Melbourne Genomics
Health Alliance

Glossary

Terminology for genomics and variant interpretation

Term	Definition
A	
Allele	Variant forms of a gene occupying the same genetic locus.
Alternative splicing	Different combinations of exons spliced together to generate more than one possible mature transcript which may lead to more than one protein product from one gene.
Amino acid	Molecules used to build a protein. Properties of amino acids (size, charge, hydrophobicity) determine protein folding, structure and function.
Annotate	Note information about the variant, such as chromosome and gene location and predicted effect on protein structure or function.
Annotation	Adding information to a variant (or other biological entity) to provide more information about structure or function.
Autosome	A chromosome that is unrelated to the sex of an organism.
B	
Bam file	Dataset of aligned reference and query DNA sequences.
Biallelic	Condition caused by having variants/mutations on both copies of the gene (i.e. on both alleles). Affected individual could be homozygous or compound heterozygous.
Bioinformatician	Bioinformatician – a scientist specialising in bioinformatics.
Bioinformatics	A field of biology that uses algorithms and software to analyse biological data, and the use of such data to make biological discoveries, construct models or make predictions.
C	

Call (a variant)	The process of identifying a variant from sequence data. The sample genome, exome or gene is sequenced, aligned to a reference genome and differences in the sample are ‘called’ as variants.
Canonical splice site	Two bases at either side of the intron 5'GU---AG3' [referred to as the donor site at 5’ end of intron and acceptor site at 3’ of intron] recognised by small ribonuclear proteins to cut the introns out of mRNA.
Cascade screening	Genetic testing of biological relatives of an individual with a pathogenic variant, to identify individuals carrying the variant and the risk of developing a condition or passing a variant on to their offspring.
cDNA	A DNA molecule that is the complementary sequence of an mRNA; a transcript of mRNA produced in a laboratory using reverse transcriptase.
Chromosome	DNA molecule coiled around histone proteins and further coiled into a compact structure visible under the microscope.
Chromosomal microarray	A molecular test to identify structural changes in chromosomes, such as aneuploidy and copy number variants.
Cis – trans	‘in cis’ – on the same strand; ‘in trans’ on different strands In relation to gene variants: Two different variants on the same allele are ‘in cis’. Two different variants on different alleles of the gene are ‘in trans’.
Codon	Group of 3 bases in messenger RNA that specifies an amino acid.
Compound heterozygous; compound heterozygote	The presence of two different variants at a locus, one on each of the paired chromosomes; having two different recessive alleles at a locus that can cause genetic disease when inherited together.
Conservation	The degree of similarity between a gene or protein sequence across species. High conservation of a region implies the sequence is essential for function; variants in a conserved region are more likely to have a major effect on gene expression or protein function.
Constraint	A limit on the ability of a DNA region to tolerate mutation/variation and be retained in the organism; e.g. some regions of a gene have few or no variants – they are ‘constrained’, the region does not tolerate change, probably because the change is deleterious.
Copy number variant (CNV)	An abnormal number of copies of a section of DNA, including large sequence duplications.

D

D	
De novo	“new”; a variant that occurs in a gamete (during meiosis), early in embryo development, or in somatic tissues is a *de novo* variant; it will be seen in the individual but not the parents.
Deletion	Deletion of one or more nucleotides from a DNA sequence.

Delins	A deletion and insertion in close proximity on a DNA strand that produces a new variant [Melbourne Genomics usage[1]].
DNA	Genetic material of life on earth. Built from 4 nucleotides – adenine (A), cytosine (C), guanine (G) and thymine (T) joined in strands by phosphodiester bonds. Exists as a double stranded molecule (double helix) of complementary base pairs A-T and C-G.
DNA sequence	The order of the nucleotide bases in a DNA molecule, usually recorded in the 5' & 3' direction.
Dominant negative	Where a variant/mutation causes a gene product to counteract or adversely affect the normal gene product in the cell
Driver mutation	In cancer, a gene with variant(s) that increase the rate of cell replication.

E

E	
Epigenetics	Heritable DNA modification that alters gene expression without changing the DNA sequence or genetic code. Commonly methyl- and acetyl-groups attached to the DNA molecule or histones.
Exome	The portion of the genome that includes all the exons of all genes (all the protein coding portions of the DNA).
Exon	Protein coding region of a gene.

F

F	
Fastq file	Data file for the raw DNA sequence.
Frameshift	A change in the 'reading frame' (groups of 3 nucleotides) of a gene. An insertion, deletion or indel that is not a multiple of 3 nucleotides will produce a frameshift.
Fusion (gene/protein)	A gene made by joining sections of two different genes; codes for a fusion protein. A common genetic variant in cancer.

G

G	
Gene	A section of DNA that carries the code for a protein or RNA molecule.
Gene expression	Gene to protein; Transcription and translation
Gene list	A list of candidate genes associated with a phenotype.

[1] Some genetics and bioinformatics terms have variable or debated meaning. Definitions given here are those used by Melbourne Genomics.

Gene structure	Elements of a gene, includes coding sequence - introns and exons, promoters, regulatory regions, untranslated regions (UTRs).
Genome	All the genetic material of an organism; all the DNA, including all the genes. The human genome is about 3 billion DNA base pairs & around 20,000 protein coding genes.
Genotype	The genetic makeup of an individual comprising all the alleles at all genetic loci.
Germline variants	Genetic variants present in gametes and potentially inherited by offspring

H

H	
Haplotype	A group of alleles or SNPs occurring close to each other on a chromosome and tend to be inherited together (linked).
Hemizygous	Having one copy of a gene as a result of having one copy of the chromosome, such as the genes on the X-chromosome in males; or loss of alleles due to deletion of a section of chromosome.
Heteroplasmy (mitochondrial)	An individual with more than one type of mitochondria, carrying different genetic sequence or different mitochondrial variants.
Heterozygous; Heterozygote	For a diploid individual, having two different alleles at a locus.
Homology	(for genes) the extent to which a DNA sequence is the same.
Homopolymer (DNA)	A repeat sequence of a single nucleotide in DNA; poly(dA), poly(dT), poly(dC) or poly(dG).
Homozygous; Homozygote	For a diploid individual, having two identical alleles at a locus.

I-J

I	
In silico tools; *in silico* scores	Online databases and computational tools to predict the effect of variants on protein structure and function, homology and conservation. Scores are calculated for variant curation.
Indel	A variation caused by an **in**sertion or **del**etion. Collective term for insertions and deletions [Melbourne Genomics usage].
Insertion	Addition of one or more nucleotides to a DNA sequence.
Intron	Intervening sequence – DNA that intervenes between two exons; regions of a gene that do not code for protein.

K-L

K

Karyotype	Arrangement of chromosomes showing the number and structure of the set of chromosomes in a species or individual.

M

M	
Mendelian (inheritance)	Inheritance patterns of characteristics due to a single gene (monogenic conditions), e.g. recessive, dominant, X-linked.
Mendeliome	Around 4000 genes known to carry variants that cause monogenic conditions (Mendelian inheritance)
Microarray	*See chromosomal microarray*
Missense	Genetic variant (nucleotide substitution) causing a change in an amino acid in the resulting protein. Also called non-synonymous.
Monogenic	Condition or phenotype caused by a variant in one gene
Mosaic variant	A variant present only in some cells of the individual.
mRNA	Messenger RNA produced by transcription of the template strand of a gene. The primary transcript or precursor (pre-mRNA) contains intron and exon sequence. Introns are sliced out to produce mature messenger RNA (mRNA).
mRNA Splicing	Editing of primary transcript/pre-mRNA to remove the intron sequences and join exons.
Multigene panel test	Laboratory test of several candidate genes known to cause a condition/phenotype; used to identify pathogenic variant.
Mutation	A change in DNA sequence; ‘permanent’ change in DNA sequence [Melbourne Genomics usage].

N-O

N	
Next generation sequencing (NGS)	High-throughput DNA sequencing technology (non-Sanger sequencing method) for genomic sequencing (whole genome, whole exome); also called massively parallel sequencing. Sequence many genes at once.
NMD – Nonsense mediated decay	Cellular pathway to breakdown mRNA carrying a non-sense variant, i.e. mRNAs with a premature stop codon. Nonsense variants downstream of the last 50 nucleotides of the second last exon may not cause nonsense mediated decay.
Nonsense	Genetic variant that causes a premature stop codon, producing a short/truncated protein product; can cause NMD (*see above*).
Non-synonymous	Genetic variant that changes a codon and results in a change of amino acid in the protein. Also called missense.

Nucleotide	Component of nucleic acid, comprised of sugar, phosphate and nitrogenous base. The base components in DNA are adenine (A), cytosine (C), guanine (G) and thymine (T); in RNA: adenine (A), cytosine (C), guanine (G) and uracil (U).
Orientation of DNA strands: plus (+) strand, minus strand (-)	For a given gene in double stranded DNA, the 5'-3' strand with the code for protein is designated the plus (+) strand, coding strand or sense strand. The complementary 3'- 5'strand for the gene is the minus (-) strand, or non-coding or anti-sense strand.

P-Q

P	
Panel	see 'multigene panel test'
Pathogenic	Disease-causing. A pathogenic variant affects cell function and causes disease.
Pedigree	Chart with symbols representing inheritance over 2 or more generations of a family.
Phasing	Distinguishing whether an allele or variant is on the maternal or paternal chromosome.
Phenotype	The physical appearance and physiology of an individual, resulting from expression of the genotype and influenced by environmental factors.
Phred score	Base call quality score; provides an estimated probability of an error in the base call at that location.
Plus (+) strand, minus (-) strand	DNA orientation. For a given gene in double stranded DNA, the 5'-3' strand with the code for protein is designated the plus (+) strand, coding strand or sense strand. The complementary 3'-5'strand for the gene is the minus (-) strand, or non-coding or anti-sense strand.
Polymorphism	Variant that occurs frequently in a population; e.g. frequency >1%
Polyploid	Cells containing more than two sets of homologous chromosomes
Proband	The individual through whom a family with a genetic disorder is ascertained. The first person in a family identified with a genetic disorder.
Protein	Molecules encoded by genes, comprised of amino acids in a sequence specified by the gene sequence. Amino acid sequence determines protein folding and function.
Pseudogene	An inactive version of a gene; originating as a functional protein-coding gene but altered by mutations through evolution.

R

R	

Reads	The sequencing copies of a DNA sequence. Many reads of the same DNA region are needed for reliable variant identification compared to a reference genome.
Reference sequence or genome	A ‘representative’ sequence of a gene or genome for comparison to individual gene or exome sequences.
Refseq	A database of reference sequences that have an empirical (rather than predicted) basis to them. Usually used in the diagnostic setting.
Regulatory gene	A gene encoding a protein that controls expression of other genes.
Regulatory sequence	DNA sequence involved in controlling when genes are expressed.
RNA processing	Modification of the primary transcript, including splicing, addition of 5’CAP and 3’ poly-A tail to produce mature mRNA.

S

S	
Sanger sequencing	Method of determining the order of nucleotides in DNA, one gene at a time. Used to confirm variants and single gene sequence.
Segregation studies	Genetic testing of parents/grandparents etc. of an individual with a pathogenic variant, to gain information on mode of inheritance of the variant, e.g. *de novo*, recessive, dominant, and pathogenicity
Sex chromosome (allosome)	In mammals X chromosome and Y chromosome.
Sex-linked	Genes located on the sex chromosomes (X or Y chromosomes).
Single gene test	Laboratory test to identify variants in one gene associated with a phenotype and clinical presentation.
Singleton	Sequencing and variant curation performed on the individual subject; as compared to trio analysis, sequencing affected individual and parents.
SNP, Single nucleotide polymorphism	A singe base pair in DNA that shows polymorphism (i.e. has alternate alleles) in a population.
SNV, Single nucleotide variant	Single base difference between individuals in a population.
Somatic variant	A change in DNA that occurs after fertilisation of egg and sperm and is not present in the germline
Splice site	Two bases at either side of the intron 5'GU---AG3' [referred to as the donor site at 5’ end of intron and acceptor site at 3’ of intron] recognised by small ribonuclear proteins to cut the introns out of mRNA.
Splice site variant	A genetic alteration in the DNA sequence at the boundary of an exon and intron (the splice site). This change can disrupt RNA splicing resulting in the loss of exons or the inclusion of introns and an altered protein-coding sequence.
Structural gene	Gene coding for an RNA or protein (but not a regulatory protein).

Structural variant (SV)	Large deletions, insertions, inversions, translocations, gene fusions and gene duplications.
Substitution	Variant where one nucleotide is replaced by one other nucleotide.
Synonymous	Genetic variant (nucleotide substitution) that changes a codon but not the amino acid in the protein (also called silent variant).

T

T	
Transcript	The RNA produced by transcription of a gene; variant forms of the gene and alternative splicing produce different transcripts.
Translation	Process of the ribosome reading the mRNA to bring correct amino acids to produce a polypeptide/protein
Trinucleotide/Triplet repeat	3 consecutive nucleotides that repeat in tandem at one location. Also called triplet repeat expansion.
Trio	Sequencing for variant curation performed on the individual subject and both biological parents.

U-V

U-V	
Uniparental disomy	In an individual, two copies of a chromosome (or part of a chromosome) come from one parent and none from the other parent.
UTR	Untranslated regions located 5' (upstream) and 3' (downstream) to a gene. Involved in regulation of gene expression.
Variant	A variation in DNA sequence as compared to a 'reference sequence'. Range from single base change to large rearrangements of DNA.
Variant classification	The result of weighing up curation evidence and categorise the confidence associated with the variant being pathogenic or benign. Classifications used are typically: 5-Pathogenic, 4-Likely Pathogenic, 3-Variant of Uncertain Significance, 2-Likely Benign and 1- Benign. Subclasses of class 3 can also be used.
Variant curation	The process of gathering evidence for and against a variant being pathogenic or benign.
Variant interpretation	Combining the clinical information with the variant classification.
VCF file	Data file format for 'called' variants.
VUS (VOUS), variant of uncertain significance	A change in DNA sequence where it is unclear whether it is disease-causing, i.e. whether it is pathogenic or benign.

W-Z

W-Z

WES, whole exome sequencing	Determining the sequence of all the exons in a genome.
WGS, whole genome sequencing	Determining the sequence of all the DNA (coding and non-coding).
X-inactivation	Inactivation of one copy of the X-chromosome in female XX mammals (placentals and marsupials).
Zygosity	The degree of similarity of the alleles at a locus, usually defined by the terms homozygous, heterozygous or hemizygous.

Online genomics glossaries

Organisation	**URL**
Scitable (Nature Education)	**https://www.nature.com/scitable/glossary**
Australian Genomics Health Alliance – Genomics Glossary	https://www.australiangenomics.org.au/for-participants/genomics-glossary/
JAMA Genomics glossary	https://jamanetwork.com/journals/jama/fullarticle/1677346
National Cancer Institute, NIH, Dictionary	https://www.cancer.gov/publications/dictionaries/genetics-dictionary/
National Centre for Biotechnology Information (NCBI), National Institutes of Health, USA	https://www.ncbi.nlm.nih.gov/projects/genome/glossary.shtml
European Bioinformatics Institute – Glossary	https://www.ebi.ac.uk/training/online/glossary#letter_b