# Construction of the Settlement Path to Hawaii Using Genomic Statistics

*Feiyang Liu[1], Javier Blanco-Portillo[2], Mark Penjueli[3], Alexander Ioannidis[1,4]*

*[1]Institute for Computational and Mathematical Engineering (ICME), Stanford University*

*[2]Department of Biology, Stanford University*

*[3]Department of Biology, New York University Abu Dhabi*

*[4]Department of Biomedical Data Science, Stanford University*

## Abstract

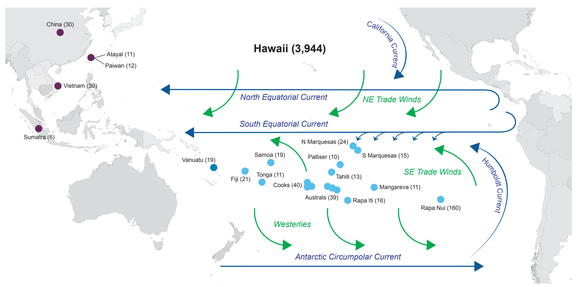
We introduce some genetic statistics, mainly the range expansion statistic $\psi$, efficiently computed from the allele frequency files of populations, to explore some perspectives of the historical founder event between populations, such as the geographical origin of a range expansion, and leverage these genetic statistics to construct a potential path of migration for Polynesian settlement of Hawaii.

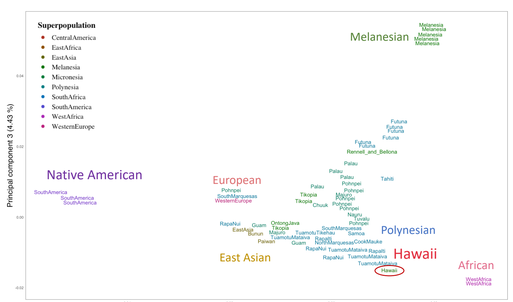
| | ALT_FRQ |
|---|---|
| A (G) | 0.13 |
| G (A) | 0.12 |
| T (C) | 0.02 |
| A (G) | 0.47 |
| C (T) | 0.45 |
| T (C) | 0.49 |
| T (C) | 0.40 |
| A (G) | 0.80 |
| G (A) | 0.84 |
| C (T) | 0.49 |
| A (G) | 0.58 |

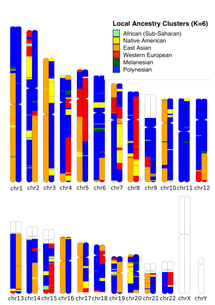
Aggregate Samples

## Dataset

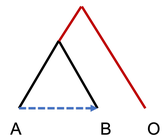
The dataset contains 3,944 modern Hawaiians as well as individuals from other Polynesian islands, and the Hawaiian genotype matrices have been masked to screen out SNPs that appear to have an ancestry that is not Polynesian (e.g., European, African) using the local ancestry classifier Gnomix.




## Genetic Statistics




- $F_2$ Genetic Drift Statistic

$$F_2(A,B) = (\hat{p}_A - \hat{p}_B)^2 - \frac{\hat{p}_A(1-\hat{p}_A)}{n_A - 1} - \frac{\hat{p}_B(1-\hat{p}_B)}{n_B - 1}$$

- $F_3$ Shared Drift Statistic

$$F_3(A,B;O) = (\hat{p}_A - \hat{p}_O)(\hat{p}_B - \hat{p}_O) - \frac{\hat{p}_O(1-\hat{p}_O)}{n_O - 1}$$

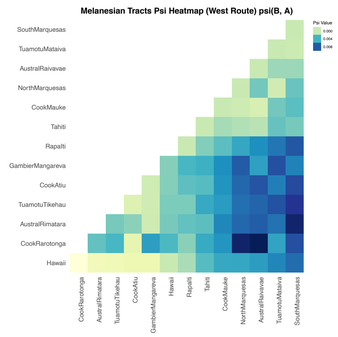
- Average Number of Pairwise Differences (Nucleotide Diversity)

$$\pi(A,B) = \hat{p}_A(1-\hat{p}_B) + \hat{p}_B(1-\hat{p}_A)$$

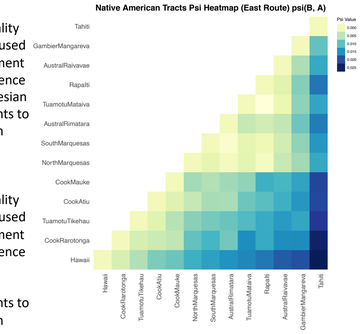
- Range Expansion Statistic (Directionality Index)

$$\psi(A,B;O) = \hat{p}_A^{(O)} - \hat{p}_B^{(O)}$$

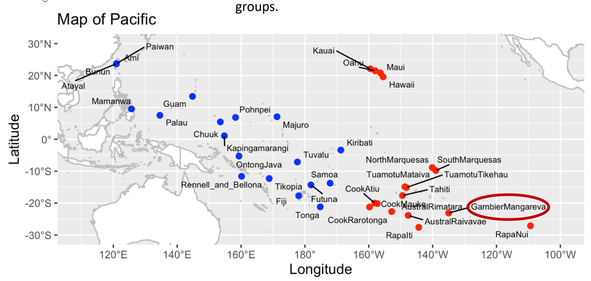
## Results




**(Left)** Directionality index ($\psi$) used for settlement path inference on Melanesian components to Polynesian groups.

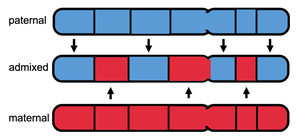


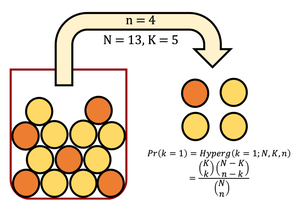

**(Right)** Directionality index ($\psi$) used for settlement path inference on Native American components to Polynesian groups.




## Methods

- For all genetic computations, the correlation of nearby SNPs (single nucleotide polymorphisms) is induced by chromosomal crossover. We use a block bootstrap to estimate standard error by dividing the data into consecutive blocks and resampling the entire blocks in the bootstrap.
- ($F_3, \psi$ only) We need to select one/more outgroup populations for computation.
- ($\psi$ only) When two populations have unequal sample sizes at locus $j$, we usually downsample the large sample to the smaller sample size, which follows a hypergeometric distribution. In reality, we pick a global downsampling size such that both populations can downsample to a fixed size at all loci.
- ($\psi$ only) We consider only rare alleles still found in both populations after downsampling, applying the filtering scheme shown on the right.




$$Pr(k=1) = Hyperg(k=1;N,K,n) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

| $k_A$ \ $k_B$ | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| 0 | 0 | -1 | -2 | -3 | -4 |
| 1 | 1 | 0 | -1 | -2 | -3 |
| 2 | 2 | 1 | 0 | -1 | -2 |
| 3 | 3 | 2 | 1 | 0 | -1 |
| 4 | 4 | 3 | 2 | 1 | 0 |

## Future Work

- We will test with different hyperparameters on outgroup populations, derived allele frequency, and downsampling size to see whether the directionality output/signs ($\psi$) is robust to changes.
- We will convert the code from R to Python via ARCH for block bootstrap and Ray for parallelism.

## Reference

- Benjamin M Peter. “Admixture, Population Structure, and F-Statistics”. In: *Genetics* 202.4 (Apr.2016), pp. 1485–1501. doi: 10.1534/genetics.115.183913. url: https://pubmed.ncbi.nlm.nih.gov/26857625.
- Benjamin M Peter and Montgomery Slatkin. “Detecting range expansions from genetic data”. In: *Evolution; international journal of organic evolution* 67.11 (Nov. 2013), pp. 3274–3289. doi: 10.1111/evo.12202. url: https://pubmed.ncbi.nlm.nih.gov/24152007.