August 22, 2017
BCB430Y

Genetic Variation in miRNA Primary Transcripts

By Moeen Bagheri
Supervisor: Zhaolei Zhang

UNIVERSITY OF TORONTO

Overview

- Project Description
  - Objective
  - Results
- Materials and Methods
  - Finding genetic variations
  - Estimating degree of variation
  - Finding characteristic SNPs
- Results
  - Pairwise degree of variation
  - Clustering & dendrogram
  - Characteristic SNPs
- Questions & Discussion

Project Description

Objectives

- Investigate the genetic variation in human populations based on data from the Phase III of 1000 genomes project.
- Identify SNPs that are unique to certain populations and, hence, could have emerged due to positive selection.

Results

- Heat-maps
- Dendrograms

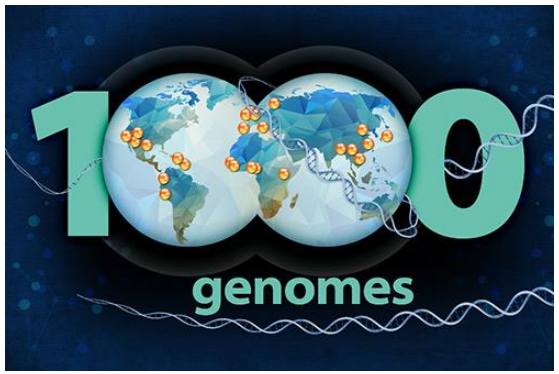
UNIVERSITY OF TORONTO

Preparing Data and Finding Genetic Variations

Materials & Methods

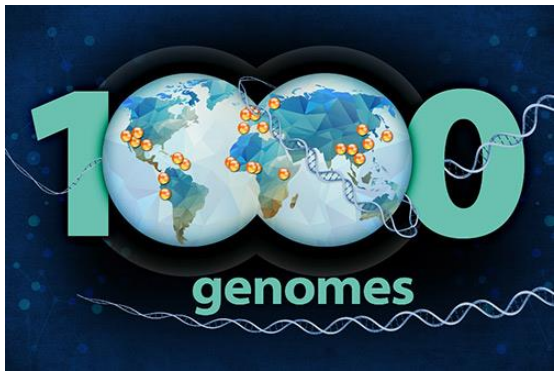
UNIVERSITY OF TORONTO

Data collection



- miRNA genome coordinates
  - obtained from miRBase v21: ***mirbase.org***
  - file format: *Generic Feature Format (*.gff3)
- Human genomes
  - obtained from phase III of 1000 genomes project: ***internationalgenome.org***
  - file format: Compressed Variant Call Format (.vcf.gz)
- Reference genome: GRCh38

Data collection



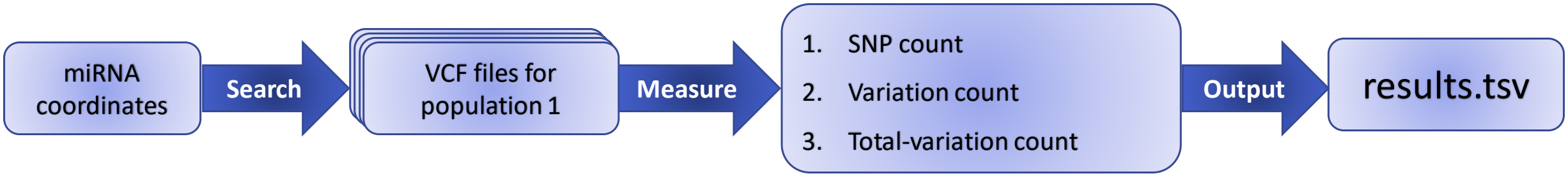
- miRBase:
  - Extract the genomic coordinates of miRNA primary transcripts.
  - A total of 1881 miRNA primary transcripts were obtained.
- 1000 genomes:
  - original VCF files were separated based on chromosome number.
  - The genomic data for each population was then separated using **BCFtools**.
  - A total of 2503 samples from 26 populations and 5 different demographic groups were obtained.

Finding genetic variations

- The coordinate range of each miRNA primary transcript was searched in the VCF files for genomic variations.
- Using R programming language
- A total of 1127 SNPs were found.

Three measures:

- SNP count
- Variation count
- Total-variation count



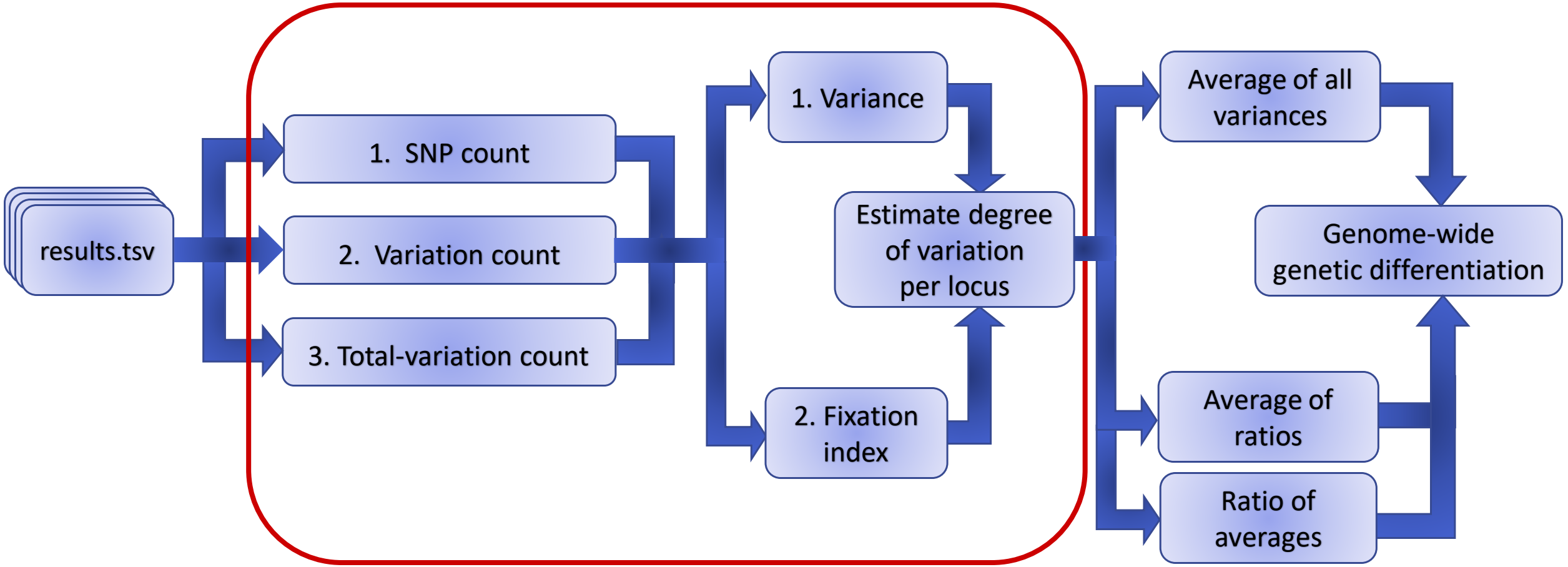
UNIVERSITY OF TORONTO

Estimating Degree of Genetic Differentiation

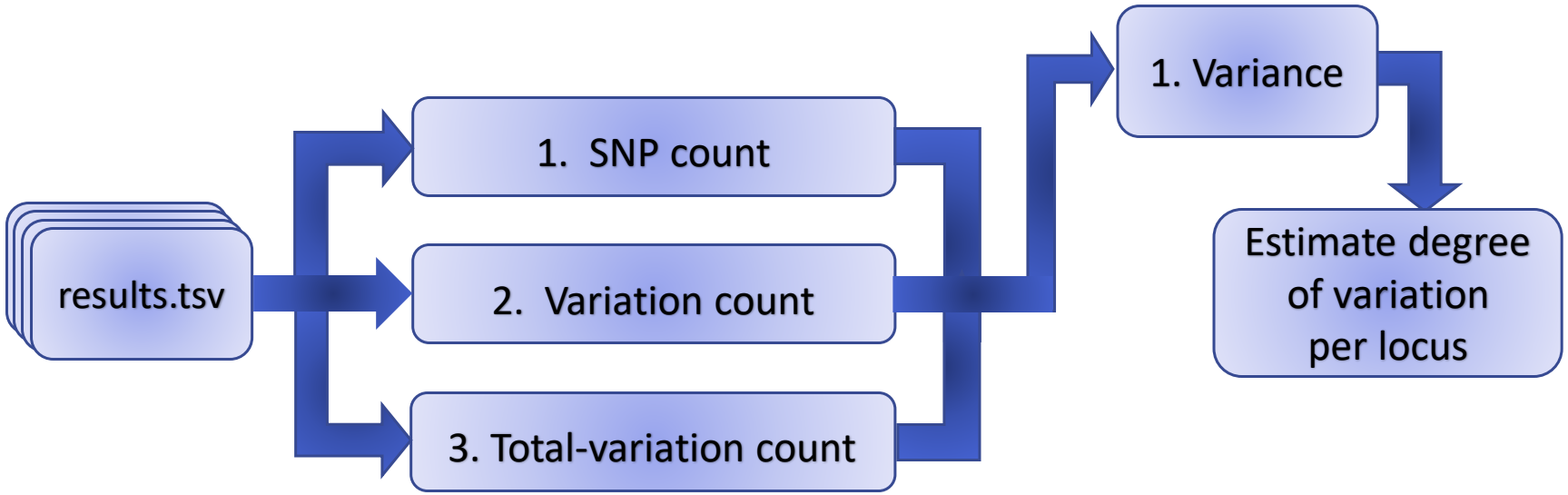
Materials & Methods

UNIVERSITY OF TORONTO

Estimating degree of variation



Variance



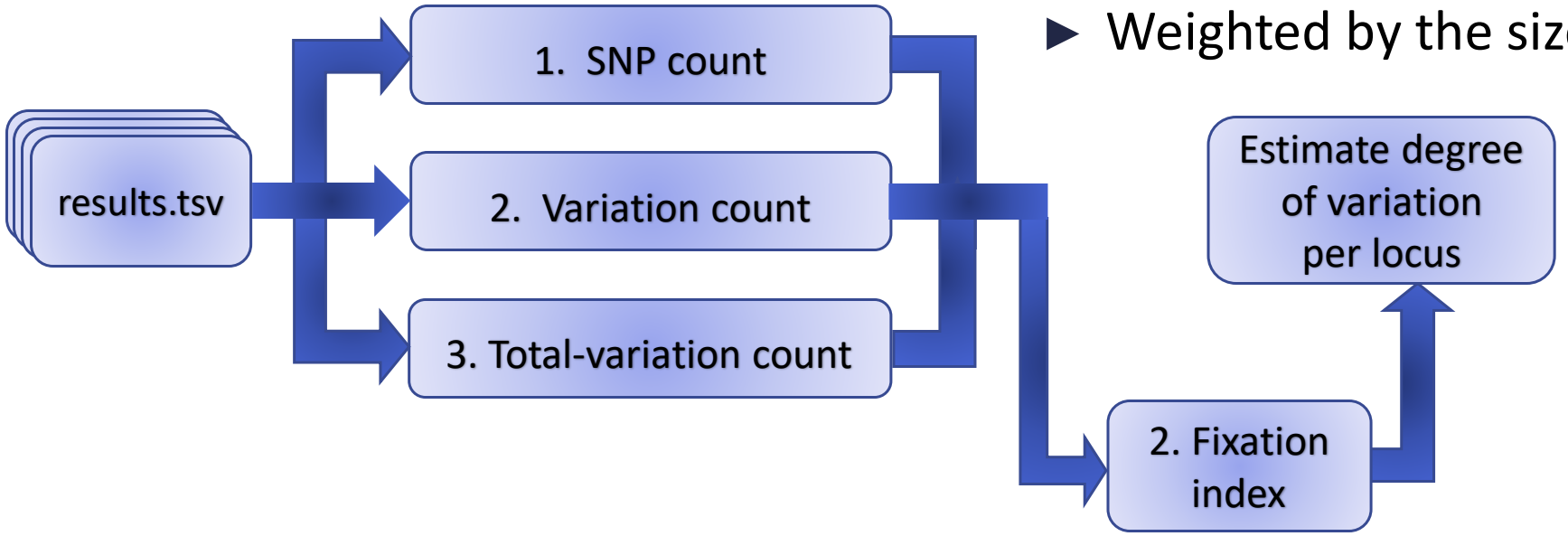
- Simply calculate the variance of the frequencies between two populations, at a specific locus.

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

$$\mu_{ij} = \frac{p_{i,k} + p_{j,k}}{2}$$

Fixation index (F_{ST})

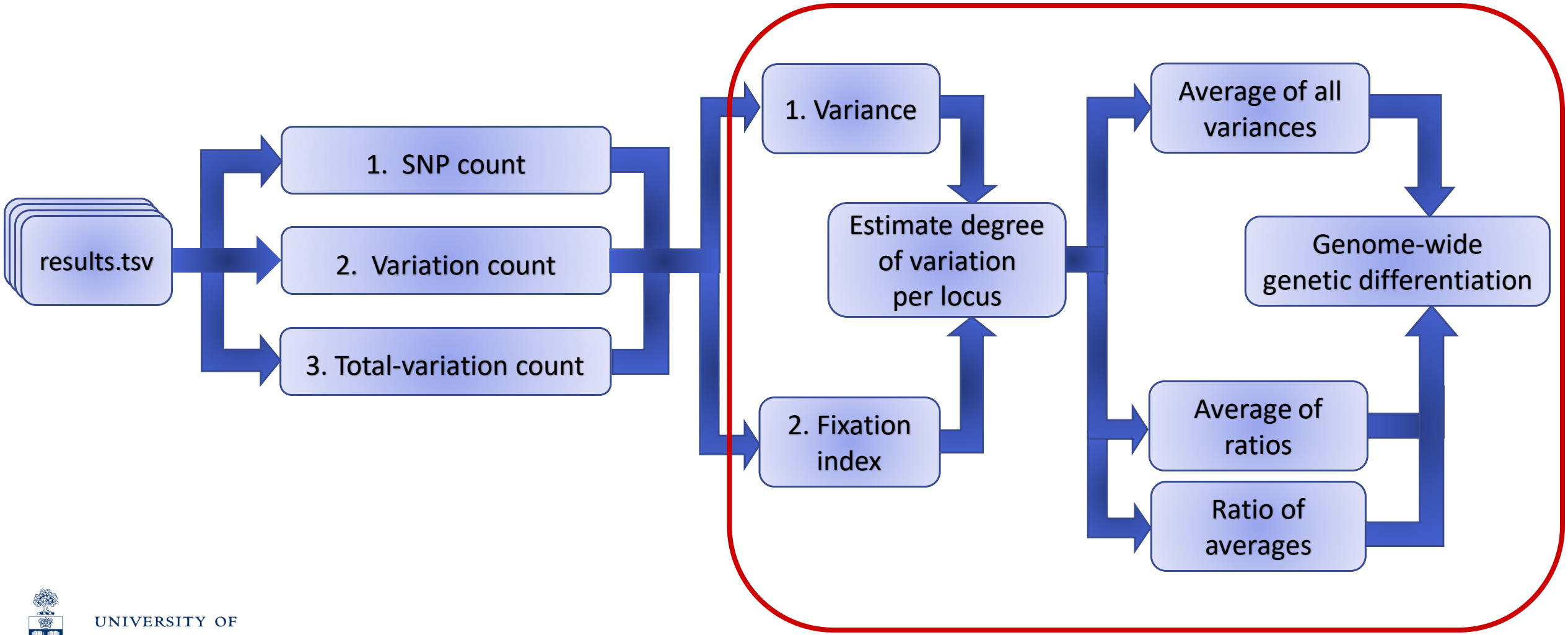
- A measure of population differentiation due to genetics.
- Weighted by the size of each population.
- Value between 0 and 1.



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

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

Genome-wide genetic differentiation

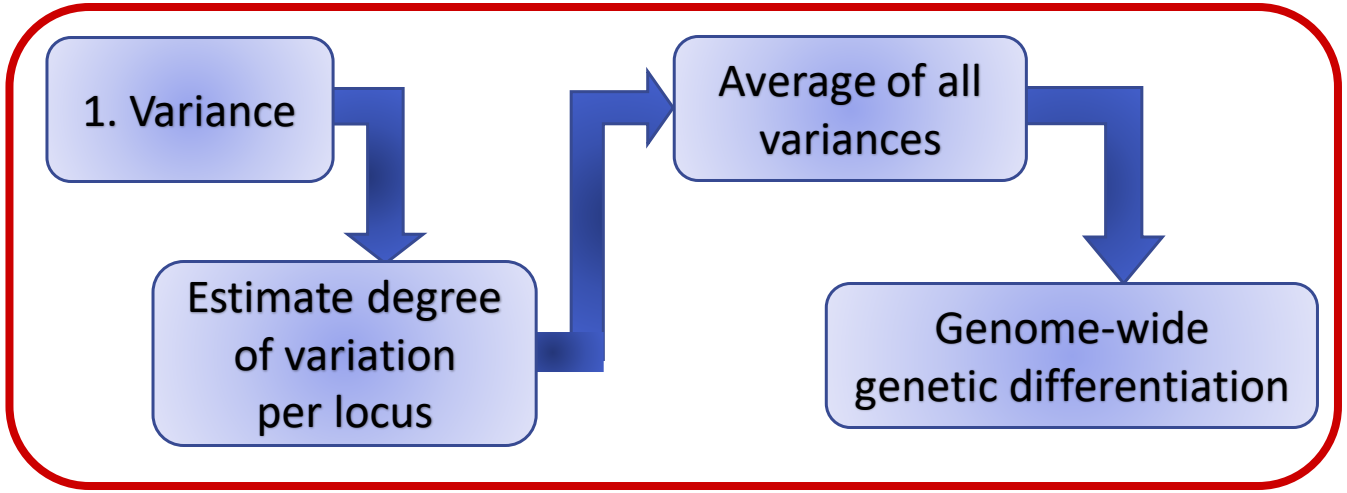


UNIVERSITY OF TORONTO

Variance

- Simply take the average of all variances across all loci.

$$\sigma_{ij}^2 = \frac{\sum_k \sigma_{k,ij}^2}{n_k}$$



Fixation index (F_{ST})

Ratio of averages

- Average variances and $\overline{p}$ values across all loci.
- Calculate a new F_{ST} value from the averages.

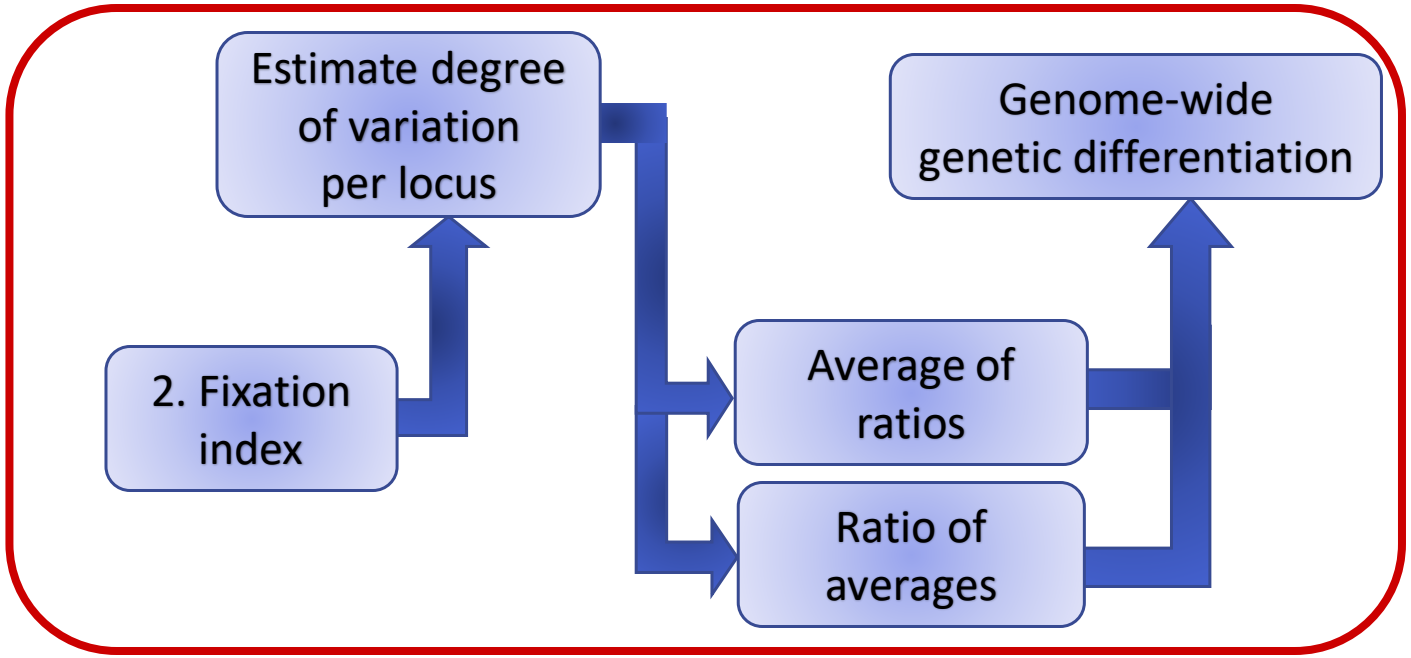
$$\sigma^2 = \frac{\sum_k \sigma_k^2}{n_k} \quad and \quad \overline{p} = \frac{\sum_k \overline{p}_k}{n_k}$$

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

Average of ratios

- Obtain genome-wide F_{ST} by taking the average across all loci.

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



Pairwise degree of genetic differentiation

- This process was performed between all possible pairs of populations to obtain a 26 × 26 matrix for each frequency measure (SNP, variation, and total-variation) and each statistical method (variance and F_{ST}).
- Plot square matrices as heat-maps using R.
- Compare heat-maps and select the superior method & measure.
- Perform clustering on the chosen matrix and plot the results as a dendrogram.
- Also, construct a 5 × 5 matrix based on demographic areas using the chosen method & measure.

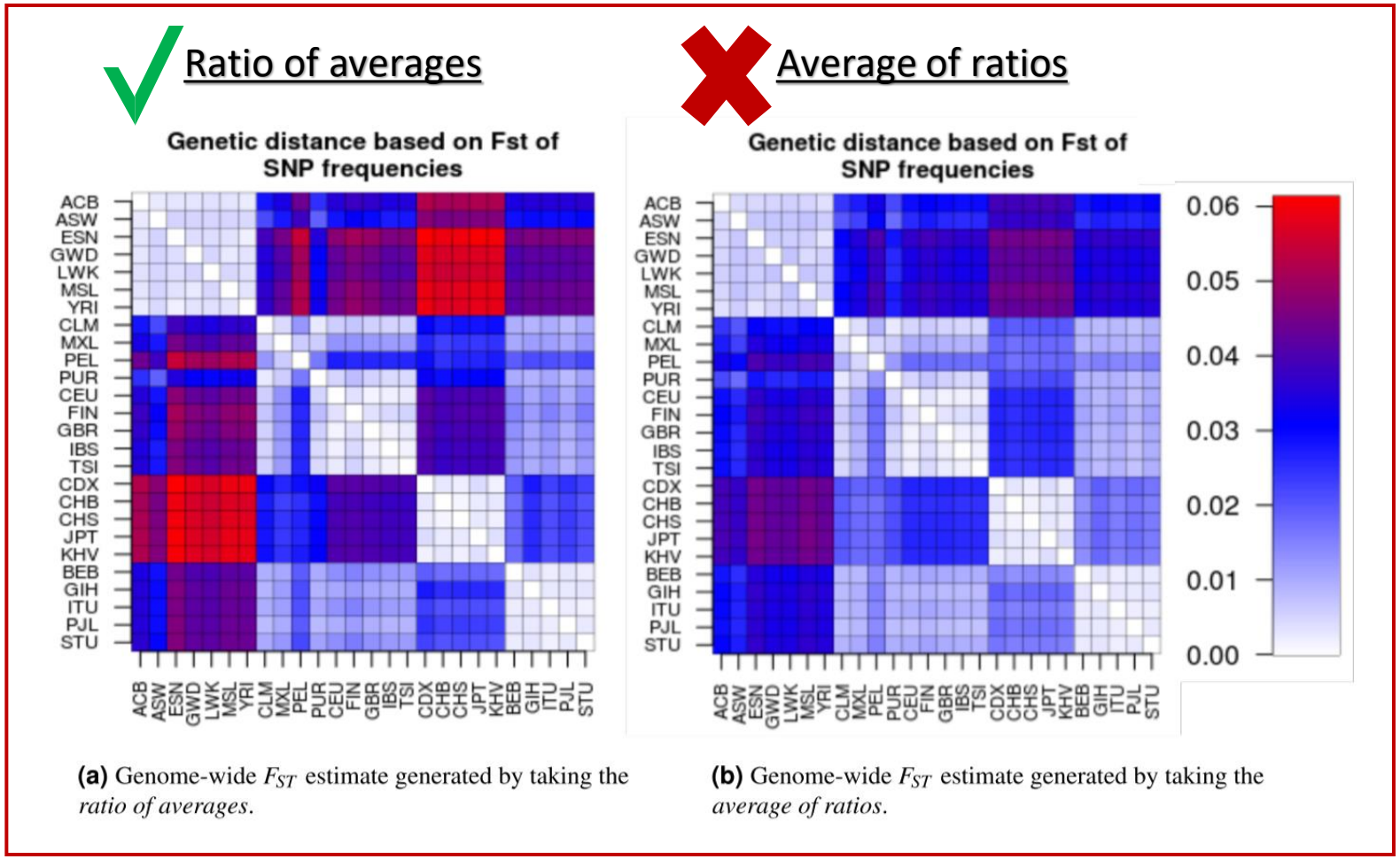
Pairwise Degree of Genetic Differentiation

Results

UNIVERSITY OF TORONTO

Genome-wide FST estimates

- Under-estimation of values when using "Average of ratios".
- Therefore, "Ratio of averages" was chosen as the method of choice for the rest of the study.



(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*.

UNIVERSITY OF TORONTO

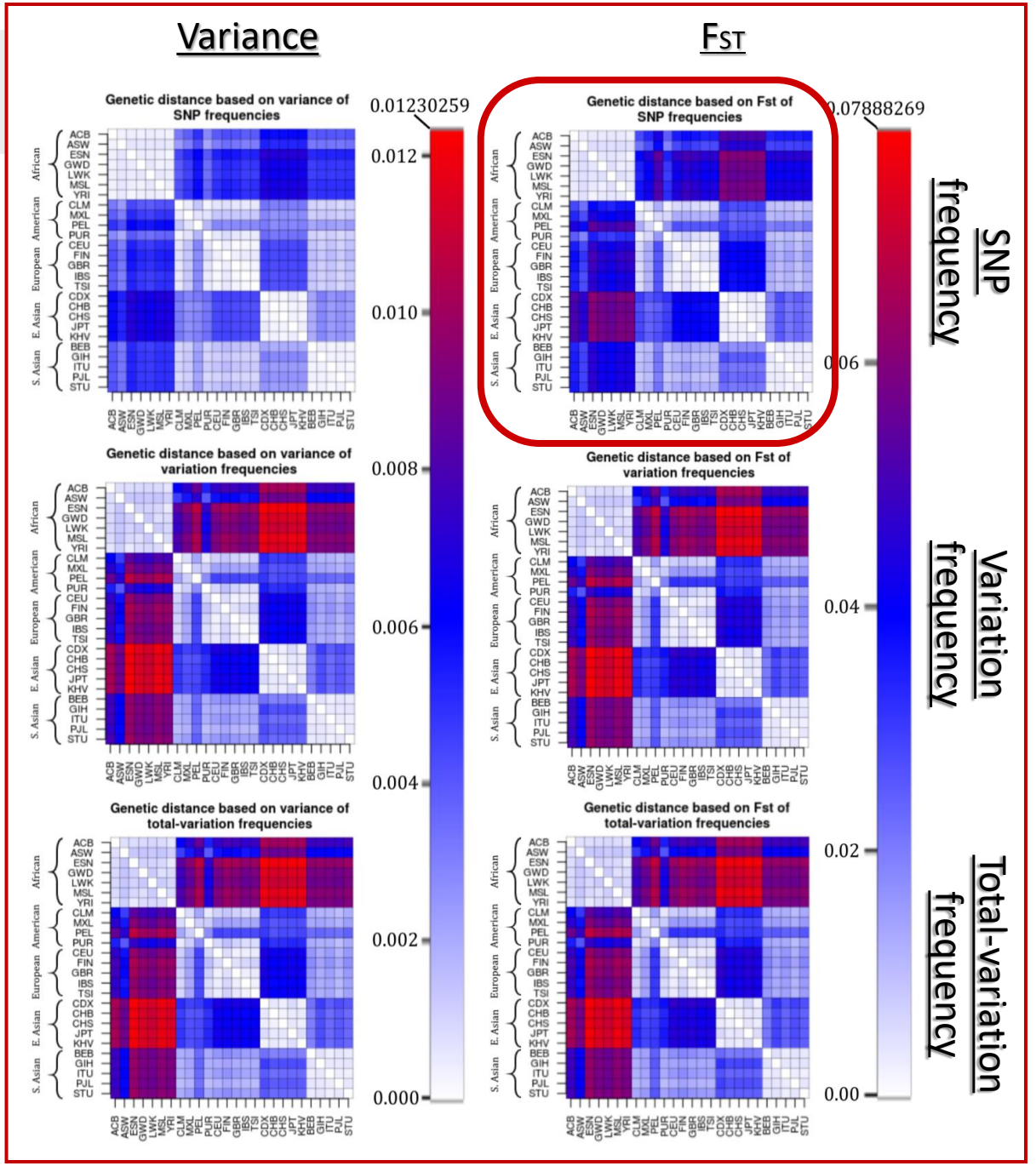
Pairwise genetic differentiation

Variance vs. F_{ST}:

- Variance and F_{ST} plots seem to agree/correlate with each other.
- However, F_{ST} takes into account the size of each population, and hence, it is a more accurate method
- Therefore, F_{ST} was chosen as the statistical method of choice for estimating the degree of genetic differentiation.

SNP vs. variation vs. total-variation frequencies:

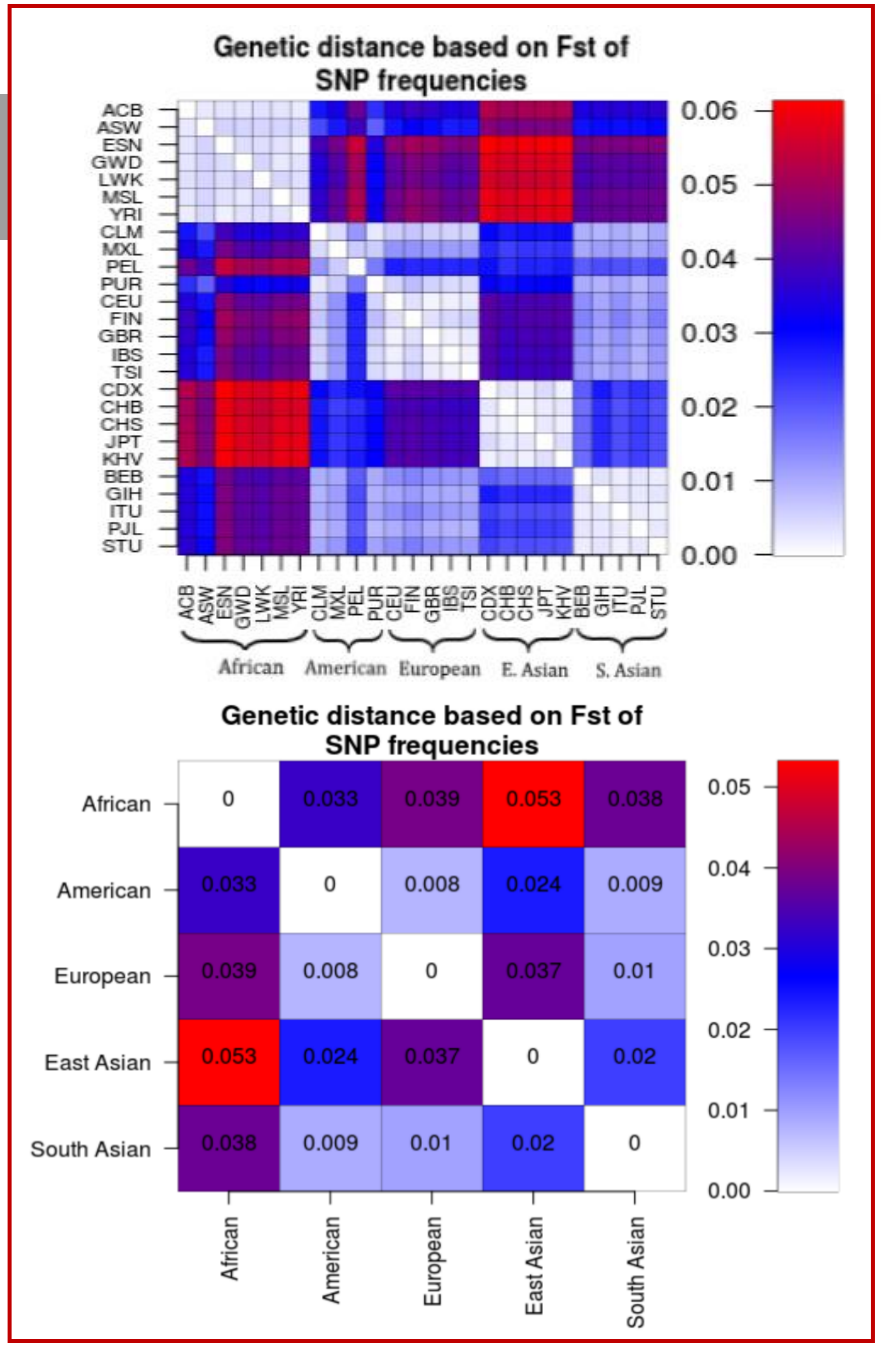
- Again, these three measures agree with each other.
- Variation and total-variation seem to emphasize the values.
- However, since using SNP frequencies is a more accurate measure and the ultimate goal of this study is to find characteristic SNPs, SNP frequencies was chosen as the measure of choice.



UNIVERSITY OF TORONTO

Heatmap of genetic differentiation values

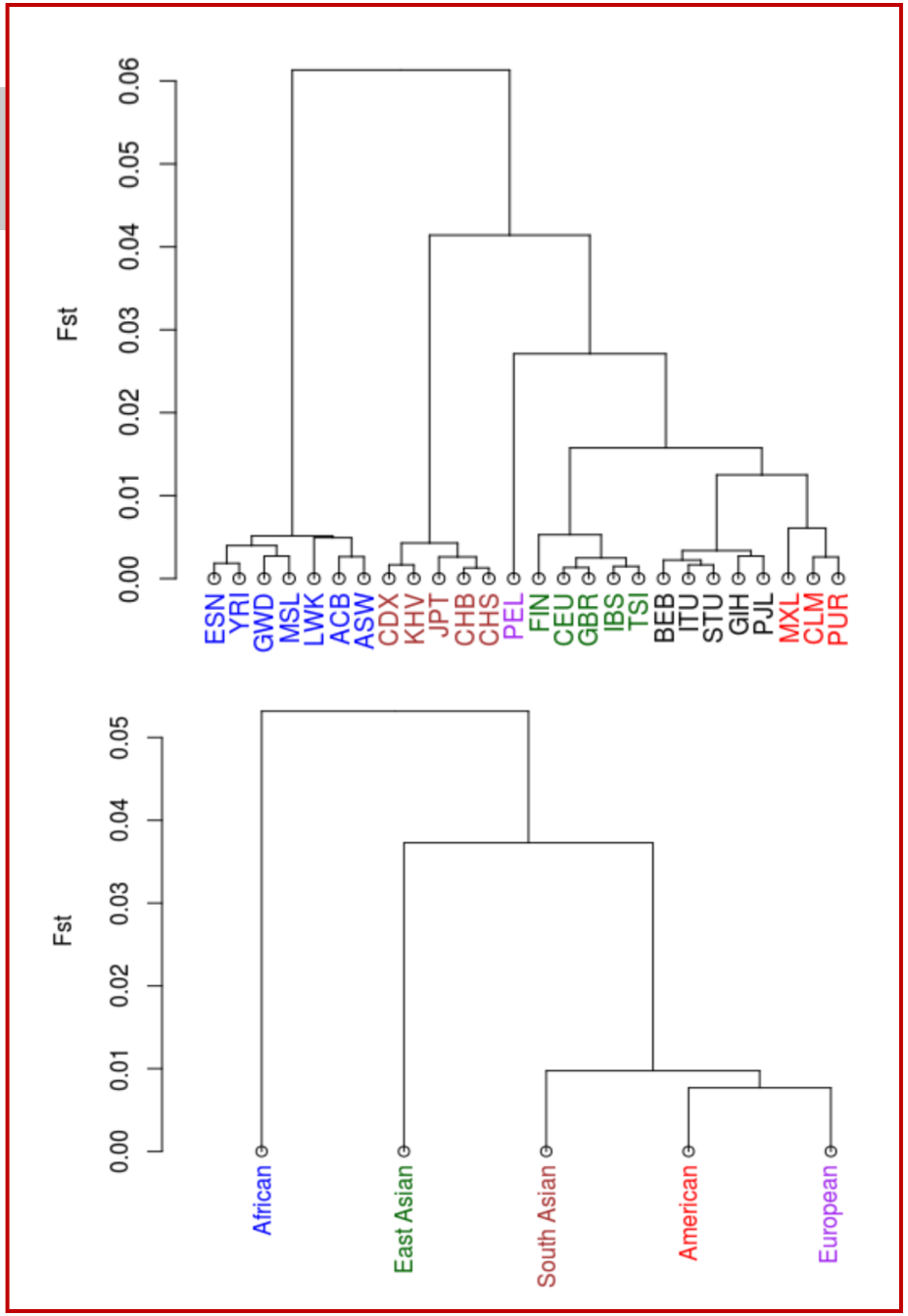
- The genetic differentiation value between populations in the same demographic group is very low.
- African populations have the highest genetic differentiation values.
- Americans, Europeans, and South Asians are close to each other.
- East Asians are considerably different than all other populations, but they are the closest to South Asians.
- The highest FST value (~0.061) occurred between the African population ESN and the East Asian population CDX.



UNIVERSITY OF TORONTO

Clustering & dendrogram

- The populations were then clustered based on the same matrices.
- A dendrogram of the results was constructed.
- All populations were clustered in their appropriate demographic group, as expected, with the exception of PEL.



UNIVERSITY OF TORONTO

Finding Characteristic SNPs

Materials & Methods

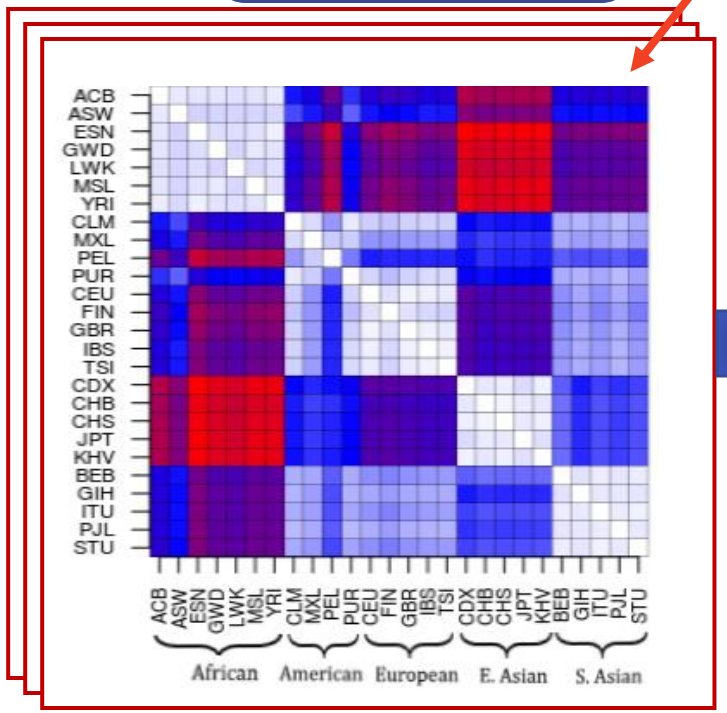
UNIVERSITY OF TORONTO

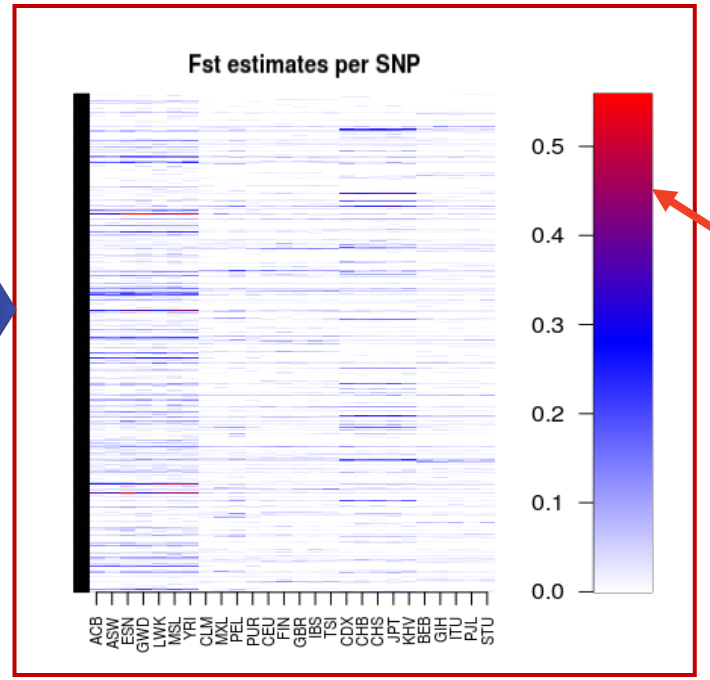
Finding characteristic SNPs

- **Definition:** SNPs that are unique to certain populations, and hence, may have emerged as a result of positive selection.

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

results.tsv → SNP frequencies → Estimate degree of variation in frequency per SNP → 1127 of 26 × 26 matrices → Median of all F_{ST} values of each population for a single SNP → Single F_{ST} value for each population at at each SNP → Repeat for all SNPs → 1127 × 26 matrix → Next Page

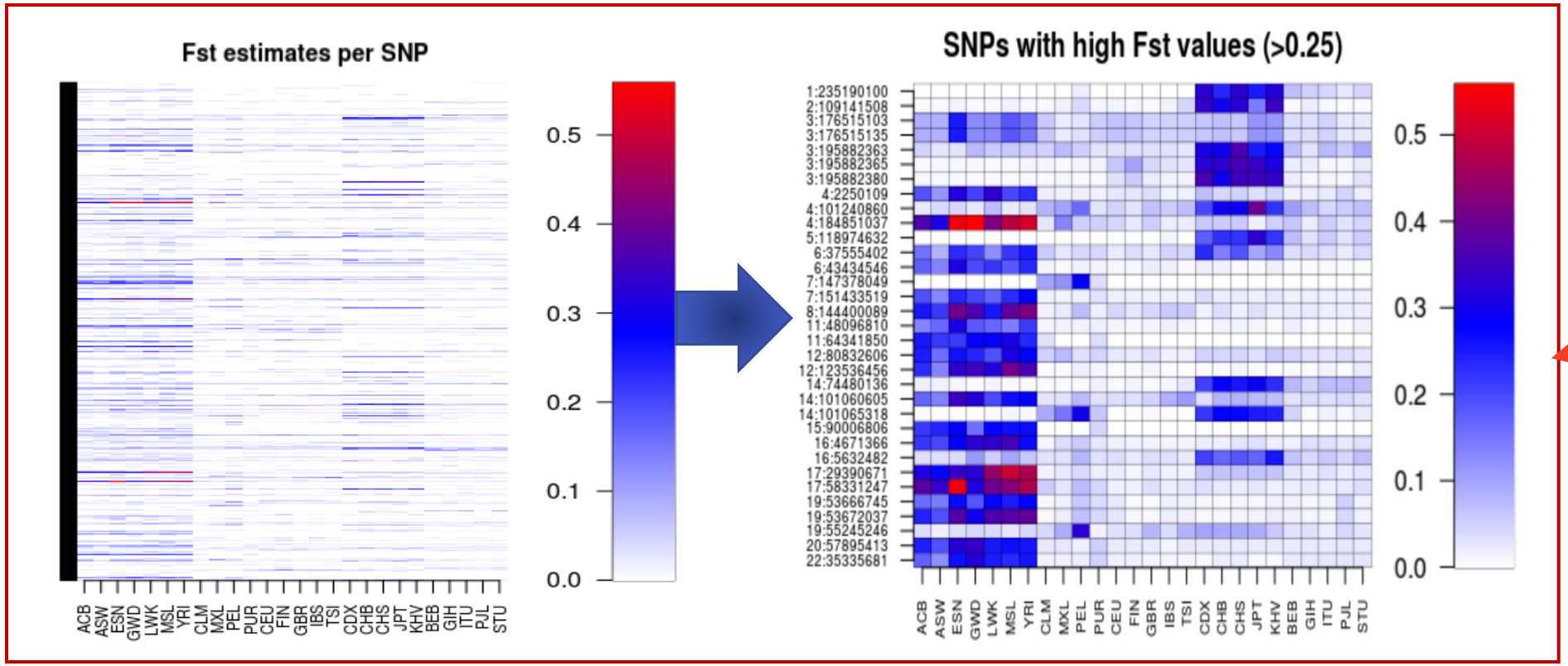




UNIVERSITY OF TORONTO

Finding characteristic SNPs

- To find characteristic SNPs, 0.25 was chosen as the cut-off between FST values that signify unique versus common SNP's.



1127 × 26 matrix

Filter rows with at least one FST value > 0.25

33 × 26 matrix

UNIVERSITY OF TORONTO

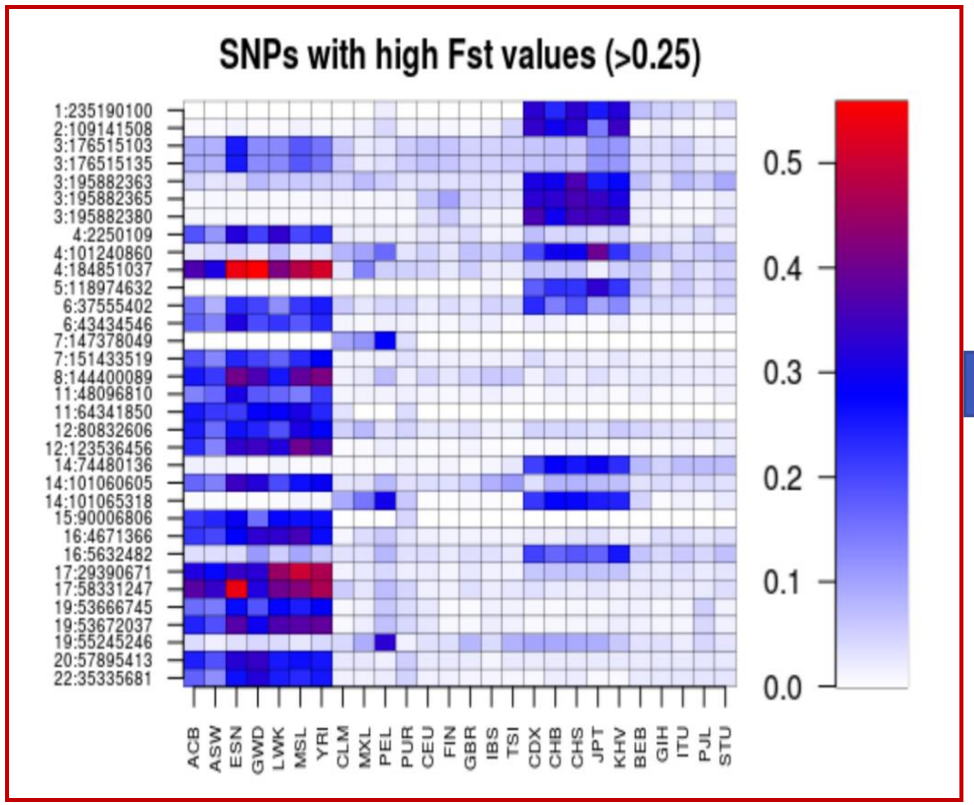
Characteristic SNPs

Results

UNIVERSITY OF TORONTO

Finding characteristic SNPs

- A total of 33 characteristic SNPs were found.



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

Genetic Variation in miRNA Primary Transcripts

Questions & Discussions

UNIVERSITY OF TORONTO

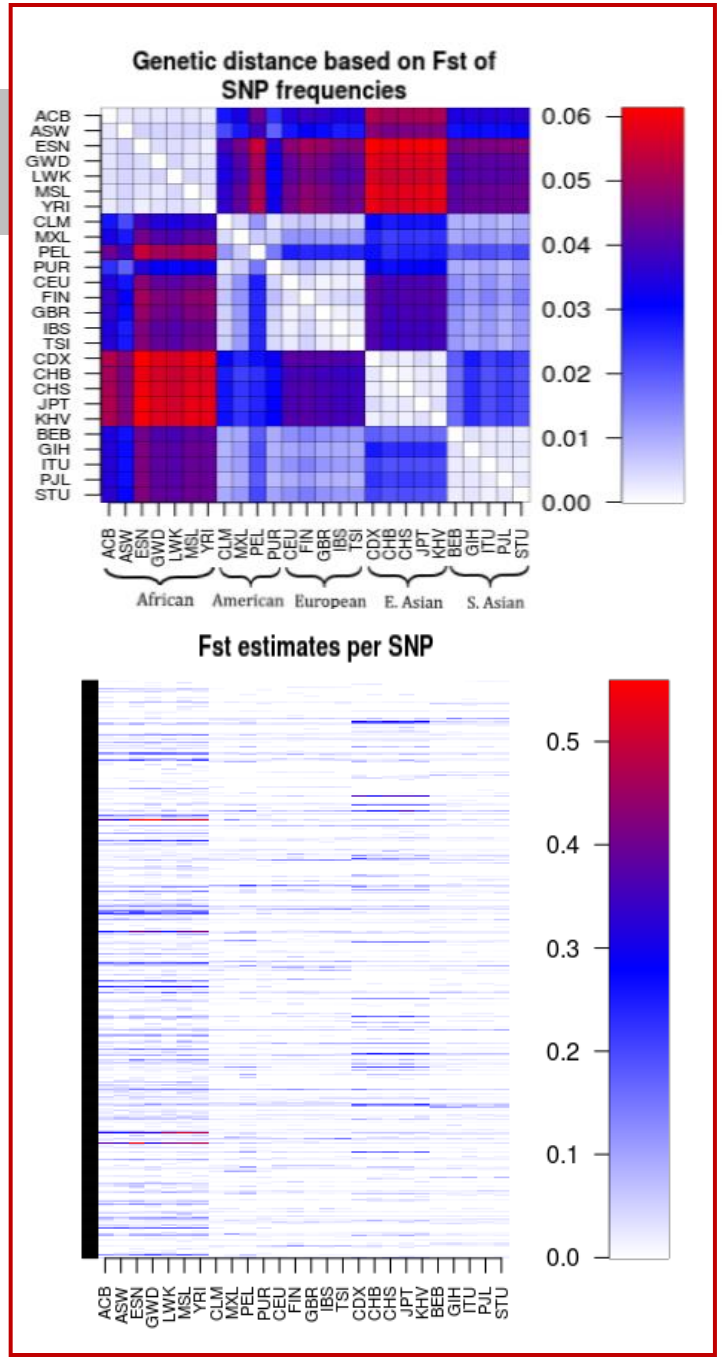
Questions & Discussions

Q) Why averaging over all SNPs rather than taking the median when calculating genome-wide genetic differentiation values?

- 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.

Q) Why taking the median of pairwise distances rather than averaging when calculating genetic differentiation per SNP?

- Median was chosen instead of the average since we want to find whether a population has a high or low genetic differentiation with *more than half* of the populations.



UNIVERSITY OF TORONTO

Questions & Discussions

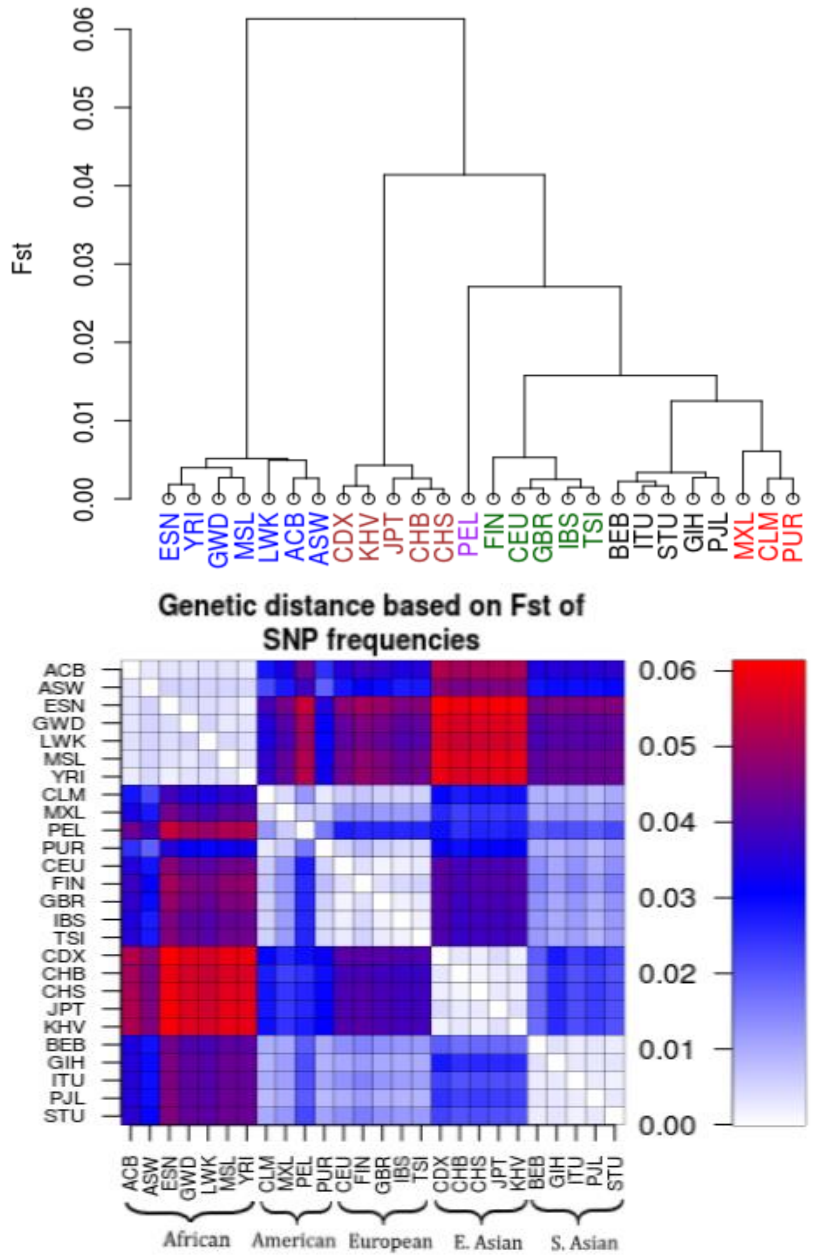
Clustering & Dendrogram:

- Surprisingly, population PEL was not clustered with any demographic group.

Q) Why did PEL not cluster with other American populations?

- Looking at the heat-map, PEL is fairly different than other American populations.
  - For example, American populations are listed to be close with Europeans and South Asians, with the exception of PEL.

- This result could imply that population PEL has been relatively isolated compared to other American populations.



UNIVERSITY OF TORONTO

Questions & Discussions

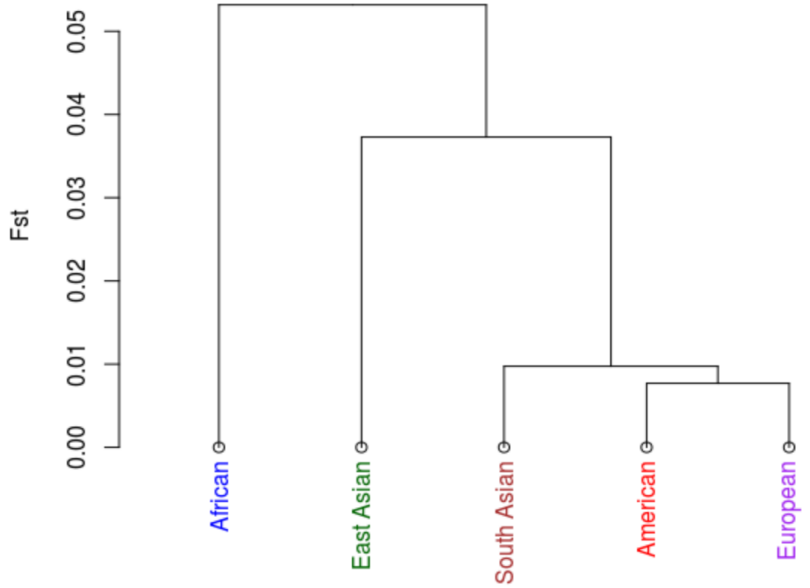
The dendrogram and heat-map of demographic groups agree with each other.

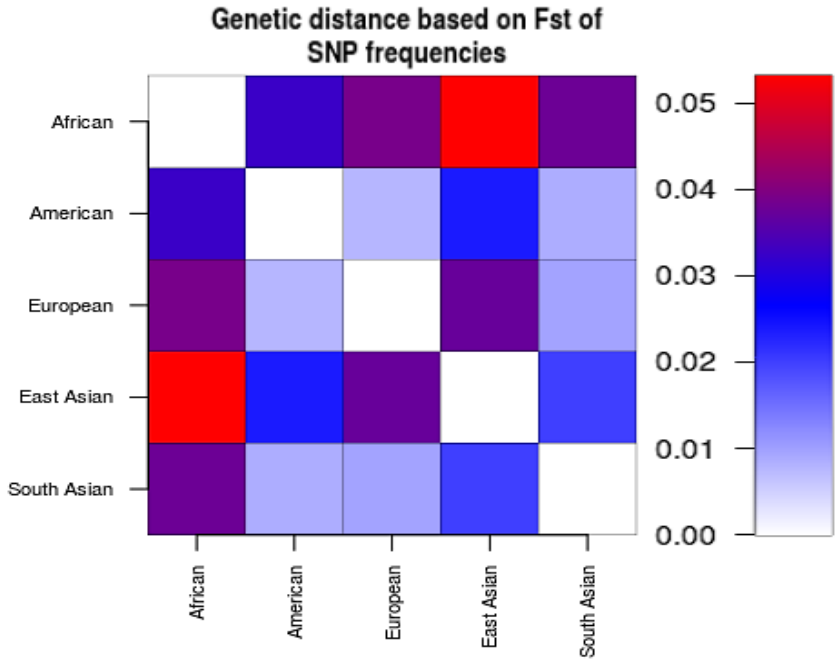
Heat-map

- Africans have the highest genetic differentiation values.
- East Asians are considerably different than others as well.
- American, European, and South Asians are fairly close to each other.

Dendrogram

- Africans were diverged first.
- East Asians were diverged next.
- Lastly, South Asians diverged from Americans and Europeans.





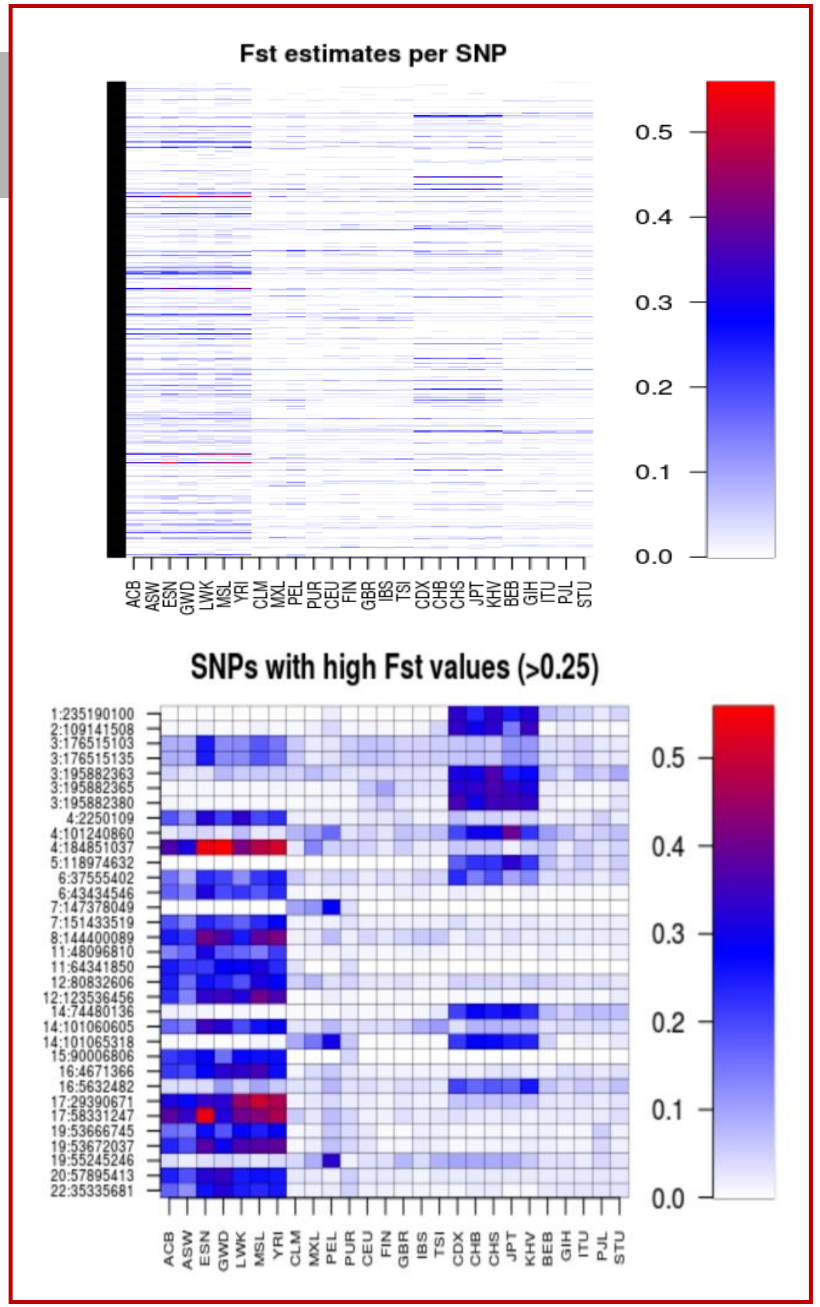
UNIVERSITY OF TORONTO

Questions & Discussions

Q) When finding characteristic SNPs, why did we choose a value of 0.25 as the cut-off?

- Consensus on interpretation of F_{ST}:
  - $F_{ST} < 0.05$: Little genetic variation,
  - $0.05 < F_{ST} < 0.15$: Moderate genetic variation,
  - $0.15 < F_{ST} < 0.25$: Great genetic variation,
  - $F_{ST} > 0.25$: Very high genetic variation.

- The value 0.25 was chosen since an F_{ST} value greater than 0.25 is considered to represent a very high degree of genetic variation.



UNIVERSITY OF TORONTO

THE END!

Thank you

UNIVERSITY OF TORONTO