Compressing HapMap Data using Sufficient Subpopulation Markov Chain Statistics

Kevin Cummiskey, Caleb Lareau, Matthew Ploenzke

May 4, 2016

Overview

1. Statistical Motivation
   - Why sufficient statistics?
2. Biological Justification
   - Why Markov Chains?
3. Compression algorithm implementation
   - Computational efficiency benefits in speed and storage
4. Testing our compression structure
   - Recapturing Subpopulations via SMCT Algorithm

Statistical approach

- A statistic is **sufficient** for an unknown parameter if "no other statistic that can be calculated from the same sample provides any additional information as to the value of the parameter."

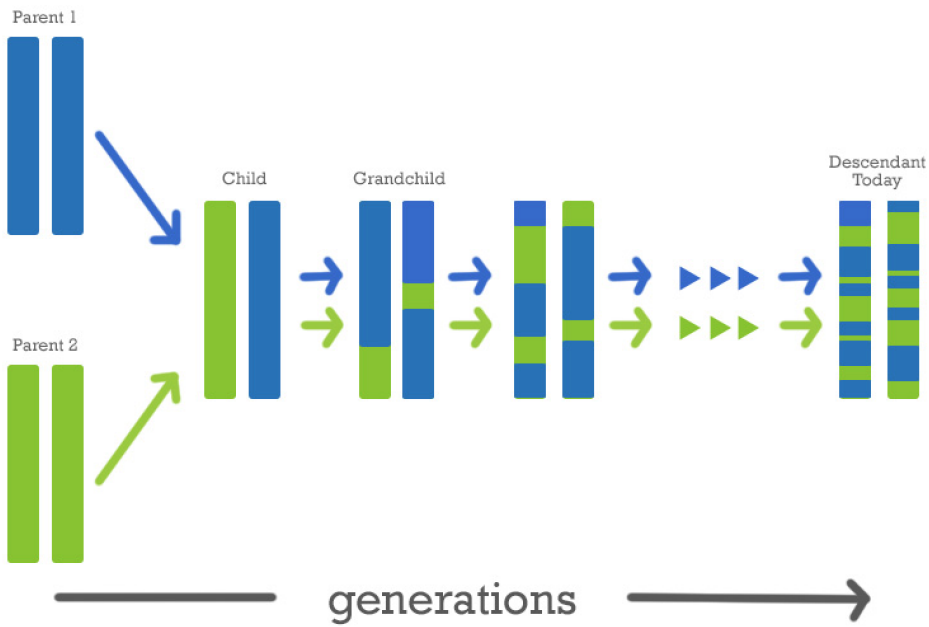
- Practically, we can "throw away" the data, and using the sufficient statistics, we could regenerate the same distribution of the data (not the data itself though).

Can we determine sufficient statistics that allows us to summarize the HapMap data?

Source: Casella, Berger

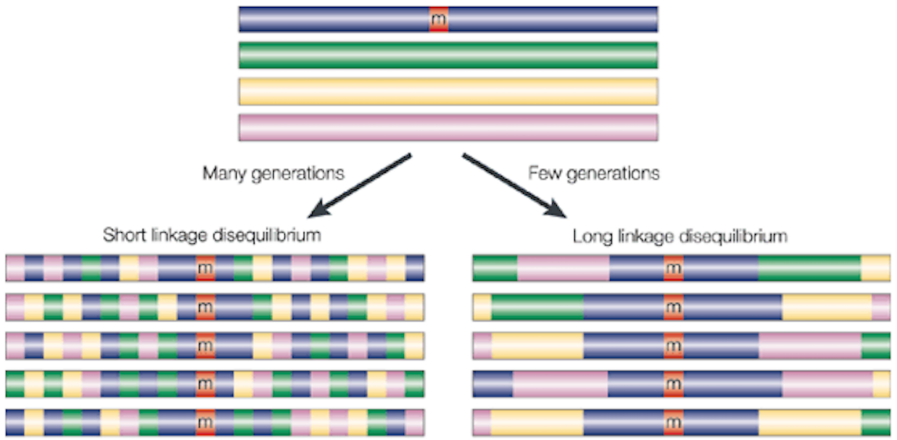
Biological justification

- Genetic recombination is defined as the production of offspring with combinations of traits that differ from those found in either parent.
- Recombination regions are not random, and there have been observed recombination "hotspots"



Linkage Disequilibrium

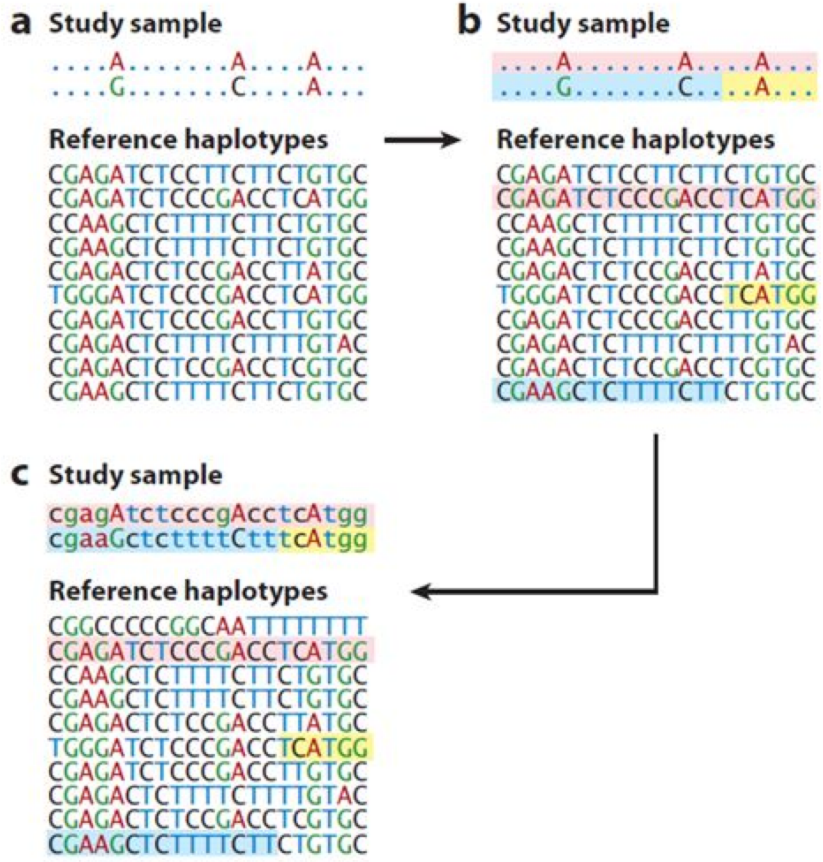
- A consequence of genetic recombination is linkage disequilibrium
- Defined as the "non-random association of alleles at different loci"
- Different from what would be expected if alleles were independently



Source: Ostrer

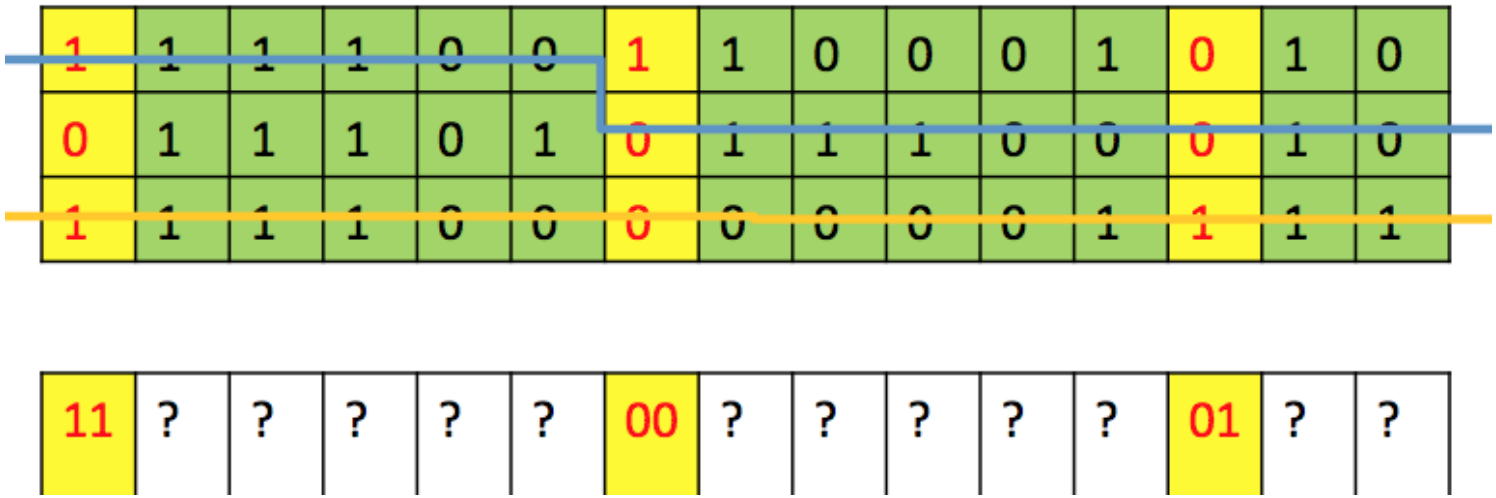
Theory of Imputation

- Motivating theory for our compression algorithm
- Individuals with tagged variants can have their haplotypes inferred from a reference panel
- Unobserved individual haplotypes are hypothesized to be the recombination of existing haplotypes in the observed population



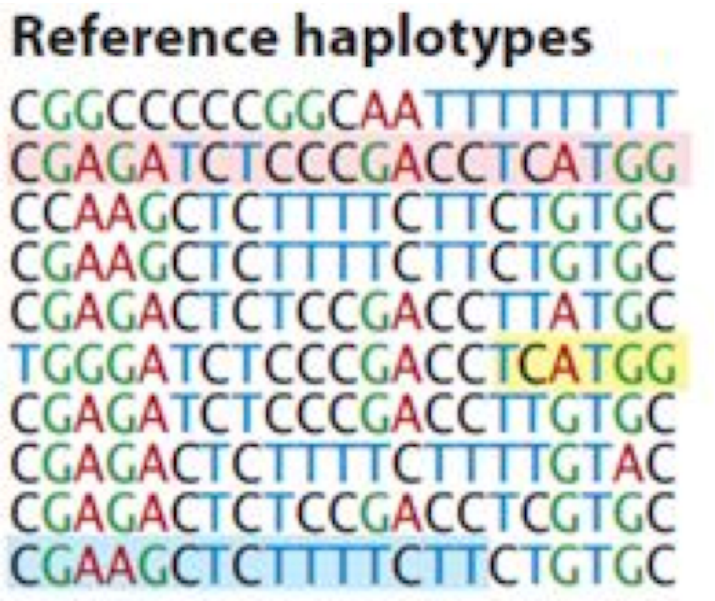
Source: Marchini

Example of imputation



Transition Probabilities

- We can summarize the observed individuals' haplotype structure by computing the transition probabilities between each pair of variants
- Recombination hotspots will result in greater transition probabilities
- Conserved regions will have low transition probabilities



Approach

Noting,

- Areas of high recombination vary between subpopulations
- A series of transition probabilities forms a Markov Chain

We hypothesize a matrix of transition probabilities provides a sufficient statistical basis to reconstruct population structure.

Implementation

1. Sufficient Markov Chain Transition Algorithm (SMCT)
2. Defining Subpopulations (Principal Components)
3. Recapturing Subpopulations via SMCT Algorithm

Sufficient Markov Chain Transition Algorithm

- Requires only simple matrix operations and random number generation
- Two main steps:
  1. Calculate and output sufficient statistics
     - Calculate the mean haplotype for each population for the first marker
     - Calculate the probability of transitioning to a new haplotype

$$\pi_{i,j} = \frac{\sum_i \text{state}_{i,j} \neq \text{state}_{i,j-1}}{\sum_i 1}$$

```
## MCT compression pseudocode

import data

foreach i in ethnicity
    MCT[1,i] = mean(haplotype[1,ethnicity==i])
    foreach j in markers
        MCT[j,i] = mean(haplotype[j]==haplotype[j-1])
    end
end

export MCT, ethnicity_key
```

Sufficient Markov Chain Transition Algorithm

- Two main steps (continued):
  2. Import sufficient statistics and randomly generate states
     - Random assignment of individuals k to ethnicity groups
     - Randomly generate first markers from mean haplotype for ethnicity i for $2 \times k$ alleles
     - Transition to new states based on the sufficient markov chain transition probabilities

$$\text{state}_{i,j,k} = (1 - \text{state}_{i,j-1,k}) \times (\text{rand}_{j,k} \leq \pi_{i,j}) + (\text{state}_{i,j-1,k}) \times (\text{rand}_{j,k} > \pi_{i,j})$$

```
## MCT expansion pseudocode

import MCT, ethnicity_key

n=number_individuals
pop = matrix(NA,markers,2*n)
eth = random(ethnicity,n)
foreach i in ethnicity
    rand = random(0,1,length=2*n[eth==i])
    start.state = rand <= MCT[1,i[eth==1]]
    foreach j in markers
        rand = random(0,1,length=2*n[eth==i])
        transitions[j] = rand <= MCT[j,i[eth==1]]
    end
end

export transitions, ethnicity_key
```

Sufficient Markov Chain Transition Algorithm

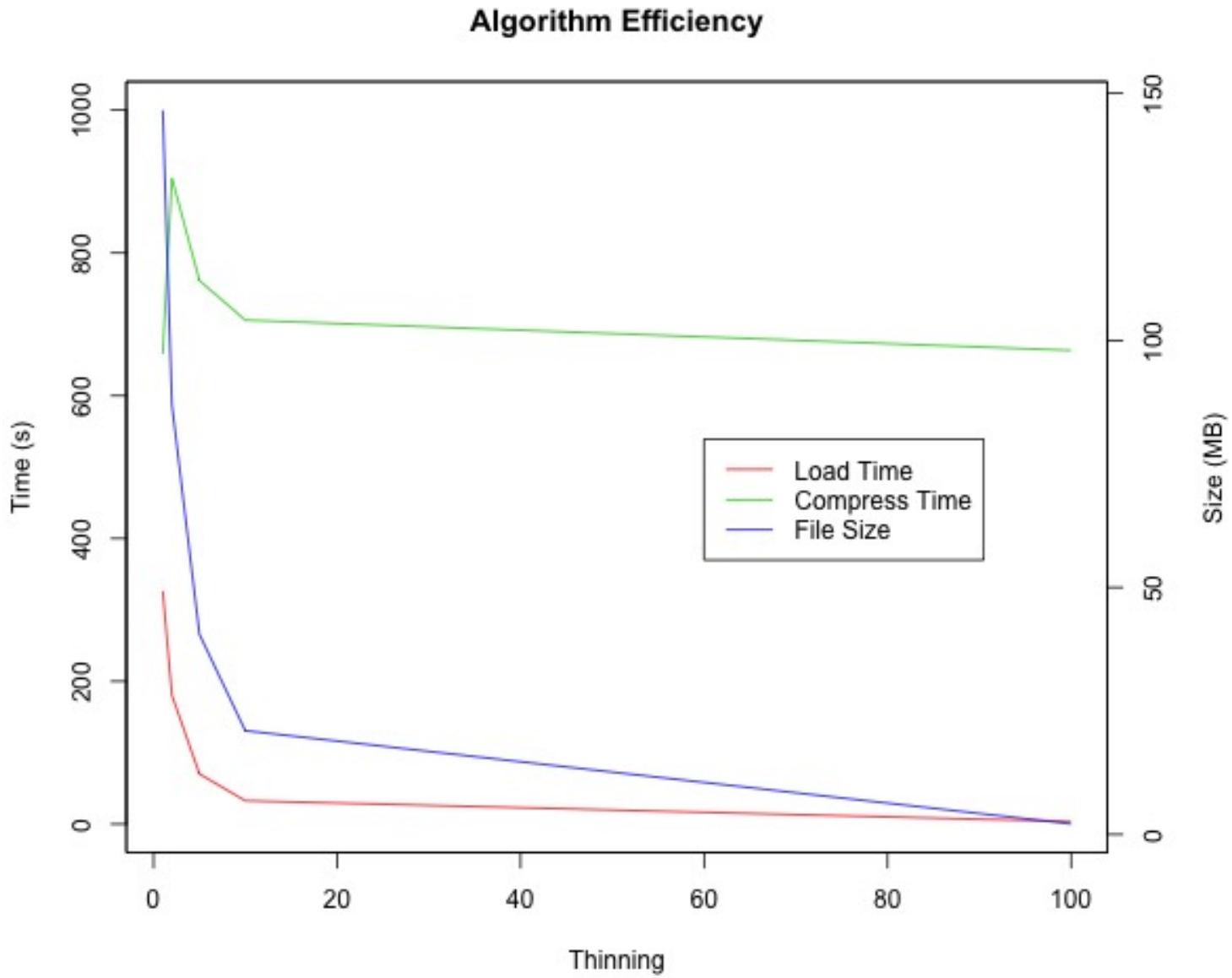
- Compression Gains:
  - File size is a direct consequence of the number of ethinicities and alleles
- Speed Gains:
  - Data regeneration via the sufficient statistics is a direct consequence of the transitions and the desired number of individuals
- Question:
  - Given a markov chain, can population substructure be regenerated?
- Concept:
  - Thinning
  - Collapsing sample to every j^{th} transition by averaging

Efficiency Analysis

File Type	Size (MB)	Relative Size	Compress Time (s)	Load Time (s)
Plain Text Document	3,609.6	1.0	NA	519.74
Zipped File	72.4	0.020	60.34	543.96
SMCT File (No Thinning)	146.6	0.041	658.23	326.35
SMCT File (Thinning=2)	87.0	0.024	905.08	180.49
SMCT File (Thinning=5)	40.5	0.011	761.23	70.25
SMCT File (Thinning=10)	21.0	0.006	705.93	32.66
SMCT File (Thinning=100)	2.1	$<$0.001	663.23	3.33

- Easily implemented in both R and julia
- Analysis based on 1.7 GHz Intel Dual-Core i7 Processor with 8 GB RAM running julia-0.3.11
- Resultant file may of course by zipped, resulting in greater storage gains

Efficiency Analysis



Principal Component Analysis (PCA)

- Consider n independent observations, $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from a p element random vector $\mathbf{x}$.
- Let $\tilde{z}_{i1} = \mathbf{a}_1' \mathbf{x}_i$ for $i = 1, \ldots, n$. The first principal component, $\mathbf{a}_1$, is

$$\arg\max_{\mathbf{a}_1} \frac{1}{n-1} \sum_{i=1}^{n} (\tilde{z}_{i1} - \bar{z}_1)^2$$

  subject to $\mathbf{a}_1' \mathbf{a}_1 = 1$. (*Intuition:* $\mathbf{a}_1$ *is the transformation of the* $\mathbf{x}_i$*'s that maximizes the variance, subject to a normalizing constraint.)*
- Additional components are solved for using the same process and the additional constraint that the $\tilde{z}_{ik} = \mathbf{a}_k' \mathbf{x}_i$ are uncorrelated with previous components.

Source: Jolliffe, Principal Component Analysis.

PCA Algorithm Analysis

- Principal components usually calculated using Singular Value Decomposition (SVD). Given an $m \times n$ matrix A, its SVD is:

$$A = U\Sigma V^T$$

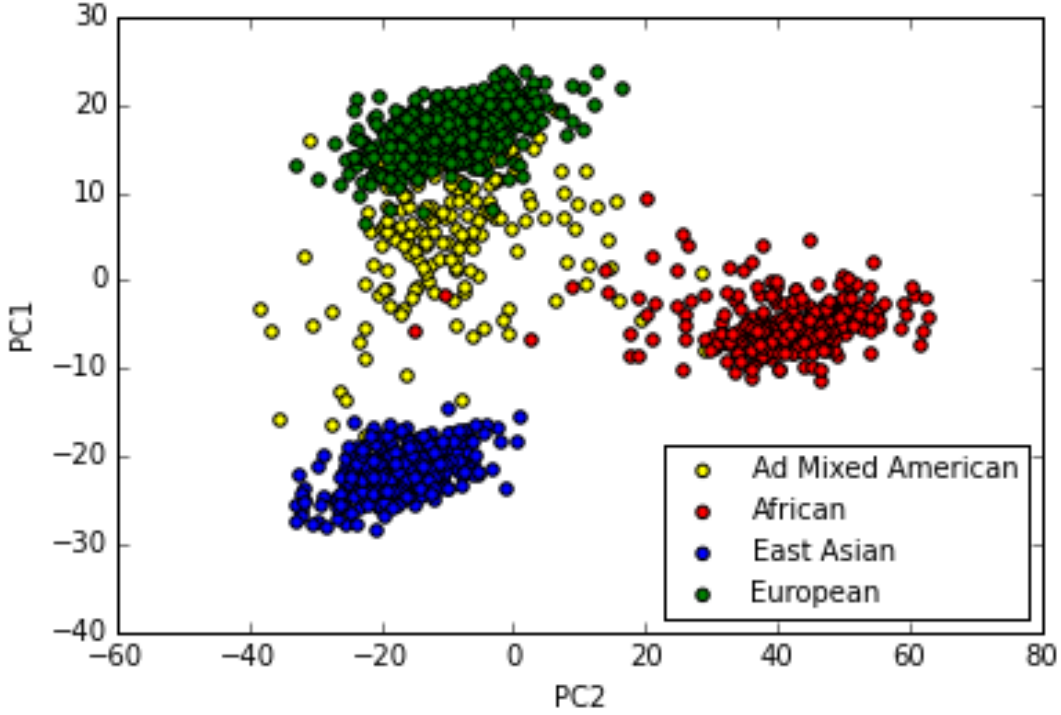
  where $U_{m\times m}$ and $V_{n\times n}$ are both orthogonal matrices and Σ is an $m \times n$ diagonal matrix with entries:

$$\sigma_1 \geq \cdots \geq \sigma_p \geq 0$$

- The σ_i's are called the singular values of A and correspond to the square root of the eigenvalues of $A^T A$.
- Numerous algorithms exist to do SVD (Golub-Kahan, for example) and are generally $O(n^3)$.

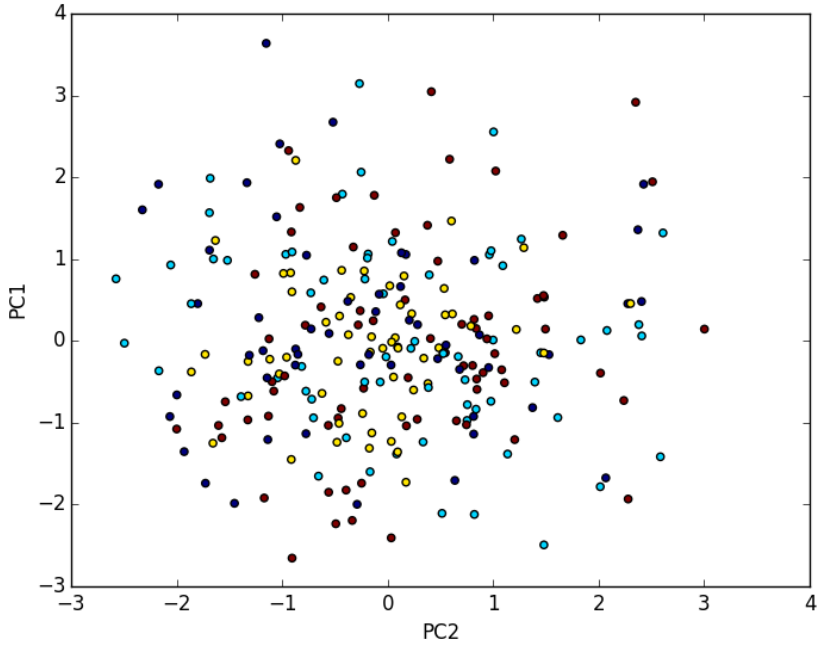
Source: Golub, Matrix Computations.

Defining Subpopulations



- PCA of individual level haplotype data ($\sim$ 8000 samples).
- First two components explain about 75% of the variation.
- Population to Super-Population mapping (i.e. CHS $\rightarrow$ EAS) available at `1000genomes.org`.

Subpopulation Regeneration



- PCA of simulated individual level haplotype data at the Super-Population mapping.
- First two components explain only about 20% of the variation (versus 75% in original data).
- No evident subpopulation clustering

Conclusion

- Despite theory, we are unable to reproduce subpopulation clusters
- Possible fixes:
  - Conditional transitions
    - Conditional on state i, probability of transition to state j
    - Opposed to the probability of any transition
    - Sacrifice compression and efficiency gains
  - Small sample size
    - $n = 500$ in simulations
    - Unable to capture recombinant patterns without large enough population
    - Only using genetic information on chromosome 22

Works Cited

- Casella G, Berger RL. Statistical inference. Pacific Grove, CA: Duxbury; 2002.
- Golub, Gene H., and Charles F. Van Loan. Matrix computations. Vol. 3. JHU Press, 2012.
- Jolliffe, Ian. Principal component analysis. John Wiley & Sons, Ltd, 2002.
- Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nature genetics. 2007 Jul 1;39(7):906-13.
- Ostrer H. A genetic profile of contemporary Jewish populations. Nature Reviews Genetics. 2001 Nov 1;2(11):891-8.
- Yeung, Ka Yee, and Walter L. Ruzzo. Principal component analysis for clustering gene expression data. Bioinformatics 17.9 (2001): 763-774.