

ADDITIONAL FILE 1 – WEB RESOURCES

Species conservation, amino acid severity, protein modeling and In silico functional analysis

Align GVGD - <http://agvgd.iarc.fr>

Align GVGD extends the Grantham matrix to incorporate multiple sequence alignments for a gene. Firstly the Grantham Variation (GV) is used to measure the extent of variation (i.e. species conservation) between residues at a position in the reference sequence alignment, then the Grantham Difference (GD) measures the distance between this group of residues and the new (variant) amino acid caused by the substitution.

- If $GD = 0$ it means that the variant falls within the range of variation in the sequence alignment so can be regarded as **neutral**.
- If $GD \neq 0$ and the GV is ≤ 61.3 then the variant is classified as **deleterious** because sequence variation is small and the variant lies outside its range.
- If $(0 < GD \leq 61.3)$ and $(GV > 61.3)$ sequence variation is greater and the variant lies close to its range, so is classified at **neutral**.
- Anything else is regarded as **unclassified**.

Details of Align GVGD application to P53 mutations can be found at <http://www-p53.iarc.fr/AGVGDmethod.html>.

The results of Align GVGD are highly dependent on the input reference sequence alignment. For this reason alignments have been carefully constructed and are available on the Align GVGD website – currently there are reference sequence alignments for BRCA1, BRCA2, CHEK2, ATM and P53. Note that these alignments are offered to different depths on the Align GVGD web site, e.g. BRCA1 and 2 can be human to frog or human to pufferfish. This can alter the result, even from deleterious to neutral, so the human to pufferfish alignment should always be used, and other gene alignments must be constructed similarly, in accordance with the guidelines given in section 4.5 above and on the Align GVGD website.

Sorting Intolerant from Tolerant (SIFT) – <http://blocks.fhcrc.org/sift/SIFT.html>

SIFT is based on reference sequence alignments and produces scores which can be classified as intolerant (0.00-0.05), potentially intolerant (0.051-0.10), borderline (0.101-0.20), or tolerant (0.201-1.00) according to the classification proposed by Ng et al. (2001) and Xi et al. (2004).

Polyphen (<http://genetics.bwh.harvard.edu/pph/>)

PolyPhen considers evolutionary conservation, physiochemical differences and the proximity of the substitution to predicted functional domains and/or structural features. Its scores can be classified as probably damaging (≥ 2.00), possibly damaging (1.50-1.99), potentially damaging (1.25-1.49), borderline (1.00-1.24), or benign (0.00-0.99)

according to the classification proposed by Xi et al. (2004). (Rudd *et al* 2005). The user's own alignment cannot be used as an input for this tool.

Splice Site Prediction

Berkley Drosophila Genome project www.fruitfly.org/seq_tools/splice.html

NetGene 2 - www.cbs.dtu.dk/services/NetGene2

Alex Dong Li's splice site finder - www.genet.sickkids.on.ca/~ali/splicesitefinder.html

GeneSplicer Web Interface www.tigr.org/tdb/GeneSplicer/gene_spl.html

Splice Sequence Finder (Montpelier) www.umd.be/SSF

Core variation databases:

HGMD - <http://www.hgmd.cf.ac.uk/ac/index.php>

This database collates published gene lesions responsible for human inherited disease. Each variant is represented only once. Silent variants and those with no phenotypic effect, somatic variants and base-level variants inferred from amino acid changes are excluded. Registration is necessary. Access is free via a web browser for academic/non-profit users but data are delayed by two years from their inclusion. Other users and those requiring up-to-date data must pay for a license to obtain access and associated software. The free version contained 53,186 entries in 2,056 loci, and the subscription version 69,734 entries in 2,577 loci, on 14 June 2007.

OMIM - <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM>

. OMIM includes selected variants with its gene entries, such as the first variant found, those with high population frequency, those with historical significance, etc. OMIM had 17,745 entries and covered 10,314 loci on 14 June 2007, though not all these have variation data present.

dbSNP - <http://www.ncbi.nlm.nih.gov/projects/SNP/>

dbSNP is a general catalogue of genome variation covering any species and part of a genome. Variants held include Single Nucleotide Polymorphisms (SNPs), Deletion Insertion Polymorphisms (DIPs), Short Tandem Repeats (STRs), Multiple Nucleotide Polymorphisms (MNPs), and NoVariations (regions that are invariant). Data are submitted from public and private laboratories and are clustered at individual genomic positions into reference SNPs (refSNPs). The build of dbSNP current on 14 June 2007 (build 127) contained 5,689,286 validated ref SNPs for the human genome.

Ensembl - <http://www.ensembl.org/index.html>

ESEfinder - <http://rulai.cshl.edu/cgi-bin/tools/ESE3/esefinder.cgi?process=home> (use with caution)

Russell - <http://www.russell.embl.de/aas/>

Publication databases:

PubMed - <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed>

Google Scholar - <http://scholar.google.com/>

The Web of Science. - <http://scientific.thomson.com/products/wos/>

Locus Specific Databases (LSDBs)

Lists of LSDBs can be found at the following sites:

HGMD (Cardiff) - http://www.hgmd.cf.ac.uk/docs/oth_mut.html. This list includes LSDBs and other databases such as disease related databases at the bottom of the table. It is not clear how up-to-date the list is and how often it is curated. The list is ordered by HUGO gene name with links to the databases and the name of the database host institution provided. The LSDB list included 306 entries on 18 May 2007.

HGVS - <http://www.hgvs.org/dblist/dblist.html>. The Human Genome Variation Society provides lists of database resources under various categories, including LSDBs, disease-centred central mutation databases, SNP databases etc. The LSDB list was added in March 2006 and currently has a 'last updated' date of March 2007. The list is ordered by HUGO gene name and provides a gene description, OMIM number link, link to the database and the name of the curator. Indexing and searching are provided. The LSDB list contained 678 entries on 18 May 2007.

BIC: BRCA1 and BRCA2 variants

Gene-Specific Mutation Databases and WWW sites

<http://wgen.eimb.relarn.ru/databases/genespec.htm>:

www.med.mun.ca HNPCC

www.insight-group.org HNPCC

Mutation Nomenclature:

www.hgvs.org

www.LOVD.nl/mutalyzer