Genetic Variation in miRNA Primary Transcripts

Moeen Bagheri[1,*]

[1]Department of Cell & Systems Biology, Department of Computer Science, University of Toronto, Toronto, Canada
[*]moeen.bagherigarekani@mail.utoronto.ca

ABSTRACT

This study was focused on the genetic variation in the primary transcripts of miRNAs between different human populations. A total of 1127 SNPs were examined in the study. These SNPs were used to estimate the degree of genetic differentiation between populations and construct a dendrogram. African populations were found to have the highest degree of genetic differentiation and according to the dendrogram, were diverged first. The next goal of this study was to find SNPs that are characteristic of certain populations and, hence, may have emerged due to positive selection. A total of 33 such SNPs were found.

Introduction

MicroRNAs (miRNA) are small, non-coding RNA molecules that are highly conserved and are involved in the regulation of gene expression.[1] MicroRNAs are known to be repressors and they regulate gene expression post-transcriptionally by binding to mRNA molecules, at the 3' untranslated region (UTR), and inhibiting protein production.[1] A single mRNA may be regulated by multiple miRNAs, and a single miRNA may regulate multiple mRNAs.[1]

MicroRNAs are negative regulators. Upon binding of miRNA to its mRNA target site, it can either induce translational repression, or cause the direct degradation of the mRNA molecule.[1] The most important factor determining the binding strength between the miRNA and mRNA molecules is the degree of base-pairing between the **seed region** of miRNA molecule, which is located at the 5' end, and the **target site** on the mRNA molecule, which is located at the 3' UTR.[1]

MicroRNAs have shown a high degree of evolutionary conservation, especially in the seed region. Moreover, orthologous miRNAs have a conserved function in the cell and recognize similar sets of mRNA target sites.[1] The importance of the conserved function of miRNAs can be seen in the fact that mutations in tumor suppressor miRNAs, known as 'oncomiRs', are associated with cancer and drive tumorigenesis.[2] Furthermore, tumor cells have a suppressed expression of miRNAs, which suggests that miRNA biogenesis impairment is a cause of cancer.[2]

Due to the important role of miRNAs as post-transcriptional regulators, they contribute fairly to the phenotypic and physiological differences between human populations. Hence, studying the variation in genes, that make up miRNAs, between different populations can help identify SNPs that are unique to certain populations. These SNPs may or may not be the result of positive selection, and can also confer population-specific risk to certain diseases.

In this study, genomes from different human populations were obtained from 1000 genomes project,[3] and two statistical measures, variance and F_{ST}, were compared in assessing the degree of genetic differentiation between these ethnicities. Then, A dendrogram was constructed from the results of the genetic differentiation. Moreover, two methods were assessed in estimating a genome-wide F_{ST} value, using F_{ST} values across all SNPs. Finally, SNPs that are characteristic of certain populations, and may contribute to the phenotypic and physiological difference between human populations were found.

Results

Populations

Different number of genomes were available in the 1000 genomes project[3] for each population (see Appendix, Table 1). A total of 2503 samples from 26 populations and 5 different demographic areas were included in the study.

Genetic differentiation

The aim of this study is to, first, estimate the degree of genetic differentiation between pairs of populations, based on the genetic variation in miRNA primary transcripts; second, find SNPs, within miRNA primary transcripts, with statistically significant difference in frequency between populations. As mentioned, these SNPs may be a result of positive selection.

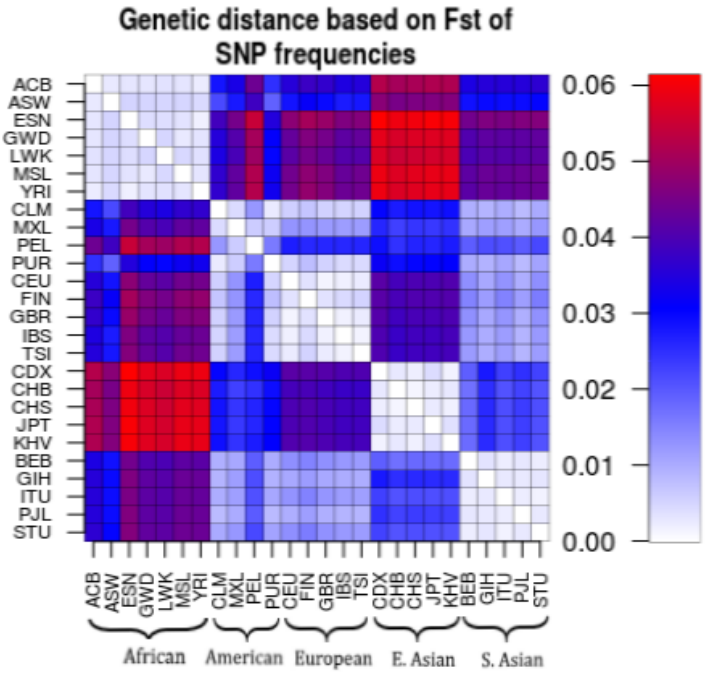


Figure 1. A heatmap of the degree of genetic differentiation between populations.

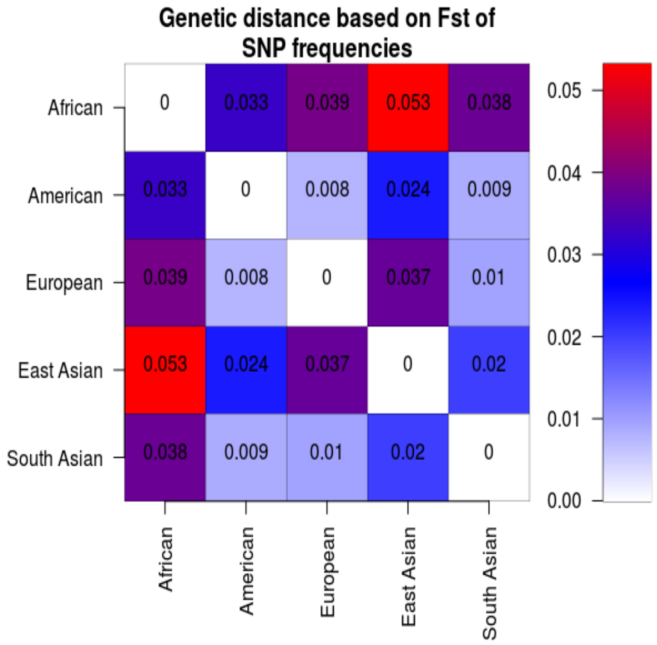


Figure 2. A heatmap of the degree of genetic differentiation between populations grouped by demographic areas.

1881 miRNA primary transcripts –hairpin precursor sequences that do not represent the full primary transcript, but rather a predicted stem-loop portion that includes the precursor miRNA– were included in the study. The study covered the somatic chromosomes (1-22) only (mitochondrial and sex chromosomes were omitted). A total of 1127 different variations with respect to the human genome (hg38) were found over many loci.

Pairwise comparison

To do a pairwise comparison, F_{st} was used to estimate the degree of genetic differentiation from SNP frequencies and a heatmap was constructed (Figure 1). Examining the figure, it is apparent that the degree of genetic differentiation between members of the same demographic group is very low. The highest F_{ST} value ($\sim$0.061) occurred between the African population ESN and the East Asian population CDX (see Table 1 for population codes). The closest populations with an F_{ST} of $\sim$0.001 were found to be the East Asian populations CHB and CHS, which makes sense since they are from the same demographic group. Furthermore, the lowest inter-demographic (between populations that are not in the same demographic group) F_{ST} value was between the American population PUR and the European population IBS ($\sim$0.004).

The pairwise comparison was then done between demographic groups (Figure 2). There were a total of 660 Africans, 347 Americans, 503 Europeans, 504 East Asians, and 489 South Asians included in the study. Africans were established to be the most genetically distant population by having the highest F_{ST} values. Americans, Europeans, and South Asians seem to be very close to each other. On the other hand, East Asians are considerably different than all other populations, but they are the closest to South Asians.

Finally, the populations were clustered and a dendrogram was constructed (Figure 3). As expected, populations within the same demographic area were clustered together. This is with the exception of PEL, which was not clustered with any demographic area. The dendrogram suggests that Africans were diverged first, followed by East Asians, Europeans, South Asians, and finally Americans.

Characteristic SNPs

This part was focused on finding **characteristic SNPs**, which we define as SNPs that are unique to certain populations, and hence, may have emerged as a result of positive selection. Such SNPs must be frequent in some populations, but rare in others. To find such SNPs, the graph in Figure 4 was constructed. The rows represent

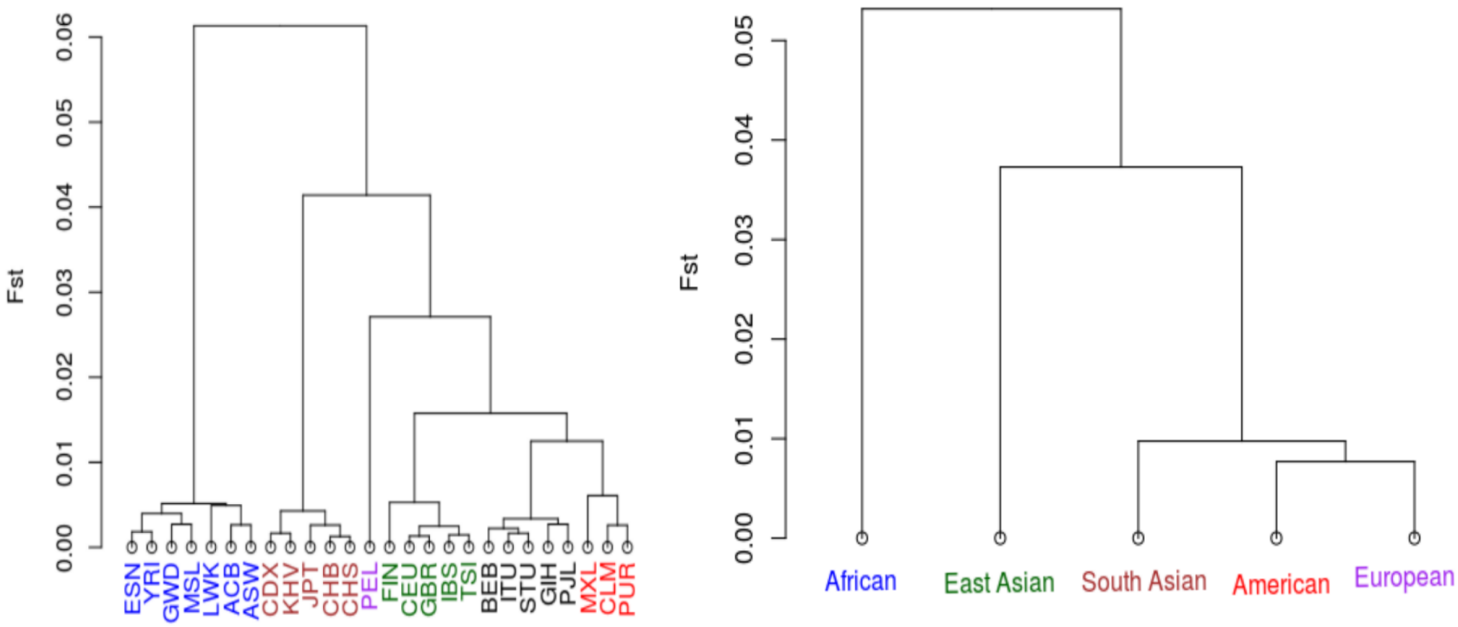


Figure 3. A dendrogram of the results of clustering of populations based on their F_{ST} estimates of genetic differentiation.

SNPs, the columns represent different populations, and the color represents F_{ST} values, which depicts how different the SNP frequency of a population is compared to all other populations.

Next, if a population contained an F_{ST} value bigger than 0.25 for a specific SNP, it was considered to have a significant difference/uniqueness in the frequency of that SNP compared to other populations. Higher F_{ST} values represent a higher degree of variation/uniqueness in the SNP frequency. Hence, all rows that contained an F_{ST} value greater than 0.25 in Figure 4 were extracted (Figure 5). As a result, a total of 33 characteristic SNPs were found and a table was created to list the genomic positions of all such SNPs, that have an F_{ST} value greater than 0.25 for at least one population (see Appendix, Table 2). The *populations* column lists the populations that have a significant difference/uniqueness in the frequency of that SNP. The last column of the table describes whether the presence or absence of the SNP is the characteristic of the populations.

Materials and Methods

Genomic data collection

Genomic data were obtained in compressed VCF format (.vcf.gz) from Phase III of the 1000 genomes project[3]. The VCF data was separated for each chromosome, with each VCF file containing the data for all populations. Since R could not handle such big files, each VCF file was divided into several smaller files by subsetting the data of each population using **BCFtools** and running the following command in bash, once for each chromosome:

```
bcftools view --min-ac=1
--force-samples -Oz -S samples.txt
chr?.vcf.gz > chr?-subset.vcf.gz
```

where *samples.txt* is a file containing the sample names to be subsetted (one sample per line), *chr?.vcf.gz* is the original compressed VCF file for a chromosome downloaded from 1000 genomes project, and *chr?-subset.vcf.gz* is the name of the output file (? replaced with chromosome number). The option *–min-ac=1* tells BCFtools to subset rows that have a variation in at least one of the samples being subsetted, which helps with reducing the size of the output file. The sample names for each population were obtained from the 1000 genomes project website[3]. However, the data for some of the samples was unavailable in the VCF files downloaded. Therefore, the option *–force-samples* was used to tell BCFtools to skip samples from *sample.txt* that are not in *chr?.vcf.gz*.

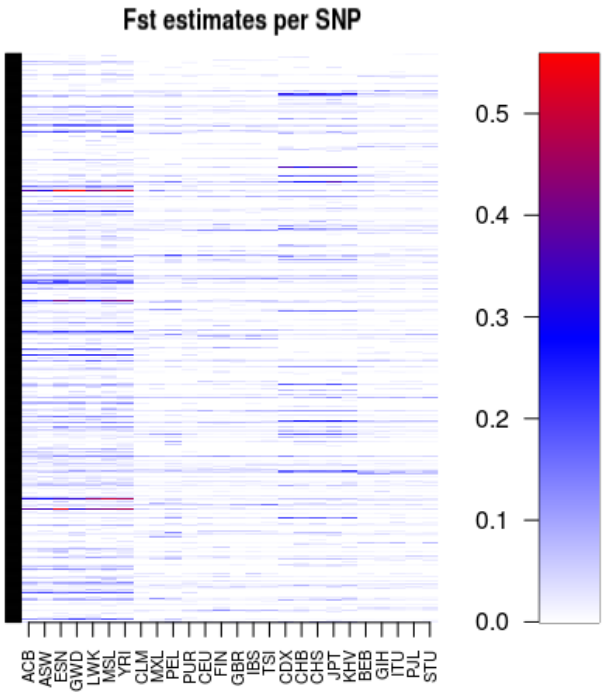


Figure 4. Heatmap of values representing the degree of difference/uniqueness of populations in the frequency of each SNP. The rows represent different SNPs.

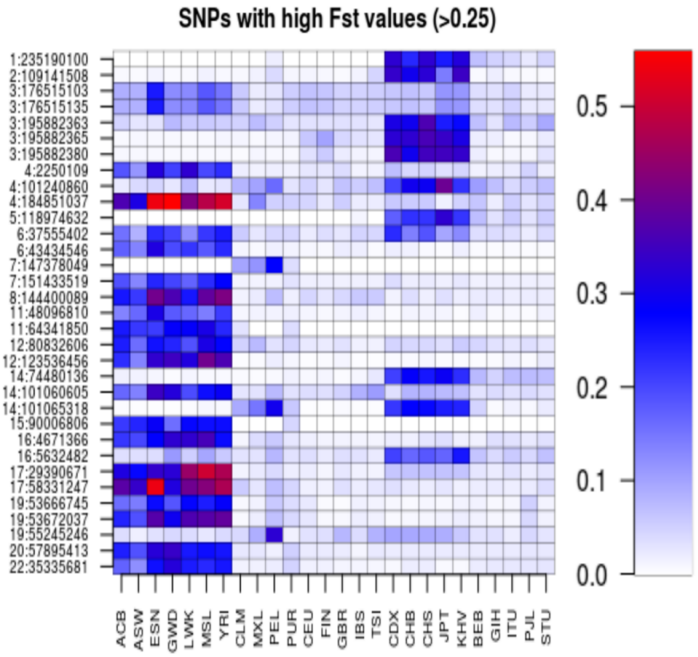


Figure 5. Heatmap of SNPs with a high degree of variation between populations.

miRNA data collection

The genome coordinates of genes coding for miRNAs were obtained from Release 21 of *miRBase*[4–9]. The original file contained genome coordinates for both primary transcripts and the mature miRNAs. Using a simple R script, this file was loaded into R as a table and the rows corresponding to *miRNA primary transcripts* and the columns corresponding to the chromosome number, start position, end position, and miRNA name were subsetted.

Finding variations

In order to find variations within miRNA primary transcripts, miRNA coordinates data and subsetted VCF files were read into R. This process was done for a single population and a single VCF file at a time. Using a for-loop, VCF files were searched for variations within the genomic range specified in the miRNA coordinates file. In each iteration of the loop, the miRNA coordinates file was subsetted to include only the data for the current chromosome number (VCF file) being examined, which helped making this process faster. Next, three quantities, SNP count, variation count, and total-variation count, were measured. SNP count is the abundance of an SNP in the population, variation count is the number of people within a population that contain a specific variation, and total-variation count is the number of people within a population that contain any variation at a specific locus. The results were saved in a .tsv file containing eight columns: *chromosome (CHR), name(NAME), position (POS), alternative allele (ALT), SNP count (altCOUNT), variation count (varCOUNT), total-variation count (totCOUNT)* and *size of population (SIZE)* (see Appendix, Figure 9). For loci with more than one alternative allele, each allele (SNP) was recorded separately in different rows. To calculate the frequency of an SNP in a population, *altCOUNT* was divided by the double of the population size, since each person contains two alleles. Variation and total-variation frequencies were simply calculated by diving the corresponding *COUNT* by the sample size.

Pairwise comparison

This part of the study was focused on comparing pairs of populations and measuring the degree of genetic differentiation between them. Two statistical measures, variance and F_{ST}, were used to estimate the the degree of genetic differentiation between two populations based

on SNP frequencies, variation frequencies, and total-variation frequencies. Also, two different methods, *ratio of averages* and *average of ratios*, were used to combine estimates of F_{ST} values across all loci to obtain a genome-wide F_{ST} value for pairs of populations.

Pairwise comparison based on variance

First, the variance (σ_s^2) in the frequency of SNPs in two populations was used to estimate the degree of genetic differentiation (GD) between the two populations.

$$GD = \sigma^2 \tag{1}$$

To do so, first, the degree of variation in the frequency of a single SNP_k between two populations was obtained by calculating the variance of the frequencies of that SNP in the two populations. For example, if $p_{i,k}$ is the frequency of SNP_k in population i, then the variance of SNP_k frequencies for two populations i and j was calculated by the formula:

$$\sigma^2_{k,ij} = \frac{(p_{i,k} - \mu_{ij})^2 + (p_{j,k} - \mu_{ij})^2}{2} \tag{2}$$

where $\mu_{ij} = \frac{p_{i,k} + p_{j,k}}{2}$ is the average of SNP frequencies in the two populations. A variance was calculated for all SNPs and all possible pairs of populations using formula (2). Then, to obtain a genome-wide estimate of genetic differentiation for the two populations, all variances of the two populations were averaged across all SNPs:

$$\sigma^2_{ij} = \frac{\sum\limits_k \sigma^2_{k,ij}}{n_k} \tag{3}$$

where n_k denotes the total number of SNPs. This was done for all pairs of populations to obtain a 26×26 matrix, which was plotted using the R function `image()`. This process was repeated two more times using variation frequencies and total-variation frequencies instead of SNP frequencies (Figure 6).

Pairwise comparison based on F_{ST}

The second statistical measure used to estimate the degree of genetic differentiation (GD) between two populations was Fixation Index (F_{ST}).

$$GD = F_{ST} \tag{4}$$

To calculate the F_{ST} between two populations based on SNP frequencies, Wright's formulation of F_{ST} was

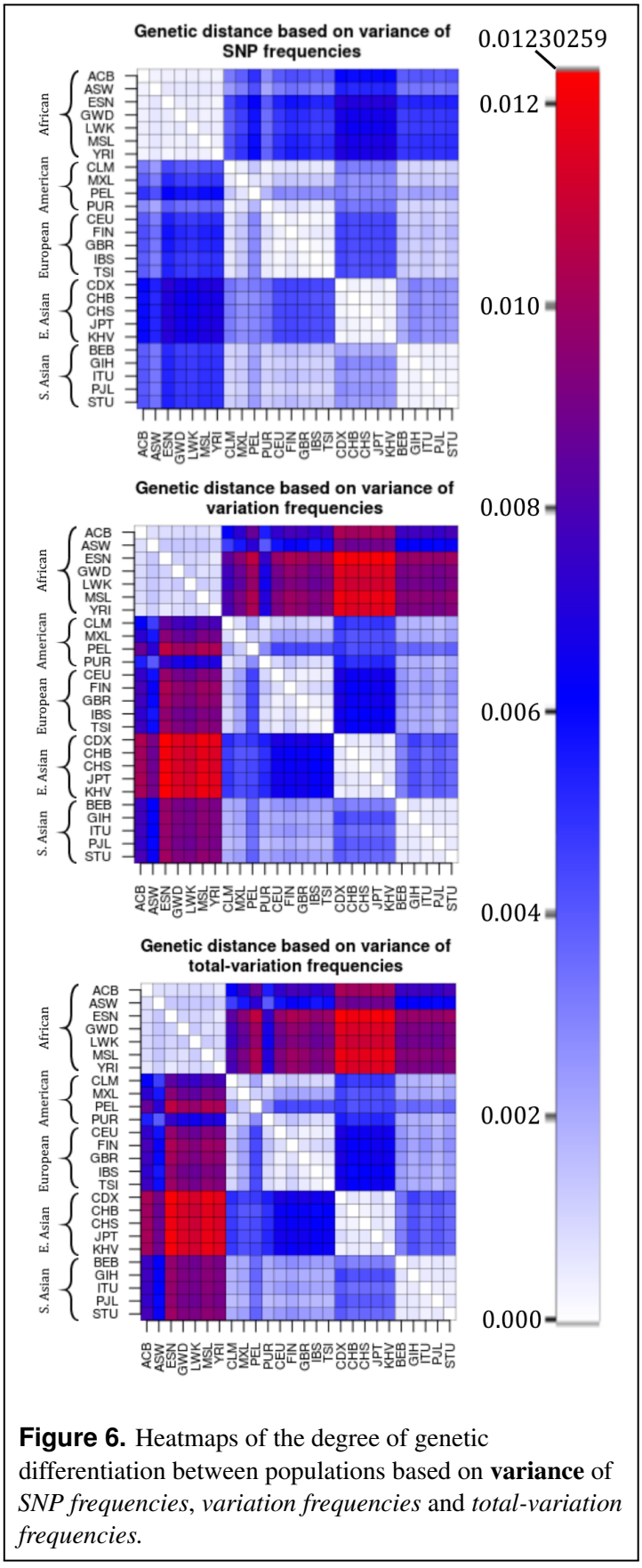


Figure 6. Heatmaps of the degree of genetic differentiation between populations based on **variance** of *SNP frequencies*, *variation frequencies* and *total-variation frequencies.*

used.[10] If s_i is the size of population i, $p_{i,k}$ is the frequency of SNP$_k$ in population i, and w_i is the proportion of the population in population i, such that:

$$w_i = \frac{s_i}{\sum_i s_i} \qquad and \qquad \sum_i w_i = 1 \tag{5}$$

then $\overline{p}_k$, the weighted frequency of SNP$_k$ in the total population, can be calculated using the formula:

$$\overline{p}_k = \sum_i w_i p_{i,k} \tag{6}$$

and then, F_{ST} can be estimated for SNP$_k$ by the following formula:[10]

$$F_{ST,k} = \frac{\sigma_k^2}{\overline{p}_k(1-\overline{p}_k)} \tag{7}$$

where σ_k^2 is the variance of the frequencies of SNP$_k$ in the two populations. Note that since:

$$\sigma_k^2 \leq \overline{p}_k(1-\overline{p}_k) \tag{8}$$

we can infer that the value of F_{ST} ranges from 0 - 1.[10] It is important to note that when calculating the F_{ST} value for two populations, the two populations were treated as if they formed the entire population. Hence, the F_{ST} value for two populations i and j for a single SNP$_k$ was calculated using the formula:

$$F_{ST,k} = \frac{var(p_{i,k}, p_{j,k})}{(w_i p_{i,k} + w_j p_{j,k})(1 - w_i p_{i,k} + w_j p_{j,k})}. \tag{9}$$

Using formula (9), F_{ST} values were calculated for all SNPs between two populations.

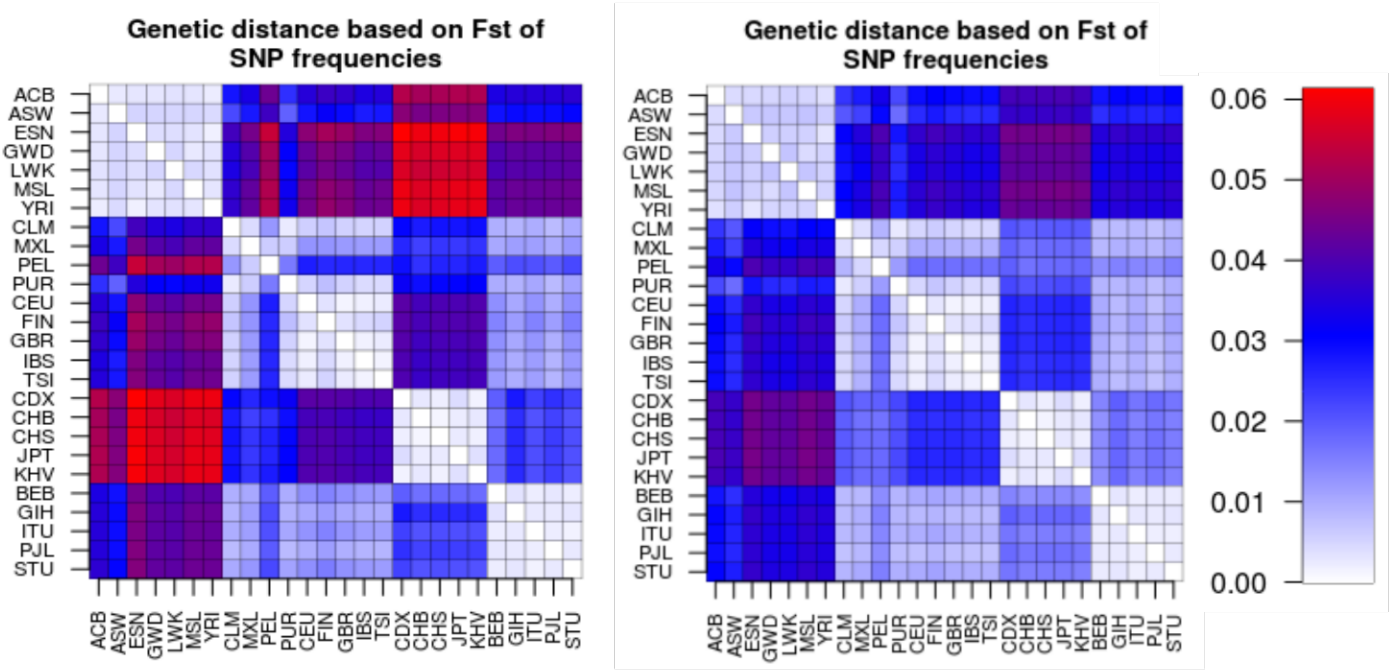
Next, to obtain a genome-wide estimation of the genetic differentiation of the two populations, two different methods were tested to obtain a genome-wide estimation of F_{ST} using the F_{ST} values of all SNPs (Figure 7).[11] The first method (*ratio of averages*) was to calculate the average of variances and $\overline{p}_k$ values, and use formula (7) to calculate a genome-wide value for F_{ST} (Figure 7a):

$$\sigma^2 = \frac{\sum_k \sigma_k^2}{n_k} \qquad and \qquad \overline{p} = \frac{\sum_k \overline{p}_k}{n_k} \tag{10}$$

then:

$$F_{ST} = \frac{\sigma^2}{\overline{p}(1-\overline{p})}. \tag{11}$$

The second method (*average of ratios*) was to average



(a) Genome-wide F_{ST} estimate generated by taking the *ratio of averages*.

(b) Genome-wide F_{ST} estimate generated by taking the *average of ratios*.

Figure 7. Heat-maps generated by using two methods for obtaining genome-wide F_{ST} estimates from all SNPs.

F_{ST} estimates for all SNPs to obtain a single genome-wide F_{ST} value (Figure 7b):

$$F_{ST} = \frac{\sum_k F_{ST,k}}{n_k} \tag{12}$$

where n_k is the total number of SNPs.

Genome-wide F_{ST} values were obtained for all pairs of populations to create 26×26 matrices and plotted using R (Figure 7). The two methods were compared and *ratio of averages* was chosen as the method of choice (See Discussion).

This process was repeated two more times using variation frequencies and total-variation frequencies instead of SNP frequencies (Figure 8). However, as stated before, the *ratio of averages* was used to calculate genome-wide estimates of F_{ST}. All variance and F_{ST} graphs were then compared and F_{ST} estimation based on SNP frequencies was chosen as the method of choice for measuring genetic differentiation (See Discussion).

Clustering and dendrogram construction

Next, the populations were clustered using F_{ST} estimations of genetic differentiation based on SNP frequencies. To do so, first, the 26×26 matrix of pairwise F_{ST} values was converted into a *dist* object using the R function `as.dist()`. Then, the R function `hclust()` was used to cluster the populations based on the *dist* object, as follows:

```
> clust = hclust(as.dist(pFstMatrix))
```

where *pFstMatrix* is the matrix of F_{ST} values. Then the object, *clust*, returned by `hclust()` was converted into a *dendrogram* object and plotted:

```
> Dendro = as.dendrogram(Clust)
> plot(Dendro)
```

Finding characteristic SNPs

To compare the frequency of an SNP between populations, the median of all pairwise F_{ST} values for a single SNP and a single population –representing the pairwise degree of variation of SNP frequencies of that population with other populations– was calculated to obtain a single F_{ST} value for each population and each SNP (median of 25 F_{ST} values). For example, if k specifies a particular SNP and i,j are populations, let $F_{ST,ijk}$ be the pairwise F_{ST} estimate for i and j for SNP_k, then:

$$F_{ST,i,k} = median_{\{\forall j | j \neq i\}}(F_{ST,ijk}) \tag{13}$$

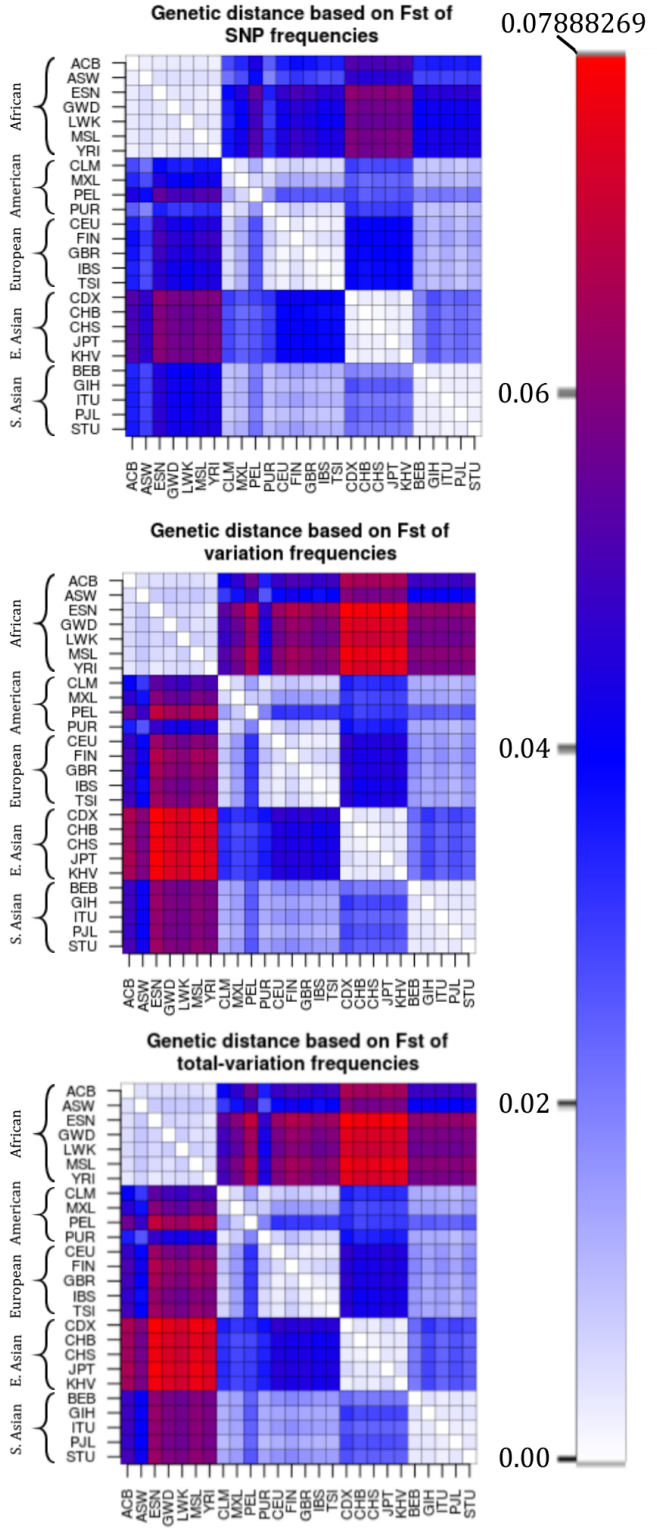


Figure 8. Heatmaps of the degree of genetic differentiation between populations based on $\mathbf{F_{ST}}$ estimates of *SNP frequencies*, *variation frequencies* and *total-variation frequencies*.

where $F_{ST,i,k}$ is the F_{ST} value for population i and SNP_k. This process was repeated for all populations and for all SNPs containing a variation to create a 1127×26 matrix (Figure 4). The rows of this matrix represent different SNPs, the columns represent the populations, and the value at row k, column i is $F_{ST,i,k}$ –the degree of difference/uniqueness in the frequency of SNP_k of population i compared with all other populations.

Then, to find characteristic SNPs –SNPs that are significantly abundant in some populations but rare in others– 0.25 was chosen as the cutoff between F_{ST} values that signify unique versus common SNP's. The value 0.25 was chosen since an F_{ST} value greater than 0.25 is considered to represent a high degree of genetic differentiation.[12] Hence, all SNPs (rows of the matrix) that contained at least one F_{ST} value greater than 0.25 for a population (columns of the matrix) were subsetted from this matrix. If the F_{ST} value for a certain population and a specific SNP is greater than 0.25, that SNP can be considered to be unique for that population. The newly constructed matrix represents all such SNPs (Figure 5). Table 2 was then constructed to summarize the results obtained from this matrix.

Discussion

Mitochondrial and sex chromosomes were not included in the study due to inconsistency in their data. For example, the data for some samples were unavailable for the sex chromosomes, making the population size smaller for these chromosomes, and making the results inconsistent. Mitochondrial chromosome does not contain any genes for miRNA's, so it was unnecessary to be included in the study.

First, I aimed to measure the genetic differentiation between different pairs of populations. To do so, first, the pairwise variance between different populations was calculated and plotted (Figure 6). These graphs represent how closely-related two populations are. The variance between two populations were found by averaging over all variances between those populations for all SNPs. Averaging was chosen over taking the median since I wanted the outliers to have a strong effect on the genetic differentiation estimate between two populations, because it is these outliers that contribute to the difference between populations. Although, the variance values have a geometric distribution and taking the median hints to be the right choice, taking the median here resulted in very low genetic differentiation values (close to 0) between all populations, which made their comparison difficult.

On the other hand, the downside of using variance is that it does not take into account the size of each population. To resolve this, F_{ST} was used to calculate the degree of genetic differentiation (Figure 8). F_{ST} also uses variance to estimate genetic differentiation, however, in F_{ST}, the frequencies of SNPs are weighted by the proportion of the population contained in each population (in the term $\overline{p}$). Looking at the pairwise variance and F_{ST} figures (Figures 6 and 8), the three different frequency measures (SNP, variation, and total-variation frequencies) seem to agree with each other, but variation and total-variation frequencies seem to emphasize the degree of genetic differentiation for both variance and F_{ST}. However, since comparing SNP frequencies is a more accurate measure for estimating the degree of genetic differentiation, and the ultimate goal of this study was to find SNPs that are characteristic of certain populations, SNP frequencies were chosen to be used for the rest of the study. Moreover, since F_{ST} weights SNP frequencies by the size of the populations, it is a more precise method for estimating the degree of genetic differentiation than variance, and hence, it was chosen as the method of choice. Furthermore, the two methods for obtaining a genome-wide F_{ST} were evaluated and since using the *average of ratios* method underestimates F_{ST} values, the *ratio of averages* was chosen as the method of choice.[11]

The first thing to notice in the heatmap (Figure 1) is the white diagonal, implying that each population is not different from itself. Next, it can be observed that the degree of genetic differentiation between members of the same demographic area is very low, which shows the fact that these populations are genetically close to each other. One important thing to note is that the highest estimated F_{ST} value (Figure 1) is close to 0.06, which is not a very high number. The consensus on interpretation of F_{ST} values is that values less than 0.05 represent little genetic differentiation, 0.05 - 0.015 is moderate, 0.15-0.25 indicates great genetic differentiation, and a value greater than 0.25 is considered to imply very high genetic differentiation[12]. This suggests that human populations are not too different from one another, and that there must be a moderate amount of interbreeding in the populations. Although African populations have the highest degree of genetic differentiation, their genome is still very similar to other populations. To confirm these relationships, the populations were clustered and

a dendrogram was constructed (Figure 3). Surprisingly, population PEL was not clustered with any demographic area. Looking at Figure 1, it can be seen that population PEL is fairly different than other American populations. For example, all American populations are shown to be close with Europeans and South Asians, with the exception of PEL. This result could imply that the population PEL has been relatively isolated compared to other American populations.

To carry out a better comparison between populations, pairwise comparison was performed among demographic areas (Figure 2). This graph confirms that Africans are the most differentiated population compared to all other populations. East Asians seem to be fairly differentiated, as well. On the other hand, Americans, Europeans, and South Asians appear to be closely-related to each other.

The next goal was to find loci that are characteristic of a certain population. The colors in Figure 4 represent the overall degree of genetic differentiation (derived from F_{ST}) of a population with the rest of the population. The F_{ST} value of a single SNP for a population was calculated by taking the median of all pairwise F_{ST} values between that population and all other populations (median of 25 F_{ST} values). In this case, the median was chosen instead of the average since I want to find whether a population has a high or low degree of genetic differentiation with more than half of the populations. Taking the average here would cause outliers to put a strong influence on the overall F_{ST} value, which is not desired. For example, for six populations $[A,B,C,D,E,F]$ and a single SNP, if populations A, B, C, and D have F_{ST} values of $[0.01,0.01,0.01,0.30,0.30]$, and populations E and F have F_{ST} values of $[0.30,0.30,0.30,0.30,0.01]$, this SNP would be a good characteristic for populations E and F, but not for A, B, C, and D (since the SNP frequency in A, B, C, and D is similar to most populations, but in E and F, it is different than most populations). With that in mind, taking the average of F_{ST} values results in an overall F_{ST} value of $\sim$0.24 for E and F, But taking the median gives a value of 0.30. In such a case, using average would not identify this SNP as a characteristic for populations E and F (with a cutoff of 0.25), whereas taking the median will. In other words, averaging F_{ST} values here would show the degree of genetic differentiation of a population with all other populations **on average**, whereas, median F_{ST} shows what the degree of genetic differentiation of the population is with **most** (more than half) of the populations, which is what is needed in this case.

The ultimate goal here was to find SNPs that are unique to certain populations and, hence, may be a result of positive selection. To do so, 0.25was chosen as the cutoff between F_{ST} values that represent common versus uncommon SNPs (Figure 5). The assumption here is that if a populations has an F_{ST} value higher than 0.25 for an SNP, then that SNP can be considered to be unique to that population. Such SNPs could have emerged as a result of positive selection and may cause population specific susceptibility to certain diseases.

Future studies can investigate the influence of these SNPs on the phenotype. These SNPs are expected to contribute to the difference between human populations, such as skin color, hair color, and eye shape. Additionally, these SNPs could contribute to the physiological difference between populations as well.

R Code

The R code used in this study can be found on GitHub[13]. github.com/bagherig/miRNA.git

References

1. Li, J. & Zhang, Z. mirna regulatory variation in human evolution. *Trends Genet.* **29(2)**, 116–124 (2013).
2. Lin, S. & Gregory, R. I. mirna regulatory variation in human evolution. *Nat. Rev. Cancer* **15**, 321–333 (2015).
3. Consortium, T. . G. P. A global reference for human genetic variation. *Nat.* **526**, 68–74 (2015).
4. mirbase. URL http://www.mirbase.org/. **Release 21**.
5. Kozomara, A. & Griffiths-Jones, S. mirbase: annotating high confidence micrornas using deep sequencing data. *Nucleic Acids Res.* **42**, D68–D73 (2014). URL +http://dx.doi.org/10.1093/nar/gkt1181. DOI 10.1093/nar/gkt1181. /oup/backfile/content_public/journal/nar/42/d1/10.1093/nar/gkt1181/2/gkt1181.pdf.
6. Griffiths-Jones, S. The microrna registry. *Nucleic Acids Res.* **32(Database issue)**, D109–11 (2004). DOI 10.1093/nar/gkh023.
7. Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A. & Enright, A. J. mirbase: microrna

sequences, targets and gene nomenclature. *Nucleic Acids Res.* **34**, D140–D144 (2006). URL +http://dx.doi.org/10.1093/nar/gkj112. DOI 10.1093/nar/gkj112. /oup/backfile/content_public/journal/nar/34/suppl_1/10.1093/nar/gkj112/2/gkj112.pdf.

8. Griffiths-Jones, S., Saini, H. K., van Dongen, S. & Enright, A. J. mirbase: tools for microrna genomics. *Nucleic Acids Res.* **36**, D154–D158 (2008). URL +http://dx.doi.org/10.1093/nar/gkm952. DOI 10.1093/nar/gkm952. /oup/backfile/content_public/journal/nar/36/suppl_1/10.1093_nar_gkm952/2/gkm952.pdf.

9. Kozomara, A. & Griffiths-Jones, S. mirbase: integrating microrna annotation and deep-sequencing data. *Nucleic Acids Res.* **39**, D152–D157 (2011). URL +http://dx.doi.org/10.1093/nar/gkq1027. DOI 10.1093/nar/gkq1027. /oup/backfile/content_public/journal/nar/39/suppl_1/10.1093_nar_gkq1027/2/gkq1027.pdf.

10. Nagylaki, T. Fixation indices in subdivided populations. *Genet.* **148(3)**, 1325–1332 (1998).

11. Bhatia, G., Patterson, N., Sankararaman, S. & Price, A. Estimating and interpreting F_{st}: The impact of rare variants. *Genome Res.* **23(9)**, 1514–1521 (2013).

12. Spencer, H. G. Population genetics. *Biom.* **2**, 317–338 (2009).

13. Github. URL https://github.com/.

Appendix

CHR	NAME	POS	ALT	altFREQ	varFREQ	totFREQ	SIZE	
1	hsa-mir-429	1169030	A	1	1	1	86	
1	hsa-mir-6726	1296153	G	147	85	85	86	
1	hsa-mir-4689	5862682	G	1	1	1	86	
1	hsa-mir-4689	5862716	T	2	2	2	86	
1	hsa-mir-4689	5862719	T	2	2	2	86	
1	hsa-mir-34a	9151750	T	69	56	56	86	
1	hsa-mir-34a	9151774	T	1	1	1	86	

Figure 9. A portion of the .tsv file produced by R after calculating allele frequencies at each position within miRNA primary transcripts.

Code	Population	Size	Demographic Area
ACB	African Caribbean in Barbados	95	African
ASW	African Ancestry in Southwest US	61	African
ESN	Esan in Nigeria	99	African
GWD	Gambian Mandinka	113	African
LWK	Luhya in Webuye, Kenya	99	African
MSL	Mende in Sierra Leone	85	African
YRI	Yoruba in Ibadan, Nigeria	108	African
CLM	Colombian in Medellin, Colombia	94	American
MXL	Mexican Ancestry in Los Angeles, California	64	American
PEL	Peruvian in Lima, Peru	85	American
PUR	Puerto Rican in Puerto Rico	104	American
CEU	Utah residents (CEPH) with Northern and Western European ancestry	99	European
FIN	Finnish in Finland	99	European
GBR	British in England and Scotland	91	European
IBS	Iberian populations in Spain	107	European
TSI	Tuscany in Italy	107	European
CDX	Chinese Dai in Xishuangbanna, China	93	East Asian
CHB	Han Chinese in Beijing, China	103	East Asian
CHS	Han Chinese South	105	East Asian
JPT	Japanese in Tokyo, Japan	104	East Asian
KHV	Kinh in Ho Chi Minh City, Vietnam	99	East Asian
BEB	Bengali in Bangladesh	86	South Asian
GIH	Gujarati Indian in Houston, TX	103	South Asian
ITU	Indian Telugu in the UK	102	South Asian
PJL	Punjabi in Lahore, Pakistan	96	South Asian
STU	Sri Lankan Tamil in the UK	102	South Asian

Table 1. The code, population name, size and demographic area of all groups included in the study.

Chromosome	Position	Allele	populations	Presence/Absence
1	235190100	T	CDX, CHS, JPT, KHV	Presence
2	109141508	C	CDX, CHB, CHS, KHV	Presence
3	176515103	A	ESN	Presence
3	176515135	A	ESN	Presence
3	195882363	T	CHS	Presence
3	195882365	T	CDX, CHB, CHS, JPT, KHV	Presence
3	195882380	A	CDX, CHS, JPT, KHV	Presence
4	2250109	C	ESN, LWK	Presence
4	101240860	T	JPT	Presence
4	184851037	A	ACB, ESN, GWD, LWK, MSL, YRI	Absence
5	118974632	C	JPT	Presence
6	37555402	C	YRI	Presence
6	43434546	A	ESN	Presence
7	147378049	A	PEL	Presence
7	151433519	G	YRI	Presence
8	144400089	T	ESN, GWD, MSL, YRI	Absence
11	48096810	C	ESN	Absence
11	64341850	A	ACB, GWD, LWK, MSL	Presence
12	80832606	T	ESN, MSL, YRI	Presence
12	123536456	CT	GWD, MSL, YRI	Absence
14	74480136	C	CHB, CHS, JPT	Presence
14	101060605	A	ESN, GWD, MSL, YRI	Absence
14	101065318	G	PEL, CHB, CHS	Absence
15	90006806	T	ESN, LWK, MSL, YRI	Presence
16	4671366	AAATT	MSL	Absence
16	5632482	T	KHV	Presence
17	29390671	A	LWK, MSL, YRI	Presence
17	58331247	A	ACB, ESN, LWK, MSL, YRI	Presence
19	53666745	C	ESN, LWK, YRI	Absence
19	53672037	C	ESN, LWK, MSL, YRI	Absence
19	55245246	A	PEL	Presence
20	57895413	C	ESN, GWD, LWK, MSL, YRI	Presence
22	35335681	G	ESN, GWD, YRI	Absence

Table 2. The genomic positions and their corresponding alleles that are unique to certain populations, which are listed in the *populations* column. The column *Presence/Absence* specifies whether the presence or absence of the allele is the characteristic of the populations.