Binomial Models of Reference Read Counts

Attila Gulyás-Kovács

March 14, 2016

1 Introduction

1.1 Goals

1.2 Improvement relative to previous approach

All inference is now based *directly on read counts* Y while previous inferences used the S statistic derived from Y (for classification) and LOI_R further derived from S (for estimation of regression parameters). Because both derivations discard some information S is not sufficient[1] and LOI_R is even less so. Thus, the direct approach using Y is more informative.

Now *global* probability models are formulated. These model the entire data (all genes and individuals) jointly and so facilitate statistical inference at all relevant levels: at the level of individual–gene pairs, of genes, of individuals, and at the level of all genes and individuals collectively. Previously only a set of local binomial models was considered, which would have allowed probabilistic inference only at the irrelevant level of heterozygous sites but not at the biologically relevant higher levels because of the missing connections among the local models.

Specifically, the above two improvements afford now

- assessment of the structure of variation both among genes and among individuals
- a better assessment of the impact of explanatory variables (more power, less bias,...)
- estimation of the expected fraction of monoallelically expressed genes per individual (and the expected variation, if any, among individuals)
- bi/monoallelic classification with error control

2 Data and local models

2.1 The modeled data: read counts

We have $i = 1, ..., I$ individuals, $g = 1, ..., G$ genes and $v = 1, ..., V$ polymorphic (SNP) sites. With the notation $v \in (i, g)$ we will express that site v is in gene g and it is heterozygous in individual i, and we distinguish v from w if $w \in (j, g)$ and if $i \neq j$ even if both v and w map to the same site in a reference genome (meaning they are homologous).

[1] https://en.wikipedia.org/wiki/Sufficient_statistic

We assume only one alternative allele at each site v, and write Y_v to denote the read count of the alternative allele at site v. We also define

$$\begin{aligned} Y_{ig} &= \{Y_v\}_{v \in (i,g)}, \qquad n_{ig} = \{n_v\}_{v \in (i,g)} & (1) \\ Y &= [Y_{ig}], \qquad n = [n_{ig}], & (2) \end{aligned}$$

where $[Y_{ig}]$ denotes a matrix whose rows are indexed by $i = 1, ..., I$ and columns by $g = 1, ..., G$. Moreover, we have an $I \times R$ design matrix $X = [x_{ir}],\ r = 0, ..., R-1$ whose columns are explanatory variables a.k.a. regressors except for the 0th column, whose entries $x_{i0} = 1$ for all i. All proposed inferences in this article will be based on Y and X.

TODO: Much of the previous inferences of the MAE project were based on the statistic $S = [S_{ig}]$. The connection between S and Y can be drawn by introducing the "higher read count" $H_v = \max(Y_v, n_v - Y_v)$ and writing $S_{ig} = \left(\sum_{v \in (i,g)} H_v\right) \times \left(\sum_{v \in (i,g)} n_v\right)^{-1}$. 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.

2.2 Local model of allelic exclusion

The probability model presented here is *local* in the sense that the global models in Section 3 will be based on this local model or a very similar one. However, even though we call this model local, it describes allelic exclusion and read counts at the biologically relevant level of (i, g) pairs in contrast with the previously considered binomial model restricted to the lower, and irrelevant, level of sites v.

We introduce (allelic) *exclusion state* $\theta_{ig}$for any given (i, g) pair such that biallelic expression of gene g in individual i is indicated by $\theta_{ig} = 0$ and monoallelic by $\theta_{ig} = 1$. Suppose p_{ig} is the expected fraction of transcripts[2] from the maternal chromosome and $1 - p_{ig}$ for the paternal chromosome, and let $q_{ig} = \max(p_{ig}, 1 - p_{ig})$ implying that $q_{ig} \geq 1/2$.

We regard q_{ig} as the single direct determinant of allelic exclusion (Figure TODO): if q_{ig} is near 1/2 we call (i, g) biallelically expressed, whereas if q_{ig} is near 1 we classify (i, g) monoallelic. Formally, let $\mathcal{P}_0 = [1/2, p')$ and $\mathcal{P}_1 = [p'', p''']$ disjoint subintervals of $[1/2, 1]$ so that $1/2 \leq p' \leq p'' \leq p''' \leq 1$.

Then we *define* exclusion state of (i, g) as follows:

$$q_{ig} \equiv \max(p_{ig}, 1 - p_{ig}) \in \begin{cases} \mathcal{P}_0 & \Leftrightarrow \theta_{ig} = 0, \text{ biallelic} \\ \mathcal{P}_1 & \Leftrightarrow \theta_{ig} = 1, \text{ monoallelic.} \end{cases} \tag{3}$$

There are some complications with this definition. First, p_{ig} is generally unknown and must be inferred from the data, which results in uncertainty about not only its exact value but also whether $p_{ig} \geq 1/2$ and therefore $q_{ig} = p_{ig}$, or else $< 1/2$ and therefore $q_{ig} = 1 - p_{ig}$. Let $\phi_{ig} = 1$ indicate the former event and $\phi_{ig} = 0$ the latter with prior probability κ and $1 - \kappa$, respectively. Thus κ quantifies the tendency of the paternal allele to be excluded. In the present models κ is not specific to individuals and genes but it is straight forward to extend the models in that direction at the expense of introducing many more parameters. It may be reasonable to set $\kappa = 1/2$.

[2] The word "expected" implies a probability distribution for maternal transcripts. This can be either binomial if the total number of transcripts is fixed, or else Poisson. In the latter case p_{ig} is to be interpreted as the relative transcription rate on the maternal chromosome.

Several further complications arise because our data consists of reads instead of full-length transcripts. We assume that the read count Y_v for the alternative allele at polymorphic site v is binomially distributed with parameters n_v (the total read counts) and p_v. However, read counts have been confounded by various measurement errors but we assume that they are proportional to allele specific transcription rates. This allows us to write $p_v = p_{ig}$ given the random event that the alternative allele is on the maternal chromosome; we denote that event with $\psi_v = 1$. Otherwise $\psi_v = 0$, which implies that $1 - p_v = p_{ig}$. We will assume $1/2$ prior probability for $\psi_v = 1$ for all v. Moreover, some reads may map to multiple polymorphic sites $v_1, v_2, \ldots$ coupling $\psi_{v_1}, \phi_{v_2}, \ldots$. We suppose this happens rarely enough to be completely ignored so that all allele configurations ψ_v for any given (i, g) can be assumed independent.

We may call (ϕ_{ig}, ψ_v) *allele configuration* at site v. With the preceding considerations the definition of exclusion state θ_{ig} can be based on p_v and the allele configuration

	$\phi_{ig} \neq \psi_v$	$\phi_{ig} = \psi_v$
biallelic, $\theta_{ig} = 0$	$1 - p_v \in \mathcal{P}_0$	$p_v \in \mathcal{P}_0$
monoallelic, $\theta_{ig} = 1$	$1 - p_v \in \mathcal{P}_1$	$p_v \in \mathcal{P}_1$

Table 1: Definition of exclusion state θ_{ig} of (i, g) based on p_v and the allele configuration (ϕ_{ig}, ψ_v) for site $v \in (i, g)$

We will symbolically represent Table 1 by writing

$$p_v = P[\theta_{ig}, \delta_{\phi_{ig}\psi_v}] \tag{4}$$

$$P = \begin{pmatrix} 1 - \mathcal{P}_0 & \mathcal{P}_0 \\ 1 - \mathcal{P}_1 & \mathcal{P}_1 \end{pmatrix}, \tag{5}$$

where δ_{ab} is the Kronecker delta function, which is 1 if $\phi_{ig} = \psi_v$ and 0 otherwise.

To see the utility of P, consider the following example. Based on the data we have some uncertain knowledge on p_v, which we want to use to infer θ_{ig}. Suppose we know the allele configuration $(\phi_{ig}, \psi_v) = (0, 1)$. Then $\delta_{\phi_{ig}\psi_v} = 0$ and so we need to consider only the first column of P. If the data supports $p_v = P[0, 0] = 1 - \mathcal{P}_0$ better than $p_v = P[1, 0] = 1 - \mathcal{P}_1$, we can conclude that $\theta_{ig} = 0$ (biallelic expression) is more likely than $\theta_{ig} = 1$ (monoallelic expression).

In general we are uncertain about the allele configuration and we need to take expectation (i.e. average) over all four configurations using the prior probabilities κ and $1/2$. Moreover, if the number s_{ig} of polymorphic sites is > 1 then we will base the inference of θ_{ig} on all $p_v : v \in (i, g)$ jointly, taking expectation over all $4^{s_{ig}}$ configurations.

3 Three alternative global models

M1 No influence of explanatory variables

TODO: plate diagram

In this model both subintervals in Eq. 3-5 consist of a single point such that $\mathcal{P}_0 = \{1/2\}$ and $\mathcal{P}_1 = \{p_1\}$, where p_1 is some fixed number, say 0.9. Then Eq. 4 remains the same but

Eq. 5 changes to

$$P = \begin{pmatrix} 1/2 & 1/2 \\ 1 - p_1 & p_1 \end{pmatrix}. \tag{6}$$

For example, if the allele configuration at site v is $(0, 0)$ and the data supports $p_v = p_1$ stronger than $p_v = 1/2$ then we conclude, based only on v, that the exclusion state $\theta_{ig} = 1$ (i.e. monoallelic expression) is more likely.

M1.1 Special cases

TODO: DAG with theta leaves

As $\nu \to 0$, the beta distribution becomes Bernoully and μ_g will be 1 with probability π and 0 with probability $1 - \pi$. For any gene g this couples the exclusion state for all individuals so that $\theta_{1g} = ... = \theta_{Ig}$. So we can replace the general structure of model M1 with a probabilistically equivalent but simpler structure by introducing $\theta_g \equiv \theta_{1g}$ and removing μ_g (Figure TODO).

In the limit $\nu \to \infty$ we have $\mu_1 = ... = \mu_G = \pi$. Therefore we can once again simplify the model structure by removing μ_g. But the effect on θ_{ig} is the opposite in that $\theta_{1g}, ..., \theta_{Ig}$ become completely uncoupled in the sense that $\{\theta_{ig}\}_{ig}$ becomes independent and identically distributed (Figure TODO).

The interpretation of the first limiting case is that individuals show no variation in exclusion status for any gene g. Thus it makes sense to speak about bi or monoallelically expressing genes population-wide without the need of looking at individuals. The second limiting case, on the other hand, means that all genes have the same population-wide tendency for bi or monoallelic expression.

M2 Regression of Y_v on explanatory variables

TODO: plate diagram

The global structure of this model is the same as the special case of M1 given by $\nu \to 0$. So, for a given gene g all individuals have the same exclusion state θ_g but the across individual variation in explanatory variables x_i induces variation in p_v. For this the local model introduced in Section 2.2 must be extended with the regression of Y_v on X.

Given that Y_v is binomial, logistic regression appears as a natural framework, although some shortcomings will be discussed below TODO. In this framework the logit function links the expected fraction p_v of Y_v to the ith row of design matrix X so that Eq. 4 modifies to

$$\begin{aligned} p_v &= \operatorname{logit}^{-1}(x_i \, b_v) \qquad (7) \\ b_v &= B[\theta_{ig}, \delta_{\phi_{ig}\psi_v}], \qquad (8) \end{aligned}$$

where b_v is the R-length vector $(b_{v0}, ..., b_{vR-1})^\top$ and plays the role of regression coefficient in Eq. 7. As Eq. 8 says, b_v is an entry of matrix B of regression parameters, which is indexed by the exclusion state θ_{ig} and the allelic configuration (ϕ_{ig}, ψ_v).

Analogously to P under M1 (Eq.6), B under the present model M2 facilitates the inference of θ_{ig} based on y_v and (ϕ_{ig}, ψ_v) but, because b_v is a vector, B has a more complex structure than P, consisting of four R-length vectors:

$$\begin{aligned} B &= \begin{pmatrix} (0, ..., 0)^\top & (0, ..., 0)^\top \\ -\beta & \beta \end{pmatrix} \qquad (9) \\ \beta &= (\beta_0, \beta_1, ..., \beta_{R-1})^\top \qquad (10) \end{aligned}$$

β is a vector of regression parameters consisting of the intercept β_0 and a "slope" parameter β_r for each explanatory variable $x_r,\ 0 < r < R$. The bottom left entry represents a reflection of the regression curve defined by the bottom right entry accross the horizontal straight line defined by $p_v = 1/2$, which is analogous to the "reflection" in P of the point p_1 across the same horizontal line resulting in $1 - p_1$. That the $1, ..., R-1$ elements of top right entry are 0 expresses the assumption that when $\theta_{ig} = 0$ (biallelic expression) then the explanatory variables have no impact on p_v (Eq. 7); that the 0th element is also 0 follows from the equality $\text{logit}^{-1}(0) = 1/2$ showing that exclusion state $\theta_{ig} = 0$ under both M1 and the present M2 is defined by $p_v = 1/2$.

The connection between M1 and M2 can be made even more explicit by considering the special case of M2 that $\beta_1, ..., \beta_{R-1} = 0$ so that explanatory variables have no impact on p_v also when $\theta_{ig} = 1$ (monoallelic expression). Furthermore, if $\beta_0 = \text{logit}(p_1)$ also holds, then M2 is probabilistically equivalent to the special case M1 given by $\nu \to 0$. So for consistency between models we should set $\beta_0 = \text{logit}(p_1)$, which has the additional advantage of having one less unknown parameters.

M2.1 Shortcomings

TODO: logit function

M3 Regression of θ_{ig} on explanatory variables

TODO: plate diagram

TODO: link function

4 Inference

4.1 Local models and classification

4.1.1 Likelihood

$$\binom{n_v}{y_v} p_v^{y_v} (1-p_v)^{n_v - y_v} = \begin{cases} f_v(y_v | n_v, P, \phi_{ig}, \psi_v, \theta_{ig}), & p_v = \text{Eq. } 4 \quad (M1, M3) \\ f_v(y_v | n_v, x_i, B, \phi_{ig}, \psi_v, \theta_g), & p_v = \text{Eq. } 7, 8 \quad (M2) \end{cases} \tag{11}$$

As mentioned in Section 2.2 the allelic configurations $\{(\phi_{ig}, \psi_v) : v \in (i,g)\}$ are neither known nor informative and so must be considered nuisance parameters that need to be removed by marginalization, taking expectation over all possible configurations. This yields the following probability mass function under model M1 and M3:

$$L_{ig}^a \equiv f_{ig}(y_{ig} | n_{ig}, P, \kappa, \theta_{ig} = a) \tag{12}$$

$$= \frac{1}{2} \sum_{\phi_{ig}=0}^{1} \kappa^{\phi_v} (1-\kappa)^{1-\phi_v} \prod_{v \in (i,g)} \sum_{\psi_v = 0}^{1} f_v(y_v | n_v, P, \phi_{ig}, \psi_v, \theta_{ig} = a), \tag{13}$$

where a is 0 or 1, and L_{ig}^a is a convenient shorthand. Under model M2 f_{ig} has the same form except that P is replaced by x_i, B and θ_{ig} by θ_g as in Eq. 11. The same shorthand L_{ig}^a shall be used model M2 as well; its specific semantics shall be clear from the context.

4.1.2 Classification

4.2 Selection of a global model, estimation of π and β

The general frequentist procedure goes as follows:

1. express the marginal likelihood L_m for π based on y and X under all models $m = 1, ...,$ by taking expectations (over nuisance parameters such as μ_{ig}, ψ_v or over unknown θ_{ig})

2. maximize L_m with respect to π obtaining the ML estimate $\hat{\pi}_m = \arg\max_\pi L_m(\pi)$

3. for each m evaluate model fit using a criterion based on the maximized likelihood (such as AIC, BIC) and select the highest scoring model m^* and the corresponding $\hat{\pi}_{m^*}$

TODO: note on Bayesian procedure

4.2.1 Marginal likelihood for π

Under model M1 the marginal likelihood $L_{M1}(\pi, \nu) \equiv f(y|n, P, \kappa, \pi, \nu)$ for π and ν is given by

$$L_{M1}(\pi, \nu) = b^{-1} \prod_g \int_0^1 \mu^{\pi\nu} (1-\mu)^{(1-\pi)\nu} \prod_i \left[(1-\mu) L_{ig}^0 + \mu L_{ig}^1 \right] \, \mathrm{d}\mu \tag{14}$$

where b is the beta function evaluated at $(\pi\nu, (1-\pi)\nu)$. L_{M1} is marginal in the sense that expectation was taken over not only ψ_v (as in Eq. 12) but also θ_{ig} and μ_g.

L_{M1} in Eq. 14 depends on the parameter ν, which may be of some interest because it quantifies the tendency of genes to be monoallelically expressed across all individuals (so that individuals tend not to vary for any given gene). We may decide not to care about ν or take it to the limit $\nu \to 0$ or $\nu \to \infty$ by recalling the two special cases of model M1. To obtain the likelihood for those cases let us denote $L_{M1}(\pi) \equiv f(y|n, p, \kappa, \pi)$ and recall Eq 12. Then Eq. 14 simplifies to

$$L_{M1}(\pi) = \begin{cases} \prod_g \left[(1-\pi) \prod_i L_{ig}^0 + \pi \prod_i L_{ig}^1 \right] & \text{if } \nu \to 0 \\ \prod_{i,g} \left[(1-\pi) L_{ig}^0 + \pi L_{ig}^1 \right] & \text{if } \nu \to \infty, \end{cases} \tag{15}$$

where L_{ig}^a is used in the sense of M1-M3 (Eq. 12).

Turning to model M2, we assume that the matrix B of regression parameters is known (preset and/or estimated). Write $L_{M2}(\pi) \equiv f(y|n, X, B, \kappa, \pi)$. It is easy to see that $L_{M2}(\pi)$ has the same form as the top case ($\nu \to 0$); of course in this case the semantics of L_{ig}^a is connected to M2 (recall remark below Eq. 12).

4.3 Estimating regression parameters from training data

This estimation we need

- to accept model M2, which implies, for each gene, uniformity of exclusion state across all individuals

- a training set of genes known to be expressed monoallelically, collected from I' individuals

$$L_{M2}(B) = \frac{1}{2} \prod_{g} \prod_{i=1}^{I'} \sum_{\phi_{ig}=0}^{1} \kappa^{\phi_v} (1-\kappa)^{1-\phi_v} \prod_{v \in (i,g)} \binom{n_v}{y_v} \sum_{\psi_v=0}^{1} p_v^{y_v} (1-p_v)^{n_v - y_v} \tag{16}$$

where p_v is given by Eq. 7-8. The outer and inner summation together reflect marginalization over all allelic configurations (ϕ_{ig}, ψ_v), whereas the three running products represent the data aggregation over individual polymorphic sites v over individuals i over monoallelically expressed genes g to the level of the complete training data set.

4.4 Classification

Neyman-Pearson lemma

Log likelihood ratio statistic W_{ig} under model M1 M3 and W_g under M2.

5 Appendix

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.???. 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 $\{f_{ig}\}_{(i,g)\in\mathcal{S}}$:

$$h_{ig}(s_{ig}|n_{ig}, p_h) = \sum_{(i,g)\in\mathcal{S}} f_{ig}(y_{ig}|n_{ig}, p_h) \tag{17}$$

$$h'_{ig}(s_{ig}|p_h) = \sum_{(i,g)\in\mathcal{S}} f_{ig}(y_{ig}|n_{ig}, p_h)\, q_{ig}(n_{ig}|p_h). \tag{18}$$

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. 18), 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 or assumption on n_{ig}. This holds regardless we want to use h_{ig} (or h'_{ig}) in simulations, in parameter estimation or in classification with error control.