# Chapter 3: The genetic history of Papua New Guinea

## 3.1 Introduction

The island of New Guinea constitutes the northern part of the Sahul continent, just south of the equator. Geographically and climatically it differs greatly from the very dry landscapes that make up much of Australia, being mostly covered in tropical rainforest, wetlands and grasslands. A mountain chain runs through the centre of the island, its highest peaks reaching over 4,000 meters but most of it situated at 1,000-2,000 meters. The archaeological evidence for human occupation in New Guinea is about as old as that for Australia, going back to ∼50 kya (Summerhayes, et al. 2010; O'Connell and Allen 2015), consistent with deriving from the same colonization event of the continent. Today, the island is politically split into two parts, the western half being part of Indonesia while the eastern part, plus a number of nearby islands of Melanesia, constitute the nation of Papua New Guinea (PNG). PNG is very culturally and linguistically diverse with approximately 850 different language groups, which is more than any other country and represents over 10% of all languages in the world. It is not known if this diversity is reflected in strong population structure on the genetic level.

New Guinea was one of the handful of places in the world where humans developed agriculture, in all cases in the Holocene, i.e. approximately the last 12 ky. The timing of this development in New Guinea is not very firmly established, but is estimated at approximately 10 kya (Denham and Barton 2006). Little is known about where within New Guinea agriculture originated, but the consensus is that it was somewhere in the highlands. Because of their elevation, the highlands are relatively cool and with plenty of rain and fertile soils. It has been hypothesized that the largest language family of New Guinea, the so called Trans-New Guinea (TNG) family, which is spoken across all of the highlands and large parts of the lowlands, spread alongside the spread of agriculture (Pawley 2005). While early influences from Southeast Asia cannot be completely ruled out, and there have been later external influences e.g. in the introduction of pigs and the sweet potato, it appears that the development of agriculture in New Guinea was indigenous and independent from developments in the rest of the world. There is therefore an opportunity to compare the population history of New Guinea with the histories of other parts of the world, in particular with respect to how it was shaped by the development of agriculture. Such a comparative view across independent histories might even allow us to say something about the "reproducibility" of human evolutionary trajectories.

In the last few thousand years, New Guinea saw an influx of people from Southeast Asia, leading to genetic admixture and the introduction of Austronesian languages. This is one of the world's

largest language families, spoken by many different groups across Southeast Asia and Oceania, and was spread by massive seaborne migrations by Southeast Asian agriculturalists (Duggan and Stoneking 2014; Skoglund, et al. 2016). In New Guinea, Austronesian languages are fairly restricted to certain coastal areas. Genetic studies so far have documented Southeast Asian admixture in the lowlands of New Guinea, but not in the highlands (Stoneking, et al. 1990). However, most studies have been small both in terms of the number of samples and the number of markers assayed, most being limited to the uniparental chromosomes, such that the extent of Southeast Asian admixture is not understood in any greater detail. In particular, it has arguably not been conclusively established if the highlands indeed have been isolated from non-New Guinean gene flow.

In this chapter, I primarily make use of two different datasets to study the population history of Papua New Guinea, both generated at the Wellcome Trust Sanger Institute. The first is a set of 39 high-coverage, whole genome sequences from Papua New Guinea, also used in Chapter 2. The second is a dataset of genome-wide array genotypes from 381 individuals from Papua New Guinea, generated specifically for this study and published in 2017 (Bergström, et al. 2017).

## 3.2 Array genotyping of Papua New Guineans

381 individuals were selected from a set of ∼800 PNG DNA samples collected in the early 1980's and stored at the University of Oxford, and genotyped on the Illumina Infinium Multi-Ethnic Global array. This array contains 1.78 million markers and its marker content has been ascertained in a less European-centric manner than most arrays commonly used in the past; however, less than 600,000 of these were polymorphic in this set of PNG samples. This is, however, still sufficient for high-resolution population genetic analyses. After quality control at the individual and site levels, 378 individuals and 529,137 variants remained for downstream analyses. The sample set is very comprehensive, covering approximately 85 different and geographically diverse language groups within PNG (Figure 3.1).

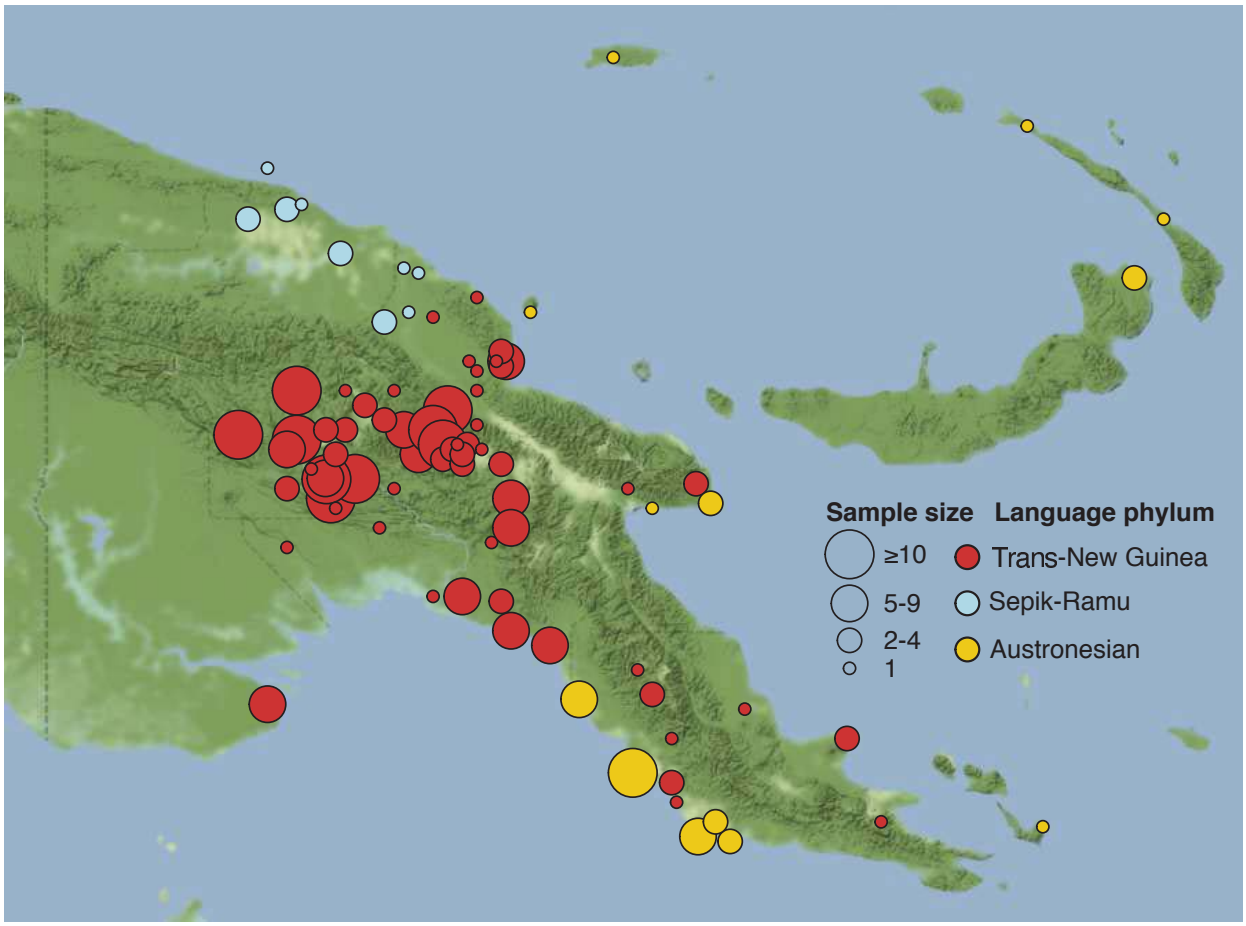


***Figure 3.1: Genotyped samples in Papua New Guinea.*** *Each sampled language group is represented by a circle, the area of which indicates the number of individuals sampled in the group and the colour of which represents the top-level language phylum/family. An additional 39 individuals are not included in this figure as their exact language group is not known or their parents are known to come from different language groups. The base map was obtained from Stamen (http://maps.stamen.com/, under Creative Commons BY 3.0. Data by OpenStreetMap, http://www.openstreetmap.org/, under Open Data Commons Open Database License (ODbL)).*

## 3.3 Genetic identity of the HGDP-CEPH Papua New Guinean individuals

The widely used HGDP-CEPH collection of DNA samples from diverse, worldwide populations (Cann, et al. 2002) contains 17 individuals from Papua New Guinea. When analysed in the context of the comprehensive set of array genotyped PNG samples, it is clear that these 17 individuals actually consist of two subsets with very different genetic affinities. One set has affinities to groups in the East Sepik region of the northern PNG lowlands, consistent with the geographical coordinates provided in the HGDP-CEPH metadata, and the other set has affinities to groups in the eastern highlands. This finding has important implications for the use of these samples in analyses, particularly as I make use of whole-genome sequencing data from 14 of them. In what follows I refer to the lowlander subset as "HGDP_L" and the highlander subset as "HGDP_H".

## 3.4 Southeast Asian admixture in PNG

The array genotype dataset reveals a very variable impact of Southeast Asian, or other non-New

Guinean, gene flow in PNG. Firstly, only two individuals displayed evidence of European ancestry, in sharp contrast with neighbouring Australia and consistent with the small impact of European colonialism on PNG. Secondly, there is a lack of Southeast Asian admixture in the highlands, as determined using ADMIXTURE as well as per-individual *D*-statistic tests of the form *D*(Yoruba, East Asian; Aboriginal Australian, PNG individual). Only four highlanders display East Asian ancestry in these analyses; however, principal component analyses of the relationships between highlanders and lowlanders show that all four of these individuals have lowlander affinities in the Papuan component of their genomes, such that they very likely reflect recent movement into the highlands from lowland groups. Consistently, analysis of the mitochondrial and Y chromosome variants included on the array showed that none of the sampled highlanders carried uniparental chromosomes of recent non-New Guinean origin. In summary, there is thus no evidence for Southeast Asian gene flow into the highlands, implying genetic independence of highlander ancestry from non-Sahul sources from the initial peopling of the continent until the present day (Figure 3.2).

Lastly, Southeast Asian admixture is present in all parts of the lowlands but in highly variable amounts across individuals and regions. Individuals speaking Austronesian languages have substantially higher amounts of Southeast Asian ancestry than those speaking non-Austronesian languages (mean of 38.7% vs 11.6%, $p = 1.4 \times 10^{-13}$, Wilcoxon rank sum test). Speakers of the Sepik-Ramu language family, which is an indigenous language family unrelated to the larger Trans-New Guinea family and spoken only in parts of the northern lowlands around the Sepik and Ramu rivers, display the lowest amounts of admixture (average of 4.3%). These results thus demonstrate that, while all coastal areas of New Guinea would have been accessible to the seafaring Southeast Asian migrants, the genetic impact was very variable.

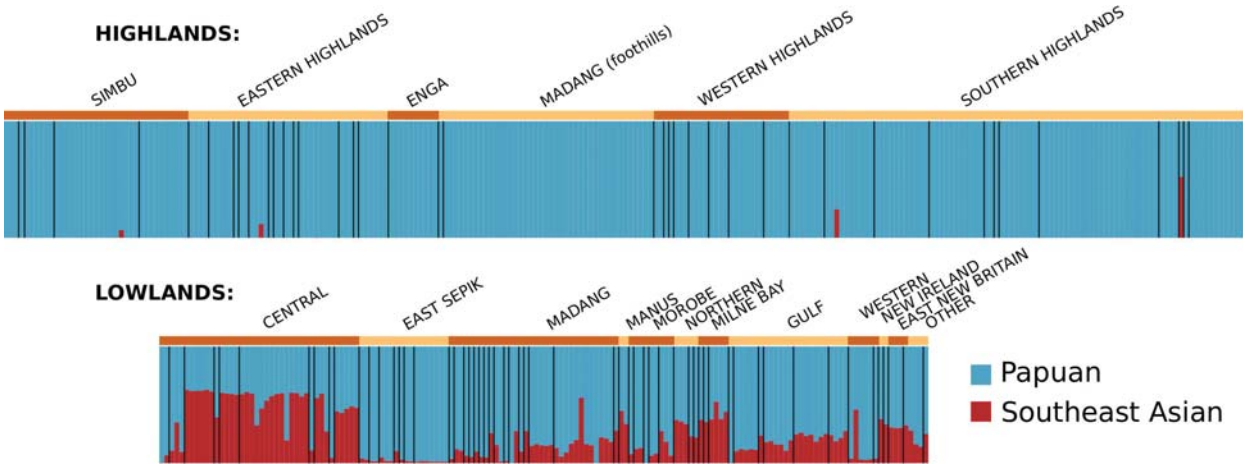


***Figure 3.2: Southeast Asian admixture in PNG.*** *ADMIXTURE was run at K=2 together with the 504 East Asian individuals from the 1000 Genomes Project (not displayed) to estimate the Papuan and Southeast Asian ancestry proportions across sampled PNG individuals. Individuals are grouped by language group, separated by vertical black lines, and then by province. The estimated ancestry proportions correlate strongly with those estimated using* $f_4$*-ratios.*

## 3.5 The relationship to Aboriginal Australians

As discussed in Chapter 2, the genetic separation between Aboriginal Australians and Papuans appears to have occurred relatively early, long before the geographical separation of Australia and New Guinea following rising sea levels. Furthermore, Aboriginal Australians across different regions of Australia displayed a largely uniform relationship to Papuans, implying shared ancestry across Australia after the separation from Papuans, and at most limited gene flow from New Guinea after that point. However, it is not known what corresponding situation is on the other side of the Torres Strait, i.e. if all Papuans are uniformly related to Aboriginal Australians, or if, for example, groups on the southern coast have higher affinity, e.g. due to gene flow or ancestral population structure. In the array genotype dataset, no individual has a directly visible higher affinity to Aboriginal Australians than the highlanders do, i.e. there are no significantly negative values of the statistic $D$(Yoruba,Aboriginal Australian;X,Highlander). However, the effect of the widespread Southeast Asian admixture present in most lowlander genomes is to decrease affinity to Aboriginal Australians, thereby counteracting the ability of a test like this to detect a higher affinity. I therefore performed analyses aiming at assessing affinity to Aboriginal Australians while accounting for the Southeast Asian admixture.

First, in a PCA constructed using only PNG highlanders, Aboriginal Australians and East Asians (using 1000 Genomes Han Chinese as a proxy for the Southeast Asian ancestry present in New Guinea), the position of a sample projected into the resulting space should be informative about its relationship to these three ancestral poles. Most individuals line up along the arc between the highlander and East Asian corners, suggesting that the Sahul component of their ancestry is not any closer to Aboriginal Australians than highlander ancestry is (Figure 3.3C). Deviating from this trend are the individuals from the large islands of New Britain and New Ireland, as well as individuals from Western Province (the southernmost samples in our dataset). This could reflect a closer relationship to Aboriginal Australians in these groups, and at least superficially appears to mirror such a relationship reported for the island population of Tonga, further out into the Pacific (Skoglund, et al. 2016).

Second, data from each sampled individual in turn was used to fit an admixture graph (Patterson, et al. 2012) modelling their ancestry as a mixture of one source related to PNG Highlanders and a second source related to East Asians (Figure 3.3A). The key aspect of this graph in the context of the relationship to Aboriginal Australians is that it shows all the Sahul-related ancestry of the modelled individual coming from the same lineage as PNG highlander ancestry, with Aboriginal Australians being a strict outgroup to this lineage. Any extra affinity to Aboriginal Australians would lead to a rejection of this simple model. This graph fit the data with no outlier ($|Z|>3$) $f$-statistics for any except three individuals. Two of the individuals for which the model did not fit had European admixture, and the last is a highlander who is an ancestry outlier relative to their language group. A similar admixture graph was also tested using whole-genome sequencing data from the

Simons Genome Diversity Project (Mallick, et al. 2016), representing an independent dataset, with the Bougainville Islanders (who are not present in the array dataset) as the test population (Figure 3.3B). The simple graph with Aboriginal Australians being an outgroup to the Sahul ancestry also fits the data for Bougainville Islanders.

Last, *D*-statistics were directly plotted against each other to allow visual examination of how East Asian admixture impacts affinity to Aboriginal Australians (Figure 3.3D). The statistic *D*(Yoruba,Vietnamese;X,Chinese) measures the amount of Sahul ancestry, as opposed to East Asian ancestry, and the statistic *D*(Yoruba,Aboriginal Australian;X,Highlander) measures the affinity to Aboriginal Australians. If the Sahul component in some individuals has a stronger affinity to Aboriginal Australians, we would expect them to depart from the trend line (i.e. a lower value of the latter *D*-statistic than expected given their value of the former *D*-statistic). No individuals departed from the trend line, suggesting a uniform relationship to Aboriginal Australians.

The lack of signal in the admixture graph and *D*-statistics analyses suggest that the pull of New Ireland and New Britain groups towards Aboriginal Australians in PCA might be an artefact, perhaps related to the likely fairly deep divergence of these populations from mainland populations. The signal reported for Tongan islanders (Skoglund, et al. 2016) thus seems to be absent from the populations sampled here, closer to the New Guinea mainland. The PCA position of the Western Province Southern Kiwai individuals, while intriguing given their geographical proximity to Australia, also has no support in the formal statistics. In summary, there is no strong evidence for any of the sampled groups displaying a closer relationship to Aboriginal Australians than anyone else. This mirrors the uniform relationship to Papuans found among Aboriginal Australians, and sets Sahul apart from many other parts of the world where genetic gradients across geographical space tends to be the norm.

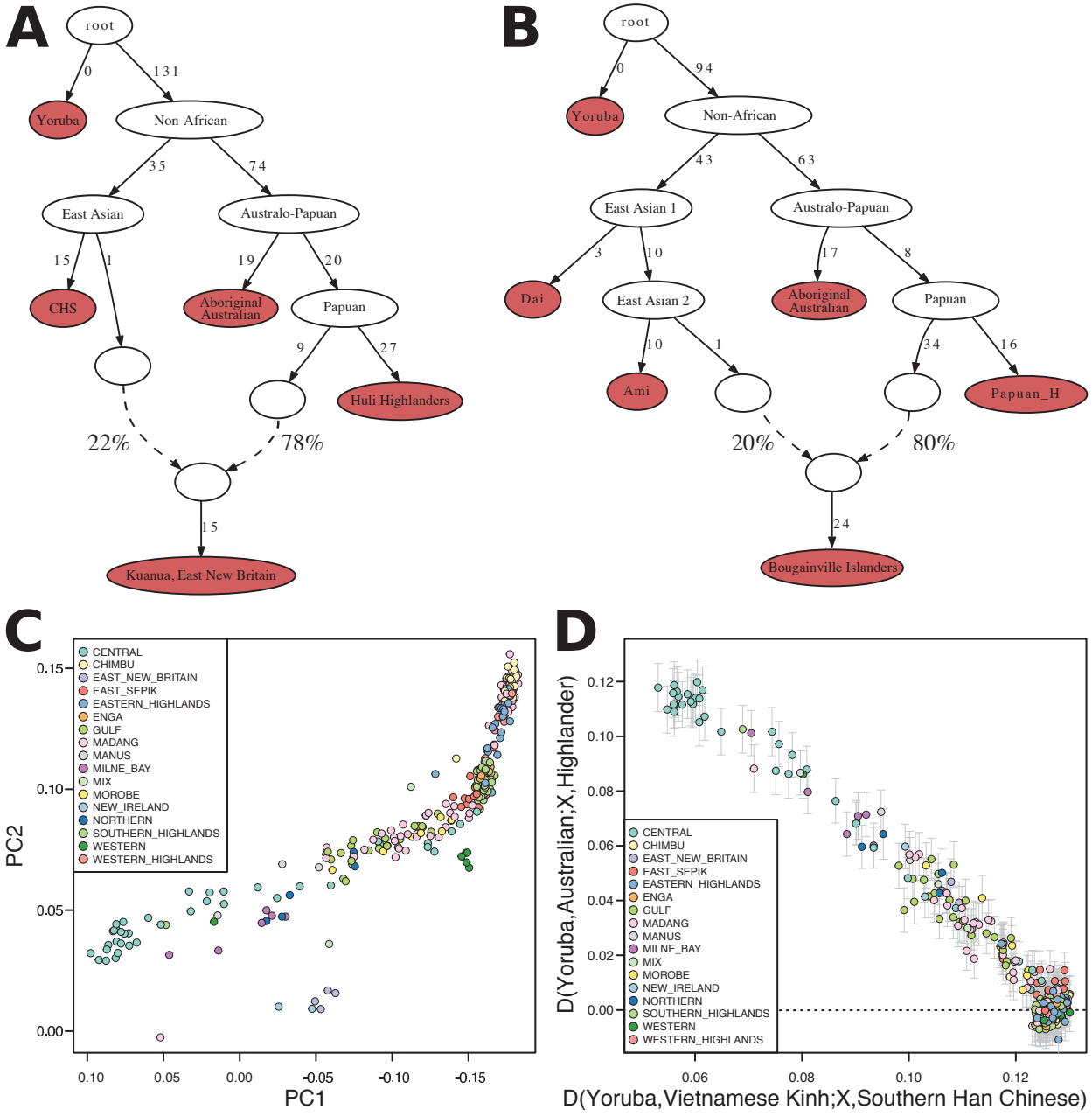


***Figure 3.3: The relationship of PNG individuals to Aboriginal Australians.*** *A) An admixture graph, modelling the ancestry of an individual as a mixture of one source related to PNG highlanders and one source related to East Asians, fits the data for all except three outliers. B) A similar admixture graph using data from the SGDP fits the Bougainville Island population. C) A principal components analysis where only two PNG Highlanders, two Aboriginal Australians and two East Asian individuals are used to define the space, defining ancestral poles in the top-right, bottom-right and left part of the plot, respectively (but not displayed on the plot themselves). The rest of the individuals are then projected into this space. D) Two* D*-statistics plotted against each other, the one on the horizontal axis measuring the amount of Sahul, as opposed to East Asian, ancestry, and the one on the vertical axis measuring affinity to Aboriginal Australians.*

## 3.6 Local ancestry inference in PNG

The widespread Southeast Asian admixture present in the lowlands of PNG would confound many analyses, including comparisons between different groups within PNG. The local ancestry classification and subsequent masking of haplotypes has proven successful in recently admixed populations such as Latin- and African Americans where the ancestry blocks are still long (Brisbin, et al. 2012), but has not been widely applied in PNG or Melanesian populations where the admixture is likely to be at least 3,000 years old. Furthermore, current reference panels for haplotype phasing, which is a requirement for most local ancestry inference methods, do not include haplotypes from this part of the world. However, phasing haplotypes against the very large HRC panel (McCarthy, et al. 2016) and inferring local ancestry using RFMix (Maples, et al. 2013), using unadmixed PNG highlanders and the 1000 Genomes East Asians as the two reference populations, proved surprisingly successful in the PNG array genotype dataset. Validation experiments in which highlanders and East Asians were held out from the reference panels and then subjected to the same local ancestry inference indicated an ancestry misclassification rate of less than 0.5%. Furthermore, the overall ancestry proportions obtained by summing the lengths of all haplotypes of each ancestry correlated strongly with those obtained from global ancestry estimation methods (Pearson correlation between RFMix and ADMIXTURE = 0.997, between RFMix and $f_4$-ratio = 0.982). The masked genotypes were therefore used in many downstream analyses, thereby avoiding the confounding effects of Southeast Asian admixture to a large extent.

## 3.7 Population structure in PNG

The strongest genetic separation within PNG appears to be that between groups on the mainland and those on the large islands of the Bismarck Archipelago, New Ireland and New Britain, consistent with previous studies hinting at a potentially fairly old split between these (Friedlaender, et al. 2008). Within the highlands, there is very clear clustering into a western cluster, an eastern cluster and one cluster corresponding to a small set of Angan language groups living in the south-eastern highlands, the latter likely a case of genetic isolation. Within the lowlands, there is clear separation between the south coast and north coast (Figure 3.4).

I performed various analyses to determine the nature of the relationship between highlanders and lowlanders. When projected into a PCA space constructed using only highlander genotypes, all lowlander individuals project into largely the same part of the plot (Figure 3.4A). It is not the case that, for example, northern lowlanders project closer to northern highlanders, and southern lowlanders closer to southern highlanders. In the inverse experiment, all highlanders similarly project into largely the same part of the space constructed using only lowlander genotypes (Figure 3.4B). In both cases there are a handful of outlier individuals who are drawn towards particular parts of the space, but these likely reflect very recent admixture between the highlands and the lowlands (in some cases confirmed by the documentation on the sample origins). It thus appears that population structure in the highlands is largely independent of that in the lowlands.

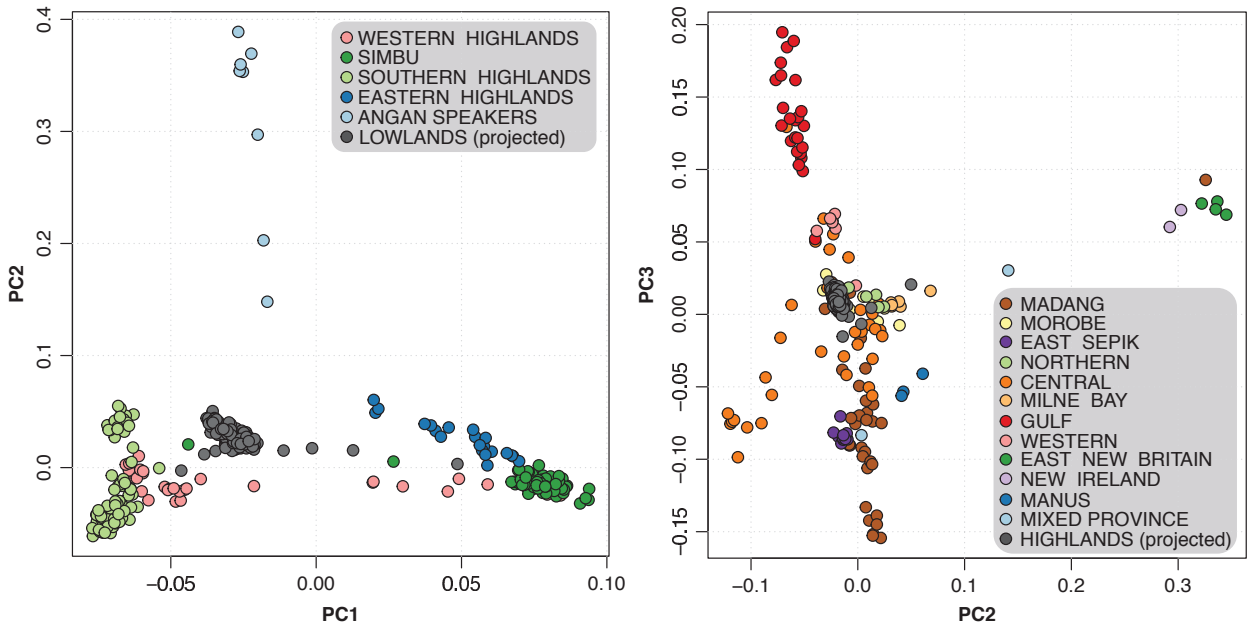


***Figure 3.4: Population structure of the highlands and the lowlands of PNG.*** *A) A PCA constructed using only highlander genotypes reveals a strong division between eastern and western groups (PC1), as well as separation of Angan speaking groups (PC2). When projected into this space, all lowlanders group largely uniformly, except a few outliers. B) A PCA constructed using only lowlander genotypes reveals a strong separation between groups on the mainland and those on the large islands of New Britain and New Ireland (PC2), as well as separation between the north and the south coast (PC3). PC1 corresponds to Southeast Asian admixture (not shown). When projected into this space, all highlanders group largely together, except a few outliers.*

To further study the relationships between highlanders and lowlanders, four sets of all-against-all *D*-statistics were computed and visualized using Q-Q (quantile-quantile) plots, comparing the Z-scores obtained to those expected by chance under a normal distribution with mean 0, given the number of tests performed. These were calculated on data locally masked for East Asian ancestry (and excluding the PCA outliers mentioned above), in all cases with the African Yoruba as an outgroup:

- *D*(Yoruba,Lowlander;Highlander1,Highlander2) (Figure 3.5A): this statistic asks for each lowlander group if it shares more with one highlander group than another. While some values of this statistic reach the seemingly significant $|Z|>4$, the close fit to the diagonal line in the Q-Q plot demonstrates that this is expected by chance given the large number of tests performed. The results thus give no evidence for any lowlander group having a higher affinity to one highlander group over another.

- *D*(Yoruba,Highlander;Lowlander1,Lowlander2) (Figure 3.5B): this statistic asks for each highlander group if it shares more with one lowlander group than another. The deviation from the diagonal in the Q-Q plot indicates that this is sometimes the case. Most of the signal is driven by highlanders being closer to mainland lowland populations than to the divergent New Ireland and New Britain island populations, but some signal remains even when excluding these.

- *D*(Yoruba,Highlander1;Highlander2,Lowlander) (Figure 3.5C): this statistic asks for each

highlander group if it is closer to another highlander group than to a lowlander group. The massive shift towards negative values indicates strong highlander genetic unity to the exclusion of lowlanders, and despite the large numbers of tests performed there is not a single test in which a highlander group is significantly closer to a lowlander than to another highlander group (largest Z=2.69).

- *D*(Yoruba,Lowlander1;Lowlander2,Highlander) (Figure 3.5D): this statistic asks for each lowlander group if it is closer to another lowlander group than to a highlander group. The deviation from the diagonal in the Q-Q plot indicates that this is not always the case. In other words, while there is highlander genetic unity, there is no lowlander genetic unity. Much of the signal is driven by most lowlanders being closer to highlanders than to the divergent New Ireland and New Britain island populations, but considerable signal remains even when excluding these. This is driven mainly by northern lowland groups being closer to highlanders than to southern lowland groups (example: *D*(Yoruba,Wagi Northern Lowlanders;Waima Southern Lowlanders,Aiya Highlanders) = 0.0095, Z = 3.074), and southern lowland groups being closer to highlanders than to northern lowland groups (example: *D*(Yoruba,Waima Southern Lowlanders;Wagi Northern Lowlanders,Aiya Highlanders) = 0.0105, Z = 3.576).

Lastly, following up on the observation that highlanders are not equally similar to all lowlander groups, outgroup $f_3$-statistics demonstrate that the set of groups that highlanders are the most similar to are those from the East Sepik area in the north-western lowlands (Figure 3.5F). This is surprising from a linguistic point of view, as it is the only sampled area where the widespread Trans-New Guinea language family is not spoken, instead being dominated by languages from the independent Sepik-Ramu family. There is, however, archaeological evidence for the transfer of items between the highlands and the Sepik region during the Holocene (Swadling, et al. 2008).

In summary, these results reveal a striking picture where all highlanders, regardless of geographic location, can be described as being a clade to the exclusion of lowlanders. As a telling example of this, Gende speakers who live on the very northern edge of the highlands are more similar to all other highlander people than to Sop speakers living in the lowlands just 40 km to the northeast (Figure 3.5E). This very sharp division between highland and lowland groups departs dramatically from the gradual isolation-by-distance patterns that are typically seen in human populations (Lao, et al. 2008; Novembre, et al. 2008). Possible explanations include a recent expansion-and-replacement episode that homogenized the ancestry of all highlanders, or possibly a more subtle, long-term process where gene-flow has remained high within the highlands but very limited between the highlands and the lowlands. Lastly, it can be noted that it's possible that there is a lack of statistical power with the current dataset to for example reject the clade-like status of highlanders (e.g. in tests of the form D(Outgroup,Lowlander;Highander 1;Highlander 2)), and that denser genotypes and larger population samples would potentially allow such rejections. However,

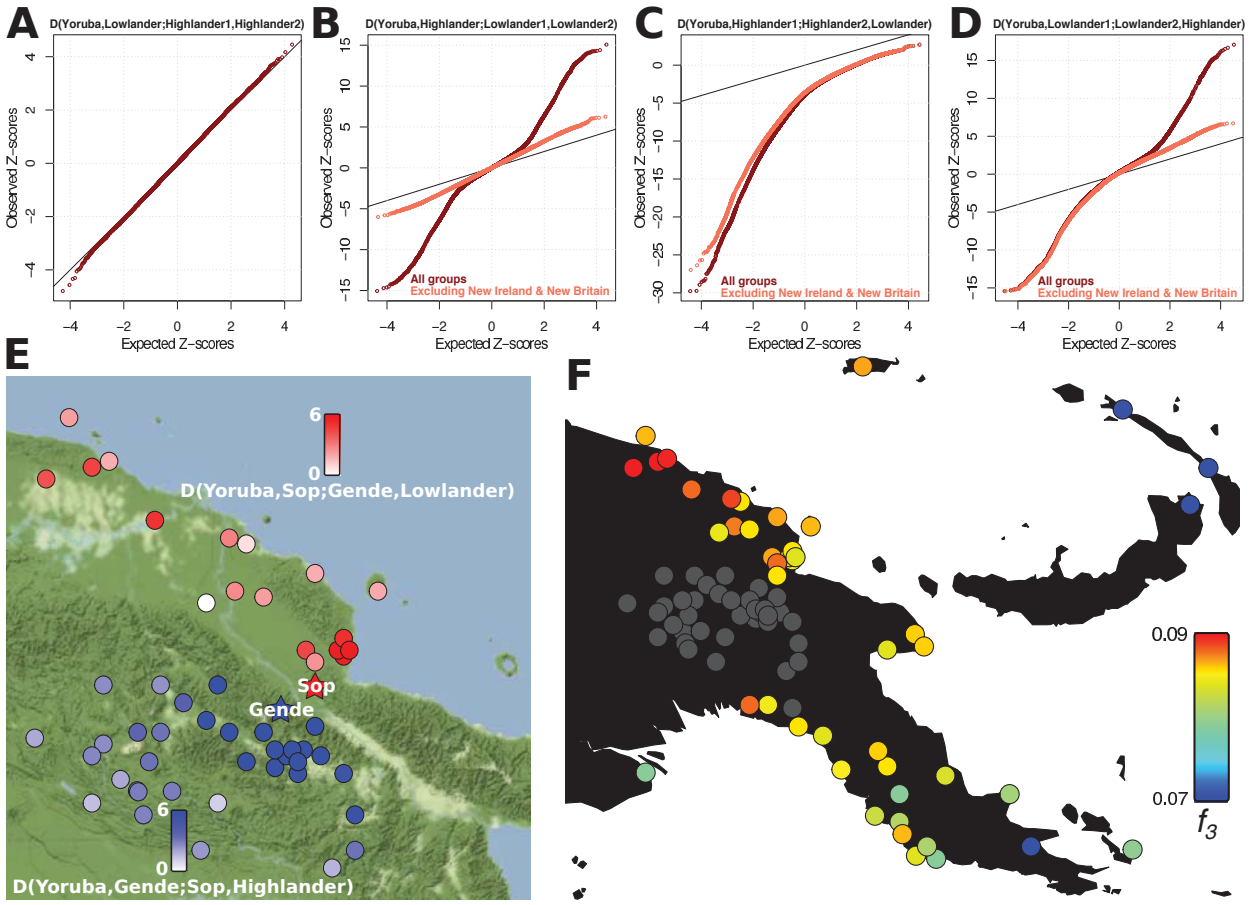


***Figure 3.5: The genetic relationships between highlanders and lowlanders.*** *Lowlander genomes were locally masked for Southeast Asian ancestry. A-C: Quantile-quantile plots comparing distributions of Z-scores from D-statistics relating highlanders and lowlanders to those expected under a normal distribution. The African Yoruba population is used as an outgroup. A) Lowlanders are equally similar to different highlander groups. B) Highlanders have stronger affinity to some lowlander groups than to others. C) Highlanders are more similar to each other than to lowlanders. D) Lowlanders are not always more similar to each other than to highlanders. E) The Z-scores of two different* D*-statistics, the first measuring if the highland Gende speakers are more similar to the lowland Sop speakers or to other highlanders (blue meaning more highlander similarity), and the second if Sop speakers are more similar to Gende speakers or to other lowlanders (red meaning more lowlander similarity). Z-scores were capped at 6. F) Genetic affinity of highlanders (treated as a single group, marked in grey) to different lowland groups measured by the outgroup* $f_3$ *statistic* $f_3$*(Highlanders,X;Aboriginal Australian) (red meaning higher affinity).*

even so, the overall picture and what it tells us about the relationships between highlanders and lowlanders would still likely remain the same.

## 3.8 The time depth of population separation in PNG

While array genotypes allow for the similarity between individuals and groups to be assessed, their use is limited when trying to measure the timing of events. Whole-genome sequencing is better-suited for this purpose, and we have generated this for 7 different PNG groups; six in the highlands and one in the lowlands (the "HGDP_L" East Sepik subset of the HGDP-CEPH PNG samples). I applied the MSMC method (Schiffels and Durbin 2014) to these data in order to estimate pairwise split times between groups, scaling results using a point mutation rate of $1.25 \times 10^{-8}$ per site per generation and a generational interval time of 30.5 years (the latter deriving from

an anthropological study of a PNG highlander group (Wood 1987)).

The MSMC method requires phased haplotypes, and it is becoming increasingly clear that insufficient phasing accuracy can have large impacts of inferences of split times (Song, et al. 2017). This is of particular concern in populations like Papuans, the ancestry of which is not represented at all in current haplotype reference panels used for statistical phasing. I took two approaches to tackle this issue here. First, using 10x Genomics linked-read technology (Zheng, et al. 2016) to physically infer phase for a subset of eight genomes from four groups. This is a library preparation technology that links a small barcode sequence to all DNA fragments that derive from the same, longer ($\sim$10-100 kb) molecule fragment, such that reads obtained from standard Illumina sequencing can then be computationally linked together and haplotype phase inferred. Second, using the haploid and therefore necessarily perfectly phased X chromosomes of males, as in Chapter 2.

The separation between highlanders and the East Sepik lowlanders appears to have occurred between 10 and 20 kya, perhaps 15 kya, on the basis of the genomes phased using linked-read technology (Figure 3.6A). With statistically phased data, the curves are substantially shifted towards older times and inferred splits would appear several thousand years older (and furthermore there is quite a large difference between results obtained using 4 haplotypes versus using 8 haplotypes, not shown), likely as a consequence of poor phasing. The X chromosome results generally suggest more recent split times, but the variable and sometimes non-monotonic behaviour of the curves makes these results more difficult to interpret.

Separations between groups within the highlands are more recent, with all appearing to have occurred $\sim$10 kya or more recently (3.6B). Similarly to above, physically phased genomes and X chromosomes result in more recent splits relative to statistically phased genomes. The oldest splits occur between groups in the eastern and western highland clusters, consistent with the population structure inferred from the array genotype dataset. A split between two groups within the eastern cluster (Gende and HGDP_H) appears to be less than 5 ky old.

In summary, the age of population structure in PNG does not date back to the initial peopling of Sahul. While there is considerable technical uncertainty surrounding the estimated dates, both methodological and related to mutation rate and generation time estimates, the overall picture is one where the separation between highlands and lowland groups occurred 10-20 kya, and highland groups then separated from each other within the last 10 kya. The use of physical phasing through the 10x Genomics linked-read technology greatly aided these analyses, providing one way to overcome the problem of haplotype phasing in diverse human populations.

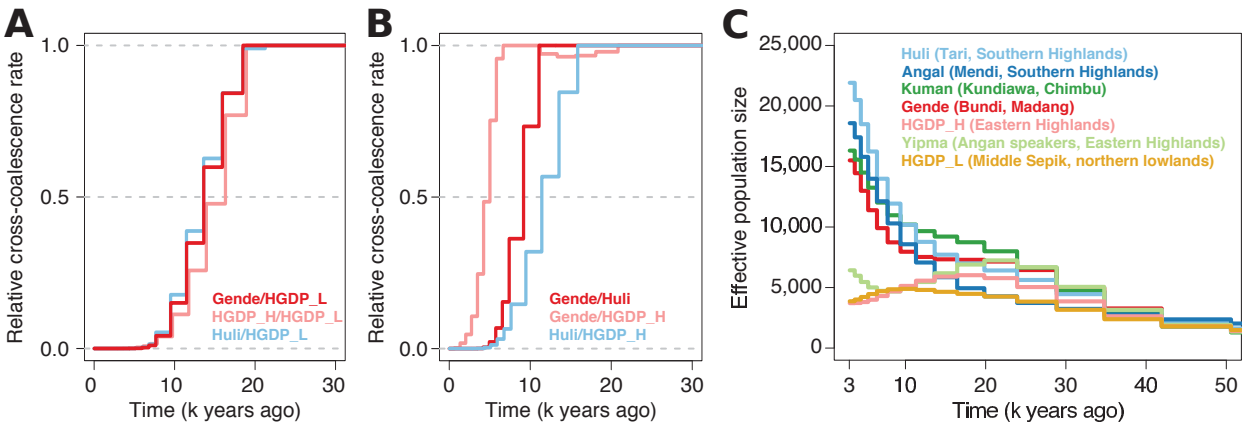


***Figure 3.6: The time depth of population separation and growth in PNG.*** *(A) MSMC relative cross-coalescence curves between highland groups and a northern lowland (East Sepik) group suggests a split time between 10 and 20 kya. (B) MSMC relative cross-coalescence curves between highland groups suggest split times within the last ∼10 ky (Huli representing the western cluster, Gende and HGDP_H the eastern). These curves were inferred from genomes physically phased using linked-read sequencing. (C) Effective population size histories of different groups as inferred using SMC++ on five high-coverage genomes per group.*

## 3.9 Population size histories in PNG

The MSMC and the conceptually related SMC++ (Terhorst, et al. 2017) methods allow for effective population size histories to be inferred from whole genome sequences. The former requires haplotype phasing when run on more than one genome, while the latter utilizes only allele frequency information from additional genomes and therefore does not require phasing. Applied to PNG groups, the two methods give qualitatively similar results, though MSMC gives very high estimates of effective population sizes in the very recent times, which might be an artefact of the poor phasing quality of these genomes (not shown). The SMC++ results might thus be more reliable. They show that most highlander groups have experienced major growth of effective population sizes, especially in the last 10 kya (Figure 3.6C). Exceptions are the Angan-speaking Yipma, who in other analyses show signs of genetic isolation and heavy drift, and the HGDP_H group (the origin of which is not actually known, and also in contrast display a recent increase in the MSMC analysis). The Sepik lowlanders do not display the same growth, instead retaining a more or less constant effective population size. This is in line with anthropological records of lower population densities in the lowlands, and might be linked to widespread malaria in these regions (Riley 1983).

## 3.10 Genetic differentiation in PNG

Genetic differentiation between pairs of populations as measured using the $F_{ST}$ statistic typically take values on the order of 1% or less between major, non-isolated populations within Europe, East Asia or South Asia (1000 Genomes Project Consortium 2015; Pagani, et al. 2016). Between European and East Asian populations, $F_{ST}$ is approximately 10%.

Calculating $F_{ST}$ between language groups in the PNG array genotype dataset after locally masking Southeast Asian ancestry (Figure 3.7), values between the western and eastern highlands clusters reach 2-3%, with all values within each of these clusters being below 2% but some being above 1%. Values between the Angan speaking groups in the south-eastern highlands and other highlander

groups reach 4-5%, which is as high as between European and South Asian populations. This is all within a sampled area within the highlands that is approximately the same size as Denmark, or the Netherlands. Within both the southern and the northern lowland areas, levels of $F_{ST}$ are as high as those in the highlands. This suggests that the mountainous terrain of the highlands, often cited as an explanation for the great cultural diversity of PNG, might not be what underlies differentiation between groups, instead raising the likely importance of cultural and linguistic factors. Between the highlands, southern lowlands and northern lowlands, $F_{ST}$ values are even higher, with many over 5%. These results thus demonstrate that the great cultural and linguistic diversity of PNG is reflected in strong genetic differentiation.

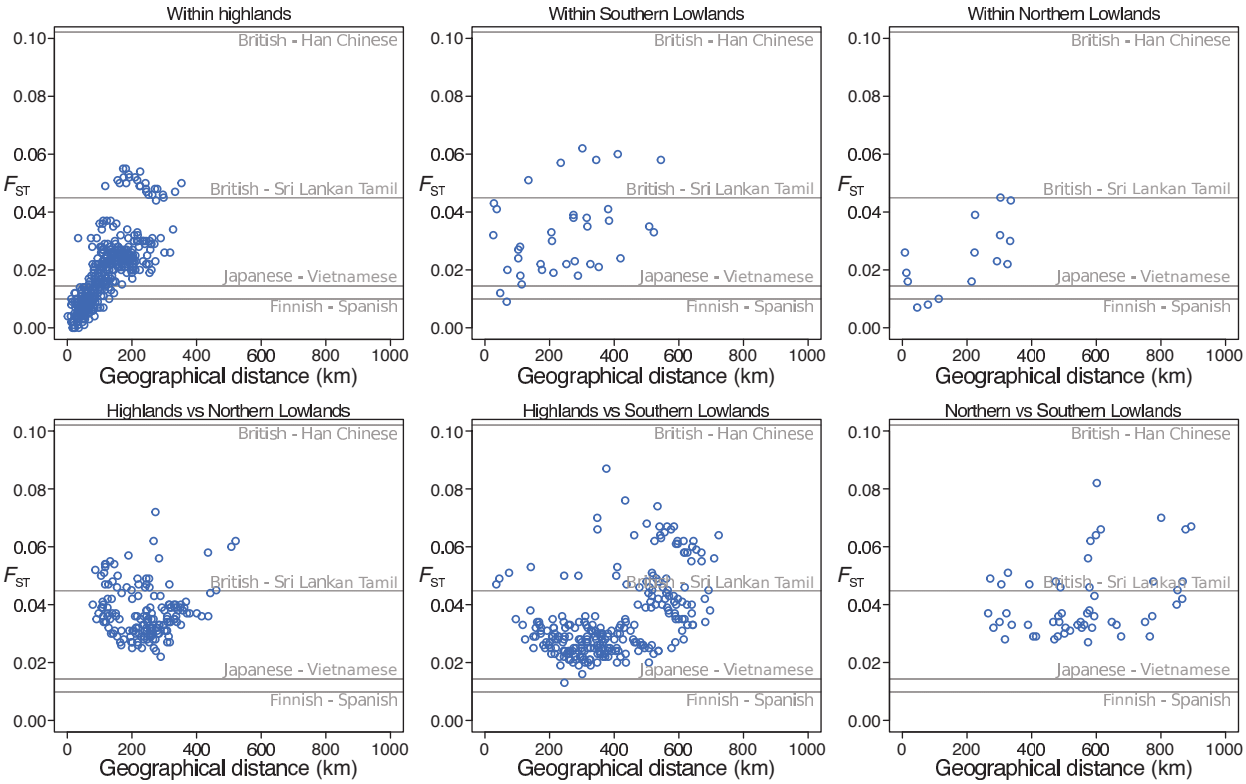


***Figure 3.7: Genetic differentiation in PNG.*** *Geographic distance between groups is plotted against* $F_{ST}$*, a measure of allele frequency differentiation. This was calculated after locally masking lowlander genomes for Southeast Asian ancestry, which otherwise has large effects on* $F_{ST}$*. Grey lines indicate a number of* $F_{ST}$ *values between selected populations from the 1000 Genomes Project for comparison.*

Population differentiation is slightly stronger for the Y chromosome than for the mitochondrial genome (p = 0.0035, Wilcoxon signed rank test for difference in the mean of a $F_{ST}$ metric for haploid loci). This implies that there is more female than male movement between groups and/or that male effective population sizes are smaller (e.g. due to larger variance in reproductive success), and is consistent with previous studies of New Guinean populations (Kayser, et al. 2003).

## 3.11 Diversity and isolation in PNG

The high differentiation between groups, the high linguistic diversity and the small present-day

size of many PNG language groups raise questions about lifestyles and interaction patterns in PNG. While the coalescent-based approaches applied above (SMC++, MSMC) can reveal demographic histories deep into the past, they have limited resolution in last few thousand years. Patterns of runs of homozygosity (RoH) in a genome can inform on mating patterns in more recent history. A RoH reflects a genomic segment inherited from the same ancestor, and its length the number of generations that have passed since that ancestor lived. Recent small effective population sizes and inbreeding will therefore lead to an increase in the number of long RoHs in a genome. Inference of RoH in the genotype array dataset reveals levels that are generally higher than those of major populations in East Asia and Europe (Figure 3.8A), likely reflecting generally lower recent effective population sizes. There are also substantial differences between different groups within PNG (Figure 3.8B). The highest levels are found among the Angan-speaking groups from the south-eastern highlands. These are the highland groups that display the highest $F_{ST}$ values to other highland groups (4-5%). The western highlands cluster has slightly lower amounts than the eastern highlands cluster. There is no overall difference between highlanders and lowlanders in their amount of long RoH (Wilcoxon signed rank test, p = 0.227).

As the rate at which $F_{ST}$ increases is dependent on effective population sizes, the small sizes of PNG groups likely contribute to the high differentiation observed. $F_{ST}$ to an outgroup can be taken as a proxy for the amount of genetic drift experienced by a given group. While neither the negative correlation between the estimated number of speakers in a language group and the $F_{ST}$ to outgroup (Figure 3.8B), nor that between the number of speakers and the amount of RoH (Figure 3.8C), is significant, a handful of groups stand out in having small numbers of speakers, high $F_{ST}$ to an outgroup and large amounts of RoH. These include the Angan speakers mentioned above as well as the Grass Koairi and the Keapara speakers, both from Central Province. These likely represent examples of recently culturally isolated groups that have experienced high levels of genetic drift in relatively short periods of evolutionary time. There is a significant positive correlation between the amount of RoH in a group and the $F_{ST}$ to an outgroup (Figure 3.8D), consistent with small effective population sizes driving the genetic drift and high differentiation in PNG.

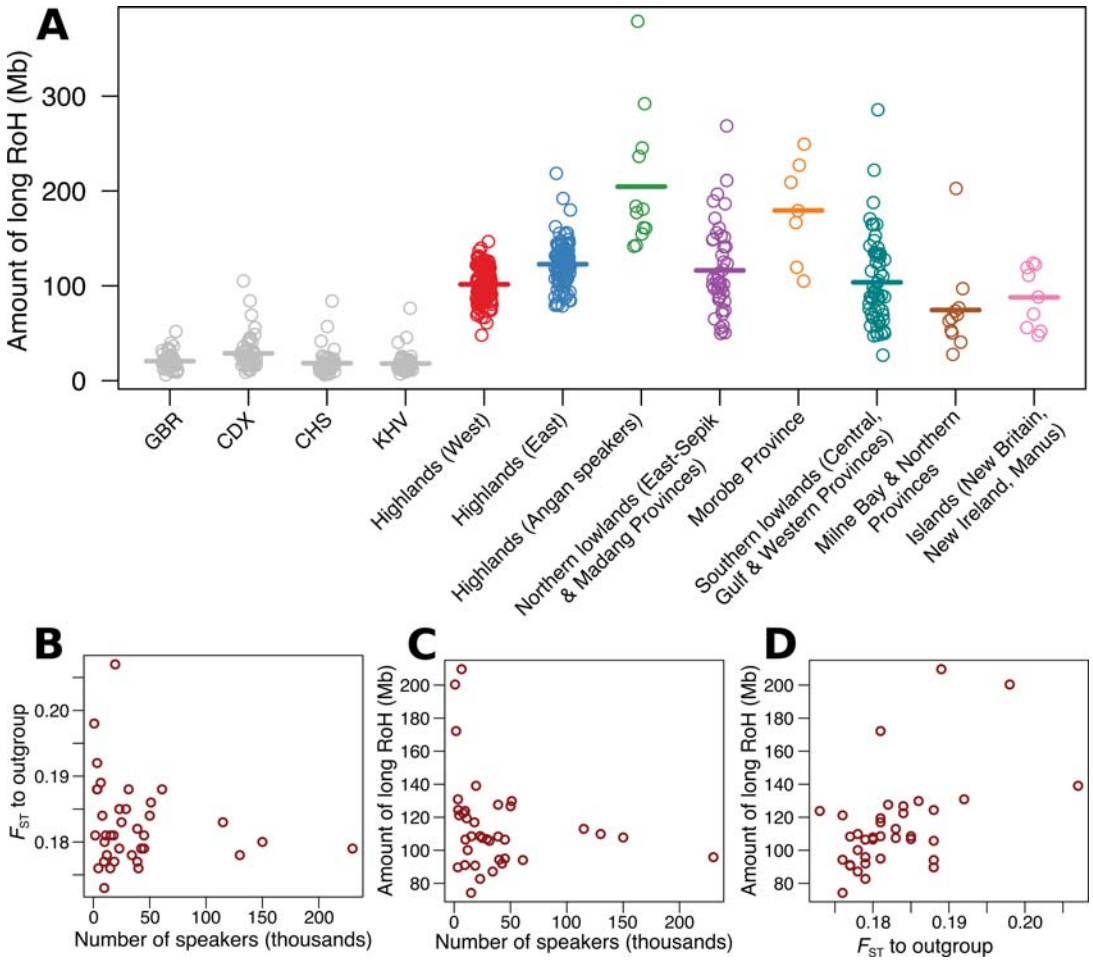


***Figure 3.8: Patterns of runs of homozygosity and genetic drift.*** *A) The sum of lengths of RoH longer than 1Mb in each individual genome, grouped by major region. Horizontal lines indicate the mean per region. Four populations from the 1000 Genomes project are included for comparison: GBR = British in England and Scotland; CDX = Chinese Dai in Xishuangbanna, China; CHS = Southern Han Chinese; KHV = Kinh in Ho Chi Minh City, Vietnam. B) The estimated number of speakers in a group plotted against* $F_{ST}$ *to an outgroup (KHV). Pearson correlation = -0.176, p = 0.2967. C) The estimated number of speakers in a group plotted against the group mean sum of lengths of RoH longer than 1 Mb. Pearson correlation = -0.233, p = 0.1656. D)* $F_{ST}$ *to an outgroup (KHV) plotted against the group mean sum of lengths of RoH longer than 1 Mb. Pearson correlation = 0.520, p = 0.0009662. In B-D, only groups with at least two individuals are included. Estimated speaker numbers for each language group were obtained from Ethnologue (Lewis, et al. 2016).* $F_{ST}$ *to KHV was calculated after locally masking lowlander genomes for Southeast Asian ancestry, and restricted to groups with at least two sampled individuals. Results are very similar if using Aboriginal Australians as outgroup instead of KHV.*

## 3.12 A model for population history in Holocene PNG

Rather than supporting a naïve view where population structure was established shortly after the peopling of New Guinea and maintained since then, the results obtained here indicate that more recent processes have overwritten earlier structure. Ultimately, ancient DNA will likely be needed to determine with more certainty what exactly these processes were, but at present one can at least speculate. One model which would be compatible with present-day patterns of genetic variation is an expansion of a single agriculturalist group in the highlands, starting approximately 10 kya. If all highlanders derive most of their ancestry from this same source population, this could explain their clade-like relationship to lowlanders. It is also compatible with the split time estimates between highlanders and lowlanders and those within the highlands, and the observation of major increases in highlander effective population sizes within the same time period. This agriculturalist expansion could also have spread the languages that have now developed into the Trans-New Guinea language

family.

If this model is correct, then it means that, similarly to west Eurasia and sub-Saharan Africa, and likely East Asia too, agriculture spread through the movement of people in PNG. Its independent history would thereby provide another data point in favour of people in the "pots vs people" debate. At present, a model where agriculture spread without genetic admixture or replacement cannot be completely ruled out – perhaps the current divergences were established while groups where still hunter-gatherers, reflecting either some pre-agricultural population transformation (the likes of which have been described in Europe (Fu, et al. 2016)) or some kind of upper limit on the divergence achievable between hunter-gatherer groups living in a fairly small geographical area. However, the close correspondence between the genetic time estimates and the archaeological evidence for the emergence of agriculture arguably lends support to an agriculturally-driven restructuring.

A striking result which sets PNG apart from other regions of the world that have also undergone agricultural transitions is the very strong present-day genetic differentiation. The high $F_{ST}$ values across all of PNG differ dramatically from the relatively homogenous genetic landscapes of Europe, East Asia, and most populations of sub-Saharan Africa. Areas of comparable size to PNG in Europe have been found to have close to no population differentiation, e.g. Great Britain (Leslie, et al. 2015), the Netherlands (Genome of the Netherlands 2014) and Denmark (Athanasiadis, et al. 2016). An important insight into this comes from ancient DNA studies in Europe and the Near East, which have documented a dramatic but gradual decline in $F_{ST}$ values in this part of the world over the last 10 ky (Lazaridis, et al. 2016), as well as higher differentiation between hunter-gatherer groups than between farmer groups (Skoglund, et al. 2014). Before the agricultural transition in west Eurasia, some hunter-gatherer groups were as different as present-day Europeans and East Asians ($F_{ST}$ of $\sim$10%). These studies show that the homogenous landscape of present-day Europe actually only emerged in the last few thousand years. While an agricultural transition might be necessary to achieve this high level of genetic homogenization, the history of PNG demonstrates that it is not sufficient: PNG also went through an agricultural transition, but present-day genetic differentiation is high.

This thus calls for another explanation for the dramatic differences in genetic structure between PNG and west Eurasia or other genetically homogenous regions. A hypothesis is that the key difference is the absence in PNG of a Bronze age, Iron age or other similar post-agricultural, cultural transformation. In west Eurasia, the Bronze Age started $\sim$4.5 kya and was driven by an expansion of herders from the Pontic–Caspian steppe who had domesticated horses, metal technology and perhaps a different social structure. This expansion resulted in dramatic genetic admixture and replacement across all of Europe, as well as replacement of the vast majority of the indigenous languages with Indo-European languages (Allentoft, et al. 2015; Haak, et al. 2015). The presence of Indo-European languages, and steppe-related ancestry (Lazaridis, et al. 2016), in South Asia

suggest that a similar expansion shaped genetic structure here too. In sub-Saharan Africa, the expansion of Bantu-speaking farmers from west Africa was associated with an Iron Age (Huffman 1982), and resulted in genetic homogenization across a large geographical area (Li, et al. 2014; Patin, et al. 2017). PNG might then be similar in genetic structure to pre-Bronze Age Europe, with small, sedentary, agriculturalist groups and relatively little gene flow between them. The genetic results also align with, and provide some insight into, the enormous linguistic diversity of PNG, and perhaps suggest that many other parts of the world also harboured greater linguistic diversity in the past, before being homogenized by large-scale cultural expansions.

These expansions have often been associated with the rapid proliferation of particular Y chromosome lineages, e.g. R1b with the steppe migration into Europe, E1b with the Bantu expansion in sub-Saharan Africa, and likely R1a in South Asia (Karmin, et al. 2015; Poznik, et al. 2016), giving rise to 'star-like' phylogenies as the rapid spread leaves little time for new mutations to accumulate on the basal branches. The explanation for this is likely a social structure with higher variance in male reproductive success, where whatever Y chromosomes are carried by a small number of high-status males expand dramatically. An extreme case of this from more recent times might be the Y chromosome lineage thought to be carried by the central Asian ruler Genghis Khan, today present in an estimated 8% of East Asian males (Zerjal, et al. 2003). Consistent with the genetic landscape of PNG being unperturbed by large-scale, post-agricultural expansions, the phylogeny of PNG Y chromosomes does not feature any star-like bursts of expansion (not shown). One striking case of Y chromosome structure can be noted, however: the Gende highland group carry a Y chromosome from haplogroup C at a frequency of 86%, but this chromosome is absent from the rest of the highland samples examined. Analysis of whole-genome sequencing data from five randomly selected Gende males indicates a common ancestor for their Y chromosomes only $\sim$1 kya. It thus seems likely that this chromosome has risen in frequency due to a recent, small-scale instance of high male reproductive variance in the Gende speakers. However, consistent with the high degree of population differentiation in PNG, this Y chromosome expansion appears to have been contained to only this single language group, without spreading more widely.

The proposed model of an agriculturalist expansion in the highlands is simple, and therefore likely to be inaccurate in the finer details. There are also aspects that it fails to explain, including the fact that lowland groups also practice agriculture, and in many areas also speak languages from the Trans-New Guinea language family. If this was the result of the same expansion of people spreading down from the highlands, more recent divergences and greater genetic continuity between highland and lowland groups would have been expected. More work is thus needed to further the test the basic outline of the model and refine it with additional detail.

## 3.13 Conclusions

The people living in the highlands of PNG appear to have been unaffected by Southeast Asian gene flow, thereby making them as genetically independent from Eurasian sources as Aboriginal

Australians, i.e. for ∼50 ky. While the notion that agriculture was an independent development in the PNG highlands largely serves as a background assumption for genetic studies, derived from archaeology, these genetic findings themselves also provide evidence for this notion.

The first detailed view of population structure in PNG, made possible by the comprehensive array genotyping dataset in combination with whole-genome sequences, hints at a complex population history. The age of present-day population structure in the highlands is not older than the development of agriculture, suggesting a reshaping following this lifestyle transition. The highlands-lowlands contrast constitutes a major barrier to gene flow, and all people across the highlands appear to form a clade relative to lowlanders. Population differentiation is high, despite divergences not being particularly ancient, instead implying that relatively recent isolation between groups is responsible. Population genetics theory also demonstrates that $F_{ST}$ will increase more rapidly in groups with smaller effective population size. A parallel can perhaps be drawn between PNG and some communities in South Asia, where founder effects and cultural isolation between groups even living in the same geographical areas has led to strong genetic differentiation (Reich, et al. 2009; Nakatsuka, et al. 2017). A hypothesis for why PNG has such strong population structure while regions such as Europe, East Asia and parts of sub-Saharan Africa do not, is that the latter regions have been recently homogenized by large-scale, technology-driven expansions such as Bronze and Iron ages. The history of PNG therefore demonstrates that human population histories are not deterministic, but can take quite different trajectories in different parts of the world.

There is currently no consensus on where within the highlands agriculture was first developed. The population-genetic analyses presented here provide little if any insight into the matter – relationships between present-day groups tell us very little about the geographical origins of their ancestors. However, the results can at least be superficially compared to predictions implied by a hypothesis put forth on anthropological and archaeological grounds, and which states that agriculture likely started in the western parts of the PNG highlands and then spread from there to the eastern parts (Feil 1987). The basis for this hypothesis is several observations suggesting higher present-day agricultural productivity in the western parts, as well as larger group sizes. The genetic results give some support for this. Firstly, the levels of runs of homozygosity are slightly lower in the west (Figure 3.8A), suggesting slightly larger population sizes. Secondly, although this might not be statistically significant given the resolution offered by the methods, inferred historical effective population sizes in the last few thousand years are also slightly higher in the western groups (Figure 3.6C).

It is worth noting that, despite the strong differentiation and low gene flow between groups, Papua New Guinea has undergone one quite substantial cultural transition in the last 500 years or so. Before this, the prominent agricultural crops in PNG were taro and yams; however, today the non-indigenous sweet potato dominates (Bourke, et al. 2009). It is not entirely clear how the sweet potato, which originated in the Americas, reached New Guinea: whether it was brought by

European explorers or by Polynesian seafarers shortly prior to the era of European colonialism. Genetic analyses of sweet potato samples seem to favour the latter hypothesis (Roullier, et al. 2013). In any case, the use of the sweet potato seems to have spread rapidly across New Guinea without signs of any substantial genetic restructuring, showing that such cultural innovations can disseminate even through human societies that are highly genetically and linguistically fragmented.

## 3.14 Materials and Methods

Array genotyping and 10x Genomics linked-read sequencing was carried out by the Wellcome Trust Sanger Institute sequencing and genotyping facilities. Analyses of these data and the 39 whole genome sequences are described in further detail in (Bergström, et al. 2017). A brief description of the methods used for the analyses described here follows.

Array genotypes were filtered for markers with ambiguous chromosomal location, indel status, high rate (>5%) of missing genotypes, low minor allele frequency (<1%), and for individuals with high rate (>10%) of missing genotypes, to produce a final variant set consisting of 529,137 variants and 378 individuals. These genotypes were combined with those from previously whole-genome sequenced PNG, Aboriginal Australian (Prufer, et al. 2014) and worldwide populations (1000 Genomes Project Consortium 2015) using the "merge" command from the bcftools software (https://samtools.github.io/bcftools/) with the "-m" argument, excluding sites that became multi-allelic after merging. Data from the Simons Genome Diversity Project (Mallick, et al. 2016) was downloaded and analysed in isolation.

Uniparental genotypes were handled separately to assign each individual to haplogroups. Array genotypes were merged with whole-genome sequencing data using Y chromosome genotypes previously called, and mitochondrial genotypes called using FreeBayes v0.9.18 (Garrison and Marth 2012) with the arguments "--ploidy 1" and "--report-monomorphic". For the Y chromosome, the patterns of genotypes across individuals were hierarchically clustered and variants defining haplogroups manually identified and checked against the ISOGG database (http://isogg.org/tree/, April 21 2016 release). Individuals were assigned to the C, M, S, O and K haplogroups. For the mitochondrial genome, the same approach was applied, using information in the PhyloTree database (build 17) (van Oven and Kayser 2009), alongside an independent, higher-resolution haplogroup prediction by the HaploGrep 2 software (Weissensteiner, et al. 2016). The two approaches agreed on the (lower-resolution) haplogroup in all cases. Individuals were assigned to the P, M, E, B4, B5b and N13 haplogroups. A uniparental $F_{ST}$ metric was calculated as ($H_T$ – $H_S$) / $H_T$, where H is haplotype diversity for the total ($H_T$) and the subpopulations ($H_S$), calculated as 1 - $\sum p^2$, where p is allele frequency.

PCA analyses were performed using EIGENSTRAT 6.0.1 (Patterson, et al. 2006), using the "-w" flag to restrict the calculation of principal components to a subset of individuals. ADMIXTURE 1.23 (Alexander, et al. 2009) was used for model-based ancestry assignment. $f_3$- and *D*-statistics,

$f_4$-ratios and admixture graph fits were calculated using ADMIXTOOLS 4.1 (Patterson, et al. 2012). $F_{ST}$ between pairs of language groups was calculated using EIGENSTRAT 6.0.1 (Patterson, et al. 2006), restricted to groups that have at least two sampled individuals. Local ancestry assignment was performed using RFMix v.1.5.4 (Maples, et al. 2013), using PNG highlanders and 1000 Genomes East Asians as reference panels, after phasing haplotypes on the Sanger Imputation Server (McCarthy, et al. 2016) using the EAGLE algorithm (Loh, et al. 2016) and the Haplotype Reference Consortium (release 1.1) reference panel. Runs of homozygosity were inferred using PLINK v1.07 (Purcell, et al. 2007) with the arguments "--homozyg-window-kb 1000 --homozyg-window-snp 50 --homozyg-density 50 --homozyg-window-het 0".

Effective population size histories were inferred on five genomes per population using SMC++ (Terhorst, et al. 2017) and on four genomes per population using MSMC (Schiffels and Durbin 2014) (as well as MSMC2), in both cases using the mappability mask recommended for the latter. MSMC was also used to estimate the relative cross-coalescence rate between populations, applying the "—skipAmbiguous" argument to exclude unphased segments and the "—fixedRecombination" argument. To make use of 10x Genomics experimental phasing data, phase information from the LongRanger v2.1.2 VCFs was lifted into the existing VCFs, setting to unphased any position where the genotypes disagreed between the two VCFs. Analyses on the X chromosome were performed as described in Chapter 1. All demographic results were scaled to units of years using a mutation rate of $1.25 \times 10^{-8}$ per site per generation and a mean generation interval time of 30.5 years (Wood 1987).

To infer the genetic identify of all the 17 Papua New Guinean individuals in the HGDP-CEPH panel, array genotype data on these (Li, et al. 2008) was merged with the newly generated PNG array genotypes, and their relationships to the latter was inferred using PCA and outgroup $f_3$-statistics.