EASEA: Easy Allele Specific Expression Analysis

a novel computational biology pipeline with direct clinical application

Daan van Beek[1], Martina Kutmon[1], Theo de Kok[2], Ilja Arts[1], Michiel Adriaens[1]
[1]*The Maastricht Centre for Systems Biology, Maastricht University, Maastricht, The Netherlands*
[2]*Toxicogenomics, Maastricht University, Maastricht, The Netherlands*

Background

The human genome is diploid, meaning that every individual possesses two copies of every gene, one maternal and one paternal. Due to mutations humans therefore often possess two different versions of the same gene, called alleles. Allele specific expression (ASE) describes the phenomenon in which the expression of one those alleles for a given gene is significantly different. ASE serves as a proxy for underlying genetic expression regulation and alternative splicing[1].

If ASE for a gene is detected significantly more often amongst cases as compared to controls, we can deduce that the underlying regulatory mechanisms play a role in the development of the phenotype of interest.

Methods

The analysis is based on the current best practices guidelines and initializes with the alignment of the RNA-sequencing data to the reference genome[1]. Subsequently, the read counts for each locus can be calculated, as well as the read counts for each of the two alleles at biallelic sites (**Figure 1**).

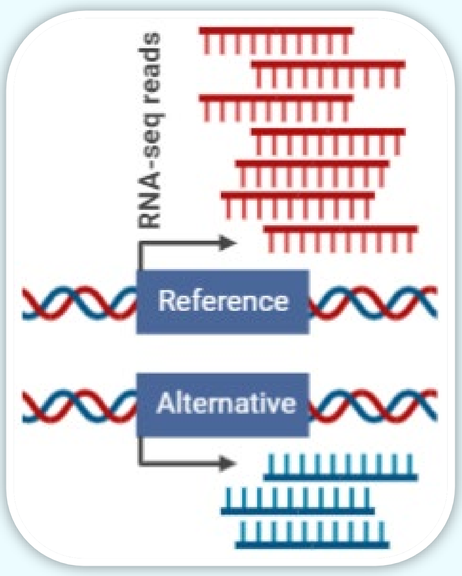

Figure 1: Detecting ASE through RNA-sequencing data

The next step comprises the actual ASE analysis (**Figure 2**). The ref:alt ratios are calculated for each site for each individual. Homozygotes are excluded since. At this point, cases and controls can be compared to pinpoint ASE loci specific to the phenotype of interest. Those loci showing significant ASE effects are then annotated with their host gene. Finally, gene-based (**Figure 3A**) and patient-based (**Figure 3B**) networks will be created. Affinity-based clustering will be performed on the patient-based network to show clusters of patients with highly similar ASE profiles[3,4].

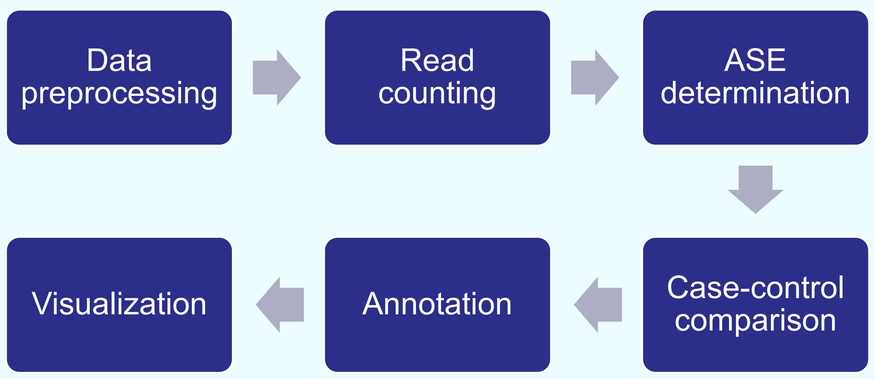

Figure 2: Abstraction of the pipeline

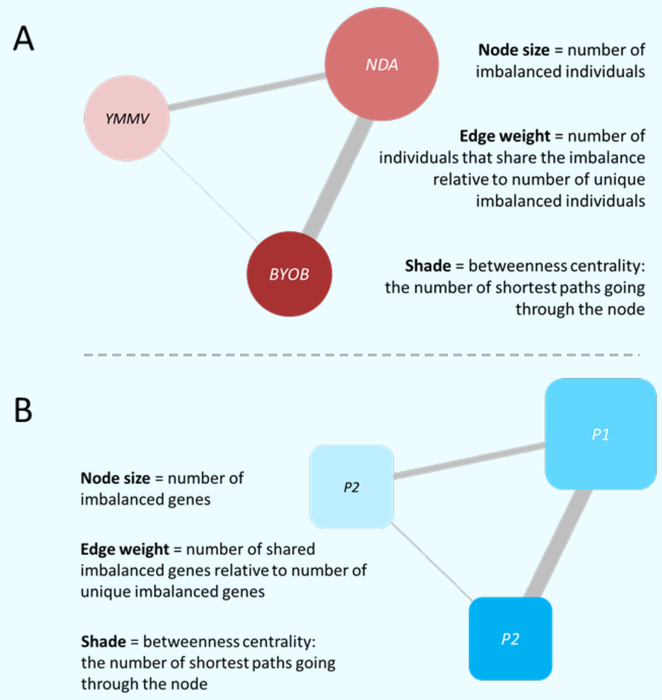

Figure 3: Gene-based (A) and patient-based (B) network visualization examples

(Preliminary) Results

As a proof of principle we will apply the pipeline to a unique, in-house case-control cohort focusing on dilated cardiomyopathy (DCM). Like other complex genetic disorders, DCM has a largely unexplained heritability factor with many identified SNPs residing in non-coding regions[4]. Thus, it can be reasonably assumed that regulatory and splicing factors play a substantial role in the pathophysiology of DCM. The preliminary results from a subset of the cohort can be seen in (**Figure 4**).

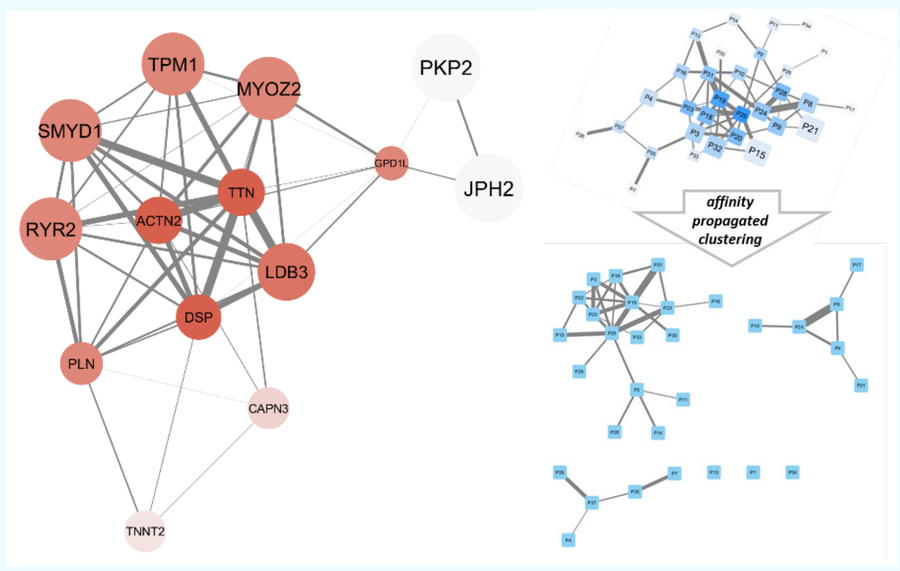

Figure 3: Gene-based (A) and patient-based (B) network visualization examples
Left: gene-based network; Right: patient-based network before and after affinity-based clustering.

Conclusion

We have developed and showcased a complete pipeline offering RNA-sequencing data analysis and visualization for ASE following best practices guidelines. Since ASE is reliable for individuals, large sample sizes, and small sample sizes, we think the pipeline is highly relevant for the study of (rare) complex genetic disorders in order to find potential genes that are differentially regulated or spliced in the disease phenotype.

References

1. *Castel S, Levy-Moonshine A, Mohammadi P et al.: Tools and best practices for data processing in allelic expression analysis. Genome Biol 2015, 16:195.*
2. *Shannon P et al.: Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003, 13(11):2498-504.*
3. *Morris JH et al.: clusterMaker: a multi-algorithm clustering plugin for Cytoscape. BMC Bioinformatics 2011, 12:436.*
4. *Heinig M, Adriaens ME, Schafer S et al.: Natural genetic variation of the cardiac transcriptome in non-diseased donors and patients with dilated cardiomyopathy. Genome Biol 2017, 18:170.*

Correspondence to:
Michiel Adriaens, PhD

michiel.adriaens@maastrichtuniversity.nl

maastrichtuniversity.nl/research/maastricht-centre-systems-biology

Correspondence to:
Daan van Beek, MSc

d.beek@maastrichtuniversity.nl

Maastricht Centre for Systems Biology
Faculty of Science and EngineerinMaastricht University
P.O. Box 616
6229 EN Maastricht, The Netherlands