BIOINFORMATICS

Vol. 01 no. 01 2016
Pages 1–6

Ancestry Mapping using PPCA & Variational Inference

Akshay Shinde[1] and Ronak Sumbaly [1*]

[1]Department of Computer Science, University of California, Los Angeles

Associate Editor: Sriram Sankararaman

ABSTRACT

Motivation: Inferring ancestry has long been recognized as a confounding factor in genetic testing. Estimation of individual ancestry from genetic data is useful in various applications such as during analysis of disease association studies, understanding human population history, interpreting personal genomic variation and also as covariates to correct for population stratification. The project from the human genetics community point of view focuses on determining and evaluating new methods of Variational Inference and Probabilistic Principal Components towards the estimation and visualization of admixture ancestral origin for an individual.

Results: We have tried to validate our framework in European, American, Asian, African populations, making use of SNPs derived from 1000 Genome project. We have compared and evaluated our approach visually with other existing techniques which performs analysis without a reference panel. We finally show how the entire framework of inferring ancestry using our workflow would lead to more effective results.

Availability: Existing development and source code in R is available for contribution and for download by the public from GitHub (www.github.com/RonakSumbaly/Ancestry-Mapping).

Contact: rsumbaly@ucla.edu

1 INTRODUCTION

Admixed individuals have genomes that can be described as arrangement of alternating segments of different ancestries. The length and origin of each of these segments basically reflects the admixture history of each individual. More importantly, these boundaries and origin of each segment can be estimated via statistical methods which examine the distribution of genetic variants along each chromosome and also the difference of allele and haplotype frequencies between ancestral populations. The relative effects of different population genetic forces on the variation of human genomic is a central focus of population genomics.

Since there are large numbers of markers they will nearly always be able to provide accurate discrimination for at least 6 or 7 geographic regions (Kenneth *et al*., 2014). Estimation of individual ancestry from genetic markers like Single Nucleotide Polymorphism (SNPs) is helpful in diverse situations, apart from performing it on individual basis, in an attempt to determine ancestral origins, it is also performed to infer bio-geographical origins or admixtures of populations for research purposes. It is a primary consideration in studies including admixture and association mapping, forensics, prediction of medical risks, wildlife management, and studies of dispersal, gene-flow, and evolutionary history (Shriver *et al*., 1997). Hence accurate inference and understanding of ancestry from genetic variation data is very critical.

Ancestry estimation on a high level is distinguished into estimation of global ancestry and local ancestry. The local ancestry paradigm (Sankararaman *et al*., 2008) imagines the human genomes to be divided into smaller chromosome segments each with its definite ancestral origin. The goal is to find these segments and to estimate its origin. This project on the other hand deals with global ancestry (J.K Pritchard, 2000). The global ancestry paradigm is concerned with estimating the proportion of ancestry from each contributing population. Literature provides two broad global ancestry mapping approaches which include: Model Based Ancestry Mapping and Algorithmic Ancestry Mapping.

Model based ancestry mapping include approaches that are based on the idea of estimating ancestry coefficients as parameters of a statistical model and evaluating the data based on these coefficients. Popular examples of these approaches include STRUCTURE (J.K Pritchard, 2000), FRAPPE (Tang *et al*., 2003) and ADMIXTURE (Alexander *et al*., 2009). Algorithmic ancestry mapping on the other hand mostly employs usage of cluster-based analysis and principal component analysis (Price A *et al*., 2006), to discover the structure within the data. The cluster-based analysis are further broadly classified as: Distance-based methods and Model-based methods. While the former works by calculating a pairwise distance matrix, the latter assumes the observation from each cluster are randomly drawn from some parametric model. The clusters are later used as markers for different ancestral clusters while PCA projections are used to explain variations between populations.

Both approaches of global ancestry have their respective caveats. STRUCTURE (J.K Pritchard, 2000) uses Markov Chain Monte Carlo (MCMC) that can analyze admixed individuals without requiring ancestral representatives, but its complex model for uncertainty, combined with the use of MCMC, makes it highly computationally intensive and thus less practical for large-scale analysis. On the other hand PCA methods (Patterson N *et al*., 2006) in large structured samples, particularly when sub-populations are closely related have difficulty assigning individuals to the correct sub-population (Intarapanich A *et al*., 2009). PCA also has a non-parametric approach which does not consider the distribution of the observed data. In order to tackle these issues two alternative approaches of Probabilistic PCA and Variational Inference have been constructed and evaluated in this project.

*to whom correspondence should be addressed.

2 APPROACH

A variety of methods have been proposed for the identification of genetic ancestry differences among individuals in a sample using genetic data. Instead of providing a single approach we purpose a framework that could be used for ancestry inference estimation. The framework entails the following techniques.

2.1 Linkage Disequilibrium (LD) Pruning

Studies have shown that, in the presence of LD, the prediction accuracy of GWAS falls short. In order to avoid any spurious results in ancestry estimation one possible solution to this problem which we perform is LD pruning which involves removing highly-correlated SNPs from the genetic data and use the pruned data as input data for ancestry estimation.

2.2 Probabilistic Principal Component Analysis

GWAS routinely applies PCA to infer population structure within a sample to correct for confounding due to ancestry (Patterson N *et al*., 2006). GWAS implementation of PCA uses tens of thousands of SNPs to infer structure, despite the fact that only a small fraction of such SNPs provides useful information on ancestry. Even after performing LD-pruning we only have a small fraction of SNPs that form a reduced set of ancestry-informative markers (AIMs) for any practical usage.

PCA can be used for dimensionality reduction to group those with similar genetic ancestry together. Construction of the principal components is straightforward and computationally efficient and the top components with the largest eigenvalues reliably capture differences in genetic variation due to sub-population status. Briefly, the PCA methods focus on the spectral decomposition of a variance covariance matrix for dimensionality reduction. But in large, highly structured samples, particularly when sub-populations are closely related or when there is a genetically distant sub-population, traditional PCA methods have difficulty assigning individuals to the correct sub-population along with it being highly computational due to the calculation of the variance covariance matrix.

In this project, we apply a different statistical method for inferring ancestry that avoids the painstaking computational complexity of PCA and further assumes an underlying probabilistic model of the genetic data. We purpose application of Probabilistic Principal Component Analysis (PPCA) (Tipping, M.E *et al*.) for the problem of ancestry inference. Unlike standard PCA methods that yields a multitude of SNPs with non-zero contributions to the principal components, PPCA method produces a limited number of influential SNPs for ancestry inferencing, with no loss in visualization of ancestry. The process is easy to implement and comprehensive in that if can allow for all SNP data to be processed.

2.3 Variational Inference

After obtaining AIMs the next task in hand is to estimate the proportions of ancestry for an individual in the database. As indicated before there are basically two broad methods of doing this task: Model based method and Algorithmic based method. For the purpose of this project we choose a model based method for ancestry mapping. Model based methods assume that observations from each population are random draws from some parametric model. Inference for parameters corresponding to each population is then done jointly with inference for the cluster membership of each individual, using standard statistical methods like maximum-likelihood or Bayesian methods.

Consider a simple admixture model in the Fig. 1 Let X denote the genotypes of the individuals, Z^a, Z^b denote the population assignment of of two copies of a locus and P denote the allele frequency at L loci. Let Q denote admixture proportion of each individual. We are interested in inferring the posterior $P(Z^a, Z^b, P, Q \mid X)$. We can take a bayesian approach to estimate global ancestry by sampling from the posterior distribution over global ancestry parameters using a Gibbs sampler (J.K Pritchard, 2000) that appropriately accounts for the conditional independence relationships between latent variables and model parameters.

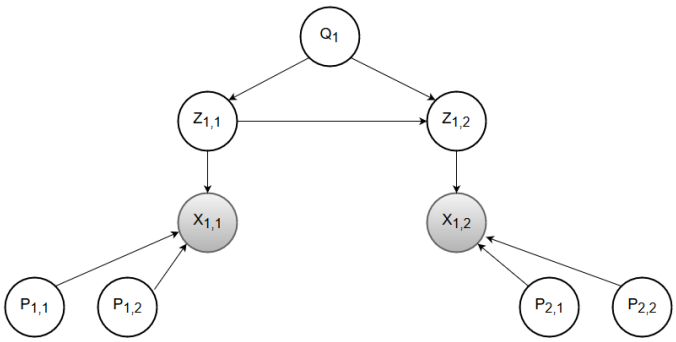

Fig. 1. Admixture Model

However, even well-designed sampling schemes need to generate a large number of posterior samples to resolve convergence and mixing issues and yield accurate estimates of ancestry proportions, greatly increasing the time complexity of inference for large genotype data sets. To provide faster estimation, FRAPPE (Tang *et al*., 2003) and ADMIXTURE (Alexander *et al*., 2009) both use a maximum-likelihood approach. More recently, (Frichot E. *et al*., 2014) described fast computational algorithms for ancestry inference that make use of sparse non-negative matrix factorization (sNMF) and least squares optimization to produce estimates of ancestry proportions. Following in the foot-steps of these approaches we purpose a Variational Bayesian inference framework to speed up the inference of population structure.

Here, we propose a novel framework that addresses the above challenges of estimating individual ancestry. We also compare our results with existing software/techniques provided above.

3 METHODS

In this project, we consider the inference of ancestry in an admixed population. Admixed populations arise from K ancestral populations $A_1, \ldots, A_K$ that have been mixing for g generations. For the purposes of simplification we focus on the analysis of SNP data that we obtained from 1000 Genome Project and not via simulation. For a given set of genotypes from the admixed population, we describe each individual genotype as a

vector g_i, where $g_{ij} \in \{0, 1, 2\}$ is the minor allele count of individual i at position j. At position j, the two alleles of individual i have descended from one or two of the K ancestral populations. We are interested in estimating these ancestral population(s) for each individual i.

Mathematically, we model the recently admixed populations as a set of K independent populations that have come together at some point in history and have been mixing (through random mating) for g generations. Our framework first performs LD pruning and then constructs a coordinate system by applying PPCA to SNP genotypes of the individuals, and then uses Variational Inference to identify their ancestral proportions which can be directly used to correct for population structure in association studies or for the various applications which were indicated before. The framework applied to the project is presented in the following subsections.

3.1 LD Pruning

The Admixture model described in Figure 1 assumes that genetic marker are in linkage equilibrium, so that aligned bases are independent given the underlying ancestry information. But often this assumption does not hold and directly applying the model could lead to many false ancestry switches. This is also a known problem for analyzing high-density genotype data.

To resolve the issue we applied a linkage disequilibrium pruning algorithm that aims to retain as much sequence information as possible while excluding markers in linkage disequilibrium with each other. We first estimated the squared correlation (r^2) between marker pairs for the population by dividing the genetic data into windows of size 500; for each window the marker pairs which recorded a r^2 value exceed a cutoff (0.7 was used in our project), we exclude those markers in our model for ancestry inference estimation. This prevents high-LD blocks from having excessive influence on the inferred ancestry of a region, while retaining a dense, informative set of SNPs.

3.2 PPCA Framework

Probabilistic PCA (PPCA) is a probabilistic formulation of PCA based on a Gaussian latent variable model and was first introduced by (Tipping, M.E *et al.*). The estimation problem was modeled using probabilistic PCA. On a high-level the PPCA model reduces the dimension of our high-dimensional genetic data by relating a p-dimensional observed data point to a corresponding q-dimensional latent variable through a linear transformation function, where $q << p$. In the PPCA model we assume there is an (unknown) affine space in higher dimension on which hidden (latent) points are generated at random according to a Gaussian distribution. The observed point which in our case is the genotype data are then generated from these latent points by addition of an isotropic Gaussian noise.

We follow the formal notation that was in (Tipping, M.E *et al.*) paper. The observed random variables t (SNP data) is given by where x are the latent (hidden) variables, W is a matrix describing a subspace and Wx are the latent points on an affine subspace. Finally ϵ is an error term, given by a Gaussian random variable with mean 0 and covariance matrix $\sigma^2 I$.

$$t_n = Wx_n + \epsilon_n \qquad \epsilon_n \sim \mathcal{N}(0,\ \sigma^2 I) \qquad x_n \sim \mathcal{N}(0,\ I) \tag{1}$$

Equivalent, our observed random variables are themselves Gaussian and are described as,

$$t_n = \mathcal{N}(\mu,\ WW^T + \sigma^2 I) \tag{2}$$

The parameters W, μ, σ^2 of the model can be estimated using maximum likelihood. Maximizing the (log) likelihood function with respect to model parameters is non-trivial; in (Tipping, M.E *et al.*) it is demonstrated that the estimates do however have closed form solutions given by,

$$\mathcal{LL} = -\frac{N}{2}\{p\ ln(2\pi) + ln|WW^T + \sigma^2 I| + tr((WW^T + \sigma^2 I)^{-1} S)\}$$
$$S = \frac{1}{N}\sum_{n=1}^{N}(t_n - \mu)(t_n - \mu)^T \tag{3}$$

Fig. 2 represents the graphical model of the PPCA model.

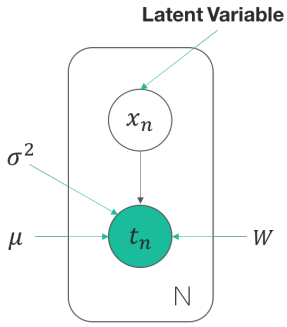

Fig. 2. Probabilistic PCA Graphical Model

The maximum likelihood estimate (MLE) of the weight matrix $\hat{W}$ in PPCA corresponds exactly to the weight matrix in conventional PCA. Hence the model output in PPCA is exactly that obtained in conventional PCA which furthers our purpose, but also with this we get additional advantages of uncertainty assessment and potential model extensions.

Implementation

The EM algorithm is used to compute the MLE's in the probabilistic model given above. The algorithm alternates between two steps until convergence: the expectation (E) step and the maximization (M) step. In the E-step, the expected values of the latent variables are estimated given the observed genetic data and the current estimates of the model parameters. In the M-step, the model parameters are re-estimated by maximizing the log likelihood function using the expected values of the latent variables derived in the previous E-step. The two steps are repeated until convergence. The weights matrix W after maximizing basically gives us the principal components of the data which is given by,

$$W_{ML} = U_q \sqrt[2]{(\lambda_q - \sigma^2 I)R} \tag{4}$$

where U are the principal components, λ are the diagonal eigenvalues and R is the orthogonal rotation matrix. The model was written in R statistical software.

3.3 Variational Inference Framework

Variational Bayesian Inference aims to redefine the problem of inference as an optimization problem rather than a sampling problem. Variational Bayesian techniques approximate the log-marginal likelihood of the data by proposing a family of tractable parametric posterior distributions over hidden variables in the model, the goal is then to find the optimal member of this family that best approximates the marginal likelihood of the data. Thus, a single optimization problem gives us both approximate analytical forms for the posterior distributions over unknown variables and an approximate estimate of the intractable marginal likelihood.

The central aim is to find an element of a tractable family of probability distributions, called variational distributions, that is closest to the true intractable posterior distribution of interest. A natural choice of distance on

probability spaces is the Kullback-Leibler (KL) divergence, defined for a pair of probability distributions q(x) and p(x) as:

$$D_{kl}(q(x) \,||\, p(x)) = \int q(x)\, log\, \frac{q(x)}{p(x)} dx \tag{5}$$

Given the asymmetry of the KL divergence, VB inference chooses $p(x)$ to be the intractable posterior and $q(x)$ to be the variational distribution; this choice allows us to compute expectations with respect to the tractable variational distribution. An approximation to the true intractable posterior distribution can be computed by minimizing the KL divergence between the true posterior and variational distribution. We will restrict our optimization over a variational family that explicitly assumes independence between the latent variables Z and parameters (P, Q); this restriction to a space of fully factorizable distributions is commonly called the mean field approximation.

KL divergence quantifies the tightness of a lower bound to the log-marginal likelihood of the data. Specifically, for any variational distribution $q(Z, P, Q)$, we have

$$\begin{aligned} log\ P(G \mid K) = \mathcal{E}[q(Z^a, Z^b, Q, P)] + \\ D_{kl}(q(Z^a, Z^b, Q, P) \,||\, p(Z^a, Z^b, Q, P \mid G)) \end{aligned} \tag{6}$$

where $\mathcal{E}$ is a lower bound to the log-marginal likelihood of the data, log $p(G \mid K)$. Thus, minimizing the KL divergence is equivalent to maximizing the log-marginal likelihood lower bound (LLBO) of the data.

Implementation

We have implemented VB inference using Accelerated Variational inference (Anil *et al.*, 2014). It treats the iterative update equations for the set of variational parameters as a deterministic map, a globally convergent algorithm with improved convergence rates can be derived by adapting the Cauchy-Barzilai-Borwein method for accelerating the convergence of linear fixed-point problems (Raydan and Svaiter 2002) to the nonlinear fixed-point problem given by our deterministic map. We have chosen a symmetric Dirichlet prior over admixture proportions, $p(Q_n) = Dirichlet(\frac{1}{K}\mathbf{1}_K)$. For population-specific allele frequencies at each locus, we have chose a flat beta-prior, $p(P_{ij}) = Beta(1, 1)$.

4 DATA

The project uses 1000 Genomes Project Phase 3 variants (1000 Genomes Project Consortium, 2015) from four different ancestral cohorts. Super populations are defined using the 1000G panel file. Only the SNPs located on Chromosome - 22 are considered. The database comprises of 1092 individual and ≈ 400K SNPs.

4.1 Database Summary

The below table summarizes the number of individuals in the database:

Super Population	Number of Individuals in Database
European (EUR)	379
African (AFR)	246
American (AMR)	181
Asian (ASR)	286

5 DISCUSSION

By using the 1000 Genomes Project genetic data for chromosome 22, we applied our LD pruning algorithm to remove sites that are effectively in linkage equilibrium. The process continues until no marker pairs in linkage disequilibrium remain in the list of sites. Note that this pruning algorithm is applied to the entire population over a window size and not each sequenced subject independently. Fig. 3 shows a section of the genetic data's LD value. The darker shade of red represents SNPs that are in LD and to be removed.

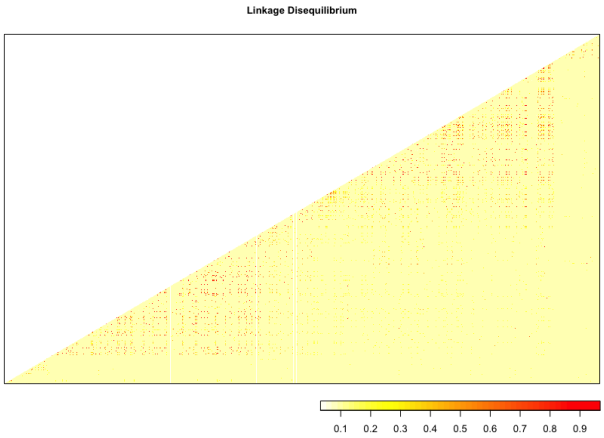

Fig. 3. LD Pruning - Section of Chromosome 22

After removing SNPs in LD we were left with ≈ $225K$ SNPs for consideration of ancestry inference. The next step of the framework involved application of PCA and PPCA to the pruned data.

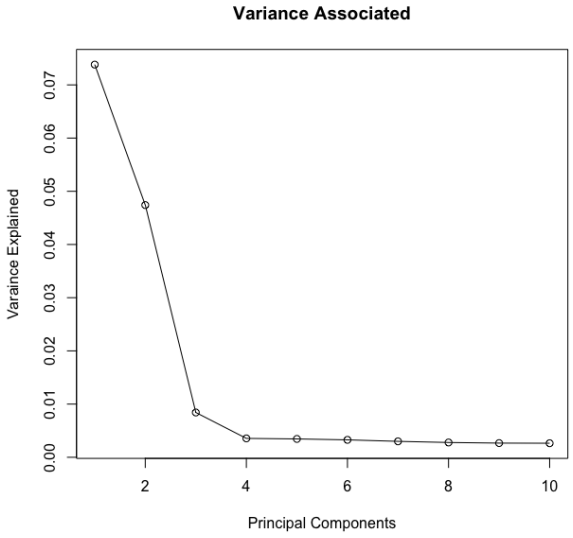

Fig. 4. LD Pruning - Section of Chromosome 22

In order to validate the difference between the PCA and PPCA approach we implemented both the techniques. flashPCA (Inouye M. *et al.*, 2014) was used to obtain the principal components of genetic data. By observing the variance between the principal components as shown in Fig. 4 we restricted the analysis to the top 3 principal components.

The PCA and PPCA results of analysis are plotted on a two-dimensional graphs, individuals of similar ancestry are shown with same color and are seen to cluster together. Fig. 5 represents PCA results.

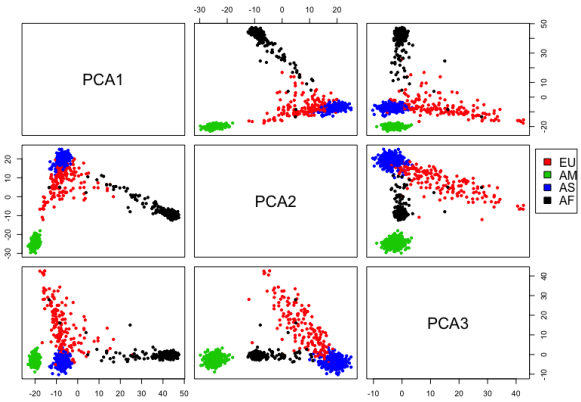

Fig. 5. PCA - Principal Components

Fig. 6 represents PPCA visualized results of the principal components.

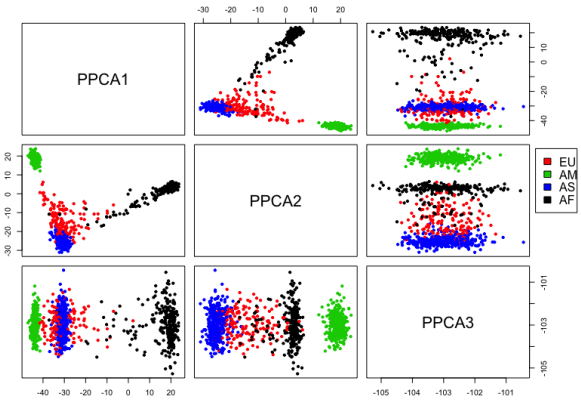

Fig. 6. PPCA - Principal Components

Looking at the results we can see that PCA and PPCA innately provide similar results for clustering of the individuals. European and Asian populations are seen to have an admixture property and a dispersion is seen in the African population towards the European and Asian population which might give evidence to the ”Single point origin” theory of modern humans and well as signify African-American populations.

Hence we can conclude that Probabilistic PCA leads to negligible loss of ancestry information compared to traditional PCA analysis of genome-wide SNP data and at the same time provide for a more less computational and error free model for ancestry inference. The next step of the work-flow involved using this low dimensional data from PPCA and applying Variational Inference to estimate ancestry proportions. Due to computational problems we limited our dataset to 100 SNPs rather than the whole data.

Similar to applying PCA as our reference we have used sNMF software to compare the results of our variational inference model. sNMF uses non-negative matrix factorization to infer the population structure. The admixture proportions from sNMF software are plotted in Fig. 7.

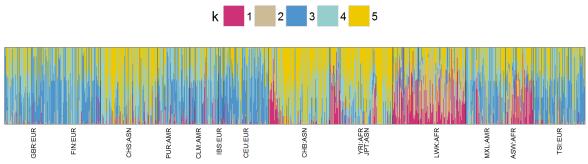

Fig. 7. Ancestry Proportions estimated by sNMF

Similarly we applied our variational inference to our data to obtain results shown in Fig. 8.

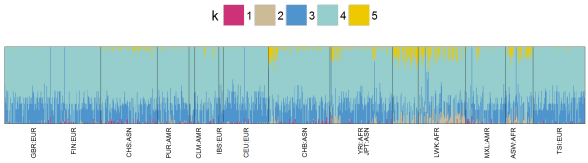

Fig. 8. Ancestry proportion estimated by Variational Inference

Looking at the ancestry proportion estimated by sNMF, we can tell that it successfully identifies 4 underlying populations whereas our model primarily identifies only 3 populations. Reason for this could be that the 100 SNPs chosen do not capture the necessary information required by the variational inference algorithm. Also due to limitation of computing resources, we had to limit the number of iterations to 200. Fig. 9 shows marginal likelihood at the end of every 20 iterations.

6 CONCLUSION

The project contributes to the field by developing a workflow, to efficiently infer ancestry in admixed individuals based on the three steps that were presented above. The major strength of the workflow

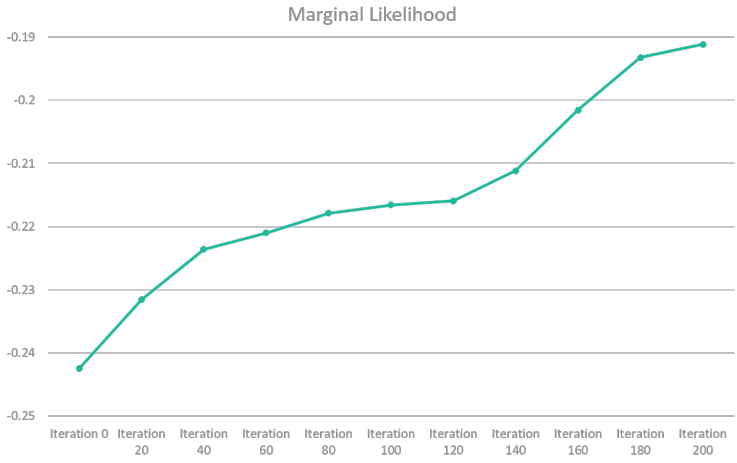

Fig. 9. Marginal likelihood versus number of iterations

is that it strengthens the assumption of linkage disequilibrium and only consider uncorrelated loci to increase the power of inference and at the same time takes into consideration the underlying probabilistic model the genetic data. While there is still work needed in optimization of our variational inference model as well a stringent LD pruning algorithm we believe that the framework does look promising.

ACKNOWLEDGEMENT

We would like to express our deep gratitude and thank Prof. Sriram Sankararaman for his support and assistance in this project and for providing us with the necessary guidance throughout the quarter.

REFERENCES

A global reference for human genetic variation, The 1000 Genomes Project Consortium, *Nature 526*, **68-74** (01 October 2015).

Abraham G, Inouye M (2014) Fast Principal Component Analysis of Large-Scale Genome-Wide Data. *PLoS ONE* **9(4): e93766.**

Alexander, D. H., J. Novembre, and K. Lange (2009) Fast model based estimation of ancestry in unrelated individuals, *Genome Res*, **19(9):** 1655-1664.

Anil Raj, Matthew Stephens, and Jonathan K. Pritchard (2014) fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets, *(Genetics)*, **197:** 573-589.

Frichot E., Mathieu F., Trouillon T., Bouchard G., Franois O., 2014. Fast inference of admixture coefficients using sparse non-negative matrix factorization algorithms. *Genetics* **196: 973983**.

Intarapanich A, Shaw PJ, Assawamakin A, et al (2009). Iterative pruning PCA improves resolution of highly structured populations. *BMC Bioinforma* **10:382.**

J K Pritchard, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, **155(2)**, 945.

K.K. Kidd, W.C. Speed, A.J. Pakstis, M.R. Furtado, R. Fang, A. Madbouly, et al. Progress toward an efficient panel of SNPs for ancestry inference *Forensic Sci. Int. Genet.*, **10 (2014), pp. 2332**

MacHugh D.E., Shriver M.D. Microsatellite DNA variation and the evolution domestication and phylogeography of taurine and zebu cattle (Bos taurus and Bos indicus) *Genetics*. **1997;146:** 10711086.

Patterson N, Price A, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genetics* **2: e190.**

Price A, Patterson N, Plenge R, Weinblatt M, Shadick N, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*. **38: 904909**.

Sankararaman S., Sridhar S., Kimmel G., Halperin E., 2008b Estimating local ancestry in ad-mixed populations. Am. J. Hum. *Genet. 82:* 290303.

Tang, H., J. Peng, P. Wang, and N. J. Risch (2005) Estimation of individual admixture: analytical and study design considerations.(Genet. Epidemiol), **28(4):** 289301.

Tipping, M. E., & Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B,* **61(3), 611622.**