Cryptic Population Substructure and Fuzzy Clustering

Senior Honors Thesis

Department of Computer Science

Author: Jacob Franco

Advisor: Sorin Istrail

Second Reader: Eli Upfal

School: Brown University

Date: May 2012

Table of Contents

Acknowledgements:

I would like to thank Prof. Istrail for introducing me to the wonders of Computational Biology. His mentorship and guidance has greatly influenced my path in life. A special thanks to Prof. Upfal for making the commitment to being my second reader. Finally, I owe great deal of gratitude to Derek Aguiar for all of his help in analyzing and finding my genomic data.

Abstract:

We describe a clustering model for large Single Nucleotide Polymorphism (SNP) arrays to assign individuals into populations. We do this using the Fuzzy C-Means (FCM) algorithm, which uses fuzzy set theory to assign individuals proportionally into populations. We are able to detect admixed individuals by looking at the proportions that they are assigned to populations. The purpose of this algorithm is to group similar individuals together for use in Genome Wide Association Studies (GWAS). It has been proven that cryptic or unknown population substructure can lead to spurious findings, and hide true correlations. Structure, the current standard in population clustering for GWAS uses Markov Chain Monte Carlo to assign individuals into populations. My thesis is to show that the FCM finds clusterings that make biological sense, and to compare those results to Structure's. This was done using three datasets used in published Structure studies, a dataset taken from the International Haplotype Map Project, and another set that includes 53 World Populations.

Introduction:

History of the Human Genome:

The completion of the first human genome sequencing by the International Human Genome Sequencing Consortium (IHGSC) (Lander et al., 2001), and by Celera (Venter et al., 2001) started the field of computational genomics. The information from the projects enabled us to compare human genomes and to quantify diversity in populations at a genomic level. However it also gave the scientific community a false

sense of optimism about our ability to predict and treat diseases based off genetic markers. Our misunderstanding of the complexities of the genome was shown through our initial guesses on the amount of protein coding genes. Before the human genome was sequenced most estimates on the number of protein coding genes for humans were over 100,000. However estimates by IHGSC, Celera, and even modern day estimates are between 20,000 and 40,000. The question of how humans can have such variability with so few genes has driven biologists to research how genes are regulated through transcription factors, and to explore how RNA can effect gene regulation and mRNA processing.

Another Biological application to the human genome is the ability to compare whole human genomes to find possible mutations that are correlated with disease. However, the human genome is far too large to compare each nucleotide. Multiple genomic surveys have suggested that only 1 in every $3.7 * 10^4$ nucleotides are different (Stephens et al., 2001). So instead researchers will want to look at points in the genome that give are the most informative about differences between people. These points are called tagging Single Nucleotide Polymorphisms (SNP). Computational Biologists use the fact that the genome consists largely of blocks of common SNPs with relatively little recombination shuffling within smaller haplotype blocks(Patil et al., 2001). Then they determine within those blocks if there is a correlation between two SNPs by a measure called Linkage Disequilibrium (LD). LD refers to the non-random association of alleles at different loci in haplotypes(Weiss & Clark, 2002). So if one tagging SNP is in high LD with many other SNPs, then when you sequence that tagging SNP you also get information about all the other SNPs that it is in LD with. Many algorithms go into defining these blocks, e.g. (Gabriel et al., 2002), and how to determine the most informative SNPs (Carlson et al., 2004). A major problem with haplotype blocks is that SNPs at the end of a block may be in LD with the SNPs at the beginning of the next block. Another tagging technique was developed by finding a neighborhood of predictive SNPs for each SNP and then maximization the informativeness measure over all SNPS (Halldorsson et al., 2004). By maximizing the informativeness of SNPs, we can minimize the number of SNPs needed and thus decrease the cost of human sequencing of those SNPs.

GWAS Studies and the Missing Heritability Problem:

Genome Wide Association Studies (GWAS) utilize these tagging SNPs in order to predict which SNPs are correlated with disease. A GWAS study has the basic case-control structure where individuals are sequenced and then the corresponding SNP frequencies between the two groups are compared. When GWAS studies first started most believed that SNPs for most genetically inherited diseases would soon be found. This, however, was not the case and this problem is called the missing heritability problem, because we cannot find where in the genome the disease is being inherited. A prime example of the missing heritability problem is type 2 diabetes (T2D). T2D has been subject to more Genome-wide Association Studies (GWAS) than any other genetically linked disease. Despite an extensive amount of studies less than 10% of the heritability of T2D can be explained by Single Nucleotide Polymorphisms (Bonnefond, Froguel, & Vaxillaire, 2010).

Mary-Claire King (McClellan & King, 2010) described the problem as, "In molecular terms, we suggest that human disease is characterized by marked genetic heterogeneity, far greater than previously appreciated. Converging evidence for a wide range of common diseases indicates that heterogeneity is important at multiple levels of causation." What she means is that there are a variety of factors that are problematic to the GWAS structure.

One such problem is that 90% of the differences between people are actually ancient (more than 50,000 years ago) polymorphisms (Cavalli-Sforza & Feldman, 2003) and that those mutations may be the ones recorded as tagging SNPS (Tishkoff & Verrelli, 2003). This problem is exacerbated by the fact that there are 175 new alleles per person (Nachman & Crowell, 2000). With an exponential growth model it is clear that there are many newer alleles that may account for disease. Also the problem exists that because of the rate of new alleles, a sickness may be caused by an extremely rare mutation.

The authors of "Missing heritability and strategies for finding the underlying causes of complex disease,"(Eichler et al., 2010) also describe several possible solutions to the missing heritability problem. Eichler claims that the heritability may not be in single SNPs but in the large duplications and deletions in the genome that Biologists cast

off as non-important because they are introns. Flint says that part of the problem may lie with groups of SNPs that compose a phenotype for the organism. The problem with this model is that without prior information about which groups of SNPs compose a phenotype the problem becomes NP-hard on the size of the size of the phenotype. Kong says GWAS may need to take into account, which SNPs come from which parents as parental origin of genes may have different recombination rates.

There are a lot of possible reasons for the missing heritability problem, but I focused on how to alleviate the effects of cryptic population substructure. Cryptic population substructure is the subtle differences in ancestry between cases and controls in GWAS. Researchers take steps to alleviate this problem by taking into account the average differences in SNP frequencies between populations and by removing individuals who are sufficiently far away from the population and deemed as outliers (Price et al., 2006; Purcell et al., 2007). Cryptic population substructures can lead to false positives, (Chakraborty & Smouse, 1988) mask true positives (Deng, 2001) in GWAS. Often people do not know the extent of their admixture and certain social and geographical populations may actually have many subpopulations in them. Therefore, this problem cannot be solved by allowing participants to mark their own population of origin (Serre et al., 2008).

I will now present two pieces of software that attempt to solve the problem of sorting individuals into populations. The first is Structure developed by the Pritchard lab in 2000 and extended three times (Falush, Stephens, & Pritchard, 2003, 2007; Hubisz, Falush, Stephens, & Pritchard, 2009; Pritchard, Stephens, & Donnelly, 2000) and my program utilizing the Fuzzy C-Means Clustering Algorithm (Dunn, 1973), extended by (Bezdek, 1984).

Structure:

Structure is a Bayesian clustering algorithm that utilizes a Markov Chain Monte Carlo (MCMC) approach to probabilistically assign individuals to populations. The goal of MCMC is to sample from a distribution P or approximate the expected value f(x). Where in this case F(x) would be the probability of a person in a population. In a standard

Monte Carlo approach E[f(x)]= $\frac{1}{n}\sum_{i=1}^{n} f(x_i)$ for large n where x_i is sampled from probability density p. The density p is not possible to model using formulas, so the next powerful tool is to try importance sampling or rejection sampling. Importance sampling requires the statistician to have an equation for a density q(x) that roughly models that of p(x). Using the definition of Monte Carlo once more we have $E[f(x)] = \int f(x)p(x)dx$ but now lets multiply the left side of the equation by $\frac{q(x)}{q(x)}$ to get $E[f(x)] = \int [f(x)\frac{p(x)}{q(x)}]q(x)dx$. What I enclosed in the brackets can be seen as the new f(x) function, lets call it g(x). Using the Monte Carlo definition we can solve that integral with $\frac{1}{n}\sum_{i=i}^{n} g(x_i)$ where x_i is drawn from q instead of p. In rejection sampling you use the basic Monte Carlo equation but sample x from q where q(x) is proportional to f(x) and also that q(x)>f(x) for all x. However, there is a probability that you do not accept the value of x and that probability is given by P(Reject x|f,q)=q(x)-f(x). For large number of samples get possible values of x and probabilistically accept or reject x from the probability above. The two main caveats with importance and rejection sampling is that both require a proposal distribution q that is either closely matches f, or is proportional to f. Since SNPs arrays are in high dimensional space proposing good q distributions is not possible.

The MCMC approach does not require a sample distribution as it instead uses a random walk principle to calculate the probabilities. The intuition behind the MCMC method is that you start somewhere on the walk. In terms of populations, you would start by initializing everyone into a population. The next step is to probabilistically choose which way to take the next step and shift people in your population. Then you continue this process for a large number of steps. The theory to this approach is that you start in some low probability space, but after a large number of time steps you enter a high probability state and then you stay in that high probability space. The Markov Chain part of the MCMC is there to state that the probability of state or iteration i, is only dependent of the (i-1) iteration.

Subscripts:

K=The population of origin, for k populations K=(0,1..,k-1)

J=The number of distinct alleles. For GWAS J=(0,1)

L=Locus. For l SNPs L=(0,1..,l-1)

I=Individual

Parameters:

1. P_{KLJ}=Frequency of allele J in Locus L in population K
2. $Q_k^{(i)}$=Proportion of individual i in population k.
3. $Z_L^{(i,a)}$=Population of origin of locus L, in individual i, for copy a. a=(0,1) for diploids.
4. $X_L^{(i,a)}$=SNP value (J) at locus L in individual i at allele copy a

Structure Algorithm:

1. Initialize $Z^{(1)}$

Repeat while m<M

2. Sample $P^{(m)}$, $Q^{(m)}$ using $Z^{(m)}$
3. Sample $Z^{(m)}$ using $P^{(m)}$, $Q^{(m)}$

Updating P:

Before we can sample from P we first must update the probabilities of P_{KLJ} from parameter 1. The first step in recalculating the new P_{KLJ} is to create a count N_{KLJ}, which is the number of matching SNP values (J), at locus (L), in population K. More formally $N_{KLJ}= \sum_{\forall(i,a)} (X_L^{(i,a)} = J) \,\&\, (Z_L^{(i,a)} = k)$, where $Z_L^{(i,a)}$ and $X_L^{(i,a)}$ is described in parameters 3 and 4 respectively. Knowing N we update the P using $\text{Prob}(P_{KL} \mid X,Z) = D(\lambda_1 + N_{KL1},...,\lambda_J + N_{KLJ})$. Where λ=1 and D is the Dirichlet distribution.

Updating Q:

Similar to P we will need a count called $M_k^{(i)}$ and then use the Dirichlet distribution. $M_k^{(i)}$ is the number of SNPs in individual i that are part of population k. Formally it is $M_k^{(i)} = \sum_{\forall(l,a)} Z_l^{(i,a)} = K$. Knowing M Q is modeled by the Dirichlet distribution with $\text{Prob}(Q^i \mid X, P) = D(\alpha + M_1^i, ..., \alpha + M_k^i)$ where $\alpha > 0$.

Updating Z:

The probability that an individual i's locus l, allele copy a is in population k is proportional to the proportion an individual is in population k times the probability that, that specific locus and allele is in population k. The formal equation is written below.

$$\text{Prob}(Z_l^{(i,a)} = k) = \frac{Q_k^i * \Pr(X_l^{(i,a)} \mid P, Z_l^{(i,a)} = k)}{\sum_{k'=0}^{K} Q_{k'}^i * \Pr(X_l^{(i,a)} \mid P, Z_l^{(i,a)} = k')}$$

Fuzzy C-Means

The Fuzzy C-Means (FCM) algorithm is an Expectation-Maximization (EM) algorithm and an extension of the K-Means clustering algorithm with the major difference that the FCM allows an object to be proportionally placed in a cluster or population, which is advantageous in sorting admixed individuals. The EM algorithm was first formalized in 1977 by Dempster (Dempster, Laird, & Rubin, 1976). The goal in an EM algorithm is to find the maximum likelihood of a set of unknown parameters like the population of origin, based off a set of known parameters like SNP arrays. In the expectation step, the missing data is estimated using your known parameters and our current estimate of the model parameters. In the M step we maximize the objective function assuming the missing data is what we calculated in the E step. By maximizing the objective function we create a new estimate of the model parameters and then repeat the E step. So in the case of sorting populations:

Let:

$\overline{X}$ = Vector of people's SNP arrays (known)

$\overline{\theta}^{(t)}$ = Probability of a person is in a population at step t (unknown)

$\overline{Z}^{(t)}$ = Distance a person is from a given population at step t

Expectation Step:

$$Z^{(t)} = E(Z \mid X, \theta^{(t-1)})$$

Maximization Step:

$$\theta^{(t)} = E(\theta \mid X, Z^{(t)})$$

The EM algorithm has become a popular tool because of its low memory needs and that it is guaranteed to converge to a local minima (Wu, 1983). The FCM itself has been used to cluster many problems in biology and medicine: For example, it was used to cluster possible cancerous cells based on gene expression (Zhang, Adamu, Lin, & Yang, 2011), and brain lesion detection from MRI images (Pham, 2010). However, I have not found a study that uses the FCM to cluster individuals based on SNP arrays. The Fuzzy C-Means algorithm is as follows.

FCM Algorithm:

Let:

W_{pc} = Weight matrix. Proportion person 'p' is in cluster 'c' with constraint $\sum_{c'=0}^{C} W_{pc'} = 1$

V_{sp} = SNP Matrix. SNP value at locus 's' for person p. V_{sp}= (0, .5, or 1)

V_{sp}=.5 if the person is heterozygous or in the case of missing data.

U_{sc} = Average Matrix. Average weighted value of all SNPs at locus 's' in cluster c.

The goal of the FCM is to minimize the objective function J which is to minimize the total distance a person is to the center of their cluster.

$$J = \sum_{p=0}^{P}\sum_{c=0}^{C} W_{pc}\sqrt{\sum_{s=0}^{S}(V_{sp} - U_{sc})^2}$$

The FCM is minimizing the Euclidean distance between an individual and a cluster and then weighting it by the fraction that the person is in the cluster. The initialization of the algorithm needs both the V_{sp} matrix, and the W_{pc} matrix, a fuzzy factor m, and the

number of clusters C. Allowing the user to input a W_{pc} matrix will increase the likelihood that the FCM's local minimum would be the global minimum. However, for most of the runs used in this paper, no prior knowledge for W_{pc} was used. Instead W_{pc} was generated randomly for each trial by randomly choosing a value in each cell of the matrix using Java's Math.random() function and then normalizing over the rows to sum to 1.

Algorithm:

While [$W^t - W^{(t-1)} > \epsilon$ & t<max iterations]

Step 1:Expectation

Step 2:Maximization

end

Step 1: Expectation step.

Update U_{sc} (Average SNP value in the population) by taking the weighted average of each SNP value over the sum of all the weights in that cluster. $U_{sc} = \frac{\sum_{p=0}^{P} V_{sp} W_{pc}}{\sum_{p=0}^{P} W_{pc}}$

Step 2: Maximization step.

Update W_{pc} (Proportion of individual in population. This step is done by calculating the distance person p is from cluster c, weighted against how far p is from each other cluster, and scaled by the fuzzy factor. This step also has the property that for a given person, the sum across the clusters will equal one.

$$W_{pc} = \frac{1}{\sum_{c'=0}^{C} \left[\frac{\sqrt{\sum_{s=0}^{S} (U_{sc} - V_{sp})^2}}{\sqrt{\sum_{s=0}^{S} (U_{sc'} - V_{sp})^2}} \right]^{\frac{2}{m-1}}}$$

It is interesting to note that there will be SNPs that are more 'informative' than other SNPs. If a SNP is highly correlated in one population and not present in other populations

then if a person with the SNP attempted to be sorted into a population without the SNP then the distance gets increased by one. However, if the SNP is present in all populations then it will not increase the distance at all because $(U_{sc}-V_{sp})=0$ for all p and c for that SNP s.

Fuzzy Factor:

The Fuzzy Factor (m) is allows you to control how "noise" in the data affects the certainty that a person is clustered into a population. The simplest explanation of how the Fuzzy Factor affects the data is shown in the case-by-case analysis below. This behavior is similar to that of the Ising model invented by physicist Wilhelm Lenz and the one dimensional model was solved by Ernest Ising in 1925. This model was developed to model electron spins for ferromagnetic materials. It was designed such that an electron has a high probability to share the same spin as its neighbors. In this model you have an N-dimensional lattice, with values either 0 or 1 to represent electron spins. However, the probability that an electron spins to a certain value is also dependent on the temperature. With a high temperature the probability of any electron being in a single state is about one half. At low temperatures the electrons have a higher probability of matching its neighbors spin and the system goes to a stable state with all zeros or all ones. There is a temperature T_{Crit} that when $T<T_{Crit}$ the system will eventually go to a single state, and when $T>T_{Crit}$ the system will never stabilize. At around the T_{Crit} the temperature component does not overwhelm the similar spin preference and clusters of spins will form. The fuzzy factor mimics this property in the assignment of probabilities to clusters. When m is close to one individuals are hard clustered into a population similar to when $T<T_{Crit}$. When m is large each individual has an even probability of being in every cluster much like how the spin neighbors are disregarded in the Ising model so too are the SNP arrays. Calculating T_{Crit} for lattices of greater than 2-Dimensions is NP-Complete (Istrail, 2000) and similarly no value of m can be decided as optimal for a dataset (Bezdek, 1984).

Interestingly, Structure has a similar feature in their use of α to determine the prior probability distribution of a SNP to a cluster. α is the pseudo count value. Pritchard says that for large α "models each individual as having allele copies originating from all C populations in equal proportions. This makes sense because if α is large then the actual count that the MCMC makes will not be as powerful as very small values of α. Small α

will restrict movement of the model, and thus eventually each individual will seem as though they are originating mostly from a single population. Structure deals with the α problem by varying it across its MCMC steps and fitting it to the data as the program runs. For FCM the fuzzy factor is an input parameter, but it can be varied across different runs. Later in this paper I will discuss how to determine when the clustering is too hard or too soft and how to adjust the fuzzy parameter accordingly.

Fuzzy Factor Intuition:

Weight Equation: for easy reference

$$W_{pc} = \frac{1}{\sum_{c'=0}^{C} \left[\frac{\sqrt{\sum_{s=0}^{S} (U_{sc} - V_{sp})^2}}{\sqrt{\sum_{s=0}^{S} (U_{sc'} - V_{sp})^2}} \right]^{\frac{2}{m-1}}}$$

Case 1: Small m, c is closest fit to p.

Assume for a person p, that cluster c is the best fit. Then the distance between p and c will be smaller than all other c'. If m is close to 1, then the exponent is large. A large exponent on a number less than 1 drives it towards zero. So for all other c' they will sum toward ε where ε is very small. Because c is also included and that fraction will equal 1, the total weight assignment for $W_{pc} = \frac{1}{1+\varepsilon} \approx 1$

Case 2: Small m, c is no closest fit to p.

Symmetrically if cluster c is not the best fit then there must be a cluster c' that is closer than c, then $\frac{|Person \text{ p to c}|}{|Person \text{ p to c'}|} > 1$. Therefore, as m approaches 1, that fraction goes to infinity. Now $W_{pc=} \frac{1}{1+\infty} \approx 0$.

Case 3: Large m

When m is large the exponent approaches zero. Therefore, all of the distance fractions raised to a zero power will be approximately 1 and thus each have an equal weight $W_{pc=}\frac{1}{\#\ \text{of clusters}}$

Whitefish Test:

In order to compare the two programs I thought it would be best to take an example straight from Structure's paper. Structures 2007 follow up (Falush et al., 2007) included a whitefish dataset with two populations: 23 normal and 24 dwarves, with 440 Amplified Fragment Length Polymorphisms (AFLP). The sample was taken from (Campbell & Bernatchez, 2004). According to NCBI "AFLPs are differences in restriction fragment lengths caused by SNPs or indels that create or abolish endonuclease recognition sites." However, an AFLP is not stored as a length but rather as a 1 or 0 depending if the endonuclease site was recognized. Thus the data obtained by an AFLP experiment can be used the same way as a SNP in the FCM algorithm and Structure program.

According to their 2007 paper, Structure found 2 dwarves, and 5 normal whitefish with mixed ancestry and classified 1 normal as a dwarf. With m=1.2 and the number of clusters K=2, I found the same 5 normal fish with mixed ancestry. The five admixed normals had admixed levels of [.30%, .23%, .23%, .67%, .54%] while the rest of the classified normals had admixture rates below .05. Both programs classified the same normal ancestry fish as a dwarf with 92% certainty. I found three dwarves of mixed ancestry not two as described in Structure's paper however they had much smaller admixture rates than the set of admixed normals. When I ran Structure with a 1000 burn-in and 10,000 iterations without allowing it to have a prior guess on population I saw that it also found the same three admixed dwarves. When I turned the USEPOPINFO flag on as described in the paper, the results from Structure did not match the output that I received and thus I have omitted those results.

I then attempted to add prior information into the FCM algorithm. I set the initial weight matrix to have a probability of 80% for a dwarf being a dwarf and vice versa for a

normal and found that the results were similar to that of the random initialization. In fact all runs with random initialization matrices gave similar results and found the same admixed individuals across runs.

Table 1: Comparison of Structure VS FCM on Whitefish Data

Individual	*Structure % Dwarf*	*Structure % Normal*	*FCM % Dwarf*	*FCM % Normal*
Dwarf 11	.84	.16	.92	.08
Dwarf 23	.83	.17	.88	.12
Dwarf 24	.84	.16	.87	.13
Normal 29	.47	.53	.30	.70
Normal 36	.17	.83	.23	.77
Normal 42	.47	.53	.23	.77
Normal 45	.99	.01	.92	.08
Normal 46	.65	.35	.67	.33
Normal 47	.24	.76	.54	.46

Fuzzy Factor on Whitefish Data:

To examine the effect of the fuzzy factor has on the data I ran the Whitefish data with different m values and K=2, and recorded the average proportions a non-admixed individual in its proper cluster. The results are shown in figure 1. When the exponent was low, thus the fuzzy factor large the certainty of an individual goes towards .5, and when the exponent is large thus a small m the certainty approaches 1. This jump takes place over a narrow range of exponents. When I ran trials on simple simulated data I found that in order to detect mild admixture levels you have to lower your certainty.

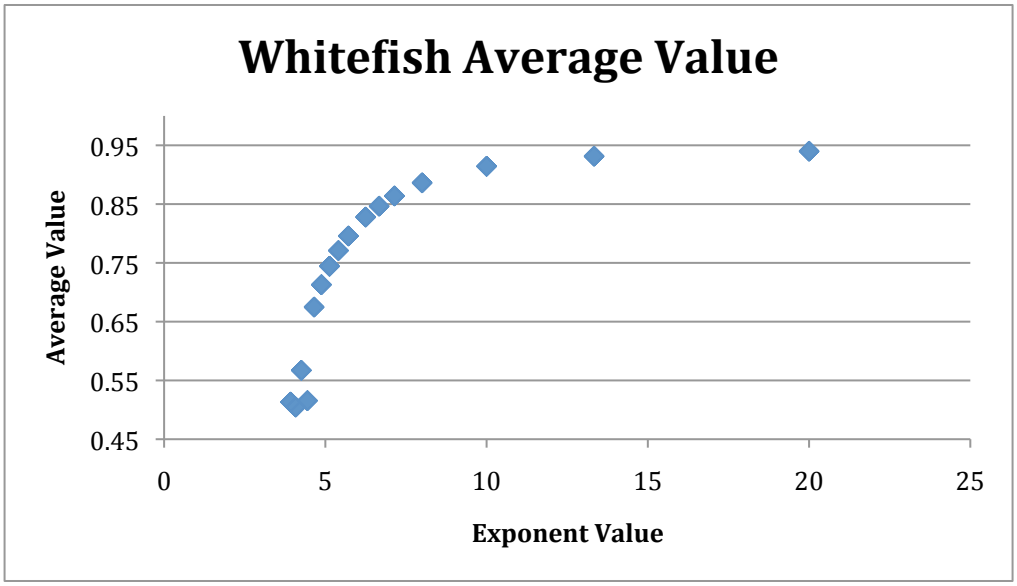

Figure 1: Fuzzy Factor Effect on Whitefish Clusters

HapMap Tests:

HapMap Data:

The International HapMap Project was started in 2002 with the goal to develop a haplotype map of the human genome and describe the common patterns of human DNA sequence variation. The HapMap Project has sequenced various populations and allows the scientific community to download the SNPs for their own use. The SNPs that I used for this project came from their 2008-2010 phaseII+III forward strand. This data is available at http://hapmap.ncbi.nlm.nih.gov/downloads/genotypes/2010-08_phaseII+III/forward/2010-08_phaseII+III/. I used their Chromosome 1 data from three populations: 86 African-American sampled from Southwestern USA (ASW), 173 Utah residents with European ancestry (CEU), and 138 Han Chinese sampled from China. I chose chromosome 1 because it is the largest chromosome and has the largest amount of SNPs. The three populations had 122,751 SNPs in common and the missing data rate was below 5%. When I encounter missing data for an individual I would assume that the individual was heterozygous for that SNP.

HapMap on Large Data:

I ran the FCM on all combinations of two for the three populations, with two different m parameters m=1.15 and m=1.1, and I ran the Structure. I used all 122,751 SNPs on the FCM as input and 1000 randomly selected SNPs for the Structure with a burnin of 1,000 and the number of iterations equal to 100,000. I chose 100,000 because I ran Structure on 20,000 I saw good results, so to further increase Structure's accuracy I increased the number of iterations to 100,000. I could only use 1000 SNPs for Structure because it would take over two weeks to complete the calculation with all 122,751 SNPs and saw fairly good accuracy with the 1,000 SNPs. The results are shown in Table 2.

Fuzzy Factor Affects on Admixture:

Table 2 shows how the fuzzy factor m can affect the admixture rates. A clear pattern can be seen across all the trials between m=1.15 and m=1.1 and that is that when m=1.1 the percentage of admixture decreases but admixture is still detected. For example in trial 1 m=1.15 there were 42 CHB that were shown to have an admixture of 25% to ASW, but with m=1.1 that admixture decreased to 8%. Although the rate decreased the 8% admixture was still a very distinguishable group in the m=1.1 because the rest were very hard clustered to 99%. This shows that with the FCM you should not predefine what is the percentage needed for admixture before you run the trial, but only after when clears groups emerge. In trial 2, m=1.15 none of the ASW individuals had a clustering percentage greater than 90% but most were close to 100% when m=1.1. For the 86 ASW individuals in trial 2 of m=1.15 I would say that all those individuals are relatively close together and a cluster despite them having only 86% of being in the ASW cluster because there were no individuals who had a high percentage of being in that specific cluster. For the two values of m, no individual was sorted in different clusters and the same number of individuals was sorted in admixed and non-admixed populations.

I then compared the m=1.15 clustering's with Structures results. Across all the trials the FCM and Structure assigned each individuals to the same dominant population, where the dominant population is the population with >50%. The max differences in admixture assignment across the populations of ASW, CEU, and CHB were 15%, 38%, and 15% respectively. In the 38% case Structure assigned an CEU individual to be 95%

in CEU, while my FCM assigned the individual to be 58% CEU. However, for this particular trial I would have gone with the m=1.1 because the clustering seemed to be too soft given that none of the ASW were hard clustered. Also it is possible that the admixture was not picked up in Structure because the admixture was not contained in the sample of 1,000 SNPs that it was provided. For the vast majority of cases Structure agreed with the FCM on the wrong assignment of individuals, and locating admixed individuals.

Table 2: HapMap Data. All Combinations of Two

		m=1.15	*m=1.15*	*m=1.1*	*m=1.1*	*Structure*	*Structure*
Number of Individuals	Pop. of origin						
Trial 1 ASW-CHB							
		ASW	CHB	ASW	CHB	ASW	CHB
86	ASW	97	0	99	0	96±3	2±1
2	CHB	92	7	96	3	78	22
42	CHB	25	75	8	92	30±6	70±6
94	CHB	1	98	0	99	3	97
Trial 2 ASW-CEU							
		ASW	CEU	ASW	CEU	ASW	CEU
1	ASW	80±7	20±7	80	20	71	29
1	ASW	80±7	20±7	88	12	79	21
84	ASW	80±7	20±7	95±5	5±5	98	2
77	CEU	40±9	60±9	10	90	25±10	75±10
9	CEU	85	15	94	5±5	75	25
85	CEU	13	87	0	99	1	99
Trial 3 CHB-CEU							
		CHB	CEU	CHB	CEU	CHB	CEU
2	CHB	13	86	4	95	27	73
42	CHB	75	25	88	11	23±10	77±10
94	CHB	96	4	99	0	1	99
9	CEU	45	54	37	62	46	53
164	CEU	10	90	0	99	3	97

Overestimation of Clusters:

I then wanted to observe what would happen if I overestimated the number of clusters. I ran three trials of each combination of populations with the expected number of clusters equal to three. I ran the FCM with m=1.1, and Structure with a burnin of 1,000 and the number of iterations equal to 100,000. The results are summarized in table 3. In

general it seems that when the estimate of the number of populations is too high, the FCM splits the more heterogeneous cluster into two clusters. This is seen in trial 1, 2 where the CHB and CEU populations split and the ASW was still sorted into its own cluster. In the third trial between CHB and CEU, we knew from the two cluster results that there was a lot of admixture. So when the number of clusters was raised to three population I found that the extra population was filled by both CHB and CEU individuals. This most likely filled the role as a place for those admixed individuals to be sorted. Only the admixture rates changed when I used m=1.12, and when m=1.15 all clusters became equally likely.

Structure gave similar results in trials 1 and 2 but instead of hard sorting individuals into the third population, those individuals were admixed. It is clearer in Structure's case that the number of populations is wrong because no individuals are ever sorted into the third population. What the "correct" sorting is for a problem like this is the researchers preference to the question, is a group of similar admixed individuals its own population?

Table 3: HapMap Data. 3 Clusters of 2 Populations

Num. of Individuals	*Pop. of Origin*						
		FCM	FCM CHB	FCM CHB	Structure	Structure	Structure
Trial 1		ASW	1	2	ASW	CHB1	CHB2
16	ASW	99	0	0	84±6	16±6	1
70	ASW	99	0	0	95±3	2±3	0
2	CHB	66	33	0	58	42	0
42	CHB	0	99	0	1	99	0
94	CHB	0	0	99	0	50±5	50±5
Trial 2		ASW	CEU1	CEU2	ASW	CEU1	CEU2
1	ASW	53	46	0	48	52	
10	ASW	80±5	15±5	0	70±5	30±5	0
26	ASW	99	0	0	75±5	15±6	0
49	ASW	99	0	0	99	0	0
9	CEU	68	30	1	51	49	0
81	CEU	0	90±	10±5	0	98±2	2±2
83	CEU	0	10±5	90±5	0	2±2	98±2

Trial 3		FCM CHB	FCM Mix	FCM CEU	Num. of Individuals	Pop Origin	Structure CHB	Structure Mix	Structure CEU
42	CHB	12	88	0	42	CHB	45±15	55±15	0
96	CHB	99	0	0	96	CEU	98±2	2±2	0
89	CEU	0	96±3	4±3	81	CEU	0	99	0
84	CEU	0	4±3	96±3	9	CEU	35	65	0
					83	CEU	0	45±5	55±5

Three Population Test:

My next test was to try and sort all three populations at the same time with the expected number of populations equal to three. To be sure that this case was fair and that the data was not skewed by the FCM having more data, I ran the FCM on the same 1,000 SNPs as Structure. I ran this for the FCM with m=1.1 and Structure to have a burnin of 1,000 and used 100,000 iterations. I again got extremely similar results. All of the individuals were sorted into the same dominant populations. Almost all the same admixed individuals were found and the largest differences in percentages across the two programs were 4% for ASW individuals, 26% for CHB and 31% CEU and those largest differences are highlighted as red. The only admixed individuals that were found by Structure and not by the FCM were the individuals who were incorrectly associated with a population. Those individuals were the 2 CHB and the 9 CEU that were incorrectly sorted into ASW, however as shown in table 2 if you then run the program with only ASW and CEU/CHB the admixture is detected.

Table 4: HapMap Data. Structure FCM Comparison

		FCM	*FCM*	*FCM*	*Structure*	*Structure*	*Structure*
Number Of Individuals	Pop. of Origin	ASW	CHB	CEU	ASW	CHB	CEU
1	ASW	95	0	4	92	0	8
1	ASW	77	0	21	79	0	20
84	ASW	88	0	11	84	0	16
29	CHB	21±4	67±3	12±1	32±10	68±10	1±1
2	CHB	90	9	0	64	1	27
109	CHB	0	99	0	0	99	0
9	CEU	93	2	3	65	33	2
1	CEU	17	0	81	38	0	61
60	CEU	18±4	0	86±4	23±10	1	77±10
103	CEU	5±5	1	95±1	4±4	1	94±4

Rosenberg, Pritchard Dataset:

To further examine admixture within populations I downloaded a dataset from the Rosenberg Lab at Stanford University. Credit for the sequencing goes to the Human Genome Diversity project, and the credit of choosing SNPs that are correlated with the populations is from (Conrad et al., 2006) and extended by (Pemberton et al., 2008). This dataset had 2,810 SNPs from 53 populations.

Clustering 53 World Populations Into Six Clusters:

In 2002, the Rosenberg and Pritchard Labs ran the Structure on this data set for all 53 populations (Rosenberg et al., 2002). They found that the optimal clustering for the data set was K=6. I decided to test my FCM results against theirs. I used an m=1.03 recorded the results are recorded in table 5. Structure divided the populations into the categories of Africa, Europe/Middle East/Central/South Asia, Kalash, East Asia, Oceania, and America. The FCM grouped the nationalities similarly except instead of Kalash being its own population I split the large second cluster into Europe/Middle East and made Kalash into Kalash/Central/ South Asia. However, the FCM did find admixture between the Europe and South Asian group and vice versa, except for the Kalash, which were hard clustered together. Rosenberg and Pritchard revisited this data set in 2005(Rosenberg et al., 2005) and had a similar clustering scheme except that instead of

Kalash getting its own cluster it was sorted with the large Europe/Middle East/ South Asia cluster and the extra cluster went to splitting the Karitiana, and Surui (Brazilian tribes) from the rest of America cluster. According to the paper Structure found this clustering once in ten tries.

Table 5: 53 World Population Clusters

		Structure	*Structure*	*FCM*	*FCM*
Area	Nationality	Main	Secondary	Main	Secondary
Africa	Bantu	1		1	
Africa	Mandenka	1		1	
Africa	Yoruba	1		1	
Africa	San	1		1	
Africa	Mbuti Pygmy	1		1	
Africa	Biaka Pygmy	1		1	
Europe	Orcadian	2		2	3
Europe	Adygei	2		2	3
Europe	Russisan	2		2	3
Europe	Basque	2		2	3
Europe	French	2		2	3
Europe	Italian	2		2	3
Europe	Sardinian	2		2	3
Middle East	Tuscan	2		2	3
Middle East	Mozabite	2	1	2	3
Middle East	Bedouin	2	1	2	3
Middle East	Druze	2		2	3
Middle East	Palestinian	2		2	3
Middle East	Balochi	2	3/4	3	2
Central/ South Asia	Brahui	2	3/4	3	2
Central/ South Asia	Burusho	2	3/4	3	2
Central/ South Asia	Hazara	2/4		4/3	2
Central/ South Asia	Kalash	3		3	
Central/ South Asia	Makrani	2	4	3	2
Central/ South Asia	Pathan	2	3	3	2
Central/ South Asia	Sindhi	2	1	3	2
Oceania	Melanesian Bougainville	5		5	
Oceania	Papuan New Guinea	5		5	
America	Columbian	6		6	
America	Karitiana	6		6	
America	Surui	6		6	
America	Maya	6	3	6	
America	Pima	6		6	
Central/ South Asia	Bengali	NA		3	2
Central/ South Asia	Tamil	NA		3	2
East Asia	Han	4		4	
East Asia	Han North China	4		4	
East Asia	Dai	4		4	
East Asia	Daur	4		4	
East Asia	Hezhen	4		4	
East Asia	Lahu	4		4	
East Asia	Miao	4		4	
East Asia	Oroquen	4		4	
East Asia	She	4		4	
East Asia	Tujia	4		4	
East Asia	Tu	4		4	
East Asia	Xibo	4		4	
East Asia	Yi	4		4	
East Asia	Mongola	4		4	

East Asia	Naxi	4	4	
East Asia	Uygur	2/4	2/4	3
East Asia	Cambodian	4	4	
East Asia	Japanese	4	4	
East Asia	Yakut	4	4	

Intermediate Clusters:

I wanted to get some intuition on how the FCM was clustering in the intermediate iterations. To do this I analyzed how much movement there was between iterations. I defined movement to be the sum of all differences from the previous weight matrix to the current. For example, if an individual moved from 90/10 pop1/pop2 to 80/20 or vice versa that would give a distance of .2. The initial proportions of individuals into populations are randomly assigned and thus there are no clear clusters and thus we would expect there to be a lot movement to happen in the first step. This is consistent with figure 2. In fact because individuals are randomly assigned, the centers of the clusters are close together making everyone have a similar proportion in each cluster. Now for the next iteration if more individuals of a certain population were assigned to the same cluster they would have a shorter distance to that cluster and be assigned to that cluster with a higher proportion. In the next step since those individuals are now assigned a higher proportion, the center of their cluster will now be closer to their true population center, which will attract other individuals that share their population of origin.

As shown in figure 3, which is a zoomed in version of the first figure, around the 20^{th} iteration there is not a lot of reassignment of individuals, but then there is a stage of rapid reassignment starting at 49^{th} iteration followed by another rapid decent. To investigate what is happening I followed four individuals, two that were of European decent, one of Asian decent, and one of Oceania and their proportion of the individual in the first population at a given iteration is shown in figure 4. Figure 4 shows how the Oceania cluster was formed. At first the European cluster, which contains the Orcadian's had people in both cluster 2 (Oceania) and 5 (Europe). In figure 3 the total movement slowed around the 11^{th} iteration with just moving a select few individuals. During this time all the Oceanian's were being moved into cluster 2 while the rest of the clusters remained fairly constant. The movement is perceived as small because there are only 27

Oceanian individuals. After the Oceanian cluster is established the remaining European's are moved into their correct cluster.

This example shows that the FCM may overcome an initial incorrect guess on cluster assignment. When datasets are run multiple times on different random weight settings the FCM has shown to give similar results. While this does not imply that it is reaching the global minimum, it does show that the space may not have a lot of distributed local minima.

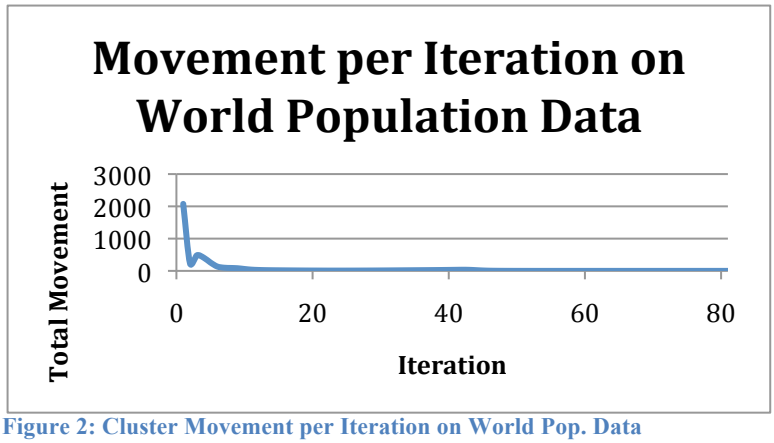

Figure 2: Cluster Movement per Iteration on World Pop. Data

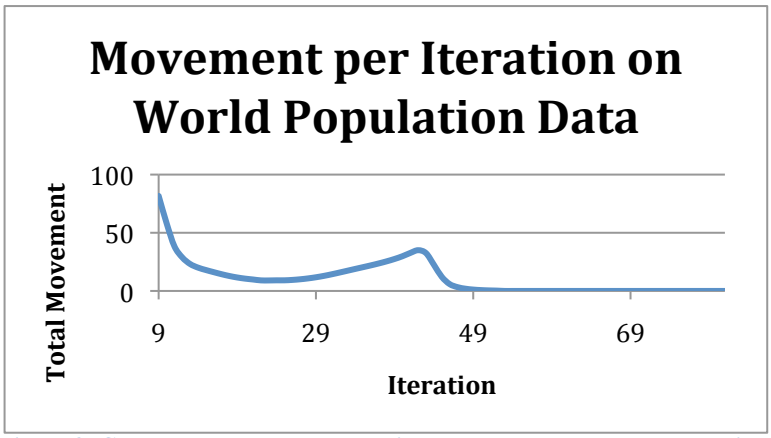

Figure 3: Cluster Movement per Iteration on World Pop. Data Zoomed in

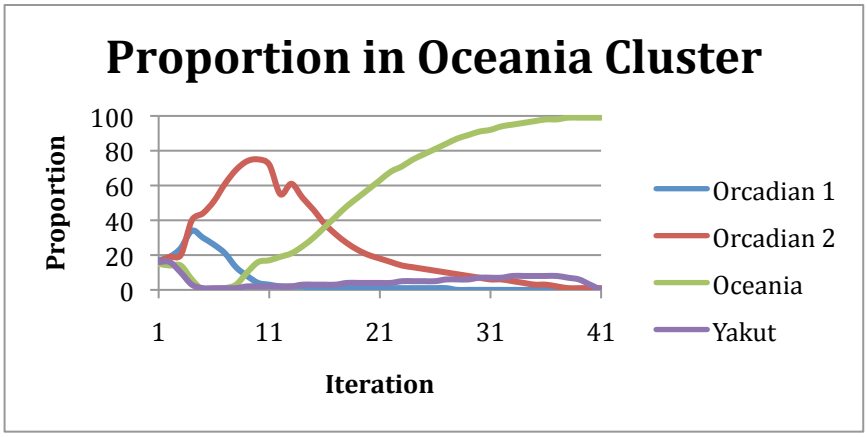

Figure 4: Percentage of Individuals in Oceania's Cluster per Iteration Step

Two Pakistani Tribes:

To test and see if the 2810 SNPs informative enough to determine population differences within the same region, I ran a test with 24 Balochi Pakistan individuals and 23 Kalash Pakistan individuals. The Balochi are Shia Muslims from Western Pakistan close to Iran, and the Kalash are non-Muslims from Chitral and thus we would not expect mating between the two groups. The FCM with m=1.05 correctly assigned both groups of individuals. However, a Balochi individual, and a Kalash individual who showed a form of admixture (60-40) were sometimes assigned to the wrong population with admixture around (40-60). It is important to remember that this does not mean that the individual of Balochi is part Kalash, but rather that he is further from the Balochi center and closer to the Kalash center than other members of his tribe. I ran the FCM 1,000 times with each trial having a different random probability matrix and found that that the Balochi individual was sorted into the Balochi population 58.8% of the time and the Kalash individual was sorted into Kalash 65.4% of the time. The 1,000 trials took less than 5 minutes to run on a standard machine. Structure with k=2, and 100,000 iterations found the admixed Baloshi individual to be 55% Baloshi and the Kalash individual to be 50% Kalash.

I realized that the inconsistency between where individuals were sorted could be because I was prematurely stopping my algorithm. I stop my algorithm when the difference between the previous and current trial is below a threshold. So I lowered the difference threshold from .1 to .005 and found that the Balochi and the Kalash individuals

were in fact sorted correctly 99.6% of the time. This gives the intuition that the admixed individuals are assigned into the correct populations in the last few iterations and the strongly correlate individuals are sorted quickly. Also the 1,000 trials show the consistency of the sorting regardless of the starting weight matrix for this data set.

Web of Eight Pakistani Tribes:

The Rosenberg data set had eight Pakistani tribes and I wanted to see which pairs of tribes I could correctly sort into distinct populations. I would expect that if two tribes were similar then I would not be able to sort individuals correctly, but if they were very distinct then I would. My results of this is in figure 5 where a blue edge means that I could not sort them and thus they were similar, a red edge means I was able to partially sort with the percentage of accurately labeled on the edge and no edge means that I could very accurately sort.

This was the first time I experienced a significantly different results on repeated trials. This could be because of the relatively small sample size of only 23 individuals per population combined with the fact that they are more genetically similar than the other tests that I have run. To compensate for the lack of a single optima, I ran each of these tests fifty times and took the trial that minimized the average distance within a cluster. When I did this I found clusterings that made the most sense from my given data showing that minimum distance is a good metric. I also found that when the fuzzy factor was raised from 1.03 to 1.05 the clusterings converged closer to optima more frequently. I believe this is happens because for a large fuzzy factor the data individuals are not as hard clustered allowing for easier movement between clusters.

To test if Structure could do any better I ran Structure for Balochi and Burusho. I found that Structure also had the two groups extremely admixed and had an accuracy of 66% same as the FCM. I also tested Balochi and Pathan with Structure and found that the clusters were indistinguishable.

Hazara and Mongolia

What I found was the Kalash and Hazara were distinct from all other tribes in Pakistan. This makes sense historically because the Kalash are a non-Islamic, isolated tribe in the northern region of Pakistan. The Hazara tribe is thought to be associated with

Asian decent, specifically Mongol decent. When I ran all 53 populations at once the Hazara came up as very strongly admixed with Asian decent. I attempted to sort the Hazara and Mongolian individuals and was unsuccessful meaning that the two tribes are very similar. For completeness, I correctly sorted Balochi and Monglia, and Burusho and Mongolia, which tell me that my association of Hazara and Mongolia is not from shared genetics from the Pakistan region.

Shia and Sunni Pakistan:

Most surprisingly I was unable to sort the predominantly Shia Balochi from the Sunni Pathan and Sindhi. My project was to develop software to detect differences that may cause false correlations in GWAS studies, so I will have to assume that because these groups are so similar, then they should not affect the results. This inference that they are similar is strengthened by the Structure tests that found the two groups to be indistinguishable.

What this graph does show is that if a GWAS was run on individuals from the Pakistan region the Kalash and the Hazara individuals should be able to be identified and removed from the study. To test if the Kalash could be identified within a mixed group I ran a test with Kalash, Makrani, Pathan, and Sindhi. With two clusters all the Kalash were together but, individuals from other populations were also sorted with the Kalash. When I increased the number of clusters to three, the Kalash were exclusively sorted into a group. This stressed the importance of being able to find metrics of how many clusters there should be, which is a target of my future work.

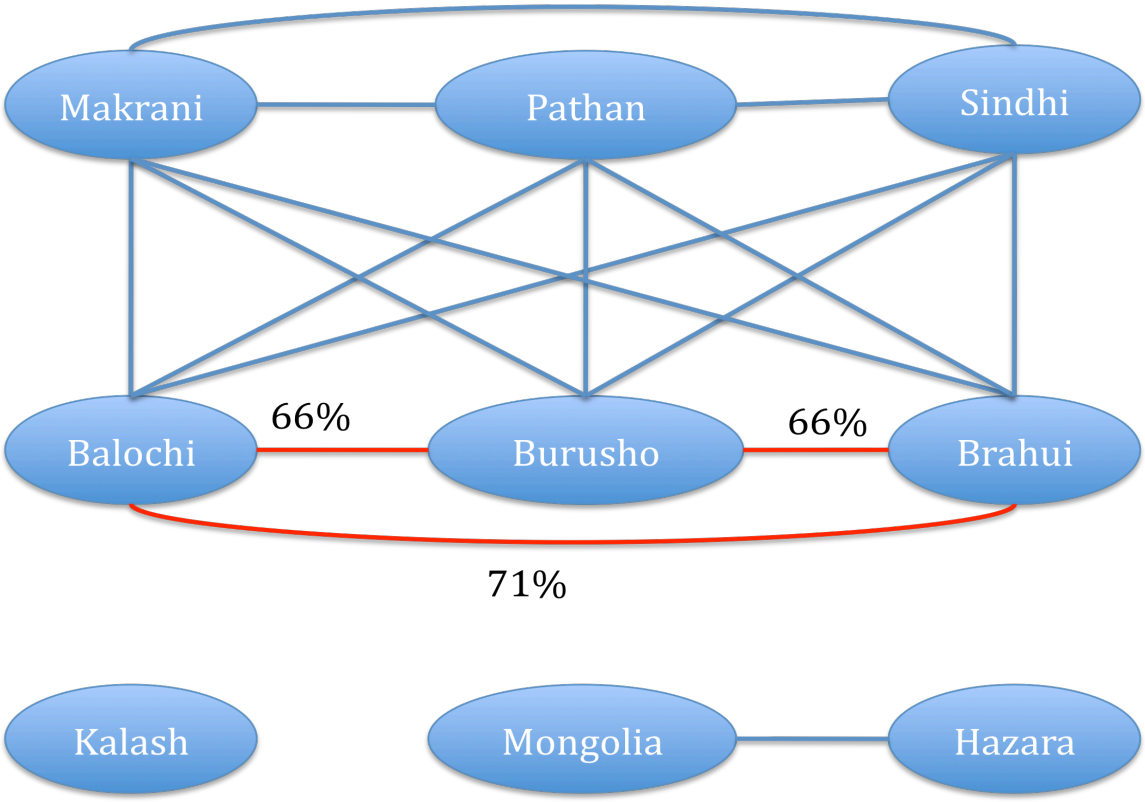

Figure 5: FCM's Ability to Distinguish Pakistani Tribes. (An arrow represents that the two could not be distinguished.)

To further show how powerful the FCM is for GWAS studies I ran the program against Maya Mexicans, and Pima Mexicans. The FCM sorted the two groups easily and consistently. This shows how large genetic differences can occur within a country that American's can easily use as samples for GWAS and how spurious that assumption is.

Choosing K:

I implemented many statistics to help me find an optimal K: Silhouette, Simplified Silhouette, Dunn's Index, Davies Bouldin, and CH Index (Bolshakova, Azuaje, & Cunningham, 2005). However, none of these metrics were consistent with my population of origin data. The reason why these metrics failed is that the average distance between an individual and the center of its cluster was usually larger than the distance between two clusters. From this we can assume that the variance within a cluster is large compared to the differences between clusters.

This large variance makes it extremely difficult to find a biologically suitable answer. Structure's authors say that the simplest algorithm to estimate the number of cluster's, "…is notoriously unstable, often having infinite variance, and thus little use in practice." They admit that their implementation uses a "dubious assumptions".

Conclusion:

I have shown that the FCM is consistent on known population data. The FCM was able to distinguish dwarf from normal Whitefish. The FCM was also able to find the same admixed Whitefish as the industry standard program Structure. I have demonstrated that the FCM is able to in a time and space efficient way to find clusterings and admixture in very high dimensional space (122,751 SNPs). And that it was also able to find similar clusterings in lower dimensional (1000 SNPs).

On the 53 World Population data set, with all 53 populations, I found regional clusterings that were historically consistent and to some degree more consistent than Structure's 2002 and 2005 clusterings. Similar historically accurate clusterings were evident in the tribes of Pakistan and two Mexican tribes. The accuracy of the tribal clusterings show the power of the FCM to detect clusters within people of similar regions that may be overlooked as a homogeneous sample in GWAS.

Future Work:

The future work on the FCM will revolve around proving the statistical validity of the clusters as I give the researcher no indication of how well the clusterings are beyond the distance within and between clusters. Structure is able to do this by assuming that the clusters are in Hardy-Weinberg equilibrium and then calculates the probability of that specific clustering given the data, and the clusters as the parameters. I could potentially take my clusters and use that as an input to Structure and compare my results based off of Structures assumptions.

Another area of work is to be able to distinguish outliers from admixed individuals. If in the two tribe Mexico example we have some admixed individuals and a single Brazilian or European individual a metric should be made to find which individuals are admixed and which do not belong in either cluster. This could be done

either by calculating the variance in the population and rejecting individuals who stray too far from the mean. Another potential approach would be to try and locate SNPs that are highly correlated with the cluster and score an individual by how many of his SNPs agree with that clusters highly correlated SNPS.

There is other published population stratification software that I was unable to compare against the FCM. PLINK (Purcell et al., 2007) by the Broad Institute uses complete linkage agglomerative clustering, based on pairwise identity-by-state (IBS) distance, but with some modifications to the clustering process. HAPAA (Sundquist, Fratkin, Do, & Batzoglou, 2008) from Stanford University uses Hidden-Markov Models to infer haplotype blocks and those haplotype blocks can then be matched to a population of origin. HapMix (Price et al., 2009) also uses haplotype blocks to detect the proportion of admixture in individuals.

Thorough reviews of all the software packages are needed to evaluate the strengths and weaknesses of each. In this review these packages should be run on GWAS data and after clustering with each software the correlated SNPs should be compared.

Final Evaluation:

In conclusion, I have shown that my clusters are fast to calculate and very accurate. The FCM scaled very well in high dimensions and was able to complete stratifications in minutes that would take Structure weeks to complete on a standard machine. When running a GWAS I suggest that both the FCM and Structure programs should be used to validate each other's results and inconsistencies across the two programs may indicate a highly heterogeneous population or that the number of clusters chosen may be inaccurate.

Table Index:

Figure Index:

Bibliography

Bezdek, J. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences, 10*(2-3), 191-203.

Bolshakova, N., Azuaje, F., & Cunningham, P. (2005). An integrated tool for microarray data clustering and cluster validity assessment. [Comparative Study

Evaluation Studies

Research Support, Non-U.S. Gov't

Validation Studies]. *Bioinformatics, 21*(4), 451-455. doi: 10.1093/bioinformatics/bti190

Bonnefond, A., Froguel, P., & Vaxillaire, M. (2010). The emerging genetics of type 2 diabetes. [Review]. *Trends Mol Med, 16*(9), 407-416. doi: 10.1016/j.molmed.2010.06.004

Campbell, D., & Bernatchez, L. (2004). Generic scan using AFLP markers as a means to assess the role of directional selection in the divergence of sympatric whitefish ecotypes. [Research Support, Non-U.S. Gov't]. *Mol Biol Evol, 21*(5), 945-956. doi: 10.1093/molbev/msh101

Carlson, C. S., Eberle, M. A., Rieder, M. J., Yi, Q., Kruglyak, L., & Nickerson, D. A. (2004). Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. [Research Support, Non-U.S. Gov't

Research Support, U.S. Gov't, P.H.S.]. *Am J Hum Genet, 74*(1), 106-120. doi: 10.1086/381000

Cavalli-Sforza, L. L., & Feldman, M. W. (2003). The application of molecular genetic approaches to the study of human evolution. [Historical Article

Review]. *Nat Genet, 33 Suppl*, 266-275. doi: 10.1038/ng1113

Chakraborty, R., & Smouse, P. E. (1988). Recombination of haplotypes leads to biased estimates of admixture proportions in human populations. [Research Support, U.S. Gov't, Non-P.H.S.

Research Support, U.S. Gov't, P.H.S.]. *Proc Natl Acad Sci U S A, 85*(9), 3071-3074.

Conrad, D. F., Jakobsson, M., Coop, G., Wen, X., Wall, J. D., Rosenberg, N. A., & Pritchard, J. K. (2006). A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. [Research Support, Non-U.S. Gov't

Research Support, U.S. Gov't, Non-P.H.S.]. *Nat Genet, 38*(11), 1251-1260. doi: 10.1038/ng1911

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1976). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, 39*(1), 1-38.

Deng, H. W. (2001). Population admixture may appear to mask, change or reverse genetic effects of genes underlying complex traits. [Research Support, Non-U.S. Gov't

Research Support, U.S. Gov't, Non-P.H.S.

Research Support, U.S. Gov't, P.H.S.]. *Genetics, 159*(3), 1319-1323.

Dunn, J. C. (1973). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. ***Cybernetics, 3*****(3), 32-57.**

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., & Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. [Research Support, N.I.H., Extramural

Research Support, Non-U.S. Gov't]. ***Nat Rev Genet, 11*****(6), 446-450. doi: 10.1038/nrg2809**

Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. [Comparative Study

Research Support, Non-U.S. Gov't

Research Support, U.S. Gov't, Non-P.H.S.

Research Support, U.S. Gov't, P.H.S.]. ***Genetics, 164*****(4), 1567-1587.**

Falush, D., Stephens, M., & Pritchard, J. K. (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. ***Mol Ecol Notes, 7*****(4), 574-578. doi: 10.1111/j.1471-8286.2007.01758.x**

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., . . . Altshuler, D. (2002). The structure of haplotype blocks in the human genome. [Research Support, Non-U.S. Gov't]. ***Science, 296*****(5576), 2225-2229. doi: 10.1126/science.1069424**

Halldorsson, B. V., Bafna, V., Lippert, R., Schwartz, R., De La Vega, F. M., Clark, A. G., & Istrail, S. (2004). Optimal haplotype block-free selection of tagging SNPs for genome-wide association studies. ***Genome Res, 14*****(8), 1633-1640. doi: 10.1101/gr.2570004**

Hubisz, M. J., Falush, D., Stephens, M., & Pritchard, J. K. (2009). Inferring weak population structure with the assistance of sample group information. ***Mol Ecol Resour, 9*****(5), 1322-1332. doi: 10.1111/j.1755-0998.2009.02591.x**

International, H. P. W. About the International HapMap Project Retrieved 4/12/2012, from http://hapmap.ncbi.nlm.nih.gov/abouthapmap.html

Istrail, S. (2000). Statistical mechanics, three-dimensionality and NP-completeness: I. Universality of intracatability for the partition function of the Ising model across non-planar surfaces (extended abstract). ***STOC '00 Proceedings of the thirty-second annual ACM symposium on Theory of computing*****.**

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., . . . Chen, Y. J. (2001). Initial sequencing and analysis of the human genome. [Research Support, Non-U.S. Gov't

Research Support, U.S. Gov't, Non-P.H.S.

Research Support, U.S. Gov't, P.H.S.]. ***Nature, 409*****(6822), 860-921. doi: 10.1038/35057062**

McClellan, J., & King, M. C. (2010). Genetic heterogeneity in human disease. ***Cell, 141*****(2), 210-217. doi: 10.1016/j.cell.2010.03.032**

Nachman, M. W., & Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. [Comparative Study

Research Support, U.S. Gov't, Non-P.H.S.]. ***Genetics, 156*****(1), 297-304.**

Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., . . . Cox, D. R. (2001). Blocks of limited haplotype diversity revealed by high-

resolution scanning of human chromosome 21. *Science, 294*(5547), 1719-1723. doi: 10.1126/science.1065573

Pemberton, T. J., Jakobsson, M., Conrad, D. F., Coop, G., Wall, J. D., Pritchard, J. K., . . . Rosenberg, N. A. (2008). Using population mixtures to optimize the utility of genomic databases: linkage disequilibrium and association study design in India. [Research Support, N.I.H., Extramural

Research Support, Non-U.S. Gov't

Research Support, U.S. Gov't, Non-P.H.S.]. *Ann Hum Genet, 72*(Pt 4), 535-546. doi: 10.1111/j.1469-1809.2008.00457.x

Pham, T. D. (2010). Brain lesion detection in MRI with fuzzy and geostatistical models. *Conf Proc IEEE Eng Med Biol Soc, 2010*, 3150-3153. doi: 10.1109/IEMBS.2010.5627188

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. [Research Support, Non-U.S. Gov't]. *Nat Genet, 38*(8), 904-909. doi: 10.1038/ng1847

Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., Ruczinski, I., . . . Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. [Evaluation Studies

Research Support, N.I.H., Extramural

Research Support, N.I.H., Intramural

Research Support, Non-U.S. Gov't]. *PLoS Genet, 5*(6), e1000519. doi: 10.1371/journal.pgen.1000519

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. [Research Support, Non-U.S. Gov't

Research Support, U.S. Gov't, P.H.S.]. *Genetics, 155*(2), 945-959.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., . . . Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. [Research Support, N.I.H., Extramural

Research Support, Non-U.S. Gov't]. *Am J Hum Genet, 81*(3), 559-575. doi: 10.1086/519795

Rosenberg, N. A., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J. K., & Feldman, M. W. (2005). Clines, clusters, and the effect of study design on the inference of human population structure. [Research Support, N.I.H., Extramural

Research Support, Non-U.S. Gov't]. *PLoS Genet, 1*(6), e70. doi: 10.1371/journal.pgen.0010070

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., & Feldman, M. W. (2002). Genetic structure of human populations. [Research Support, Non-U.S. Gov't

Research Support, U.S. Gov't, Non-P.H.S.

Research Support, U.S. Gov't, P.H.S.]. *Science, 298*(5602), 2381-2385. doi: 10.1126/science.1078311

Serre, D., Montpetit, A., Pare, G., Engert, J. C., Yusuf, S., Keavney, B., . . . Anand, S. (2008). Correction of population stratification in large multi-ethnic association studies. [Multicenter Study

Research Support, Non-U.S. Gov't]. *PLoS One, 3*(1), e1382. doi: 10.1371/journal.pone.0001382

Stephens, J. C., Schneider, J. A., Tanguay, D. A., Choi, J., Acharya, T., Stanley, S. E., . . . Vovis, G. F. (2001). Haplotype variation and linkage disequilibrium in 313 human genes. [Comparative Study]. *Science, 293*(5529), 489-493. doi: 10.1126/science.1059431

Sundquist, A., Fratkin, E., Do, C. B., & Batzoglou, S. (2008). Effect of genetic divergence in identifying ancestral origin using HAPAA. [Research Support, N.I.H., Extramural

Research Support, Non-U.S. Gov't]. *Genome Res, 18*(4), 676-682. doi: 10.1101/gr.072850.107

Tishkoff, S. A., & Verrelli, B. C. (2003). Patterns of human genetic diversity: implications for human evolutionary history and disease. [Research Support, Non-U.S. Gov't

Research Support, U.S. Gov't, Non-P.H.S.

Review]. *Annu Rev Genomics Hum Genet, 4*, 293-340. doi: 10.1146/annurev.genom.4.070802.110226

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., . . . Zhu, X. (2001). The sequence of the human genome. [Research Support, Non-U.S. Gov't]. *Science, 291*(5507), 1304-1351. doi: 10.1126/science.1058040

Weiss, K. M., & Clark, A. G. (2002). Linkage disequilibrium and the mapping of complex human traits. [Research Support, U.S. Gov't, P.H.S.]. *Trends Genet, 18*(1), 19-24.

Wu, J. (1983). ON THE CONVERGENCE PROPERTIES OF THE EM ALGORITHM. *The Annals of Statistics, 11*(1), 95-103.

Zhang, M., Adamu, B., Lin, C. C., & Yang, P. (2011). Gene expression analysis with integrated fuzzy C-means and pathway analysis. *Conf Proc IEEE Eng Med Biol Soc, 2011*, 936-939. doi: 10.1109/IEMBS.2011.6090211