Master 2 Bioinformatique

Semestre 10
Université de Bordeaux

Project report: TITLE

Aurélien Luciani

Supervisor : Murray P. Cox

Year 2014-2015

Contents

Introduction

Two groups of populations can be identified in the Islands of South-East Asia (ISEA), one is composed of the Melanesians, whose ancestors settled in these islands during the first human settlement, around 45 thousand years ago. The other arrived more recently during a period often called the Austronesian expansion, between 5 and 4 thousand years ago, when people from mainland China settled in the islands. Nowadays people living in this area have mixed genomic ancestry and markers can be identified and defined as either from an Asian ancestry or a Melanesian one. These markers are based on signle nucleotid polymorphisms located in different chromosomes and 52 markers can be used to define accurately the admixture of the Asian ancestry in every individuals [1].

The choice of these SNPs is a result of previous studies at Massey University, the University of Arizona, the Santa Fe Institute, and the Eijkman Institute that sequenced 1430 individuals from 60 populations. This set of SNPs is defined as highly informative and allows to define the ancestry of a person based on a small quantity of markers that are highly discriminant.

Two specific patterns can be seen, one is the non linear gradient of Asian admixture when observing individuals in the different islands when looking along the longitudinal axis, corresponding more or less the wave the settling might have happened. The second is the difference of admixture when looking at specific parts of the genomes associated with male or female ancestry, implying a gender-biased expansion.

Chapter 1

Project presentation

The project consists of developing a model of the Austronesian expansion throughout the ISEA that could reproduce the same two patterns observed in the real data. The first pattern can be seen in the figure (REF!!!) and is a non-linear gradient in Asian admixture that declines abruptly around the eastern part of Indonesia. The second one, represented in the figure (REF!!!), is a

Chapter 2

Implementation

2.1 Run management

The fact that many simulations are required to be able to infer meaningful information from the model imply that they cannot be run on a single desktop computer. The only case

2.2 Data processing, storage and query

2.3 Analysis

2.3.1 Comparisons

To be able to compare two different scenarios, one has to define comparison functions that will be able to provide a value of similarity or dissimilarity when provided with observable values that characterise the scenarios. In this case, the comparison is done between the real observed data and the output of one simulation. Every simulation output has to be treated to be of the same form than the observed data so that it can be compared. The observed data does not include values for specific islands that are included in the simulations, namely Alor, Tanimbar and Aru in eastern Indonesia. The admixture values for mitochondrion and Y-chromosome are also not available for every island. Those values in the models can then be dropped as they cannot be compared.

Many different comparison functions have been tested. Two functions have been selected that can hold different information about a comparison. The mean square distance (MSD) and the partial Mantel test have been selected as they do not have a good correlation when applying them to random admixture values, meaning that they do not carry the same information, they are complementary, and using them both gives a better overview of the comparison.

Mean square distance

The mean square distance is the mean value of all the distances between values of admixture for every islands.

$$MSS = \frac{\sum_{i=1}^{n}(Real_i - Simulated_i))^2}{n} \tag{2.1}$$

It gives a distance value, with 0 meaning that the two observed values are absolutely identical. In this specific context, with the fact that an admixture value is compared and because of the values observed in the real data, the higher possible value will be around 0.8.

Partial Mantel test

The Mantel test has been developed to be able to compare two matrices with the same information, the partial version of this test also uses a third matrix, holding geographical distances for the cells of the matrices to be able to weight the values according to the actual geographical distance of the points [2].

In this case, the matrices contain the values of distances of admixtures between every islands in the graph and a matrix M is calculated as

$$M = \begin{bmatrix} D(A_0, A_0) & \cdots & D(A_0, A_n) \\ \vdots & \ddots & \vdots \\ D(A_n, A_0) & \cdots & D(A_n, A_n) \end{bmatrix} \tag{2.2}$$

were D is the function returning the distance between the two arguments and A the admixtures of the n islands of the graph.

the partial Mantel test returns a correlation value that is between -1 and 1 with a value of 0 meaning that the two matrices are not correlated at all and 1 that they are completely correlated.

2.3.2 Visualisations

Chapter 3

Results

3.1 Grid search analysis

3.2 ABC Framework

Chapter 4

Discussion

Conclusion

Bibliography

[1] M.P. Cox, T.M. Karafet, Lansing J.S., Sudoyo H., and Hammer M.F. Autosomal and x-linked single nucleotide polymorphisms reveal a steep asian-melanesian ancestry cline in eastern indonesia and a sex bias in admixture rates. *Proceedings of The Royal Society*, 2010.

[2] Smouse P.E., Long J.C., and Sokal R.R. Multiple regression and correlation extensions of the mantel test of matrix correspondence. *Systematic Zoology*, 1986.