Supplementary Information for "Assessing allele specific expression across multiple tissues from RNA-seq read data"

Matti Pirinen, Tuuli Lappalainen, Noah A. Zaitlen,
GTEx Consortium, Emmanouil T. Dermitzakis, Peter Donnelly,
Mark I. McCarthy and Manuel A. Rivas

July 17, 2014

S1 Priors for the groups

Our goal is to classify observed allelic read counts at each site and each tissue into one of the three groups. We want the groups to represent (i) no ASE (group $\mathcal{N}$) where both alleles are (almost) equally expressed, (ii) strong ASE (group $\mathcal{S}$) where one of the alleles is expressed very little if at all, and (iii) moderate ASE (group $\mathcal{M}$) that represents everything in between the first two groups. In the main text we propose the following priors for the reference allele read count frequencies of these groups:

$$\begin{aligned}\theta(\mathcal{N}) &\sim \mathrm{Beta}(2000, 2000),\\ \theta(\mathcal{M}) &\sim \frac{1}{2}\,\mathrm{Beta}(36, 12) + \frac{1}{2}\,\mathrm{Beta}(12, 36),\\ \theta(\mathcal{S}) &\sim \frac{1}{2}\,\mathrm{Beta}(80, 1) + \frac{1}{2}\,\mathrm{Beta}(1, 80).\end{aligned}$$

Figure S1 shows the densities of these priors together with the regions of the read count frequency space where each of the group is dominating the other two by at least a factor of 10. We see that our choices for prior parameters satisfy our goal since:

- (i) group $\mathcal{N}$ dominates in the small region (0.47,0.53) around 0.5,
- (ii) group $\mathcal{S}$ dominates at extreme frequencies of ≤ 0.07 and ≥ 0.93,
- (iii) group $\mathcal{M}$ dominates at nearly all the remaining frequencies: (0.10,0.46) and (0.54,0.90).

Truncated prior. Our implementation allows to truncate each Beta-distribution on a user-specified interval in order to make the support of the different groups non-overlapping. This is useful especially when one-sided priors are used. For example, if we are studying non-sense mediated decay and want that ASE is called only if the reference allele shows read count frequency over 0.5, we could use the following one-sided truncated priors:

$$\begin{aligned}\theta(\mathcal{N}) &\sim \mathrm{Beta}(2000, 2000) I_{[0,0.52)},\\ \theta(\mathcal{M}) &\sim \mathrm{Beta}(36, 12) I_{[0.52,0.95)},\\ \theta(\mathcal{S}) &\sim \mathrm{Beta}(80, 1) I_{[0.95,1.0]},\end{aligned}$$

where $I_{[a,b)}$ denotes truncation of the distribution on the interval $[a, b)$.

Independent tissues. Our implementation allows relaxing the assumption that all tissues in one group have exactly the same reference allele read count frequency. This is done by modeling each tissues-specific θ_s as an independent draw from the corresponding prior for the group. This is useful when we have informative data with a large number of reads for each tissue and the tissues within one group do not have exactly the same value for θ. On the other hand, with a small number of reads per tissue the basic GTM (without independence assumption) is our default choice because it allows borrowing strength across the tissues in the same group.

S2 Gibbs sampler for GTM

We use a Gibbs sampler algorithm to explore the posterior distribution of configuration $\overline{\gamma} \in \{\mathcal{N}, \mathcal{M}, \mathcal{S}\}^T$, where T is the number of tissues. We denote by π_H the (fixed) prior probability of heterogeneity states. (In the main text we use $\pi_H = 0.25$.) As in the main text, we denote by $\boldsymbol{y}$ the observed read count data at one site and across all tissues.

We fix the number of iterations $n_{\text{iter}} = 2,000$ and the number of burn-in iterations $n_{\text{burn}} = 10$ and run the following Gibbs sampler.

1. Initialize $\overline{\gamma} = \left(\gamma_1^{(0)}, \ldots, \gamma_T^{(0)}\right)$ with a random configuration.

2. Repeat for $t = 1, 2, \ldots, (n_{\text{burn}} + n_{\text{iter}})$:
   For $s = 1, 2, \ldots, T$:

   - Compute probability vector

     $$p_s^{(t)} = \left(p_s^{(t)}(\mathcal{N}), p_s^{(t)}(\mathcal{M}), p_s^{(t)}(\mathcal{S})\right),$$

     where for each group $G \in \{\mathcal{N}, \mathcal{M}, \mathcal{S}\}$,

     $$p_s^{(t)}(G) \propto f\left(\boldsymbol{y}; \overline{\gamma}_s^{(t)}(G)\right) \pi\left(\overline{\gamma}_s^{(t)}(G)\right).$$

     Here $f(\boldsymbol{y}; \overline{\gamma})$ is the beta-binomial marginal likelihood for the data given the group indicators $\overline{\gamma}$ and the prior distributions for θ parameters of each group; $\pi(\overline{\gamma})$ is the prior probability of the configuration $\overline{\gamma}$, which is determined by π_H together with the distance $d(\overline{\gamma})$; and

     $$\overline{\gamma}_s^{(t)}(G) = \left(\gamma_1^{(t)}, \ldots, \gamma_{s-1}^{(t)}, G, \gamma_{s+1}^{(t-1)}, \ldots, \gamma_T^{(t-1)}\right).$$

- Generate

$$\gamma_s^{(t)} \sim \begin{cases} \mathcal{N}, & \text{with probability } p_s^{(t)}(\mathcal{N}) \\ \mathcal{M}, & \text{with probability } p_s^{(t)}(\mathcal{M}) \\ \mathcal{S}, & \text{with probability } p_s^{(t)}(\mathcal{S}). \end{cases}$$

S3 Hierarchical model (GTM*)

We extend the grouped tissue model (GTM) defined in the main text to the case where many variants with similar properties (such as protein truncating variants) are analyzed simultaneously. We add one level of hierarchy to the model by introducing vector $\boldsymbol{\pi} = (\pi_N, \pi_M, \pi_S, \pi_{H0}, \pi_{H1})$ that determines the proportion of variants in each of the five states defined in the main text (N=NOASE, M=MODASE, S=SNGASE, H0=HET0 and H1=HET1). Denote by $\mathbf{y}^{(\ell)} = \left(\left(y_{1,1}^{(\ell)}, y_{2,1}^{(\ell)} \right), \ldots, \left(y_{1,T_\ell}^{(\ell)}, y_{2,T_\ell}^{(\ell)} \right) \right)$ the reference (1) and non-reference (2) allele counts for variant ℓ over available T_ℓ tissue types, and by $\mathbf{y} = (\mathbf{y}^{(\ell)})_{\ell=1}^{L}$ all the data over all L variants.

This extension, called GTM*, is the following model, over variants $\ell = 1, \ldots, L$ and tissues $s = 1, \ldots, T_\ell$:

$$\begin{aligned} \theta^{(\ell)}(\mathcal{N}) &\sim \text{Beta}(2000, 2000) \\ \theta^{(\ell)}(\mathcal{M}) &\sim \frac{1}{2}\text{Beta}(36, 12) + \frac{1}{2}\text{Beta}(12, 36) \\ \theta^{(\ell)}(\mathcal{S}) &\sim \frac{1}{2}\text{Beta}(80, 1) + \frac{1}{2}\text{Beta}(1, 80) \\ \boldsymbol{\pi} &\sim \text{Dirichlet}(1, 1, 1, 1, 1) \\ \left(\overline{\gamma^{(\ell)}} = \overline{\gamma} \right) |\boldsymbol{\pi} &\sim \begin{cases} \overline{\gamma} = \text{NOASE}, & \text{with probability } \pi_N \\ \overline{\gamma} = \text{MODASE}, & \text{with probability } \pi_M \\ \overline{\gamma} = \text{SNGASE}, & \text{with probability } \pi_S \\ \overline{\gamma} \in \text{HET0}, & \text{with probability } \frac{\pi_{H0}}{(T_\ell - \lceil T_\ell/3 \rceil) h_0(d(\overline{\gamma}))} \\ \overline{\gamma} \in \text{HET1}, & \text{with probability } \frac{\pi_{H1}}{\lfloor T_\ell/2 \rfloor h_1(d(\overline{\gamma}))} \end{cases} \\ y_{1,s}^{(\ell)} \left| \gamma_s^{(\ell)}, \theta^{(\ell)} \right. &\sim \text{Bin}\left(y_{1,s}^{(\ell)} + y_{2,s}^{(\ell)}; \theta^{(\ell)} \left(\gamma_s^{(\ell)} \right) \right), \end{aligned}$$

where $d(\overline{\gamma})$ is the distance of configuration $\overline{\gamma}$ from homogeneity (see the main text) and $h_0(d)$ is the number of configurations belonging to state HET0 and having distance d from homogeneity, (similarly $h_1(d)$ for HET1 configurations). The values $(T_\ell - \lceil T_\ell/3 \rceil)$ and $\lfloor T_\ell/2 \rfloor$ are the maximum distances among all configurations in HET0 and HET1, respectively. In other words, we directly model the probability of the three homogeneous states by π_N, π_M and π_S and we distribute the probability (π_{H0} and π_{H1}) among each heterogeneous state uniformly with respect to the distance, and also uniformly among the configurations with the same distance. This model is slightly different from our original GTM as the probabilities π_{H0} for HET0 and π_{H1} for HET1 states have been separated from each other. In settings where we want to follow the exact prior structure of GTM, our implementation also makes it possible to run GTM* parameterized with a single heterogeneity probability $\pi_H = \pi_{H0} + \pi_{H1}$. This mode can be invoked by simply specifying the Dirichlet prior for $\boldsymbol{\pi}$ with four parameters instead of five.

We have implemented GTM* through a Gibbs sampler, which follows the algorithm given above for GTM with an additional Gibbs update for $\boldsymbol{\pi}$ with

$$\boldsymbol{\pi} \sim \text{Dirichlet}(n_{\text{NOASE}} + 1, n_{\text{MODASE}} + 1, n_{\text{SNGASE}} + 1, n_{\text{HET0}} + 1, n_{\text{HET1}} + 1),$$

where each n_S denotes the number of variants currently assigned to state S.

An advantage of GTM* over variant specific analyses using GTM is that the posterior distribution of $\boldsymbol{\pi}$ is available. We expect that the posterior of $\boldsymbol{\pi}$ using GTM* is more accurate than averaging the variant specific posteriors from GTM, and, importantly, properly accounts for uncertainty in these estimates. However, when read counts are not very small, (say we have 30 or more reads per tissue per variant), we expect that the two approaches give fairly similar estimates. We next give some comparisons between GTM* and GTM approaches to inference about $\boldsymbol{\pi}$.

S4 Comparing GTM and GTM*

First we analysed the simulated data of the main text with GTM* (1,000 data sets per $T = 5, 10, 30$ tissues and $n = 10, 50$ reads and each of the nine scenarios). We present the posterior expectation of $\boldsymbol{\pi}$ from GTM* in Figure S2, together with the original GTM results from the main text, which average the individual state posteriors across the 1,000 data sets.

The results show that with 50 reads GTM* correctly infers the true state even in scenarios which were not completely solved by GTM. Also for 10 reads, GTM* improves the proportion estimate compared to GTM in most cases. A notable exception is scenario 5, which according to GTM* is almost completely in HET1 state whereas the data sets were simulated with a HET0 state. This phenomenon happens because the prior probability of HET1 state has been separated from HET0 in GTM* and thus, under GTM*, any one tissue-specific configuration in HET1 state has a higher prior probability than a tissue-specific configuration in HET0 state (as there are fewer such configurations in HET1 than in HET0). Thus, if data have little information to distinguish between a configuration in HET0 and another one in HET1, then GTM* tends to prefer the HET1 state. On the other hand, GTM gives the same prior probability for every tissue-specific configuration, whether it belongs to HET0 or HET1 state. When the latter property of the model is considered more appropriate, one can run our GTM* implementation parameterized with combined heterogeneity probability $\pi_H = \pi_{H0} + \pi_{H1}$ by simply specifying the Dirichlet prior for $\boldsymbol{\pi}$ with four parameters instead of five. More importantly, when the amount of information increases, the small differences between the two prior specifications become insignificant, as shown by the results with 50 reads in Figure S2.

The above comparison shows how much GTM* estimation of $\boldsymbol{\pi}$ differs from GTM in an extreme case where all the variants analysed belong to the same underlying state. More realistically, variants would represent different states, and in that case we expect that the difference between GTM* and GTM decreases. To compare the approaches on such a setting we randomly subsampled from among our simulated data sets for $T = 10$ tissues and for both 10 and 50 read counts per tissue, 50 collections of 200 variants with the following proportions of states: 10% NOASE (from scenario 1), 30% MODASE (from scenario 2), 40% HET0 (from scenario 8) and 20% HET1 (from scenario 9). The 50 point estimates of the proportions by GTM* and GTM together with the true values are show in Figure S3.

For 10 reads per tissue, both GTM* and GTM underestimate the proportion of heterogeneous variants and overestimate the homogeneous one. This is in line with the principle that with insufficient information we prefer homogeneous states. GTM* is notably more accurate than GTM with MODASE and HET1 states while the opposite is true with NOASE and HET0 states.

For 50 reads per tissue, both approaches give accurate estimates for practical purposes, but GTM* is more accurate than GTM.

We conclude that when many variants are available and we are interested in the state proportions $\boldsymbol{\pi}$, we should apply GTM* to estimate $\boldsymbol{\pi}$ together with its uncertainty. However, GTM is both an essential building block for

GTM* and an important model on its own, since it is quick to run, easy to understand and requires data on only a single variant. For these reasons, we have devoted the main text of this work to GTM.

S5 Combinatorics of configurations

Consider T tissues and a configuration $\overline{\gamma} = (\gamma_1, \ldots, \gamma_T)$ where each $\gamma_s \in \{\mathcal{N}, \mathcal{M}, \mathcal{S}\}$. All together there are 3^T configurations of which 3 are homogeneous and $3^T - 3$ are heterogeneous. Total number of HET1 configurations is $2^T - 2$ and hence the number of HET0 configurations is $3^T - 2^T - 1$.

Consider the configurations at distance d from homogeneity, where

$$d = T - \max\{\ell_N, \ell_M, \ell_S\} \text{ with } \ell_G = \#\{s : \gamma_s = G\}$$

being the number of tissues in group $G \in \{\mathcal{N}, \mathcal{M}, \mathcal{S}\}$. Denote the three counts (ℓ_N, ℓ_M, ℓ_S) in ascending order by $i \leq d - i \leq T - d$ whence

$$\max\{0, 2d - T\} \leq i \leq \lfloor d/2 \rfloor.$$

The number of heterogeneous configurations at distance d is

$$h(d) = \sum_{i=\max\{0,2d-T\}}^{\lfloor d/2 \rfloor} \binom{T}{i \; (d-i) \; (T-d)} \frac{3!}{(4 - \#\{i, d-i, T-d\})!}$$

where the first term in the sum is the multinomial coefficient telling how many ways there are to split T tissues among the given group counts, and the second term multiplies by 6, 3 or 1 according to whether all three counts are different, exactly two of the counts are equal to each other or all of the counts are equal.

The number of HET1 configurations at distance $d = 1, \ldots, \lfloor T/2 \rfloor$ is

$$h_1(d) = \binom{T}{d} \frac{2!}{(3 - \#\{d, T-d\})!}.$$

Using the above derived formulae, the number of HET0 configurations at distance d is $h_0(d) = h(d) - h_1(d)$.

S6 The GTEx Consortium

Laura Barker
Margaret Basile
Alexis Battle
Joy Boyer
Debra Bradbury
Jason P. Bridge
Amanda Brown
Robin Burges
Christopher Choi
Deborah Colantuoni
Nancy Cox
Emmanouil T. Dermitzakis
Leslie K. Derr
Michael J. Dinsmore
Kenyon Erickson
Barbara A. Foster
Timothee Flutre
Eric R. Gamazon
Gad Getz
Bryan M. Gillard
Roderic Guigo
Kenneth W. Hambright
Pushpa Hariharan
Rick Hasz
Hae K. Im
Scott Jewel
Ellen Karasik
Manolis Kellis
Susan Koester
Daphne Koller
Anuar Konkashbaev
Tuuli Lappalainen
Roger Little
Jun Liu
Edmund Lo
John T. Lonsdale

Chunrong Lu
Daniel G. MacArthur
Julian B. Maller
Yvonne Marcus
Deborah C. Mash
Mark I. McCarthy
Bernadette Mestichelli
Mark Miklos
Jean Monlong
Magboeba Mosavel
Michael T. Moser
Sara Mostafavi
Dan L. Nicolae
Jonathan Pritchard
Liqun Qi
Kimberly Ramsey
Manuel A. Rivas
Barnaby E. Robles
Daniel C. Rohrer
Mike Salvatore
Michael Sammeth
Saboor Shad
John Seleski
Laura A. Siminoff
Matthew Stephens
Jeff Struewing
Timothy Sullivan
Susan Sullivan
David Tabor
Mehran Taherian
Jorge Tejada
Gary F. Temple
Jeffrey A. Thomas
Alexander W. Thomson
Denee Tidwell
Heather M. Traino
Zhidong Tu
Dana R. Valley

Simona Volpi
Gary D. Walters
Xiaoquan Wen
Wendy Winckler
Shenpei Wu
Nancy Young
Jun Zhu

Figure S1: The top panel shows the densities for the prior distributions of the reference allele for the three groups: $\mathcal{N}$, $\mathcal{M}$ and $\mathcal{S}$. The lower panel shows the regions where each of the densities is dominating the other two by a factor of at least 10 and 95% highest probability regions for each of the prior distributions.

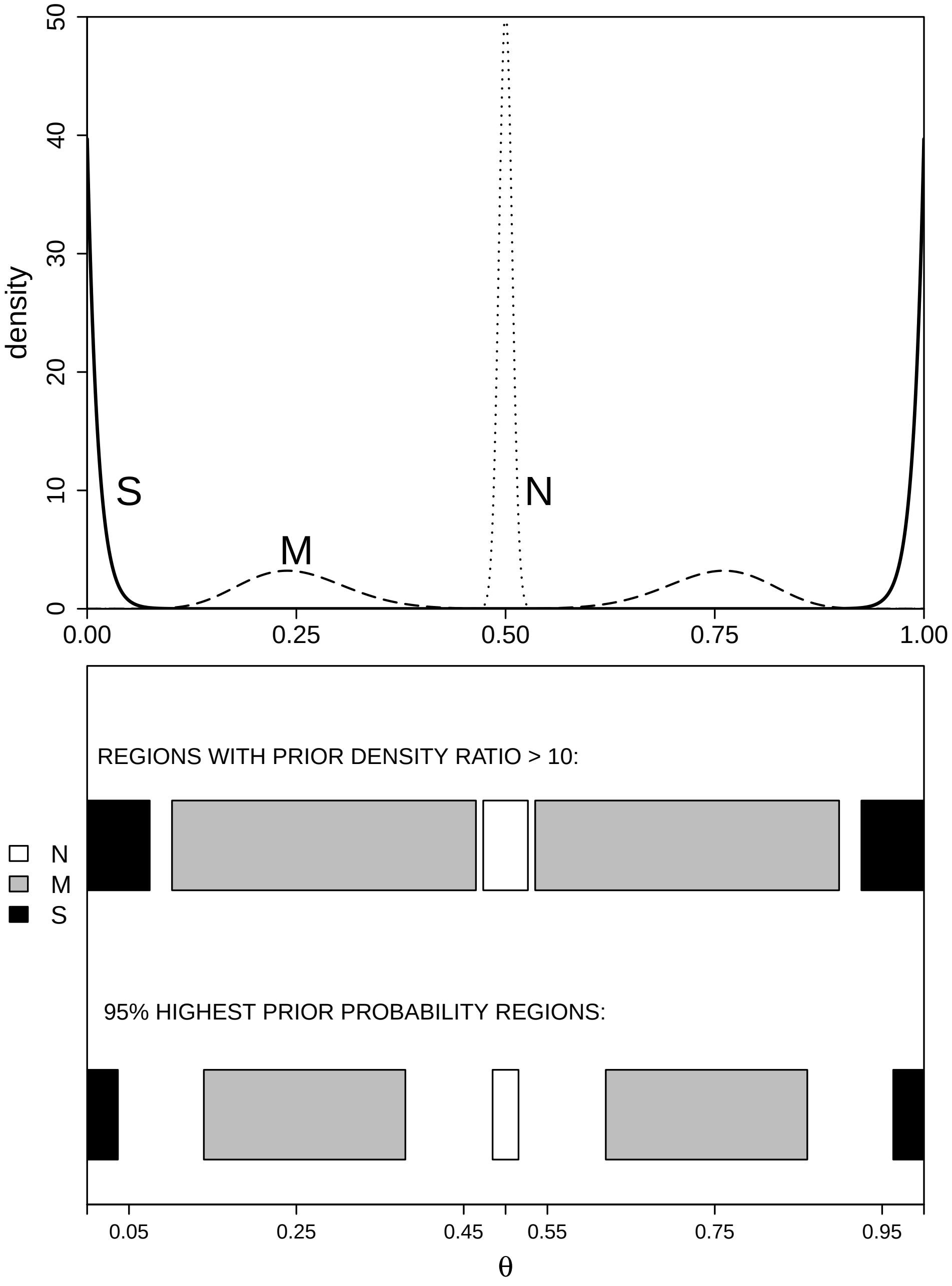
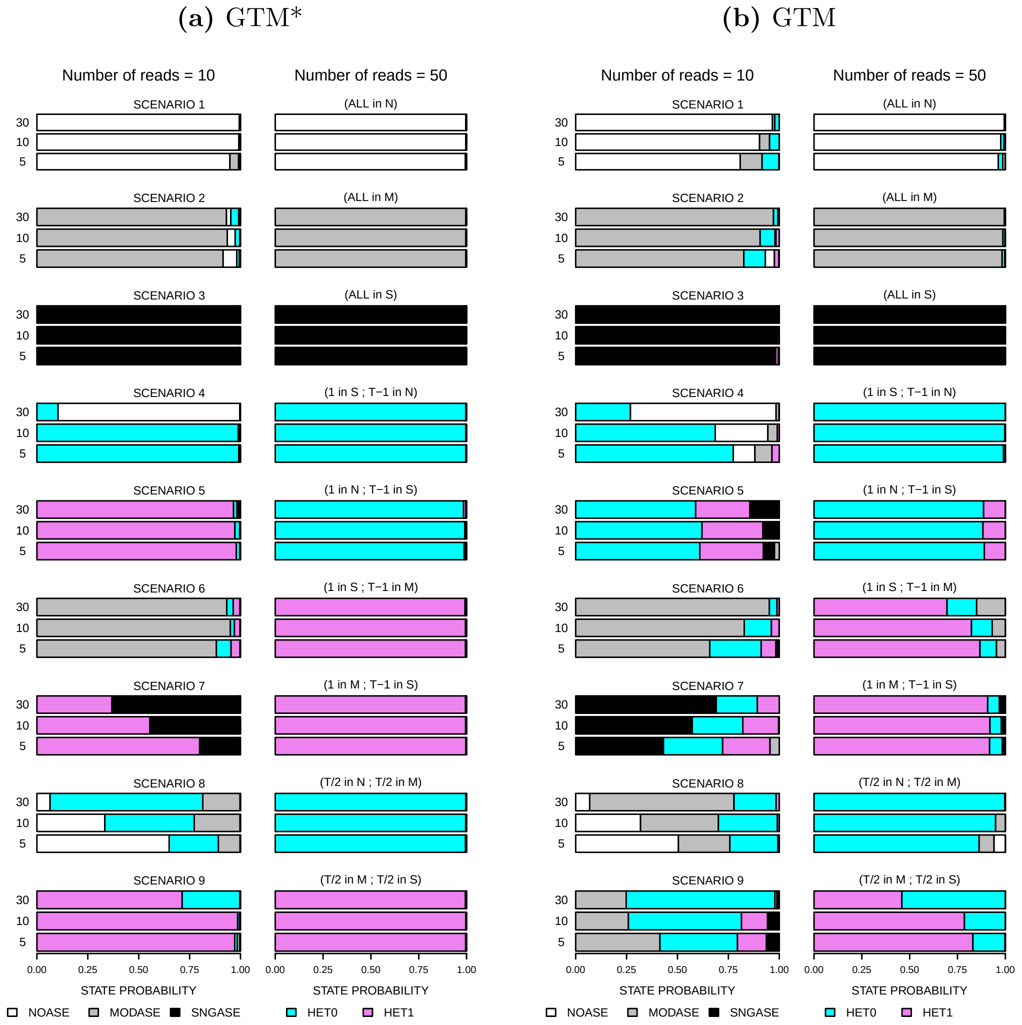
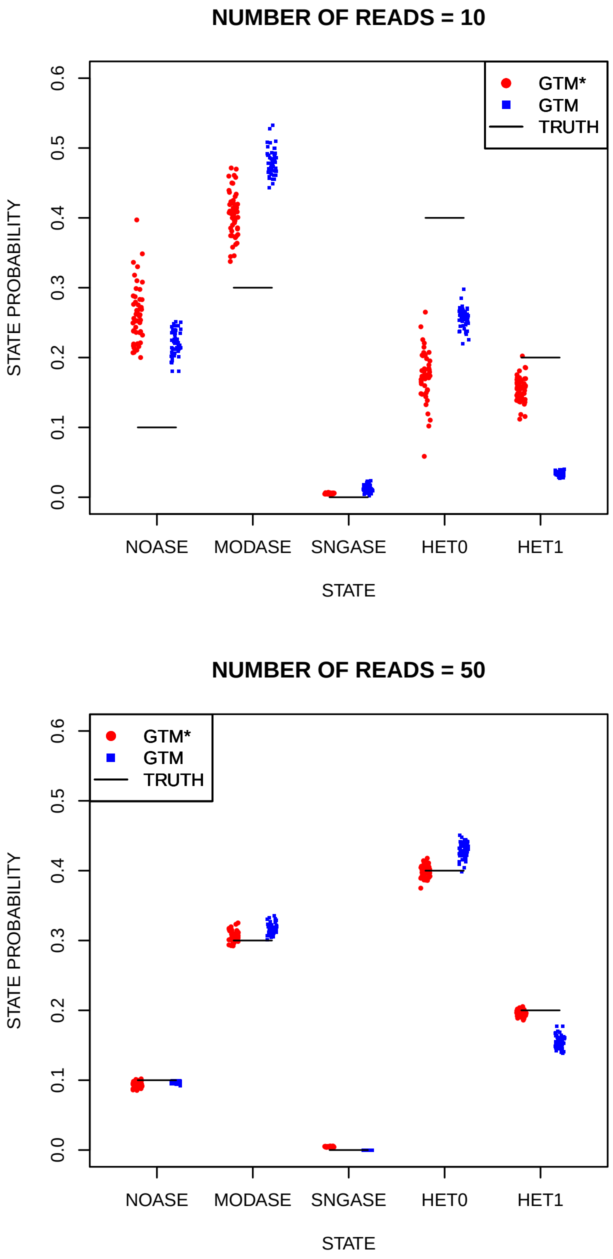


Figure S2: Results of GTM* and GTM on the simulated data sets of the main text. Each of the nine simulation scenarios (Table 1 in the main text) is represented by three numbers of tissues (5, 10, 30) and two values for number of reads (10, left columns and 50, right columns). Each bar is divided into five colors (map given at the bottom) according to the posterior expectation of the state probabilities, $\boldsymbol{\pi}$, for GTM* and the (average) posterior probability of the five states for GTM.


(a) GTM*

(b) GTM

Figure S3: Fifty collections of 200 variants with 10 tissue types were analysed and the estimates of the proportions of variants in each of the five states are shown for GTM* (posterior expectation of $\boldsymbol{\pi}$) and for GTM (average over variant specific state posteriors). The true proportions are shown with horizontal lines. The analyses were done for both 10 and 50 reads per tissue per variant.



NUMBER OF READS = 10

NUMBER OF READS = 50