Melbourne Genomics
Health Alliance

Clinical Genomics: the Background Biology

Global knowledge. Individual care.

melbournegenomics.org.au

Alliance members

The Royal Melbourne Hospital
The Royal Children's Hospital Melbourne
The University of Melbourne
Walter+Eliza Hall Institute of Medical Research
murdoch children's research institute
CSIRO
agrf
Peter Mac Peter MacCallum Cancer Centre Victoria Australia
Austin Health
Monash Health

Supported by

Victoria State Government

Melbourne Genomics
Health Alliance

Global knowledge. Individual care.

This book was prepared for clinicians completing the Melbourne Genomics Health Alliance blended learning course **Genomics in the Clinic: An Introduction for Medical Professionals**.

View this book in your pdf viewer to be able to access the accompanying hyperlinked videos identified by the symbol  . You can also download to print a hard copy.

By accepting a place in this free course, you agree not to distribute the course resources, including this book.

Authors and acknowledgements

Author: Fran Maher, Education Officer, Melbourne Genomics Health Alliance

Acknowledgements: Thanks go to Taryn Charles, Dr Zoe McCallum and Dr Chloe Stutterd and Natalie Thorne for review, suggestions and comments for improving this resource; also to Nat Thorne, Sebastian Lunke, Tiong Tan, Ain Roesley and Andrew Fellowes for the use of slides, used as starting points for several images for this book and modified for presentation/video clips appearing as hyperlinks for this book and used in the accompanying online course.

Contents

1 Introduction to the human genome and genomic variants

The human genome comprises all the genetic information in a human cell. This includes the DNA in the chromosomes that reside in the cell nucleus and in the mitochondrial DNA (Figure 1).

The human genome contains 3,000,000,000 DNA nucleotide pairs, which includes 20-22,000 protein-coding genes distributed across a set of chromosomes.

The genome also contains genes for non-protein-coding RNA, including ribosomal RNA, transfer RNA and microRNAs. Interspersed between the genes is a large amount of non-coding DNA.

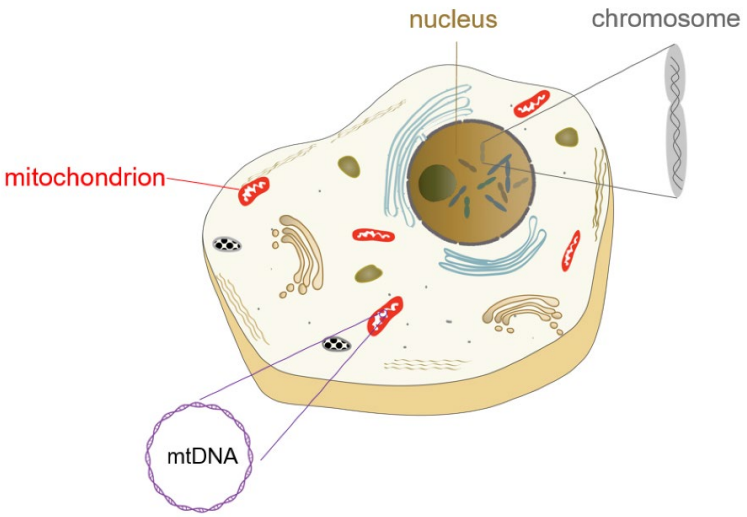


Figure 1 Eukaryotic cells have DNA in the form of linear chromosomes located in the nucleus and circular mitochondrial DNA (mtDNA) located in the mitochondria

Genomic Variants – an overview

Variation occurs at different levels throughout the genome, from changes to whole chromosome number, to changes in smaller regions within chromosomes affecting one or several genes, down to single nucleotide changes (Figure 2). Large scale changes to chromosomes usually result in clinically relevant conditions. This is not always the case for variants at the single nucleotide level. We all have many single nucleotide variants. Most small changes simply contribute to the normal range of variation in the population. However, some variants adversely alter the proteins encoded by genes, affecting cell function and causing disorders.

Range of genomic variation

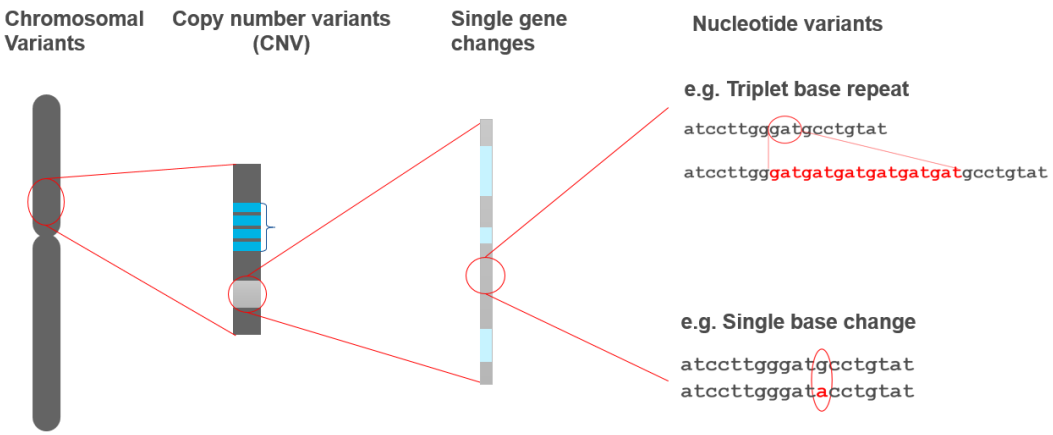


Figure 2 Genomic variation occurs at different levels, from whole chromosome to single nucleotide changes

Genomic variants all start as a mutation event in the DNA (note that we usually now use the term variant to describe the changes in DNA). Depending on the developmental stage and cell type where the change occurs, variants are present either in germline cells and gametes (eggs and sperm) and are potentially inherited (the so-called germline variants), or present only in somatic cells and are not inherited.

New *(de novo)* variants can arise very early in development before all the tissues and organs are differentiated. These early variants (so-called post-zygotic *de novo* variants) may affect only one or a few cell lineages; the range of cell types affected depends of the timing of the variant during development. For example, some forms of epilepsy are caused by this type of genetic change. *De novo* variants can also arise in mature somatic cells (so-called somatic *de novo* variants) and some of these lead to cancer. Somatic cancer variants are not inherited.

2 DNA and Chromosomes

DNA structure

The DNA double helix is built from paired nucleotides, forming nuclear chromosomes and mitochondrial DNA.

Nucleotides are composed of smaller units: a phosphate group, a deoxyribose sugar, and a nitrogen-containing base. DNA uses four bases: adenine (A), guanine (G), cytosine (C) or thymine (T). In a double stranded DNA molecule, the bases form complementary base pairs (bp) by hydrogen bonding; A-T and C-G, forming the complementary strands of the double helix (Figure 3). The length of a DNA molecule is described in terms of the number of base pairs (bp).

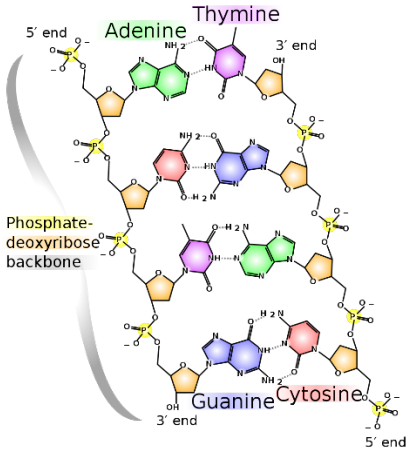


Figure 3 Chemical structure of DNA[1]

Chromosomes long and short:
Chromosome 1, the longest with ~249,000,000 bp and ~2100 genes;
Chromosome 21, the shortest, with ~47,000,000 bp long and 200-300 genes
Genes big and small:
The TTN gene is 281,434 bp and codes for a very large protein, titin,
The INS gene is 1,430 bp and codes for insulin, a small protein

Chromosomes in the human genome

A chromosome is one linear DNA double-helix molecule that contains many genes (protein coding regions). The DNA is wound around proteins called histones, to condense the DNA. The **centromere** is where the chromosome attaches to the mitotic spindle during cell division. The **telomeres** stabilise and protect the ends of the chromosome and prevent the ends sticking together.

The set of human chromosomes

- human cells have 46 chromosomes (2 sets of 23, 1 set from each parent)
- one set of 23 chromosomes has 22 autosomes (non-sex-determining chromosomes) and 1 sex-determining chromosome (X or Y)

[1] Image source: Madeline Price Ball, Wikimedia Commons CCO https://commons.wikimedia.org/wiki/File:DNA_chemical_structure.svg

- chromosomes vary in size and number of protein-coding genes (see ⓘ box)
- the chromosomes are contained in the nucleus of eukaryotic cells
- human somatic cells have 2 sets of 23 chromosomes - they are diploid
- gametes (ova and sperm) contain only one set of chromosomes - they are haploid

Karyotypes and Ideograms

Chromosomes can be seen under the microscope after staining. This G-banding staining method for karyotype analysis is performed on dividing cells when the chromosomes replicate and condense to form the characteristic 'X' shape frequently shown in diagrams. A karyotype (Figure 4) shows the set of chromosomes arrangement in **homologous pairs** (i.e. one of each chromosome from each parent; homologous chromosomes have the same genes at the same loci) and in order of size from chromosome 1 to 22 (the autosomes), followed by the sex chromosomes, X and Y.

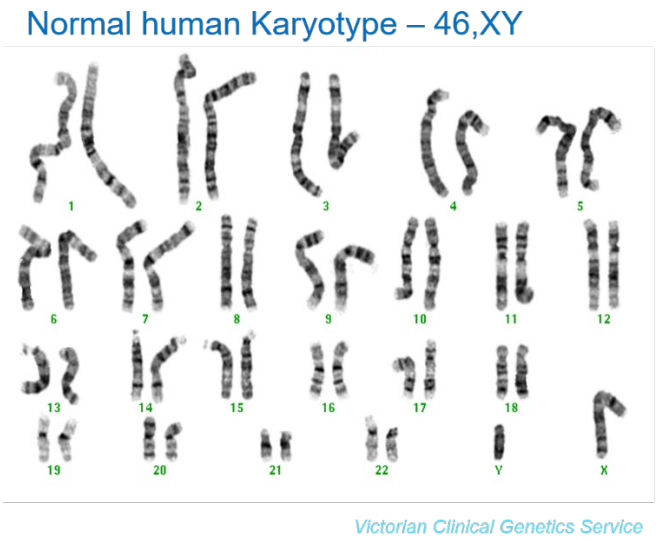


Figure 4 In a karyotype, chromosomes are arranged in homologous pairs, in order of size and position of the centromere, with the shorter 'section of the chromosome (the p arm) uppermost.

Nomenclature: chromosome number and gene location

Chromosome nomenclature (Figure 5a) includes:

- Chromosome number and length in DNA base pairs (bp)
- Short arm (p) and long arm (q)
- Cytogenetic locations - numbered sections defined by the G-banding pattern

The molecular location of a gene gives the nucleotide position within the chromosome (e.g.Figure 5b):

- the gene for amyloid precursor protein (APP) is located on chromosome 21
- Chromosome 21 length = 46,709,983 base pairs (bp)
- Cytogenetic Location: 21q21.3 = q arm of chromosome 21 at position 21.3
- Molecular Location: base pairs 25,880,550 to 26,171,128 on chromosome 21

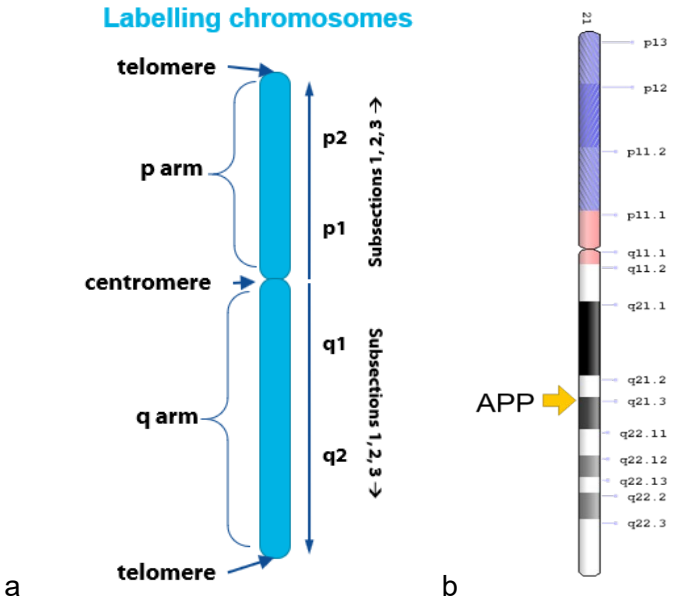


Figure 5 Chromosome structure and labelling: (a) 'p arm'– short arm, 'q arm'– long arm, Centromere – where chromosomes attach to spindle fibres during mitosis and meiosis (cell division). Chromosome numbering goes from centromere towards telomere for each arm. Telomeres are chromosome ends. (b) Ideogram[2] of chromosome 21 showing the location of the APP gene at 21q21.3 (long arm, section 2, subsection 1.3).

Mitochondrial DNA (mtDNA)

Mitochondria are the energy generating organelles in eukaryotic cells. They contain small circular DNA molecules, mtDNA (Figure 6). The genes on mtDNA are all essential for normal mitochondrial function. Nuclear chromosomes also carry genes necessary for mitochondrial function. Thus, a genetic cause of mitochondrial disfunction could be the result of chromosomal or mitochondrial DNA mutations.

Cells can carry both 'normal' and mutated mitochondria at the same time. This feature, called heteroplasmy, can present challenges in detecting disease-related mitochondrial DNA variants.

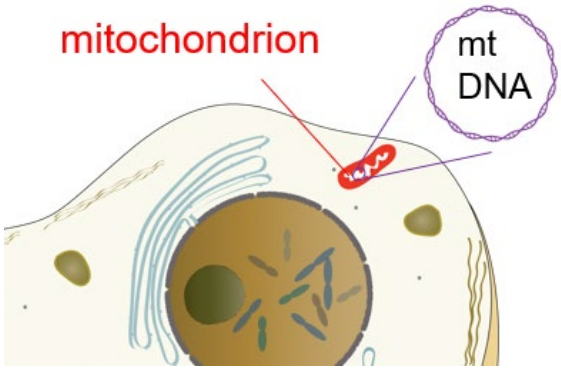


Figure 6 Mitochondria contain multiple copies of the circular mitochondrial DNA (mtDNA). mtDNA replicates independently of the cell. Human mtDNA is 16,569 bp long and carries 37 genes (13 protein-coding genes and 24 non-protein-coding genes).

[2] Ideogram source: https://ghr.nlm.nih.gov/gene/APP#location

3 Genes

DNA controls what happens in cells because it contains the code for all cellular proteins (proteins are built from amino acids). The sequence of nucleotide bases in a gene specifies the amino acid sequence of a protein. Genes contain coding and non-coding regions, and regulatory elements which control the timing and cell specificity of gene expression (Figure 7).

Gene structure

- Within the coding region of a gene are **exons** and **introns**
- Exons have the nucleotide sequence that codes for the protein
- Introns, also called intervening sequences, do not carry information for the protein; they are spliced out of the RNA before the protein is produced
- Mitochondrial and bacterial genes lack introns
- Regulatory elements include promoters and enhancers, which can be upstream or downstream of the coding sequence
- Untranslated regions (UTR, not illustrated here) lie either side of the coding sequence.

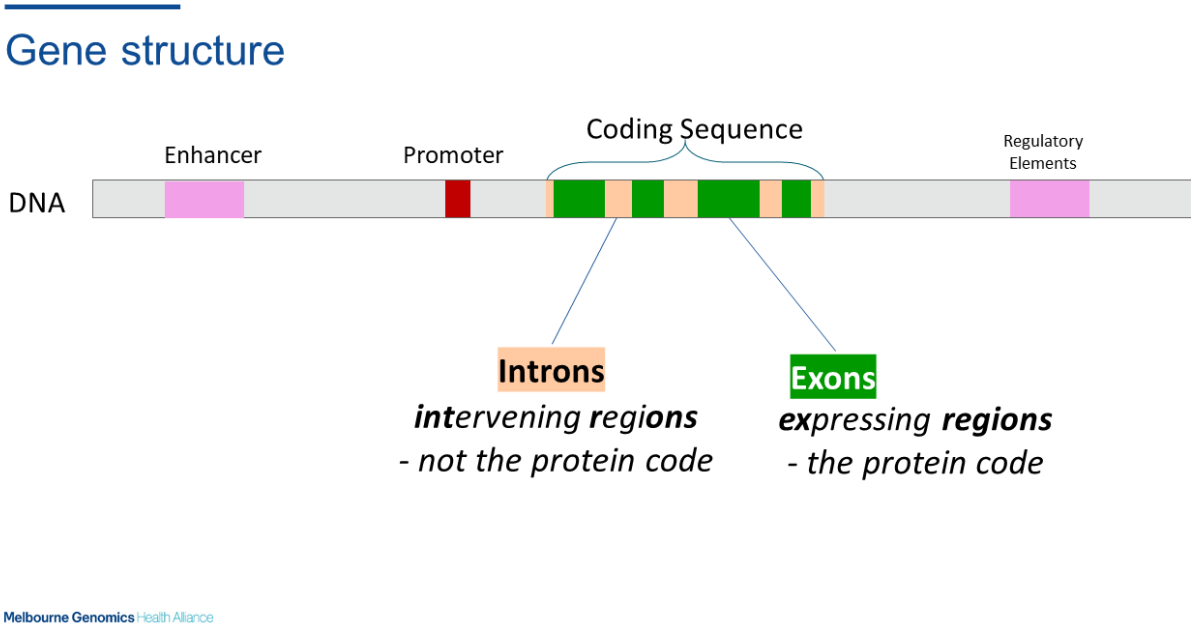


Figure 7 Structure of a gene. The coding region contains exons (protein coding sequence) and introns (intervening sequence). Regulatory regions upstream and/or downstream of the coding region include the promoter and enhancers.

3 Gene structure

4 Reading the Code

Gene Expression

Gene expression refers to the process of 'turning on' a gene so that the code in the DNA can be read and translated to produce a protein. The DNA remains in the nucleus. An RNA copy enters the cytoplasm for protein synthesis to occur. The process involves **transcription** – copying the DNA sequence into RNA, followed by **translation** – reading the RNA and translating the nucleotide code to amino acids to build the correct protein. The overall process is summarised below and Figure 8.

DNA (gene) —*transcription*→ mRNA —*translation*→ Protein

The steps in gene expression (Figure 8) include the following:

- Transcription occurs in the cell nucleus. One strand of the DNA double helix is transcribed into RNA, the primary transcript
- The primary transcript of RNA is edited and modified (see 'Splicing' below) to produce a mature messenger RNA (mRNA)
- mRNA moves through pores in the nuclear membrane into the cell cytoplasm and interacts with a ribosome. The sequence of bases in the mRNA is read as groups of 3-bases (codons)
- Transfer RNA molecules (tRNA) bring the appropriate amino acid to the ribosome to match the 3-base codons in the mRNA; this matching occurs by the 'anticodon' on tRNA forming complementary base pairs with the codon on the mRNA (RNA base pairing is A-U and C-G)
- Amino acids are linked by peptide bonds, forming a polypeptide chain
- The polypeptide chain then folds into a functional protein.

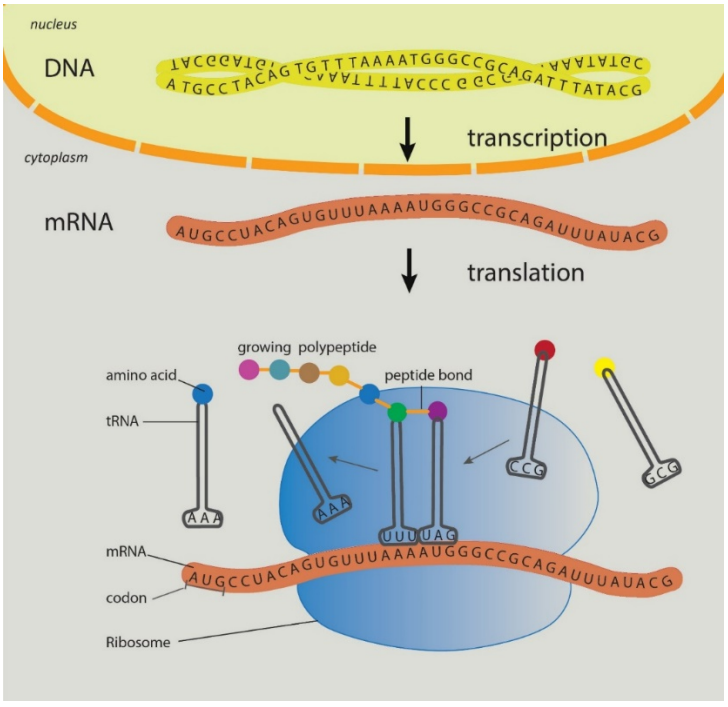


Figure 8 Gene expression - transcription and translation

4.1 Reading the code

Splicing

Recall that genes have exons and introns. The first transcript from the DNA, the primary transcript, contains both intron and exon sequence. The introns are removed in a process called **splicing**, leaving only the protein coding exons in the messenger RNA (Figure 9). Nucleotides on either side of the intron-exon junctions form a **splice site**. A 'molecular machine' in the nucleus called the spliceosome recognises the splice sites, loops out the introns, cuts the RNA and re-joins the exons.

mRNA undergoes other changes to stabilise it for protein synthesis to occur. They include addition of methyl-G at one end (called the 5'-CAP) and a string of A nucleotides at the other end (called the 3' poly-A tail) (not shown in the diagrams below).

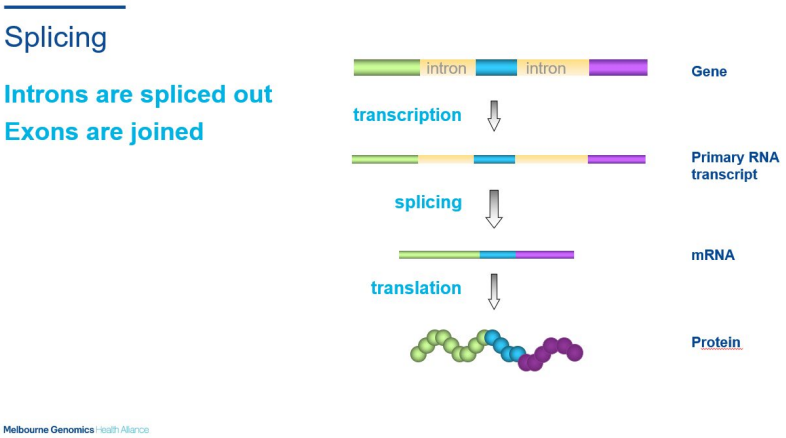


Figure 9 Splicing removes introns from the primary transcript and joins the exons.

Alternative splicing

The human genome has around 20-22,000 protein-coding genes, but many more proteins than this exist in the body. Thus, one gene can code for more than one protein. One mechanism for this is alternative splicing. Alternative splicing of the primary transcript can produce more than one type of transcript with different combinations of exons from one gene, and therefore different proteins (Figure 10). Alternative splicing can occur in a tissue specific manner. Genetic variants near the splice sites can alter normal splicing patterns and potentially contribute to genetic conditions.

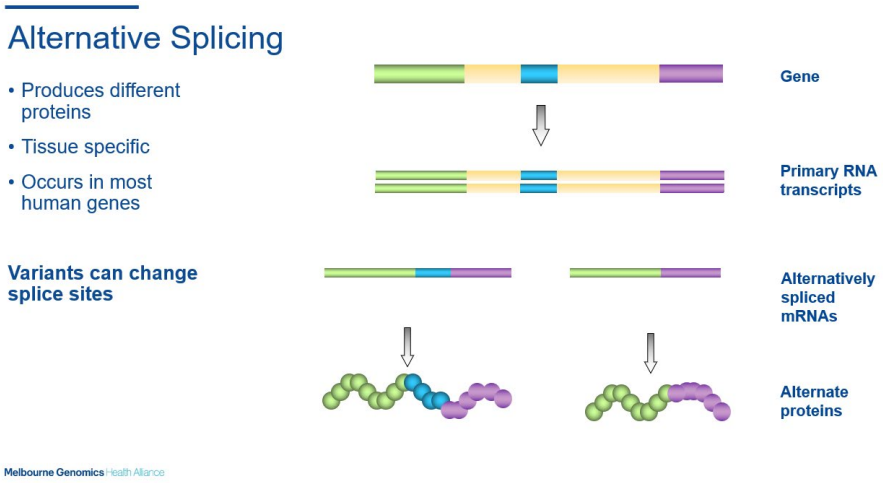


Figure 10 Alternative splicing produces alternate mRNAs and proteins.

4.2 Splicing

5 Proteins

Proteins are molecules that provide structure and function to cells. Some proteins are produced in, and important for, a wide range of cell types. Malfunction of these proteins can have wide-ranging effects in the body. Other proteins are produced in, or act on, a narrow range of cells. Malfunction of such proteins might display more limited effects on the body.

Types of proteins (examples)

Enzymes

Antibodies

Ion channels

Transcription factors

Extracellular matrix proteins

Nutrient transporters and carriers

Cell signalling molecules: peptide hormones, cytokines

Receptors for hormones, cytokines and neurotransmitters

Cytoskeleton proteins – e.g. microtubules and microfilaments

Structural elements of proteins

Proteins are translated as a string of amino acids, the polypeptide chain, which then folds into a functional protein. The amino acid sequence of a protein determines its folding pattern, and the final shape is essential to the function of many proteins[3].

The primary structure is the amino acid sequence of the protein. It forms secondary structures (alpha-helices and beta-pleated sheets) and then folds into a complex tertiary structure with defined structural and functional domains. Some proteins require more than one polypeptide to form the functional protein; this is called quaternary structure. Examples of proteins with quaternary structure are haemoglobin with two alpha-globin and two beta-globin polypeptides and the insulin receptor with two alpha-subunits and two beta subunits.

Protein Domains

Proteins fold to produce the functional and structural domains needed for their cellular location and function (Table 1). The amino acid sequence determines folding, therefore DNA changes that alter the amino acid sequence can alter structural and functional domains.

Table 1 Protein domains determine location or function of a protein

Tissue location determined by structural domains	Functional domains (examples)
• Cytoplasm, e.g. metabolic enzymes • Secreted, e.g. hormones • Plasma membrane, e.g. ion channels, hormone receptors, nutrient transporters • Mitochondria, e.g. cytochromes • Intracellular cytoskeleton • Extracellular matrix	• The substrate binding site of enzymes • The ion binding sites of ion channels • The hormone binding domain of receptors • The antigen binding domain of antibodies • DNA binding domain of transcription factors • Transmembrane domains of membrane-spanning proteins

5 Proteins

[3] Further reading about amino acids and proteins: video at the RCSBProteinDataBank https://www.youtube.com/watch?v=wvTv8TqWC48

6 Genomic Variants

6.1 Genomic Variants

Genomic variants are small or large changes in the DNA that can affect proteins and potentially alter our characteristics, or phenotype. In Figure 2 we introduced the range of genomic variation, from very large alterations of chromosome number and structure, to copy number variants involving tens of thousands of base pairs that can involve many genes, changes in whole gene number, or changes at the nucleotide level, such as triplet nucleotide repeats or single nucleotide changes.

Mutations occur at the DNA level when errors occur during DNA replication, or due to chemical or radiation damage and failure of the cellular DNA repair mechanisms. This produces variant forms of the gene (or new alleles). New variants are called ***de novo***.

Germline and Somatic Variants

Variants present in gametes and inherited from parents are called germline variants. New variants that arise during gamete formation, so called pre-zygotic *de novo* variants, will be heritable. *De novo* variants that arise early in development (post-zygotic *de novo* variants) might affect only some cell lineages. The type and extent of affected tissues depends of the stage in development when the new variant arises. Variants occurring in mature somatic cells (acquired variants) called can lead to cancer.

Structural variants

6.2 Structural Variants

Structural variants (SV) and structural rearrangements (SR) are large-scale changes in chromosomes.

Aneuploidy is variation in the number of an individual chromosome, such as an extra copy of chromosome 21, trisomy 21, in individuals with Down Syndrome, or an extra X chromosome in males with Klinefelter syndrome. Aneuploidy can be the loss of a chromosome, such as loss of an X chromosome in females with Turner syndrome, and loss of individual chromosomes in cancer cells.

Polyploidy is when an individual has a whole extra set of chromosomes. Polyploid humans don't survive to birth, however polyploidy can occur in some cells and tissues, and is a common feature of some cancers.

Aneuploidy and polyploidy are detected by karyotype analysis, fluorescence *in situ* hybridisation (FISH) and molecular karyotyping (chromosomal microarrays).

Structural variants at individual chromosome level (Figure 11) include:

- Translocations – the reciprocal exchange of DNA between chromosomes
- Inversions of chromosome segments
- Deletions
- Insertions
- Expansions

Balanced structural variants involve rearrangement of genetic information, such as translocations and inversions, without gain or loss of DNA.

Unbalanced SVs involve a gain or loss of genetic information, such as deletions, duplications and triplet repeat expansions.

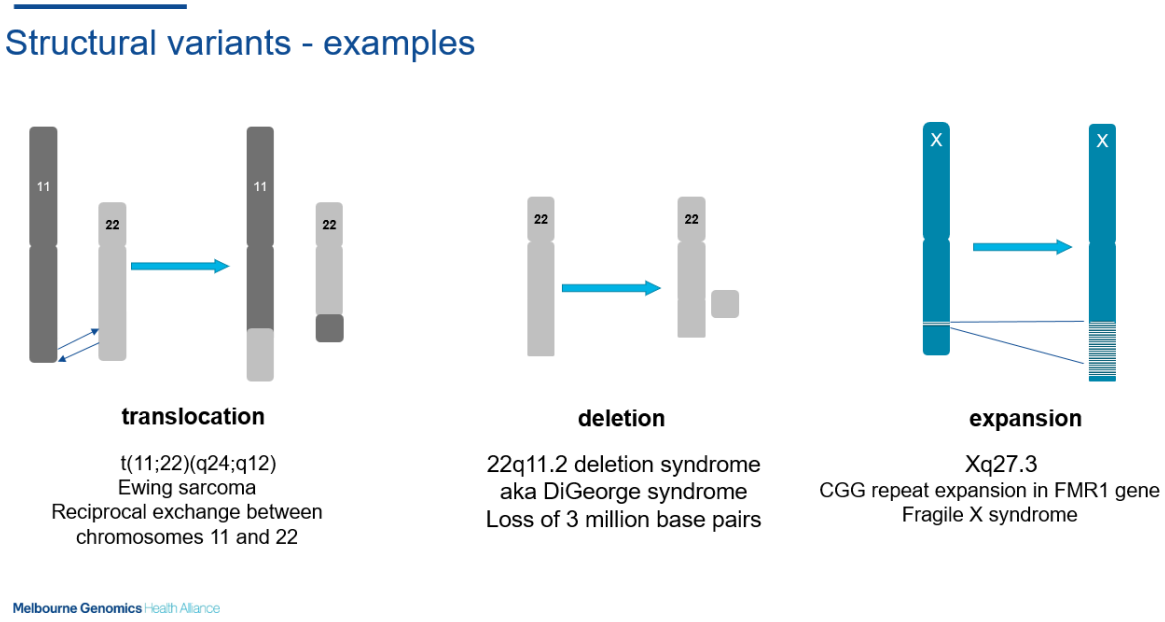


Figure 11 Structural variants (examples)

Structural variants and chromosome rearrangements play a significant role in cancer. Consequences of chromosome events in cancer include:

- Amplification - overexpression of oncogenes (e.g. ERBB2, myc)
- Deletion - Loss of tumour suppressor functions (e.g. mutated BRCA1)
- Translocation - deregulated gene expression; new functional protein product, e.g. transcription factors or kinase activation (e.g. BCR-ABL)
- Inversion - gene activation or inactivation (e.g. ALK)

Gene fusions

6.3 Gene fusion

Gene fusions occur during chromosomal rearrangements such as deletions and translocations. Breakage and re-joining of DNA brings different genetic elements together, leading to chimeric transcripts and proteins (Figure 12). They are common type of structural variant in cancer.

Consequences of gene fusions include:

- loss of protein function (e.g. loss of tumour suppressor activity)
- chimeric protein with oncogenic action (e.g. permanent activation of kinases, driving cell growth)
- deregulated gene expression (e.g. fusing a strong promoter to a proto-oncogene gene; activation of two genes in the Ewing sarcoma translocation t(11;22)(q24;q12)) (see Figure 12).

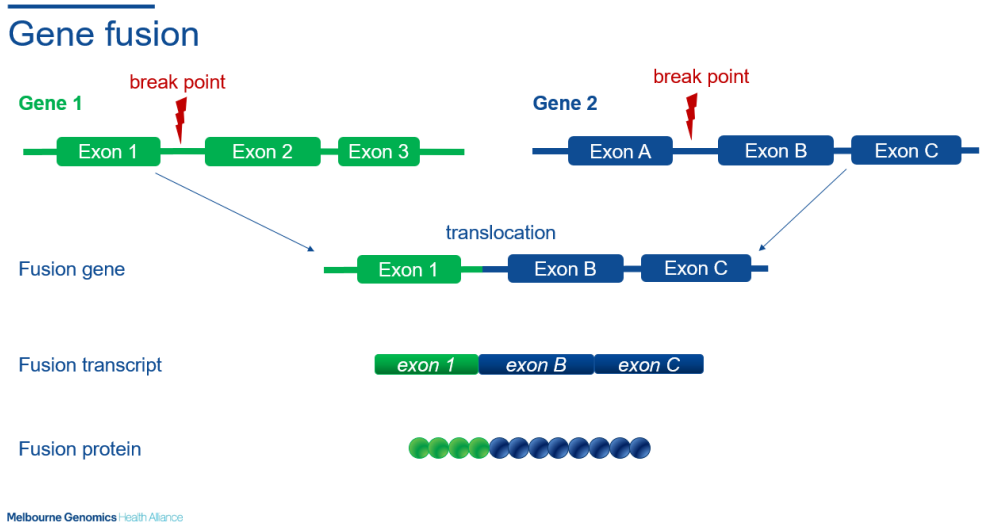


Figure 12 Gene fusion can occur during chromosomal rearrangements such as translocation

Copy number variants (CNV)

6.4 CNV

Copy number variants are the loss or gain of a region of DNA which can involve one or several genes, typically range from 1000 to 3,000,000 bp and may be too small to detect with karyotype analysis (Figure 13). They are detected with molecular karyotyping - chromosomal microarrays.

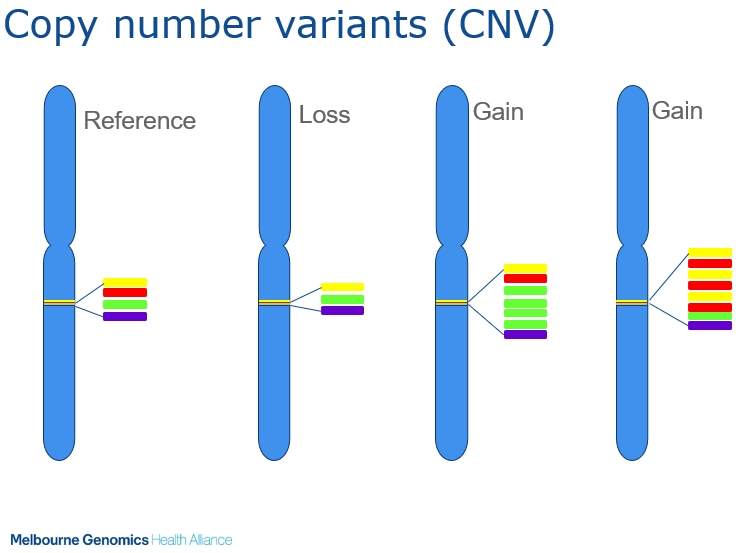


Figure 13 Copy number variants (CNV) are gains or losses of genetic information. Small CNVs are detected with molecular karyotyping.

Triplet repeats

A **triplet repeat expansion** is when three nucleotides repeats many times in tandem, causing expansion of the region.

An example is the CGG repeat in the *FMR1* gene on the X chromosome, with > 200 repeats causing Fragile X syndrome. Another example is the CAG repeat in the *HTT* gene for huntingtin protein (Figure 14). The normal huntingtin gene has up to about 26 CAG repeats, while more than around 40 repeats causes Huntington disease. Triplet repeats are detected by PCR methods and Southern blotting. They are not well detected by genomic sequencing or chromosomal microarrays.

Figure 14 (a) Trinucleotide repeat illustration; (b) the CAG trinucleotide repeat in the HTT gene - normal alleles carry up to 30 repeats, between 36 and 39 repeats may or may not develop Huntington disease; >39 repeats causes Huntington disease.

Single nucleotide variants (SNPs and SNVs)

6.5 SNV_SNP

A **single nucleotide variant** is a change in a single base pair. They include base substitutions, insertions or deletions (called indels) and duplications (Figure 15). Single nucleotide variants occurring at low frequency in a population (**SNV**) are the variants of interest in the hunt for genetic changes that cause clinically relevant conditions. Common variants, those occurring at a frequency of more than 1% in a population, are called **single nucleotide polymorphism**, or **SNP** ('SNiP'), and are usually not a direct cause of disease.

SNVs and SNPs are detected by sequencing of single genes or of whole exomes or genomes, or by PCR methods in the case of triplet repeats. SNPs are also detected by SNP array technology.

Figure 15 Single nucleotide variants

7 Consequence of variants

Effect of variants on proteins

Earlier we saw that during translation the nucleic acid code directs which amino acids come together to build the protein. Genetic variants may or may not change the amino acid sequence. Understanding codons and the genetic code can help understand the potential impact of genetic change.

Codons, the groups of 3 bases in the mRNA, specify the sequence of amino acids in a protein. The reading frame begins with a 'start' codon (AUG) which specifies the amino acid methionine (met) and ends with a 'stop' codon (UAA, UAG or UGA). A gene can have more than one start codon, and therefore more than one reading frame(Figures 16 and 17).

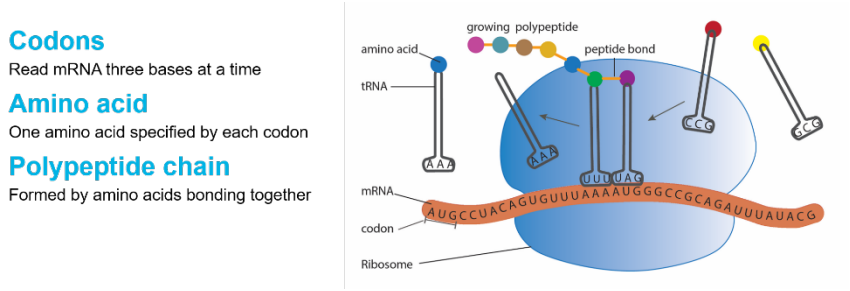


Figure 16 mRNA codons specific the amino acid placed in the protein. Changes in DNA are copied into the mRNA and may alter the amino acid sequence.

The genetic code

	AGA									UUA
	AGG									UUG
GCA	CGA						GGA			CUA
GCC	CGC						GGC		AUA	CUC
GCG	CGG	GAC	AAC	UGC	GAA	CAA	GGG	CAC	AUC	CUG
GCU	CGU	GAU	AAU	UGU	GAG	CAG	GGU	CAU	AUU	CUU
Ala	Arg	Asp	Asn	Cys	Glu	Gln	Gly	His	Ile	Leu
A	R	D	N	C	E	Q	G	H	I	L

				AGC					
				AGU					
			CCA	UCA	ACA			GUA	
			CCC	UCC	ACC			GUC	UAA
AAA		UUC	CCG	UCG	ACG		UAC	GUG	UAG
AAG	AUG	UUU	CCU	UCU	ACU	UGG	UAU	GUU	UGA
Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	stop
K	M	F	P	S	T	W	Y	V	

Figure 17 The genetic code table shows mRNA codons (groups of 3 bases) with the corresponding 3 letter and single letter amino acid abbreviations/symbols.

DNA variants can alter the code. Loss or gain of 1 or more bases (other than in multiples of 3) changes the reading frame. Substitution of one base can change the amino acid. The following mRNA sequence (Figure 18 and paragraph illustrates consequences of variants on the protein. Let's consider the consequences of changing the 6th base in the mRNA sequence. Refer to the Genetic Code (Figure 17) to find the amino acid change.

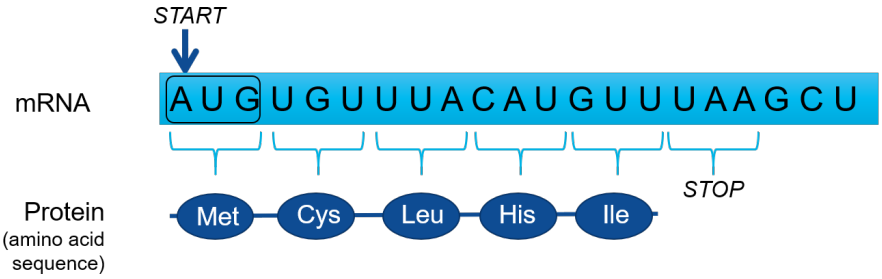


Figure 18 Codons in mRNA specify the amino acid in the protein sequence. AUG is a start codon for translation. UAA, UAG and UGA are stop codons.

The second codon, UGU, encodes cysteine, which forms a disulphide bond with another cysteine. Disulphide bridges are important in protein folding.

- If the 6th base changes from U to C (UGU to UGC), the amino acid will still be cysteine (UGC - Cys)
- If the 6th base changes from U to G (UGU to UGG), the amino acid will be tryptophan (UGG - Trp), a significant change as Trp does not form disulphide bonds.
- If the 6th base changes from U to A (UGU to UGA), a stop codon is formed (UGA – stop). This produces a truncated protein. The impact of a premature stop codons depends on how much of the protein is missing.
- Deleting the 6th base shifts the reading frame. The new base sequence is AUG UGU UAC AUG UUU AAG CU_, encoding **met-cys-tyr-met-phe-lys-…** All the amino acids after the base change are altered.

Key terms describing the effects of genetic variants on protein

The main variant types are briefly described below. These terms are used in the variant descriptions of genomic test reports. Follow the hyperlinks to view animations for each variant. They are also illustrated and summarised in Figure 19.

Synonymous or **silent** variants are when a nucleotide change does not change the amino acid sequence but can affect regulation or splicing.	*7.1 'Synonymous variants' - animation*
Non-synonymous or **Missense** variants are when a change in the DNA causes an amino acid change	*7.2 'Non-synonymous missense variants' - animation*
Frameshift variants are nucleotide substitution, deletion or insertion (other than in multiples of 3) that alter the reading frame of the mRNA, thus altering the amino acid sequence of the protein after the point of change. Depending on where the frameshift occurs the impact can be loss or reduction of protein function, production of an early stop codon, gain of function or dominant negative activity.	*7.4 'Frameshift variants' - animation*
Nonsense (stop-gain) variants are a base substitution, deletion or insertion that produces an early stop codon, also called a premature termination codon (PTC). Depending on location, the impact on the protein can be loss of function, gain of function, dominant negative activity (where the protein product adversely affects a normal version of the protein) or loss of protein synthesis due to activation of non-sense mediated decay of mRNA *	*7.3 Nonsense variants - animation*

***Nonsense-mediated decay (NMD)**
Nonsense variants produce a premature stop codon (also called a premature termination codon or PTC). The shortened or truncated protein could be deleterious to the cell, such as developing a dominant negative action or deleterious gain of function. A premature stop codon that occurs before the last codon is likely to trigger NMD to breakdown all the mRNA molecules of that type. This mechanism prevents the cell producing potentially deleterious proteins.

Variants: effects on protein

Variant	Change		Possible effect on protein
'Normal' protein			
Silent, Synonymous	No change		None
Missense, Non-synonymous	Amino acid change		Minor to major change depending on location and type of amino acid change
Nonsense, Stop-gain	Premature stop codon		Truncated protein; loss of protein; loss or gain of function; dominant negative activity; activation of non-sense mediated decay of mRNA
Frameshift	Altered amino acid sequence		Loss or reduction of protein function, production of an early stop codon, gain of function or dominant negative activity.

Melbourne Genomics Health Alliance

Figure 19 Summary of types of variants and the effect on protein. The single letter amino acid code is used in the illustrations. Note – while a synonymous variant does not alter the amino acid sequence, if the nucleotide change occurs near a splice site, splicing intron splicing may be altered.

8 Genomic testing

The clinical care of patients with a suspected genetic condition involves some type of DNA test. Different genetic and genomic test methods are used for different types of variants. All tests have their strengths and limitations. Common testing methods are summarised in Figure 20 and Table 2. This section focusses on the **sequencing technologies** to identify nucleotide variants such as base substitutions and small insertions and deletions (indels).

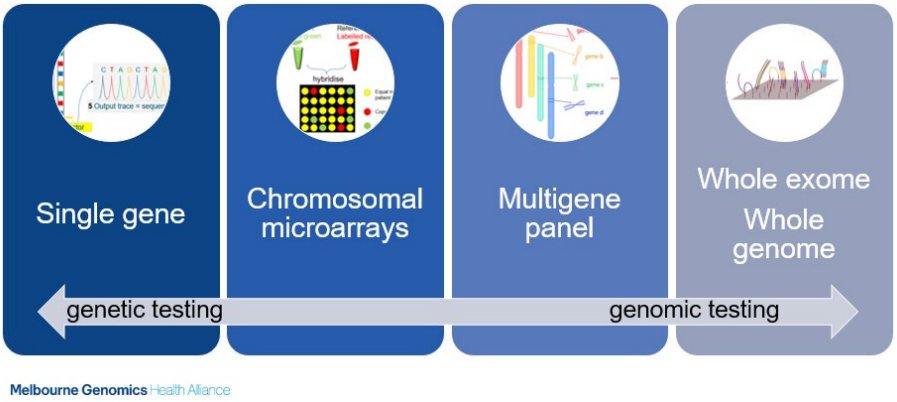


Figure 20 Spectrum of genetic to genomic testing.

Table 2 Summary of Genetic and Genomic tests

	Genetic/genomic test	What is analysed	What is detected
GENETIC TESTING	Cytogenetics, Karyotype	Whole chromosomes, G-banding method	Aneuploidy, polyploidy, some structural rearrangements
	Chromosomal microarray; SNP/CGH array	Chromosomal DNA	CNVs, some structural rearrangements
	PCR methods, MLPA,	Targeted DNA regions	Triplet repeat expansions, CNVs - deletions, duplications
	Sanger sequencing	Single gene sequence	Nucleotide sequence of one gene known to cause the condition
GENOMIC TESTING	Mitochondrial genome sequencing	Mitochondrial DNA	single nucleotide variants, deletions, duplications
	NGS - Multigene panel	Nucleotide sequence of a targeted group of genes; designed to target genes associated with a condition	Small nucleotide variants, gene fusions
	NGS - Whole Exome	Sequence of all protein-coding regions of genes (exons only)	Single nucleotide variants, small deletions, insertions, duplications
	NGS - Whole Genome	Nucleotide sequence of all genes (introns and exons), regulatory regions, mitochondrial DNA, non-coding DNA, non-protein coding genes	Single nucleotide variants, small deletions, duplications

8 'Genetic and genomic tests'- illustrating a range of genetic and genomic tests (video-non-narrated)

Using single gene test

For a well characterised **monogenic** condition caused by only one gene, such as cystic fibrosis (*CFTR* gene) or beta-thalassaemia (*HBB* gene), a single gene test is conducted to confirm the clinical diagnosis. Sanger Sequencing (see 'Sequencing Technology' / Figure 22) is used for a single gene sequence. PCR-based methods are also used to identify common variants.

review Monogenic, polygenic, multifactorial conditions

Using multigene panel test

For a monogenic condition known to be caused by one of several genes, such as dilated cardiomyopathy (>30 genes) or non-syndromic hearing loss (>90 genes), a targeted multigene panel test can be performed (Figure 21). Next Generation Sequencing (NGS) technology (see 'Sequencing Technology' / Figure 23) is used to sequence the group of genes to identify the relevant variant in one of those genes.

One limitation of the multigene panel test is that sequencing is confined to a small group of genes. If a pathogenic variant is not found, a new sample and further sequencing of more genes is often required, as there is no other sequence data available for immediate reanalysis.

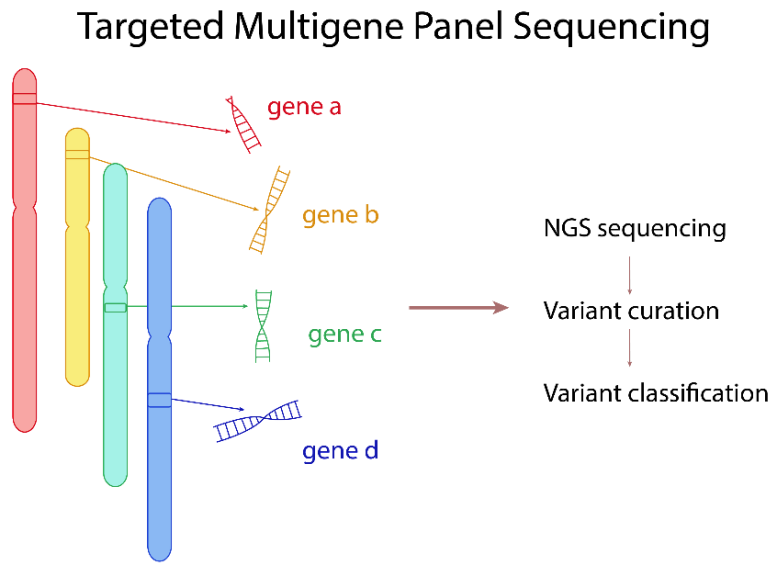


Figure 21 NGS can target a defined set of genes known to be associated with a genetic condition, to identify which one of these genes is responsible in an individual patient

Genomic testing in cancer mainly uses targeted multigene panels, such as a Comprehensive Cancer Panel, or panels designed for different types of cancer. Cancer panels may target particular exons with known cancer-related variants or mutation 'hotspots. Testing for blood cancers employs panels targeting inherited variants and acquired variants, as blood cancers can arise from one or both types of variant.

Mitochondrial genome sequencing

Next generation sequencing platforms are used for sequencing the 16,569 bp of the mitochondrial genome. Human mtDNA carries 37 genes (13 protein-coding genes and 24 non-protein-coding genes), all essential for mitochondrial function. For potential mitochondrial conditions, mtDNA sequencing can be done in parallel with whole exome sequencing to capture nuclear genes involved mitochondrial function.

Using whole exome (WES) or whole genome (WGS) sequencing

If the condition is likely to be monogenic but could be caused by one of many genes, for example conditions with a complex, non-specific phenotype and/or features such as intellectual disability, developmental delay or syndromic features, and infectious or environmental agents are ruled out, then Next Generation Sequencing of the whole exome or genome is recommended, if available.

Sequencing the whole genome produces vastly more data than sequencing the exome (exome is ~2% of genome), and therefore requires more time and computational capacity for data analysis and storage. As most currently known genetic conditions are caused by changes in the protein coding regions of the genome, sequencing the exome may be the fastest and most cost-effective option. However, each method has technical limitations, advantages and disadvantages, which may vary over time as methodologies develop and costs change.

Features of whole exome and genome sequencing are listed below.

WES can detect:

- Single nucleotide changes in coding regions (frameshift, truncating, missense)
- Small deletions, duplications, insertions in coding regions
- Some splice site variants (WES captures a small amount of sequence into the introns at the exon-intron junction)

WES does not (easily) detect:

- Structural variants
- Copy number variants
- Deletions/duplications >15-30 bp
- Gross chromosome changes
- Triplet repeat expansions
- Changes in non-coding regions (including regulatory elements, introns)
- Mitochondrial genes

WGS detects:

- Single nucleotide changes in coding and non-coding regions (frameshift, truncating, missense)
- Small deletions, duplications, insertions in coding and non-coding regions
- Nuclear and mitochondrial DNA
- Splice site variants

Sequencing technology

Sanger sequencing

Sanger sequencing was developed in the 1970s to determine the sequence of a single gene. Despite advances in sequencing technology with Next Generation Sequencing (NGS) (see below) Sanger sequencing remains in use for single gene sequencing and confirmatory testing of variants identified in whole exome or genome sequencing.
The method incorporates fluorescent-labelled nucleotides to enable detection of the order of bases in the sequence (Figure 22).

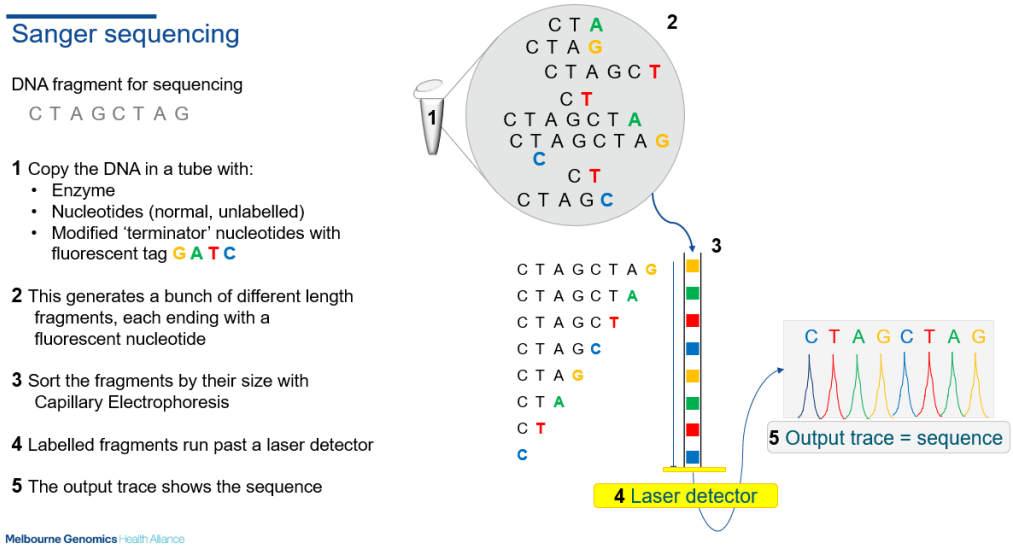


Figure 22 Illustration simplifying the Sanger Sequencing method for sequencing single genes.

Next generation sequencing (NGS)

Next generation sequencing (NGS), also called massively parallel sequencing (MPS) enables the sequencing of many genes at once (Figure 23). This can mean sequencing the whole genome, whole exome or targeted gene panels.

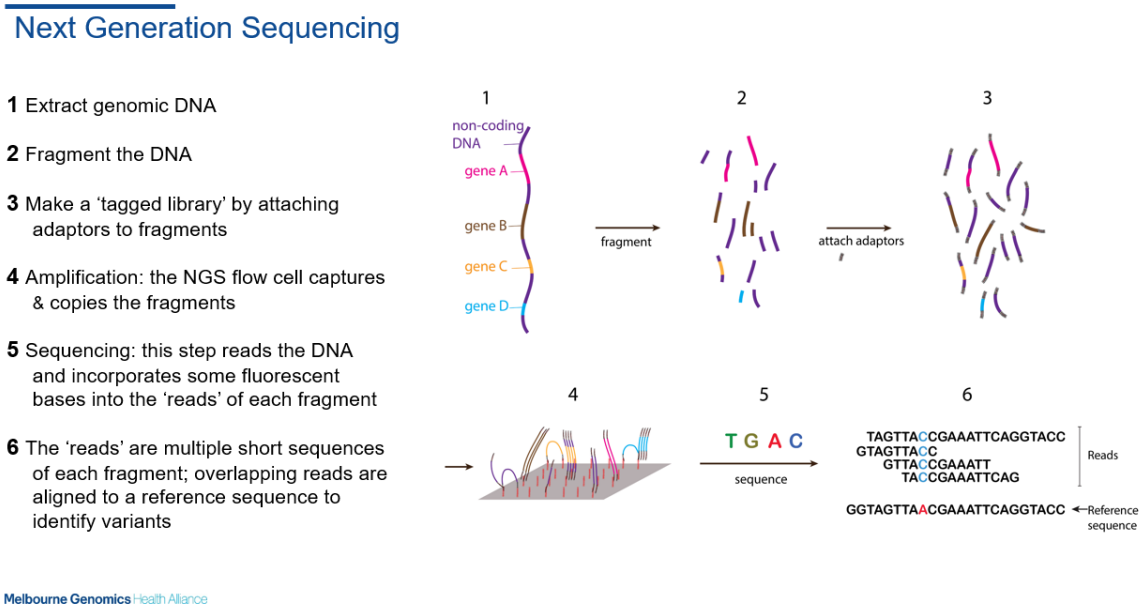


Figure 23 Illustration of key steps in a Next Generation Sequencing method used in sequencing for multigene panels, mitochondrial genome, whole exome and whole genome sequencing.

Technical / sequencing limitations affecting variant analysis

Coverage

Sequence coverage in genomic tests refers to the amount of the genome accurately captured in the sequencing data, that is, how much of the sequence aligns to the reference sequence.

Coverage includes:

- Breadth - percentage of bases sequenced
- Depth - number of reads; e.g. target read depth = 100x for WES

Coverage varies in different sequencing tests and can be a factor in the choice of test. Sequencing fewer genes, as in a multigene panel test gives better coverage than larger scale genomic tests. There are also variations on NGS methods; for example, NGS for WES produces short 'reads' of the DNA (about 150bp) and good coverage relies on having many overlapping reads to cover the full sequence and to give confidence in the identity of each base (Figure 24). Gaps in coverage are difficult to interpret, e.g. is it poor sequencing (missing data) or is it a deletion variant?

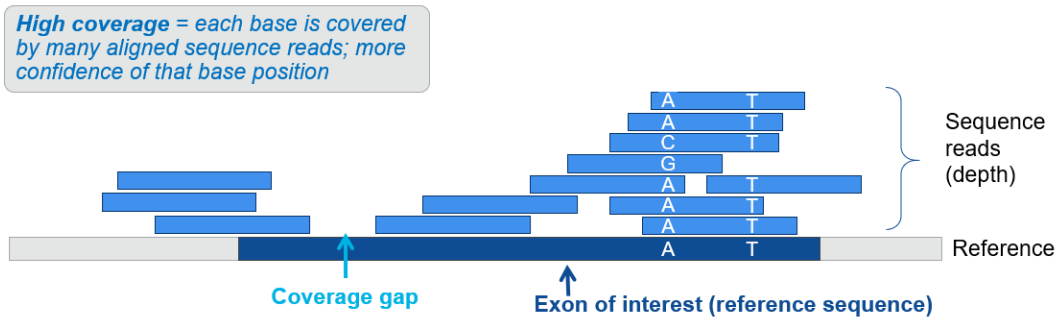


Figure 24 Coverage means the breadth and depth of the target DNA that aligns to the reference sequence. Good coverage needs many overlapping reads for each nucleotide.

Coverage in different tests

- Sanger sequencing of single gene – the best coverage for most genes
- NGS for multigene panel – good coverage with fewer genes sequenced
- NGS for WES – uses 'short read' technology, i.e. reads of ~150bp. Sequence determination in NGS relies on computational reconstruction of overlapping reads. WES typically has more gaps and lower coverage than multigene panels; some regions are harder to piece together; overall the sensitivity for detecting a variant could be lower
- NGS for WGS – as WGS does not target coding regions, it is better at detecting structural variants and CNVs, and has improved overall coverage; gap problems remain as for WES; some methods use 'long read' technology which leaves fewer gaps and may provide greater sensitivity for detecting variants

Coverage limitations, and other technical limitations, might be noted on a test report you receive from the lab. Laboratory reports may note regions of expected low coverage, or overall low coverage requiring a new sample or re-sequencing of the original sample (Figure 25).

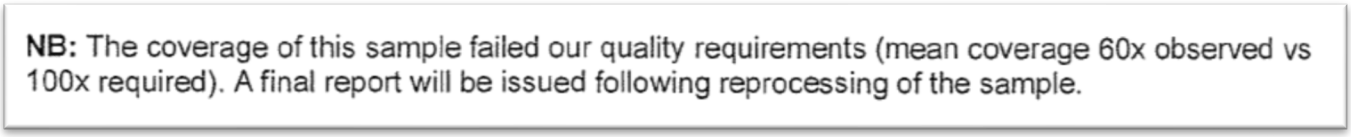


Figure 25 Low coverage, failed quality control for exome sequence, requiring resequencing of the sample

Low variant allele frequency

Coverage and variant detection are limited with a low variant allele frequency (VAF). For example, a tumour tissue sample might have a very low proportion of cancer cells and, therefore, a very low frequency of variants. Combined with coverage gaps, the variant(s) can be harder to detect. Cancer samples therefore need more sequencing reads for reliable detection of the variant(s).

Sequencing limitations affect certain regions of the genome

Some genes are hard to sequence due to the nature of the DNA sequence, thus limiting variant identification. Specialised tests may be required. Examples include:

- Regions with a high GC content
    - e.g. exon 1 of a gene is frequently GC-rich; limited detection by WES and WGS; panel sequencing test or Sanger sequencing is best.
- Repetitive regions
    - e.g. *HTT* gene (in Huntington Disease); triplet repeats in an exon are poorly detected by WES and WGS; PCR testing is best.
- Presence of pseudogenes
    - e.g. Congenital adrenal hyperplasia (CAH) due to 21-hydroxylase deficiency is caused by mutations in the CYP21A2 gene which has a pseudogene next to it; panel sequencing or Sanger sequencing is best.

9 Variant Identification and Interpretation

Let's say a whole exome sequence (WES) test is ordered. In this test, exons of the protein coding regions of all 20-22,000 genes are sequenced by NGS technology. To identify nucleotide variant(s) causing the condition you don't usually need to analyse all the genes, as most will be irrelevant. , Rather, you can analyse genes that are associated with the phenotype and have previously been identified as relevant to the condition.

Variant interpretation is the process of interpreting the effect of a genomic variant, typically in a diagnostic context, to determine whether the variant causes the patient's condition. The process requires the expertise of bioinformaticians, medical scientists, clinical geneticists and genetic counsellors. Analysis is guided by the detailed clinical and phenotypic information provided by the referring physician. For our current discussion of variant interpretation, we include three stages:

- variant identification and selection
- variant curation
- variant classification

Identify variants

The first stages involve **bioinformatics** analysis of the DNA sequence. The exome sequence is compared to a reference sequence to identify (call) variants. The reference sequence used for identifying germline variants is a sequence compiled from many human genome sequences and sourced from an international database. In addition, in cancer, the reference sequence for analysing variants in solid tumours is the patient's own germline genome sequence from non-tumour cells. As described in Section 8, good sequence coverage is required for reliable identification of a sequence variant.

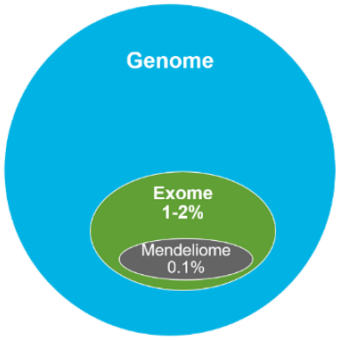
Filters and gene lists

All genes have variants that contribute to the normal range of human variation. Clinical testing is interested only in the variants relevant to the patient's clinical phenotype, so the data is filtered to selectively 'ignore' polymorphisms (common SNPs) and variants in untranslated, intronic and other non-coding regions. Filtering also allows you to focus on variants inherited in a recessive, dominant, autosomal or sex-linked pattern.

The '**incidentalome'** (incidental findings unrelated to the condition under investigation; see text box, right) can also be applied as a filter. However, specific incidentalome genes can be included in the analysis if they are relevant to the patient's phenotype.

The 'incidentalome' refers to pathogenic variants that may be identified in a genome sequence as an 'incidental' finding, that is, unrelated to the condition under investigation, or variants causing late onset conditions with no effective treatment.

Analysts then apply one or more **gene lists** (sets of genes associated with the patient's phenotype) to identify sequence variants in the genes most relevant to the patient. If the cause of the condition is not found using the selected gene lists,, exome sequence data can be re-analysed with other gene lists, or with the **Mendeliome**, essentially a very big gene list with around 4000 genes each known to cause monogenic (Mendelian) conditions (Figure 26).



Genome *3x10^9 base pairs, includes all coding and non-coding DNA*

Exome *Exons of the 20,0000 genes*

Mendeliome *Around 4000 genes known to cause monogenic conditions, i.e. conditions inherited in a Mendelian fashion (autosomal, sex-linked, recessive, dominant, codominant)*

Figure 26 Genomic sequencing is typically used for analysis of the whole genome, the whole exome or the genes associated with monogenic conditions and Mendelian inheritance pattern (Mendeliome)

Gene lists - sets of genes with known association to a phenotype or disorder

They vary in number of genes; they are periodically updated with new information. Examples…

- *Short QT syndrome – 3 genes*
- *Cardiomyopathy – 55 genes*
- *Dementia - 64 genes*
- *Immunological disorders – 298 genes*
- *Intellectual Disability syndromic and non-syndromic – 1064 genes*

(examples from the VCGS 2018 clinical exome gene list catalogue)

These steps of comparing to a reference sequence, filtering and applying gene lists typically identifies several variants of potential relevance to the patient's phenotype. They are now curated.

Curate variants

Variant curation is a process of gathering evidence about a variant to determine whether it is or is not the cause of a condition. A team of experts (clinical geneticist, bioinformation, genetic counsellor, medical scientist) **prioritise** the variants based on rarity, potential to affect the protein and relevance to the patient's phenotype. Variants are then **curated**; this involves gathering and assessing evidence about the known or predicted effects, and phenotype associations of the variant.

Evidence and databases for curation

The search for evidence uses a range of databases to address the following questions:

- Does the variant disrupt the gene?
- How does the gene disruption affect the protein?
- Is the variant common or rare; is it in population databases; if so, at high or low frequency?
- Has the variant previously been associated with disease?
- Does the altered protein relate to the phenotype?
- Has the variant previously been classified as pathogenic?

The strength and reliability of the evidence, in relation to the phenotype or condition being analysed, is weighed up to arrive at a classification for the variant (Figure 27).

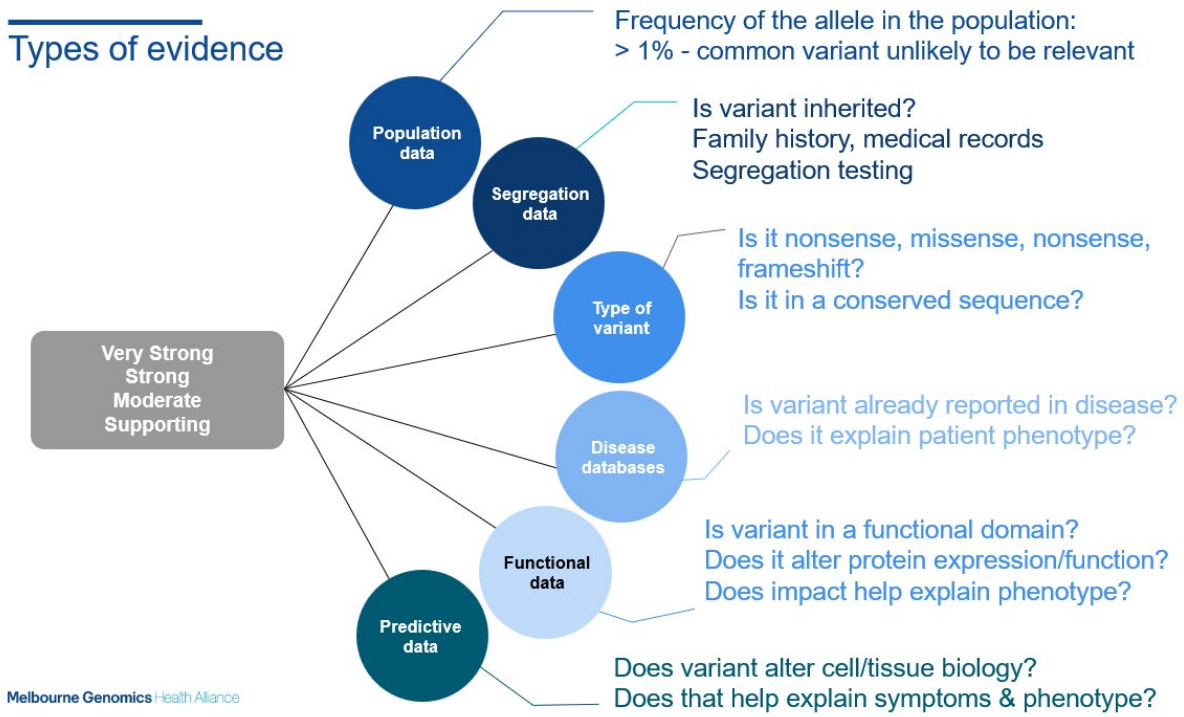


Figure 27 Schematic of types of evidence collected for variant classification and interpretation. Evidence is sourced from multiple databases (blue circles) and weighted from 'very strong' to 'supporting' (grey box) to assess an overall classification as benign or pathogenic.

For a broad appreciation of the depth of evidence required for variant interpretation by medical scientists and clinical geneticists, some key aspects are summarised in Table 3 (also see Appendix 2 for database links). These aspects are included in genomic test reports and the relevant databases are referenced. General physicians usually don't need to use the databases, but you might find occasion to seek more information in ClinVar, OMIM or other sites.

Table 3 Evidence and databases for germline variant curation - summary

Information required	Key aspects and terminology	Databases
Reference sequences	Identify variants compared to a reference	• Human Genome Reference Consortium (HGRC) • NCBI or UCSC RefSeq • Cancer Genome Atlas
Inheritance pattern	Homozygous, heterozygous, hemizygous, compound heterozygous, Sex-linked, X-linked, De novo, Dominant, Recessive, Penetrance, Expressivity. Zygosity can determine pathogenicity, e.g. a variant might be pathogenic only in the homozygous state, so an individual with one copy of such a variant is a heterozygous carrier	• OMIM - Online Mendelian Inheritance in Man • Clinvar
Type of variant	DNA level - Substitution, Insertion or deletion (indel), deletion & insertion together (delins) Protein level - Synonymous, Missense, Nonsense, Frameshift	• HGRC • NCBI RefSeq • TCGA
Disease databases	Variants present in a disease database due to previous association with a condition, are significant ACMG guidelines: a variant classified as pathogenic in one condition must be subsequently classified as pathogenic in the same condition in another individual	• Clinvar • HGDP - Human Genome Diversity Project • HGMD - Human Genome Mutation Database • OMIM - Online Mendelian Inheritance in Man
Population databases	Variants at high frequency in population databases (e.g >1%) are considered common polymorphisms (normal variation) If a variant is absent from a population database, it is more likely to be a new variant, less likely to be common benign variant	• gnomAD • dbSNP • 1000G
Conservation	Conserved sequences are the same across species, which implies that the region or specific sequence of the gene/ protein has an important biological function; e.g. a variant in that position may alter protein function Variants in regions of high conservation may have a greater effect on the protein and more potential for pathogenicity. The impact also depends on location, e.g. functional domain, which are often conserved.	• 100 vertebrates • UCSC
In silico predictions	Molecular biology, bioinformatics and computational tools to predict the effects of a variant on gene expression and protein function. For example: whether an amino acid change is minor or major is assessed by amino acid properties (size, charge, polarity) and location, such as occurring in a functional domain.	• SIFT • Polyphen • CADD • Mutation T@ster • DECIPHER • PDB • PROVEAN

Classify variants

The strength of the evidence collected is weighed up (very strong, strong, moderate, supporting) to arrive at an overall classification.

Classification follows the American College of Genetics and Genomics (ACMG) Guidelines[4] scale of pathogenic, likely pathogenic, uncertain significance, likely benign, benign. Some laboratories categorise them as Class 5 to Class 1, and/or may also subclassify the variants of uncertain signficance (VUS) as Class 3a, 3b and 3c according to whether the evidence tends towards pathogenic or benign (as practised by the Victorian Clinical Genetics Services (VCGS) (Figure 28).

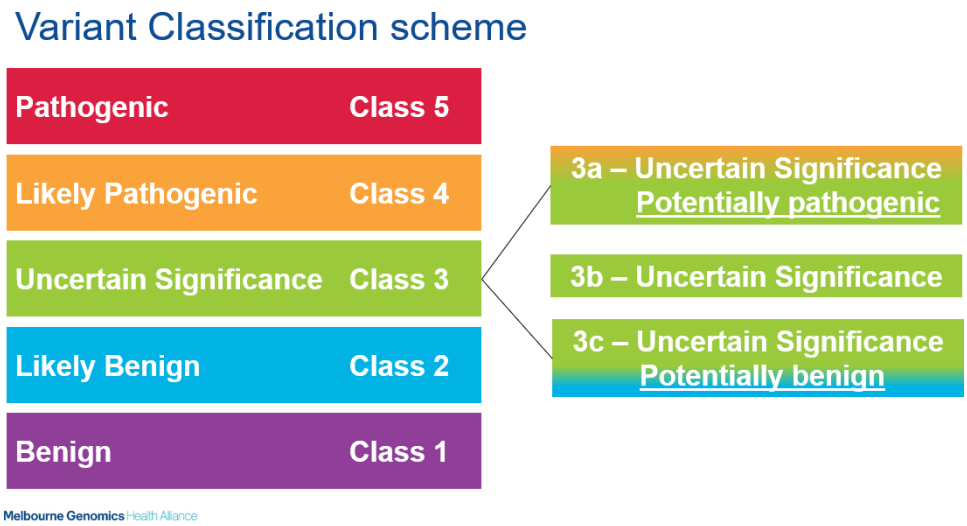


Figure 28 The variant classification scheme used by VCGS, based on ACMG Guidelines with additional sub-classification of variants of uncertain significance (VUS, Class 3)

Clinical actions from genomic test report

The recommended clinical actions, determined by variant classification, are summarised in Table 4.

Table 4 Clinical actions recommended clinical actions based on the reported variant interpretation

Type of Variant	Class	Implications for patients
Pathogenic	5	Cause for condition identified Can be used to direct management Can be used for family planning Can be used for predictive testing
Likely pathogenic	4	Cause for condition likely been identified May be used to direct management May be used for family planning May be used for predictive testing in other family members
Uncertain significance*	3	Cause for condition still unclear Cannot be used to direct management Cannot be used for family planning Cannot be used for predictive testing in other family members **Class 3a perform segregation studies
Likely benign	2	As for VUS - class 3 except no segregation studies
Benign	1	As for VUS - class 3 except no segregation studies

[4] Richards et. al., Interpretation of sequence variants, Genetics in Medicine 2015; https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4544753/

Classifying cancer variants

Curation of cancer variants follows a similar path to that of germline variant interpretation, with additional classification related to the **diagnostic, therapeutic,** and **prognostic** outcomes.

Differences in cancer variant curation and classification include:

- The use of cancer specific databases - OncoKB, cBioPortal, NIH national Cancer Institute GDC Data Portal, Jackson Lab JAX-CKB, CIViC and COSMIC.
- Classification guidelines are available from AMP[5] (Association for Molecular Pathology) for interpreting cancer variants, and evidence is grouped into four tiers, summarised in Table 5 (see AMP guidelines[5] for full classification details)

Table 5 Summary of AMP Classification tiers for cancer variants (see [5])

Tier 1	Tier 2	Tier 3	Tier 4
Variants of strong clinical significance	**Variants of potential clinical significance**	**Variants of unknown clinical significance**	**Benign or likely benign variants**
Therapeutic, prognostic, diagnostic value	Therapeutic, prognostic, diagnostic value	N/A	N/A
FDA -approved therapy or well powered studies published	FDA -approved therapy for some tumours; small studies published or preclinical trials	Low frequency of the variant in cancer databases; No clear evidence of cancer association	Variant at significant frequency in general population and databases; no published evidence of cancer association

[5] Li et.al., Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer, J Mol Diag 19 (1) 4-23 2017 https://jmd.amjpathol.org/article/S1525-1578(16)30223-9/fulltext

Genomic test reports

Reports from different laboratories vary in the amount of detail in variant description and classification. Some laboratories report all VUSs, without subclassification, requiring the clinician to make a judgement about their value, further investigation or actionability.

Elements of a genomic test report

- Test requested
- Reason for referral
- Result summary
- Interpretation of the findings
- Findings related to phenotype: chart listing gene(s), variant(s), zygosity, classification, inheritance pattern (if known)
- Variant description: detailed description with references
- Methods: lab & technical details; includes prioritised gene lists for exome analysis

Genomic test report examples

View the following extracts from a clinical exome test report for a young adult attending an immunology clinic with a non-specific phenotype and undiagnosed condition (note: variant coordinates have been changed for anonymity).

1

Test Requested: Clinical Exome Analysis
Clinical Details: Immune disorder

Results: **A heterozygous variant associated with this patient's condition was detected.**

Interpretation: This patient is heterozygous for a likely pathogenic variant in the *CTLA4* gene. Heterozygous pathogenic variants in the *CTLA4* gene are associated with autoimmune lymphoproliferative syndrome. This finding is consistent with this patient's phenotype.
First degree relatives of this patient are at 50% risk of inheriting this mutation; testing is available for this patient's family.
Genetic counselling for this family is being provided by our Genetics Counselling Service.

Findings related to phenotype:

Gene	Phenotype (Inheritance): OMIM	Genomic Location (hg19)	Variant	Zygosity	Classification
CTLA4	Autoimmune lymphoproliferative syndrome, type V (AD): 616100	chr2:200000000	c.000C>T	Heterozygous	Likely pathogenic Class 4

Legend: **AD" Autosomal Dominant, AR" Autosomal Recessive, XLD = X-Linked Dominant, XLR = X-Linked Recessive**
Variants with a frequency of >1% that have unknown clinical significance were Identified in a number of phenotype-specific genes but not reported (available on request).

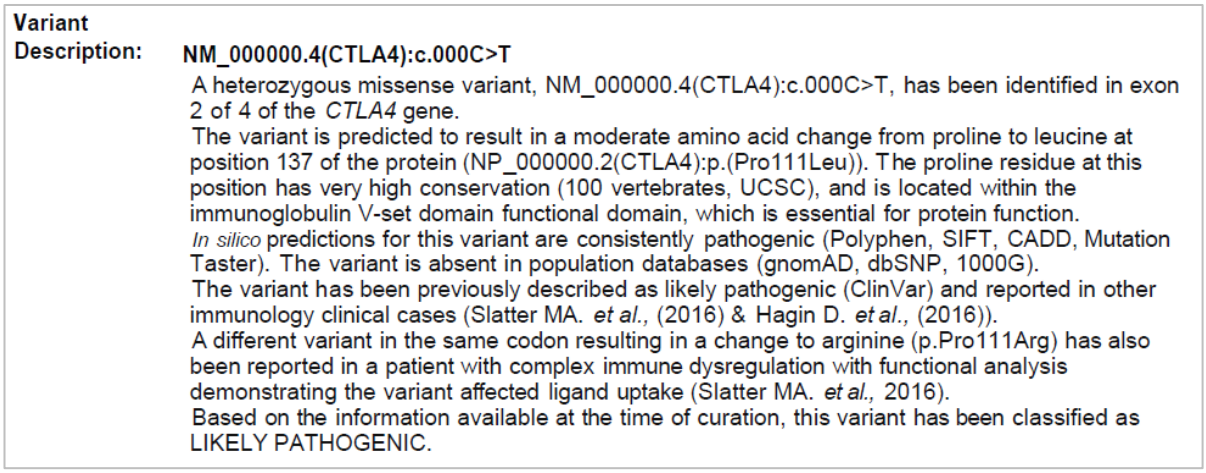
Figure 29 Exome test report – result summary and interpretation

The report starts with a summary of the results, including zygosity (heterozygous, the patient has one copy of the variant) and whether a variant associated with, or a cause for, the condition is found (Figure 29). The interpretation paragraph includes the name of the gene (*CTLA4*), the variant classification (likely pathogenic), previous links of this gene and/or the specific variant to the patient's condition (associated with autoimmune lymphoproliferative disorders), or other disorders. If parent have been tested, the inheritance (maternal or paternal) or presence of a new (*de novo*) variant in the patient is given. In this case, no parental testing has been done so inheritance is unknown.
Other information may be provided, such as recommendations for segregation or cascade testing for other family members, inheritance risk and genetic counselling.

Only the variants relevant to the phenotype are reported. The location and coordinates of the variant are given: the chromosome (Genomic location, chromosome 2; 'chr2:') and the coding sequence ('c.') and the type of variant (single nucleotide substitution, C replaced by T; c.000C>T). The coordinates

enable you or the clinical geneticist to do further research on this variant if necessary. See more about naming and coordinates in ‘Nomenclature’ (page 31).

2



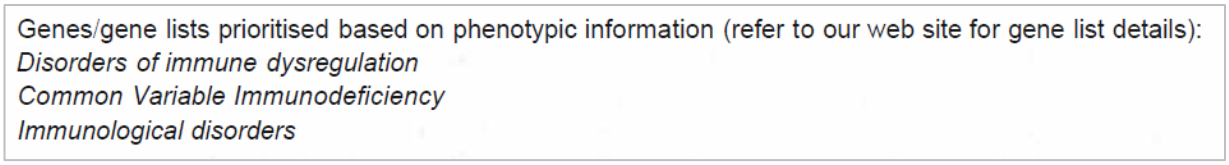
Variant Description: **NM_000000.4(CTLA4):c.000C>T**

A heterozygous missense variant, NM_000000.4(CTLA4):c.000C>T, has been identified in exon 2 of 4 of the *CTLA4* gene.
The variant is predicted to result in a moderate amino acid change from proline to leucine at position 137 of the protein (NP_000000.2(CTLA4):p.(Pro111Leu)). The proline residue at this position has very high conservation (100 vertebrates, UCSC), and is located within the immunoglobulin V-set domain functional domain, which is essential for protein function.
In silico predictions for this variant are consistently pathogenic (Polyphen, SIFT, CADD, Mutation Taster). The variant is absent in population databases (gnomAD, dbSNP, 1000G).
The variant has been previously described as likely pathogenic (ClinVar) and reported in other immunology clinical cases (Slatter MA. *et al.,* (2016) & Hagin D. *et al.,* (2016)).
A different variant in the same codon resulting in a change to arginine (p.Pro111Arg) has also been reported in a patient with complex immune dysregulation with functional analysis demonstrating the variant affected ligand uptake (Slatter MA. *et al.,* 2016).
Based on the information available at the time of curation, this variant has been classified as LIKELY PATHOGENIC.

Figure 30 Exome test report – variant description

The variant description (Figure 30) summarises the evidence. In this case, the substitution variant causes a change in one amino acid of the protein (missense variant). While this amino acid difference is only moderate, it is present in a highly conserved region in a functional domain, so likely to have a large effect on protein function, as predicted by the *in silico* tools. The absence of the variant from population databases tells you it is not a common polymorphism. The summation of the evidence, including previous clinical reports on this variant, concludes this variant is likely pathogenic, but the evidence is not strong enough to confirm it as the cause of the condition.

3



Genes/gene lists prioritised based on phenotypic information (refer to our web site for gene list details):
Disorders of immune dysregulation
Common Variable Immunodeficiency
Immunological disorders

Figure 31 Exome test report - gene lists

Genomic reports also provide technical information about limitations of the tests, coverage of the sequence, information not reported, and the gene lists (Figure 31) used in analysis. In this case, gene lists covering a wide range of immunological disorders; these gene lists have overlap in the genes covered but may have all been selected because some gene lists might have been more recently updated.

The amount and presentation of technical information in reports differs with the lab. Some labs report several VUSs, without sub-classification, and perhaps without recommendations for clinical actions. Contact your clinical geneticist for further guidance in these situations.

Nomenclature [6]

It is helpful to be familiar with the standard nomenclature for genes, variants and proteins used in genomic test reports.

Accession numbers:	• the accession number for the genomic sequence is recorded as 'NG___' • the accession number for the coding/mRNA sequence is recorded as 'NM___'	**Examples *CFTR* [7] gene** • e.g. NG_016465.4 • e.g. NM_000492.3
Naming the location and type of variant:	• the genomic DNA sequence is recorded as 'g.' • the coding DNA (or mRNA) sequence is recorded as 'c.' • The protein sequence is recorded as 'p.' • The mitochondrial DNA sequence is recorded as 'm.'	**Example: *CFTR* most common deletion** • g.98809_98811delCTT • c.1521_1523delCTT • p.Phe508del

Naming variants at DNA and protein levels

The example in the blue box (below) shows a short sequence of nucleotides in the DNA coding sequence (c., positions 1-12) and the corresponding amino acids in the protein (p., codon numbers 1-4). The table that follows gives examples of how a variant is reported (the nomenclature).

```
Coding nt number (c.)     1   4   7   10
Nucleotides               ATG AGC CCT GGT
Amino acids (p.)          met ser pro gly
Codon/aa number           1   2   3   4
```

The change	The nomenclature
A substitution of nucleotide 1 (A) for T	c.1A>T; codon 1 ATG>TTG
A deletion of AG at bases 4&5	c.4_5delAG
A duplication of C at nucleotide 8	c.8dupC
A substitution at nucleotide 10 (G) for A, resulting in amino acid number 4, glycine, being replaced by serine	c.10G>A; codon 4 GGT>AGT; p.gly4ser

[6] **References for nomenclature**
Other Accession numbers may be used, referring to predicted transcripts
Standardised naming system - HGNC guidelines: https://www.genenames.org/about/guidelines/
Molecular sequences are compared to a reference sequence: https://www.ncbi.nlm.nih.gov/refseq/
Example of CFTR nomenclature: Ogino et. al. J Mol Diagn. 2007 Feb; 9(1): 1–6 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1867422/
Variant nomenclature: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1867422/ ; http://varnomen.hgvs.org/recommendations/DNA/

[7] Clinvar https://www.ncbi.nlm.nih.gov/clinvar/RCV000007523/

10 Inheriting Germline Variants

Heritable germline variants associated with monogenic conditions usually have identifiable patterns of inheritance. These patterns provide information relating to the type of variant, help identify *de novo* variants and determine pathogenicity. However, some conditions display unpredictable patterns due to incomplete penetrance (not everyone with the variant shows the phenotype) and variable expressivity (different individuals with the same variant show different degree of phenotypic change).

Terminology for describing inheritance patterns include the following terms (also see the Glossary[8]). To review inheritance patterns and terminology, follow the hyperlink provided below.

- Homozygous, heterozygous, hemizygous
- Compound heterozygous
- Sex-linked, X-linked
- *De novo*
- Dominant, Recessive
- Penetrance
- Expressivity

Collecting and interpreting family genetic history, including accurate pedigree charts, aids clinical genetics and genomics analysis (Figure 32).

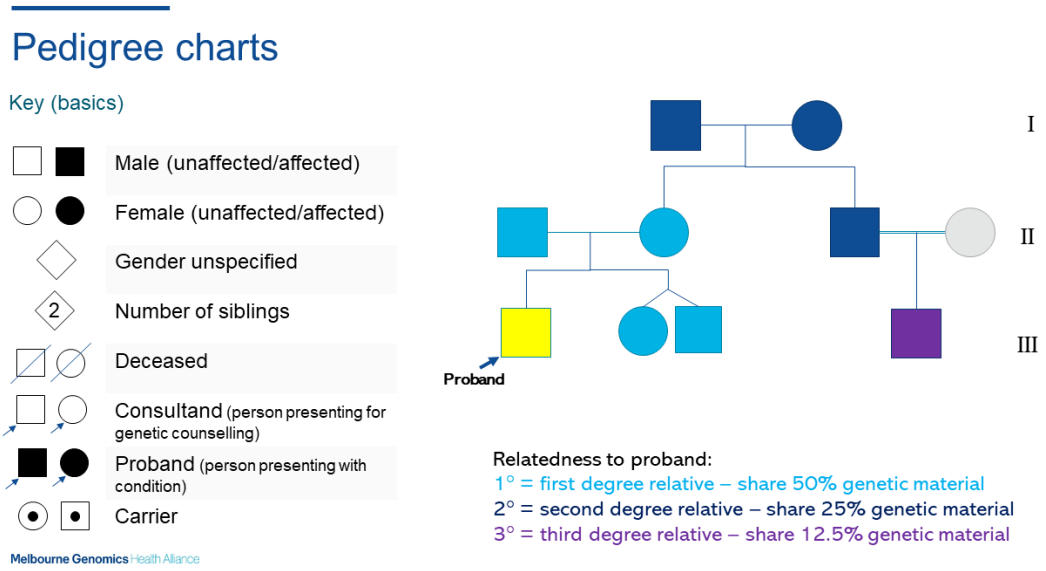


Figure 32 Key features of a pedigree chart with some standard symbols for males and females, with 1st, 2nd and 3rd degree relatives of the proband indicated (colour code).

Review Inheritance Patterns *10 Inheritance patterns_PDF*

[8] Also see Glossary-Appendix 1

11 Somatic variants and cancer

Variants that arise in mature somatic cells (somatic *de novo* variants) can lead to cancer. These somatic variants are not inherited (Figure 33).

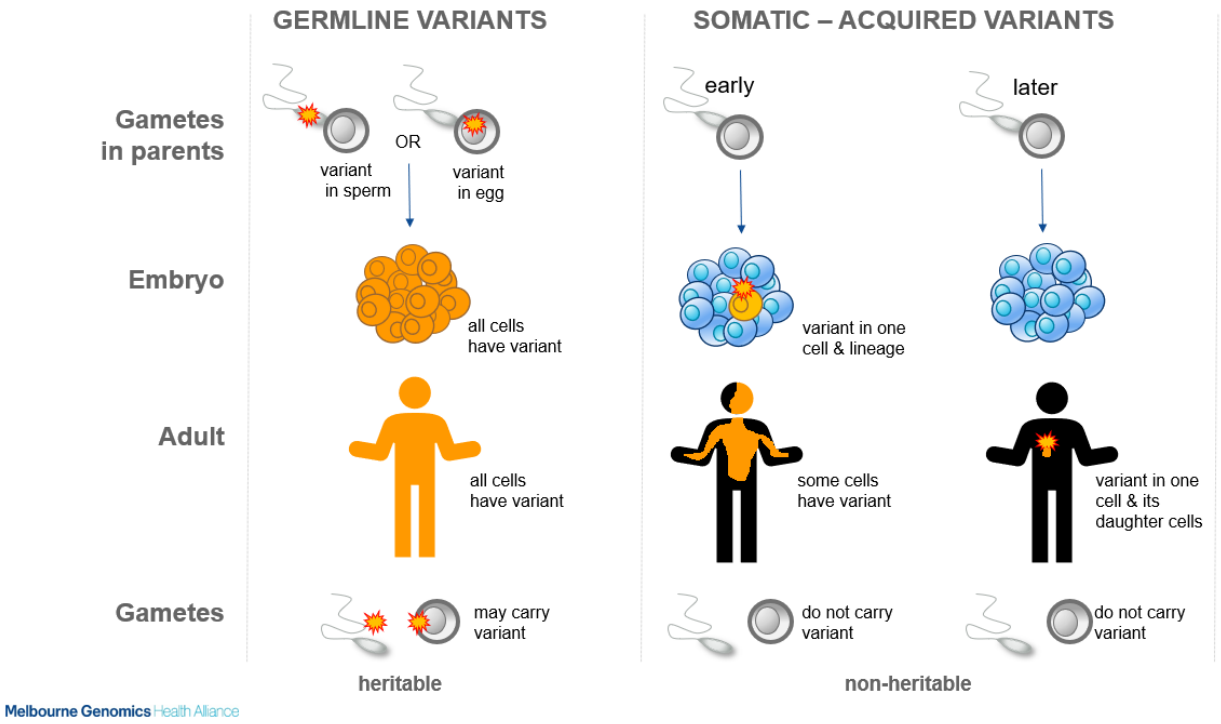


Figure 33 Comparison of germline and somatic variants. Some germline variants are associated with cancers. Most cancers are caused by somatic – acquired mutations.

Genomics analysis in cancer is concerned with the variant(s) in the tumour cells. Cancer can arise from a single variant or more often from a series of mutation events. The testing and analysis of cancer variants shares the principles of germline variant analysis but diverges in some aspects of testing and the variant interpretation process, as noted in earlier sections (Sections 8 and 9).

For purposes of this book we will summarise the key terms and principles applying to acquired cancer variants and refer you to the **online Module 4 Somatic Variants** for further information. The focus is on solid tumours. Blood cancers may be further complicated by having both inherited and acquired variants. Testing and analysis of haematological cancers therefore differs in testing for both goups of variant.

See this introductory video from Genomics Education Program UK: How is genomics used in cancer care

Background – cancer genetics and cell biology

Key features of cancer cell biology and genetic changes contributing to cancer are summarised below. See **online Module 4** for further information on each topic.

Genomic changes: A wide range of genomic alterations and biological effects are the hallmarks of cancer; variants include single nucleotide substitutions, insertions and deletions, structural rearrangements, copy number variants and gene fusions. Cancers often acquire more variants over time, leading to complex genetic profiles.

Cancer-causing genes: Genes commonly involved in cancer include tumour suppressor genes and proto-oncogenes. Tumour suppressor genes normally prevent unregulated cell growth.

Variants that inactivate the genes (loss of function variants) lead to excessive cell proliferation. Proto-oncogenes usually code for cell signalling molecules that promote cell growth. Variants that activate these genes (gain of function variants) lead to excess or un-regulated cell proliferation.

Genetic alterations can influence one or several steps in a cellular biochemical pathway. Understanding where in a pathway a genetic change is acting is a foundation for the therapeutic implications of a genomic test result.

Tumour heterogeneity: Tumours have distinct morphological, phenotypic and genetic profiles. Solid tumours contain a diverse collection of cells and sub-populations with distinct genetic alterations. Analysis of tumour variants needs to account for the cellular profile, for example the proportion of cancerous cells in a tumour sample, and the range of variants acquired as a tumour develops.

Tumour mutation burden: cancers tend to accumulate more genetic changes over time. The total load of variants, the tumour mutation burden (TMB), reflects the evolution of the tumour and can potentially provide predictive biomarkers for therapy.

Tumour signatures: Characteristic combinations of mutation types, i.e. specific types of base change arising due to errors in DNA replication and repair, or the action of genotoxins; they tend to occur in recognisable combinations in different types of cancer, providing a genetic 'signature'.

Procedures and reporting in cancer genomics

Tissue sample for DNA preparation: Appropriate tissue sampling and preparation are essential for DNA quality for NGS. Tumour purity, shown by histopathology, is required for variant analysis.

Genomic analysis: NGS for cancer genome testing generally uses gene panels established for known cancer variants and types of cancer; these include a comprehensive cancer panels and smaller targeted panels. Sequencing for blood cancers usually employs panels for both acquired and inherited variants. Identification of specific variants in solid tumours may involve sequencing of both the tumour and a blood sample (non-cancer cells) for a comparison of the cancer cell genome to the patient's own germline genome (a tumour-normal analysis). Cancer panels are designed to detect the wide range of genetic changes common in cancers. WES and WGS are also used in cancer genomic testing.

Range of genetic variants identified and reported: Variant heterogeneity and accumulation of variants within a tumour, combined with the variety of genetic changes that can occur in cancers, often results in several variant and different types of variant identified, curated and reported, including somatic SNVs, CNVs, gene fusions, mutation signatures, and germline variants.

Classification and therapeutic implications: Variant classification employs ACMG guidelines and additional AMP guidelines to consider diagnostic, prognostic and therapeutic implications (see pages 29-30).

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

Melbourne Genomics Health Alliance

melbournegenomics.org.au

C/- Walter and Eliza Hall Institute
1G Royal Parade, Parkville, Victoria 3052

Tel: +61 3 9936 6499

Email: enquiries@melbournegenomics.org.au

Alliance members

The Royal Melbourne Hospital

The Royal Children's Hospital Melbourne

The University of Melbourne

Walter+Eliza Hall Institute of Medical Research

murdoch children's research institute

CSIRO

agrf

Peter Mac Peter MacCallum Cancer Centre Victoria Australia

Austin Health

Monash Health

Supported by

Victoria State Government