JAMES RENNIE BEQUEST

REPORT ON EXPEDITION / PROJECT / CONFERENCE

Expedition/Project/ Conference Title: Summer internship in Prof Irene Gallego Romero's group (Human Genomics and Evolution)

Travel Dates: July 8 – September 13, 2024

Location: St Vincent's Institute of Medical Research, Melbourne, Australia

Group member(s): Alice Groudko

Aims: To infer the traces of archaic human introgression in low-coverage genomes from Indonesia, develop skills in human genomic data analysis and methods in evolutionary genomics

Photography consent form attached: *(please refer to your award letter)*

☐ Yes
☒ No

OUTCOME (a minimum of 500 words):-

During my internship in Prof. Irene Gallego Romero's group (Human Genomics and Evolution) in St Vincent's Institute of Medical Research in Melbourne this summer, I found and quantified the traces of Neanderthal and Denisovan admixture in the sample of more than 200 modern human genomes from Indonesia and investigated selection signatures in a single genomic region. This area of the world is characterised by vast genetic and cultural diversity of local human populations; however, it is disproportionally understudied(1). My results will be used in the further analysis of this dataset, in particular, single cell RNA sequencing data, possibly revealing the background for local adaptation (e.g. to infectious diseases), and will help contribute to characterising the diversity of human populations in this region.

Introgressed genomic material from admixture can be an important source of variation driving adaptive evolution. Adaptive effect of introgression has been shown in various systems, from beneficial mimetic wing patterns in *Heliconius* butterflies to the EPAS1 variant in a Denisovan-like haplotype in Tibetans which is associated with high altitude adaptation(2,3). The genomes of Oceanians bear the traces of contact with at least two diverged Denisovan populations which are distinct from the archaic contribution into East Asians(4), and have the highest percentage of Denisovan introgression in the world(5).

For inferring the most information about the individuals from their genomes, whole genome sequencing (WGS) is a preferred method over genotyping. WGS is now often performed in low coverage (that is, every position in the genome is covered by fewer reads on average) as this allows to minimise the costs and drastically expand sample size(6). To mitigate sequencing errors, gaps, and uncertain bases in low-coverage data, such datasets are imputed against a reference panel, so the low-confidence variants are compared to the haplotypes occurring in the population. However, this might introduce bias, especially if reference panel includes few individuals from the source population. Here, I analysed the dataset of >200 imputed low-coverage genomes from four villages in Papua Highlands, Lowlands, and Bali to test if low-coverage data is suitable for inferring archaic introgression and to expand the view on genomic diversity in these populations.

I used HMM-based method (hmmix) to identify candidate fragments of archaic origin(7). Hmmix detects the regions with elevated density of the variants not seen in the non-admixed population (here, sub-Saharan African genomes from 1KGP) (7,8). The archaic fragments may derive from incomplete lineage sorting rather than admixture. In this case, the archaic haplotype will have more time (since the common ancestor of humans and Neanderthals/Denisovans) to be broken up by recombination, so the average length of such fragments is expected to be shorter. I performed filtering for the possibility of incomplete lineage sorting by accounting for the fragment length and recombination rate in the genomic region(3,9). Then, I assigned the ancestry of each fragment to Neanderthal, Denisovan, or unknown based on the number of variants matching the archaic human genomes (Figure 1) (10–12). To process the data, I combined several hmmix execution steps and post-processing in a Nextflow pipeline which can be run on a high-performance computer. I used bedtools and wrote R scripts to find the intersections between the introgressed fragments in high and low coverage data from the same 4 individuals and found a substantial overlap of more than 90%, which supports the usage of the method in low-coverage data(13). I visualised the distribution and prevalence of the introgressed segments in the sampled individuals on the chromosome heatmap, registering deserts and high-prevalence “marker” introgressed variants. As expected, the highest proportion of Denisovan introgression was observed in Papua Highlanders, compared to the villages in Papua Lowlands and Bali.

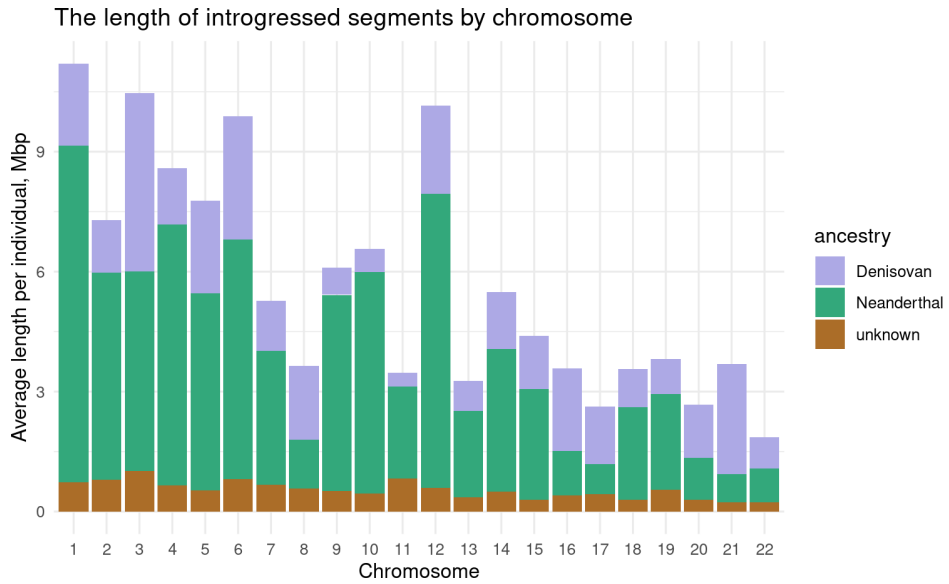


Figure 1 – **The average length of archaic genomic material of Denisovan, Neanderthal, or unknown ancestry on each chromosome.** The data shows higher percentage of Neanderthal introgressed variation; however, this is possibly due to huge evolutionary distance of the individual that provided Denisovan reference genome from the admixed populations(4).

I also investigated a set of ancestry-specific variants associated with differential splicing of DNM3 gene in the Korowai in the RNAseq dataset from across Island Southeast Asia.

Human populations across Island Southeast Asia exhibit genomic cline of Asian to Papuan genetic ancestry from west to east(14). The lab has previously investigated alternative splicing patterns in the whole blood RNA-seq data from three traditional populations (of Asian-like, Papuan-like, and admixed ancestry) spanning the west-east cline of genetic diversity in Indonesian archipelago (15). Among the variants affecting gene splicing that correlate with Papuan genetic ancestry, a splicing quantitative trait locus (sQTL) associated with the differential splicing of DNM3 gene was identified. DNM3 codes for dynamin-3 –

endocytosis mediating protein which plays different cellular roles such as megakaryocyte development. The identified sQTL is co-inherited with several other SNPs in this genomic region, all are derived from the same genomic ancestry, and also correlates with haemoglobin concentration (15). Increased haemoglobin concentration is a known adaptation to high altitudes. Considering this and the fact that Papuan highlands was one of the main routes for the settlement of Papua New Guinea Island by humans (16), the region could possibly have undergone past selection.

The investigated sQTLs variants appeared to have notable patterns of allele frequencies globally, with exceptionally high percentage in Korowai and low percentage in Europe and East Asia. I phased the data with SHAPEIT5, calculated F_{ST} and selection scores in the region, and predicted possible splicing sites with SpliceAI. I found a 70 kbp block of elevated F_{ST} covering the parts of DNM3, PIGC, and C1orf105 genes, which falls into 8% of the highest F_{ST} values on chromosome 1, indicating high level of genetic differentiation between East Asian and Korowai populations in this genomic region (Figure 2). Most these variants also show slower haplotype decay around than ancestral variant (extended haplotype homozygosity, EHH), which may be a signature of a faster beneficial haplotype spread due positive selection.

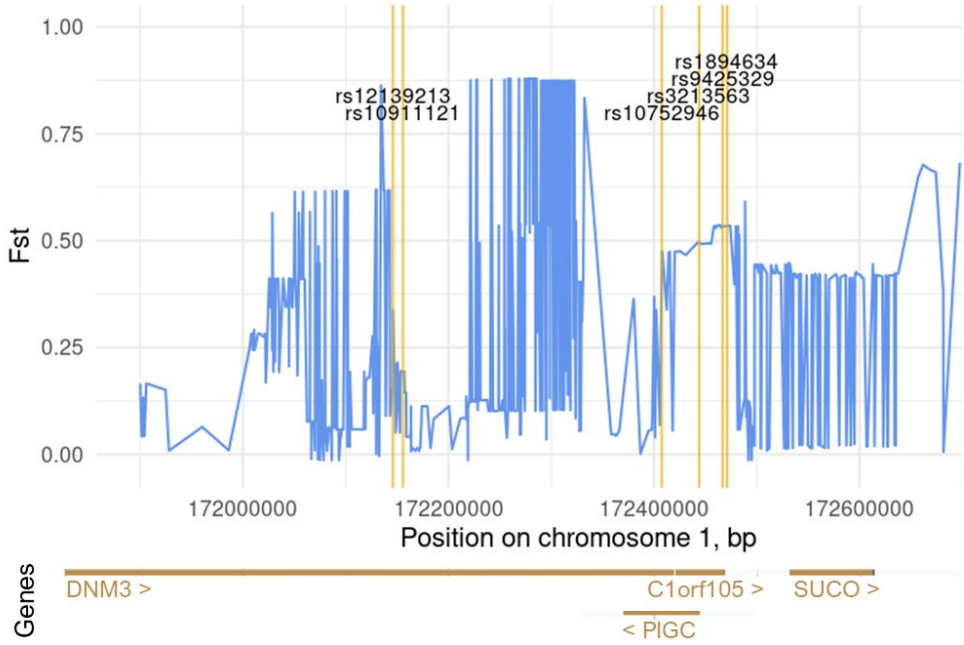
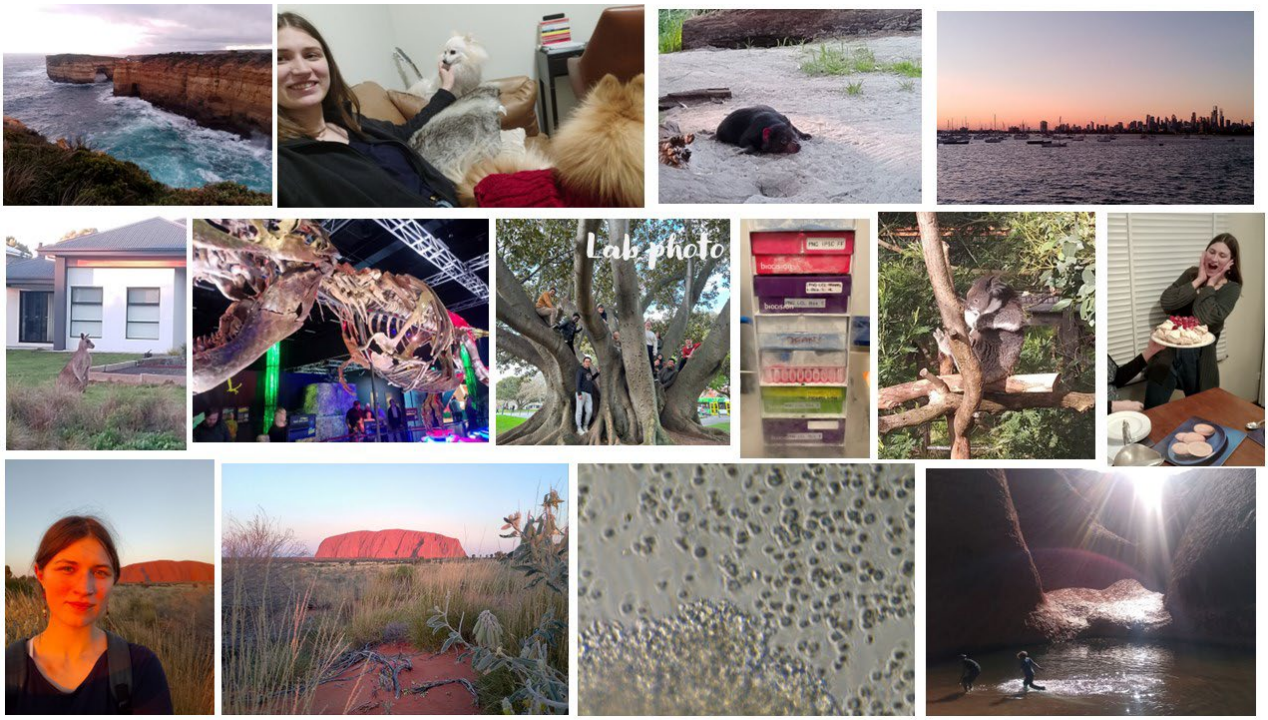


Figure 2 – **Fst scores for the variants associated with differential splicing in Korowai**. The populations compared are East Asian and Indonesian from 1KGP and HGDP(8,17). The position of the variants is marked with yellow vertical lines.

By working on the project, I learnt to manipulate human variant data and got acquainted with the methods for introgression detection and selection inference using genomic information. I actively communicated with the collaborators from multiple research groups, discussing the methods and the dataset. I gave two presentations on my work in a lab meeting and a joint meeting with another research group. I also acquired functional genomics perspective on human variation from the work of other group members and by attending the weekly seminars in the institute. Besides, I helped in recording and sorting the old samples from the cryostorage, learnt how to handle cell lines and carried out cell culture passaging.

Figure 3 – **Collage of the photos from my stay in Australia and work in the lab**.



Reference list

1. Natri, H. M. *et al.* Genetic architecture of gene regulation in Indonesian populations identifies QTLs associated with global and local ancestries. *The American Journal of Human Genetics* **109,** 50–65 (2022).
2. Pardo-Diaz, C. *et al.* Adaptive Introgression across Species Boundaries in Heliconius Butterflies. *PLOS Genetics* **8,** e1002752 (2012).
3. Huerta-Sánchez, E. *et al.* Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512,** 194–197 (2014).
4. Jacobs, G. S. *et al.* Multiple Deeply Divergent Denisovan Ancestries in Papuans. *Cell* **177,** 1010-1021.e32 (2019).
5. Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468,** 1053–1060 (2010).
6. Terhorst, J., Kamm, J. A. & Song, Y. S. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat Genet* **49,** 303–309 (2017).
7. Skov, L. *et al.* Detecting archaic introgression using an unadmixed outgroup. *PLOS Genetics* **14,** e1007641 (2018).
8. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
9. Halldorsson, B. V. *et al.* Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science* **363,** eaau1043 (2019).
10. Meyer, M. *et al.* A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science* **338,** 222–226 (2012).
11. Mafessoni, F. *et al.* A high-coverage Neandertal genome from Chagyrskaya Cave. *Proceedings of the National Academy of Sciences* **117,** 15132–15136 (2020).
12. Prüfer, K. *et al.* A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358,** 655–658 (2017).
13. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).

14. Hudjashov, G. *et al.* Complex Patterns of Admixture across the Indonesian Archipelago. *Molecular Biology and Evolution* **34,** 2439 (2017).
15. Ibeh, N. *et al.* Profiling genetically driven alternative splicing across the Indonesian archipelago. *The American Journal of Human Genetics* (2024). doi:10.1016/j.ajhg.2024.09.004
16. Brucato, N. *et al.* Papua New Guinean Genomes Reveal the Complex Settlement of North Sahul. *Molecular Biology and Evolution* **38,** 5107–5121 (2021).
17. Bergström, A. *et al.* Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367,** eaay5012 (2020).