Log book for CYSF:

Title: A Phylogenetic Study of Contemporary Mainland Southeast Asian Populations Using Single Nucleotide Polymorphisms (SNP) - by Aaron George
Field of Study: Genetics, Biological Anthropology, Life Sciences, Bioinformatics

November:

11/27/25:

- Interested in Southeast Asian population genetics, and originally thought of using mtDna data (mitochondrial DNA) in order to track maternal DNA, hence gaining an idea of human migration into southeast asia
- Decided CYSF to present potential findings
- Researched topics and methods in which Southeast Asian population genetics may be depicted, going through scientific research papers from universities such as Khon Kaen university, from scientists who specialize in the region, such as Changmai et.al.
- Delved into genome libraries such as GenomeAsia 100k and the Singapore Genome Variation Project - a lot of these have MSEAN populations missing

December:

12/4/25:

- My project has not been entered into the CYSF platform yet, nor has my school started providing any Science Fair information but started deciding how to conduct the project

- Considered Y-chromosomal DNA to track paternal DNA
- Wanted to incorporate the idea of haplogroups as genetic markers for common ancestors
- MtDNA was too broad to start a project, and resources were vague, and after some research from studies like: https://pmc.ncbi.nlm.nih.gov/articles/PMC3525085/, SNP data (single nucleotide polymorphisms) seems to offer many more sources that I may use, providing a unique insight into population genetics in Southeast Asia
- Many genomic libraries offer SNPs and their components

12/5/25:

- Researched more information about SNPs, and their role within a broader genetic context: differences between SNPs and alleles, how SNPs distinguish alleles from other alleles,
- Clarified understanding of alleles, genes, and polymorphisms
- Tried thinking about ways to represent data through ways such as bioinformatics
- Used Google Colab and experimented with ways to represent data into a CSV file for later use.

January:

1/12/26:

- Entered the CYSF online project area with the help of my science fair coordinator
- Read through previous projects in order to gain an idea of how I should structure my online portion of the project
- Mainly spent the day exploring the platform

1/15/26:

- Researched genome libraries that I may find Asia-Pacific populations to sample, and found Ensembl (https://useast.ensembl.org/Homo_sapiens/Info/Index), which uses populations from the 1000 Genomes Project.
- Used Chinese Dai from Xishuangbanna, Kinh from Ho Chi Minh City Vietnam, Bengali from Bangladesh, Sri Lankan Tamil, Southern Han Chinese, and Northern Han Chinese from Beijing as the initial populations.
- Lack of other MSEAN populations, so I will need to find other external sources to gain them from.

1/17/26:

- Researched technology and applications in which I could construct a phylogenetic tree
- Originally tried utilizing POPGENE made by Francis Yeh from the University of Alberta: played around with the website but it was a bit hard to use
- Next tried POPMLVIS https://popmlvis.qcri.org/ in order to construct dendrograms and other ways to map population genetics, as it appealed for its more modern interface and seemingly easier navigation, but it turned out to be even harder than the older one, and no data would load
- Finally decided to use POPTREE2, developed by Naoko Takezaki et al. at Kagawa University https://www.med.kagawa-u.ac.jp/~genomelb/takezaki/poptreew/ , played around with it and initially had trouble understanding how to construct the populations, but it turned out to be exponentially easier - I will be using this application in the future for constructing dendrograms with SNP data

1/18/26:

- Read through manuals of POPTREE, how to use, how to format SNP data and alleles so the program can accurately assess and draw conclusions to format the phylogenetic tree, considering branch length, nodes, clades, etc
- What types of SNPs should I use, and which ones are redundant for the construction of a visual representation?

1/26/26:

- Added the Yoruba from Ibadan, Nigeria population after more research
- In the context of a phylogenetic tree, adding an outlier population is necessary in order to depict how East-Eurasian populations diverge from other non-Asian populations
- Added the respective allele frequencies for SNPs in Yoruba individuals into my table

1/31/26:

- Researched various ways in order to represent proxy data of SNP polymorphisms, and discovered PCA scatterplots. PCA (Principal Component Analysis) plots can visualize complex genetic data into a graph in order to cluster certain ethnic groups based on genetic similarities. PCAs traditionally use individual genotypes, resulting in the large amount of points on the plot. For the sake of this project, and considering the fact that I am using allele frequency data, each point will represent one entire population, hence leading to only a few points
- Used this website (https://builtin.com/machine-learning/pca-in-python) in order to gain an idea of how to create a PCA table/plot using Python along with my own understanding of Python.
- Continued conducting manual analysis of SNP data from the https://useast.ensembl.org/Homo_sapiens/Info/Index website, finding samples of SNP data from various studies and the SNPedia: Wangkumhang, P., Shaw, P. J., Chaichoompu, K., Ngamphiw, C., Assawamakin, A., Nuinoon, M., ... & Tongsima, S. (2013). Insight into the peopling of Mainland Southeast Asia from Thai population genetic structure. *PLoS One*, *8*(11), e79522., ,
- Finished problem statement on CYSF platform

February:

2/7/26:

- Continued taking SNP data for allele frequencies from Ensembl, and continued putting the data into a separate document so I can later input all of my SNP data into POPTREEW for phylogenetic tree construction.
- Mainly repetitive work, so very little else to say

2/8/26:

- Added Japanese from Tokyo, Japan as a distinctly East Asian group, in order to compare certain Sinicized Southeast Asian groups such as the Kinh with East Asian populations in order to evaluate where they will lie
- JPT is assumed to be the East Asian population least related to Southeast Asian populations, due to Jomon admixture with Yayoi populations separate from Southeast Asian lineages

2/13/26:

- Same as 2/7/26
- More research into the late Pleistocene period of Southeast Asia as well the Neolithic periods, tracking human migration into the region

- Research on Indian and Chinese migration into Southeast Asia, and the effects on the genetics of modern/contemporary southeast asian populations and ethnic groups such as the Dai and Kinh
- Completed my research section, discussing the definition of SNPs as nucleotide substitutions for certain alleles and Southeast Asian migration history, with primary focus on Mainland Southeast Asia

2/14/26:

- Clarified parts of the research section again, ensuring coherence, along with including more data on ethnolinguistic groups in MSEA

2/15/26:

- More SNP entries for the phylogenetic tree, collected data for SNP allele frequencies in a separate document for organization
- Used the Fondation Jean Dausett to find a Cambodian population, which was originally missing from the Southeast Asian population section.
- Cambodian population is essential, considering that they were one of the first contemporary Austroasiatic groups in MSEA
- Same as 2/7/26
- Very repetitive tasks such as retrieving rsIDs (names of SNPs) and loci (location of the allele on a chromosome) from genomic libraries like Ensembl

2/18/26:

- More research into SNPs again, including ancestry informative markers and how to accurately assess which SNP frequencies to use
- Genetic changes accumulate due to populations reproducing in certain areas, leading to a higher concentration of particular alleles within that region
-

2/20/26:

- Began inputting SNP for analysis into python (PCA plot), filling in the placeholders in the dataframe meant for them. Reviewed the code for generating the PCA plot to ensure its efficacy
- Entered 23 rsIDs (SNP names essentially) and their allele frequencies
- Generated the scatter plot and verified its accuracy with other credible sources
- Continued working on online project, such as the method

2/28/26:

- Fully finished the “creation” components of the project (phylogenetic tree and scatterplot)
- Dedicated to finishing the method and data sections of the online project
- Finished method, finished half of the data section

- Inputted the results of the scatterplot and tree into the data

March:

3/1/26:

- Aiming to finish full online project for submission
- Finished the data section, along with the inferences and conclusions that can be made from the results, such as the Cambodian population serving as a genetic bridge between southeast/east and South Asia, along with the chronological order or divergence from a common ancestor for each population
- Made final conclusions, and discussed the limitations of my project, such as the lack of inclusion for underrepresented populations in Southeast Asia, such as hill tribes and other indigenous groups that portray southeast asian genetic diversity in high resolution
- Did the declarations, and aiming to email school coordinator about submission of project

3/2/26:

- Nearing the firm due date for online submissions - quickly trying to polish online portion
- Started citing all sources used
- Will continue working on logbook, but due to submission deadline of the online project, along with submitting the logbook into the online portion, the logbook will end here when viewed from the online page
- This will continue after, when creating the physical poster this month.