1 Corrected supplementary methods for Vireo paper [1]

In this supplementary section, we will first re-introduce the notation we use, then derive the detailed computation of the lower bound of the variational distribution $\mathtt{L}(q)$ in Eq(7) in the main text, and lastly derive the updates of each variational component in equations Eq(9-11) in the main text. By leveraging the read counts of alternative alleles A and both alleles (namely depth) D from N variants in M cells, Vireo aims to estimate the joint posterior distribution of sample identity Z for each cell j from each sample k, the genotype G for variant i in each sample k, and the corresponding alternative allele rate $\boldsymbol{\theta}$ for each genotype $t \in \{0, 1, 2\}$. As described in the main text, we used multinomial priors for the categorical variables Z and G with hyper-parameters $\boldsymbol{\pi}$ and U, respectively, and by default both take uniform multinomial priors. We used beta priors for the parameter of the alternative allele rate $\boldsymbol{\theta}$, and we took the hyper-parameters $(\alpha_t^{(0)}, \beta_t^{(0)}), t \in \{0, 1, 2\}$ that generally fit well to highly expressed germline variants in standard scRNA-seq data set (not multiplexed). Specifically, the default prior distribution are: $\theta_0 \sim \mathtt{beta}(0.3, 29.7)$, $\theta_1 \sim \mathtt{beta}(3, 3)$, and $\theta_2 \sim \mathtt{beta}(29.7, 0.3)$.

Next, the lower bound $\mathtt{L}(q)$ in Eq(7) can be written as follows

$$\begin{aligned}\mathcal{L}(q) =& \sum_{Z}\sum_{G}\int_{\boldsymbol{\theta}} q(Z, G, \boldsymbol{\theta}) \log\left\{\frac{p(A, D, Z, G, \boldsymbol{\theta})}{q(Z, G, \boldsymbol{\theta})}\right\} \mathrm{d}Z\mathrm{d}G\mathrm{d}\boldsymbol{\theta} \\ =& \mathbb{E}_{G,Z,\boldsymbol{\theta}}[\log p(A, D, Z, G, \boldsymbol{\theta})] - \mathbb{E}_{G,Z,\boldsymbol{\theta}}[\log q(Z, G, \boldsymbol{\theta})] \\ =& \mathbb{E}_{G,Z,\boldsymbol{\theta}}[\log p(A, D|Z, G, \boldsymbol{\theta})] + \mathbb{E}_{Z}[\log p(Z|\boldsymbol{\pi})] + \mathbb{E}_{G}[\log p(G|U)] + \\ & \mathbb{E}_{\boldsymbol{\theta}}[\log p(\boldsymbol{\theta}|\boldsymbol{\alpha}^{(0)}, \boldsymbol{\beta}^{(0)})] - \mathbb{E}_{Z}[\log q(Z)] - \mathbb{E}_{G}[\log q(G)] - \mathbb{E}_{\boldsymbol{\theta}}[\log q(\boldsymbol{\theta})]\end{aligned} \tag{S1}$$

where each part is expressed below.

$$\mathbb{E}_{G,Z,\boldsymbol{\theta}}[\log p(A, D|Z, G, \boldsymbol{\theta})] = \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{K}\sum_{t\in T}\left\{\tilde{r}_{j,k}\tilde{g}_{i,k,t}\left[\log\binom{d_{i,j}}{a_{i,j}} + a_{i,j}\varphi(\tilde{\alpha}_t) + b_{i,j}\varphi(\tilde{\beta}_t) - d_{i,j}\varphi(\tilde{\alpha}_t + \tilde{\beta}_t)\right]\right\} \tag{S2}$$

$$\mathbb{E}_{\boldsymbol{\theta}}[\log p(\boldsymbol{\theta}|\boldsymbol{\alpha}^{(0)}, \boldsymbol{\beta}^{(0)})] = \sum_{t\in T} -(\alpha_t^{(0)} + \beta_t^{(0)} - 2)\varphi(\tilde{\alpha}_t + \tilde{\beta}_t) + (\alpha_t^{(0)} - 1)\varphi(\tilde{\alpha}_t) + (\beta_t^{(0)} - 1)\varphi(\tilde{\beta}_t) - \log(\mathrm{B}(\alpha_t, \beta_t)) \tag{S3}$$

$$\mathbb{E}_{\boldsymbol{\theta}}[\log q(\boldsymbol{\theta}|\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}})] = \sum_{t\in T} -(\tilde{\alpha}_t + \tilde{\beta}_t - 2)\varphi(\tilde{\alpha}_t + \tilde{\beta}_t) + (\tilde{\alpha}_t - 1)\varphi(\tilde{\alpha}_t) + (\tilde{\beta}_t - 1)\varphi(\tilde{\beta}_t) - \log(\mathrm{B}(\tilde{\alpha}_t, \tilde{\beta}_t)) \tag{S4}$$

$$\mathbb{E}_{Z}[\log p(Z|\boldsymbol{\pi})] = \sum_{j=1}^{M}\sum_{k=1}^{K}\{\tilde{r}_{j,k}\log(\pi_k)\}, \quad \mathbb{E}_{G}[\log p(G|U)] = \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{t\in T}\{\tilde{g}_{i,k,t}\log(u_{i,k,t})\} \tag{S5}$$

$$\mathbb{E}_{Z}[\log q(Z)] = \sum_{j=1}^{M}\sum_{k=1}^{K}\{\tilde{r}_{j,k}\log(\tilde{r}_{j,k})\}, \quad \mathbb{E}_{G}[\log q(G)] = \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{t\in T}\{\tilde{g}_{i,k,t}\log(\tilde{g}_{i,k,t})\} \tag{S6}$$

Note, the variables with tilde hat are the estimated parameters otherwise are fixed hyper parameters, including α_t and β_t. Same below.

Then, following the general updating rule in the mean-field variational inference (see main text Eq(8)), we can update the parameters in each component alternately while fixing all other components of the variational distributions and reach the finalized equations Eq(9-11) in the main paper.

First, by using the distributions of genotype G and alternative allele rate $\boldsymbol{\theta}$ that are estimated from a previous step in the iteration, we can analytically update the distribution of the sample

[1] This is a corrected version of the Additional file 1: `https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1865-2#Sec20` by Yuanhua Huang on 28/02/2021 with proofreading from Rongting Huang.

assignment Z by a categorical distribution.

$$\log q^*(Z) = \mathbb{E}_{G,\boldsymbol{\theta}}[\log p(A, D, Z, G, \boldsymbol{\theta})] + \texttt{const.}$$
$$= \sum_{j=1}^{M}\sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{t\in T} Z_{j,k}\left\{\log(\pi_k) + \tilde{g}_{i,k,t}[a_{i,j}\varphi(\tilde{\alpha}_t) + b_{i,j}\varphi(\tilde{\beta}_t) - d_{i,j}\varphi(\tilde{\alpha}_t + \tilde{\beta}_t)]\right\} + \texttt{const.} \tag{S7}$$

where $\varphi(\cdot)$ is the digamma function, the same below. As $q(Z_j)$ for any j follows a multinomial distribution, we can therefore have the updated parameter $r_{j,k}$, namely the probability of cell j from component k as follows,

$$r_{j,k} = \frac{\pi_k \exp \sum_{i=1}^{N}\sum_{t\in T}\left\{\tilde{g}_{i,k,t}[a_{i,j}\varphi(\tilde{\alpha}_t) + b_{i,j}\varphi(\tilde{\beta}_t) - d_{i,j}\varphi(\tilde{\alpha}_t + \tilde{\beta}_t)]\right\}}{\sum_{h=1}^{K}\pi_h \exp \sum_{i=1}^{N}\sum_{t\in T}\left\{\tilde{g}_{i,h,t}[a_{i,j}\varphi(\tilde{\alpha}_t) + b_{i,j}\varphi(\tilde{\beta}_t) - d_{i,j}\varphi(\tilde{\alpha}_t + \tilde{\beta}_t)]\right\}} \tag{S8}$$

Second, with a similar procedure, the analytical updates for the genotype distribution can be written in the form of a categorical distribution as follows,

$$\log q^*(G) = \mathbb{E}_{Z,\boldsymbol{\theta}}[\log p(A, D, Z, G, \boldsymbol{\theta})] + \texttt{const.}$$
$$= \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{t\in T}\sum_{j=1}^{M} G_{i,k,t}\left\{\log(u_{i,k,t}) + \tilde{r}_{j,k}[a_{i,j}\varphi(\tilde{\alpha}_t) + b_{i,j}\varphi(\tilde{\beta}_t) - d_{i,j}\varphi(\tilde{\alpha}_t + \tilde{\beta}_t)]\right\} + \texttt{const.} \tag{S9}$$

where the updated probability of variant i in component k equals to t can be expressed as follows,

$$g_{i,k,t} = \frac{u_{i,k,t} \exp \sum_{j=1}^{M}\left\{\tilde{r}_{j,k}[a_{i,j}\varphi(\tilde{\alpha}_t) + b_{i,j}\varphi(\tilde{\beta}_t) - d_{i,j}\varphi(\tilde{\alpha}_t + \tilde{\beta}_t)]\right\}}{\sum_{h\in T} u_{i,k,h} \exp \sum_{j=1}^{M}\left\{\tilde{r}_{j,k}[a_{i,j}\varphi(\tilde{\alpha}_h) + b_{i,j}\varphi(\tilde{\beta}_h) - d_{i,j}\varphi(\tilde{\alpha}_h + \tilde{\beta}_h)]\right\}}. \tag{S10}$$

Lastly, the analytical updates of the distribution of the alternative allele rate $\boldsymbol{\theta}$ can be expressed in the form of a beta distribution as follows,

$$\begin{aligned}\log q^*(\boldsymbol{\theta}) &= \mathbb{E}_{G,Z}[\log p(A, D, Z, G, \boldsymbol{\theta})] + \texttt{const.}\\ &= \sum_{t\in T}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{K}\{\tilde{r}_{j,k}\tilde{g}_{i,k,t}[a_{i,j}\log(\theta_t) + (d_{i,j} - a_{i,j})\log(1-\theta_t)]\} + \\ &+ \sum_{t\in T}\left[(\alpha_t^{(0)} - 1)\log(\theta_t) + (\beta_t^{(0)} - 1)\log(1-\theta_t)\right] + \texttt{const}\\ &= \log(\texttt{beta}(\theta_t|\tilde{\alpha}_t, \tilde{\beta}_t)).\end{aligned} \tag{S11}$$

where the parameters for this beta distribution are

$$\tilde{\alpha}_t = \alpha_t^{(0)} + \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{K}\{\tilde{r}_{j,k}\tilde{g}_{i,k,t}a_{i,j}\}\,;\tilde{\beta}_t = \beta_t^{(0)} + \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{K}\{\tilde{r}_{j,k}\tilde{g}_{i,k,t}(d_{i,j} - a_{i,j})\}\,. \tag{S12}$$

Now, by updating these parameters iteratively, we can achieve the maximized lower bound of $\texttt{L}(q)$, and equivalently the minimized $\mathrm{KL}(q(Z, G, \boldsymbol{\theta})||p(Z, G, \boldsymbol{\theta}|A, D))$.