Characterizing genetic variation in three Southeast Asian populations

Ilana Arbisser
ilanama@stanford.edu

Jonathan Kang
jtlkang@stanford.edu

Introduction

Singapore is a small island-nation situated at the tip of the Malay Peninsular in Southeast Asia. It existed as a British colony from 1819 to 1959, interspersed by a period of Japanese occupation from 1942 to 1945 during World War II. Throughout its colonial history, Singapore has been the recipient of extensive immigration from various parts of Asia, and the same diversity is still reflected in Singapore's ethnic composition today.

Population of Singapore from the end of 1823 to beginning of 1833.

Classes.	1823.	1824.	1825.	1826.	1827.	1828.	1832.
Europeans	74	84	111	87	108	122	119
Native Christians	74	132	206	188	193	272	300
Armenians	16	9	18	19	25	24	35
Arabs	15	10	17	18	17	32	96
Natives of Coromandel and Malabar	390	690	605	777	1095	1440	1819
Natives of Bengal and other parts of Hindostan.	366	226	384	244	294	455	400
Indo-Britons	—	—	—	—	—	—	96
Buggies, Balanese, &c.	1851	1704	1442	1242	1252	1360	1726
Malays	4580	5130	5697	4790	5336	5750	7131
Javanese	—	38	146	267	355	634	595
Chinese	3317	3828	4279	6088	6210	7575	8517
African Negroes	—	—	2	5	—	—	37
Total	10683	11851	12905	13725	14885	17664	20917

Figure 1 Census data from colonial Singapore showing the population count as broken down into various "classes". (Martin, 1839)

As can be seen from Figure 1, three ethnic groups constituted the majority of the migrant population in British Singapore: the Chinese, the Malays, and the Indians. Today, the population of Singapore stands at 5.34 million, of which 3.87 million are residents. Of the latter group, the racial breakdown is as follows: 74.26% Chinese, 13.35% Malay, and 9.12% Indian (Department of Statistics, Singapore, 2014), reflecting a multicultural landscape that is a legacy of the country's history over the past two centuries.

Yet, it likely that there exists uncharacterized substructure within these three ethnic populations in Singapore. Most Singaporean Chinese originated from Southern China, but belong to various dialect groups, likely representing different ethnic groups. India was at that time also a British colony. Immigrants from India to Singapore hailed mostly from the southern part of the country, but with a significant minority (such as the Sikhs) coming from the north as well. It has also been found that many Singaporean Indians have a higher fraction of European ancestry as compared to native Indian populations. Malays are the indigenous people of the Malay Archipelago. The native Malays of the Singapore made up just a small proportion of the total Malay population on the island, the majority of whom migrated from Malaya (now Malaysia) and the Dutch East Indies (now Indonesia).

Characterizing the population substructure of Singapore can not only help illuminate its cultural history, it also has potential medical implications for Singaporeans. For example, uncovering population substructure could prevent spurious results in genetic association studies for disease. While many studies have previously explored the question of substructure in a global context, few have done so on a population that exists in such close geographic proximity, and therefore shares more genetic similarities.

The Singapore Genome Variation Project

With the advent of modern-day genomic sequencing, and the rapidly decreasing cost of the technology, it has now become feasible to collect large amounts of genetic data from populations that span different geographic areas of the world. For example, the 1000 Genomes Project is an international research effort to establish the most detailed database of worldwide human genetic variation, consisting of 697 individuals from 7 regions (The 1000 Genomes Project Consortium, 2010). The availability of such information has not only contributed significantly to our under-

standing of the genetic structure that underlie the differences between human populations, but has also aided in the search for genetic variants that may be associated with certain diseases or traits.

In a similar vein to the 1000 Genomes Project, the Singapore Genome Variation Project (SGVP) aims to provide a publicly-available resource for cataloging genetic variation within the Singaporean population, especially in the context of the three major ethnic groups. The data can obtained online at `http://www.statgen.nus.edu.sg/~SGVP`. After some initial cleaning up, SGVP contains data for a total 96 Chinese, 89 Malay, and 83 Indian individuals at about 1.5 million single nucleotide polymorphisms, or SNPs (Teo *et al.*, 2009). SNPs, representing a single base (A, C, G, T) change at a single locus, is the most common way of measuring variation within a genome.

In this project, we plan to first apply several unsupervised machine learning techniques on the SNP data to see if we can reveal any underlying population structure. Using this information, we can then also attempt to devise a prediction algorithm that can classify an individual into the correct ethnic group based on his or her genetic data.

Methods

Data processing

The initial step of this project involves processing the data into a format that is suitable for further analysis. Since humans are diploid organisms, SGVP has data on SNPs found on both copies of each individual's chromosomes. Thus, we have a total of 192 Chinese, 178 Malay, and 166 Indian haplotypes. For each autosome (non-sex chromosome), we first identify and retain only the SNPs that can be found in all three ethnic groups in order to ensure a common ground for comparison. The SNPs are represented by 1's and 0's, where 1 denotes the presence of a specific nucleotide base, and 0 denotes the presence of an alternative base. Here, all the SNPs are biallelic (only two possible variants exist).

We then concatenate all 22 autosomes into a single long string consisting of a total 1,369,502 SNPs. This is taken to be a single haplotype. Following this, we want to compare the pairwise distances between all possible haplotype pairs by calculating a distance matrix. The metric we choose to consider is the Hamming distance, which is given by the number of loci at which the two haplotypes do not share the same SNP, normalized by the total number of SNPs. Mathematically, the Hamming distance H between haplotypes a and b is

$$H_{ab} = H_{ba} = \frac{\sum_{i=1}^{m} \mathbf{1}\{a[i] \neq b[i]\}}{m},$$

where in this case $m = 1,369,502$. For example, the Hamming distance between two haplotypes that have the same SNP at all loci is 0, whereas the Hamming distance between two haplotypes that have different SNPs at all loci is 1.

Unsupervised learning methods

With the Hamming distance matrix calculated, we can now proceed to apply some unsupervised learning techniques to further analyze the data. We first use k-medoids clustering (Kaufman & Rousseeuw, 1987) with $k = 3$ to see if haplotypes belonging to the three ethnic groups can be correctly clustered together. k-medoids clustering is conceptually similar to k-means clustering, with the main difference being that in the former, each cluster's "center" must be chosen from the original set of data points, whereas there is no such restriction in the latter. Here, we choose to use k-medoids instead of k-means because the Hamming distance is a non-Euclidean distance measure, and the notion of a "mean" is less clearly defined within such a context. We use the `kcca` function found in the `flexclust` package in `R` to perform the clustering (Leisch, 2006).

In addition, we also want to examine the data using classical multidimensional scaling (MDS) on the Hamming distance matrix. Classical MDS, also known as principal coordinates analysis, finds a distance matrix D' between a set of points in some reduced dimensional space, that is as close as possible to the original distance matrix D in the full dimensional space. The advantage of this method over principal components analysis (PCA) is that it does not require a Euclidean distance measure, which the Hamming distance is not. Of interest is the fact that classical MDS yields the exact same results as PCA when the distance matrix is Euclidean

(Cox & Cox, 2001). We use the `cmdscale` function in `R` to implement classical MDS.

Finally, we consider a third unsupervised learning algorithm: hierarchical clustering. This is an agglomerative method that seeks to come up with a hierarchy of clusters of the haplotypes. The algorithm initializes with each haplotype in its own cluster. Pairs of clusters most similar to each other, as measured by the Hamming distance, are then sequentially fused, until only a single cluster remains. The distance between two clusters is taken to be the maximum of the pairwise distances between the haplotypes in each cluster, otherwise known as complete linkage. We use the `hclust` function in `R` to perform the clustering, and the `ColorDendrogram` method found in the `sparcl` package to plot the dendrogram (Witten & Tibshirani, 2010).

Results

k-medoids clustering

Using k-medoids clustering with $k = 3$, we found that one Chinese, Malay, and Indian haplotype were each respectively selected as the medoids. For the remainder of the haplotypes, we can define a classification based on the cluster they each belong to. For example, if a haplotype is assigned to a cluster with a "Chinese" medoid, then it is classified as Chinese. The confusion matrix arising from this process is shown in Table 1.

		predicted		
		Chinese	Malay	Indian
actual	Chinese	**191**	1	0
	Malay	8	**170**	0
	Indian	0	0	**166**

Table 1 Confusion matrix showing the 536 haplotypes as broken down into actual and predicted ethnic groups using k-medoids clustering.

As we can see from Table 1, k-medoids clustering gives relatively good performance when attempting to classify the three ethnic groups from this data set. All 166 of the Indian haplotypes are correctly classified, while the same is true for 99.5% and 95.5% of the Chinese and Malay haplotypes respectively. All misclassified Malay haplotypes are classified as Chinese, and vice versa.

k-medoids classification with unseen data

To see how k-medoids classification as described in the previous section works with previously unseen data, we randomly split the haplotypes into training and test sets. Overall, 134 Chinese, 125 Malay, and 116 Indian haplotypes are used for training, with the remainder reserved for testing. Each test haplotype is assigned to an ethnic group according to the medoid it is closest to. The average results over 10 trials is shown in Table 2.

		predicted		
		Chinese	Malay	Indian
actual	Chinese	**57.63**	0.37	0
	Malay	2.28	**50.72**	0
	Indian	0	0	**50**

Table 2 Confusion matrix showing the average performance, over 10 iterations, of k-medoids classification on unseen data.

Multidimensional scaling

We make use of MDS to obtain a better visualization of our data in lower dimensions. Figure 1 shows a plot of the haplotypes in 2D coordinate space.

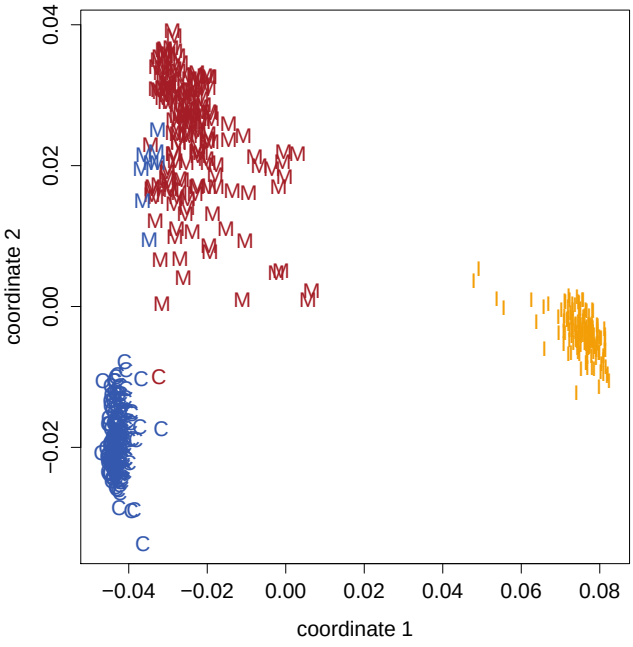

Figure 1 The 536 haplotypes in 2D space. Each individual point is labeled according its true ethnic group (C = Chinese, M = Malay, I = Indian), and is colored based on the cluster it is assigned to by the k-medoids algorithm, as per Table 1.

We see from the MDS plot that there is a good clustering of the three different ethnic groups in 2D coordinate space. The first (and major) coordinate separates out the Indians from the Chinese and Malays, and the second coordinate further differentiates the latter two groups from one another.

In addition, we can also observe, based on the how the points are colored, where the haplotypes that are placed into the incorrect cluster by k-medoids lie on the MDS plot. The 8 Malay haplotypes that are misclassified as Chinese are all on the left of coordinate 1, placing them closer to the bulk of the Chinese haplotypes. Similarly, the lone Chinese haplotype wrongly labeled as Malay is proximate to Malay haplotypes. Overall, the structure of the plot suggests that the position of a haplotype in 2D coordinate space is a strong predictor of the ethnic group it belongs to.

Hierarchical clustering

The dendrogram arising from hierarchical clustering is shown in Figure 2. Branches that fuse near the bottom of the tree are more similar to each other than branches that fuse closer to the top of the tree. Based on the tree topography, the appropriate value of the number of clusters, k, can then be decided upon *a posteriori*.

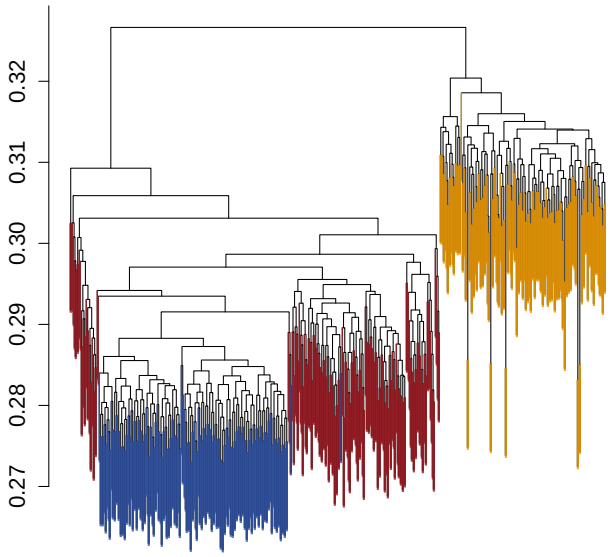

Figure 2 The 536 haplotypes in a dendrogram from hierarchical clustering, colored according to their true ethnic group (blue = Chinese, red = Malay, yellow = Indian). The vertical axis shows the Hamming distance at which clusters fuse together.

For this particular case, the choice of $k = 2$, which corresponds to taking the first split at the very top of the dendrogram, results in the separation of the Indian haplotypes from the Chinese and Malay haplotypes, which is the expected outcome given the results from k-medoids and MDS. However, when we take $k = 3$, the second top-most split occurs within the Indians. Therefore, hierarchical clustering appears to give an inferior performance as compared to the previous methods we have attempted.

Nevertheless, the fact that the Indian haplotypes have subdivisions that occur close to the top of the tree suggests potential presence of substructure within the Singaporean Indian population. To further investigate this claim, we repeat k-medoids clustering with $k = 4$.

k-medoids clustering with $k = 4$

Figure 3 shows a similar plot as Figure 1, but with the haplotypes now colored according to the clusters generated by k-medoids when $k = 4$.

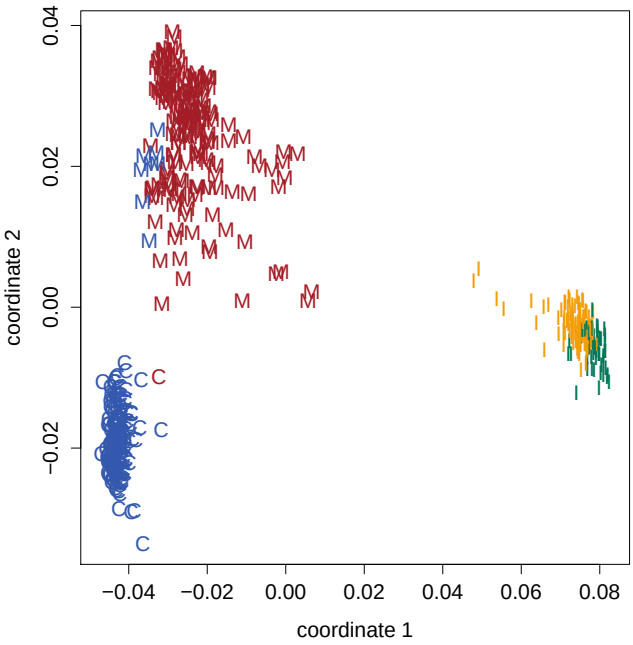

Figure 3 The 536 haplotypes in 2D space. labeled according their true ethnic groups, but now colored based on clusters produced by the k-medoids algorithm with $k = 4$.

We observe that the Chinese and Malay clusters remain the same, while the Indian haplotypes are split into two clusters, supporting the notion that there exists some substructure within Singaporean Indi-

ans. Also, the exact position of an Indian haplotype within the yellow cluster when $k = 3$ is a relatively good predictor of whether it gets assigned to the yellow or green cluster when $k = 4$, indicating consistency between the results from k-medoids and MDS.

Discussion and future work

In this project, we are interested in studying how genetic data can be used to identify Chinese, Malay, and Indian Singaporeans, as well as finding out if any additional population substructure exists beyond the ternary labels assigned. Previous studies have used PCA on raw SNP data to understand population substructure (Patterson e*t. al.*, 2006, Price *et. al.* 2008, Novembre *et al.* 2008). However, in our case, we condense the information found within the SNP data into a Hamming distance matrix. This reduces the number of dimensions our methods have to operate on, thereby improving runtimes.

We can see from the results of our MDS plot that coordinate 1 is the principal axis of variation, and represents the separation of the Indian haplotypes from the Chinese and Malay. This is corroborated by the correct classification of all Indian individuals using k-medoids. The MDS plot also shows a smaller separation between the Chinese and Malay haplotypes along coordinate 2. Accordingly, k-medoids exhibits a small number of misclassifications between these two groups. These results suggest that the three ethnic groups are genetically distinct enough such that a relatively straightforward method like k-medoids can achieve good classification performance. Furthermore, even as k-medoids clustering is an unsupervised learning method, once the medoids have been established using the training set, the test haplotypes can be assigned to a cluster based on their Hamming distances to the medoids, without the need to repeat the entire algorithm.

In a previous study (Novembre *et al.* 2008), a PCA plot of genetic variation within Europe corresponds to the actual map of Europe with remarkable accuracy, indicating that the principal components from genetic data can represent actual geographic distances between populations. Among China, the Malay Archipelago, and India, the latter is the most geographically distant relative to the other two locations. It is therefore unsurprising that our results show Indian haplotypes as being more genetically distinct. Furthermore, when k-medoids is performed with $k = 4$, a split within the Indian haplotypes is observed. While we do not have the geographic ancestry of the individuals in our data set apart from their reported ethnic group, we hypothesize that this split could correspond to Singaporean Indians originating from either Northern or Southern India. The large latitudinal distance of the country could explain why hierarchical clustering revealed substructure in the Indian haplotypes that is not present within the Chinese and Malays.

If it were true that the observed substructure within the Indians is due to latitude, then our MDS plot could be revealing an east–west split along coordinate 1 and a north–south split along coordinate 2. In future work, we could explore this question by introducing individuals known to be from Northern and Southern India, and observe if k-medoids clustering produces the expected patterns.

References

Cox, T. F., M. A. A. Cox, 2001. *Multidimensional Scaling* (pp. 43–44). London: Chapman & Hall.

Department of Statistics, Singapore, 2014. *Singapore in Figures* (p. 5). Singapore.

Kaufman, L., P. J. Rousseeuw, 1987. Clustering by means of medoids. In Y. Dodge (Ed.), *Statistical Data Analysis Based on the L1-Norm and Related Methods* (pp. 405–416).

Leisch, F., 2006. A toolbox for k-centroids cluster analysis. *Comput Stat Data An.* 51: 526–544.

Martin, R. M., 1839. *Statistics of the Colonies of the British Empire* (p. 410). London: W. H. Allen and Co..

Novembre, J. *et. al.*, 2008. Genes mirror geograpgy within Europe. *Nature* 456: 98–101.

Patterson, N. *et. al.*, 2006. Population structure and eigenanalysis. *PLoS Genet.* 2: e190.

Price, A. L. *et. al.*, 2008. Discerning the ancestry of European Americans in genetic association studies. *PLoS Genet.* 4: e236.

Teo, Y.-Y. *et. al.*, 2009. Singapore Genome Variation Project: A haplotype map of three Southeast Asian populations. *Genome Res.* 19: 2154–2162.

The 1000 Genomes Project Consortium, 2010. A map of human genome variation from population-scale sequencing. *Nature* 467: 1061–1073.

Witten, D. M., R. Tibshirani, 2010. A framework for feature selection in clustering. *J. Am. Statist. Assoc.* 105: 713–726.