LoF update

Geuvadis RNA Sequencing Project

Annotation

The variant annotation follows the 1000 Genomes Project Style

Coding annotation is under the VA flag of the INFO field, and contains the following categories:
SNPs:
NONSYNONYMOUS, SYNONYMOUS, PREMATURESTOP, REMOVEDSTOP, REMOVEDSTART, SPLICEOVERLAP
Indels:
SPLICEOVERLAP, ENDOVERLAP, STARTOVERLAP, DELETIONFS, INSERTIONFS, DELETIONNFS, INSERTIONNFS

Transcripts are annotated using Gencode V7 transcript set for consistency. We should discuss updating the transcript set and freezing the version number.

Update from D. Macarthur on 1KG LoF annotation

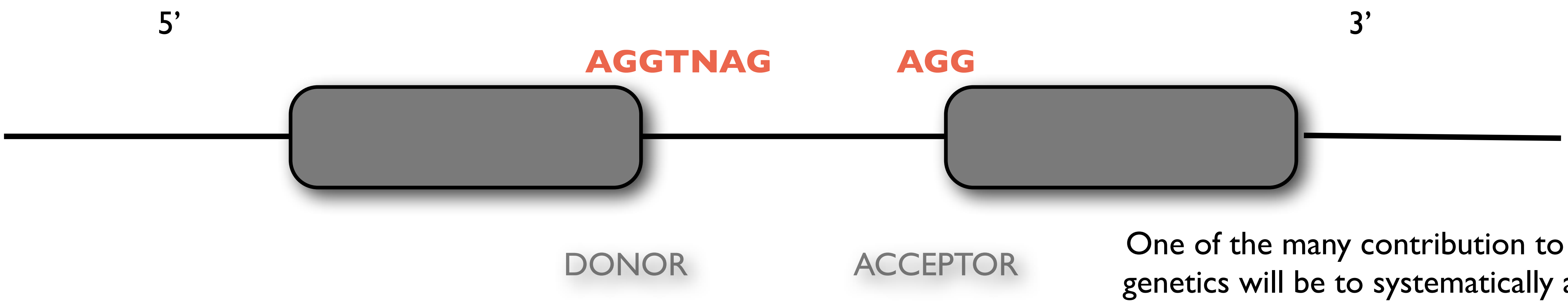
Described in Geuvadis RNASeq Wiki Page (Tuuli)
22/04/2012

Annotation

We propose the following additions/changes for the severe Loss of Function/Protein Truncating variant analysis:

Indels (currently annotated based on positional effect of indel - need more on sequence context work in progress):

1. DONOR_IN2 - Splice Donor Variants
2. DONOR_IN_45AG - Splice Donor Variants
3. DONOR_EX2_AG - Exonic Splice Variants
4. ACCEPTOR_IN2 - Splice Acceptor Variants
5. ACCEPTOR_EX1_G - Exonic Acceptor Splice Variants



5'

3'

AGGTNAG

AGG

DONOR

ACCEPTOR

One of the many contribution to the field of medica genetics will be to systematically assess the relevanc of predicted annotated splice variants: Exon Skipping Intron Retention, etc?
sQTLs will add additional categories that were probably excluded from analysis.

Additions discussed by Manuel Rivas and Tuuli L.

Categories follow Faustino et al. (2003) : Pre-mRNA splicing and human disease, and relevant to human disease: Jordan et al. (2012): PSOR2 is due to mutations in CARD14.

Annotation

Additional meta-information

1. worst=FLAG
2. nmd=BOOL (0 or 1) For frameshift indels and nonsense variants
3. ofptv=BOOL (0 or 1) For splice variants assuming exon-skipping maintain frame?
4. fsX=pepnumber (int) How many downstream peptides for next Stop.
5. pepsize=orig_new (string) For frameshift indels, nonsense, startlost, and readthrough variants: original peptide length, new peptide length.
6. exin=INT For splice variants: Nearest exon number in transcript. e.g. IVS11+1G>C
7. HGVS=string e.g. c.IVS11+1G>C

Implemented in v0.9 version of PLINK/SEQ software.

Additions discussed by Manuel Rivas and Tuuli L.

http://atgu.mgh.harvard.edu/plinkseq/

Annotation

LoF/PTV variant	Effect on Transcription	Detection
Nonsense	Nonsense Mediated Decay (NMD)	ASE, Exon Quantification, Transcript Quantification
Splice	NMD, Exon Skipping, Intron Retention	ASE (for exonic splice variants), Splice junction quantification, Exon + Intron Quantification, Transcript Quantification
Frameshift Indels	NMD	Transcript Quantification, Exon Quantification
SV/CNV	Gene Dosage Compensation, eQTL	Transcript Quantification, Exon Quantification
Start Lost	ASE/eQTL	ASE, Transcript Quantification, Exon Quantification
Readthrough	ASE/eQTL	ASE, Transcript Quantification, Exon Quantification

Quantifications

1. Exon Quantification

Early data upload made available by Tuuli with BWA mappings.

i. Quantifications from UNIGE are raw counts of reads over exons, calculated by in-house script made by Tuuli.

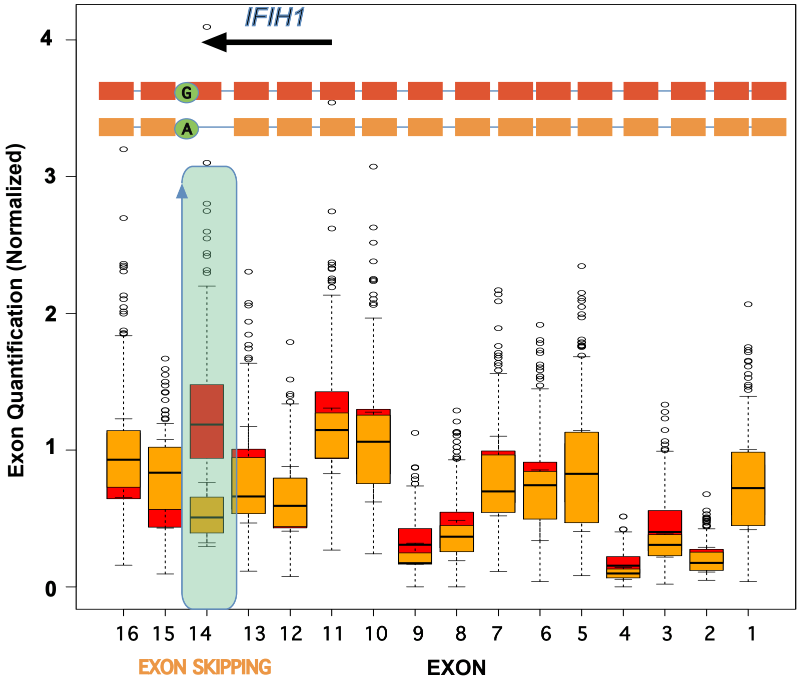
1 we take only uniquely mapped properly paired reads with bwa MAPQ>10 for both mates

2 we count reads only in protein-coding and linc-RNA transcripts of the annotation, because we're sequencing a poly-A library

3 what is an "exon": we merge all overlapping exons of a gene into a meta-exon, ID: ENSG000001_exonstartcoord_exonendcoord. This is to avoid problems with reads that map to several exons that overlap. Note though in the case of partially overlapping exons, our quantification units are not real exons but kind of meta-exons.

4. we count reads over these exons without using information of read pairing, except that we exclude reads where the pairs map to two different genes. We count a read in an exon if either it's start or end coordinate overlaps an exon.

5. in the case of split reads, we count the exon overlap of each split fragment, and add counts per read as 1/(number_of_overlapping_exons_per_gene) . I.e. if a read is split and the two parts map to 2 different exons of a gene, we count it as 0.5 in each.



Early p.o.p. of dataset interpretation demonstrates that exon-quantifications are reliable.

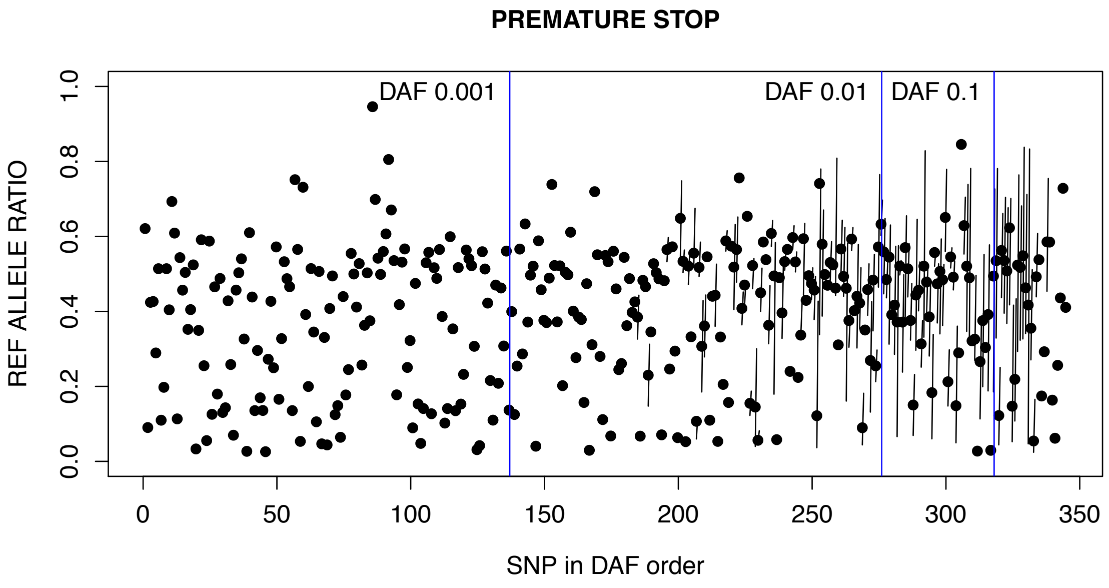
Taken from Geuvadis RNA-Seq wiki page (Tuuli).

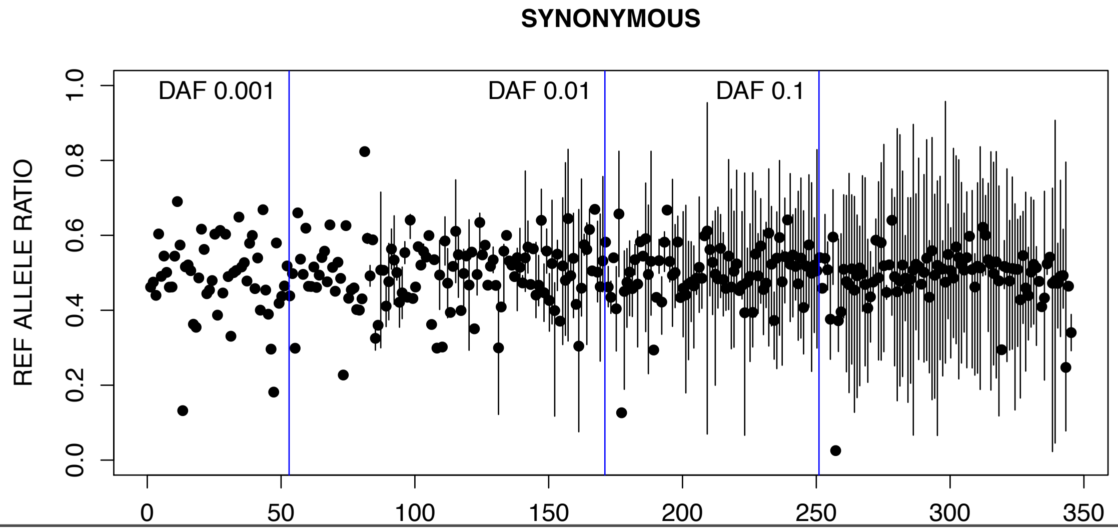
Methods/Analysis

Early analysis demonstrate strong ASE signal for nonsense variants.

Propose the following additional analyses:

1. Add layer of Nonsense Mediated Decay Predictions based on “50 bp termination code” rule.
2. Identify additional predictors of ASE/NMD signal:
   - i. Relative position to start
   - ii. Relative position to last splice junction
   - iii. Protein Size or Number of Exons
   - iv. Absolute Size of new transcript
   - v. Conservation Score
   - vi. Length of the Poly-A tail? (Is this measured)
   - vii. Proximity of New/Original Stop to 3’ UTR Start.
   - viii. Additional predictors?





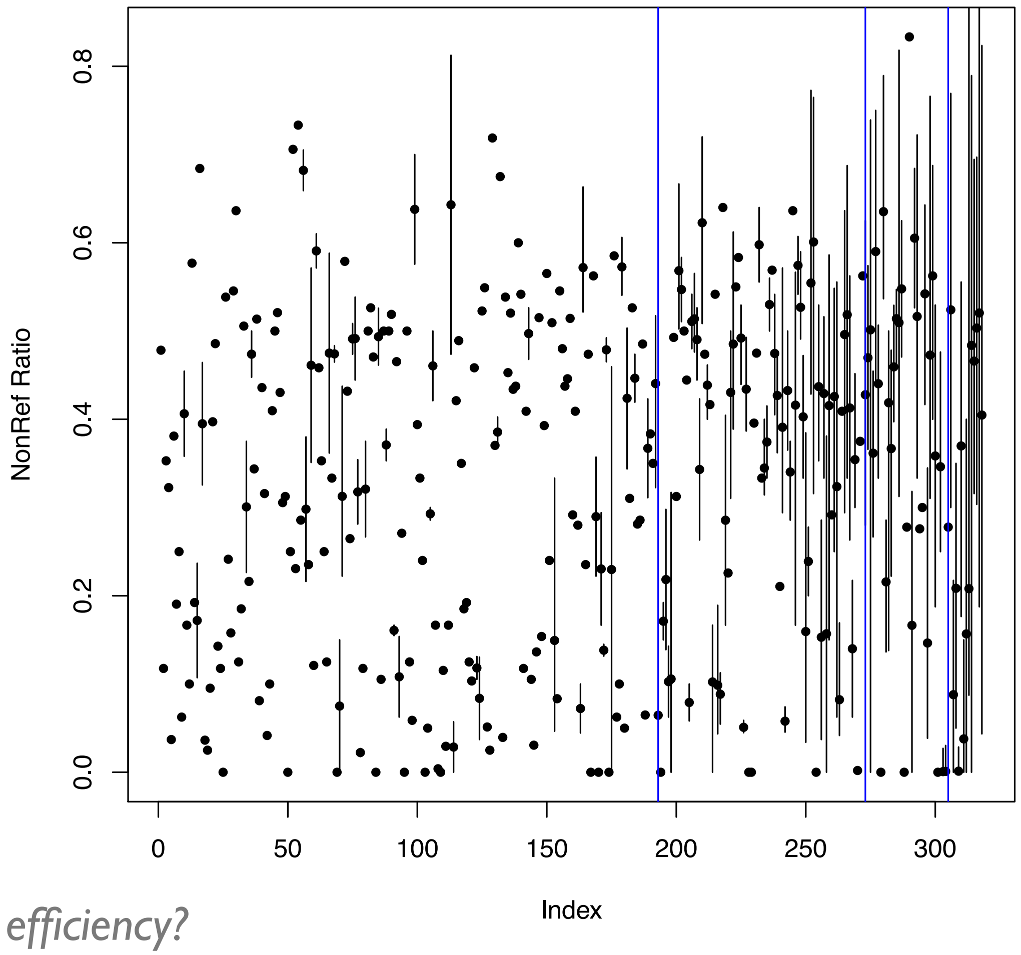
Taken from Geuvadis BOG Poster.

Methods/Analysis

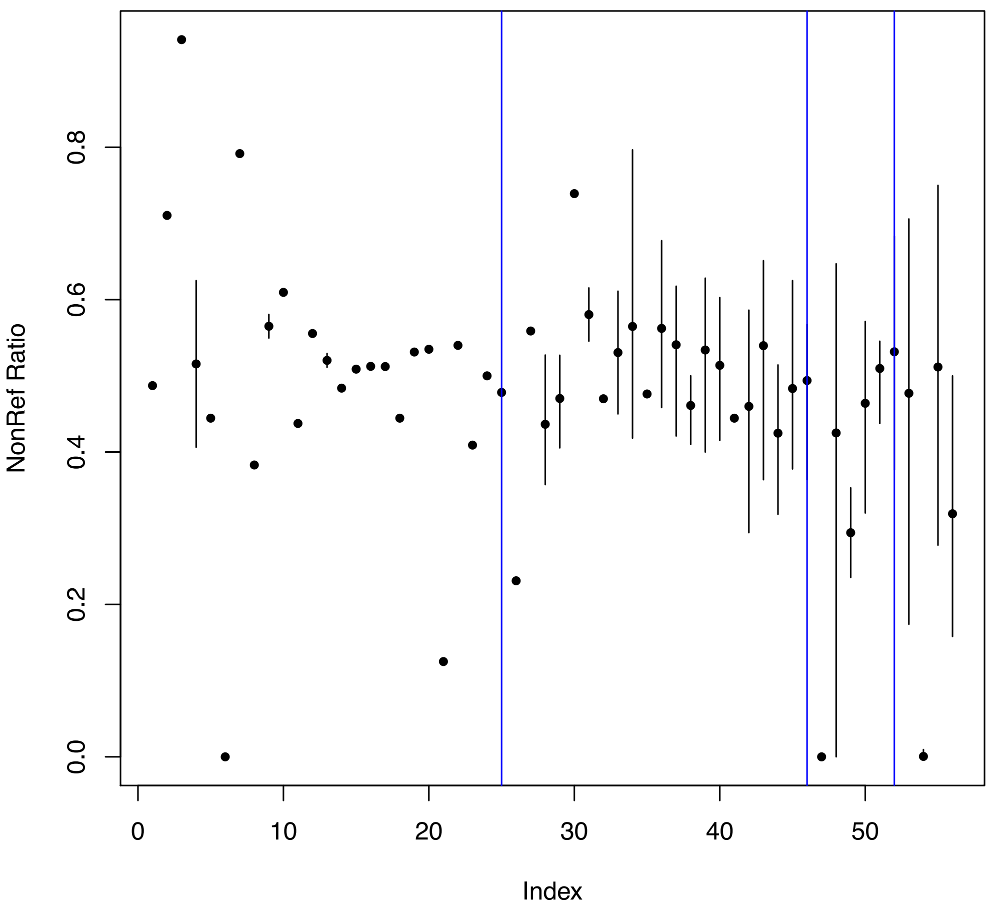
Early analysis demonstrate strong ASE signal for nonsense variants.

Propose the following additional analysis:
1. Add layer of Nonsense Mediated Decay Predictions based on “50 bp termination code” rule.

Predicted to Trigger NMD



Predicted to Escape NMD



Can we estimate NMD efficiency?