**Things to look at**

* What we are looking for is how demographic histories of many different populations relate to each other
  + What does the African Demographic history tell us about the European demographic history
* TMRCA determined from heterozygosity of recent samples tells us information about the demographic histories (Allows us to infer effective population sizes) of populations
  + Could this be used to validate or infer events like migration?
* Figure out how to perform PCA on the data that Erik gives me
* Figure out how a VAE would work on Ne graphs (if they are even appropriate)
* [Mediods](https://en.wikipedia.org/wiki/Medoid#:~:text=Medoids%20are%20representative%20objects%20of,members%20of%20the%20data%20set.)?
* Weather forecasting time series PCA/dimensionality reduction thing that Erik was talking about
* Web interface for clustering on the server

**Resolved kinda**

* How to determine similarity of squiggles?
* How to determine differences between squiggles if similarly shaped squiggles can come from completely different organisms
  + Sounds like covariance might be useful somehow?
  + Consider variational autoencoders or PCA for reducing the dimensionality of the PSMC curves

**Human Demographic History PSMC Model Paper**

Passages

1. Stating model and the data it is applied on (various human genomes)
2. Distr of time since TMRCA between 2 alleles -> Info on pop. size over time

The sequenced samples were binned into 100bp regions where each bin was given the states

1. 0 - Homozygous
2. 1 - Heterozygous
3. . - Missing

Where it was then used as input for the PSMC model to infer TMRCA

Vocab

* Coalescent times - time to the last coalescent event (TMRCA)
* Markovian - next state depends on current state

**Intro Paper**

* The more recent the coalescence event (observed heterogeneity in a population), the more likely a population size was small recently, vice versa (more homogeneity indicates a bottleneck may have happened very long ago or not recently)
* Coalescent Theory -> relates Coalescence rate to Ne
  + Time periods with higher Coalescence rate have lower Ne
* **SMC models**
  + Split loci in a genome apart to trace out evolutionary history
  + Track coalescence events between the two alleles (diploid) for each locus
    - From this, infer the number of coalescence events that occurred across the genome with a given time interval
      * Use inference on number of coalescence events to reconstruct the Ne’s in the time interval
      * Inference also includes timing of expansions and bottlenecks
  + **Limitations**
    - Inference on Ne changes over time points may be thrown off by our population not meeting assumptions needed to convert coalescence rates to effective population sizes
      * Confounders can be natural selection and nonrandom mating
    - Confusion arises when differentiating changes caused by shifts in population size and those caused by other demographic params (migration rate, inbreeding, physical barriers between population groups)
  + **Use for Studying Changes in Pop. Structure**
    - The multispecies coalescent can be used to infer timing of divergence between closely related populations or species
      * Unlike SMCs which can incorporate info from whole genomes, this method is not feasible for large numbers of loci
    - Similar lines on SMC plots may indicate shared population history
    - PSMC can only provide a maximum bound on divergence time since inferred population size tends towards infinity as the coalescence rate of reproductively isolated populations becomes 0
      * Due to this, hPSMCs are more often used for dating cessation of gene flow rather than population divergence
      * MSMCs tend to work better than PSMCs for these scenarios
  + **Use for estimating mutation rate**
    - Can work but only under strict assumptions, Max Likelihood is said to be more accurate
  + Notes
    - Time interval size seems to be set at 64 by default (no greater reason)

**Polar Bear Paper**

PCA on time series data (SSA?)

How to differentiate time series curves

Variational Autoencoder on seq data

Variational Autoencoder based Anomaly Detection using Reconstruction Probability