Supplemental Appendix 1 - inferring SNVs from aligned reads using a Bayesian mixture model

Sohrab P. Shah, Ryan D. Morin and Samuel Aparicio

1 Introduction and problem statement

Given a set of aligned short reads from next generation sequence data, it is of interest to identify single nucleotide variants (SNVs), defined as nucleotide positions in the genome (or transcriptome) where at least one allele is different from the reference base. SNVs predicted from the short read alignments become candidate germline or somatic mutations that are important to identify in order to catalogue the mutational spectrum of the genome (or transcriptome) under investigation. We developed a Bayesian mixture model called SNVmix for this purpose that takes as input the complete set of aligned reads and produces as output the probability of an SNV at each position represented in the data.

2 SNVmix model specification

The proposed model for SNVmix is shown as a probabilistic graphical model in Figure 1. We will now describe the model in detail. Consider \( G_i = k, k \in \{aa, ab, bb\} \) to be a multinomial random variable representing the genotype at nucleotide position \( i \), where \( aa \) is homozygous for the reference allele, \( ab \) is heterozygous and \( bb \) is homozygous for the non-reference allele. We represent the observed allele frequency \( X_i = [a_i, b_i]^T \) as a vector of counts of the reference and non-reference alleles at position \( i \) such that \( N_i = a_i + b_i \) is the observed read depth at position \( i \). Figure 2 shows an example of the input and how aligned reads are transformed into allelic counts. The central idea is that we assume the counts were generated by a class conditional density, such that conditioned on \( G_i = k \), \( X_i \sim Binom(a_i | \mu_k, N_i) \) where \( \mu_k \) is the parameter of a Binomial distribution for genotype \( k \). \( \mu_k \) is a key quantity that models the expectation that for a given genotype \( k \) a randomly sampled allele will be the reference allele. Intuitively we should expect \( \mu_{aa} \) to be close to 1, \( \mu_{ab} \) to be close to 0.5 and \( \mu_{bb} \) to be close to 0. Given a value for \( \mu_k \) the probability density function for \( Binom(a_i | \mu_k, N_i) \) is given by:

\[
\begin{pmatrix} N_i \\ a_i \end{pmatrix} \mu_k^{a_i}(1 - \mu_k)^{b_i}
\]

(1)

Thus, the key intuition is that for genotype \( k = aa \), the Binomial distribution defined by \( \mu_{aa} \), should be highly skewed toward the reference allele. Similarly for \( k = bb \), the distribution defined by \( \mu_{bb} \) would be skewed toward the non-reference allele. For \( k = ab \), the distribution would be much more uniform. We represent the prior probability of observing genotype \( k \) at any position with a multinomial variable \( \pi \), such that for the components of \( \pi \), \( 0 \leq \pi_k \leq 1 \ \forall k \) and \( \sum_k \pi_k = 1 \). Therefore, we assume a classical generative mixture model to explain the observed data. As such, for a given position \( i \), the marginal distribution of \( X_i \) is generated from a convex combination of the class conditional Binomial densities, weighted by the multinomial \( \pi \):

\[
p(X_i) = \sum_{k=1}^{K} \pi_k Binom(X_i | \mu_k, N_i)
\]

(2)

where \( K = 3 \) is the number of possible genotypes.
**Figure 1:** SNVmix model for detecting SNVs from next generation sequence data shown as a probabilistic graphical model (see [1]). Random variables are shown as round circles. Rounded rectangles indicate user settable parameters (or hyperparameters). Observed quantities (or those known at time of inference) are shaded. Unobserved quantities (or those we wish to infer) are unshaded. Arrows indicate a probabilistic dependency of the head variable on the tail variable. The conditional probability distributions are shown on the right. Where Dir is the Dirichlet, Mult is the Multinomial, Binom is the Binomial and Beta is the Beta distribution.

\[
p(\pi | \delta) \sim \text{Dir}(\pi | \delta)
\]
\[
p(G_i = k | \pi) \sim \text{Mult}(G_i = k | \pi)
\]
\[
p(X_i | G_i = k, \mu_k) \sim \text{Binom}(X_i | \mu_k)
\]
\[
p(\mu_k | \alpha_k, \beta_k) \sim \text{Beta}(\mu_k | \alpha_k, \beta_k)
\]

**Figure 2:** Schematic diagram of input data to SNVmix. We show how allelic counts (bottom) are derived from aligned reads (top). The reference sequence is shown bolded in blue. The arrow indicates a position containing an SNV. The non-reference bases are shown in red.
Therefore the complete data log-likelihood is given by:

$$\log p(X_{1:T} | \mu_{1:K}, \pi) = \sum_{i=1}^{T} \log \sum_{k=1}^{K} \pi_k \text{Binom}(X_i | \mu_k, N_i)$$

(3)

where $T$ is the number of observed positions in the input.

The parameters $\theta = (\pi, \mu_{1:K})$ are fit to data using maximum a posteriori (MAP) expectation maximization (EM). If the true genotype is known, $\theta$ can be learned in a supervised setting using the true genotypes as training data. Our goal is to infer the genotype state given the model parameters and the data. We can make use of Bayes rule and express $\gamma_i(k) = p(G_i = k | X_{1:N}, \pi, \mu_k)$ as the marginal probability of the genotype for position $i$ given all the data and the model parameters:

$$\gamma_i(k) = \frac{\pi_k \text{Binom}(X_i | \mu_k, N_i)}{\sum_{j=1}^{K} \pi_j \text{Binom}(X_i | \mu_j, N_i)}$$

(4)

3 Prior distributions

We assume that $\pi$ is distributed according to a Dirichlet distribution parameterized by $\delta$, the so-called pseudocounts. We set delta to be skewed towards $\pi_{aa}$ under the assumption that most positions will be homozygous for the reference allele. $\mu_k$ is conjugately distributed according to a Beta distribution: $\mu_k \sim \text{Beta}(\alpha_k, \beta_k)$. We set $\alpha_{aa} = 1000, \beta_{aa} = 1; \alpha_{ab} = 500, \beta_{ab} = 500$ and $\alpha_{bb} = 1, \beta_{bb} = 1000$ assuming that the $\mu_{aa}$ should be near close to 1, $\mu_{ab}$ should be close to 0.5 and $\mu_{bb}$ should be close to 0.

4 Model fitting and parameter estimation

We fit the model using the expectation maximization (EM) algorithm. In order to begin EM the model parameters must be initialised. We initialise $\mu_k$ by taking the mean of 1000 random samples from the Binomial parameterized by $\mu_k = \frac{\alpha_k}{\alpha_k + \beta_k}.$ $\pi(k)$ is initialised to $\frac{\delta(k)}{N_\delta}$ where $N_\delta = \sum_k \delta(k)$.

The EM algorithm iterates between the E-step where we assign the genotypes using Equation 4 and the M-step where we re-estimate the model parameters. At each iteration we evaluate Equation 3 and the algorithm terminates when this quantity no longer increases.

The M-step equations are standard conjugate updating equations:

$$\pi_{new}(k) = \frac{\sum_{i=1}^{T} I(G_i = k) + \delta(k)}{\sum_{j=1}^{K} \sum_{i=1}^{T} I(G_i = j) + \delta(j)}$$

(5)

where $I(G_i = k)$ is an indicator function to signal that $G_i$ is assigned to state $k$ as position $i$, and:

$$\mu_{new}^{aa} = \frac{\sum_{i=1}^{T} I(G_i = k) + \alpha_k}{\sum_{i=1}^{T} \sum_{j=1}^{K} I(G_i = k) + \alpha_k + \beta_k - 2}$$

(6)

5 Evaluation of model

We evaluated the performance of SNVmix by running it to convergence using 15000 coding positions determined by using data from an orthogonal assay, namely genotype calls from an Affymetrix SNP 6.0 genotyping array. We limited the positions by only considering coding positions where the CRLMM algorithm [2] predicted genotype with $> 0.99$ confidence. We therefore defined any position that was predicted to be heterozygous or homozygous for the non-reference allele to be a true positive (TP) and any position predicted to be homozygous for the reference allele a true negative (TN). We then computed $p(SNV_i) = \gamma_i(ab) + \gamma_i(bb)$ representing the posterior marginal probability of an SNV at position $i$ as predicted by the model. This allowed us to compute standard receiver operator characteristic (ROC) curves in order to quantitatively assess performance. We fit the model separately for WTSS-PE and WGSS-PE.
WGSS-PE AUC = 0.985  
WTSS-PE AUC = 0.976  

Figure 3: ROC curves of SNVMix predictions compared against high confidence genotype calls from an Affymetrix SNP 6.0 array. The WGSS-PE (left) and WTSS-PE (right) indicate the algorithm is highly concordant with the orthogonally derived results from the SNP chip.

using the set of positions described above and computed $p(SNV_{1:T})$ for each set of data. The resulting ROC curves are shown in Figure 3 and demonstrate the algorithm is highly concordant with the high-confidence genotype calls from the SNP chip for both WGSS-PE (left) and WTSS-PE (right).

6 Computation of SNVs and selection criteria for validation

Using the converged parameter estimates from the model fits (described above) we evaluated Equation 4 for all the data for WTSS-PE, WGSS-PE and WGSS-PRI. We chose to call an SNV as such by determining threshold probability $t$ at the 0.01 false positive rate. This was $t = 0.53$ for WTSS-PE and $t = 0.77$ for WGSS-PE. We applied these thresholds so that locations where $P(SNV_i) > t$ were ‘called’ SNVs. We then filtered out any SNVs that were present in dbSNP v129, Jim Watson, Craig Venter, the Yoruban male and data from the 1000 genomes project as described in the Supplemental Methods. The remaining SNVs were categorized according whether or not they induced a non-synonymous change. All non-synonymous changes were selected for validation by PCR amplicon sequencing.

7 Theoretical bounds

Figures 4 and 5 shows how the algorithm behaves under different depths over the distribution of allelic counts. Two observations that can be made is as depth increases, the highest probability increases. For example at $N = 2$ (top left) there are 0 points at probability = 1. By contrast, there are almost no positions after $N = 20$ for which the prediction for the most probable genotype is not 1. This is intuitive in that given more data, the model will be more certain in its prediction.
Figure 4: Theoretical behaviour of SNVmix at depths of 2, 3, 5, 10, 15, 20, 35, 50 and 100. The plot show how the distribution of marginal probabilities changes with the number of reference alleles given the model parameters fit the WGSS-PE data.
Figure 5: Theoretical behaviour of SNVmix at depths of 2, 3, 5, 10, 15, 20, 35, 50 and 100. The plot show how the distribution of marginal probabilities changes with the number of reference alleles given the model parameters fit the WTSS-PE data.
References
