Description of the Sequenza algorithm

Francesco Favero[*], Aron C. Eklund[†]

October 8, 2014

1 Aims

The goal is to use the sequencing data to estimate the following unknown integer-valued parameters at each genomic position i of the tumor: the copy number n_i, the minor allele copy number m_i. The copy number n_i is defined as the total number of alleles present at position i. The minor allele copy number m_i is defined as the smaller of the two allele-specific copy numbers ($m_i \leq n_i/2$).

Estimation of these parameters requires estimation of two real-valued meta-parameters that may also be biologically or clinically informative: the cellularity ρ, defined as the fraction of tumor cells in the sample, and the ploidy ψ, defined here as twice the ratio of tumor DNA mass to normal DNA mass.

In addition we define the normal genome constants at all positions: copy number n_{0_i} which is 2 for the autosomes, the minor allele copy number m_{0_i} equal to 1, and the ploidy of the human genome which is 2.

2 Input

The input data consists of two *pileup* files derived from the aligned, and possibly filtered and/or trimmed, sequencing reads of the tumor specimen and of the matched normal specimen. The pileup files provide the observed bases corresponding to each genomic position i. For every position we extract the number of reads covering that position, read depth, in the tumor τ_i and in the normal ν_i. Only the positions present in both the normal and tumor samples and where the sum of τ_i and ν_i is greater than a defined threshold (20 by default) are included in further analysis. We define the *unnormalized depth ratio* as $r'_i = \tau_i/\nu_i$.

From the normal pileup we identify the subset of genomic positions that are heterozygous, defined as those where two bases are detected with the less abundant representing at least 25% of the reads. For each heterozygous position, we extract the tumor B allele frequency b_i, defined as the lower of the two base frequencies in the tumor.

[*] favero@cbs.dtu.dk

[†] eklund@cbs.dtu.dk

At each homozygous position, we compare the two pileup files and identify those positions with variant reads in the tumor, and we calculate the mutant allele frequency F_i as the fraction of bases in the tumor that differ from the normal at position i. The values F_i are used only in the graphical output, not for the estimation of the parameters. Nevertheless F_i indirectly provides an independent visual validation of the cellularity and ploidy estimation by comparing the detected mutant allele frequencies to the expected allele frequencies.

2.1 Depth ratio normalization

Variation in the local abundance of G-C base pairs is a factor known to cause uneven sequencing coverage along the genome. Also the different library size between the normal and the tumor sequencing will effect the average depth ratio value. In order to reduce bias resulting from this effects, we assume that copy number is independent of GC content and perform GC-normalization as follows: First, for every position i in the reference genome, we calculate the GC fraction g_i in a 50-base (by default) window surrounding the position. Then, for each of the (50 by default) possible values of g_i, we define λ_k as the mean unnormalized depth ratio over all positions i such that $g_i = k$. Finally, we define the normalized depth ratio r_i (hereafter simply *depth ratio*) as: $r_i = r'_i/\lambda_{g_i}$. This process takes care of the normalization of the GC-content effect for each depth ratio value, and simultaneously, the library size normalization, adjusting the average depth ratio to 1.

3 Segmentation

The parameters n_i and m_i typically remain constant over a large number of consecutive genomic positions; therefore we can reduce the n_i and m_i specified at each genomic position to a set of genomic segments of defined start and end positions, each with a copy number N_s and minor allele copy number M_s. The segment boundaries are estimated directly from the r and b using allele-specific copy number segmentation implemented in the `copynumber` package from Nilsen et al[1]. For each segment s, we record the length in megabases L_s, and we calculate the mean and the standard deviation for the depth ratio (R_s and S_{R_s}) and for the B allele frequency (B_s and S_{B_s}).

We also define the normal genome constants at all segments: the copy number N_{0_s} (2 in autosomes) and the minor allele copy number M_{0_s} (1 in autosomes).

4 Probabilistic model

We take as observables the output from the segmentation step: R_s, B_s, S_{R_s}, S_{B_s}, and L_s for each segment s. For convenience, we introduce bold symbols $\boldsymbol{R}$, $\boldsymbol{B}$, $\boldsymbol{S_R}$ and $\boldsymbol{S_B}$, each indicating the set, over all segments, of the respective segment-level parameters. Also, we use $\boldsymbol{x}$ as shorthand to indicate all observables.

The model includes the following parameters: Two meta-parameters ρ and ψ, as well as segment-level parameters N_s, and M_s for each segment s, as described previously. Additionally, we introduce parameters characterizing the dispersion in depth ratio and B allele frequency in each individual segment: σ_{R_s} and σ_{B_s}. For convenience we use $\boldsymbol{N}$, $\boldsymbol{M}$, $\boldsymbol{\sigma_R}$, and $\boldsymbol{\sigma_B}$ for the respective sets of parameters over all segments, as well as $\boldsymbol{\theta}$ to represent the entire set of parameters and θ_s to represent the set of parameters for a specific segment.

Under our model the probability of generating the observations $\boldsymbol{R}$ and $\boldsymbol{B}$, corresponding to the measure of depth ratio and B allele frequency of each segment s, given a set of parameters $\boldsymbol{\theta}$, which indicate a specific cellularity and ploidy state, and a vectors $\boldsymbol{N}$ and $\boldsymbol{M}$, which indicate the copy number and the number of minor alleles for all segments, is described by the formula:

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_s p(R_s, B_s|\theta_s) \tag{1}$$

Where the probability of each segment is given by the product of the probabilities of the observed depth ratio and the observed B allele frequency of the segment:

$$p(R_s, B_s|\theta_s) = p(R_s|\theta_s)p(B_s|\theta_s) \tag{2}$$

Both of these probability densities are modeled with a non-standardized Student's t distribution with the estimated dispersion parameters σ_{R_s} and σ_{B_s} as scale parameters (σ), the expectation values $E[R_s|\theta_s]$ and $E[B_s|\theta_s]$ of the depth ratio and the B allele frequency with a given set of parameters θ_s as location parameter (μ), and the degrees of freedom (ν) set to 5:

$$p_t(x|\mu,\sigma) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})}\left(\frac{\sigma^2}{\pi\nu}\right)^{\frac{1}{2}}\left(1+\frac{\sigma^2(x-\mu)^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{3}$$

Therefore, the probability of observing a given depth ratio is equivalent to:

$$p(R_s|\theta_s) = p_t(R_s|E[R_s|\theta_s], \sigma_{R_s}) \tag{4}$$

And for the B allele frequency:

$$p(B_s|\theta_s) = p_t(B_s|E[B_s|\theta_s], \sigma_{B_s}) \tag{5}$$

4.1 Expected depth ratio

For each segment s, the expected depth ratio $E[R_s|\theta_s]$ is given by the equation:

$$E[R_s|\theta_s] = \frac{\bar{\tau}}{\bar{\nu}}\big[1-\rho+\big(\rho\frac{N_s}{N_{0_s}}\big)\big]\frac{1}{(\psi\rho)2(1-\rho)} \tag{6}$$

where the ratio of $\bar{\tau}$ to $\bar{\nu}$, the average depth of the tumor and the average depth of the normal, is equal to 1 after normalization, and N_{0_s} is the copy number of the segment s in the germline specimen.

4.2 Expected B allele frequency

In a simplified model in which the B allele frequency is defined *a priori*, the expected B allele frequency value β_s is given by the equation:

$$\beta_s = \frac{(M_s\rho) + M_{0_s}(1-\rho)}{(N_s\rho) + N_{0_s}(1-\rho)} \tag{7}$$

To calculate the expected B allele frequency $E[B_s|\theta_s]$ for a segment s, we have to consider that to calculate the average value B_s, we use only values below or equal 0.5. This selection cause a systematic effect in the expected values of the model.

In order to account for this effect, for each expected value we define a distribution $f(x)$:

$$f(x) = \begin{cases} p_t(x|\beta, \bar{\sigma}_B) + p_t(x|1-\beta, \bar{\sigma}_B) & \text{if} \quad 0 \leq x \leq 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

and the expected B allele frequency for a segment is calculated by:

$$E[B_s|\theta] = \frac{\int f(x)\, x \,\mathrm{d}x}{\int f(x) \,\mathrm{d}x} \tag{9}$$

In other words $E[B_s|\theta_s]$ is the mean value from the distribution generated by the function $f(x)$.

5 Maximum *a posteriori* estimation

5.1 Definition of priors

Based on our analyses of absolute copy number in solid tumors, the copy number 2 usually occurs at least twice as frequently as any other copy number state (data not shown). Considering each segment s we define prior probabilities of copy number p_{N_s}, by default favoring the solution where $N_s = 2$:

k	0	1	2	3	4	5	6	7
weight	1	1	2	1	1	1	1	1
$p_{N_s=k}$	0.11	0.11	0.22	0.11	0.11	0.11	0.11	0.11

5.2 Estimation of the parameters

To simplify computation, we assume the dispersion parameters σ_{R_s} and σ_{B_s} can be estimated from the observed data as follows:

$$\hat{\sigma}_{R_s} = \frac{S_{R_s}}{\sqrt{L_s}} \tag{10}$$

$$\hat{\sigma}_{B_s} = \frac{S_{B_s}}{\sqrt{L_s}} \tag{11}$$

We consider two distinct categories of parameters: The segment-specific parameters: N_s, M_s, σ_{R_s}, σ_{B_s} and the overall meta-parameters: ρ and ψ. The meta-parameters are descriptive of an overall state of a sample, and given values for the meta-parameters it is possible to estimate the corresponding segment-specific parameters using the equation:

$$\begin{aligned} &\hat{N}_s(\psi,\rho), \hat{M}_s(\psi,\rho) = \underset{k,j}{\arg\max}\, p(R_s, B_s \,|\psi, \rho, N_s = k, M_s = j, \hat{\sigma}_{R_s}, \hat{\sigma}_{B_s})\; p_{N_s=k} \\ &\text{where} \quad k \in \{0, 1, ..., 7\}, \quad j \in \mathbb{Z} \quad \text{and} \quad j \leq k/2 \end{aligned} \tag{12}$$

In order to estimate $\hat{\psi}$ and $\hat{\rho}$ we use a comprehensive grid-based search, considering a given range of the two parameters and apply a maximum *a posteriori* approach, where the conditional posterior probability is expressed by the equation:

$$P(\psi, \rho|\boldsymbol{x}) = \prod_s p(R_s, B_s \,|\psi, \rho, \hat{N}_s(\psi,\rho), \hat{M}_s(\psi,\rho), \hat{\sigma}_{R_s}, \hat{\sigma}_{B_s})\; p_{N_s=\hat{N}_s(\psi,\rho)} \tag{13}$$

By default the grid is defined by ρ ranging from 0.1 to 1 in steps of 0.01, and ψ ranging from 1 to 7 in steps of 0.1. The estimated parameters $\hat{\psi}$ and $\hat{\rho}$ are chosen by the set maximizing the equation:

$$\hat{\psi}, \hat{\rho} = \underset{\psi,\rho}{\arg\max}\, P(\psi, \rho|\boldsymbol{x}) \tag{14}$$

Once the parameters $\hat{\psi}$ and $\hat{\rho}$ are estimated we re-estimate the copy number parameters $\hat{N}_s$ and $\hat{M}_s$ with Equation 12, but using an increased maximum limit for k; by default up to 20 copies.

References

[1] Gro Nilsen, Knut Liestøl, Peter Van Loo, Hans Kristian Moen Vollan, Marianne B Eide, Oscar M Rueda, Suet-Feung Chin, Roslin Russell, Lars O Baumbusch, Carlos Caldas, Anne-Lise Børresen-Dale, and Ole Christian Lingjaerde. Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. *BMC genomics*, 13:591, January 2012.