RNA-Seq data integration:
Mapping eQTLs using a LoF approach and identification of additional LoF variants

Manuel Rivas (University of Oxford) in collaboration with
Tuuli Lappalainen and the Dermitzakis group at University of Geneva

Aims

- Mapping cis eQTL using a LoF approach.
- Identify splice variants that create truncated transcripts.
  - Now we only focus on Essential splice variants (+/- 2bp intronic), Frameshift, Nonsense mutations.
  - We would like to include additional variants into other projects based on empirical data that demonstrate similar effect.
  - Ideas (1. Empirical -> Motif identification 2. Computational prediction -> empirical)

Motivation for focus on LoF

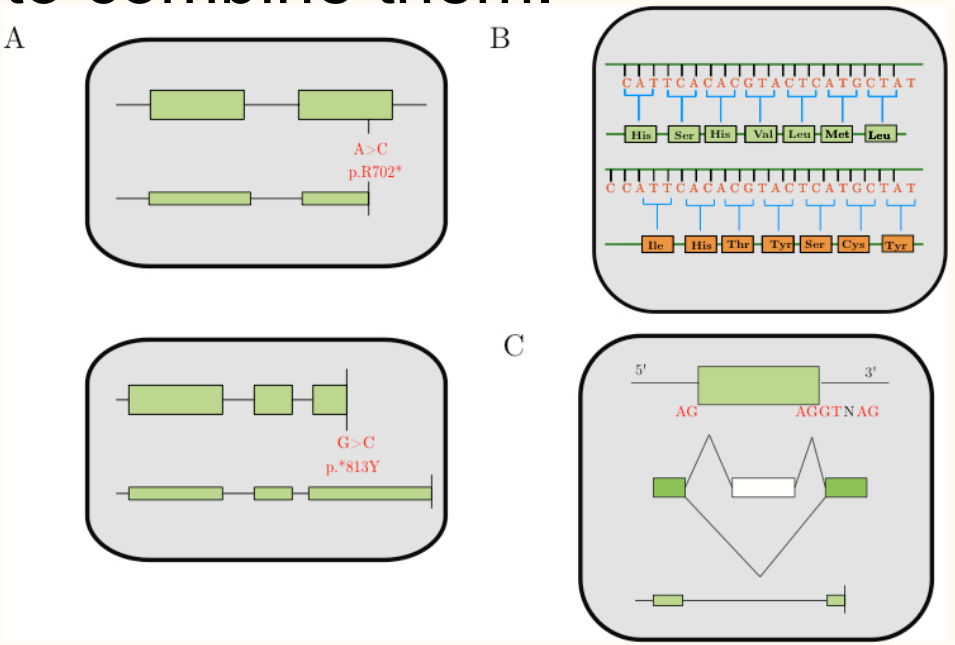
- Unlike GWAS, we are unlikely to have power to test rare and low-frequency variants individually.
- So, look for a signal by combining evidence across variants within genes (or perhaps within pathways,....)
- Challenge is how to do the combining – how to measure “burden”! Various methods (C-alpha, VT, KBAC etc) available. Weights could depend on ..... functional prediction, MAF, how unusual that type of mutation is, ...

Basic Idea

- We focus only on Loss of Function (LoF) variants: assume that all LoF variants have the same effect on gene, so straightforward to combine them.

- (Putative) LoF variants:

- Splice site variants
- Nonsense mutations
- Frameshift indels.



Interested in using this strategy to map disease and quantitative trait genes. However, knowledge about splice variants that lead to different transcripts is limited. In addition, how often nonsense and frameshift indels escape NMD is unknown (at least to me). RNAseq data is perfect experimental dataset to learn more about LoF variants.

Human Gene Mapping

- Has been helpful for disease and qt mapping.
  - Breast Cancer ~5 genes discovered so far with this approach.
  - IBD (Rivas et al. 2011; *CARD9,CUL2*) protective splice variants
  - T1D (Nejentsev et al. 2009; *IFIH1)* LoF variants at *IFIH1* confer protection to disease
  - CAD (Cohen et al., 2008-2010; *PCSK9)* LoF variants at *PCSK9* reduce risk to CAD and reduced cholesterol levels.

Methods Intuition for QTL data

- Take individuals with LoF mutations in a particular gene, and ask whether the QT values for these individuals are extreme. (Typically only a few individuals will have LoF variants in any particular gene.)

- If QT values have been transformed to standard normal (for actual QT, or for residuals in an appropriate model), then under the null hypothesis that LoF in the gene in question has no effect, then the QT values of individuals with LoF will just be a sample from a standard normal distribution.

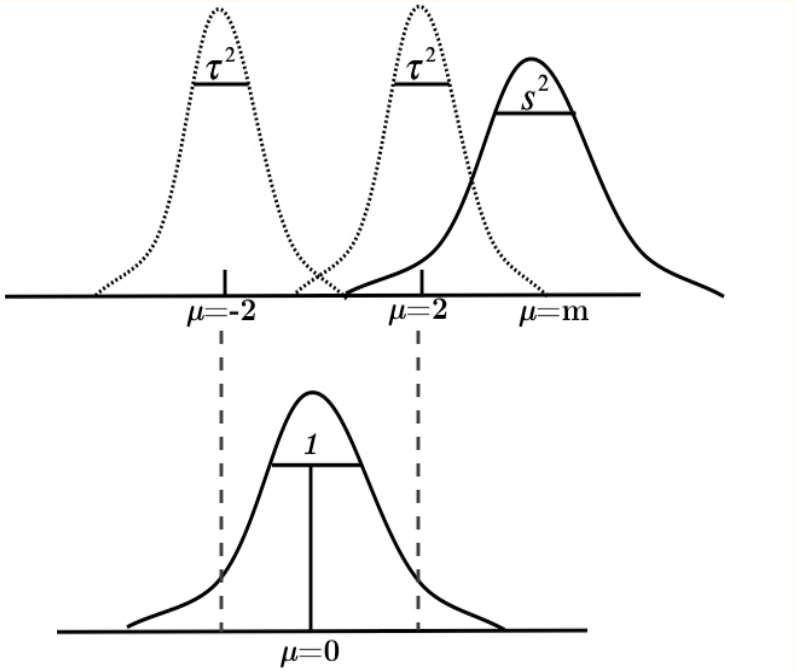
- A natural alternative hypothesis is that the QT values for individuals with LoF in that gene are normally distributed with a different mean.

- Consider a Bayesian approach to comparing these two hypotheses.

Method

Under the null hypothesis phenotypes of individuals with LoF mutations in a gene are samples from a standard normal with mean equal to 0 variance equal to 1 (Lower Panel).

Under the alternative hypothesis phenotypes of individuals with LoF mutations in a gene are drawn from a normal distribution at the tails. Our prior belief is that LoF mutations will have strong effects on observed phenotypes.



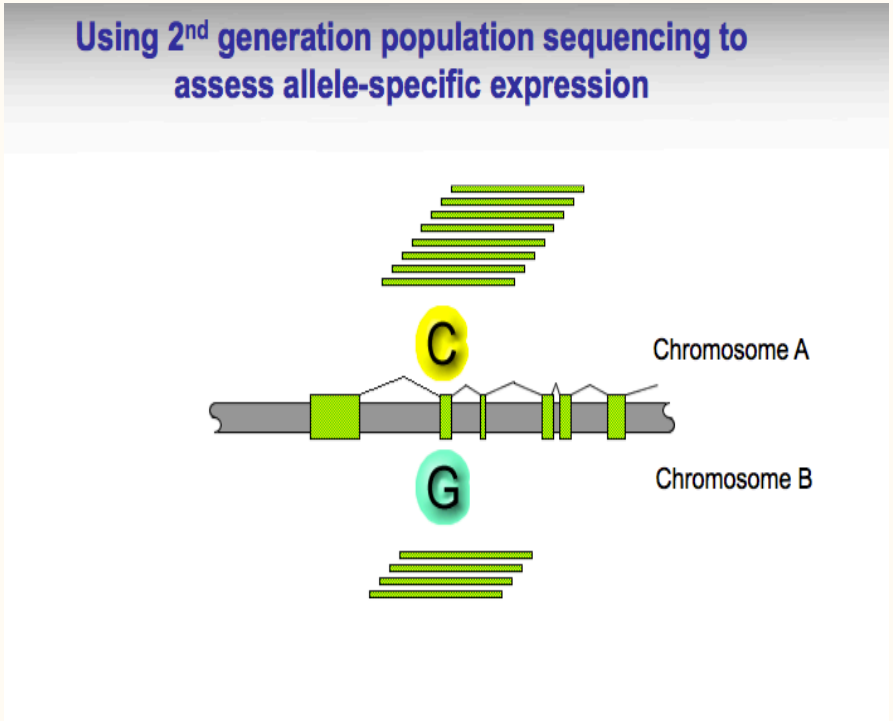
Approach is being applied to clinically relevant quantitative traits.
Not necessarily preferred way to analyze expression data.

Manuel Rivas, Matti Pirinen, and Peter Donnelly

ASE approach for analyzing expression data

- Montgomery et al (*PloS Genetics 2011)*.

For LoF variants may well have greater power to identify effects on expression.
With expression levels we can safely assume that LoF variants will decrease expression of transcript. Apply method that combines signal from LoF variants within a gene.



Manolis Dermitzakis

Could we use this approach to quantify how often nonsense variants escape NMD?

Identify alternative splicing motifs

In disease studies focus is on +/- 2bp of exon sequence boundaries as candidate splice variants

Identify alternative splicing motifs

Identify motifs that lead to intron retention and other alternative splicing events using empirically supported data.

What sequence motifs are good predictors?

At the moment not clear to me we know the answer.

Identify alternative splicing motifs

By integrating knowledge from the group and empirical data that supports motif enrichment for alternative splicing events may well be translated to disease sequencing based studies.