

Statistical model for one reciprocal cross (autosomal)

For a particular gene of interest, let n_i be the total number of AS reads in the i -th sample, and let n_{iB} be the number of AS reads mapped to strain B in the i -th sample. We model n_{iB} by a beta-binomial distribution, which is an extension of a binomial distribution to allow for possible over-dispersion. Specifically, let n_{iB} follow a binomial distribution with the number of trials n_i , and the probability of success p_S . If p_S follows a beta distribution with parameters α and β , the resulting distribution for n_{iB} is a beta-binomial distribution

$$h(n_{iB}; n_i, \alpha_i, \beta_i) = \binom{n_i}{n_{iB}} \frac{B(n_{iB} + \alpha_i, n_i - n_{iB} + \beta_i)}{B(\alpha_i, \beta_i)}. \quad (1)$$

For ease of modeling, we adopt a commonly used strategy to parameterize a beta-binomial distribution by $\pi_i = \alpha_i / (\alpha_i + \beta_i)$ and $\phi = 1 / (\alpha_i + \beta_i)$ (?):

$$h(n_{iB}; n_i, \pi_i, \phi) = \binom{n_i}{n_{iB}} \frac{\prod_{k=0}^{n_{iB}-1} (\pi_i + k\phi) \prod_{k=0}^{n_i-n_{iB}-1} (1 - \pi_i + k\phi)}{\prod_{k=1}^{n_i-1} (1 + k\phi)}, \quad (2)$$

where π_i is the expected proportion of AS reads from strain B . If there is no over-dispersion, then $\phi = 0$ and n_{iB} follows a binomial distribution. We further model the relation between π_i and paternal/maternal status by

$$\log \left(\frac{\pi_i}{1 - \pi_i} \right) = b_0 + b_1 x_i \quad (3)$$

where $x_i = 1$ if strain B is the paternal strain, and $x_i = -1$ if strain B is the maternal strain. Now we can test for strain effect and parent of origin effect as follows

$$\text{Strain effect: } H_0 : b_0 = 0 \quad \text{vs.} \quad H_1 : b_0 \neq 0 \quad (4)$$

$$\text{Parent of origin effect: } H_0 : b_1 = 0 \quad \text{vs.} \quad H_1 : b_1 \neq 0 \quad (5)$$

Let $\mu_B^{(p)}$ and $\mu_B^{(m)}$ be the expected expression of strain B in one cell when it is the paternal and maternal allele, respectively. Similarly define $\mu_A^{(p)}$ and $\mu_A^{(m)}$. Then the above parameterization can be written as

$$\log \left(\mu_B^{(p)} / \mu_A^{(m)} \right) = b_0 + b_1, \quad \text{and} \quad \log \left(\mu_B^{(m)} / \mu_A^{(p)} \right) = b_0 - b_1, \quad (6)$$

Therefore

$$b_0 = \log \left(\sqrt{\frac{\mu_B^{(p)} \mu_B^{(m)}}{\mu_A^{(p)} \mu_A^{(m)}}} \right) \text{ and } b_1 = \log \left(\sqrt{\frac{\mu_B^{(p)} \mu_A^{(p)}}{\mu_B^{(m)} \mu_A^{(m)}}} \right). \quad (7)$$

We further assume

$$\frac{\mu_B^{(p)}}{\mu_A^{(p)}} = \frac{\mu_B^{(m)}}{\mu_A^{(m)}} = \exp(b_0), \text{ and } \frac{\mu_B^{(p)}}{\mu_B^{(m)}} = \frac{\mu_A^{(p)}}{\mu_A^{(m)}} = \exp(b_1).$$

Next we consider the modeling of Total Read Counts (TReC). Throughout this paper, we denote a cross by Maternal Strain \times Paternal Strain. We consider four groups of mice: inbred strain A ($A \times A$), inbred strain B ($B \times B$), F1 cross of $A \times B$, and F1 cross of $B \times A$. Without any other covariates, the expected expression (in terms of Total Read Counts (TReC)) in the four groups of mice can be written as

$$\mu_{A \times A} = \mu_A^{(m)} + \mu_A^{(p)} = \mu_A^{(p)} \{1 + \exp(-b_1)\}, \quad (8)$$

$$\mu_{B \times B} = \mu_B^{(p)} + \mu_B^{(m)} = \mu_A^{(p)} \{\exp(b_0) + \exp(b_0 - b_1)\}, \quad (9)$$

$$\mu_{B \times A} = \mu_A^{(p)} + \mu_B^{(m)} = \mu_A^{(p)} \{1 + \exp(b_0 - b_1)\}, \quad (10)$$

$$\mu_{A \times B} = \mu_B^{(p)} + \mu_A^{(m)} = \mu_A^{(p)} \{\exp(b_0) + \exp(-b_1)\}. \quad (11)$$

There is a linear dependence among $\mu_{A \times A}$, $\mu_{B \times B}$, $\mu_{A \times B}$, and $\mu_{B \times A}$: $\mu_{A \times A} + \mu_{B \times B} = \mu_{A \times B} + \mu_{B \times A}$. Thus $\mu_{A \times A}$, $\mu_{B \times B}$, $\mu_{A \times B}$, and $\mu_{B \times A}$ only account for three independent observations. Furthermore, in real data analysis, we need to consider at least one covariate, the total number of reads per sample, denoted by κ_i . Therefore the total read counts themselves vary across samples due to other covariates and only the relative ratios among $\mu_{A \times A}$, $\mu_{B \times B}$, $\mu_{A \times B}$, and $\mu_{B \times A}$ are identifiable. Let η_i be the log ratio of the TReC of the i -th sample vs. the TReC of a “comparable” sample, except that it is from strain A . Here “comparable” means

all the covariates except strain have the same values. From the above equations, we have

$$\eta_i = \begin{cases} 0 & \text{if sample } i \in A \times A \\ b_0 & \text{if sample } i \in B \times B \\ -b_1 + \log \{1 + \exp(b_0 + b_1 x_i)\} - \log \{1 + \exp(-b_1)\} & \text{if sample } i \in A \times B \\ \log \{1 + \exp(b_0 + b_1 x_i)\} - \log \{1 + \exp(-b_1)\} & \text{if sample } i \in B \times A \end{cases}$$

Finally, we can model the TReC n_i by a negative binomial distribution with mean μ_i and over-dispersion parameter φ :

$$n_i \sim NB(\mu_i, \varphi), \quad \log(\mu_i) = \beta_0 + \beta_1 \kappa_i + \eta_i. \quad (12)$$

Statistical model for one reciprocal cross (X chromosome)

The above formula are for a single cell. We apply the formula to RNA-seq data from autosome while implicitly assuming the RNA-seq data are extracted from a homogenous cell population. However, this assumption is not valid for X chromosome. In each cell, only one copy of the X chromosomes is expressed. In F1 mice of AxB, let $\tau_{i,A}$ and $\tau_{i,B}$ be the proportions of cells where the A allele of X chromosome is expressed at individual i . Thus $\tau_{i,A} + \tau_{i,B} = 1$. Let $u_{i,B}^{(p)}$ and $u_{i,B}^{(m)}$ be the expression of B allele (across a large number of cells) at the i -th individual when B allele is paternal or maternal allele, respectively. Similarly we can define $u_{i,A}^{(p)}$ and $u_{i,A}^{(m)}$. Let ρ_A be the escaping ratio of the gene expression for allele A while the inactivated copy of X chromosome is from strain A . Then for individual i of $A \times B$:

$$\log \left(\frac{u_{i,B}^{(p)}}{u_{i,A}^{(m)}} \right) = \log \left(\frac{\tau_{i,B} \mu_B^{(p)} + \rho_B \tau_{i,A} \mu_B^{(p)}}{\tau_{i,A} \mu_A^{(m)} + \rho_A \tau_{i,B} \mu_A^{(m)}} \right) = \log \left(\frac{\tau_{i,B} + \rho_B \tau_{i,A}}{\tau_{i,A} + \rho_A \tau_{i,B}} \right) + b_0 + b_1, \quad (13)$$

and for individual i of $B \times A$,

$$\log \left(\frac{u_{i,B}^{(m)}}{u_{i,A}^{(p)}} \right) = \log \left(\frac{\tau_{i,B} \mu_B^{(m)} + \rho_B \tau_{i,A} \mu_B^{(m)}}{\tau_{i,A} \mu_A^{(p)} + \rho_A \tau_{i,B} \mu_A^{(p)}} \right) = \log \left(\frac{\tau_{i,B} + \rho_B \tau_{i,A}}{\tau_{i,A} + \rho_A \tau_{i,B}} \right) + b_0 - b_1. \quad (14)$$

$$(15)$$

Then for a single cell from X chromosome, the expected expression (in terms of Total Read

Counts (TReC)) in the four groups of mice can be written as

$$\begin{aligned}
u_{i,A \times A} &= 0.5\mu_A^{(m)} + .5\rho_A\mu_A^{(m)} + 0.5\mu_A^{(p)} + .5\rho_A\mu_A^{(p)} \\
&= 0.5(1 + \rho_A)\mu_A^{(p)} \{1 + \exp(-b_1)\}, \\
u_{i,B \times B} &= 0.5\mu_B^{(m)} + .5\rho_B\mu_B^{(m)} + 0.5\mu_B^{(p)} + .5\rho_B\mu_B^{(p)} \\
&= 0.5(1 + \rho_B)\mu_B^{(p)} \{\exp(b_0) + \exp(b_0 - b_1)\}, \\
u_{i,A \times B} &= \tau_{i,B}\mu_B^{(p)} + \rho_B\tau_{i,A}\mu_B^{(p)} + \tau_{i,A}\mu_A^{(m)} + \rho_A\tau_{i,B}\mu_A^{(m)} \\
&= \mu_A^{(p)} \{(\tau_{i,B} + \rho_B\tau_{i,A}) \exp(b_0) + (\tau_{i,A} + \rho_A\tau_{i,B}) \exp(-b_1)\}, \\
u_{i,B \times A} &= \tau_{i,B}\mu_B^{(m)} + \rho_B\tau_{i,A}\mu_B^{(m)} + \tau_{i,A}\mu_A^{(p)} + \rho_A\tau_{i,B}\mu_A^{(p)} \\
&= \mu_A^{(p)} \{\tau_{i,A} + \rho_A\tau_{i,B} + (\tau_{i,B} + \rho_B\tau_{i,A}) \exp(b_0 - b_1)\}.
\end{aligned}$$

Note that in the above equation, we assume that in inbred mouse strains, 50% of activated X chromosomes are from maternal strain and 50% of activated X chromosomes are from paternal strain with same escaping inactivation ratio.

Then we have

$$\eta_l = \begin{cases} 0 & \text{if sample } l \in \text{strain } A \\ b_0 + \log \left\{ \frac{1+\rho_B}{1+\rho_A} \right\} & \text{if sample } l \in \text{strain } B \\ -b_1 + \log \left\{ 1 + \frac{(\hat{\tau}_{iB}/\hat{\tau}_{iA} + \rho_B)}{(1+\rho_A\hat{\tau}_{iB}/\hat{\tau}_{iA})} \exp(b_0 + b_1x_i) \right\} \\ \quad + \log \left\{ \frac{1+\rho_A\hat{\tau}_{iB}/\hat{\tau}_{iA}}{1+\rho_A} \right\} + \log \{2\hat{\tau}_{iA}\} - \log \{1 + \exp(-b_1)\} & \text{if sample } l \in A \times B \\ \log \left\{ 1 + \frac{(\hat{\tau}_{iB}/\hat{\tau}_{iA} + \rho_B)}{(1+\rho_A\hat{\tau}_{iB}/\hat{\tau}_{iA})} \exp(b_0 + b_1x_i) \right\} \\ \quad + \log \left\{ \frac{1+\rho_A\hat{\tau}_{iB}/\hat{\tau}_{iA}}{1+\rho_A} \right\} + \log \{2\hat{\tau}_{iA}\} - \log \{1 + \exp(-b_1)\} & \text{if sample } l \in B \times A \end{cases}$$

Statistical model for ASE in three reciprocal crosses

Now we extend our notation with superscript $^{(AB)}$ indicating a cross from strains A and B , could be either $A \times B$ or $B \times A$. Denote the three strains as A , B , and C . For strain effect, we have three situations:

(S1) there is no strain effect:

$$b_0^{AB} = b_0^{BC} = b_0^{AC} = 0.$$

(S2) there is consistent strain effect only:

$$b_0^{AB} \neq 0, b_0^{BC} \neq 0, \text{ and } b_0^{AC} = b_0^{AB} + b_0^{BC}.$$

(S3) there is in-consistent strain effect:

$$b_0^{AB} \neq 0, b_0^{BC} \neq 0, \text{ and } b_0^{AC} \neq 0.$$

Similarly, there are three situations for parent of origin effect:

(P1) there is no parent of origin effect:

$$b_1^{AB} = b_1^{BC} = b_1^{AC} = 0.$$

(P2) there is consistent parent of origin effect:

$$b_1^{AB} = b_1^{BC} = b_1^{AC} \neq 0.$$

(P3) there is inconsistent parent of origin effect:

$$b_1^{AB} \neq 0, b_1^{BC} \neq 0, \text{ and } b_1^{AC} \neq 0.$$

Different hypotheses can be tested by comparing models under different situations. For example, if we want to ask whether strain background affects the parent of origin effect. We can compare models of situation (P2) vs. situation (P3). While choose of the the situations (S1), (S2), and (S3) for strain effect.

Similarly to the previous section, the total read count (TReC) of strains A , B , C , and all the 6 reciprocal crosses can be modeled by negative binomial distributions, after introducing parameters β_0 , β_1 , b_2 and φ .

[Table 1 about here.]

Table 1
Power Analysis

b_0	b_1	b_2	$A \times B$	$B \times A$	TReC		ASE	TReCASE	
					(RC)	(RCI)		(RC)	(RCI)
0	0	0	2	2	0.009	0.016	0.051	0.056	0.051
0	0	0.3	2	2.7	0.004	0.012	0.054	0.054	0.054
0	0.3	0	1.7	2.3	0.008	0.017	0.199	0.193	0.202
0	0.3	0.3	1.7	3.2	0.009	0.013	0.203	0.209	0.219
0.3	0	0	2.3	2.3	0.032	0.080	0.214	0.251	0.311
0.3	0	0.3	2.3	3.2	0.020	0.078	0.215	0.249	0.305
0.3	0.3	0	2	2.8	0.040	0.095	0.422	0.464	0.525
0.3	0.3	0.3	2	3.8	0.042	0.111	0.382	0.431	0.490