Supplementary for “MixClone: a mixture model for inferring tumor subclonal populations”

By Yi Li

1 Detecting heterozygous SNP sites

The probabilistic model of MixClone incorporates allele frequency information of each genomic segment by modelling the read counts information of the heterozygous SNP sites within each segment. To identify heterozygous SNP sites, we filter out sites without significant heterozygosity in the normal sample, which is defined as

$$\frac{b_{ij}^N}{d_{ij}^N} < 0.40 \text{ or } \frac{b_{ij}^N}{d_{ij}^N} > 0.60$$

2 Curating the baseline segments

To determine the absolute value of read counts mapped to each segment $j$, which is denoted as $\lambda_j$, we curate a list of segments which contain no loss of heterozygosity (LOH) according to their allele frequencies information. Specifically, we quantify the significance of LOH of each segment in the tumor sample, which is often resulted from somatic copy number alterations (SCNAs). The LOH significance of each segment provides additional information of the allelic configuration of the segment, e.g. segments with significant LOH must have unequal copy numbers for each allele. To quantify the LOH significance of each segment in the tumor sample, we calculate the fraction of LOH sites in the segment. We determine whether a site is LOH based on its one-tailed p-value of the binomial test

$$p_{ij}^T = \text{Binomial}_\text{cdf}(\min(d_{ij}^T - b_{ij}^T, b_{ij}^T)|d_{ij}^T, 0.5)$$

A site is defined as LOH if $p_{ij}^T < 0.025$. Then based on the fraction of LOH sites within each segment we assign an LOH status to the segment

- $\text{LOH}_{\text{fraction}} < \text{LOH}_{\text{min}}$, $\text{LOH}_{\text{status}} = \text{FALSE}$
- $\text{LOH}_{\text{min}} \leq \text{LOH}_{\text{fraction}} < \text{LOH}_{\text{max}}$, $\text{LOH}_{\text{status}} = \text{UNCERTAIN}$
- $\text{LOH}_{\text{fraction}} \geq \text{LOH}_{\text{max}}$, $\text{LOH}_{\text{status}} = \text{TRUE}$

We choose $\text{LOH}_{\text{min}} = 0.09$, $\text{LOH}_{\text{max}} = 0.11$ in practice.

Those segment with LOH status “FALSE” are curated as the initial list of baseline segments. We further remove “outlier” segments from the list if their copy numbers, which are correlated with the read depth ratios between the tumor and the normal samples, are different from the bulk of the segments in the list. The remaining segments are defined as the final set of baseline segments.

3 EM algorithm

We find the MLE of the model parameters using the Expectation-Maximization (EM) algorithm [1]. Since we have two sets of latent variables $H_j, Z_j$ for segment $j$, we define two auxiliary variables $\psi_{jh}, \kappa_{jk}$ to denote the conditional probability of the latent variable $H_j, Z_j$, conditional on all the observed data and the model parameters $\Theta$.

$$\psi_{jh} = \mathbb{P}(H_j = h|D_j^T, \{b_{ij}^T\}_{i=1}^J, \Theta)$$

$$\kappa_{jk} = \mathbb{P}(Z_j = k|D_j^T, \{b_{ij}^T\}_{i=1}^J, \Theta)$$

Then, the expected complete log-likelihood is specified as

$$Q = \sum_{j=1}^J \sum_{k=1}^K \sum_{h \in \mathcal{H}} \psi_{jh} \left( \log \pi_k + \log \rho_{jh} + [D_j^T \log \lambda_j - \lambda_j - \log D_j^T] \right) + \sum_{i=1}^I \log \sum_{g \in \mathcal{G}} Q_{gh} \left( \frac{d_{ij}^T}{b_{ij}^T} \mu_{ij} (1 - \mu_{ij}) d_{ij}^T - b_{ij}^T \right)$$

and the subsequent E-step and M-step of the EM algorithm can be derived as follows:

E-step
We calculate the conditional probabilities of the latent variables given the observed data and previously estimated model parameters

$$
\psi_{jh}^{(t+1)} = \mathbb{P}(H_j = h| D_j^T, \{\psi_{ij}^{(t)}\}_{i=1}^{T}, \Theta^{(t)})
$$

$$
\kappa_{jk}^{(t+1)} = \mathbb{P}(Z_j = k| D_j^T, \{b_{ij}^{(t)}\}_{i=1}^{T}, \Theta^{(t)})
$$

**M-step**

We estimate the model parameters based on the conditional probabilities calculated in E-step. For \(\rho_j\), the MLE is given by

$$
\rho_{jh}^{(t+1)} = \psi_{jh}^{(t+1)}
$$

For \(\pi_k\), the MLE is given by

$$
\pi_k^{(t+1)} = \frac{\sum_{j=1}^{J} \kappa_{jk}^{(t+1)}}{J}
$$

For each subclonal cellular prevalence \(\phi_k\), ideally we need to take the derivative of \(Q\) in respect to each \(\phi_k\) to find the value that maximize \(Q\) as the MLE. However, the solution of \(\frac{\partial Q}{\partial \phi_k} = 0\) is not in closed-form. Thus, we numerically search each \(\phi_k\) to find the value that maximize \(Q\). Additionally, the computation for the binomial distribution part of \(Q\), which accounts for the sequence information from heterozygous SNP sites, is time-consuming when applying iteratively numerical search. However, the log-likelihood contribution of the binomial distribution part to \(Q\) is negligible, comparing the to the Poisson distribution part, which accounts for the sequence information from SCNAs. So practically, we numerically search the value of each \(\phi_k\) that only maximize the Poisson distribution part of \(Q\), denoted as \(Q_{SCNAs}\), as the MLE for \(\phi_k\), which is specified as

$$
Q_{SCNAs} = \sum_{j=1}^{J} \sum_{k=1}^{K} \kappa_{jk} \sum_{h \in \mathcal{H}} \psi_{jh} \{ \log \pi_k + \log \rho_{jh} + \left[ D_j^T \log \lambda_j - \lambda_j - \log D_j^T \right] \}
$$

### 4 Sequencing data simulation

We used the method described in [2] to simulate reads from artificial paired normal-tumor samples. First, we created the genomic sequences of the artificial normal genome and tumor genome. We inserted SNP sites into the sequence of the reference genome NCBI Build 36.1 (hg18) to generate the sequences of the artificial normal genome with respect to the two alleles. Based on the SNP frequency observed in the human population (one SNP every 1.9 kilobases [3]), we randomly selected 10% SNP sites of dbSNP 132 [4] for each allele, and substituted the base of the reference sequence with the variant base of each SNP site. To generate the sequences of the two alleles of the artificial tumor genome, we inserted the somatic point mutation sites reported in PR2832 [5] into the artificial normal genome, and substituted the base of the artificial normal genome with the variant base of each somatic point mutation site.

Next, based on the genomic sequences of the artificial tumor genome, we created ten heterogeneous tumor genomes that differ in their segmentations, allelic configurations and subclonal cellular prevalences. Among the ten heterogeneous tumor genomes, five of them contain two subclonal populations and the other five contain three subclonal populations. We only used chromosome 1 for all the simulated datasets to reduce computational time. We randomly segmented chromosome 1 (excluding the centromere, heterochromatin and 1q21.1) into 20 segments for tumor genomes with two subclonal populations and 30 segments for tumor genomes with three subclonal populations. For one half of the segments of each tumor genome, we assigned their allelic configuration as \(PM\), as we assume the bulk of the baseline segments are diploid. For the other half of the segments, we randomly assigned them non-diploid allelic configurations with respect to both alleles based on a prior distribution, which is specified according to the edit distance between the non-diploid allelic configuration and the diploid allelic configuration \(PM\). Specifically, we denote the edit distance between allelic configuration \(h\) and \(PM\) as \(d_h\), e.g. the edit distance between \(h = PPMN\) and \(PM\) is 2. Large edit distance is less likely to happen as it represents large somatic copy number alterations. Therefore, the probability of sampling allelic configuration \(h\) for a given segment \(j\) is specified as

$$
P(H_j = h) = \frac{\alpha^{d_h}}{\sum_{h \in \mathcal{H}} \alpha^{d_h}}
$$

in which \(\alpha < 1\) is a shrinkage coefficient to make \(P(H_j = h)\) decrease exponentially with the increase of the edit distance \(d_h\). We choose \(\alpha = 0.5\) in practice, and the set of allelic configurations \(\mathcal{H}\) was formulated with maximum copy number 4 for all the simulated datasets.

Finally, for each subclonal population within each tumor genome, we randomly picked its cellular prevalence from the set \({0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9}\), and each cellular prevalence is at least 0.2 away from other cellular prevalences within the same tumor genome. We uniformly sampled reads within each segment, assuming the sequencing process has no error and all the segments share the same mappability. Thus, the factors that affect the probability of sampling a read from a specific segment are only its copy number, subclonal cellular prevalence and genomic length. We simulated 150 million reads of single-end and 100-bp length for each read set, which is roughly 60X coverage for chromosome 1.
5 Sequencing data preprocessing

NCBI Build 36.1 (hg18) was used as the reference genome for all the datasets. We used bowtie-0.12.8 [6] with the option “-n 3 -k 2” to align the reads of simulated datasets. We used BWA-0.7.5a [7] with the default setting to align the reads of the breast cancer datasets. We used segmentations based on both ground truth and BIC-seq-1.2.1 [8] for simulated datasets, and segmentations based on BIC-seq for the breast cancer datasets. For simulated datasets with segmentations based on BIC-seq, we used segments with genomic length longer than 1 million bp as the input. We used SAMtools-0.1.18 [9] to convert reads alignment from SAM format to BAM format with the option “-Sbh -q 1” and also eliminate duplicates. We used GATK-2.4-7 [10] for base quality score recalibration and indel realignment with dbSNP 132 [4] for the breast cancer datasets. For more details about the usage of the breast cancer datasets, please refer to [2].

6 Parameter setting details

6.1 MixClone details

We used “--max_copynumber 4” for the simulated datasets, and “--max_copynumber 6” for the breast cancer datasets.

6.2 PyClone details

We used the beta-binomial model of PyClone with the default setting to run all the analyses on the simulated datasets.

7 Supplementary figure 1-2


Figure S1: The subclonal cellular prevalences estimated by MixClone, the tumor purities estimated by PyLOH, THetA [2], and the tumor purities estimated by ABSOLUTE [11] reported in [12] of sample MB-45.
Figure S2: The subclonal cellular prevalences estimated by MixClone, the tumor purities estimated by PyLOH, THetA [2], and the tumor purities estimated by ABSOLUTE [11] reported in [12] of sample MB-123.
References


