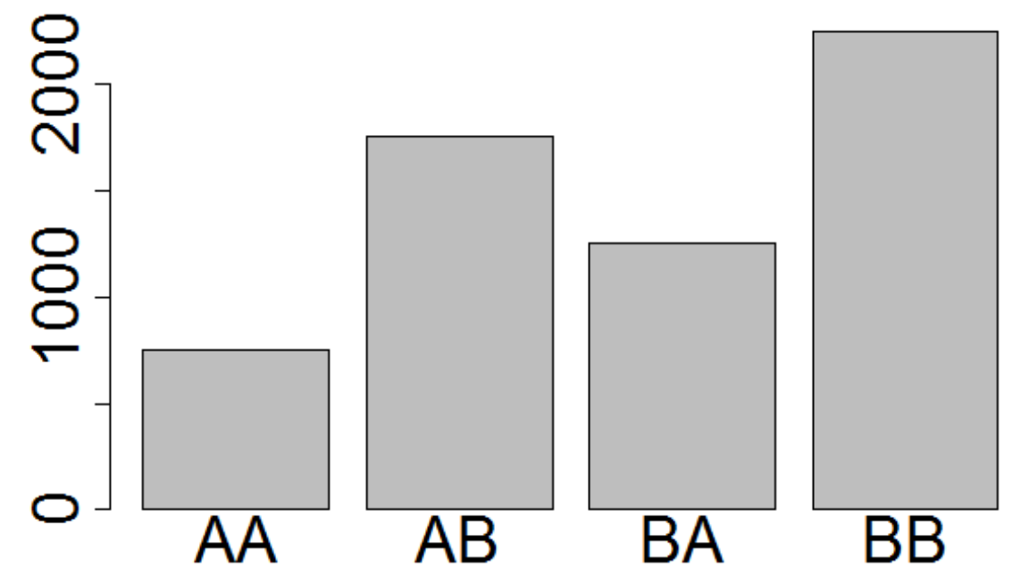
TReCASE: a powerful method to detect differential gene expression using total and allele specific RNA-seq data

Fei Zou[1], Wei Sun[1,2], James J Crowley[2], Vasyl Zhabotynsky[1], Patrick F Sullivan[2] , Fernando Pardo-Manuel de Villena[2] , James Xenakis[1] , Paola Giuisti[2]

[1]Department of Biostatistics, UNC, [2]Department of Genetics, UNC

Differential gene expression through total and allele specific counts

ASE	mat	pat
A	250	500
B	750	1500



Motivating example: for a certain gene, for which Haplotype B is expressed 3 times as haplotype A and maternal reads are twice less expressed compared to paternal we can get the counts as shown in the table.

So the differential expression can be inferred from both total read counts (TReC) and allele specific expression (ASE).

In the data produced by RNA-seq we may consider all the reads that were mapped to the gene as TReC, plus based on SNP or indel information part of these reads in F1 can be attributed either to mother or father.

Autosome Model

To allow over-dispersion assume that allele specific reads follow beta-binomial distribution:

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

Where π_i – expected proportion of the strain B allele specific reads is linked with additive and parent of origin effects with the following link:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right)=(b_{0F}+b_{1F}x_i)sex_i+(b_{0M}+b_{1M}x_i)(1-sex_i)$$

For total read counts we assume negative binomial distribution

$$m_l \sim f_{NB}(m_l;\mu_l,\varphi),\quad for\ l=1,2,\ldots,K_1+K_2,\quad with$$
$$\log(\mu_l)=\beta_0+\beta_1\kappa_l+\beta_2 sex_l+\beta_3 dom_l+\beta_4 dom_l\times sex_l+\eta_l$$

which is connected with the above defined beta-binomial through η_i, using the stated to the right assumption.

$$\log\left(\frac{\mu_{F,B}^{(p)}}{\mu_{F,A}^{(m)}}\right)=b_{0F}+b_{1F}$$

$$\log\left(\frac{\mu_{F,B}^{(m)}}{\mu_{F,A}^{(p)}}\right)=b_{0F}-b_{1F}$$

Which leads to the expected TReC due to additive or parent of origin effect expression for each female mouse would be as stated below:

$$\begin{cases}\mu_{F,A}^{(m)}+\mu_{F,A}^{(p)}=\mu_{F,A}^{(p)}\{1+\exp(-b_{1F})\} & for\ A\times A\\ \mu_{F,B}^{(m)}+\mu_{F,B}^{(p)}=\mu_{F,A}^{(p)}\{\exp(b_{0F})+\exp(b_{0F}-b_{1F})\} & for\ B\times B\\ \mu_{F,A}^{(m)}+\mu_{F,B}^{(p)}=\mu_{F,A}^{(p)}\{\exp(b_{0F})+\exp(-b_{1F})\} & for\ A\times B\\ \mu_{F,B}^{(m)}+\mu_{F,A}^{(p)}=\mu_{F,A}^{(p)}\{1+\exp(b_{0F}-b_{1F})\} & for\ B\times A\end{cases}$$

Thus, taking AA cross as a reference, we defines η_i. Same procedure is done for male. Finally, we can define joint likelihood of the combined F1 and inbred mice.

$$L(\Theta)=\prod_{i=1}^{K_1} f_{BB}(n_{iB};n_i,\pi_i,\phi)\prod_{l=1}^{K_1+K_2} f_{NB}(m_l;\mu_l,\varphi)$$

And testing additive and parent of origin effect with LR test as follows:

$$H_0: b_{0F}=b_{0M}=0 \quad H_0: b_{1F}=b_{1M}=0$$

X Chromosome

For X chromosome we need to take in account the chromosome-wide *Xce* effect (τ_B) so the link modifies to:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right)=\log\left(\frac{\tau_{i,B}}{\tau_{i,A}}\right)+b_{0F}+b_{1F}x_i$$

Which leads to η_i (for female samples)

$$\eta_l=\begin{cases}0 & l\in A\times A\\ b_{0F} & l\in B\times B\\ \log\left\{1+\exp\left(\log\left(\frac{\tau_{l,B}}{\tau_{l,A}}\right)+b_{0F}+b_{1F}\right)\right\} & l\in A\times B\\ \quad -\log\{1+\exp(b_{1F})\}+\log\{2\tau_{l,A}\} & \\ \log\left\{1+\exp\left(\log\left(\frac{\tau_{l,B}}{\tau_{l,A}}\right)+b_{0F}-b_{1F}\right)\right\} & l\in B\times A\\ \quad -\log\{1+\exp(-b_{1F})\}+\log\{2\tau_{l,A}\} & \end{cases}$$

And for male samples

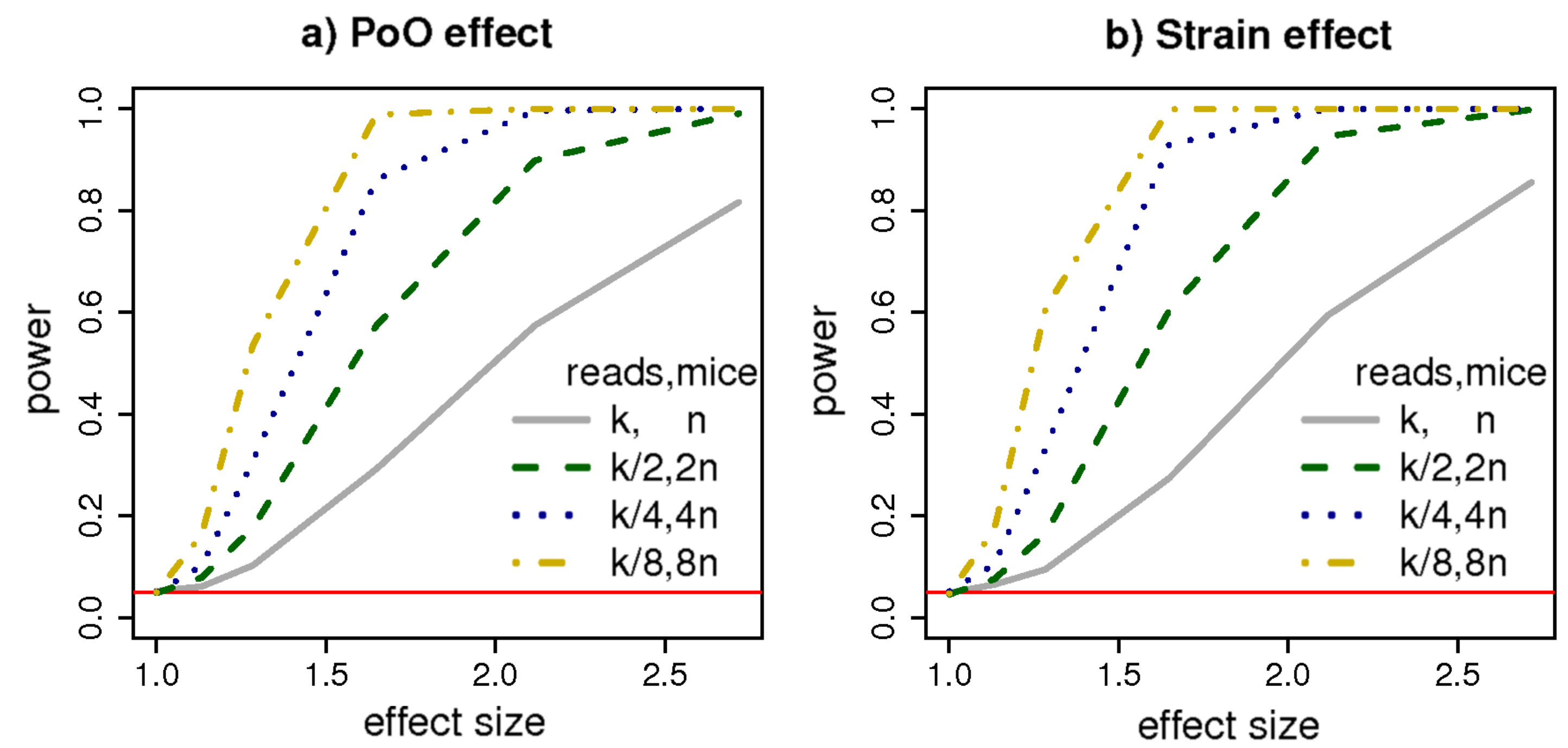
$$\eta_l=\begin{cases}\log\{2\} - \log\{1+\exp(-b_{1F})\} & l\in A\times A\ or\ A\times B\\ b_{0M}+\log\{2\} - \log\{1+\exp(-b_{1F})\} & l\in B\times B\ or\ B\times A\end{cases}$$

Simulations

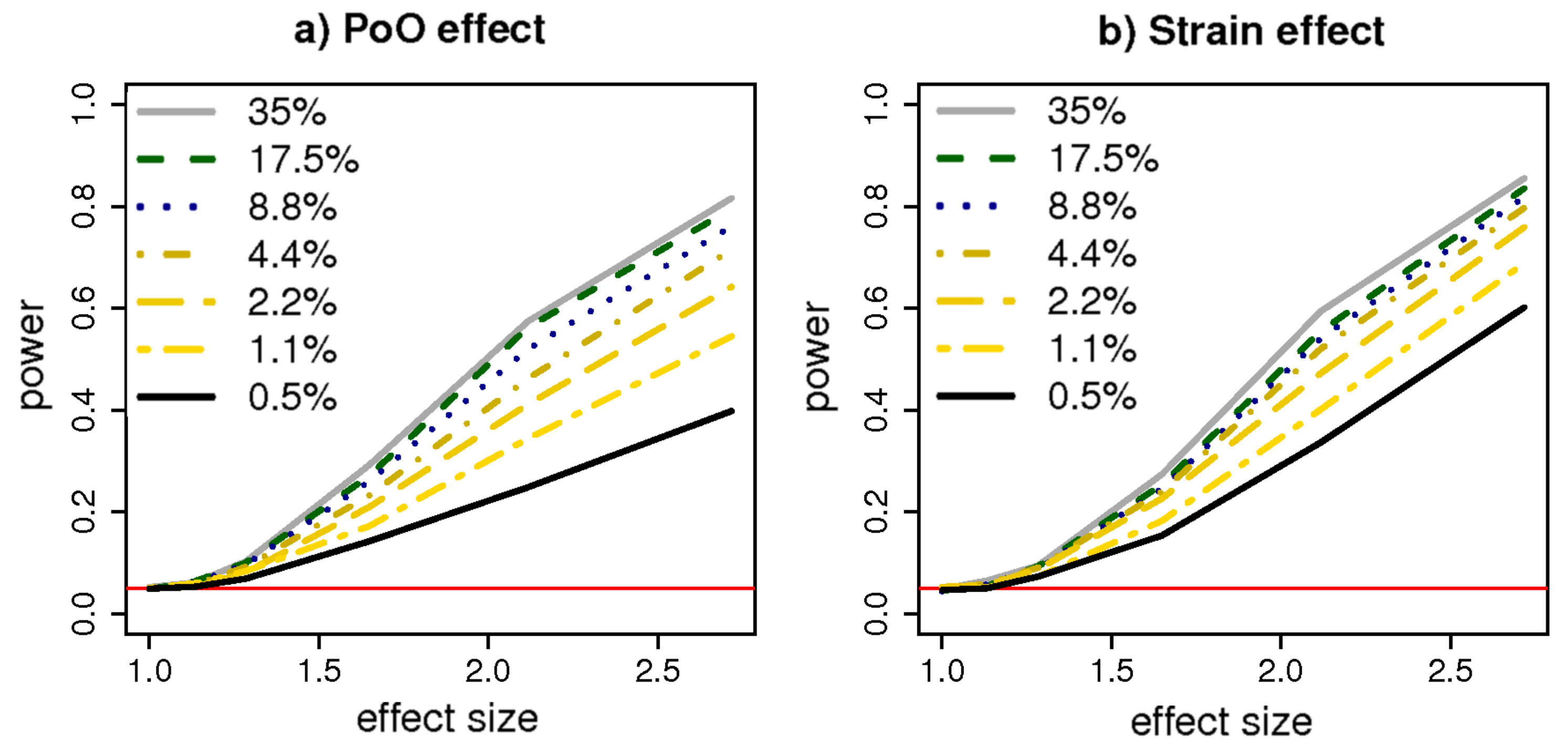
Comparing the model with simple Poisson for ASE and binomial fit TReC we can see that in contrast to our model such simple fit doesn't control properly for type 1 error both when we simulate using the model (left table) or using flux-simulator (right table)

TReCASE		Simple			TReCASE		Simple	
Strain	PoO	Strain	PoO	alpha	Strain	PoO	Strain	PoO
4.65E-02	4.80E-02	2.24E-01	2.14E-01	5E-02	4.65E-02	4.80E-02	2.24E-01	2.14E-01
8.54E-03	9.96E-03	1.42E-01	1.32E-01	1E-02	8.54E-03	9.96E-03	1.42E-01	1.32E-01
9.36E-04	1.01E-03	8.57E-02	8.11E-02	1E-03	9.36E-04	1.01E-03	8.57E-02	8.11E-02
7.19E-05	1.08E-04	5.85E-02	5.43E-02	1E-04	7.19E-05	1.08E-04	5.85E-02	5.43E-02
5.65E-06	1.28E-05	4.17E-02	3.94E-02	1E-05	5.65E-06	1.28E-05	4.17E-02	3.94E-02

Simulations also show that increasing number of reads to increase power is less efficient than increasing sample size:



Finally, we can see an increase in power as long as we have decent allele specific counts: even if ASE is 4% total read counts greatly improve power. This is especially clear in the parent of origin effect.



Work in progress: modelling RIX lines

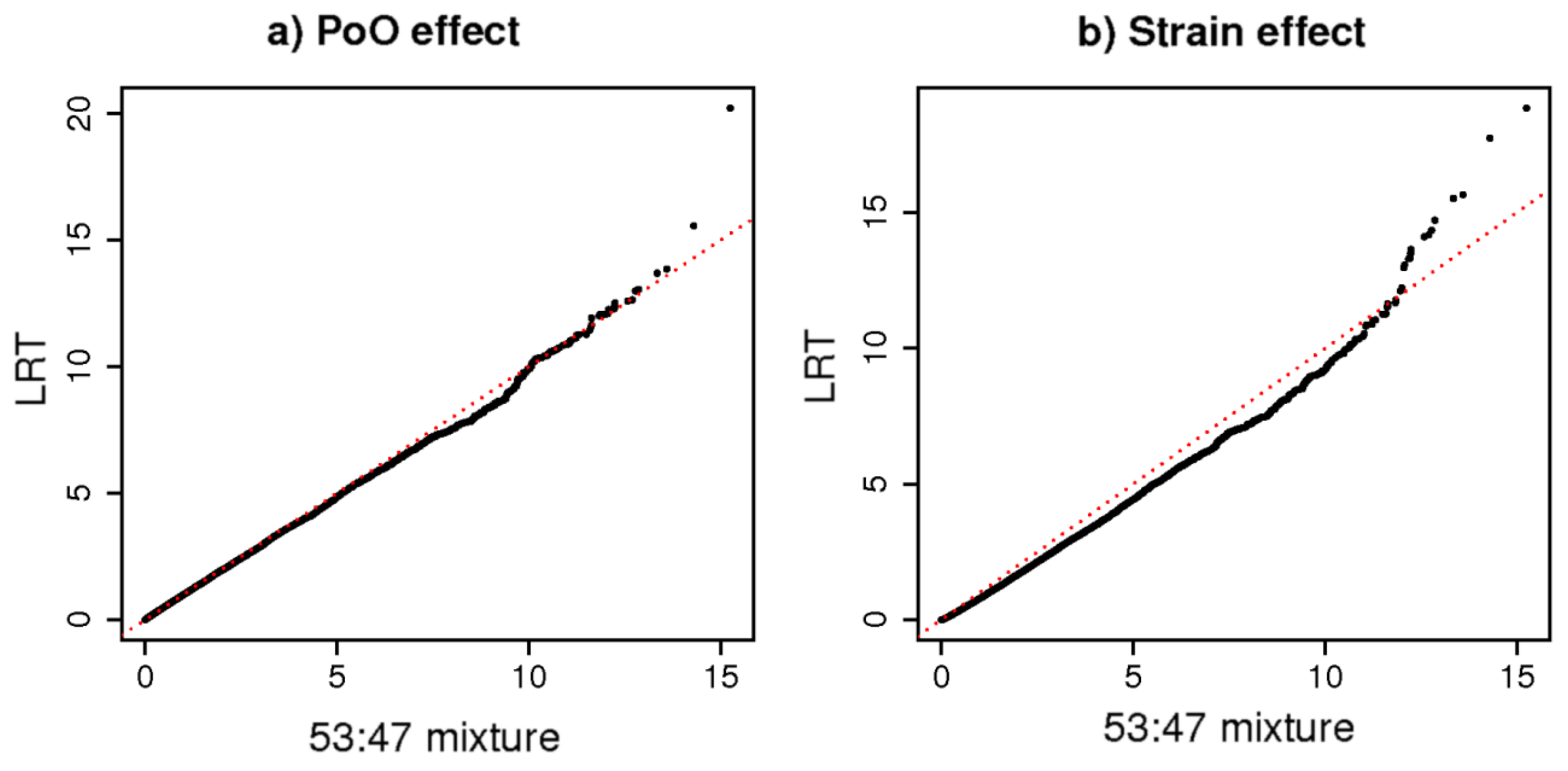
The experiment goal is to analyze mice reaction to haloperidol. Each RIX line each crossed with the next one in the loop design, with target of 6 male and 6 female mice. Furthermore, 3 mice of each sex get haloperidol treatment and 3 mice get placebo treatment. Each cage contains one treated and one untreated mouse. We fit linear mixed model

$$y=\beta_0+\beta_1 sex+\beta_2 treatment+\beta_3 sex\times treatment+\beta_4 pretreat+$$
$$Z_1\alpha_1+Z_2\alpha_2+Z_3\alpha_3+Z_4\alpha_4+Z_5\alpha_5+\epsilon$$

The random effects are additive, parent of origin, additive*treatment, parent of origin*treatment and batch effect. Note, that due to structure of the loop design, Z1 is defined by 1 for each RIX line present in this cross, Z2 as 1 for present maternal line and -1 for present paternal line. Treatment interactions are produced from Z1 and Z2

$$Z_1=\begin{pmatrix}1&1&0&\ldots&0&0\\0&1&1&\ldots&0&0\\\ldots&\ldots&\ldots&\ldots&\ldots&\ldots\\0&0&0&\ldots&1&1\\1&0&0&\ldots&0&1\end{pmatrix}\quad Z_2=\begin{pmatrix}1&-1&0&\ldots&0&0\\0&1&-1&\ldots&0&0\\\ldots&\ldots&\ldots&\ldots&\ldots&\ldots\\0&0&0&\ldots&1&-1\\-1&0&0&\ldots&0&1\end{pmatrix}$$

Simulations for the loop design with reciprocal crosses based on 1E5 simulations shows that for such sample size the mixture χ^2 distributions with 0 and 1 df varies around 0.53:0.47 and 0.60-0.40



References:
Fei at al. A Novel Statistical Approach for Jointly Analyzing RNA-Seq Data from F1 Reciprocal Crosses and Inbred Lines 2014 Genetics