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

Contents

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:

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

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).

Motivation for allowing some deviation (0.47,0.53) from 0.5 for no ASE group is given by Supplementary Figure S28 of Lappalainen et al. (2013) that shows that such amount of variation can be caused by technical biases.

For choosing the priors for the groups $\mathcal{M}$ and $\mathcal{S}$, our principle has been that if the proportion of RNA from one of the alleles is less than 5% then strong ASE is called, whereas if both alleles have proportions $\geq 10\%$ then moderate ASE is called.

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:

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

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 tissue-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 (typically $\mathcal{M}$) 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.

Only two groups. Our implementation allows GTM to be run without the $\mathcal{S}$ group. This is useful when we do not want to make a difference between variants showing strong or moderate ASE. See section "Modest ASE" below for details.

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.

In the examples of this paper, we use the number of iterations $n_{\text{iter}} = 2,000$ and the number of burn-in iterations $n_{\text{burn}} = 10$, since these values gave good results in Table 2 of the main text. In general, one should test the convergence of the Gibbs sampler by running it several times on the same data and comparing the results.

Here is our 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.

A further extension of GTM* could have an additional level of hierarchy to learn the parameters of the beta distributions. This could be useful when many variants and/or many tissues per variant were present but, contrary to our current approach, would not lead to directly comparable proportions across different sets of variants. For example, the data driven frequency range for, say, group $\mathcal{S}$ could be quite different for nonsynonymous variants from that of protein truncating variants.

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 Repeatability

We consider 5 samples from the GEUVADIS data set (Lappalainen et al., 2013) for which we had access to 7 or 8 technical replicates of the RNA-seq on lymphoblastoid cell lines (LCLs). For each individual, we applied GTM* to several thousands of heterozygous synonymous SNVs by treating each technical replicate of the RNA-seq counts as a separate "tissue" in the model. In Table S1 we report individual-specific estimates of the proportion of the variants that showed heterogeneity. The maximum point estimate for heterogeneity (HET0+HET1) across the five individuals is low (<1.5%) and HET1 proportion alone is negligible (<0.2%). These results give confidence that the method is not overestimating the heterogeneity.

IND	SNVs	READS	HET0	HET1
NA20527	2664	41	0.013 (0.007,0.024)	0.001 (0.0,0.003)
NA19095	3550	41	0.013 (0.006,0.032)	0.001 (0.0,0.003)
NA06986	2337	34	0.014 (0.006,0.038)	0.001 (0.0,0.002)
HG00117	2493	39	0.009 (0.002,0.031)	0.001 (0.0,0.003)
HG00355	2696	42	0.011 (0.005,0.038)	0.002 (0.001,0.004)

Table S1: IND, id in GEUVADIS data; SNVs, the number of heterozygous synonymous SNVs considered; READS, median read count; HET0, estimated proportion of SNVs in HET0 class (95% region); HET1, estimated proportion of SNVs in HET1 class (95% region);

S6 Varying read counts

Smaller read count means more uncertainty about the actual allelic ratio. We use the following example to see how this transforms to uncertainty about the group membership.

We consider 5 tissues of which one has a small read count (10) and the other four have the same higher read count (30 or 50). We simulated 100 replicates for each homogeneous state, NOASE, MODASE and SNGASE, that were characterised by non-reference allele frequencies of 0.5, 0.25 and 0.01, respectively.

In Table S2 we compare the posterior probabilities of belonging to the true underlying group both for the lower read count tissue and for the higher read count tissues. We see that, as expected, the lower read count leads to more uncertainty about the group assignment, but for these read counts, that cover a realistic range of counts, the difference remains small in practice.

STATE	10	30	10	50
NOASE	0.933	0.952	0.980	0.990
MODASE	0.985	0.992	0.984	0.999
SNGASE	0.998	1.000	0.998	0.999

Table S2: STATE, state used for data simulation where all 5 tissue had the same allelic ratio; The columns 2 to 5 show probability of belonging to the correct group (averaged over 100 simulations) for the tissue with read count mentioned in the top row (10, 30 or 50). The columns 2 and 3 are from the scenario where one tissue had 10 reads and the other 4 tissues had 30 reads; the columns 4 and 5 are from the scenario where one tissue had 10 reads and the other 4 tissues had 50 reads. The columns 3 and 5 show median value over the 4 tissues with the corresponding read count.

S7 Comparing GTM and Q-statistic (Figure S5)

In Figure 3 of the main text we show ROC curves comparing GTM and Q-statistic to assess heterogeneity in two situations where homogeneous and heterogeneous data sets are only slightly different: 1 tissue out of 30 belongs to a separate group and read count is small (10 reads) for all tissues. In Supplementary Figure S5 we show more details about these comparisons as well as extend them to another case where underlying configurations of homogeneous and heterogeneous data sets are much more different from each other. Next we explain each panel of Figure S5 and continue with a discussion of these results.

Figure S5 has 3 columns and 5 rows. Each **column** shows results of comparison between GTM and Q-statistic applied to assess heterogeneity in 2,000 data sets from Table 1 of the main text where number of tissues $T = 30$ and number of reads $n_s = 10$ for all tissues s. In each column, 1,000 data sets represent homogeneous scenarios and 1,000 data sets represent heterogeneous scenarios. The comparisons are between scenarios 1 and 4 (left column), 3 and 7 (middle column) and 2 and 7 (right column).

- **Row 1:** ROC curves for detecting heterogeneity where the heterogeneity measures are posterior probability of heterogeneity (HET0+HET1) for GTM and empirical P-value for the Q-statistic. The values in legends are percent concordance measure, that is the same as area and the curve (AUC) measure, where larger value means better ability to separate between the heterogeneous and homogeneous configurations.

- **Row 2:** QQ-plots for GTM, plotting the ordered heterogeneity probabilities of 1,000 homogeneous data sets on x-axis against the ordered heterogeneity probabilities of 1,000 heterogeneous data sets on y-axis. The diagonal line is y=x.

- **Row 3:** QQ-plots for the Q-statistics, plotting the ordered heterogeneity $-\log_{10}$(P-values) of 1,000 homogeneous data sets on x-axis against the ordered heterogeneity $-\log_{10}$(P-values) of 1,000 heterogeneous data sets on y-axis. The diagonal line is y=x.

- **Row 4:** Rank concordance plots, plotting the ranks of 1,000 homogeneous data sets among all 2,000 data sets for GTM on x-axis against the ranks of the same data sets for the Q-statistics on y-axis. The diagonal line is y=x.

- **Row 5:** Rank concordance plots, plotting the ranks of 1,000 heterogeneous data sets among all 2,000 data sets for GTM on x-axis against the ranks of the same data sets for the Q-statistics on y-axis. The diagonal line is y=x.

Discussion on Figure S5. The ROC curves show that in the left and middle columns, where the difference between homogeneous and heterogeneous configurations is small (1 tissue out of 30), GTM is slightly better in detecting heterogeneity. On the rightmost column, where homogeneity is represented by moderate ASE (scenario 2) and heterogeneity by 29 tissues having strong ASE and 1 tissue moderate ASE (scenario 7), the situation is more complex. From the ROC curve we see that GTM detects better the first half of the heterogeneous data sets than the Q-statistic, but for the latter half the opposite is true and overall the Q-statistic has clearly higher percent concordance than GTM. This can be explained by the following two properties. (1) In around half of the data sets of scenario 7 with 10 reads, the tissue having moderate ASE has ≤ 2 reads from the non-reference allele. (2) GTM is built in such a way that it prefers homogeneity over heterogeneity when only a little information is available, and therefore GTM assigns only a small heterogeneity probability for these data sets. We claim that this is a reasonable behaviour in general: When 29 tissues consistently show strong ASE, we want to be conservative in calling a configuration heterogeneous unless we truly have clear evidence for the last tissue deviating from strong ASE. And such a clear evidence would require more than 10 reads in these cases where the point estimate for non-reference allele frequency is ≤ 0.2. Furthermore, these heterogeneity probabilities are small also relative to most of the heterogeneity probabilities of data sets from scenario 2. This is because, with read count of only 10, data sets from scenario 2 cannot be completely excluded from belonging to HET0 state (Supp Fig. S4).

We note that with higher read counts the scenario 7 data sets are more consistently assigned to the true state of HET1 (Figure S4), which enhances percent concordance of GTM on analogous comparisons with higher read counts.

S8 Modest ASE

Our main interest in this work has been to charaterize effects of variants that can lead to strong ASE. In other applications, where only modest ASE effects are expected, we suggest two changes to our default models. First, we adjust the prior for moderate ASE group $\mathcal{M}$ to better cover the region of modest ASE. For example, we could focus on the region $(0.40, 0.45) \cup (0.55, 0.60)$ by setting prior of θ in $\mathcal{M}$ group to $\frac{1}{2}\text{Beta}(2300, 1700) + \frac{1}{2}\text{Beta}(1700, 2300)$ whose lower tail has expectation $1700/4000 = 0.425$ and standard deviation of 0.008. Second, we remove the strong ASE group $\mathcal{S}$ from the model. This removes many heterogeneous configurations from the configuration space, and hence increases the prior probability of those heterogeneous configurations that remain. We did the following experiment to compare this model to our default GTM as introduced in the main text.

For 4 read counts per tissue (50, 100, 200, 500), 11 scenarios were generated where from 0 to 10 out of 10 tissues belonged to the ASE group and the remaining belonged to NO ASE group ($\theta = 0.5$). The parameter θ for the ASE group was picked from a uniform distribution on $(0.40, 0.45)$. The two versions of GTM were applied and the posterior probabilities of the states are shown (averaged over 100 data sets for each scenario) in Supp Fig. S6.

We see that the model for modest ASE calls more of the data sets as HET0 and MODASE than the default GTM. This is as expected since

- the prior of the ASE group in the modest ASE model better captures the region of modest ASE than the default prior of GTM that was designed for more prominent ASE effects;
- each heterogeneous configuration between groups $\mathcal{N}$ and $\mathcal{M}$ has higher prior probability in the modest ASE model than in the default GTM.

Even though the modest ASE model seems overall better suited for these data sets than the default GTM, a downside is that the modest ASE model will call more heterogeneity than the default GTM also in those cases where no heterogeneity is actually present (see the bars for 0 and 10 tissues showing ASE in Fig. S6).

S9 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 $h_0(d) = h(d) - h_1(d)$.

Only two groups. When there are only two groups, $\mathcal{N}$ and $\mathcal{M}$, then there are 2^T configurations of which 2 are homogeneous and $2^T - 2$ heterogeneous. The number of heterogeneous configurations at distance d is $2\binom{T}{d}$ when $1 \leq d < T/2$ and $\binom{T}{d}$ when $d = T/2$.

S10 Supplementary Figures

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.

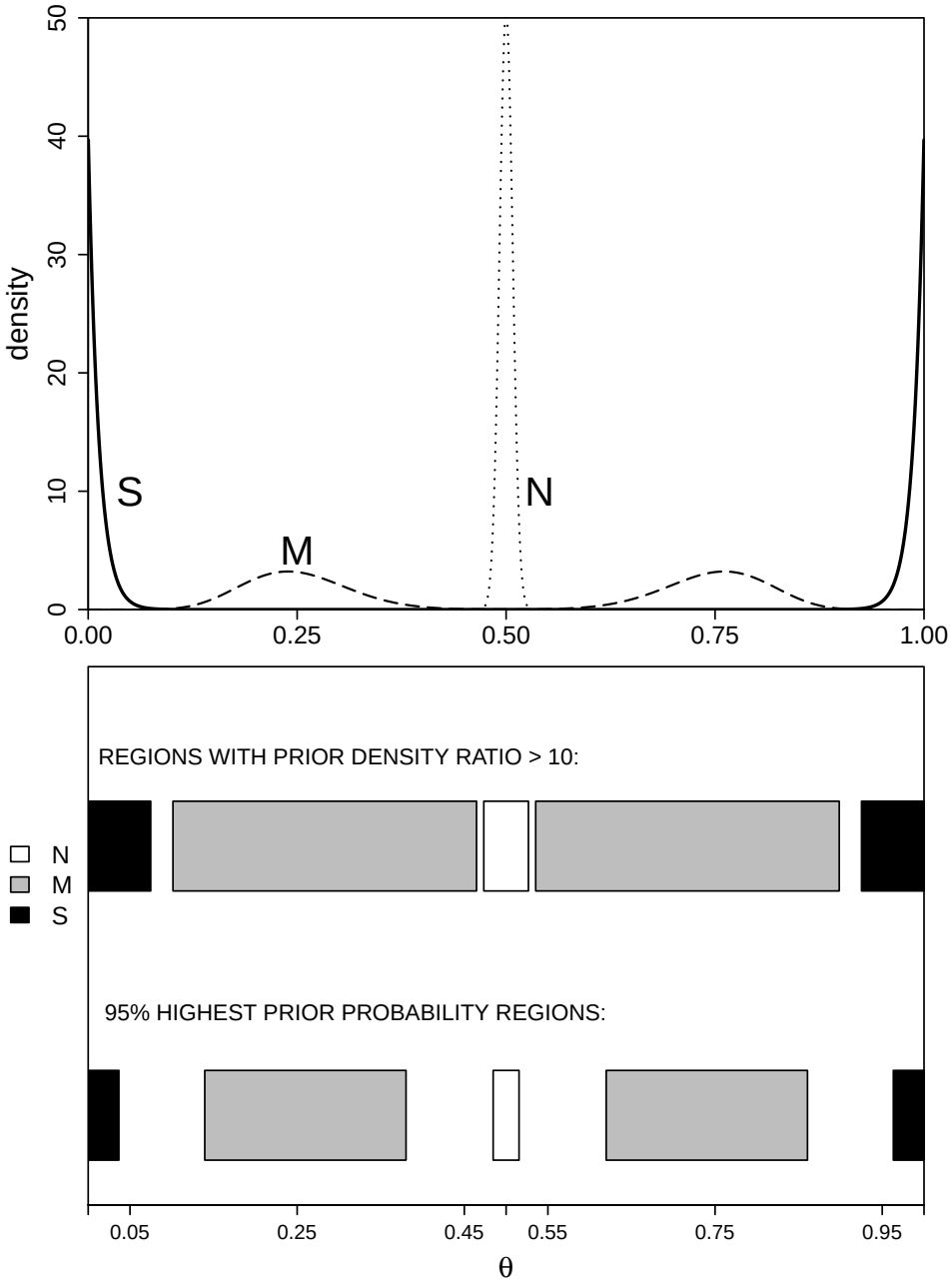

Figure S2: Results of GTM* and GTM on the simulated data sets of the main text. Each of the nine simulation scenarios (Table 1 in the main text) is represented by three numbers of tissues (5, 10, 30) and two values for number of reads (10, left columns and 50, right columns). Each bar is divided into 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.

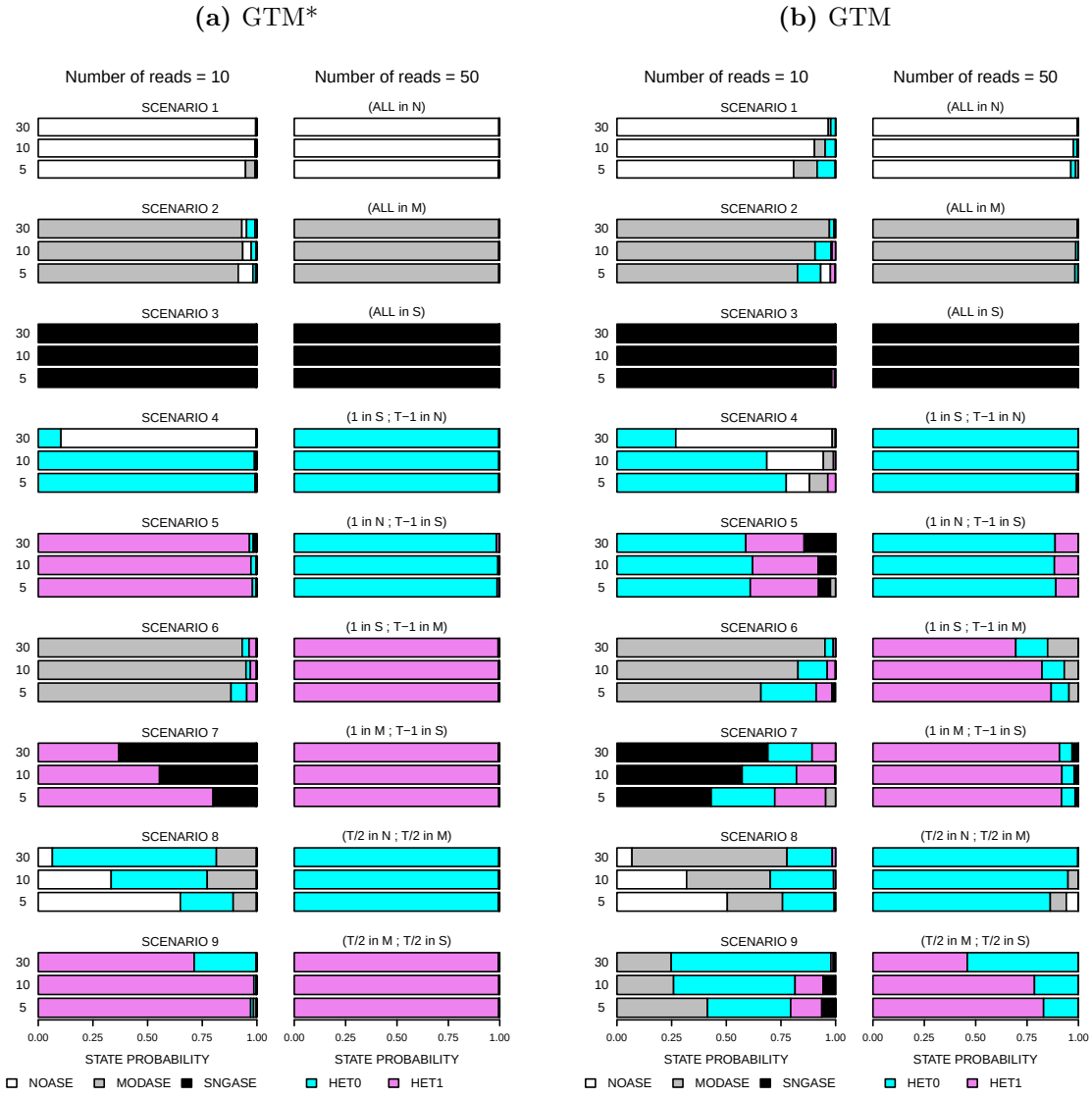

Figure S3: Fifty collections of 200 variants with 10 tissue types were analysed and the estimates of the proportions of variants in each of the five states are shown for GTM* (posterior expectation of $\boldsymbol{\pi}$) and for GTM (average over variant specific state posteriors). The true proportions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

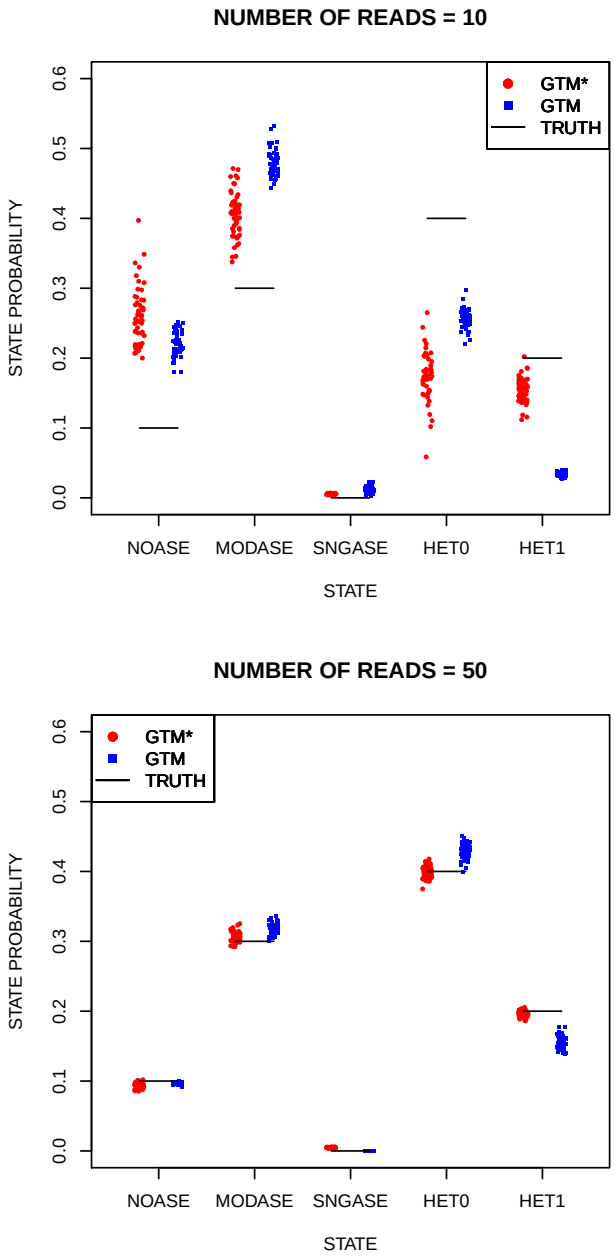

Figure S4: Results of GTM in Figure 2 of the main text (read counts 10 and 50) complemented with analogous simulations for read counts 30 and 200. Each of the nine simulation scenarios (Table 1 in the main text) is represented by three numbers of tissues (5, 10, 30) and four values for number of reads (10, 30, 50 and 200). Each bar is divided into five colors (map given at the bottom) according to the (average) posterior probability of the five states for GTM.

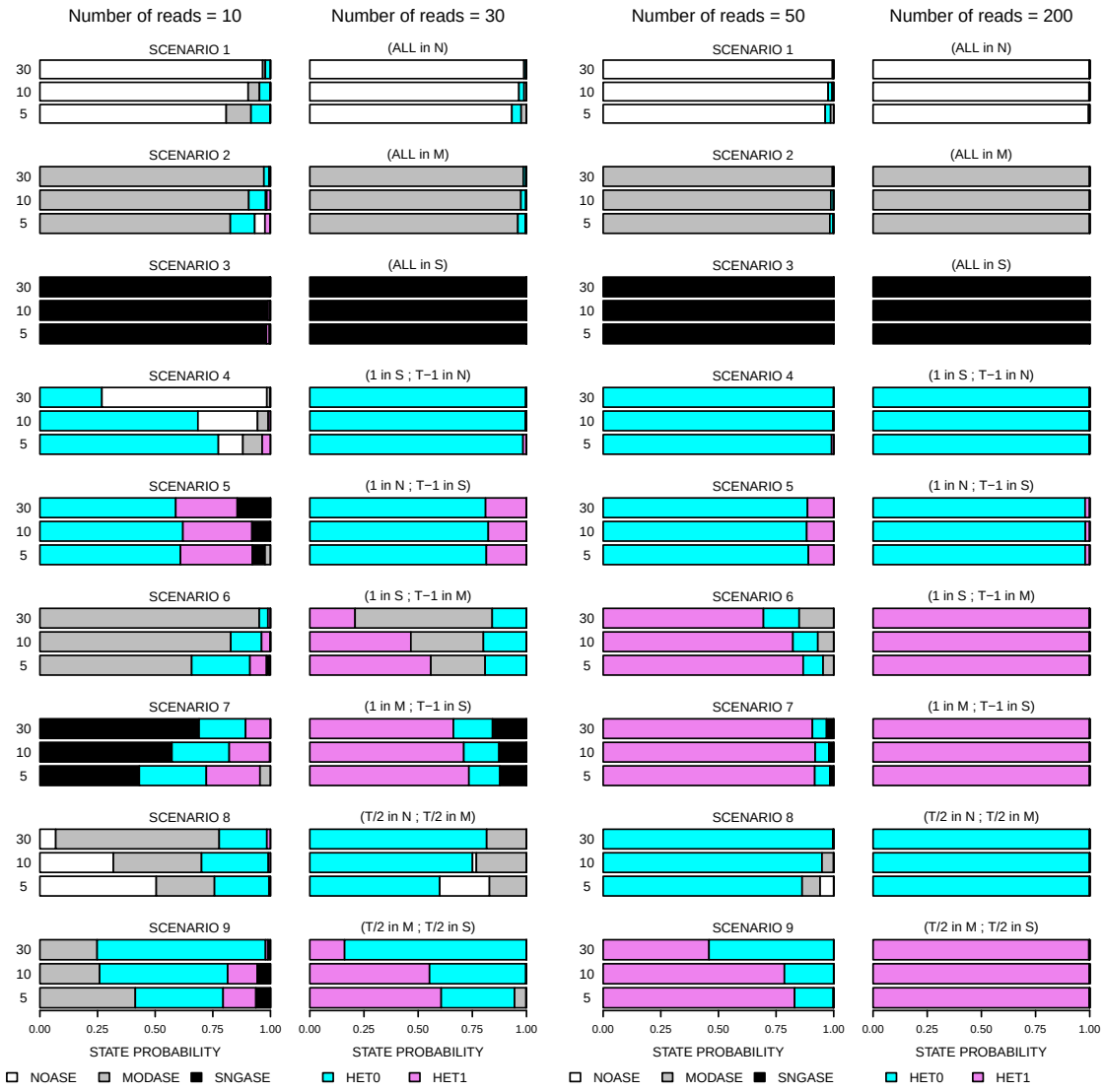

Figure S5: For legend, see section S7.

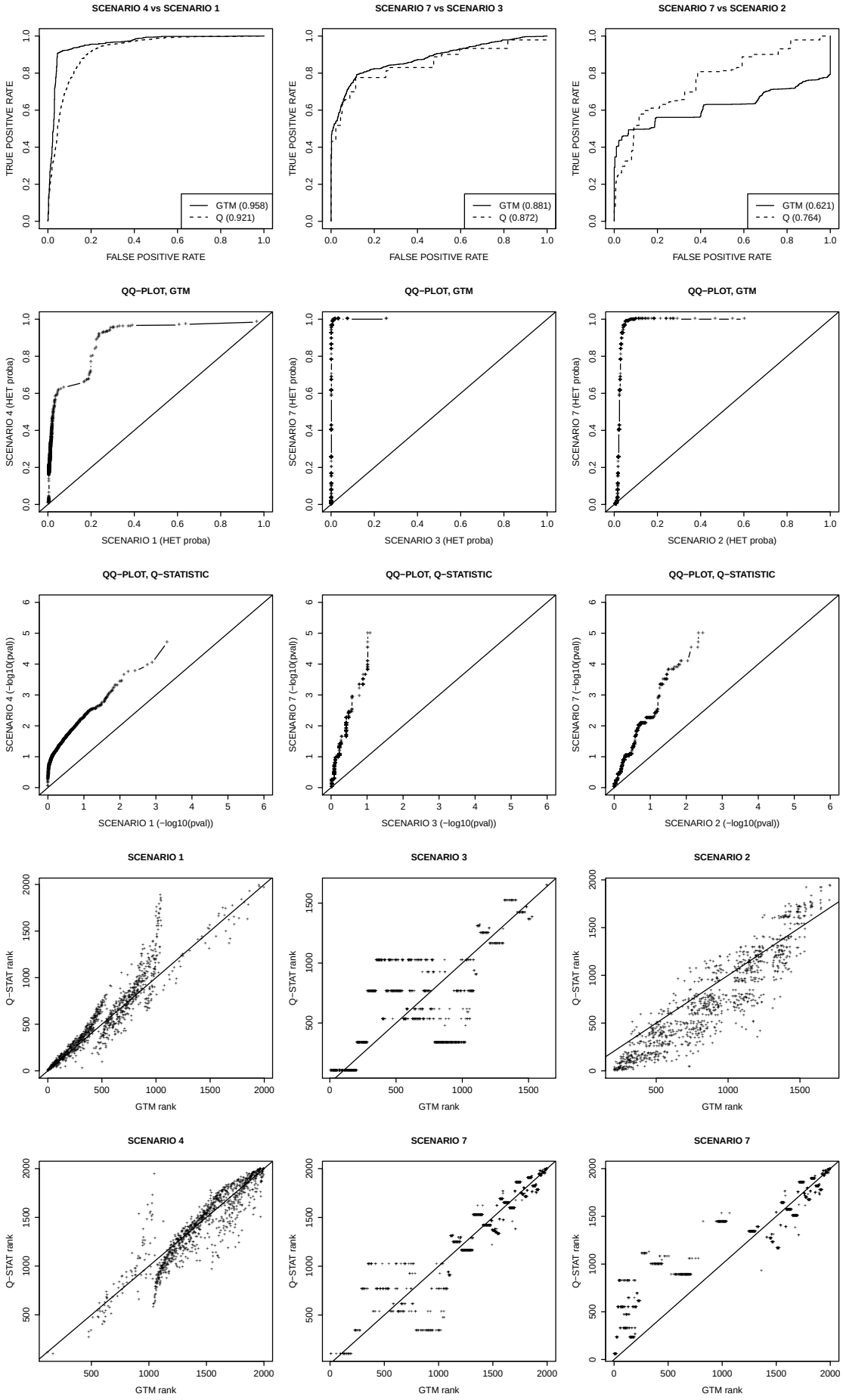

Figure S6: Modest ASE effects. For 4 read counts per tissue (50, 100, 200, 500), 11 scenarios were generated where from 0 to 10 out of 10 tissues belonged to the ASE group and the remaining belonged to NO ASE group ($\theta = 0.5$). The parameter θ for the ASE group was picked from a uniform distribution on $(0.40, 0.45)$. Two versions of GTM were applied and the posterior probabilities of the states are shown (averaged over 100 data sets for each scenario). Left-side bars show results for our default GTM model. Right-side bars show results when GTM was resticted to have only two groups (no $\mathcal{S}$ group included) and when the prior for the ASE group was Beta$(2300, 1700)$ to reflect modest ASE effects.