Notes on GBStools algorithms

Tom Cooke

July 30, 2013

1 Notations

Suppose there are N sites for n diploid individuals, and each site is composed of a restriction site with alleles {+, -}, and a SNP with alleles {A, a}. SNP alleles on the same haplotype as the '+' allele are sampled by GBS, but alleles on the same haplotype as the '-' allele are not. The '-' allele (and any 'A' or 'a' allele associated with it) cannot be observed directly, but can be observed indirectly because reduced sampling causes reduced sequencing coverage. Therefore let {A, a, -} be the set of observable alleles, and let {AA, Aa, aa, A-, a-, - -} be the set of observable genotypes.

Let $\mathbf{G} = (\vec{G_1}, ..., \vec{G_N})^\intercal$ be the observable genotypes with vector $\vec{G_s} = (\vec{G_{s,1}}, ..., \vec{G_{s,n}})$ representing the observable genotypes at site s, and $G_{s,i,1}$, $G_{s,i,2}$, $G_{s,i,3}$ representing the number of 'A', 'a', and '-' alleles for individual i. For convenience, we may drop the position subscript s when we are looking at only one locus. Let $\vec{\phi} = (\phi_1, \phi_2, \phi_3)$ be the site allele frequencies for the observable alleles.

Let $\mathbf{D} = (\vec{D_1}, ..., \vec{D_N})^\intercal$ be the data matrix with vector $\vec{D_s} = (\vec{D_{s,1}}, ..., \vec{D_{s,n}})$ representing the read data at site s. Let $\vec{d_s} = (|\vec{D_{s,1}}|, ..., |\vec{D_{s,n}}|)$ be a vector of the number of reads for each sample. Let λ be the site mean coverage for samples with genotypes {(2, 0, 0), (1, 1, 0), (0, 2, 0)} (i.e. for '++' samples).

Let $\mathbf{Z} = (\vec{Z_1}, ..., \vec{Z_N})^\intercal$ be a matrix of variables indicating success (1) or failure (0) of the restriction digest, where $\vec{Z_s} = (Z_{s,1}, ..., Z_{s,n})$. If $Z_{s,i} = 0$, then $d_{s,i} = 0$ regardless of the genotype of the i-th sample. Let δ be the site failure rate.

2 Estimating the site allele frequency

We aim to find $\vec{\phi}$, λ, and δ that maximize $\Pr\{\vec{D}|\vec{\phi}, \lambda, \delta\}$. We have:

$$\begin{aligned}\log \Pr\{\vec{D}, \vec{g}, \vec{z}|\vec{\phi}, \lambda, \delta\} &= \log \prod_{i=i}^{n} \Pr\{\vec{D_i}|g_i, d_i\} \Pr\{d_i|g_i, z_i, \lambda\} \Pr\{g_i|\vec{\phi}\} \Pr\{z_i|\delta\} \\ &= \log \prod_{i=i}^{n} \prod_{j=1}^{d_i} \Pr\{D_{i,j}|g_i\} \Pr\{d_i|g_i, z_i, \lambda\} \Pr\{g_i|\vec{\phi}\} \Pr\{z_i|\delta\}\end{aligned}$$

$$= C + \sum_{i=i}^{n} \log \Pr\{d_i|g_i, z_i, \lambda\} \Pr\{g_i|\vec{\phi}\} \Pr\{z_i|\delta\}$$

Let $m_i = 2 - g_{i,3}$ be the observable ploidy for the i-th individual (i.e. the number of '+' alleles it carries), and let $\vec{r} = (r_1, ..., r_n)$ be a vector of read count normalization factors, where

$$r_i = \frac{\sum_{s=1}^{N} d_{s,i}}{\frac{1}{n}\sum_{j=1}^{n}\sum_{s=1}^{N} d_{s,j}}$$

We assume that the sample read count, d_i follows a negative binomial distribution with mean $\mu = \lambda z_i r_i \frac{m_i}{2}$ and size parameter ψ:

$$\Pr\{d_i|g_i, z_i, \lambda\} = \frac{\Gamma(d_i + \psi)}{\Gamma(d_i + 1)\Gamma(\psi)} \left(\frac{\psi}{\lambda z_i r_i \frac{m_i}{2} + \psi}\right)^{\psi} \left(\frac{\lambda z_i r_i \frac{m_i}{2}}{\lambda z_i r_i \frac{m_i}{2} + \psi}\right)^{d_i}$$

Let $disp(\mu) = a\mu + 1$ be a function chosen to model the dispersion in the normalized read counts, $\vec{d} \circ \vec{r}$. The negative binomial variance is $\mu + \frac{\mu^2}{\psi}$. Therefore ψ is constant across all N sites and $\psi = \frac{1}{a}$.
We assume Hardy-Weinberg equilibrium for the observable genotypes:

$$\Pr\{G_i = g_i|\phi\} = \binom{2}{g_{i,1}, g_{i,2}, g_{i,3}} \phi_1^{g_{i,1}} \phi_2^{g_{i,2}} \phi_3^{g_{i,3}}$$

And the probability of the digest success/failure state for the i-th individual is:

$$\Pr\{Z_i = z_i|\delta\} = (1 - \delta)^{z_i} \delta^{1 - z_i}$$

Given estimates $\vec{\phi_t}, \lambda_t, \delta_t$ at the t-th iteration, the $Q(\vec{\phi}, \lambda, \delta|\vec{\phi_t}, \lambda_t, \delta_t)$ function of EM is:

$$Q(\vec{\phi}, \lambda, \delta|\vec{\phi_t}, \lambda_t, \delta_t) = \sum_{\vec{z}} \sum_{\vec{g}} \Pr\{\vec{g}, \vec{z}|\vec{D}, \vec{\phi_t}, \lambda_t, \delta_t\} \log \Pr\{\vec{D}, \vec{g}, \vec{z}|\vec{\phi}, \lambda, \delta\}$$

$$= C + \sum_{\vec{z}} \sum_{\vec{g}} \prod_{i=1}^{n} \Pr\{g_i, z_i|\vec{D_i}, \vec{\phi_t}, \lambda_t, \delta_t\} \sum_{j} \log \Pr\{d_j|g_j, z_j, \lambda\} \Pr\{g_j|\vec{\phi}\} \Pr\{z_j|\delta\}$$

$$= C + \sum_{i=1}^{n} \sum_{z_i=0}^{1} \sum_{g_i} \Pr\{g_i, z_i|\vec{D_i}, \vec{\phi_t}, \lambda_t, \delta_t\} \log \Pr\{d_i|g_i, z_i, \lambda\} \Pr\{g_i|\vec{\phi}\} \Pr\{z_i|\delta\}$$

$$= C' + \sum_{i=1}^{n} \sum_{z_i=0}^{1} \sum_{g_i} \Pr\{g_i, z_i|\vec{D_i}, \vec{\phi_t}, \lambda_t, \delta_t\} \Big[d_i \log(\lambda) - (d_i + \psi) \log(\lambda z_i r_i \frac{m_i}{2} + \psi) +$$

$$g_{i,1} \log(\phi_1) + g_{i,2} \log(\phi_2) + g_{i,3} \log(\phi_3) + z_i \log(1 - \delta) + (1 - z_i) \log(\delta) \Big]$$

Thus

$$\frac{\partial Q}{\partial \phi_1} = \sum_{i=1}^{n} \sum_{z_i=0}^{1} \sum_{g_i} \Pr\{g_i, z_i | \vec{D}_i, \vec{\phi}_t, \lambda_t, \delta_t\} \frac{g_{i,1}}{\phi_1}$$

$$\frac{\partial Q}{\partial \phi_2} = \sum_{i=1}^{n} \sum_{z_i=0}^{1} \sum_{g_i} \Pr\{g_i, z_i | \vec{D}_i, \vec{\phi}_t, \lambda_t, \delta_t\} \frac{g_{i,2}}{\phi_2}$$

$$\frac{\partial Q}{\partial \phi_3} = \sum_{i=1}^{n} \sum_{z_i=0}^{1} \sum_{g_i} \Pr\{g_i, z_i | \vec{D}_i, \vec{\phi}_t, \lambda_t, \delta_t\} \frac{g_{i,3}}{\phi_3}$$

$$\frac{\partial Q}{\partial \delta} = \sum_{i=1}^{n} \sum_{z_i=0}^{1} \sum_{g_i} \Pr\{g_i, z_i | \vec{D}_i, \vec{\phi}_t, \lambda_t, \delta_t\} \left[\frac{1 - z_i}{\delta} - \frac{z_i}{1 - \delta} \right]$$

$$\frac{\partial Q}{\partial \lambda} = \sum_{i=1}^{n} \sum_{z_i=0}^{1} \sum_{g_i} \Pr\{g_i, z_i | \vec{D}_i, \vec{\phi}_t, \lambda_t, \delta_t\} \left[\frac{d_i}{\lambda} - \frac{z_i r_i \frac{m_i}{2} (d_i + \psi)}{\lambda z_i r_i \frac{m_i}{2} + \psi} \right]$$

and using a first-order Taylor expansion about the point $\lambda = \lambda_t$

$$\begin{aligned} \frac{\partial Q}{\partial \lambda} \approx & \sum_{i=1}^{n} \sum_{z_i=0}^{1} \sum_{g_i} \Pr\{g_i, z_i | \vec{D}_i, \vec{\phi}_t, \lambda_t, \delta_t\} \left[\frac{d_i}{\lambda_t} - \frac{z_i r_i \frac{m_i}{2} (d_i + \psi)}{\lambda_t z_i r_i \frac{m_i}{2} + \psi} \right] \\ & + (\lambda - \lambda_t) \sum_{i=1}^{n} \sum_{z_i=0}^{1} \sum_{g_i} \Pr\{g_i, z_i | \vec{D}_i, \vec{\phi}_t, \lambda_t, \delta_t\} \left[\frac{(z_i r_i \frac{m_i}{2})^2 (d_i + \psi)}{(\lambda_t z_i r_i \frac{m_i}{2} + \psi)^2} - \frac{d_i}{\lambda_t^2} \right] \end{aligned}$$

To calculate the updated parameter estimates we set each partial derivative equal to 0 and solve for ϕ_1, ϕ_2, ϕ_3, λ, and δ. Because of the constraint $\phi_1 + \phi_2 + \phi_3 = 1$ we introduce a Lagrange multiplier:

$$\rho = \sum_{i=1}^{n} \sum_{z_i=0}^{1} \sum_{g_i} \Pr\{g_i, z_i | \vec{D}_i, \vec{\phi}_t, \lambda_t, \delta_t\} (g_{i,1} + g_{i,2} + g_{i,3}) = 2n$$

Thus

$$\phi_{1(t+1)} = \frac{1}{2n} \sum_{i=1}^{n} \sum_{z_i=0}^{1} \sum_{g_i} \Pr\{g_i, z_i | \vec{D}_i, \vec{\phi}_t, \lambda_t, \delta_t\} g_{i,1}$$

$$\phi_{2(t+1)} = \frac{1}{2n} \sum_{i=1}^{n} \sum_{z_i=0}^{1} \sum_{g_i} \Pr\{g_i, z_i | \vec{D}_i, \vec{\phi}_t, \lambda_t, \delta_t\} g_{i,2}$$

$$\phi_{3(t+1)} = \frac{1}{2n} \sum_{i=1}^{n} \sum_{z_i=0}^{1} \sum_{g_i} \Pr\{g_i, z_i | \vec{D}_i, \vec{\phi}_t, \lambda_t, \delta_t\} g_{i,3}$$

and

$$\delta_{t+1} = \frac{1}{n}\sum_{i=1}^{n}\sum_{z_i=0}^{1}\sum_{g_i}\Pr\{g_i, z_i|\vec{D}_i, \vec{\phi}_t, \lambda_t, \delta_t\}(1-z_i)$$

$$\lambda_{t+1} = \lambda_t - \frac{\sum_{i=1}^{n}\sum_{z_i=0}^{1}\sum_{g_i}\left[\frac{d_i}{\lambda_t} - \frac{z_i r_i \frac{m_i}{2}(d_i+\psi)}{z_i r_i \frac{m_i}{2}\lambda_t + \psi}\right]\Pr\{g_i, z_i|\vec{D}_i, \vec{\phi}_t, \lambda_t, \delta_t\}}{\sum_{i=1}^{n}\sum_{z_i=0}^{1}\sum_{g_i}\left[\frac{(z_i r_i \frac{m_i}{2})^2(d_i+\psi)}{(z_i r_i \frac{m_i}{2}\lambda_t+\psi)^2} - \frac{d_i}{\lambda_t^2}\right]\Pr\{g_i, z_i|\vec{D}_i, \vec{\phi}_t, \lambda_t, \delta_t\}}$$

thus

$$\phi_{1(t+1)} = \frac{1}{2n}\sum_{i=1}^{n}\frac{1}{C_i}\sum_{z_i=0}^{1}\sum_{g_i} g_{i,1}\Pr\{\vec{D}_i, g_i, z_i|\vec{\phi}_t, \lambda_t, \delta_t\} \tag{1}$$

$$\phi_{2(t+1)} = \frac{1}{2n}\sum_{i=1}^{n}\frac{1}{C_i}\sum_{z_i=0}^{1}\sum_{g_i} g_{i,2}\Pr\{\vec{D}_i, g_i, z_i|\vec{\phi}_t, \lambda_t, \delta_t\} \tag{2}$$

$$\phi_{3(t+1)} = \frac{1}{2n}\sum_{i=1}^{n}\frac{1}{C_i}\sum_{z_i=0}^{1}\sum_{g_i} g_{i,3}\Pr\{\vec{D}_i, g_i, z_i|\vec{\phi}_t, \lambda_t, \delta_t\} \tag{3}$$

$$\delta_{t+1} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{C_i}\sum_{z_i=0}^{1}\sum_{g_i}(1-z_i)\Pr\{\vec{D}_i, g_i, z_i|\vec{\phi}_t, \lambda_t, \delta_t\} \tag{4}$$

$$\lambda_{t+1} = \lambda_t - \frac{\sum_{i=1}^{n}\frac{1}{C_i}\sum_{z_i=0}^{1}\sum_{g_i}\left[\frac{d_i}{\lambda_t} - \frac{z_i r_i \frac{m_i}{2}(d_i+\psi)}{z_i r_i \frac{m_i}{2}\lambda_t + \psi}\right]\Pr\{\vec{D}_i, g_i, z_i|\vec{\phi}_t, \lambda_t, \delta_t\}}{\sum_{i=1}^{n}\frac{1}{C_i}\sum_{z_i=0}^{1}\sum_{g_i}\left[\frac{(z_i r_i \frac{m_i}{2})^2(d_i+\psi)}{(z_i r_i \frac{m_i}{2}\lambda_t+\psi)^2} - \frac{d_i}{\lambda_t^2}\right]\Pr\{\vec{D}_i, g_i, z_i|\vec{\phi}_t, \lambda_t, \delta_t\}} \tag{5}$$

where

$$\Pr\{\vec{D}_i, g_i, z_i|\vec{\phi}_t, \lambda_t, \delta_t\} = \Pr\{\vec{D}_i|g_i, d_i\}\Pr\{d_i|g_i, z_i, \lambda_t\}\Pr\{g_i|\vec{\phi}_t\}\Pr\{z_i|\delta_t\}$$

and

$$C_i = \sum_{z_i=0}^{1}\sum_{g_i}\Pr\{\vec{D}_i, g_i, z_i|\vec{\phi}_t, \lambda_t, \delta_t\}$$

3 The distribution of site '-' allele count

At site a let $\vec{X}$ be a vector of allele counts, where $X_1 = \sum_i G_{i,1}$ is the number of 'A' alleles, $X_2 = \sum_i G_{i,2}$ is the number of 'a' alleles, and $X_3 = \sum_i G_{i,3}$ is the number of '-' alleles. Define $Y = \sum_i 1 - Z_i$ to be the number of restriction digest failures. The probability that there are no '-' alleles segregating at the site will be our measure of GBS site quality. We aim to calculate

$$\Pr\{X_3 = 0|\vec{D}, \Phi, \Delta\} = \frac{\sum_y \sum_{j,k} \Pr\{\vec{D}|\vec{X} = (j,k,0), y\} \Pr\{\vec{x}|\Phi\} \Pr\{y|\Delta\}}{\sum_y \sum_{j,k,l} \Pr\{\vec{D}|\vec{X} = (j,k,l), y\} \Pr\{\vec{x}|\Phi\} \Pr\{y|\Delta\}} \quad (6)$$

where Φ is the tri-allelic site frequency spectrum, Δ is the site digest failure spectrum. The likelihood can be re-written as

$$\Pr\{\vec{D}|\vec{x}, y\} = \sum_{\vec{z}} \sum_{\vec{g}} \Pr\{\vec{D}|\vec{g}, \vec{d}\} \Pr\{\vec{d}|\vec{g}, \vec{z}, \lambda(\vec{g}, \vec{z})\} \Pr\{\vec{g}|\vec{x}\} \Pr\{\vec{z}|y\} I(\vec{x})$$

where λ depends on $\vec{g}$ and $\vec{z}$

$$\lambda(\vec{g}, \vec{z}) = \frac{\sum_i d_i 2^{\mathbb{1}_1(g_{i,3})}/r_i}{n - \sum_i \mathbb{1}_0((2 - g_{i,3})(1 - z_i))} \quad (7)$$

and the indicator function $I(\vec{X})$ equals 1 if $\vec{X} = (j,k,l)$, and 0 otherwise. Assume that each of the possible configurations $(\vec{g}, \vec{z})$ is equally likely when $\vec{x}$ and y are given (c.f. section 4.2.2 in [**?**]). Thus

$$\Pr\{\vec{D}|\vec{x}, y\} = \sum_{\vec{z}} \sum_{\vec{g}} \prod_i^n \Pr\{\vec{D}_i|g_i, d_i\} \Pr\{d_i|g_i, z_i, \lambda(\vec{g}, \vec{z})\} \frac{\prod_j \binom{2}{g_{j,1}, g_{j,2}, g_{j,3}}}{\binom{2n}{k+l}\binom{n}{y}} I(\vec{x})$$

$$= \frac{1}{\binom{2n}{k+l}\binom{n}{y}} \sum_{\vec{z}} \sum_{\vec{g}} \prod_i^n \Pr\{\vec{D}_i|g_i, d_i\} \Pr\{d_i|g_i, z_i, \lambda(\vec{g}, \vec{z})\} \binom{2}{g_{j,1}, g_{j,2}, g_{j,3}} I(\vec{x}) \quad (8)$$

Direct evaluation of Eq. (**??**) is made difficult by the dependence of λ on $\vec{g}$ and $\vec{z}$, and because there are potentially 12^n combinations of genotypes and digest failure states $(\vec{g}, \vec{z})$. The numerator in Eq. (**??**) can be approximated, however, by the probability of the most likely configuration of $\vec{x}$, y, $\vec{z}$ and $\vec{g}$, given $x_3 = 0$. The denominator in Eq. (**??**) can likewise be approximated by a sum over l of the probability of the most likely configuration given $x_3 = l$. We use a best-first search algorithm to find these configurations [**?**]. We define the initial configuration in the search to be $\vec{g} = ((2,0,0), ..., (2,0,0))$ and $\vec{z} = (1, ..., 1)$.

Algorithm 1 Best-first search for most likely $(\vec{z}, \vec{g})$

function SELECTCONFIG(C_{x_3}, l)
    If l is unspecified, choose one individual, i, and add 1 to $g_{i,2}$, $g_{i,3}$ or z_i, and return the new configuration $(\vec{g}, \vec{z})$. To choose the best individual and configuration calculate the new likelihood after each possible increment according to Eq. (**??**), multiply by the priors for $\vec{x}$ and y, and select the most probable one. If l is specified, perform the same function, but do not increment $g_{i,3}$ (i.e. do not change the number of '-' alleles).
end function
procedure BESTFIRSTSEARCH($(\vec{g}, \vec{z}), l$)
    $\vec{x} \leftarrow (\sum_{i=1}^{n} g_{i,1}, \sum_{i=1}^{n} g_{i,2}, \sum_{i=1}^{n} g_{i,3})$ ▷ Allele counts
    $y \leftarrow \sum_{i=1}^{n} z_i$ ▷ Digest failure counts
    $\lambda \leftarrow \lambda(\vec{g}, \vec{z})$ ▷ See Eq (**??**)
    $C_{x_3} \leftarrow (\vec{g}, \vec{z})$ ▷ Best configuration, given x_3
    $M_{x_3} \leftarrow \sum_{i=1}^{n} \ell(\vec{D}_i, g_i, z_i | \vec{x}, y; \lambda) + \log \Pr\{\vec{x}|\Phi\} \Pr\{y|\Delta\}$ ▷ log of likelihood times prior for configuration C_{x_3}
    while $x_1 > 0$ **do**
        $(\vec{g}, \vec{z}) \leftarrow$ SELECTCONFIG(C_{x_3}, l)
        $\lambda \leftarrow \lambda(\vec{g}, \vec{z})$ ▷ update λ
        $y \leftarrow \sum_{i=1}^{n} z_i$
        $\vec{x'} \leftarrow (\sum_{i=1}^{n} g_{i,1}, \sum_{i=1}^{n} g_{i,2}, \sum_{i=1}^{n} g_{i,3})$
        $M' \leftarrow \sum_{i=1}^{n} \ell(\vec{D}_i, g_i, z_i | \vec{x'}, y; \lambda) + \log \Pr\{\vec{x'}|\Phi\} \Pr\{y|\Delta\}$
        if $M' > M_{x_3}$ **then**
            $\vec{x} \leftarrow \vec{x'}$
            $C_{x_3} \leftarrow (\vec{g}, \vec{z})$
            $M_{x_3} \leftarrow M'$
        else
            return (C_{x_3}, M_{x_3})
        end if
    end while
end procedure

References

[1] Li H. Mathematical notes on SAMtools algorithms. `www.broadinstitute.org/gatk/media/docs/Samtools.pdf`

[2] DePristo M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequence data. *Nature Genetics* 43, 491-498 (2011).