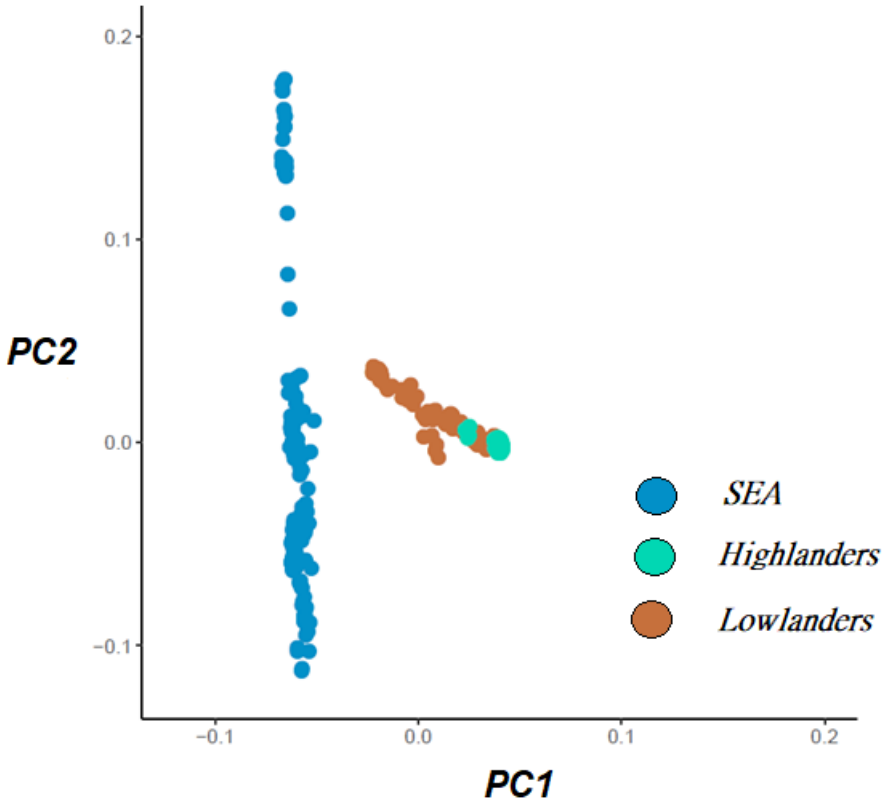

Fig.4. A ***PCA plot of Papuans and Southeast Asians.*** *The following groups most relevant for the aim of this study are plotted: present-day Southeast Asians (SEA), Papuan LLs and HLs. This plot illustrates a cline composed of LLs and formed by widespread SEA admixture among LLs. A plot of two principal components (PC1 vs. PC2) is shown.*

Whereas previous studies (Bergström et al., 2017; Lipson et al., 2017; Posth et al., 2018; Skoglund et al., 2017; Skoglund & Mathieson, 2018) presented very simple admixture graphs only, this thesis was aimed at building a more complex graph incorporating several LL and HL branches and a number of intra-Papuan gene flow events.

Six TNG-speaking HL groups (East Highlanders, South Highlanders, West Highlanders, Enga, Madang, Simbu); and 12 LL 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).

The process of admixture graph building was for simplicity split into three main phases: first, searching for a best-fitting tree including HLs only (Fig. 5); second, the searching for a best-fitting tree composed of LLs only (Fig. 6); and third, constructing a tree incorporating both HLs and LLs and including intra-Papuan admixture events (Fig. 7). 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.

The first tree (Fig. 5) includes five unadmixed HLs meta-populations (Enga, South Highlanders, East Highlanders excluding Angan speakers, Madang, and Simbu). We tested all possible tree topologies for these five groups, and this topology was significantly different from all the other topologies according to a likelihood ratio threshold of e^3. The graph does not fit the data perfectly according to a commonly accepted threshold of 3 standard errors (Z-score = 3.6).

However, adding intra-highlander admixture events or any gene flow from Southeast Asia did not improve the worst residual and did not improve likelihood significantly (likelihood ratio $>e^4$). The tree in Fig. 5 revealed two main clades within HLs: the western clade including Southern HLs and Enga, and the eastern branch (East HLs, Simbu and Madang).

This subdivision of HLs into two main clusters supports the results previously published by (Bergström et al., 2017)Furthermore, 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.

Another task focused on the incorporation of West HL: According to my previous results, those of Western_TNG populations should occupy an apical position close to Highlanders. Therefore, a topology of that part tree was (Morobe, (Toaripi, (ESepik, Highlanders))). As described in Tab. II., four of those differently branching, tested topologies had the best (although not fitting) likelihood, but none of them appeared to be fitting (as discussed in chapter 9).

Tab. II. *Gene flow testing parameters and final statistic in the combined West HL tree.*

Western HL position	**Gene flow**	**f4 statistic**				**Std. err.**	**Z-Score**
Prior to the split of E Sepik	Unidirec.	Mor	Toa	Wes	Mad	0.000601	4.303
On the ESepik branch	Unidirec.	Toa	Mad	Wes	SHi	0.000345	5.079
On the Toaripi branch	Bydirec.	Mor	SHi	ESe	Wes	0.000612	3.774
Before the Toaripi split	Unidirec.	Mor	Toa	Wes	Mad	0.000601	4.493

Having found a tree for HLs, I started building models for Lowlander populations, which resulted in one tree significantly different from the other ones but not fitting the data (Z-score = 4.7), presented in Fig. 6. By default I added into each LL branch a gene flow from a SEA source (Kankanaey), to account for SEA admixture found in nearly all LL groups (Fig. 4, Bergström et al., 2017).

I found that the best fit is observed when the SEA source is on the Kankanaey, and not on the Ami branch. The Kankanaey people are an indigenous population of the Northern Philippines that belong to Austroneasian speakers who migrated about 3500 years ago from Taiwan (Mörseburg et al., 2016). Whereas aboriginal groups from Taiwan here represented by Ami were influenced by more recent Han Chinese migrations to Taiwan, the Kankanaey are thought to have remained an isolated relict population (Mörseburg et al., 2016).

Therefore it is not surprising that Kankanaey represent a better-fitting ancestry source for Papuans. I tested several sources and according to the lowest Z-score, Kankanaey population was selected the best one.

First, I tested 840 trees with 4 successive Lowlander branches, and with Enga as a representative HL group. Second, I tested 2,520 trees with 5 successive LL branches, selected a group of best models according to the worst residual and aligned the branching orders in these trees.

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, (E, (F, (HL)))))). The branching order is as follows: Australians, Papuans from the Bismarck Archipelago, Northern LLs, Central LLs, Morobe LLs, Gulf LLs, East Sepik LLs, HLs.

Populations from the Bismarck Archipelago (New Ireland and New Britain) split first from the Papuan clade, and have a lot of SEA ancestry (23%), which is not surprising since ancient Austronesians were expert mariners and settled the islands around PNG first (Lipson et al. 2018, Posth et al. 2018). The second and third splits in the Papuan clade are formed by the Northern and Central LLs. Morobe LLs split next, and displays the second lowest fraction of SEA ancestry (5%) after East Sepik LLs who have just 1%. East Sepik LL is a closest sister-group for HL, and like HL this group is almost devoid of SEA genetic influence.

Since the basic topologies were profiled with the best likelihood of both, HLs *(Z-score = 3.64)* and LLs *(Z-score = 4.71)*, these two models were combined giving a very basic HL/LL template *(Z-score = 4.789)*. In the next step, nine additional templates of intra-Papuan admixture events were selected based on the best-likelihood-criterion (Tab. III.), and furthermore in a systemic fashion way, all of those templates were combined against each.

Tab. III. *Gene flow testing of parameters of nine intra-Papuan admixture templates including the final statistics based on they were selected for being templates.*

Gene flow	Population1	Population2	Z-Score
Unidirec	Kuanau	Morobe	4.775
Unidirec	Kuanau	ESepik	4.752
Unidirec	Northern_TNG	ESepik	3.876
Unidirec	Sinaugoro	Toaripi	4.795
Unidirec	Sinaugoro	ESepik	3.891
Bydirec.	ESepik	Morobe	3.875
Bydirec.	ESepik	Toaripi	3.847
Bydirec.	Morobe	Northern_TNG	4.769
Bydirec.	Northern_TNG	Toaripi	4.784

After establishing of the combinatory pattern, the final number of 21 combinations was run, and their final Z-scores calculated via f-4 statistics. The combinatory results are listed in Tab. IV., following the decrease in Z-Score.

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

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

Combining both of them, and by reducing selected meta-populations to populations, I finally received one improved HL-LL-mixed topology accepting the threshold of 3 standard errors, the best fitting likelihood *(Z-score = 2.97)*.

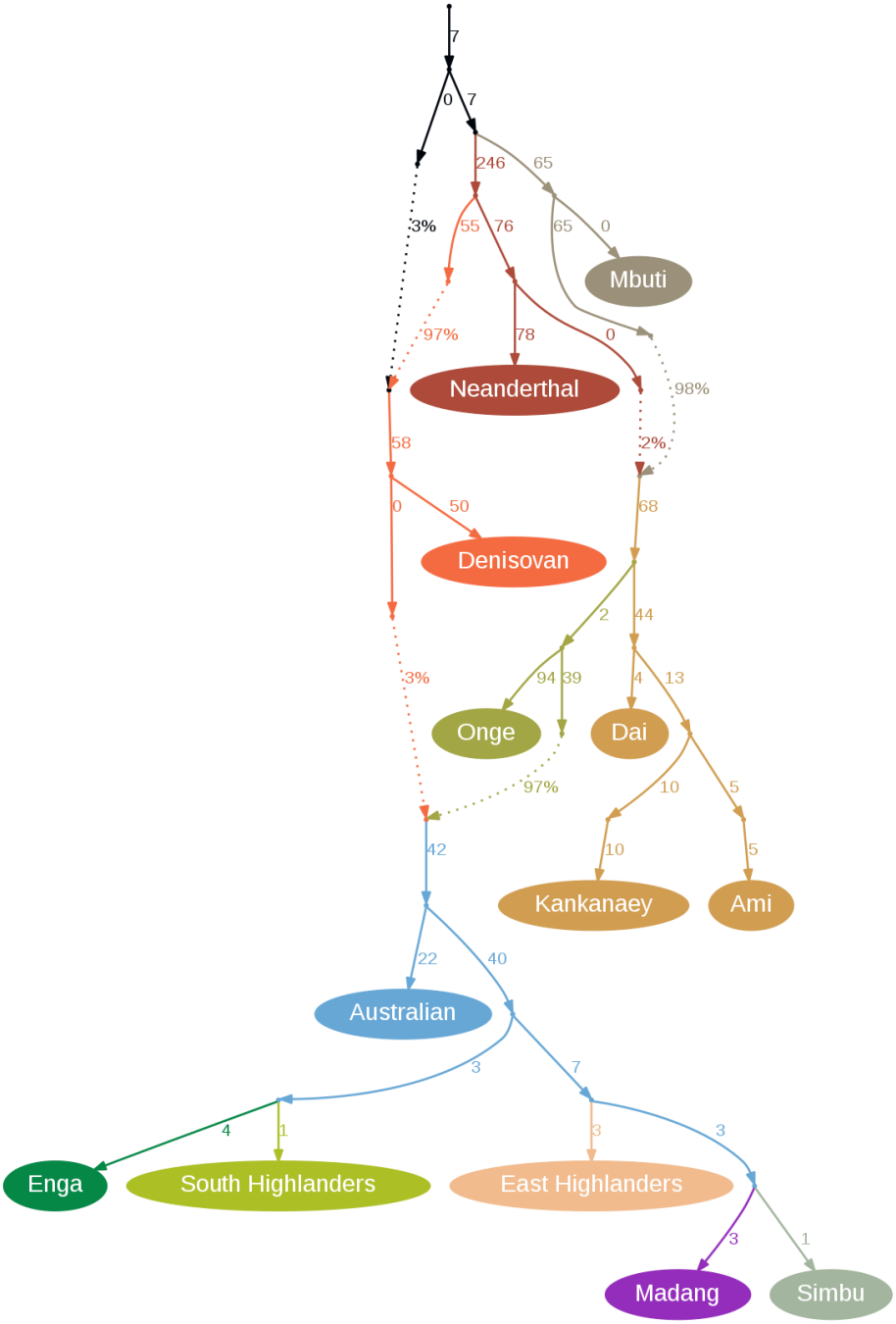

Fig. 5. ***An admixture graph for the HLs clade****. The best tree for HLs includes five unadmixed HL meta-populations (Enga, South HLs, East HLs, Madan,g and Simbu HLs). No LLs are included here. The graph does not fit the data perfectly according to a commonly accepted threshold of 3 standard errors (Z-score = 3.64).*

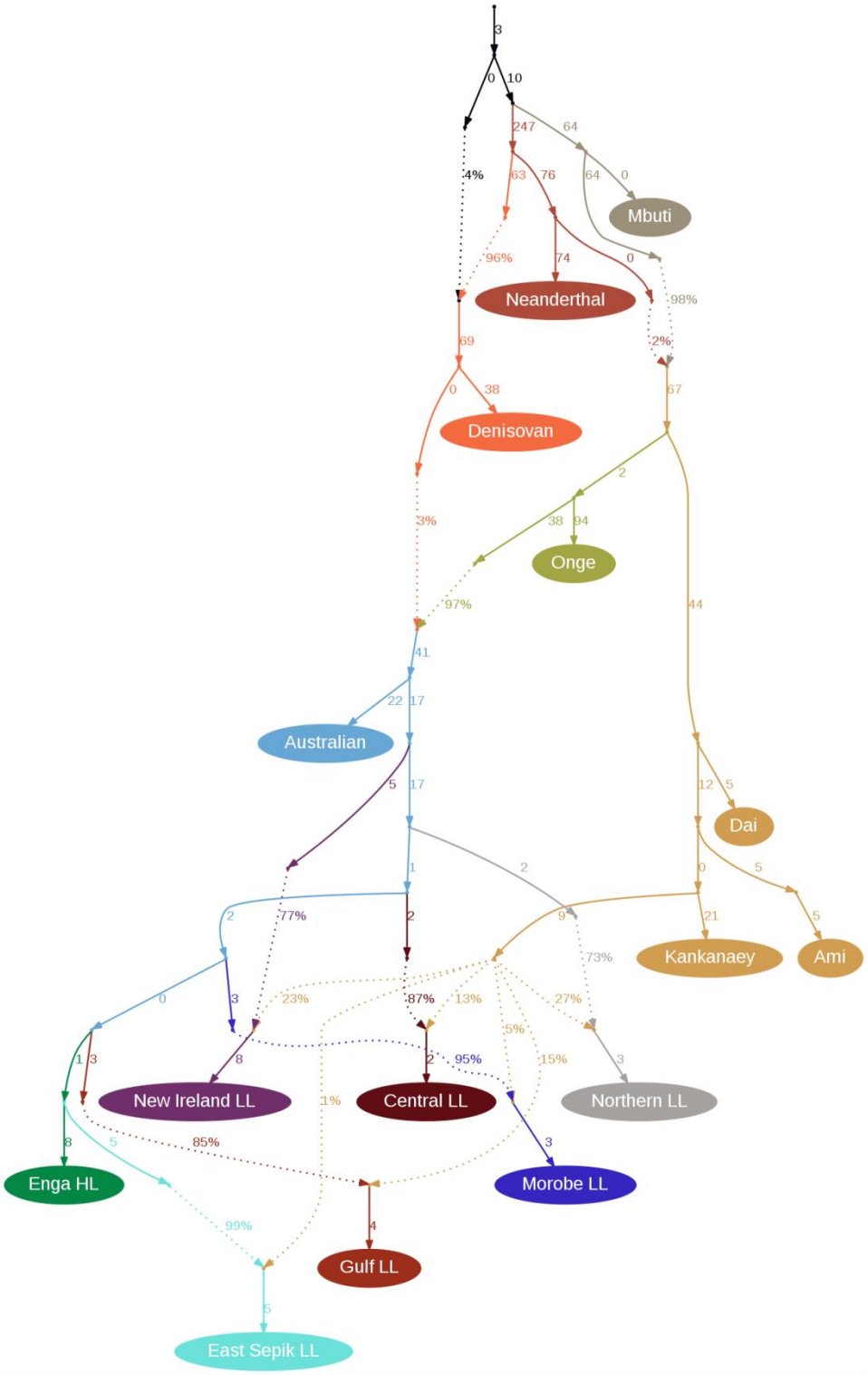

Fig. 6. ***An admixture graph for LLs.*** *The best tree for LLs includes six LL meta-populations (Table 1): East Sepik, New Ireland, Gulf, Central, Morobe, and Northern LLs. One unadmixed HL population (Enga) is included here. All LLs were modelled as receiving at least a low level of gene flow from a Kankanaey-related source (SEA). The graph does not fit the data perfectly according to a commonly accepted threshold of 3 standard errors (Z-score = 4.71).*

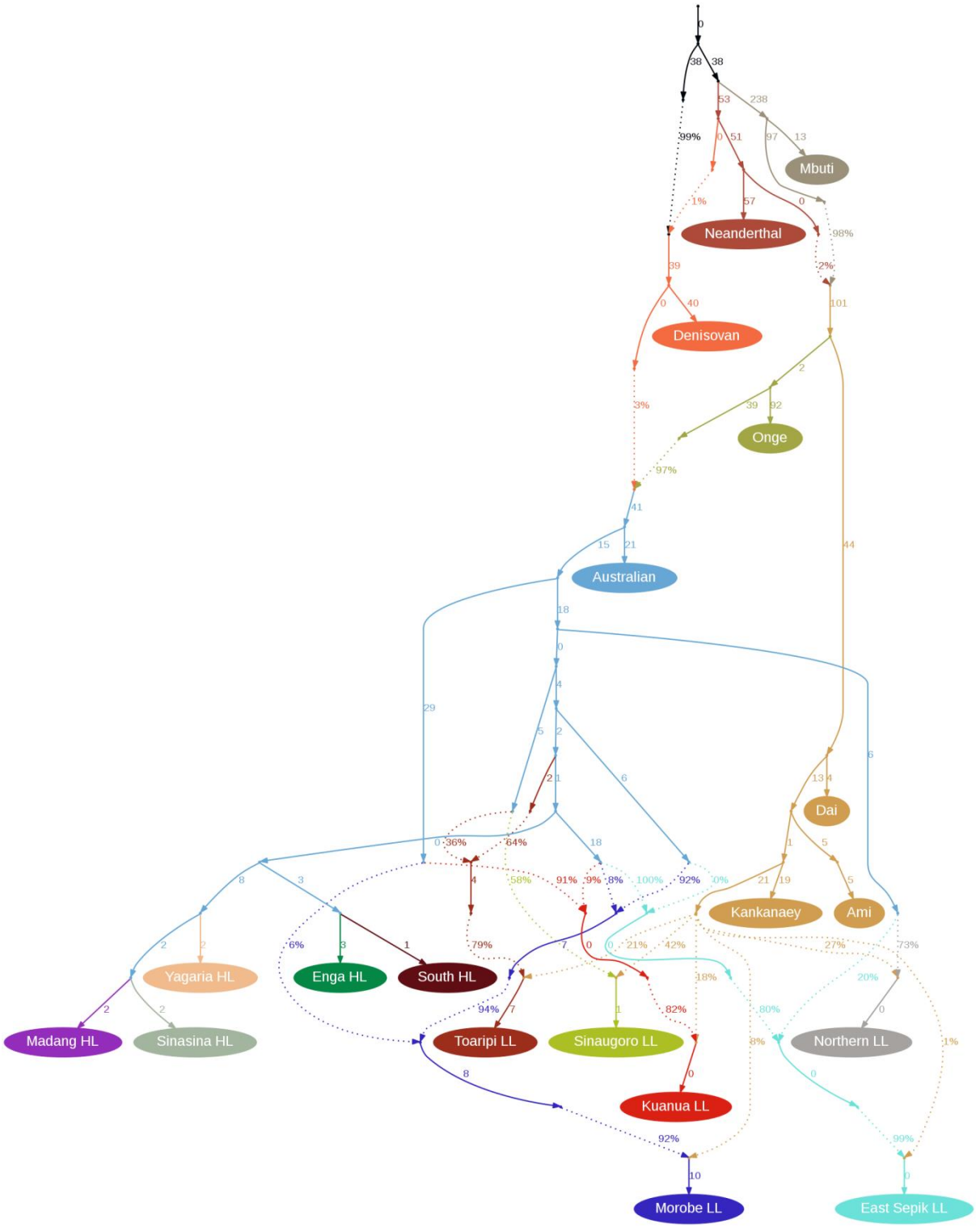

Fig. 7. *The final topology of the Papuan sub-graph obtained after optimizing meta-population composition and adding five intra-lowlander admixture events. The worst Z-score = 2.97, i.e. the model fits the data. The model presented here includes 5 unadmixed HLs and 6 admixed LLs groups.*

Recent studies (Lipson et al., 2017) suggest that all Oceanians have at least 25% of Papuan ancestry today. Furthermore, every attempt has been made to locate such an origin of the Papuan migration into Remote Oceania - so far believed to be in the N. Britain and N. Ireland (Lipson et al., 2017) . However, being poorly studied in detail, mainland Papuans was in most studies considered as a single entity, and sometimes divided into HLs and LLs.

Our results confirm results by (Bergström et al., 2017) that all highlanders form a clade within a wider diversity of lowlanders and lack SEA admixture. A new result is a lack of gene flows between HLs and LLs. The HLs show all signs of a recent massive population expansion from a previously isolated source, driven probably by the adoption of farming and herding (Bergström et al., 2017). It is notable that nearly HLs speak languages of the TNG phylum. The western (Enga, South HLs) and eastern parts (East HLs, Madang HLs, Simbu HLs) of the Highlands are genetically distinct, which again confirms the results by (Bergström et al., 2017) and is expected due to a mountain barrier separating these regions.

Bergström et al. (2017) further commented that highlanders show a slightly higher affinity to the Sepik River populations in contrast to other lowlanders. Even archaeological evidence suggests that during Holocene contacts between the East Sepik region and the Highlands regions was noticeable (Swadling et al., 2008). Indeed, East Sepik LLs represent the closest sister-group for HLs according to our model.

It is notable that no European admixture has ever been detected in Papuan samples. (Bergström et al., 2017), having included 503 European and 489 South Asian individuals from the 1000 Genomes Project, found no European admixture in Papuans.

9. Conclusions

The aim of this thesis (to generate a detailed admixture graph for mainland Papuans) was fulfilled by testing thousands of alternative topologies. Bergström et al. (2017) concluded that HLs form a "tight" clade, that LLs are more diverse and differentially related to HLs. These results were based on interpretations of simple *f*-statistics.

Our final graph topology is in perfect agreement with these earlier 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 refrained from adding those groups onto the final model, which already includes intra-lowlander admixture events. Previous inferences about the Papuan pre-history were confirmed, but a much more detailed graph was constructed for both highlanders and lowlanders.

Although having revealed a serious part of its history, far more remains to be investigated in these variable Papuan regions; and hopefully early will be.

10. List of acronyms

ABO	ABO-antigen-presence-based blood group system
AMH	Anatomically modern humans
BNG	British New Guinea
dbSNP	The Single Nucleotide Polymorphism Database
HapMap	International HapMap Project
HEVR	Human endogenous retrovirus
HL	Highlanders
HLA-B	Human Leukocyte Antigen B
HUGO	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

11. List of 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

12. References

Aghakhanian, F., Yunus, Y., Naidu, R., Jinam, T., Manica, A., Hoh, B. P., & Phipps, M. E. (2015). Unravelling the genetic history of Negritos and Indigenous populations of Southeast Asia. *Genome Biology and Evolution*, *7*(5), 1206–1215. https://doi.org/10.1093/gbe/evv065

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Allen, J. (2001). Australia and New Guinea, Archaeology of. *International Encyclopedia of the Social & Behavioral Sciences*, 952–956. https://doi.org/10.1016/b0-08-043076-7/02035-0

Anderson, S. (1981). Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Research*, *9*(13), 3015–3027. https://doi.org/10.1093/nar/9.13.3015

Ashraf, Q., & Galor, O. (2013). The "Out of Africa" hypothesis, human genetic diversity, and comparative economic development. *American Economic Review*, *103*(1), 1–46. https://doi.org/10.1257/aer.103.1.1

Barton, T. F. (1965). The Island of New Guinea. *Journal of Geography*, *64*(7), 308–309. https://doi.org/10.1080/00221346508985146

Becht, E., McInnes, L., Healy, J., Dutertre, C. A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., & Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, *37*(1), 38–47. https://doi.org/10.1038/nbt.4314

Belmont, J. W., Boudreau, A., Leal, S. M., Hardenbol, P., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Gao, Y., Hu, H., Hu, W., Li, C., Lin, W., Liu, S., Pan, H., Tang, X., Wang, J., Wang, W., … Stewart, J. (2005). A haplotype map of the human genome. *Nature*, *437*(7063), 1299–1320. https://doi.org/10.1038/nature04226

Bergström, A., Oppenheimer, S. J., Mentzer, A. J., Auckland, K., Robson, K., Attenborough, R., Alpers, M. P., Koki, G., Pomat, W., Siba, P., Xue, Y., Sandhu, M. S., & Tyler-Smith, C. (2017). A Neolithic expansion, but strong genetic structure, in the independent history of New Guinea. *Science*, *357*(6356), 1160–1163. https://doi.org/10.1126/science.aan3842

Bourke, R. M., & Harwood, T. (2009). Food and Agriculture in Papua New Guinea. *Food and Agriculture in Papua New Guinea*. https://doi.org/10.22459/fapng.08.2009

Bray, M., & Smith, P. (1985). *Education and Social Stratification in Papua New Guinea*. Longman Cheshire.

Brown, P., Sutikna, T., Morwood, M. J., Soejono, R. P., Jatmiko, Saptomo, E. W., & Due, R. A. (2004). A new small-bodied hominin from the Late Pleistocene of Flores, Indonesia. *Nature*, *431*(7012), 1055–1061. https://doi.org/10.1038/nature02999

Campbell, L., & Poser, W. J. (2008). Language classification: History and method. *Language Classification: History and Method*, *9780521880*(January 2008), 1–536. https://doi.org/10.1017/CBO9780511486906

Cavalli-Sforza, L., & Edwards, A. W. F. (1963). *Analysis of human evolution*. *29*(1965), 923–933.

Clark, G.J., Anderson, A. J., & Vunidilo, T. (2000). The archaeology of Lapita dispersal in Oceania. In P. Book (Ed.), *Papers from the Fourth Lapita Conference* (Vol. 6, Issue 2). Pandanus Books, Research School of Pacific and Asian Studies, The Australian National University, Canberra ACT 0200 Australia.

Clark, P. U., Dyke, A. S., Shakun, J. D., Carlson, A. E., Clark, J., Wohlfarth, B., Mitrovica, J. X., Hostetler, S. W., & McCabe, A. M. (2009). The Last Glacial Maximum. *Science*, *325*(5941), 710–714. https://doi.org/10.1126/science.1172873

Demeter, F., Shackelford, L., Westaway, K., Barnes, L., Duringer, P., Ponche, J. L.,

Dumoncel, J., Sénégas, F., Sayavongkhamdy, T., Zhao, J. X., Sichanthongtip, P., Patole-Edoumba, E., Dunn, T., Zachwieja, A., Coppens, Y., Willerslev, E., & Bacon, A. M. (2017). Early modern humans from tam pà ling, laos fossil review and perspectives. *Current Anthropology*, *58*(December), S527–S538. https://doi.org/10.1086/694192

Denaro, M., Blanc, H., Johnson, M. J., Chen, K. H., Wilmsen, E., Cavalli-Sforza, L. L., & Wallace, D. C. (1981). Ethnic variation in Hpa I endonuclease cleavage patterns of human mitochondrial DNA. *Proceedings of the National Academy of Sciences of the United States of America*, *78*(9 II), 5768–5772. https://doi.org/10.1073/pnas.78.9.5768

Denham, T., & Haberle, S. (2008). Agricultural emergence and transformation in the Upper Wahgi valley, Papua New Guinea, during the Holocene: Theory, method and practice. *Holocene*, *18*(3), 481–496. https://doi.org/10.1177/0959683607087936

Diamond, J. (1997). Guns, germs & steel: The fate of human societies. In *Guns, Germs & Steel: The Fate of Human Societies*. https://doi.org/10.4324/9781912128273

Diaz-Papkovich, A., Anderson-Trocmé, L., Ben-Eghan, C., & Gravel, S. (2019). UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLoS Genetics*, *15*(11), 1–24. https://doi.org/10.1371/journal.pgen.1008432

Duggan, A. T., Evans, B., Friedlaender, F. R., Friedlaender, J. S., Koki, G., Merriwether, D. A., Kayser, M., & Stoneking, M. (2014). Maternal history of oceania from complete mtDNA genomes: Contrasting ancient diversity with recent homogenization due to the Austronesian expansion. *American Journal of Human Genetics*, *94*(5), 721–733. https://doi.org/10.1016/j.ajhg.2014.03.014

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., & Foll, M. (2013). Robust Demographic Inference from Genomic and SNP Data. *PLoS Genetics*, *9*(10). https://doi.org/10.1371/journal.pgen.1003905

Flegontov, P., Altınışık, N. E., Changmai, P., Rohland, N., Mallick, S., Adamski, N., Bolnick, D. A., Broomandkhoshbacht, N., Candilio, F., Culleton, B. J., Flegontova, O., Friesen, T. M., Jeong, C., Harper, T. K., Keating, D., Kennett, D. J., Kim, A. M., Lamnidis, T. C., Lawson, A. M., … Schiffels, S. (2019). Palaeo-Eskimo genetic ancestry and the peopling of Chukotka and North America. *Nature*, *570*(7760), 236–240. https://doi.org/10.1038/s41586-019-1251-y

Flegontov, Pavel, Altınışık, N. E., Changmai, P., Rohland, N., Adamski, N., Bolnick, D. A., Broomandkhoshbacht, N., Candilio, F., Culleton, B. J., Flegontova, O., Friesen, T. M., Jeong, C., Harper, T. K., Keating, D., & Kennett, D. J. (2019). *Paleo-Eskimo genetic ancestry and the peopling of Chukotka and North America*. *570*(7760), 236–240. https://doi.org/10.1038/s41586-019-1251-y.Paleo-Eskimo

Foley, B. (2003). *More on Papuan Languages*.

Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., … Stewart, J. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, *449*(7164), 851–861. https://doi.org/10.1038/nature06258

Friedlaender, J. S., Friedlaender, F. R., Reed, F. A., Kidd, K. K., Kidd, J. R., Chambers, G. K., Lea, R. A., Loo, J. H., Koki, G., Hodgson, J. A., Merriwether, D. A., & Weber, J. L. (2008). The genetic structure of Pacific Islanders. *PLoS Genetics*, *4*(1), 0173–0190. https://doi.org/10.1371/journal.pgen.0040019

Gilliam, A. (1988). *Anthropology, geopolitics, and Papua New Guinea*. 37–51.

Glusman, G., Cox, H. C., & Roach, J. C. (2014). Whole-genome haplotyping approaches and genomic medicine. *Genome Medicine*, *6*(9), 1–16. https://doi.org/10.1186/s13073-014-0073-7

Gray, R. D., & Atkinson, Q. D. (2003). Language-tree divergence times support the

Anatolian theory of Indo-European origin. *Department OfPsychology, University OfAuckland, Private Bag 92019, Auckland 1020, New Zealand*, *3*(7), 1–18. https://doi.org/10.1029/2001gc000192

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H. Y., Hansen, N. F., Durand, E. Y., Malaspinas, A. S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H. A., … Pääbo, S. (2010). A draft sequence of the neandertal genome. *Science*, *328*(5979), 710–722. https://doi.org/10.1126/science.1188021

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K., Fu, Q., Mittnik, A., Bánffy, E., Economou, C., Francken, M., Friederich, S., Pena, R. G., Hallgren, F., Khartanovich, V., … Reich, D. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*, *522*(7555), 207–211. https://doi.org/10.1038/nature14317

Haberle, S. G., & David, B. (2004). Climates of change: Human dimensions of Holocene environmental change in low latitudes of the PEPII transect. *Quaternary International*, *118–119*, 165–179. https://doi.org/10.1016/S1040-6182(03)00136-8

HapMap. (2003). The International HapMap project. *Nature*, *426*, 789–796. https://doi.org/10.1007/978-1-4419-9863-7_100710

Harney, É., Patterson, N., Reich, D., & Wakeley, J. (2020). Assessing the Performance of qpAdm: A Statistical Tool for Studying Population Admixture. *BioRxiv*, 2020.04.09.032664. https://doi.org/10.1101/2020.04.09.032664

Harvey, R. G. (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, Biological Sciences*, *268*(893), 279–292. https://doi.org/10.1098/rstb.1974.0031

Hellenthal, G., Busby, G. B. J., Band, G., Wilson, J. F., Capelli, C., Falush, D., & Myers, S. (2014). A genetic atlas of human admixture history. *Science*, *343*(6172), 747–751. https://doi.org/10.1126/science.1243518.A

Heywood, P. F. (1983). Growth and Nutrition in Papua New Guinea. *Journal of Human Evolution*, *12*, 133–143. https://doi.org/10.1016/S0047-2484(83)80018-9

Hope, G. S., & Haberle, S. G. (2005). The history of the human landscapes of New Guinea. *Pacific Linguistics, Papuan Pasts: Cultural, Linguistic and Biological Histories of Papuan-Speaking Peoples*, 541–554. http://palaeoworks.anu.edu.au/publ2005.html%0Ahttp://palaeoworks.anu.edu.au/publ2005.html%0Ahttps://www.academia.edu/17017355/18_The_history_of_the_human_landscapes_of_New_Guinea

Hurst, G. D. D., & Jiggins, F. M. (2005). Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: The effects of inherited symbionts. *Proceedings of the Royal Society B: Biological Sciences*, *272*(1572), 1525–1534. https://doi.org/10.1098/rspb.2005.3056

Jacobs, G. S., Hudjashov, G., Saag, L., Kusuma, P., Darusallam, C. C., Lawson, D. J., Mondal, M., Pagani, L., Ricaut, F. X., Stoneking, M., Metspalu, M., Sudoyo, H., Lansing, J. S., & Cox, M. P. (2019). Multiple Deeply Divergent Denisovan Ancestries in Papuans. *Cell*, *177*(4), 1010-1021.e32. https://doi.org/10.1016/j.cell.2019.02.035

JICA (Japan International Cooperation Agency). (2002). Country Profile on Environment Planning and Evaluation Department Japan International Cooperation Agency. *Environment*, *February*.

Jobling, M., Hollox, E., Hurles, M., Kivisild, T., & Tyler-Smith, C. (2014). *Human Evolutionary Genetics* (Second). Garland Science, Taylor & Francis Group, LLC This.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis : a review and recent

developments. *Phil.Trans.R.Soc.A*, *374*(20150202), 1–16.
Kamm, J., Terhorst, J., Durbin, R., & Song, Y. S. (2019). Efficiently Inferring the Demographic History of Many Populations With Allele Count Data. *Journal of the American Statistical Association*, *0*(0), 1–16. https://doi.org/10.1080/01621459.2019.1635482
Kayser, M. (2010). The Human Genetic History of Oceania: Near and Remote Views of Dispersal. *Current Biology*, *20*(4), R194–R201. https://doi.org/10.1016/j.cub.2009.12.004
Kayser, M., Brauer, S., Weiss, G., Schiefenhövel, W., Underhill, P., Shen, P., Oefner, P., Tommaseo-Ponzetta, M., & Stoneking, M. (2003). Reduced Y-chromosome, but not mitochondrial DNA, diversity in human populations from West New Guinea. *American Journal of Human Genetics*, *72*(2), 281–302. https://doi.org/10.1086/346065
Kumar, S., Banks, T. W., & Cloutier, S. (2012). SNP discovery through next-generation sequencing and its applications. *International Journal of Plant Genomics*, *2012*(June). https://doi.org/10.1155/2012/831460
Lahr, M. M., & Foley, R. (1994). Multiple dispersals and modern human origins. *Evolutionary Anthropology: Issues, News, and Reviews*, *3*(2), 48–60. https://doi.org/10.1002/evan.1360030206
Lander, E. S., Linton, L. M., Birren, B. B., Nusbaum, C., & Zody, M. C. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*, 860–921. https://doi.org/10.1038/35087627
Lawson, D. J., van Dorp, L., & Falush, D. (2018). A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature Communications*, *9*(1), 1–11. https://doi.org/10.1038/s41467-018-05257-7
Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., & Mallick, S. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, *513*(7518), 409–413. https://doi.org/10.1126/science.1249098.Sleep
Lipson, M., Cheronet, O., Mallick, S., Rohland, N., Oxenham, M., Pietrusewsky, M., Pryce, T. O., Willis, A., Matsumura, H., Buckley, H., Domett, K., Nguyen, G. H., Trinh, H. H., Kyaw, A. A., Win, T. T., Pradier, B., Broomandkhoshbacht, N., Candilio, F., Changmai, P., … Reich, D. (2018). Ancient genomes document multiple waves of migration in Southeast Asian prehistory. *Science*, *361*(6397), 92–95. https://doi.org/10.1126/science.aat3188
Lipson, M., Reich, D., & Townsend, J. P. (2017). A working model of the deep relationships of diverse modern human genetic lineages outside of Africa. *Molecular Biology and Evolution*, *34*(4), 889–902. https://doi.org/10.1093/molbev/msw293
List, J. M., Greenhill, S. J., & Gray, R. D. (2017). The Potential of Automatic Word Comparison for Historical Linguistics. *PLoS ONE*, *12*(1), 1–18. https://doi.org/10.1371/journal.pone.0170046
Lohmann, K., & Klein, C. (2014). Next Generation Sequencing and the Future of Genetic Diagnosis. *Neurotherapeutics*, *11*(4), 699–707. https://doi.org/10.1007/s13311-014-0288-8
Ma, X., Shao, Y., Tian, L., Flasch, D. A., Mulder, H. L., Edmonson, M. N., Liu, Y., Chen, X., Newman, S., Nakitandwe, J., Li, Y., Li, B., Shen, S., Wang, Z., Shurtleff, S., Robison, L. L., Levy, S., Easton, J., & Zhang, J. (2019). Analysis of error profiles in deep next-generation sequencing data. *Genome Biology*, *20*(1), 1–15. https://doi.org/10.1186/s13059-019-1659-6
Malaspinas, A. S., Westaway, M. C., Muller, C., Sousa, V. C., Lao, O., Alves, I., Bergström, A., Athanasiadis, G., Cheng, J. Y., Crawford, J. E., Heupink, T. H., MacHoldt, E., Peischl, S., Rasmussen, S., Schiffels, S., Subramanian, S., Wright, J. L., Albrechtsen, A.,

Barbieri, C., … Willerslev, E. (2016). A genomic history of Aboriginal Australia. *Nature*, *538*(7624), 207–214. https://doi.org/10.1038/nature18299
Malcolm, L. A. (1969). Growth and development of the Kaiapit children of the Markham Valley, New Guinea. *American Journal of Physical Anthropology*, *31*(1), 39–51. https://doi.org/10.1002/ajpa.1330310106
Malcolm, L. A. (1970). Growth and development in New Guinea - a study of the Bundi people of the Madang district. *Institute of Human Biology of PNG*, *1*, 105.
Mallick, S., Li, H., Lipson, M., & Mathieson, I. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, *538*(7624), 201–206. https://doi.org/10.1016/j.physbeh.2017.03.040
Margulis, L. (1972). Origin of Eukaryotic Cells. Evidence and Research Implications for a Theory of the Origin and Evolution of Microbial, Plant, and Animal Cells on the Precambian Earth. *The Bryologist*, *75*(1), 115. https://doi.org/10.2307/3241538
Mason, A. S. (2013). Molecular cytogenetics. *Biotechnology of Crucifers*, *9781461477*, 13–22. https://doi.org/10.1007/978-1-4614-7795-2-2
Matisoo-Smith, E. (2015). Ancient DNA and the human settlement of the Pacific: A review. *Journal of Human Evolution*, *79*, 93–104. https://doi.org/10.1016/j.jhevol.2014.10.017
Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Biotechnology (Reading, Mass.)*, *74*(2), 560–564.
McColl, H., Racimo, F., Vinner, L., Demeter, F., Gakuhari, T., Moreno-Mayar, V., van Driem, G., Wilken, U. G., Seguin-Orlando, A., de la Fuente Castro, C., Wasef, S., Shoocongdej, R., Souksavatdy, V., Sayavongkhamdy, T., Saidin, M. M., Allentoft, M. E., Sato, T., Malaspinas, A. S., Aghakhanian, F. A., … 31, 32, Jean-Luc Ponche33, Laura Shackelford34, Elise Patole-Edoumba35, Anh Tuan Nguyen18, Bérénice Bellina-Pryce36, Jean-Christophe Galipaud37, Rebecca Kinaston38, 39, Hallie Buckley38, Christophe Pottier40, Simon Rasmussen41, Tom Higham29, Robert A. Fol, E. W. (2018). *The prehistoric peopling of Southeast Asia*. *92*(July), 88–92.
McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, *3*(29), 861. https://doi.org/10.21105/joss.00861
McVean, G. (2009). A genealogical interpretation of principal components analysis. *PLoS Genetics*, *5*(10). https://doi.org/10.1371/journal.pgen.1000686
Menozzi, P., Piazza, a., & Cavalli-Sforza, L. (1978). of Human Gene Synthetic Frequencies Europeans. *Science (New York, N.Y.)*, *201*(4358), 786–791. https://doi.org/10.1126/science.356262
Meyer, K. D., Saletore, Y., Zumbo, P., Elemento, O., Mason, C. E., & Jaffrey, S. R. (2012). Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons. *Cell*, *149*(7), 1635–1646. https://doi.org/10.1016/j.cell.2012.05.003
Moffatt, A. (2012). *Papua New Guinean Cultural Profile*. 1–25.
Mona, S., Tommaseo-Ponzetta, M., Brauer, S., Sudoyo, H., Marzuki, S., & Kayser, M. (2007). Patterns of Y-chromosome diversity intersect with the trans-New Guinea hypothesis. *Molecular Biology and Evolution*, *24*(11), 2546–2555. https://doi.org/10.1093/molbev/msm187
Morauta, L., Chowning, A., Gilliam, A. M., Hafer, F., Kayongo-Male, D., Levine, H. B., Maher, R. F., Nakleh, K., Simet, J. L., Strathern, A. J., Valentine, C. A., Valentine, B., & Waiko, J. (1979). Indigenous Anthropology in Papua New Guinea [and Comments and Reply]. *Current Anthropology*, *20*(3), 561–576. https://doi.org/10.1086/202325
Mörseburg, A., Pagani, L., Ricaut, F. X., Yngvadottir, B., Harney, E., Castillo, C., Hoogervorst, T., Antao, T., Kusuma, P., Brucato, N., Cardona, A., Pierron, D., Letellier, T., Wee, J., Abdullah, S., Metspalu, M., & Kivisild, T. (2016). Multi-layered population

structure in Island Southeast Asians. *European Journal of Human Genetics*, *24*(11), 1605–1611. https://doi.org/10.1038/ejhg.2016.60

Mörseburg, A., Pagani, L., Ricaut, F., Yngvadottir, B., Harney, E., Castillo, C., Hoogervorst, T., Antao, T., Kusuma, P., Brucato, N., Cardona, A., Pierron, D., Letellier, T., Wee, J., Metspalu, M., & Kivisild, T. (2016). *Multi-layered population structure in Island Southeast Asians*. *April*, 1605–1611. https://doi.org/10.1038/ejhg.2016.60

Mourant, A. E., Kopec, A. C., & Domaniewska-Sobczak, K. (1976). Book reviews: Book reviews. In *The Distribution of the Human Blood Groups and other Polymorphisms* (Second, Issue 3). https://doi.org/10.1111/1467-8489.00221

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., & Erlich, H. (1986). Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. 1986. *Biotechnology (Reading, Mass.)*, *24*(Table 1), 17–27.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., Stephens, M., & Bustamante, C. D. (2008). Genes mirror geography within Europe. *Nature*, *456*(7218), 98–101. https://doi.org/10.1038/nature07331

O'Connell, J. F., & 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. https://doi.org/10.1016/j.jas.2015.02.020

Olson, M. V. (1993). The human genome project. *Proceedings of the National Academy of Sciences of the United States of America*, *90*(10), 4338–4344. https://doi.org/10.1073/pnas.90.10.4338

Oppenheimer, S. (2009). The great arc of dispersal of modern humans: Africa to Australia. *Quaternary International*, *202*(1–2), 2–13. https://doi.org/10.1016/j.quaint.2008.05.015

Owen, R. (2000). Karl Landsteiner and the first human marker locus. *Genetics*, *155*(3), 995–998.

Pagani, L., Lawson, D. J., Jagoda, E., Mörseburg, A., Eriksson, A., Mitt, M., & Saag, L. (2017). *Europe PMC Funders Group Genomic analyses inform on migration events during the peopling of Eurasia*. *538*(7624), 238–242. https://doi.org/10.1038/nature19792.Genomic

Pagani, Lucia, Clair, P. A. S., Teshiba, T. M., Service, S. K., Fears, S. C., Araya, C., Araya, X., Bejarano, J., Ramirez, M., Castrillón, G., Gomez-Makhinson, J., Lopez, M. C., Montoya, G., Montoya, C. P., Aldana, I., Navarro, L., Freimer, D. G., Safaie, B., Keung, L. W., … Freimer, N. B. (2016). Genetic contributions to circadian activity rhythm and sleep pattern phenotypes in pedigrees segregating for severe bipolar disorder. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(6), E754–E761. https://doi.org/10.1073/pnas.1513525113

Palmer, B. (Ed.). (2018). *The Languages and Linguistics of the New Guinea Area*.

Parr, R. L., & Martin, L. H. (2012). Mitochondrial and nuclear genomics and the emergence of personalized medicine. *Human Genomics*, *6*(1), 1–8. https://doi.org/10.1186/1479-7364-6-3

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., & Reich, D. (2012). Ancient admixture in human history. *Genetics*, *192*(3), 1065–1093. https://doi.org/10.1534/genetics.112.145037

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, *2*(12), 2074–2093. https://doi.org/10.1371/journal.pgen.0020190

Pawley, A., & Hammarström, H. (2017). 2. The Trans New Guinea family. In *The Languages and Linguistics of the New Guinea Area* (Issue 1524). https://doi.org/10.1515/9783110295252-002

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space.

Philosophical Magazine, *2*, 559–572.
Pemberton, T. J., Sandefur, C. I., Jakobsson, M., & Rosenberg, N. A. (2009). Sequence determinants of human microsatellite variability. *BMC Genomics*, *10*, 31–36. https://doi.org/10.1186/1471-2164-10-612
Peter, B. M. (2016). Admixture, population structure, and f-statistics. *Genetics*, *202*(4), 1485–1501. https://doi.org/10.1534/genetics.115.183913
Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, *7*(5). https://doi.org/10.1371/journal.pone.0037135
Posth, C., Nakatsuka, N., Lazaridis, I., Skoglund, P., Mallick, S., Lamnidis, T. C., Rohland, N., Nägele, K., Adamski, N., Bertolini, E., Broomandkhoshbacht, N., Cooper, A., Culleton, B. J., Ferraz, T., Ferry, M., Furtwängler, A., Haak, W., Harkins, K., Harper, T. K., … Reich, D. (2018). Reconstructing the Deep Population History of Central and South America. *Cell*, *175*(5), 1185-1197.e22. https://doi.org/10.1016/j.cell.2018.10.027
Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., Filippo, C. De, Li, H., Mallick, S., Dannemann, M., Fu, Q., Kircher, M., Kuhlwilm, M., Lachmann, M., Meyer, M., Ongyerth, M., … Eichler, E. E. (2014). *HHS Public Access*. *505*(7481), 43–49. https://doi.org/10.1038/nature12886.The
Rasmussen, M., Guo, X., Wang, Y., Lohmueller, K. E., Rasmussen, S., Albrechtsen, A., Skotte, L., Lindgreen, S., Metspalu, M., Jombart, T., Kivisild, T., Zhai, W., Eriksson, A., Manica, A., Orlando, L., De La Vega, F. M., Tridico, S., Metspalu, E., Nielsen, K., … Willerslev, E. (2011). An aboriginal Australian genome reveals separate human dispersals into Asia. *Science*, *334*(6052), 94–98. https://doi.org/10.1126/science.1211177
Reich, D., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R., & Lander, E. S. (2001). Linkage disequilibrium in the human genome. *Nature*, *411*, 199–204. https://doi.org/10.1038/35075590
Reich, D., Thangaraj, K., & Patterson, N. (2009). Reconstructing Indian Population History. *Nature*, *461*(7263), 489–494. https://doi.org/10.1038/nature08365
Reich, David, Patterson, N., Campbell, D., 4, Arti Tandon1, 2, Ste´phane Mazieres3, 5, N. R., , Maria V. Parra3, 7, Winston Rojas3, 7, Constanza Duque3, 7, Natalia Mesa3, 7, L. F. G., Triana7, O., Blair7, S., Maestre7, A., Dib8, J. C., Claudio M. Bravi3, 9, G. B., , Daniel Corach10, Ta´bita Hu¨nemeier3, 11, Maria Ca´tira Bortolini11, F. M. S., Marı´a Luiza Petzl-Erler12, Victor Acun˜a-Alonzo13, Carlos Aguilar-Salinas14, Samuel Canizales-Quinteros15, 16, T. T.-L., Laura Riba15, Maricela Rodrı´guez-Cruz17, Mardia Lopez-Alarco´n17, Ramo´n Coral-Vazquez18, T. C.-C., Irma Silva-Zolezzi20{, Juan Carlos Fernandez-Lopez20, Alejandra V. Contreras20, G. J.-S., Maria Jose´ Go´mez-Va´zquez21, Julio Molina22, A´ ngel Carracedo23, Antonio Salas23, Carla Gallo24, G. P., David B. Witonsky25, Gorka Alkorta-Aranburu25, Rem I. Sukernik26, Ludmila Osipova27, Sardana A. Fedorova28, R. V., Mercedes Villena29, Claudia Moreau30, Ramiro Barrantes31, David Pauls32, Laurent Excoffier33, 34, G. B., Francisco Rothhammer35, Jean-Michel Dugoujon36, Georges Larrouy36, William Klitz37, Damian Labuda30, J. K., & Kenneth Kidd38, Anna Di Rienzo25, Nelson B. Freimer39, Alkes L. Price2, 40 & Andre´s Ruiz-Linares. (2012). Reconstructing Native American Population History. *Nature*, *488*(7411), 370–374. https://doi.org/:10.1038/nature11258. Reconstructing
Riley, I. D. (1983). Population change and distribution in Papua New Guinea: an epidemiological approach. *Journal of Human Evolution*, *12*(1), 125–132.

https://doi.org/10.1016/S0047-2484(83)80017-7
Riley, T. J. (1977). *Sahul: prehistoric studies*. *2*, 391–392. https://doi.org/10.1002/ajpa.1330490312
Ring, A., & Scragg, R. (1973). A demographic and social study of fertility in rural new guinea. *Journal of Biosocial Science*, *5*(1), 89–121. https://doi.org/10.1017/S002193200000897X
Ross, M. (2018). *For Pawley, Attenborough, Hide and Golson, eds,*. *June*.
Rudd, M. K., Wray, G. A., & Willard, H. F. (2006). The evolutionary dynamics of α-satellite. *Genome Research*, *16*(1), 88–96. https://doi.org/10.1101/gr.3810906
Sanger, F., & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, *94*(3), 441–448. https://doi.org/10.1016/0022-2836(75)90213-2
Schiefenhövel, W. (2014). *Human origin sites and the World Heritage Convention in Asia* (N. Sanz (Ed.)).
Schiffels, S., Haak, W., Paajanen, P., Llamas, B., Popescu, E., Loe, L., Clarke, R., Lyons, A., Mortimer, R., Sayer, D., Tyler-Smith, C., Cooper, A., & Durbin, R. (2016). Iron Age and Anglo-Saxon genomes from East England reveal British migration history. *Nature Communications*, *7*, 1–9. https://doi.org/10.1038/ncomms10408
Seeburg, P. H., Shine, J., Martial, J. A., Ullrich, A., Baxter, J. D., & Goodman, H. M. (1977). Nucleotide sequence of part of the gene for human chorionic somatomammotropin: Purification of DNA complementary to predominant mRNA species. *Cell*, *12*(1), 157–165. https://doi.org/10.1016/0092-8674(77)90193-3
Shaw, B., Field, J. H., Summerhayes, G. R., Coxe, S., Coster, A. C. F., Ford, A., Haro, J., Arifeae, H., Hull, E., Jacobsen, G., Fullagar, R., Hayes, E., & Kealhofer, L. (2020). Emergence of a Neolithic in highland New Guinea by 5000 to 4000 years ago. *Science Advances*, *6*(13), eaay4573. https://doi.org/10.1126/sciadv.aay4573
Shields, E. D., Décary, F., & Russell, A. D. (1986). Genetic distance: probing the origin of a Papua New Guinea isolate. *International Journal of Anthropology*, *1*(4), 307–321. https://doi.org/10.1007/BF02442060
Shiflett, A. M., & Johnson, P. J. (2010). Mitochondrion-Related Organelles in Eukaryotic Protists. *Annual Review of Microbiology*, *64*(1), 409–429. https://doi.org/10.1146/annurev.micro.62.081307.162826
Skoglund, P., & Mathieson, I. (2018). Ancient Human Genomics: The First Decade. *Annu. Rev. Genom. Hum. Genet*, *198*(April), 1–824. https://doi.org/10.1146/annurev-genom-083117
Skoglund, P., Posth, C., Sirak, K., Spriggs, M., Valentin, F., Bedford, S., Clark, G. R., Reepmeyer, C., Petchey, F., Fernandes, D., Fu, Q., Harney, E., Lipson, M., Mallick, S., Novak, M., Rohland, N., Stewardson, K., Abdullah, S., Cox, M. P., … Reich, D. (2016). Genomic insights into the peopling of the Southwest Pacific. *Nature*, *538*(7626), 510–513. https://doi.org/10.1038/nature19844
Skoglund, P., Posth, C., Sirak, K., Spriggs, M., Valentin, F., Bedford, S., Clark, G., Reepmeyer, C., Petchey, F., Fernandes, D., Fu, Q., Harney, E., Lipson, M., Mallick, S., Novak, M., Rohland, N., Stewardson, K., Abdullah, S., Cox, M. P., … Reich, D. (2017). Ancient Genomics and the Peopling of the Southwest Pacific. *Nature*, *538*(7626), 510–513. https://doi.org/10.1038/nature19844.Ancient
Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of Next Generation Sequencing technologies (and bioinformatics) in cancer. *Molecular Biology*, *122*(1), 1–15. https://doi.org/10.1002/cpmb.59.Overview
Soukup, M. (2010). Anthropology in Papua New Guinea: History and Continuities. *Anthropologia Integra*, *1*(2), 45. https://doi.org/10.5817/ai2010-2-45

Stoneking, M., Jorde, L. B., Bhatia, K., & Wilson, A. C. (1990). Geographic variation in human mitochondrial DNA from Papua New Guinea. *Genetics*, *124*(3), 717–733.

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. *Journal de La Société Des Océanistes*, *126–127*, 271–292. https://doi.org/10.4000/jso.2942

Tan, G., Opitz, L., Schlapbach, R., & Rehrauer, H. (2019). Long fragments achieve lower base quality in Illumina paired-end sequencing. *Scientific Reports*, *9*(1), 1–7. https://doi.org/10.1038/s41598-019-39076-7

Torroni, A., Schurr, T. G., Yang, C. C., Szathmary, E. J. E., Williams, R. C., Schanfield, M. S., Troup, G. A., Knowler, W. C., Lawrence, D. N., Weiss, K. M., & Wallace, D. C. (1992). Native American mitochondrial DNA analysis indicates that the Amerind and the Nadene populations were founded by two independent migrations. *Genetics*, *130*(1), 153–162.

Venter, C. J., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., … Zhu, X. (2001). The sequence of the human genome. *Science*, *291*(5507), 1304–1351. https://doi.org/10.1126/science.1058040

Vetter, J. (2006). Wallace's other line: Human biogeography and field practice in the eastern colonial tropics. *Journal of the History of Biology*, *39*(1), 89–123. https://doi.org/10.1007/s10739-005-6543-4

Vines, A. P., & Booth, P. B. (1963). *Highlanders of New Guinea and Papua: a blood group survey*.

Vries, L. De, Amsterdam, V. U., & Amsterdam, V. U. (2019). *The Greater Awyu language family of West Papua Special Issue 2012 Harald Hammarström & Wilco van den Heuvel (eds .) History , contact and classification of Papuan languages Part One*. *January 2012*.

Wakeley, J., Nielsen, R., Liu-Cordero, S. N., & Ardlie, K. (2001). The discovery of single-nucleotide polymorphisms—And inferences about human demographic history. *American Journal of Human Genetics*, *69*(6), 1332–1347. https://doi.org/10.1086/324521

Wallace, D. C. (2007). Why Do We Still Have a Maternally Inherited Mitochondrial DNA? Insights from Evolutionary Medicine. *Annual Review of Biochemistry*, *76*(1), 781–821. https://doi.org/10.1146/annurev.biochem.76.081205.150955

Waters, M. R. (2019). Late Pleistocene exploration and settlement of the Americas by modern humans. *Science*, *365*(6449). https://doi.org/10.1126/science.aat5447

Wollstein, A., Lao, O., Becker, C., Brauer, S., Trent, R. J., Nürnberg, P., Stoneking, M., & Kayser, M. (2010). Demographic history of oceania inferred from genome-wide data. *Current Biology*, *20*(22), 1983–1992. https://doi.org/10.1016/j.cub.2010.10.040

Papua New Guinea Economic Upadte, (2019).