# Dirichlet Process based model for GWAS

Govind Kothari, Avinash Das

May 16, 2012

**Abstract**

Human population have large genetic variation. Genetic variations are usually characterized by Single Nucleotide Polymorphism (SNP). These are the location of variation in the genome sequence of two individuals. Identification of SNPs affecting human phenotype, especially leading to risks of complex disorders, is one of the key problems of medical genetics. In this project we build a probabilistic model which can determine the deleterious mutation which can lead to heart disease in humans.

## 1 Introduction

The process of finding alleles, i.e. SNPs at different location, of an individual with respect to standard human genome is called genotyping. Complex organisms like human are bi-allelic (paternal and maternal). The genotyping technology does not give phase information of SNP, i.e whether it corresponds to maternal or paternal haplotype. The process of inferring phase of SNP is called haplotype inference. In genetic epidemiology, a genome-wide association study (GWA study, or GWAS), also known as whole genome association study (WGA study, or WGAS), is an examination of many common genetic variants in different individuals to see if any variant is associated with a trait. GWAS typically focus on associations between single-nucleotide polymorphisms (SNPs) and traits like major diseases. Most of the current approaches involves case-control model which compares two large set of individuals belonging to healthy control group and diseased cases group. The individuals are genotyped for commonly know SNPs. A differential analysis is done between control and cases to find the variation in allele frequency. Our model is a bayesian approach.

We have access to genotype data for around 313 failing heart and control cases. Our goal is to analyze the data to find the genetic variation that causes the heart failure. Our aim is to simultaneously infer and classify haplotypes for heart failure cases and control cases from genotype data. In addition, we also want to capture the genetic variation in these two classes by finding the ancestral haplotypes that generate founder population for each class. Previous work [Xing et al., 2004] [Sohn and Xing, 2009] [Xing et al., 2007] has used HDP model [Teh et al., 2006] for inferring the

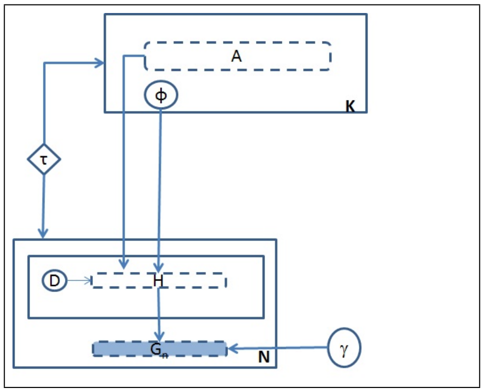

Figure 1: Model for Healthy

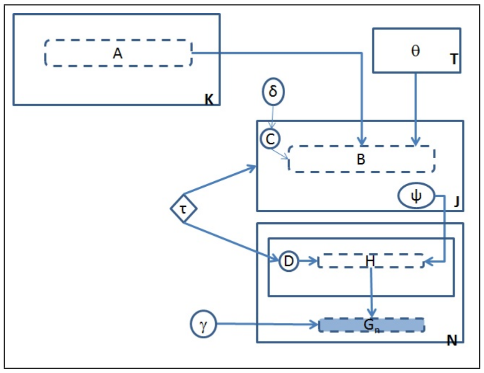

Figure 2: Model for Bad

haplotypes for multi-population. Their generative model defines a set of common ancestor haplotypes that generates founder haplotypes for each population. They then sample paternal and maternal haplotypes from ancestral haplotype that generates observed genotype.We use Dirichlet process [Teh, 2010] based model determining mutations.

In our generative model, we assume that healthy haplotypes (H) are generated by Ancestral population (A) fig 3, while diseased ancestral haplotypes (B) are produced by certain mutation events ($\phi$) in ancestral haplotypes. These diseased haplotypes in process generates the paternal and maternal haplotypes for a given observed diseased genotype (shaded). Each genotypes is generated by paternal and maternal haplotypes. Our model will be able to answer following queries: Given a genotype what is probability that it is coming from the diseased or healthy population? What is variation of a disease haplotype? What are specific mutation that causes a healthy haplotype to be converted into a disease haplotype?

# 2 Statistical Model

At first we give brief introduction to Dirichlet Process [Ferguson, 1973] and Dirichlet Distribution.

## 2.1 Dirichlet Procee Mixture Model

Consider an urn that at the outset contains a ball of a single color. At each step we either draw a ball from the urn and replace it with two balls of the same color, or we are given a ball of a new color which we place in the urn. One can see that such a scheme leads to a partition of the balls according to their color. Mapping each ball to a haploid individual and each color to a possible haplotype, this partition is equivalent to the one resulted from the coalescence-with-IMA process [Hoppe, 1984], and the probability distribution of the resulting allele spectrum - the numbers of colors (i.e., haplotypes) with every possible number of representative balls (i.e., decedents)is captured by the well known Ewens sampling formula [Ewens and Tavaré, 1998].

Letting parameter $\alpha$ define the probabilities of the two types of draws in the above mentioned Polya urn scheme, and viewing each (distinct) color as a sample from $Q_0$, and each ball as a sample from Q, Blackwell and MacQueen24 showed that this Polya urn model yields samples whose distributions are those of $Q_0$ the marginal probabilities under the *Dirichlet process.* Formally, a random probability measure Q is generated by a DP if for any measurable partition $_1, ..., A_k$ of the sample space (e.g., the partition of an unbounded haploid population according to common haplotype patterns), the vector of random probabilities $Q(A_i)$ follows a Dirichlet distribution $(Q(A_1), ..., Q(A_k)) \sim \mathtt{Dir}(\alpha Q_0(A_1), ..., \alpha Q_0(A_k))$, where $\alpha$ denotes a *scaling parameter* and $Q_0$ denotes a base measure. The Polya urn construction of DP makes explicit an order-independent sequential sampling scheme to draw samples from a DP. Specifically, having observed n samples with values $(\phi_1, ..., \phi_n)$ from a Dirichlet process

$\mathsf{DP}(\alpha, Q_0)$, the distribution of the value of the (n + 1)th sample is given by:

$$\phi_{n+1}|\phi_1, ...., \phi_n.\alpha, Q_0 \sim \sum_{k=1}^{K} \frac{n_k}{n_k + 1} \delta_{\phi_k^*}(.) + \frac{\alpha}{n+\alpha} Q_0(.)$$

where $\delta_{\phi_k^*}()$ denotes a point mass at a unique value $\delta_{\phi_k^*}$, $n_k$ denotes the number of samples with value $\delta_{\phi_k^*}$, and $K$ denotes the number of unique values in the $n$ samples drawn so far. This conditional distribution is useful for implementing Monte Carlo algorithms for haplotype inference under DP-based models.

## 2.2 Our Model

We have designed a non parametric Bayesian generative model built on Dirichlet Process which has a well formed statistical framework to handle our problem. In our generative process we assume that given healthy haplotypes, the Ancestral haplotypes are independent of disease haplotypes. Thus inferencing of our model is divided into two parts as shown in Figure 1 and 2. The generative process in Figure 1 is, given an ancestral haplotype and mutation rate $\Phi$, a healthy haplotype is generated. The observed Genotype of healthy individual is generated from two healthy haplotype with $\gamma$ being the observational noise.

Once we obtain the Ancestral population from the above process, we generate the diseased ancestral haplotype population and corresponding individuals as shown in Figure 2. Given a healthy ancestor A and unhealthy mutation rate $\theta$ a bad ancestor B shown in Figure 2 is generated. Each diseased individual haplotype H is generated from bad ancestor with mutation rate $\Psi$. Two bad ancestors (representing paternal and maternal haplotypes) generate individual genotype. Each individual genotype is observed with observational noise rate $\gamma$.

Our implementation builds upon and significantly changes the code used in paper [Xing et al., 2004].

The basic generative structure of the model is as follows

$$\begin{aligned}
\gamma &\sim \Gamma(\alpha_1, \beta_1) \\
\Psi &\sim \Gamma(\alpha_2, \beta_2) \\
\Theta &\sim \Gamma(\alpha_3, \beta_3) \\
\delta &\sim Dir(\alpha, K) \\
C &\sim Categorical(\delta) \\
b_j|A, C, \Theta, H, \tau &\sim P(.|A, C, \Theta, H, \tau) \\
G &\sim DP(\alpha, G_0) \\
\kappa &\sim G \\
D &\sim Categorical(\kappa) \\
h_{i_e}|D, \Psi &\sim P_h(.|D, \Psi) \\
g_i|h_{i_0}, h_{i_1} &\sim P_g(.|h_{i_0}, h_{i_1}, \gamma)
\end{aligned}$$

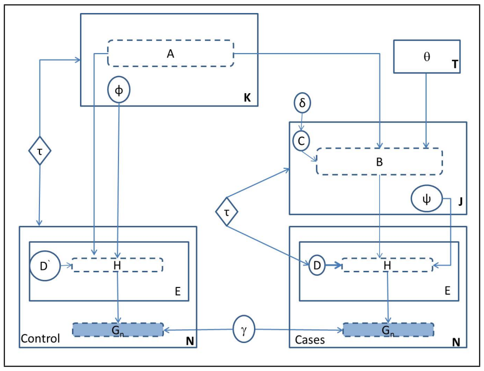

Figure 3: Model

The probability distribution for each of the random variables in our model can be described in the following set of equation.

## 2.3 double locus mutation model

$P(b_t|a_t, \theta_{kt}) = \theta_{kt}^{I(b_t=a_t)}(\frac{1-\theta_t}{B-1})^{I(b_t \neq a_t)}$where, B = num of alleles

The joint conditional distribution of bad haplotype instances $\mathbf{b} = \{\mathbf{b_j}, \mathbf{j} \in \{\mathbf{1}, ..\mathbf{J}\}\}$ and parameter instances $\mathbf{\Theta} = \{\theta_{kt} : k \leq K, t \leq T\}$, given the ancestor equivalence class indicator $\mathbf{c}$ of haplotype instances and the set of ancestors $\mathbf{a} = a_1, ..., a_k$, can be written as:

$$\begin{aligned} P(\mathbf{b}|\mathbf{c}, \mathbf{a}, \Theta) &\propto \prod_k \prod_t \prod_j \theta_{kt}^{I(b_{jt}=a_{kt})I(c_j=k)}(1-\theta_{kt})^{I(b_{jt}\neq a_{kt})I(c_j=k)} \\ &= \prod_k \prod_t \theta_{kt}^{\sum_j I(b_{jt}=a_{kt})I(c_j=k)}(1-\theta_{kt})^{\sum_j I(b_{jt}\neq a_{kt})I(c_j=k)} \\ &= \prod_k \prod_t \theta_{kt}^{m_{kt}}(1-\theta_{kt})^{m'_{kt}} \end{aligned}$$

Where $m_{kt} = \sum_j I(b_{jt} = a_{kt})I(c_j = k)$, is the sufficient statistics of $\theta_{tk}$.

The joint distribution of $\mathbf{b}, \mathbf{\Theta}$ can be written as:

$$\begin{aligned} P(\mathbf{b}, \Theta|\mathbf{c}, \mathbf{a}) &\propto \prod_k \prod_t \theta_{kt}^{m_{kt}}(1-\theta_{kt})^{m'_{kt}} P(\theta_{tk}) \\ &\propto \prod_k \prod_t \theta_{kt}^{m_{kt}+\alpha_h-1}(1-\theta_{kt})^{m'_{kt}+\beta_h-1} \end{aligned}$$

$\mathbf{\Theta}$ can be then integrated out as:

$$\begin{aligned} \delta m_{kt} &= P(\mathbf{b}|\mathbf{c}, \mathbf{a}) \\ &\propto \prod_k \prod_t \frac{\Gamma(\alpha_h + m_{kt})\Gamma(\beta_h + m'_{kt})}{\Gamma(\alpha_h + m_{kt} + \beta_h + m'_{kt})} \end{aligned} \tag{1}$$

The probability distribution of the $\mathbf{h}$ can be found in the similar manner:

$$\begin{aligned} P(h_{it}|b_{jt},\psi_j) &= \psi_j^{I(h_{it}=b_{jt})}(1-\psi_j)^{I(h_{it}\neq b_{jt})} \\ \text{Therefore, } P(h_{it}|b_{jt},\psi_j,d_i=j) &= \psi_j^{I(h_{it}=b_{jt})I(c_i=j)}(1-\psi_j)^{I(h_{it}\neq b_{jt})I(c_i=j)} \end{aligned}$$

$$\begin{aligned} p(\mathbf{h},\mathbf{\Psi}|\mathbf{d},\mathbf{b}) &\propto \prod_j\prod_i\prod_t \psi_j^{I(h_{it}=b_{jt})I(d_i=j)}(1-\psi_j)^{I(h_{it}\neq b_{jt})I(d_i=j)} \\ &= \prod_j\prod_t \psi_j^{\sum_i I(h_{it}=b_{jt})I(d_i=j)} \\ &= \prod_j\prod_t \psi_j^{l_{jt}+\alpha_h-1}(1-\psi_j)^{l'_{jt}+\beta_h-1} \\ \text{where, } l_{jt} &= \sum_i I(h_{it}=b_{jt})I(d_i=j) \end{aligned}$$

$$p(\mathbf{h}|\mathbf{d},\mathbf{b}) = \prod_j\prod_t R(\alpha_h,\beta_h)\frac{\Gamma(\alpha_h+l_{jt})\Gamma(\beta_h+l'_{jt})}{\Gamma(\alpha_h+\beta_h+l_{jt}+l'_{jt})}$$

Now we will give the gibbs sampling algorithm of each of variables $(\mathbf{d},\mathbf{h},\mathbf{b},\mathbf{c})$

## 2.4 Sampling d

In our model $\mathbf{d}_i$ is the equivalence class for sample $x_i$.

The sampling of $\mathbf{d}$ will proceed in a manner similar to [Xing et al., 2004].

$$\begin{aligned} P(d_i=j|\mathbf{d}_{-i},h_i,\mathbf{h}_{-i},\mathbf{b}) &\propto P(d_i=j|d_{-i},\mathbf{h}_{-i},\mathbf{b})P(h_i|d_i=j,\mathbf{d}_{-i},\mathbf{h}_{-i},\mathbf{b}) \\ &\propto P(d_i=j|d_{-i},\beta(j))\frac{P(\mathbf{h}|b_j,\mathbf{d})}{P(\mathbf{h}_{-i}|b_j,\mathbf{d}_{-i})} \\ &\propto P(d_i=j|d_{-i},\beta(j))P(h_i|b_j,l_{j[-i]}) \\ &\propto \chi(j)P(h_i|b_j,l_j[-i]) \\ &\propto \chi(j)\prod_t\frac{\eta(l_{jt[i]})}{\eta(l_{jt[-i]}} \end{aligned} \tag{2}$$

Where, $\chi=(n_1+\alpha_0,...,n_j+\alpha_0..,\alpha_0)$ and $\eta(l_j)=\frac{\Gamma(\alpha_h+l_{jt})\Gamma(\beta_h+l'_{jt}}{\Gamma(\alpha_h+l_{jt}+\beta_h+l'_{jt})}$. Further, $l_{jt[-s]}=\sum_{i:i\neq s}\sum_j I(h_{it}=b_{jt})I(d_i=j)$ and $l_{jt[s]}=l_{jt[-s]}+I(h_{st}=b_{jt})$

## 2.5 Sampling c

In our model $\mathbf{c}$ is the equivalence class for bad haplotype $\mathbf{b}$ We use a dirichlet prior $Dir(\alpha_0)$ on the $\mathbf{c}$. The hyperparameter can be further thought to be coming from a beta distribution.

$$
\begin{aligned}
P(c_j = k|\mathbf{c}_{-j}, \mathbf{a}, \mathbf{b}) &\propto P(c_j = k|\mathbf{c}_{-j})P(b_j|c_j = k, a_k, \mathbf{c}_{-j}, \mathbf{b}_{-j}) \\
&\propto (n_k + \alpha_0)P(b_j/a_k, m_{jt}[-j])) \\
&\propto (n_k + \alpha_0)\prod_t \frac{\Gamma(\alpha_b + \bar{m}_{kt})\Gamma(\beta_b + \bar{m}'_{kt})}{\Gamma(\alpha_b + \bar{m}_{kt} + \beta_b + \bar{m}'_{kt})} \\
&\propto (n_k + \alpha_0)\prod_t (\alpha_b + \bar{m}_{kt})^{I(b_{jt}=a_{kt})} \\
&\quad (\beta_b + \bar{m}'_{kt})^{I(b_{jt}\neq a_{kt})}
\end{aligned} \tag{3}
$$

## 2.6 Sampling of the b

We will take following two cases:

### 2.6.1 Old b

If old class is sampled in the **d**, the $b_{jt}$ will be sampled by following equation:

$$
\begin{aligned}
P(b_{jt}|\mathbf{b}_{-(jt)}, \mathbf{a}, \mathbf{c}, \mathbf{h}, \mathbf{d}) &\propto \prod_{i|d_i=j} P(h_{it}|b_{jt}, l_{jt})P(b_{jt}|a_{kt}, c_j = k, m_{kt}[-jt]]) \\
&\propto \frac{\Gamma(\alpha_h + \bar{l}_{jt})\Gamma(\beta_h + \bar{l}'_{jt})}{\Gamma(\alpha_h + \beta_h + \bar{l}_{jt} + \bar{l}'_{jt})} \\
&\quad \frac{\Gamma(\alpha_b + \bar{m}_{kt})\Gamma(\beta_b + \bar{m}'_{kt})}{\Gamma(\alpha_b + \bar{m}_{kt} + \beta_b + \bar{m}'_{kt})} \\
&\propto \frac{\Gamma(\alpha_h + \bar{l}_{jt})\Gamma(\beta_h + \bar{l}'_{jt})}{\Gamma(\alpha_h + \beta_h + \bar{l}_{jt} + \bar{l}'_{jt})} \\
&\quad (\alpha_b + \bar{m}_{kt})^{I(b_{jt}=a_{kt})}(\beta_b + \bar{m}'_{kt})^{I(b_{jt}\neq a_{kt})}
\end{aligned} \tag{4}
$$

Where, $\bar{m}_{kt} = \sum_j I(b_{jt} = a_{kt})I(c_j = k)$, is the sufficient statistics of $\theta_{tk}$ considering the $b_{jt}$ is current bad ancestor. Where $\bar{l}_{jt}$ is the number of allelic instances that are consistent with ancestor $b_j$ having pattern $b_{jt}$.

### 2.6.2 New b

If new class is sampled in the **d**, we don't know $c_{J+1}$. Therefore, we have to marginalize over all **c**. The $b_{jt}$ will be then sampled by eqn. 5.

$$
\begin{aligned}
P(b_{jt}|b_{-(jt)}, \mathbf{a}, \mathbf{c}, \mathbf{h}, \mathbf{d}) &= P(h_{it}|b_{jt})P(b_{jt}|a_{kt,c_j=k}, m_{kt}[-jt]]) \\
&\propto \frac{\alpha_h}{\alpha_h + \beta_h}^{I(h_{it}=b_{jt})} \frac{\beta_h}{\alpha_h + \beta_h}^{I(h_{it}\neq b_{jt})} \\
&\quad (\alpha_b + \bar{m}_{kt})^{I(b_{jt}=a_{kt})}(\beta_b + \bar{m}'_{kt})^{I(b_{jt}\neq a_{kt})}
\end{aligned} \tag{5}
$$

The prior on $c_j$ is dirichlet distribution.

### 2.7 Sampling $\mathbf{h}$

The Gibbs sampling for $\mathbf{h}$ is similar to the [Xing et al., 2004].

## 3 Implementation and Result

Our implementation builds upon and significantly changes the code used in paper [Xing et al., 2004]. We also parallelize the code using OpenMP [Chapman et al., 2007] and ran it on the 8 clusters, each node having 8 processor. We were able to parallelize only the first part of the inferencing problem i.e. identifying Ancestral population. Our implementation has more than 5000 lines of C code.

The size of our dataset is massive in comparison to previous work and takes considerable time to run. Besides that implementing the complete code (joint inferencing of bad haplotypes) required considerable amount of effort and we are still trying to run the complete code. After that we are planning to parallelize it.

To access the quality of result, We performed following test.

### 3.1 Motif enrichment analysis

Motifs are sequence where transcription factor binds that eventually effect the expression of near genes. If a mutation lie in one of this motifs in the genome, it is likely it will affect its binding affinity either increasing or decreasing its affinity toward its particular transcription factor. Since we are studying cardio-vascular disease in particular, we expect many of the region where these mutations exists i.e mutation rate $\theta$ is high, will be enriched in the heart related motifs. So that it can disrupt the cardio-vascular regulatory pathway thus explaining the mechanism by which a mutation causes heart disease. To test this hypothesis we sorted the SNP location based on their estimated mutation rate. We declared to top 10% of this sorted list as our foreground set and rest of SNP as background set and using 300bps flanking region around this SNP, performed motif enrichment analysis.

We first ensured the GC content between foreground and background; we divided both foreground and background into bins as with similar GC content and randomly sampled from bins so that proportion of GC content remains same between foreground and background. We then, performed PWM scan on these region with TRANSFAC vertebrate motifs. We found that foreground are enriched in motifs which have at least heart associated function. This validates that mutation found by this method are not random and are in fact heart related and are good candidates for being deleterious mutation.

### 3.2 Comparison with GWAS studies

Once we can run the program on complete genome, we can further perform comparison of our canditates deleterious mutation with those reported by traditional GWAS studies. We are also planning to do a differential analysis on the input genotype dataset using logistic regression based approach to select only those set of SNPs which have high probability of being related to heart disease based on p-value.

# References

[Chapman et al., 2007] Chapman, B., Jost, G., and Pas, R. v. d. (2007). *Using OpenMP: Portable Shared Memory Parallel Programming (Scientific and Engineering Computation).* The MIT Press.

[Ewens and Tavaré, 1998] Ewens, W. and Tavaré, S. (1998). Ewens sampling formula. *Encyclopedia of Statistical Sciences.*

[Ferguson, 1973] Ferguson, T. (1973). A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.

[Hoppe, 1984] Hoppe, F. (1984). Pòlya-like urns and the ewens' sampling formula. *Journal of Mathematical Biology*, 20(1):91–94.

[Sohn and Xing, 2009] Sohn, K. and Xing, E. (2009). A hierarchical dirichlet process mixture model for haplotype reconstruction from multi-population data. *The Annals of Applied Statistics*, 3(2):791–821.

[Teh, 2010] Teh, Y. (2010). Dirichlet process. *Encyclopedia of Machine Learning, Springer.*

[Teh et al., 2006] Teh, Y., Jordan, M., Beal, M., and Blei, D. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

[Xing et al., 2007] Xing, E., Jordan, M., and Sharan, R. (2007). Bayesian haplotype inference via the dirichlet process. *Journal of Computational Biology*, 14(3):267–284.

[Xing et al., 2004] Xing, E., Sharan, R., and Jordan, M. (2004). Bayesian haplo-type inference via the dirichlet process. In *Proceedings of the twenty-first international conference on Machine learning*, page 111. ACM.