Binomial Models of Reference Read Counts

Attila Gulyás-Kovács

February 22, 2016

1 Preliminaries

We have $i = 1, ..., I$ individuals, $g = 1, ..., G$ genes and $v = 1, ..., V$ polymorphic (SNP) sites that occur at least one (i, g) pair in heterozygous form. For each (i, g) we test hypothesis $\mathcal{H}_0$ against $\mathcal{H}_1$:

$$(i, g) \in \mathcal{H}_h : \begin{cases} (i, g) \text{ biallelically expressed} & \text{if } h = 0 \\ (i, g) \text{ monoallelically expressed} & \text{if } h = 1 \end{cases} \tag{1}$$

Assuming only one alternative allele at each v, let A_v denote the read count for the alternative allele and n_v the count of all reads. Thus, the read count for the reference allele is $n_v - A_v$. In the context of all models to follow, we will consider n_v as observed and fixed parameter while A_v as an observed *random variable* with unknown mean (expected value) $\mathrm{E}[A_v]$.

We define

$$Z_v = \begin{cases} A_v & \text{if } \mathrm{E}[A_v] \geq n_v - \mathrm{E}[A_v] \\ n_v - A_v & \text{otherwise.} \end{cases} \tag{2}$$

In words, Z_v is the read count for the allele with the higher expected read count.

Since the mean counts in Eq. 2 are unknown, Z_v is a *latent (unobserved) variable* in the sense that we don't know for sure whether Z_v corresponds to the reference or the alternative allele. But it will be much more straight-forward to express all models in Section 2 using the *expected fraction* $p_v = \mathrm{E}[Z_v/n_v]$ instead of the expected fraction of A_v in n_v.

Thus Z_v is latent; but any statistical analysis (parameter inference and hypothesis testing/classification) must be based on *observed variables.* To that end we could use A_v; but to be consistent with the previous work of the MAE project, we define

$$\begin{aligned} Y_v &= \max(Z_v, n_v - Z_v) && (3) \\ Y_{ig} &= \{Y_v\}_{v \in (i,g)}, \qquad n_{ig} = \{n_v\}_{v \in (i,g)} && (4) \\ Y &= \{Y_{ig}\}_{ig}, \qquad n = \{n_{ig}\}_{ig}. && (5) \end{aligned}$$

The random variable Y_v[1] is the *higher read count* at polymorphic site v. The notation $v \in (i, g)$ means all heterozygous sites v in individual i and gene g.

[1] The symbol H was used previously in the MAE project but conventions in statistics and information theory as well as other considerations motivated me to replace it with Y.

Much of the previous analysis of the MAE project was based on S_{ig}

$$S_{ig} = \frac{\sum_{v \in (i,g)} Y_v}{\sum_{v \in (i,g)} n_v} = \frac{||Y_{ig}||_1}{||n_{ig}||_1}. \tag{6}$$

The scalar S_{ig} aggregates the vectors Y_{ig} and n_{ig} and, as we will see, the information lost in that aggregation has an impact on all statistical analysis based on the models below.

2 Models

2.1

The most basic model

- fixed expected fraction Z_v/n_v

$$\begin{aligned}
P\left((i,g) \in \mathcal{H}_h\right) &= \pi_h \quad \textit{a priori} & (7)\\
\pi_h & \quad \text{fixed} & (8)\\
Z_v &\sim \text{Binom}(p_h, n_v) \quad v \in (i,g),\ (i,g) \in \mathcal{H}_h & (9)\\
p_h & \quad \text{fixed} & (10)
\end{aligned}$$

2.2

Uncertain expected fraction Z_v/n_v.

$$\begin{aligned}
Z_v &\sim \text{Binom}(p'_h, n_v) \quad v \in (i,g),\ (i,g) \in \mathcal{H}_h & (11)\\
p'_h &\sim \text{Beta}(\mu_h, \nu_h) & (12)
\end{aligned}$$

To obtain Model 2.1, take $\mu_h = p_h$ from Eq. 9-10 and let $\nu_h \to \infty$.

2.3

Influence of explanatory variables x_i on expected fraction Z_v/n_v.

$$\begin{aligned}
p'_h &\sim \text{Beta}(\mu'_{hi}, \nu_h) & (13)\\
\text{logit}(\mu'_{hi}) &= x_i \beta_h & (14)
\end{aligned}$$

Model 2.2 is obtained by taking $\beta_{h,0} = \mu_h$ from Eq. 12 and setting $\beta_{h,1} = ... = \beta_{h,p-1} = 0$.

2.4

Prior to observing the RNA-seq data there is evidence Ev_{ig} for/against $(i,g) \in \mathcal{H}_h$ such as

- distance of g from known imprinted genes
- cis-eQTLs of (i,g)
- confidence in calling (i,g) heterozygous at v

$$P\left((i,g) \in \mathcal{H}_h \,|\, \mathrm{Ev}_{ig}\right) \quad = \quad \pi'_h(\mathrm{Ev}_{ig}), \tag{15}$$

where π'_h is some function of the evidence Ev_{ig}. For instance, Ev_{ig} may be gene g's distance $d(g)$ from the nearest imprinted gene, and $\pi'_h(\mathrm{Ev}_{ig}) = \gamma + \exp(-d(g)/\tau)$, where τ is a length constant measured in bases. To obtain Model 2.3 let pi'_h be constant by setting $\pi'_h = \pi_h$ from Eq. 7-8 regardless of the evidence.

3 Likelihood functions

We will derive the likelihood function[2] f of the full model under the basic Model 2.1. Extensions to more complex models will follow. f fill be derived piece-wise based on the set of functions $\{f_{ig}\}_{ig}$, where each f_{ig} in turn is derived from $\{f_v\}_{v \in (i,g)}$. For all models, f will be required to infer parameters based on the observed value y of random variable Y and on the observed n. Classification of some (i,g) pair (or g in regression models) will require only f_{ig} (or f_g in regression models) because of the independencies of the model at hand.

$$f_v(y_v|n_v, p_h) \quad = \quad \frac{1}{2}\binom{n_v}{y}\left[p_h^{y_v}(1-p_h)^{n_v-y} + p_h^{n_v-y_v}(1-p_h)^{y}\right] \tag{16}$$

$$f_{ig}(y_{ig}|n_{ig}, p_h) \quad = \quad \prod_{v \in (i,g)} f_v(y_v|n_v, p_h) \tag{17}$$

$$f(y|n, p_0, p_1, \pi_1) \quad = \quad \prod_{i,g}\left[f_{ig}(y_{ig}|n_{ig}, p_1)\pi_1 + f_{ig}(y_{ig}|n_{ig}, p_0)(1-\pi_1)\right] \tag{18}$$

If we want to base inference on the scalar S_{ig} instead of the vector Y_{ig}, we need to derive likelihood functions for S_{ig} using Eq. 17. Let $\mathcal{S} = \{(i,g) : n_{ig}s_{ig} = y_{ig}\}$, that is the set of all (i,g) pairs leading to the observed s_{ig}. Then the likelihood functions h_{ig} and h'_{ig} for S_{ig} can be expressed in terms of $\{f_{ig}\}_{(i,g)\in\mathcal{S}}$:

$$h_{ig}(s_{ig}|n_{ig}, p_h) \quad = \quad \sum_{(i,g)\in\mathcal{S}} f_{ig}(y_{ig}|n_{ig}, p_h) \tag{19}$$

$$h'_{ig}(s_{ig}|p_h) \quad = \quad \sum_{(i,g)\in\mathcal{S}} f_{ig}(y_{ig}|n_{ig}, p_h)\, q_{ig}(n_{ig}|p_h). \tag{20}$$

The difference between h_{ig} and h'_{ig} is whether or not we condition the distribution of S_{ig} on the observed n_{ig}. If we don't take advantage of the observations on n_{ig} (Eq. 20), we must then treat it as a random variable and specify a distribution for it, say q_{ig}. In either case we need *some* kind of information on n_{ig}. This holds regardless we want to use h_{ig} (or h'_{ig}) in simulations, in parameter inference or in classification with error control.

4 Inference of parameters

5 Classification

[2] The notion of probability mass/density function $f(y|\theta)$ of statistic y given parameters θ is so closely related to the likelihood function $L(\theta; y)$ of θ given y that the two are often used interchangeably in the literature. Here I also use f to refer to both kinds of function.