eQTL Analysis Using Human RNA-seq Data with TReCASE and RASQUAL

Vasyl Zhabotynsky[1], Yi-Juan Hu[5], Fei Zou[1,2,6], Wei Sun[1,3,4]

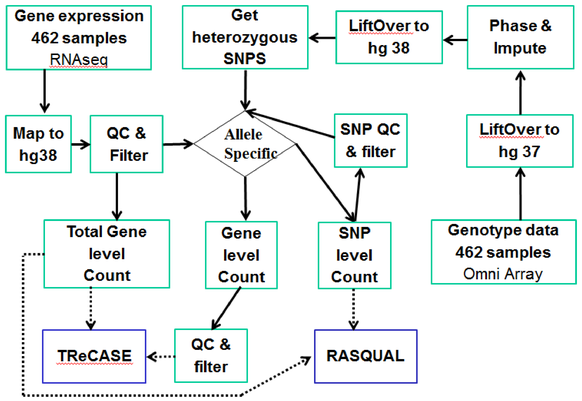
[1]Department of Biostatistics, University of North Carolina, Chapel Hill; [2]Dept of Genetics, University of North Carolina, Chapel Hill
[3]Public Health Science Division, Fred Hutchison Cancer Research Center; [4]Department of Biostatistics, University of Washington, Seattle, WA
[5]Departmentt of Biostatistics and Bioinformatics, Emory University, Atlanta, GA; [6]Department of Biostatistics, University of Florida, Gainesville, FL

Summary

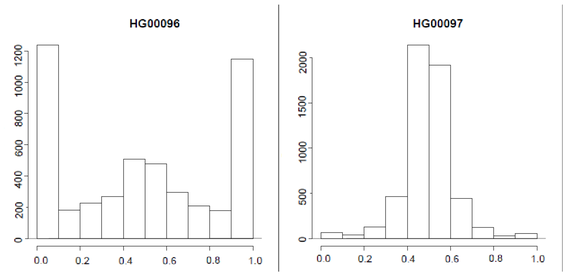
To fully utilize benefits of RNA-seq data one needs to combine total and allele specific expression derived from RNA-seq data. It, however, requires multiple steps of careful data processing.

We present a protocol for such data processing and evaluate under various assumptions two top performing methods: TReCASE and RASQUAL.

Sample processing



- RNA-seq and genotype data of 462 samples from Geuvadis project
- Phasing and imputation is done with shapeit v.2 impute v.2 and according to their pipeline and recommended settings
- Mapping was done with tophat v.2
- SNP level count performed using GATK/ASEReadCounter
- Read level allele-specific count: asSeq/extractASReads
- Gene level count: GenomicAlignments/summarizeOverlaps
- Example of the filtered out samples:



Note, to test one of the major assumptions in the real data we consider additional dataset of 30 individuals. This dataset has higher number of allele-specific SNPs due to availability of parental genotype information. Otherwise it is processed in the same fashion as the main dataset.

TReCASE introduction

Total expression model is set up on gene-level.
For individuals i=1…M

y_i ~ Negative Binomial(μ_i, ϕ_1)

$$\log(\mu_i) = \sum_{k=1}^{p} \beta_k c_{ik} + \eta_{ij},$$

$$\eta_{ij} = \begin{cases} 0 & \text{if } g_{ij} = 0(AA) \\ \log\{1+\exp(b_0)\} - \log\{2\} & \text{if } g_{ij} = 1(AB) \\ b_0 & \text{if } g_{ij} = 2(BB) \end{cases}$$

Allele specific expression model is set up on gene-level
$n_i = n_{iA} + n_{iB}$
n_{iB} ~ Beta Binomial(π_i, ϕ_2)

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = b_0$$

RASQUAL introduction

Total expression model is set up on gene-level.
For individuals i=1…M

y_k ~ Negative Binomial(μ_i, ϕ)

$$\mu_i = \begin{cases} 2(1-\pi)\lambda K_i & \text{if } g_i = 0 \\ \lambda K_i & \text{if } g_i = 1 \\ 2\pi\lambda K_i & \text{if } g_i = 2 \end{cases}$$

K_i - sample specific offset, estimated *a priory*
λ – scale parameter for mean gene expression

Allele specific expression model is set up on SNP level
$n_i = n_{iA} + n_{iB}$
n_{iB} ~ Beta Binomial(π_i, ϕ)

$\{\pi_{iB}, \pi_{iA}\}$: $\{\pi, 1\text{-}\pi\}$ or $\{0.5, 0.5\}$ for a given SNP

Major differences in assumptions

The major difference of the two methods is their approach to model allele-specific reads within an individual.

- TReCASE assumes within sample counts are distributed binomially and beta variation comes from between sample differences
- RASQUAL treats each SNP as independent Beta-Binomial with the same over-dispersion as between-individual over-dispersion
- RASQUAL also assumes the over-dispersion parameters for both total counts and allele specific counts are the same.

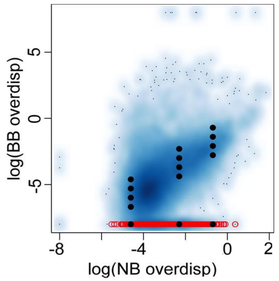
Potential issues

1. Common over-dispersion for total and allele-specific counts
2. Both methods avoid estimating within sample over-dispersion between SNPs:

- TReCASE assumes there is no such over-dispersion. In case of large inter-sample over-dispersion it will spill to between-sample over-dispersion leading to it's over-estimation.
- RASQUAL assumes that such over-dispersion is the same within sample as between samples. Since we expect such over-dispersion to be smaller than between sample over-dispersion it would underestimate overall over-dispersion

4. SNP level double-counting of allele-specific counts

Observed over-dispersion (OD)



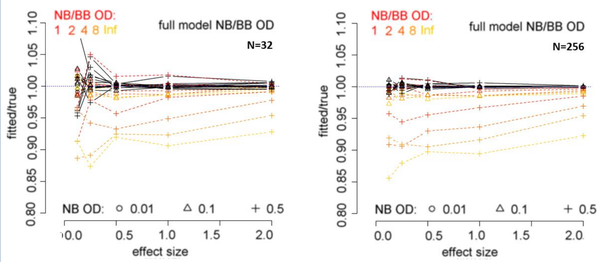
Note, that in observed data-set we see notable variation of over-dispersion parameters.

1. Beta-binomial over-dispersion in most of the cases is lower than Negative-binomial over-dispersion

2. There is a notable fraction of genes for which counts are distributed as Binomial

We simulated several setups marked by circles to study the effects of such discrepancies

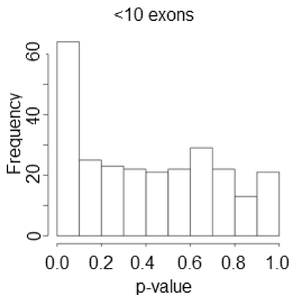
Consequences of common OD



We observe that larger ratio of Negative-Binomial and Beta-Binomial over-dispersion parameters leads to larger bias in eQTL estimate.

This bias is persistent even for large sample sizes and is especially pronounced for larger over-dispersion parameters.

Within sample OD



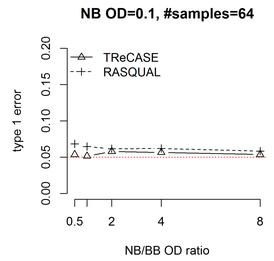
We considered a subset of genes, having multiple SNPs and used a score statistic developed by Tarone (1979) to test for a deviation from binomial distribution assumption.

Since we don't have too many SNPs this statistic is not normal, so we performed parametric bootstrap to calculate p-value

We see notable enrichment in significant p-values. It is likely due to presence of multiple isoforms and the degree of allelic imbalance may vary across isoforms.

It also leaves a possibility of other within sample over-dispersion

Consequences of OD misspecification

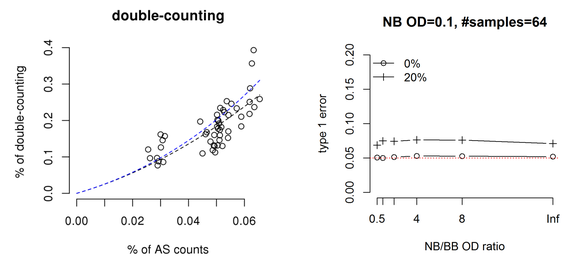


We considered a case with within sample OD of about half magnitude of between sample OD

TReCASE model: no within sample OD, but separate NB a BB OD

RASQUAL model: within sample BB = between sample BB = NB OD

Double-counting consequences



For variety of over-dispersion ratios we observe notable inflation of type 1 error.

Conclusions

- Common over-dispersion in Negative-Binomial and Beta-Binomial distribution is not typical for most of the genes and biases both OD estimate and eQTL estimate.
- Among a subset of multiple SNPs (typically multi-exonic genes) we observe about 20% of genes not satisfying the assumption of constant proportion (or within sample OD)
- We considered a worst case scenario when there is a within sample OD: it has a moderate impact on type 1 errors in both methods.
- Double-counting leads to a notable inflation of type 1 error. To avoid it we plan to add a function to asSeq package that would provide one-read per SNP functionality for RASQUAL model
- RASQUAL tends to under-estimate over-dispersion for small #SNPs

Citations

1 Sun, Wei. "A statistical framework for eQTL mapping using RNA-seq data." *Biometrics* 68.1 (2012): 1-11.
2 Hu, Yi-Juan, et al. "Proper use of allele-specific expression improves statistical power for cis-eQTL mapping with RNA-seq data." *JASA* 110.511 (2015): 962-974.2
3 Kumasaka, Natsuhiko, Andrew J. Knights, and Daniel J. Gaffney. "Fine-mapping cellular QTLs with RASQUAL and ATAC-seq." *Nature genetics* 48.2 (2016): 206.
https://www.ebi.ac.uk/arrayexpress/experiments/E-GEUV-1/
https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRP106527