This directory contains lists of LoFs and functional (missense or inframe indels) from the current ELGH callset, released in November 2017.

The files contain genotype counts after the variants have either been through basic GATK filtering at the variant level (gatk\_PASS) or, subsequently, have also been through genotype-level filtering (gatk\_PASS.FS\_30.DP\_0.GQ\_20.AB\_0.01 🡪 removing sites with strand bias FS>30, setting to missing genotypes with GQ<20 or allele balance p-value <0.01). This genotype-level filtering is not optimal (being probably too strict on homozygotes in low-coverage regions), so will be improved at a later stage.

Most of the column headers come straight from the VEP annotation. See description of these here: <https://www.ensembl.org/info/docs/tools/vep/vep_formats.html> .

The files only contain variants that were present in at least one ELGH sample. The columns of main interest should be N\_Hom\_ELGH, N\_Het\_ELGH, AN\_ELGH, AF\_ELGH, which are the number of homozytoes, number of heterozygous, allele number (2\* number of individuals with non-missing genotypes at that variants), and allele frequency in ELGH, respectively. Also see N\_Hom\_BiB, N\_Het\_ BiB, AN\_ BiB, and AF\_ BiB, which are the equivalent numbers for Born in Bradford, and

N\_Hom\_Birm, N\_Het\_ Birm, AN\_ Birm, and  AF\_ Birm, which are the equivalent numbers for Birmingham.

For the LoFs, there are pairs of files. The \*all\_transcripts\_printed.txt files contain the annotations across ALL transcripts for which the variant is a LoF. In these files, take a look at these columns:

* which.exon and n.exons: indicate the exon that the variant falls in within this transcript (if it’s exonic), and the number of exons in the transcript (these can also be found in the EXON column, but it tends to be malformatted in Excel, so I have separated it into two extra columns).
* which.intron and n.introns: indicate the intron that the variant falls in within this transcript (if it’s intronic), and the number of introns in the transcript
* in.last.exon.or.intron: indicates whether the variant is in the last exon or last intron in this transcript (1=yes, 0=no)
* N\_coding\_transcripts: indicate the number of coding transcripts for this gene
* N\_coding\_transcripts\_with\_LoF: indicates the number of coding transcripts for this gene that have the LoF
* not.all.transcripts.have.LoF: indicates whether N\_coding\_transcripts== N\_coding\_transcripts\_with\_LoF (1=yes, 0=no)
* top.transcript indicates the first transcript for the gene for which not.all.transcripts.have.LoF==0 and in.last.exon.or.intron==0. You can use this to filter for one transcript per gene that fulfils these criteria.

The files \*annotation\_not\_in\_last\_exon\_and\_present\_in\_all\_transcripts.txt contain just the LoFs for which top.transcript==1 in the \*all\_transcripts\_printed.txt files.