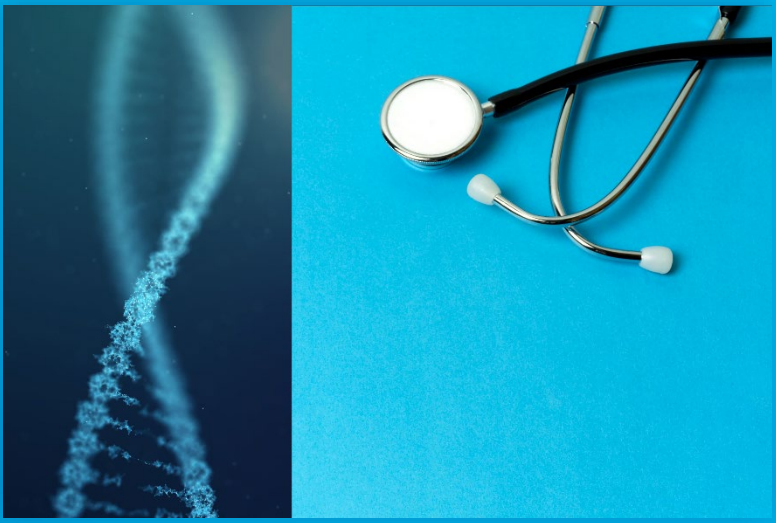
Melbourne Genomics
Health Alliance

# Clinical Genomics:
## Background Biology



Alliance members

The Royal Melbourne Hospital
The Royal Children's Hospital Melbourne
The University of Melbourne
WEHI brighter together
murdoch children's research institute
CSIRO
agrf
Peter Mac Peter MacCallum Cancer Centre Victoria Australia
Austin Health
Monash Health

Supported by

Victoria State Government

This ebook is prepared as a foundation resource for clinicians and other health professionals upskilling in genomics for clinical practice. It is one resource provided for those participating in a Melbourne Genomics Health Alliance clinical genomics workshop or engaging with our self-directed online learning modules in clinical genomics.

The biological depth is more complex than for a general reader but less complex than the level used by molecular biologists, bioinformaticians and geneticists investigating clinical genomic variation.

View this book in your pdf viewer to access the accompanying hyperlinked animations/videos identified by the symbol 📽. You can also print a hard copy to add your own notes.

## Authors and acknowledgements

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

Acknowledgements: Thanks to Taryn Charles, Dr Zoe McCallum, Dr Chloe Stutterd and Dr Natalie Thorne for review, suggestions and comments for improving this resource. Nat Thorne, Dr Sebastian Lunke, A/Prof Tiong Tan, Dr Ain Roesley and Dr Andrew Fellowes provided advice on sequencing technology, and germline and somatic variant interpretation, and kindly shared slides adapted for some diagrams and animations. Dr Kortnye Smith advised on key points for cancer genomics testing. Unless otherwise specified, illustrations were drawn by FM using PowerPoint, Adobe Illustrator or BioRender.

© Melbourne Genomics 2023 (2nd edition)

https://www.melbournegenomics.org.au/

Contact education@melbournegenomics.org.au with feedback or suggestions about this ebook

# Table of Contents

# 1 Overview - 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 about 20,000 protein-coding genes distributed across a set of 23 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.

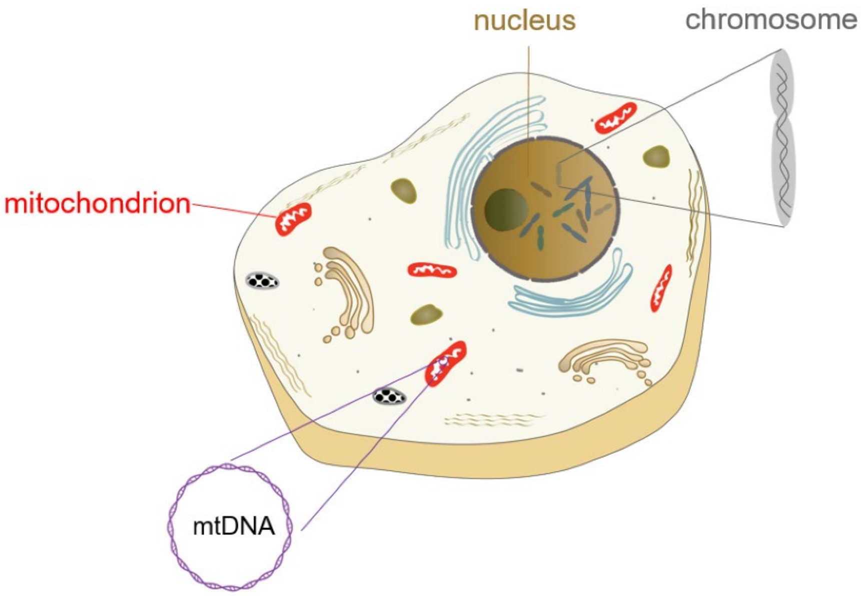


*Figure 1 Eukaryotic cells have DNA in the form of linear chromosomes located in the nucleus and circular mitochondrial DNA (mtDNA) located in the mitochondria*

## Genomic Alterations

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 small variants cause clinically relevant conditions because they adversely alter the proteins encoded by genes or affect timing or level of gene expression, affecting cell function and causing disorders.

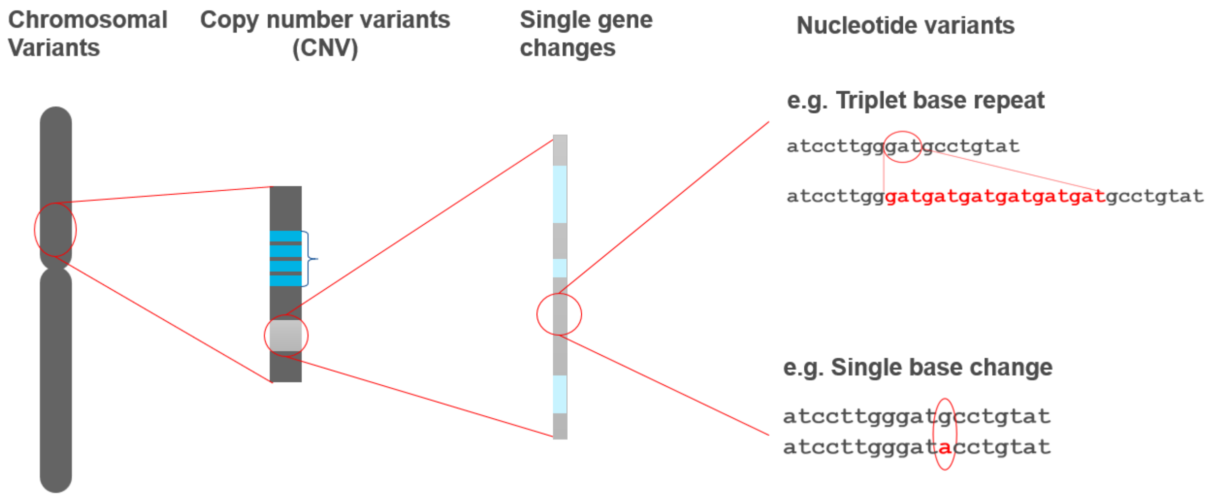


*Figure 2 Genomic variation occurs at different levels, from whole chromosome to single nucleotide changes.*

Once identified through genomic testing, a variant is classified as pathogenic, benign or of uncertain significance, according to the likelihood that it is the cause of the condition (Figure 3).

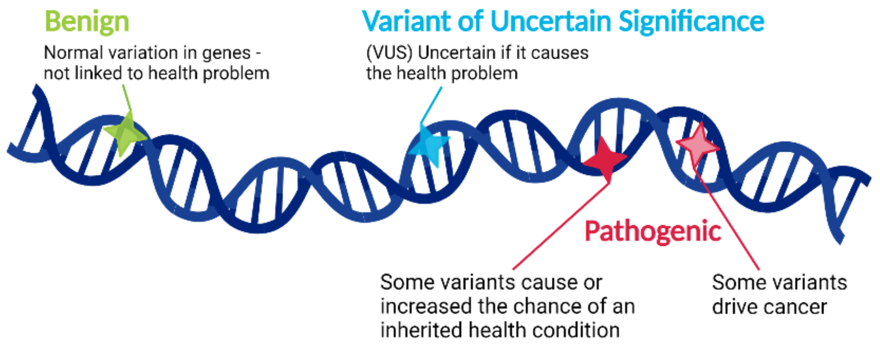

Benign
Normal variation in genes - not linked to health problem
Variant of Uncertain Significance
(VUS) Uncertain if it causes the health problem
Pathogenic
Some variants cause or increased the chance of an inherited health condition
Some variants drive cancer

*Figure 3 Variants are classified on a scale of benign (not disease causing) to pathogenic (the cause of the condition).*

## Germline and Somatic Variants

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 (Figure 4).

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 on the timing of the genetic change during development. Some forms of epilepsy, for example, 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.

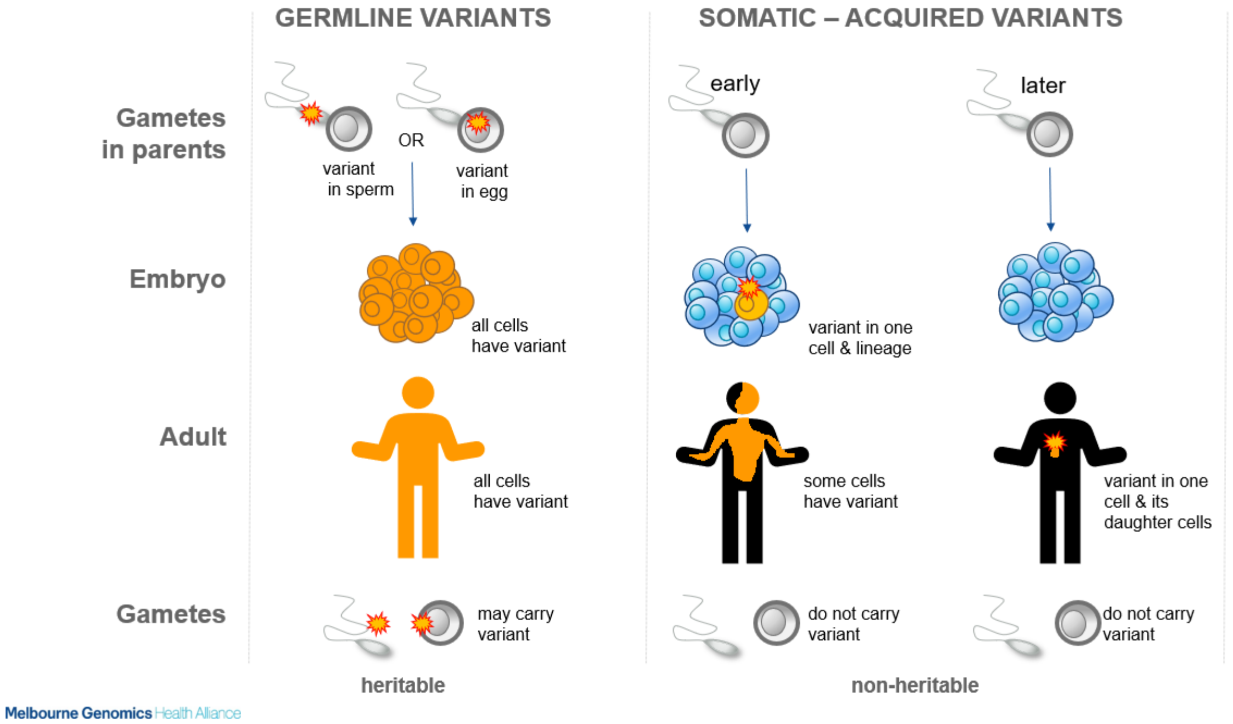


*Figure 4 Comparison of germline and somatic variants. Germline variants are the cause of heritable/familial cancer predisposition syndromes. Most cancers are caused by acquired somatic mutations.*

## Monogenic and Polygenic Conditions

A monogenic condition is one caused by a pathogenic variant in one gene. Monogenic conditions are caused by variants in a single gene and segregate with disease in families according to the Mendelian principles of inheritance. However, variants can also arise *'de novo'* in an individual (a new variant showing no family history). There could be more than one candidate gene responsible for a monogenic condition; this is known as genetic heterogeneity.

Polygenic conditions have more than one contributing genetic factor, and multifactorial conditions are caused by genetic and environmental factors.

*Review Monogenic, polygenic, multifactorial conditions*

## Inheriting Germline Variants

Clinicians should be familiar with taking detailed family histories to look for possible inherited conditions. This information is important to include on test request forms as it helps geneticists and genome analysts identify variants relevant to the patient's condition.

Heritable germline variants associated with monogenic conditions typically have recognisable patterns of inheritance, e.g. characteristic of recessive or dominant traits. This information is important in genomic analysis, for example to 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 degrees of phenotypic change).

Terminology for describing inheritance patterns include the following (also see the Glossary[1]). 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 a detailed family history, including accurate pedigree charts, aids clinical genetics and genomics analysis (Figure 5).

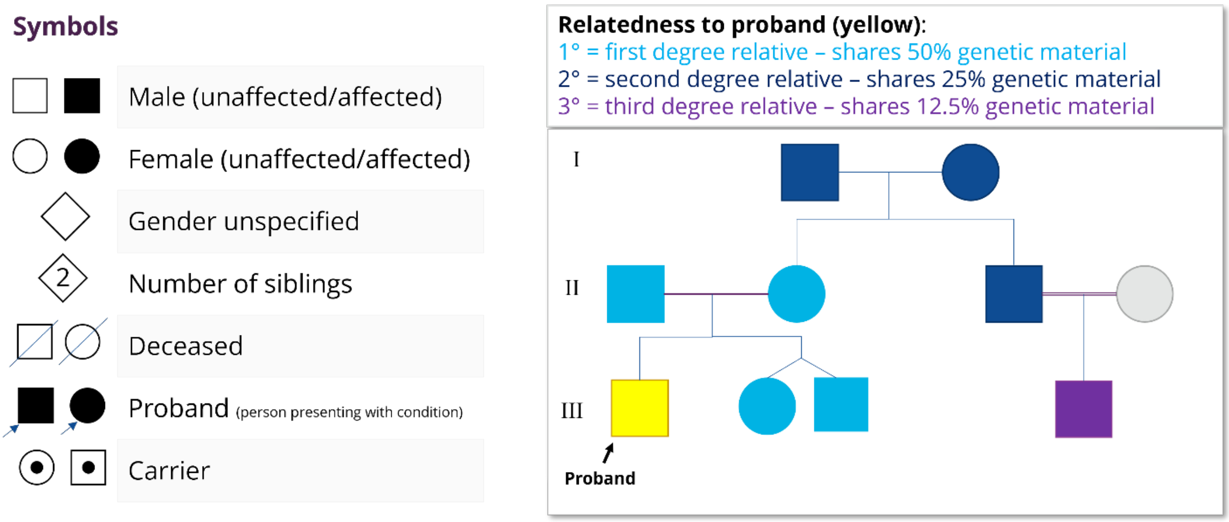

Symbols
Male (unaffected/affected)
Female (unaffected/affected)
Gender unspecified
2
Number of siblings
Deceased
Proband (person presenting with condition)
Carrier
Relatedness to proband (yellow):
1° = first degree relative – shares 50% genetic material
2° = second degree relative – shares 25% genetic material
3° = third degree relative – shares 12.5% genetic material
I
II
III
Proband

*Figure 5 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).*

***Refresh your memory about inheritance patterns*** *Inheriting variants*

## Genetic and cellular alterations in cancer

*View an introductory video from Genomics Education Program UK: How is genomics used in cancer care*

Most cancer is 'somatic' cancer, occurring due to non-heritable genetic alterations in somatic cells. Genomic testing of somatic cancer aims to find the genetic alterations present in the cancer that are not in the individual's normal germline genetic information.

Some cancers occur as part of a hereditary or familial cancer susceptibility syndrome caused by heritable germline variants. These syndromes confer a higher-than-normal risk of developing certain types of cancer. They include Lynch Syndrome, Li Fraumeni Syndrome and hereditary breast and ovarian cancer syndrome.

Genetic changes occurring in cancer alter cell growth and function. Changes include:

[1] Also see Glossary-Appendix 1

## Genomic alterations

A range of genomic alterations occur in cancer, including single nucleotide substitutions, insertions and deletions, chromosomal rearrangements, copy number variants and gene fusions. Cancers typically acquire more variants over time, leading to complex genetic profiles. Cells within a tumour acquire different genetic variants, so tumours develop cellular and genetic heterogeneity.

Some cancer-causing genetic alterations affect DNA repair mechanisms, turning on genes that promote cell growth or turning off genes that limit cell growth.

## Biochemical changes

Genetic alterations in cancer can influence one or several steps in a cellular biochemical pathway. Understanding where in a pathway a genetic change is acting informs the development of therapeutics and can help in understanding clinical implications of a genomic test results.

## Cancer-causing genes

Genes commonly involved in cancer include tumour suppressor genes (TSG) and proto-oncogenes.

- Tumour suppressor genes normally prevent unregulated cell growth; inactivation due to mutation (loss of function variants) lead to excessive cell proliferation. Mutations in tumour suppressor genes are the basis of many cancer susceptibility syndromes.
- 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 unregulated cell proliferation.
- Genes and proteins involved in DNA repair. These include mismatch repair (MMR) genes and homologous recombination (HR) genes. Deficiencies in these DNA repair pathways can underpin heritable cancer susceptibility syndromes and somatic cancers.

# 2 DNA and Chromosomes

## DNA structure

The DNA double helix is built from paired deoxyribonucleotides (we will refer to them as 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 6). The length of a DNA molecule is described in terms of the number of base pairs (bp).

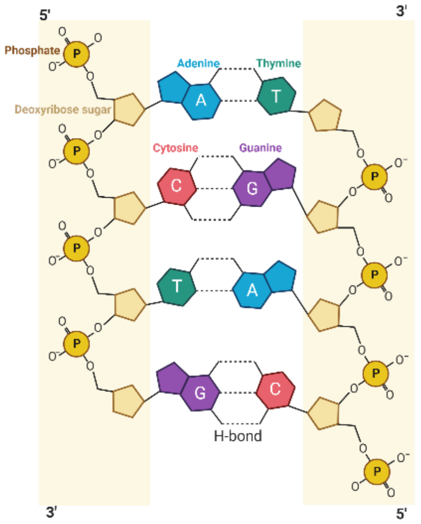


*Figure 6  Chemical structure of DNA*

## RNA Structure

Ribonucleic acids (RNA) are involved in expression and regulation of genomic information. Like DNA, RNA is composed of a chain of nucleotides, but in RNA the sugar is a ribose sugar, and the base uracil (U) replaces thymine, so the four RNA bases are A, C, G and U. Types of RNA found in cells are:

- ribosomal RNA (rRNA), a component of ribosomes
- transfer RNA (tRNA)
- messenger RNA (mRNA)

## Chromosomes in the human genome

A chromosome is one linear DNA double-helix molecule that contains many genes. 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)
- chromosomes vary in size and number of protein-coding genes (see ⓘ box)
- the chromosomes reside in the nucleus of eukaryotic cells
- human somatic cells have two sets of 23 chromosomes - they are diploid
- gametes (ova and sperm) contain one set of 23 chromosomes - they are haploid

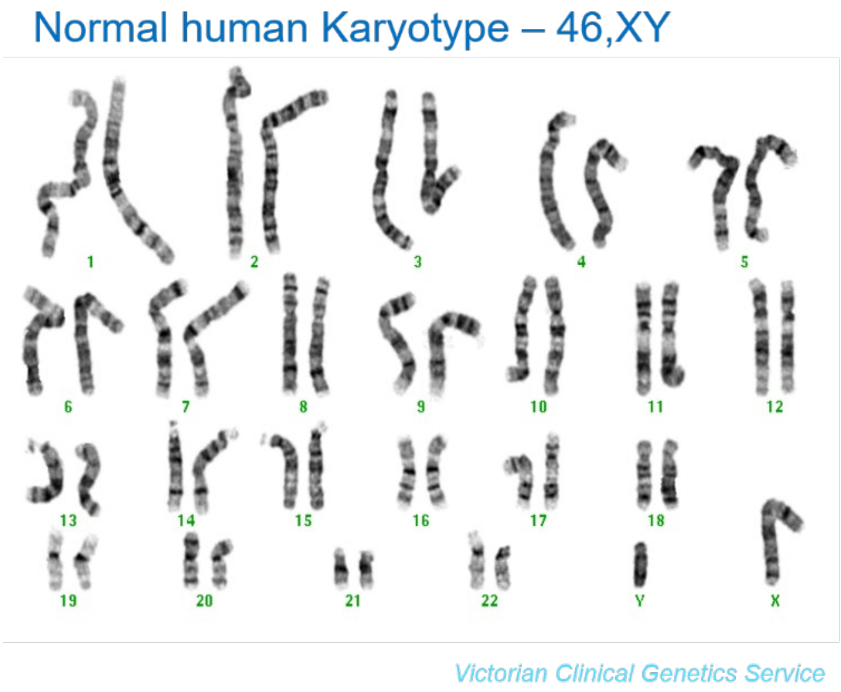
***ⓘ Chromosomes - the long and the short***

*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*

## Karyotypes and Ideograms

Chromosomes can be seen under the microscope after staining. The 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 7) shows the set of chromosomes arranged 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. Chromosomes can also be visualised by fluorescent staining methods.



Normal human Karyotype – 46,XY

Victorian Clinical Genetics Service

*Figure 7 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 8a) 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 following example describes the molecular location of a gene within a chromosome (Figure 8b):

- The gene for amyloid precursor protein (APP) is located on chromosome 21
- Chromosome 21: length = 46,709,983 base pairs (bp)
- Cytogenetic Location of APP: 21q21.3 = q arm of chromosome 21, section 2, subsection 1.3
- Molecular Location of APP: base pairs 25,880,550 to 26,171,128 (bp numbering not shown on diagram)

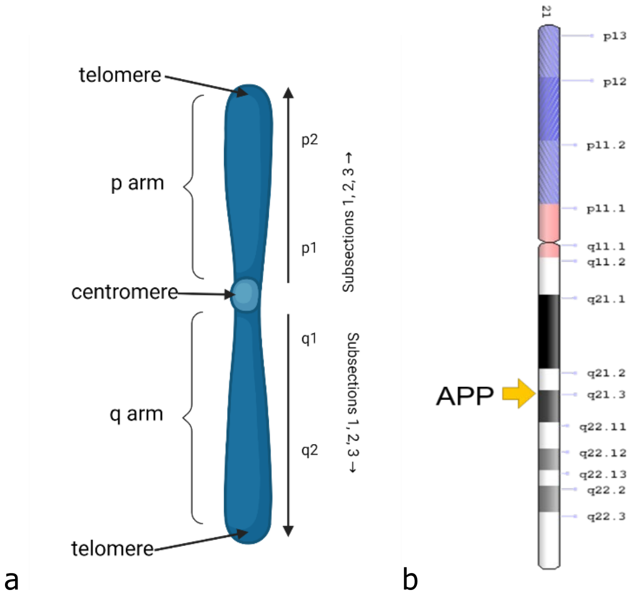


*Figure 8 Chromosome structure and labelling: (a) 'p arm' = short arm, 'q arm = long arm, Centromere – where chromosomes attach to spindle fibres during mitosis and meiosis (cell division). Chromosome section numbering goes from centromere towards telomere for each arm. Telomeres are chromosome ends. (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 (abbreviated as mtDNA) (Figure 9). 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.

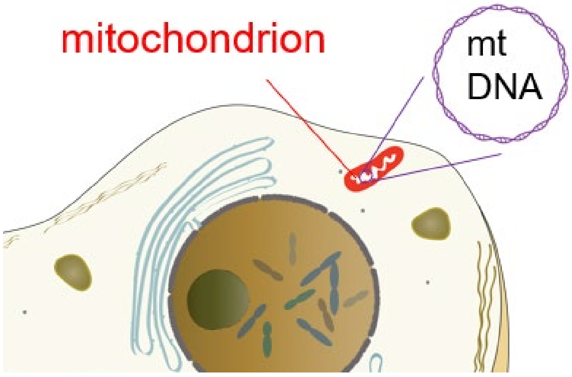


*Figure 9 Mitochondria contain multiple copies of the circular mitochondrial DNA (mtDNA). mtDNA replicates independently of the cell. Human mtDNA is 16,569 bp long and carries 37 genes (13 protein-coding genes and 24 non-protein-coding genes).*

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

Cells can carry mitochondria with only normal/'wild type' mitochondrial DNA, called homoplasmy, or mitochondria with a mixture of wild type and variant mtDNA, called heteroplasmy[3] (Figure 10a). Also, individual mitochondria can carry different proportions of wild type and variant mtDNA molecules. The proportions of different mtDNA molecules vary depending on the effect of the variant on mtDNA replication rate and mitochondrial replication rate. Different tissues within an individual can display different proportions of wild type and variant mitochondria (Figure 10b).

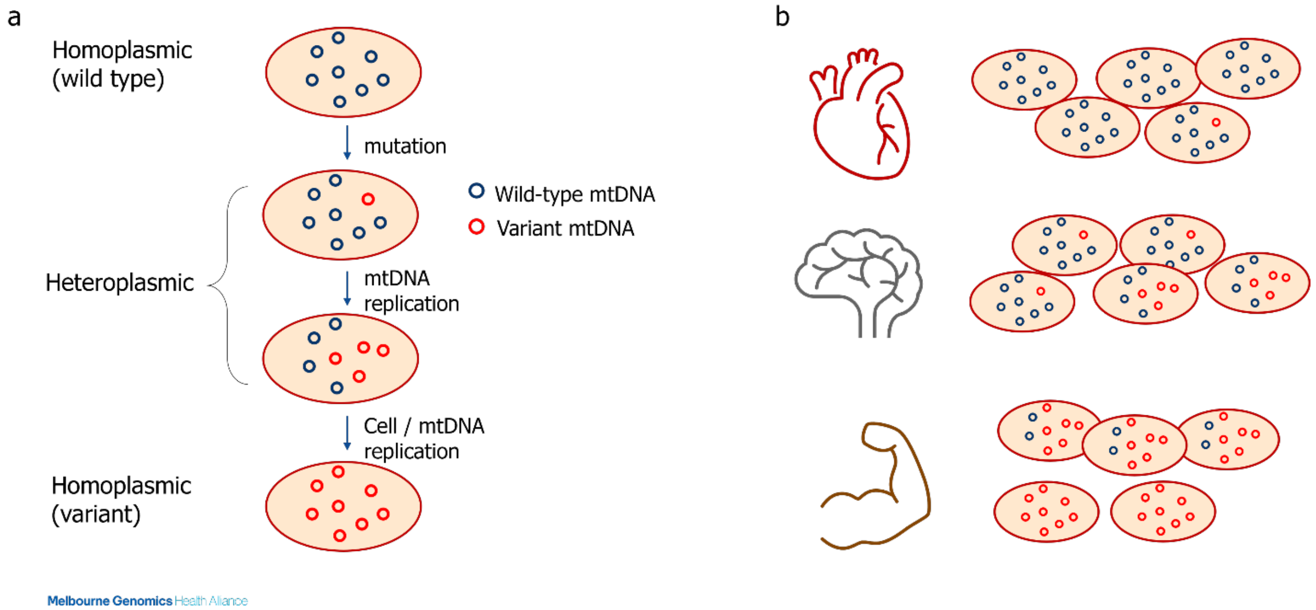


*Figure 10 (a) Mitochondria can carry one type of mtDNA (homoplasmy) or a mixture of normal and variant mtDNA (heteroplasmy). (b) Cells and tissues can vary in their mitochondrial populations.*

[3] Stewart and Chinnnery 2016 Nature Reviews 16:530-542 doi:10.1038/nrg3966

# 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 11).

*Genes big and small*

*The TTN gene is 281,434 bp and codes for titin, a very large protein*

*The INS gene is 1,430 bp and codes for insulin, a small protein*

## 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) lie either side of the coding sequence

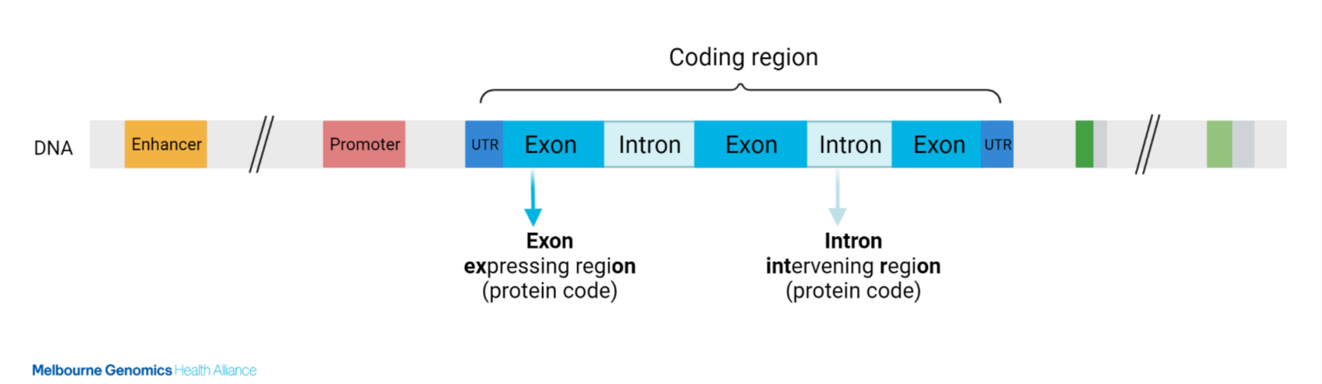


*Figure 11 Structure of a gene. The coding region contains exons (protein coding sequence) and introns (intervening sequence). Regulatory regions upstream and/or downstream of the coding region include the promoter and enhancers.*

*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. The process involves **transcription** – copying the DNA sequence into RNA, which moves into the cytoplasm, 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.

$$\text{DNA (gene)} \xrightarrow{\textit{transcription}} \text{mRNA} \xrightarrow{\textit{translation}} \text{Protein}$$

The steps in gene expression are shown in Figure 12, with a closer view of the action at the ribose in Figure 13.

- Transcription occurs in the cell nucleus. One strand of the DNA double helix is transcribed into RNA; this is called the primary transcript
- The primary RNA transcript 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

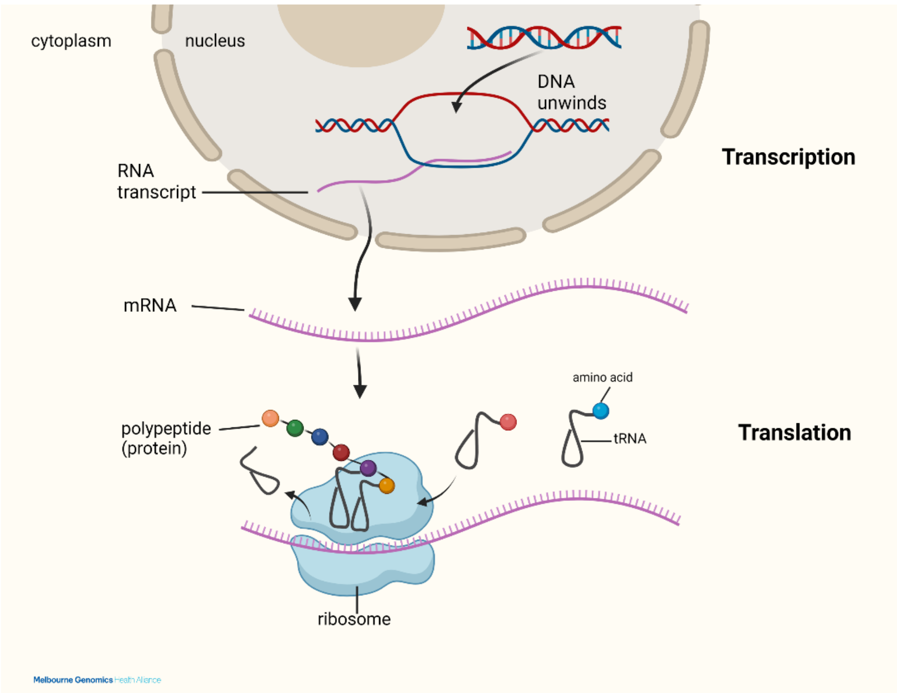


*Figure 12  Gene expression - transcription and translation*

A closer look at the action in the ribosome during translation.

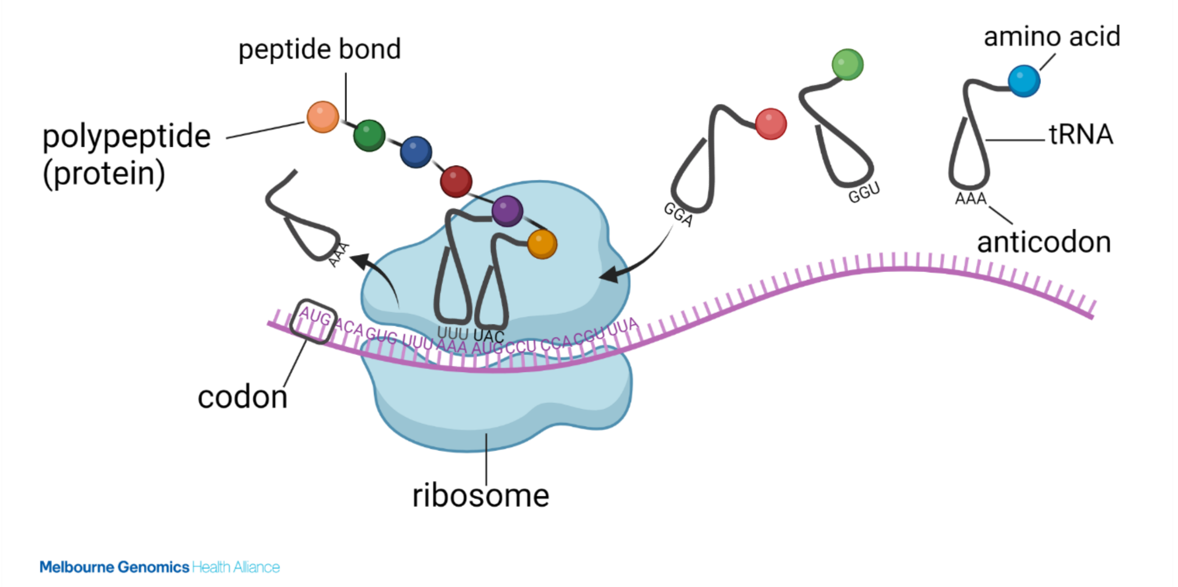


*Figure 13 Translation. Transfer RNAs, each loaded with an amino acid specific for the anticodon sequence, enter the ribosome docking stations and match up with the appropriate codon, e.g. mRNA codon UUU pairs with tRNA anticodon AAA, which carries phenylalanine. When tRNAs dock together a peptide bond forms between the amino acids, the tRNA releases the amino acid and returns to the cytoplasm to collect another amino acid and do the job again.*

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 14). 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.

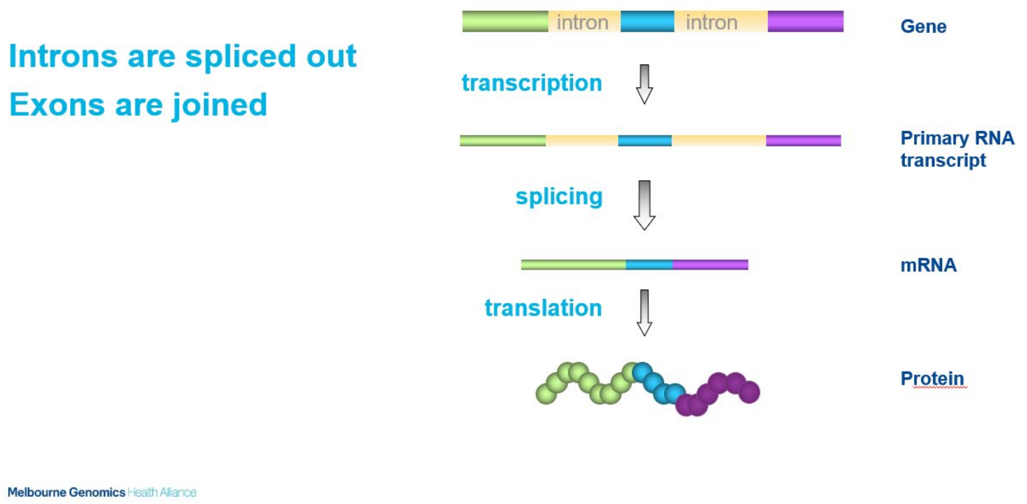
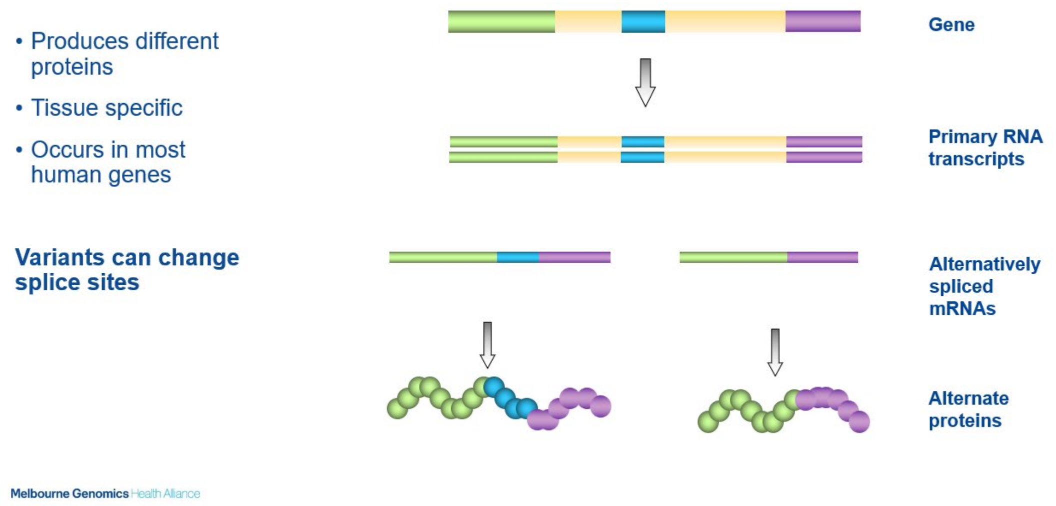


*Figure 14  Splicing removes introns from the primary transcript and joins the exons.*

**Other modifications**: mRNA is stabilised by chemical modifications, including additions 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 these diagrams).

## 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 15). 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.



- Produces different proteins
- Tissue specific
- Occurs in most human genes

**Variants can change splice sites**

Gene
Primary RNA transcripts
Alternatively spliced mRNAs
Alternate proteins
Melbourne Genomics Health Alliance

*Figure 15  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.

## 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[4].

The 'primary structure' refers to 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; the insulin receptor with two alpha-subunits and two beta subunits; the tumour suppressor protein p53 which forms a tetramer.

### Types of proteins

Enzymes

Antibodies

Ion channels

Transcription factors

Extracellular matrix proteins

Nutrient transporters

Cell signalling molecules: peptide hormones, cytokines

Receptors for hormones, cytokines and neurotransmitters

Cytoskeleton proteins

## 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<br>• Secreted, e.g. hormones<br>• Plasma membrane, e.g. ion channels, hormone receptors, nutrient transporters<br>• Mitochondria, e.g. cytochromes<br>• Intracellular cytoskeleton<br>• Extracellular matrix | • The substrate binding site of enzymes<br>• The ion binding sites of ion channels<br>• The hormone binding domain of receptors<br>• The antigen binding domain of antibodies<br>• DNA binding domain of transcription factors<br>• Transmembrane domains of membrane-spanning proteins |

*5 Proteins*

[4] Further reading about amino acids and proteins: video from RCSB Protein Data Bank 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***.

## 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 do not 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 16) 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** structural variants involve a gain or loss of genetic information, such as deletions, duplications and triplet repeat expansions.

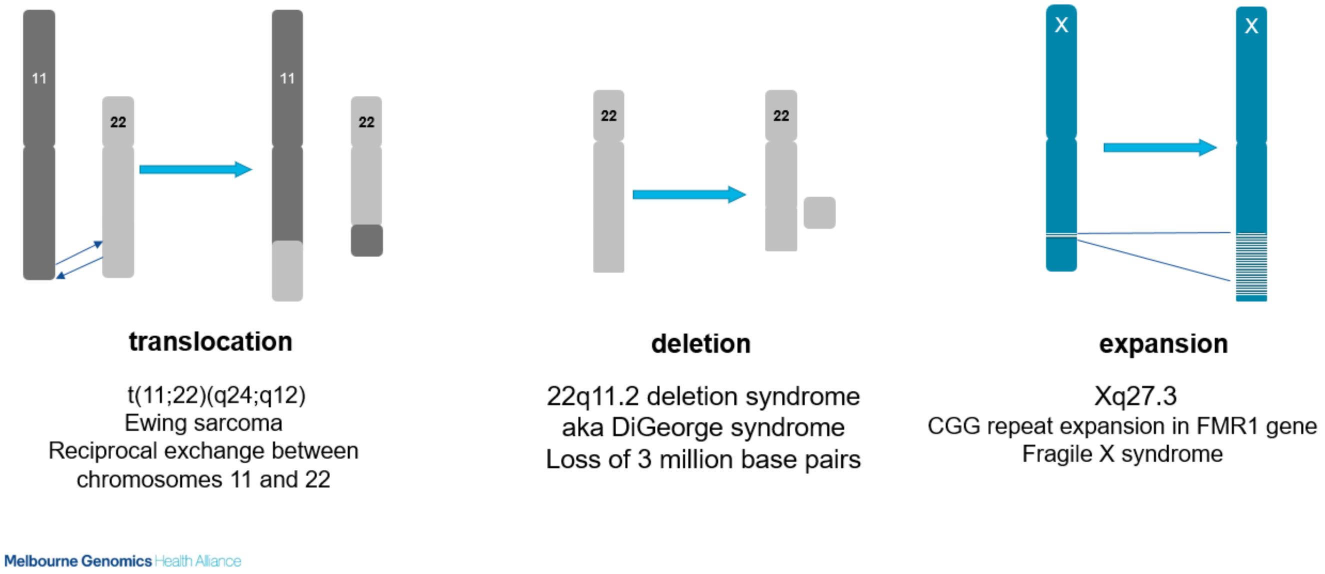

11
22
11
22
translocation
t(11;22)(q24;q12)
Ewing sarcoma
Reciprocal exchange between
chromosomes 11 and 22
22
22
deletion
22q11.2 deletion syndrome
aka DiGeorge syndrome
Loss of 3 million base pairs
X
X
expansion
Xq27.3
CGG repeat expansion in FMR1 gene
Fragile X syndrome
Melbourne Genomics Health Alliance

*Figure 16 Structural variants (examples associated with disease)*

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 - gene fusion, deregulated gene expression; new protein function (e.g. *BCR-ABL*)
- Inversion - gene activation or inactivation (e.g. *ALK*)

## Gene fusion

*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 17). They are common type of structural variant in cancer. Gene fusions are difficult to detect by current whole genome sequencing methods and are detected by FISH, targeted panels, or RNA sequencing of transcripts (the latter not yet clinically validated).

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))

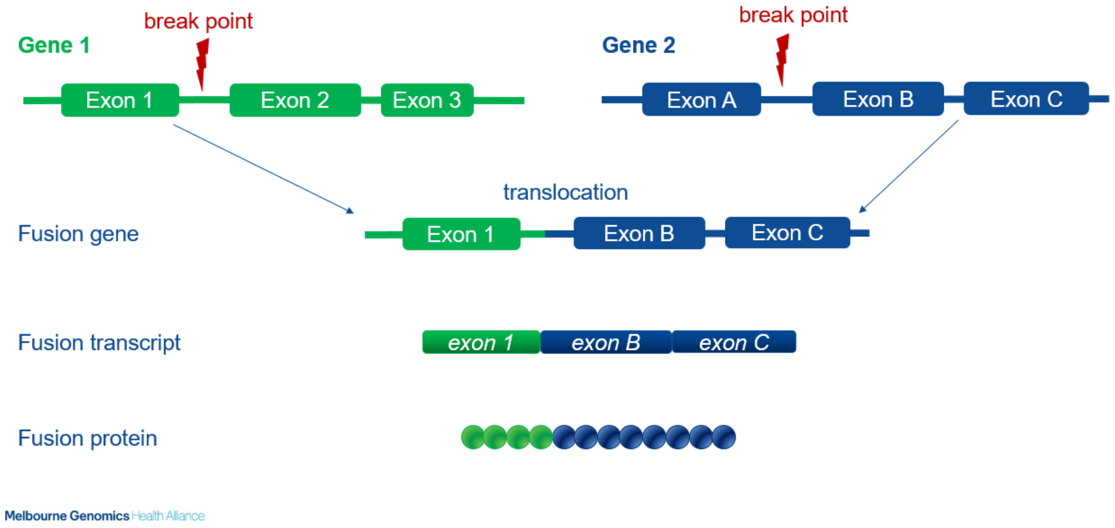


*Figure 17 Gene fusion can occur during chromosomal rearrangements such as translocation*

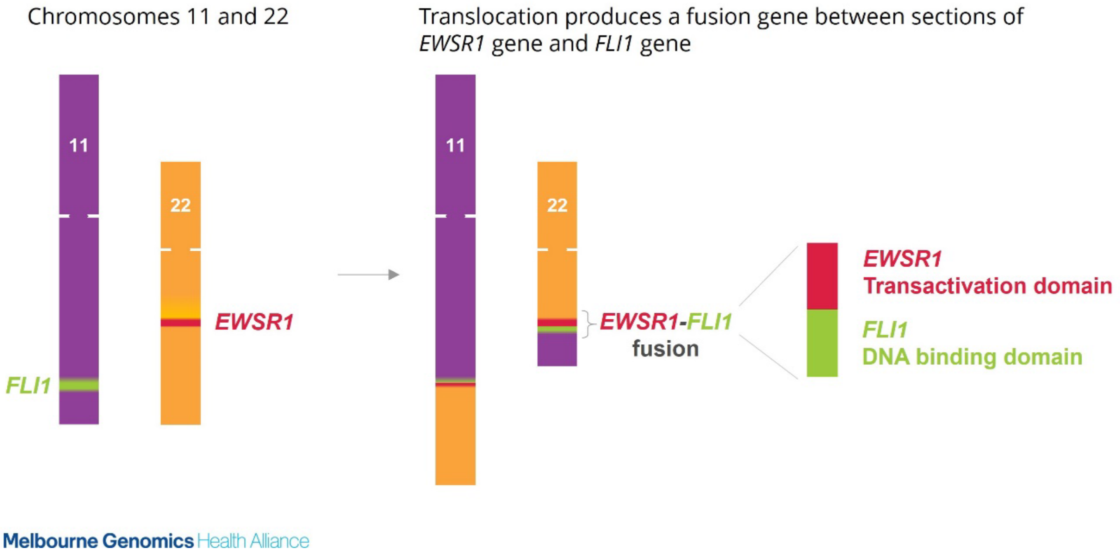

Chromosomes 11 and 22
Translocation produces a fusion gene between sections of *EWSR1* gene and *FLI1* gene
11
22
EWSR1
FLI1
11
22
EWSR1-FLI1
fusion
EWSR1
Transactivation domain
FLI1
DNA binding domain
Melbourne Genomics Health Alliance

*Figure 18 Illustration representing the translocation between chromosomes 11 and 22 leading to fusion of portions of the FLI1 gene from chromosome 11 and the EWSR1 gene on chromosome 22. This is one of several fusion proteins associated with Ewing sarcoma[5].*

## 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, and typically range from 50bp to 3,000,000 bp (3Mb) (Figure 19). They may be too small to detect with karyotype analysis and are detected with molecular karyotyping with chromosomal microarrays or other specialised tests (e.g. MLPA).

[5] Jo, VY (2020) EWSR1 fusions: Ewing sarcoma and beyond, Cancer Cytopathology 128(4):229-231 **https://doi.org/10.1002/cncy.22239**

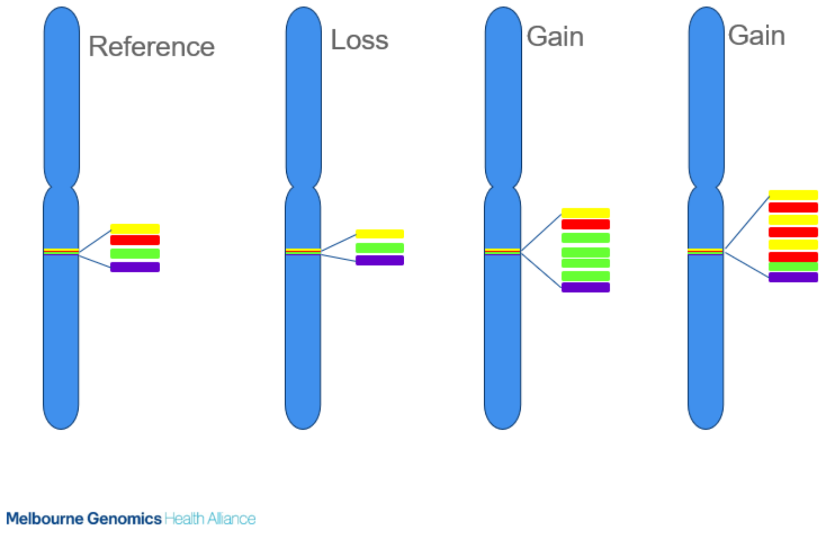


*Figure 19 Copy number variants (CNV) are gains or losses of genetic information. Small CNVs are detected with molecular karyotyping.*

## Triplet repeats

A **triplet repeat expansion** is when a group of three nucleotides repeat many times in tandem on the same stretch of chromosome, causing expansion of the region.

Examples include:

- The CGG repeat in the *FMR1* gene on the X chromosome, with > 200 repeats causing Fragile X syndrome
- The CAG repeat in the *HTT* gene for huntingtin protein (Figure 20). The normal huntingtin gene has up to about 26 CAG repeats, while more than around 40 repeats causes Huntington disease.

Repeat expansions are not confined to 3 base pairs; they typically occur for 1 to 6 base pairs.

Triplet repeats are detected by PCR methods and Southern blotting. They are not well detected by genomic sequencing or chromosomal microarrays.

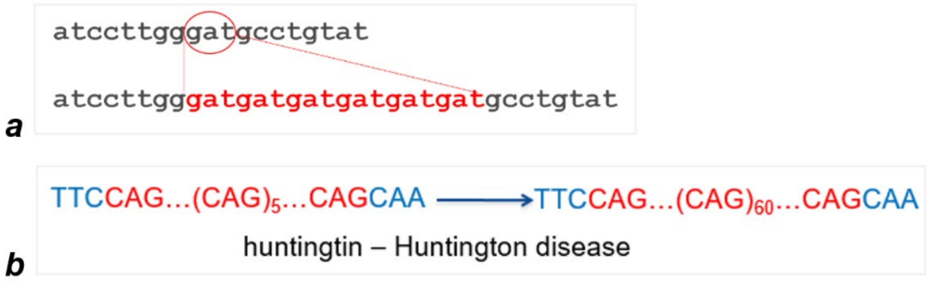

atccttgggatgcctgtat
atccttgggatgatgatgatgatgatgatgcctgtat
a
TTCCAG…(CAG)5…CAGCAA
TTCCAG…(CAG)60…CAGCAA
huntingtin – Huntington disease
b

*Figure 20 (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 (SNVs and SNPs)

6.5 SNV_SNP

A **single nucleotide variant** is a change in a single base pair, including base substitutions, insertions or deletions (called indels) and duplications (Figure 21). 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** (**SNP**, pronounced ‘SNiP’) and are usually not a direct cause of disease.

Other small variants may involve a small number of nucleotides. A variant involving both a deletion and insertion is called a delins.

SNVs and other small variants are detected by PCR-based methods targeting specific known variants or by sequencing of single genes, exomes or whole genomes.

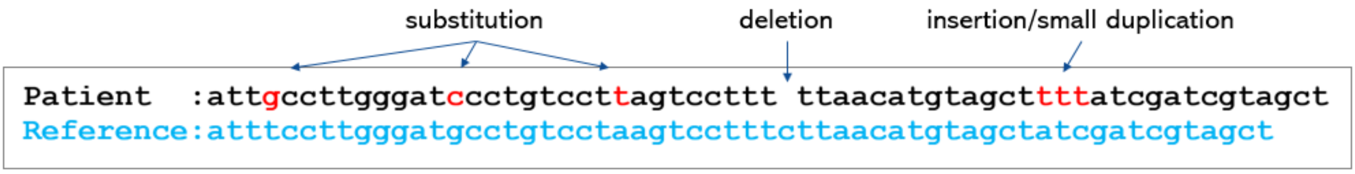

substitution
deletion
insertion/small duplication
Patient   :attgccttgggatccctgtccttagtccttt ttaacatgtagctttttatcgatcgtagct
Reference:atttccttgggatgcctgtcctaagtcctttcttaacatgtagctatcgatcgtagct

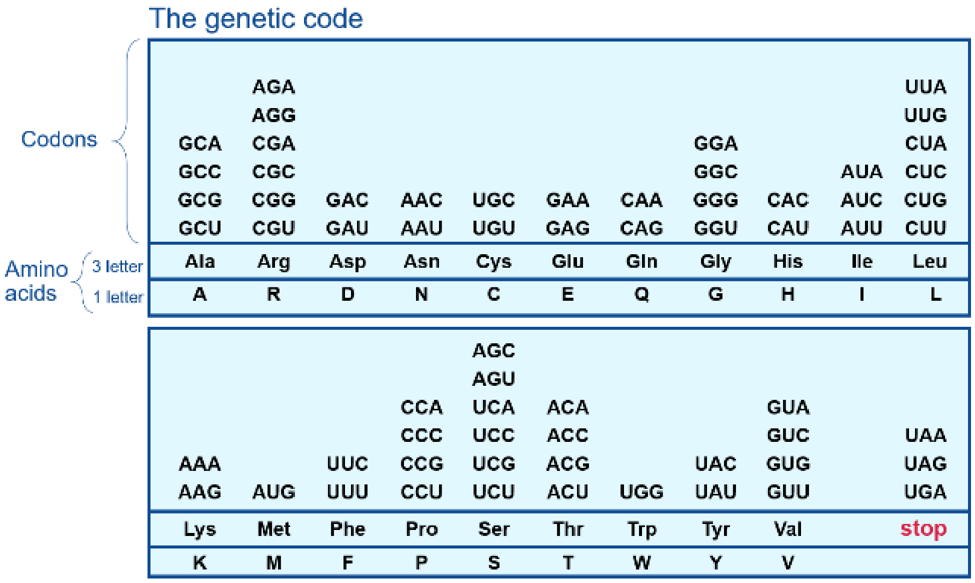
*Figure 21  Single nucleotide variants*

# 7 Consequence of variants

## Effect of variants on proteins

Earlier we saw that during translation the mRNA directs which amino acids come together to build the protein (see Figure 12 and Figure 13). Genetic changes in the DNA sequence are present in the mRNA and, depending on their position in a codon and the type of alteration, may or may not change the amino acid sequence. This is where understanding codons and the genetic code can help understand the potential impact of a variant.

Codons, the groups of 3 bases in the mRNA, specify the sequence of amino acids in a protein, referred to as the genetic code (Figure 22). 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) (Figure 23). A gene can have more than one start codon, and therefore more than one reading frame.



The genetic code

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Codons | GCA<br>GCC<br>GCG<br>GCU | AGA<br>AGG<br>CGA<br>CGC<br>CGG<br>CGU | GAC<br>GAU | AAC<br>AAU | UGC<br>UGU | GAA<br>GAG | CAA<br>CAG | GGA<br>GGC<br>GGG<br>GGU | CAC<br>CAU | AUA<br>AUC<br>AUU | UUA<br>UUG<br>CUA<br>CUC<br>CUG<br>CUU |
| Amino acids 3 letter | Ala | Arg | Asp | Asn | Cys | Glu | Gln | Gly | His | Ile | Leu |
| 1 letter | A | R | D | N | C | E | Q | G | H | I | L |
| | AAA<br>AAG | AUG | UUC<br>UUU | CCA<br>CCC<br>CCG<br>CCU | AGC<br>AGU<br>UCA<br>UCC<br>UCG<br>UCU | ACA<br>ACC<br>ACG<br>ACU | UGG | UAC<br>UAU | GUA<br>GUC<br>GUG<br>GUU | | UAA<br>UAG<br>UGA |
| | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val | | stop |
| | K | M | F | P | S | T | W | Y | V | | |

*Figure 22 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 23) and paragraph illustrates consequences of variants on the protein. Let's consider the effect of different changes to the 6th base in the sequence. Refer to the Genetic Code table to find the amino acid change.

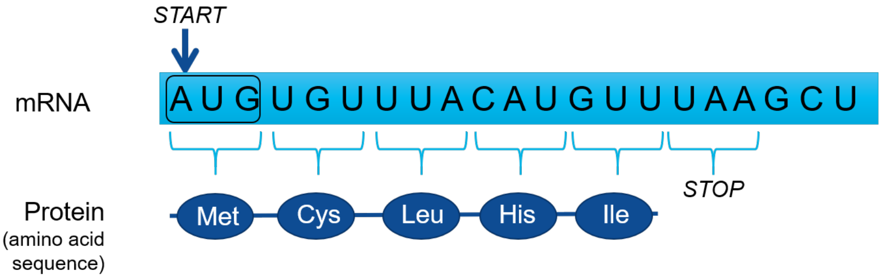


*Figure 23 Codons in mRNA specify the amino acid in the protein sequence. AUG is a start codon for translation. UAA, UAG and UGA are stop codons.*

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.

## 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 an animation for each variant. They are also illustrated and summarised in Figure 24 (next page).

**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.

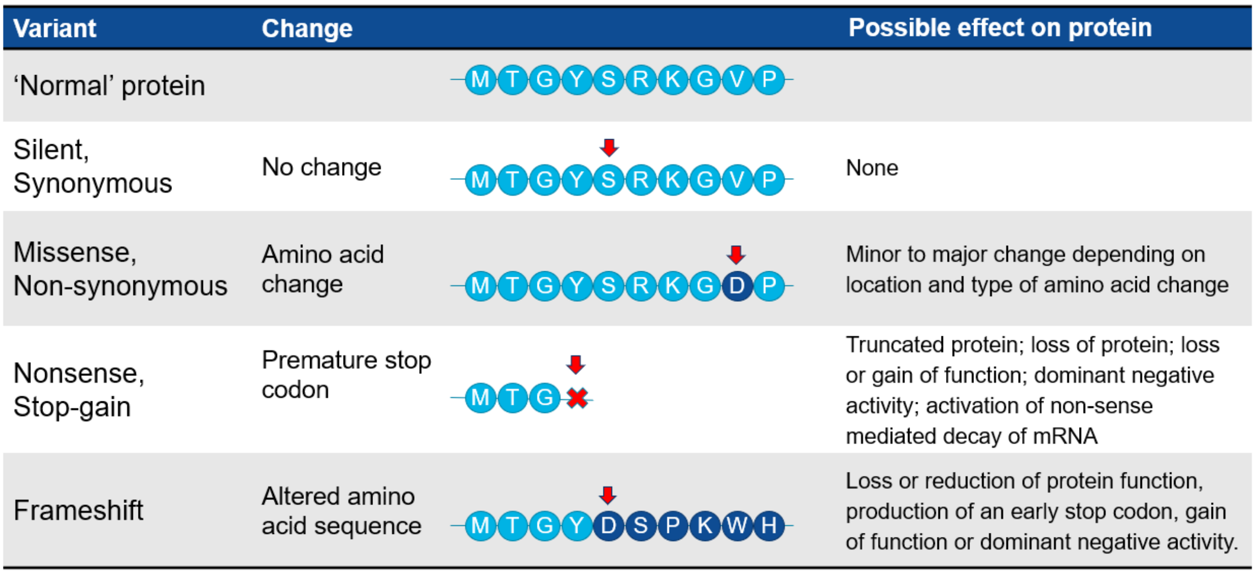
*7.4 'Frameshift variants' - animation*

**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 nonsense mediated decay of mRNA (*see NMD box below)

*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. 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.*



| Variant | Change | | Possible effect on protein |
|---|---|---|---|
| 'Normal' protein | | M T G Y S R K G V P | |
| Silent, Synonymous | No change | M T G Y S R K G V P | None |
| Missense, Non-synonymous | Amino acid change | M T G Y S R K G D P | Minor to major change depending on location and type of amino acid change |
| Nonsense, Stop-gain | Premature stop codon | M T G ✖ | 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 | M T G Y D S P K W H | Loss or reduction of protein function, production of an early stop codon, gain of function or dominant negative activity. |

Melbourne Genomics Health Alliance

*Figure 24 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 25 and Table 2. This section focusses on the **sequencing technologies** to identify nucleotide variants such as base substitutions and small insertions and deletions (indels).

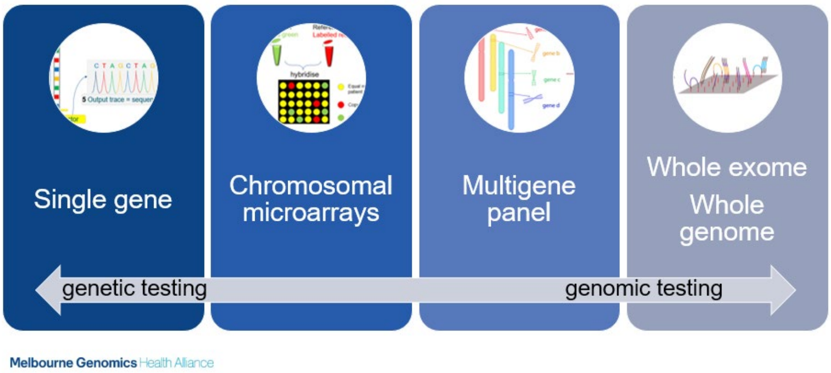


*Figure 25 Spectrum of genetic to genomic testing.*

*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, unbalanced structural rearrangements |
| | PCR-based methods: e.g. PCR plus Southern blot, ddPCR, qPCR | Targeted DNA regions | Triplet repeat expansions, CNVs, SNVs, gene expression |
| | MLPA | Targeted genes | Copy number variants |
| | Sanger sequencing | Single gene sequence | Nucleotide sequence |
| GENOMIC TESTING | Mitochondrial genome sequencing | Mitochondrial DNA | Single nucleotide variants, deletions, duplications |
| | NGS - Multigene panel | A targeted group of genes associated with a condition | Small nucleotide variants, gene fusions |
| | NGS – Exome Sequencing | Protein-coding regions of genes (exons only) | Single nucleotide variants, small deletions, insertions, duplications |
| | NGS - Whole Genome Sequencing | All genes (introns and exons), regulatory regions, mitochondrial DNA, non-coding DNA, non-protein coding genes | Single nucleotide variants, small deletions, duplications |
| OTHER | RNA sequencing | Sequencing of mRNA | Fusion transcripts (e.g. in cancer); Gene expression |
| | Methylation testing; MS-MLPA (methylation-specific MLPA) | Targeted genes | Presence of methyl groups on genes (epigenetic modification) |
| | Microsatellite instability (MSI) testing | DNA, short repetitive regions | Changes in microsatellite length |

*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 variants in only one known 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') is used for a single gene sequence. PCR-based methods are used to identify common variants.

## Using multigene panel test

For a monogenic condition known to be caused by one of several genes, that is, a genetically heterogeneous condition such as dilated cardiomyopathy (>30 genes) or non-syndromic hearing loss (>90 genes), a targeted multigene panel test can be performed (Figure 26). Next Generation Sequencing (NGS) technology (see 'Sequencing Technology') 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 sequence data is limited to a small group of genes. If a pathogenic variant is not found, a new sample and further sequencing of more genes is required.

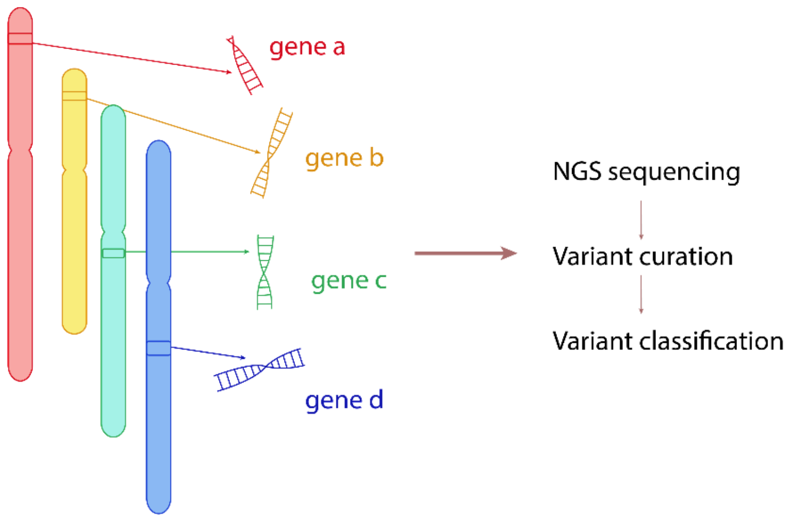


*Figure 26 NGS can target a defined set of genes known to be associated with a genetic condition, to identify which one of these genes is responsible in an individual patient.*

Genomic testing in cancer generally uses targeted multigene panels on the tumour tissue, such as a Comprehensive Cancer Panel, or small targeted panels designed for different types of cancer. Cancer panels may target specific exons with known cancer-related variants or mutation 'hotspots. Testing for bone marrow failure syndromes and for blood cancers employs panels for both inherited variants and acquired variants, as they can involve one or both types of variant.

## Mitochondrial genome sequencing

Next generation sequencing platforms are used for sequencing 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 exome sequencing (ES/WES[6]) and whole genome sequencing (WGS)

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 exome or whole 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 exome and whole genome sequencing are listed below.

**ES/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 intron sequence near the exon-intron junctions)

**ES/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

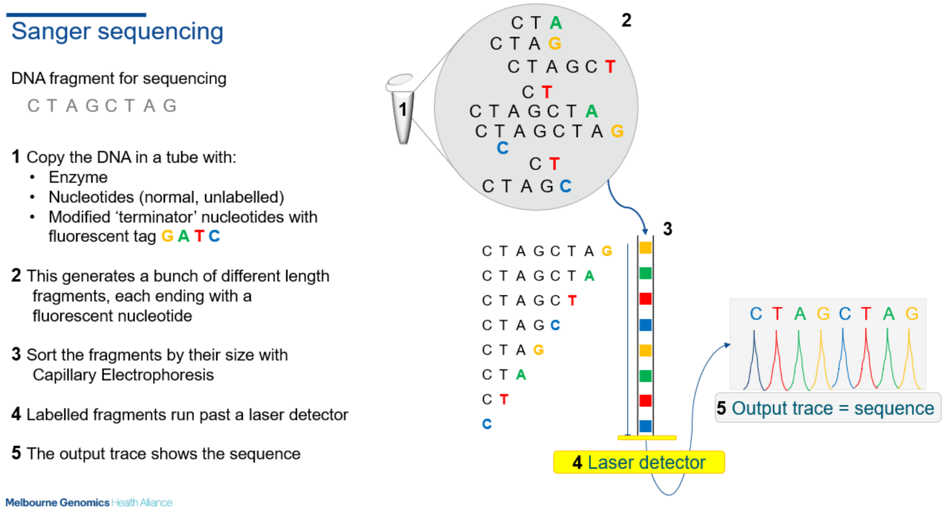
[6] * Terminology - exome sequencing (ES) and whole exome sequencing (WES) are both used

## 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), 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 27).



Sanger sequencing

DNA fragment for sequencing

C T A G C T A G

**1** Copy the DNA in a tube with:
- Enzyme
- Nucleotides (normal, unlabelled)
- Modified 'terminator' nucleotides with fluorescent tag G A T C

**2** This generates a bunch of different length fragments, each ending with a fluorescent nucleotide

**3** Sort the fragments by their size with Capillary Electrophoresis

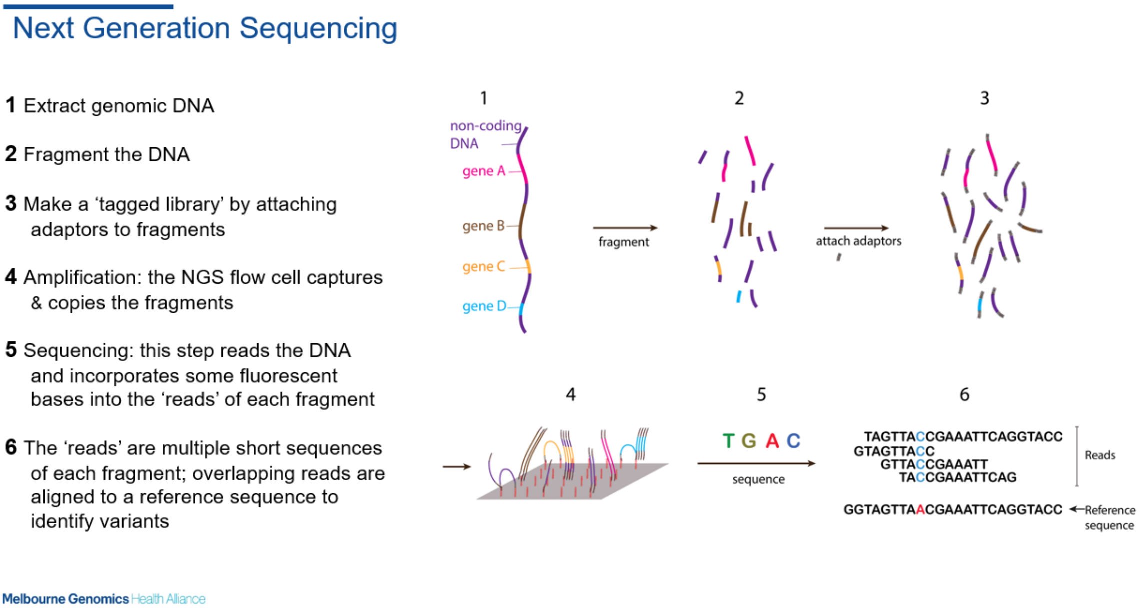
**4** Labelled fragments run past a laser detector

**5** The output trace shows the sequence

*Figure 27 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 28). This can mean sequencing the whole genome, whole exome or targeted gene panels.



Next Generation Sequencing

**1** Extract genomic DNA

**2** Fragment the DNA

**3** Make a 'tagged library' by attaching adaptors to fragments

**4** Amplification: the NGS flow cell captures & copies the fragments

**5** Sequencing: this step reads the DNA and incorporates some fluorescent bases into the 'reads' of each fragment

**6** The 'reads' are multiple short sequences of each fragment; overlapping reads are aligned to a reference sequence to identify variants

*Figure 28 Illustration of key steps in a Next Generation Sequencing method (based on the Illumina method) used in sequencing for multigene panels, mitochondrial genome, whole exome and whole genome sequencing.*

## Technical 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. The NGS method produces 'reads' (as described above) which then undergo computational reconstruction of the overlapping reads to determine the correct 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. Different NGS methods generate different read lengths and issues for analysis. 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 29). Gaps in coverage are difficult to interpret, for example, being uncertain whether a gap is poor sequencing/no reads (missing data) or a deletion variant.

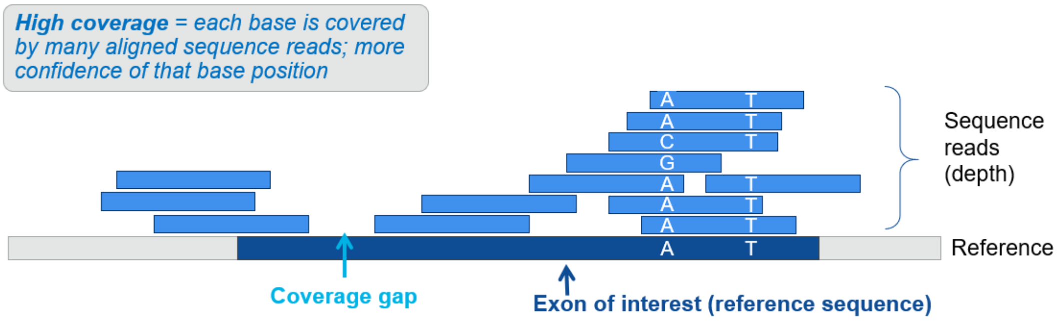
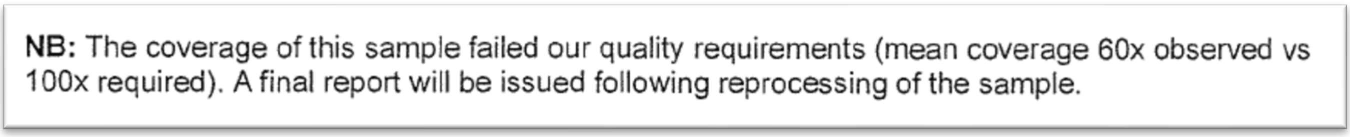

*High coverage* = each base is covered by many aligned sequence reads; more confidence of that base position
A
A
C
G
A
A
A
A
T
T
T
T
T
T
T
Sequence reads (depth)
Reference
Coverage gap
Exon of interest (reference sequence)

*Figure 29 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: Provides the best coverage for most genes
- NGS for multigene panel: Good coverage with fewer genes sequenced
- NGS for WES: To target coding regions, 'short read' sequencing technology (reads of ~150bp) is used. WES typically has more gaps and lower coverage than multigene panels, therefore some regions are harder to piece together and overall the sensitivity for detecting a variant may be lower than for panels
- 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. Methods using 'long read' technology leaves fewer gaps and may provide greater sensitivity for detecting some variant types.
- Advances in technology, including long reads and different detection methods, as well as developments in analysis algorithms, are improving coverage and detection of challenging variant types, such as repeat expansions.

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 30).



**NB:** The coverage of this sample failed our quality requirements (mean coverage 60x observed vs 100x required). A final report will be issued following reprocessing of the sample.

*Figure 30 Excerpt from an exome sequencing report noting low coverage and failed quality control requiring resequencing of the sample.*

## Low variant allele frequency

Variant allele frequency (VAF) is the percentage of sequence reads containing the variant. When VAF is low, such as in a tumour sample with only a small proportion of tumour cells in the tissue, reliable detection of the disease-causing variant is limited. Cancer samples generally need more sequencing reads for reliable detection of variants.

## 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
  - pseudogenes accumulate variants and it is challenging to distinguish the relevant variants in the functional gene and pseudogene sequences from NGS data
  - 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

## Genomic testing in germline conditions

Genomic testing for germline conditions is typically conducted on blood or saliva samples. Samples should be taken and delivered to the testing laboratory according to instructions to ensure the DNA is of high quality for DNA sequencing.

Let's say that you order a whole exome sequence (WES) test. 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. Detailed clinical and phenotypic information provided by the referring physician guides the analysis. 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, 'tumour-normal' testing. As described in Section 8, good sequence coverage is required for reliable identification of a sequence variant.

### Filters

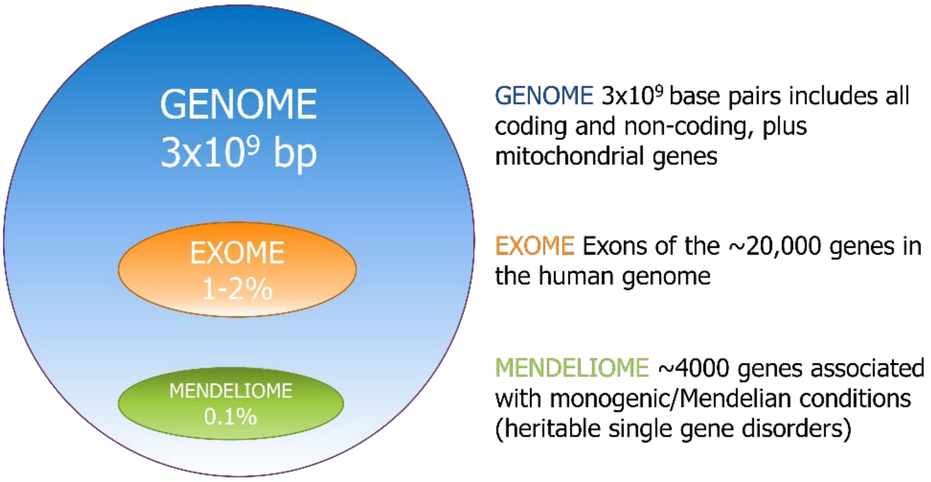
All genes have variants that contribute to the normal range of human variation. Clinical testing is interested only in the variants relevant to the patient's clinical phenotype, so the data is filtered to selectively 'ignore' polymorphisms (common SNPs) and variants in untranslated, intronic and other non-coding regions. Variants in these regions can alter gene expression, so they may be investigated further if non relevant coding variants are identified. Filtering also allows you to focus on variants according to the inheritance pattern in the family that suggest recessive, dominant, autosomal or sex-linked inheritance.

The '**incidentalome'** (incidental findings unrelated to the condition under investigation; see 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.*

## Gene Lists and PanelApp

Analysts then apply one or more lists of gene 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 of the ~4000 genes known to cause monogenic (Mendelian) conditions (Figure 31).



GENOME $3x10^9$ base pairs includes all coding and non-coding, plus mitochondrial genes

EXOME Exons of the ~20,000 genes in the human genome

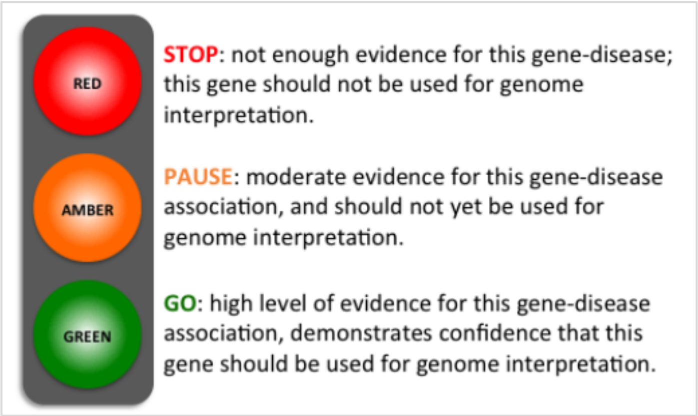
MENDELIOME ~4000 genes associated with monogenic/Mendelian conditions (heritable single gene disorders)

Melbourne Genomics Health Alliance

*Figure 31 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, the sets of genes with a known association to a phenotype of disorder, vary in the number of genes and are periodically updated based on new information.

PanelApp Australia (an instance of PanelApp developed by Genomics England) is an open access source repository of curated gene lists/panels for genomic testing. Genes are rated on a traffic light system (Figure 32). Several examples are listed below.



STOP: not enough evidence for this gene-disease; this gene should not be used for genome interpretation.

PAUSE: moderate evidence for this gene-disease association, and should not yet be used for genome interpretation.

GO: high level of evidence for this gene-disease association, demonstrates confidence that this gene should be used for genome interpretation.

*Figure 32 Traffic light rating systems for curated genes in PanelApp*

*Gene lists - examples from PanelApp*

- Short QT syndrome – 3 genes
- Cardiomyopathy_Adult_SuperPanel (combines Dilated, Hypertrophic and Arrhythmogenic cardiomyopathy panels) – 58 genes
- Early onset Dementia - 58 genes
- Cancer predisposition_paediatric – 92 genes
- Immunological disorders_SuperPanel (combines 16 panels) – 523 genes
- Intellectual Disability syndromic and non-syndromic – 1514 genes

*Examples from PanelApp Australia, January 2023 (green-rated genes)*

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. The variants are then 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**, which involves gathering and assessing evidence about the known or predicted effects and phenotypic 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 gene alteration 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 33).

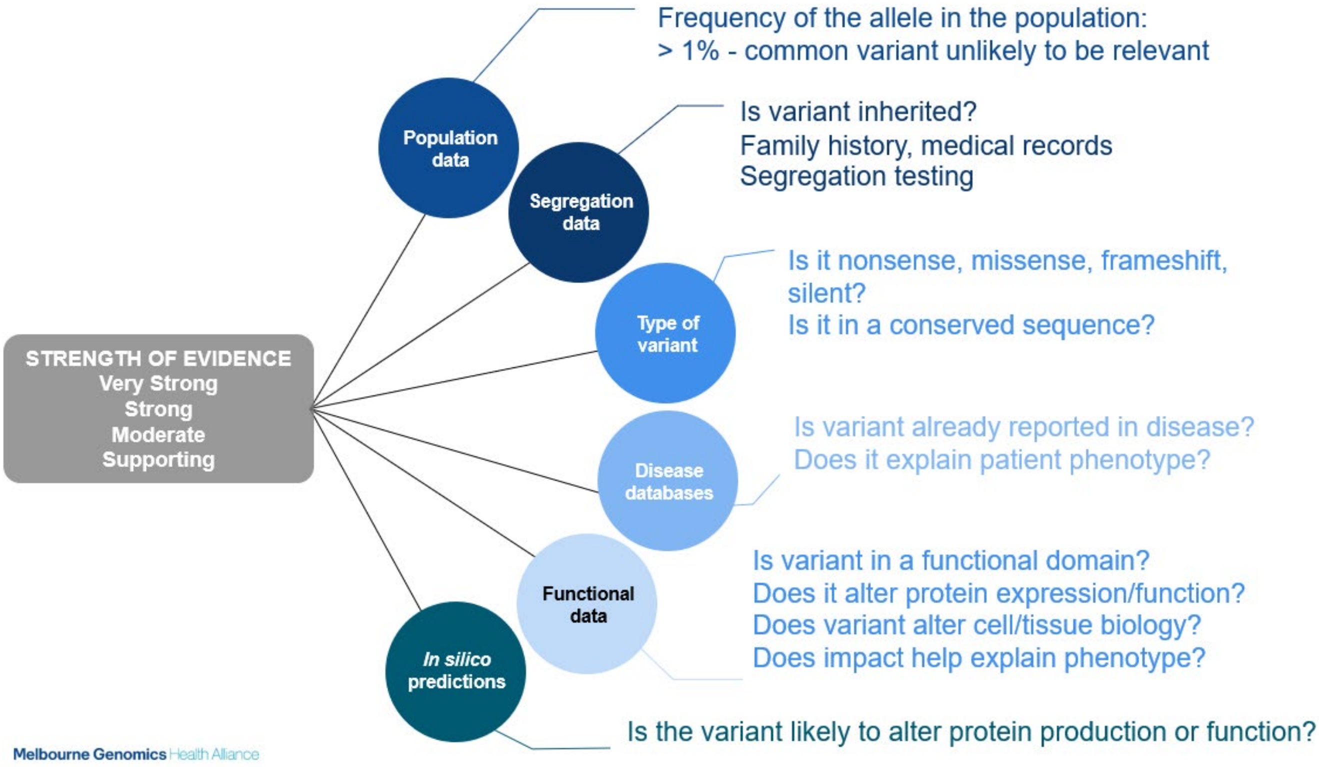

Frequency of the allele in the population:
> 1% - common variant unlikely to be relevant
Population data
Segregation data
Is variant inherited?
Family history, medical records
Segregation testing
Type of variant
Is it nonsense, missense, frameshift, silent?
Is it in a conserved sequence?
STRENGTH OF EVIDENCE
Very Strong
Strong
Moderate
Supporting
Disease databases
Is variant already reported in disease?
Does it explain patient phenotype?
Functional data
Is variant in a functional domain?
Does it alter protein expression/function?
Does variant alter cell/tissue biology?
Does impact help explain phenotype?
In silico predictions
Is the variant likely to alter protein production or function?
Melbourne Genomics Health Alliance

*Figure 33 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.*

More on the evidence required for variant interpretation is summarised in Table 3. This information is included in genomic test reports, for reporting requirements. It can be useful for non-genetic physicians to understand more about the data required to help understand the reports, participate in MDT meetings. or seeking more information from online sources such as OMIM, ClinGen, ClinVar or OncoKB.

*Table 3 Types of Evidence collected during variant curation.*

| Information required | Key aspects and terminology |
|---|---|
| **Reference sequences** | Identify variants compared to a reference. |
| **Zygosity and Inheritance pattern** | Homozygous, heterozygous, hemizygous, compound heterozygous.<br>Sex-linked, X-linked, *De novo*; Dominant, Recessive; Penetrance, Expressivity.<br>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. |
| **Type of variant** | DNA level: Substitution, insertion or deletion (indel), deletion & insertion together (delins)<br>Protein level: Synonymous, Missense, Nonsense, Frameshift |
| **Gene-disease association**<br>**Variant-disease association** | Genes and variants present in disease databases due to previous association with a condition are significant.<br>ACMG guidelines: a variant classified as pathogenic in one condition must be subsequently classified as pathogenic in the same condition in another individual. |
| **Population frequency** | Variants present at high frequency in population databases (e.g. >1%) are considered common polymorphisms (normal variation)<br>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. |
| **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.<br>Variants in regions of high conservation may have a greater effect on the protein and more potential for pathogenicity. Variants in functional domains, which are often conserved, can have significant impact. |
| ***In silico* predictions** | Molecular biology, bioinformatics, and computational tools to predict the effects of a variant on gene expression and protein function.<br>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. |

*See appendices for databases and online information sources used in variant curation: Appendix 2 Germline; Appendix 3 Somatic.*

## Classify variants

The strength of the evidence is weighed up (very strong, strong, moderate, or supporting) to arrive at an overall classification. Classification follows the American College of Genetics and Genomics (ACMG) Guidelines for reporting variants as pathogenic, likely pathogenic, uncertain significance, likely benign, benign (Class 5 to Class 1, respectively). Some laboratories also subclassify the variants of uncertain significance (VUS) as Class 3a (potentially pathogenic), 3b and 3c (potentially benign) (Figure 34). The five classes correspond to the probability that the variant is a cause of the condition or phenotype, e.g. the threshhold for pathogenic is 99%.[7]

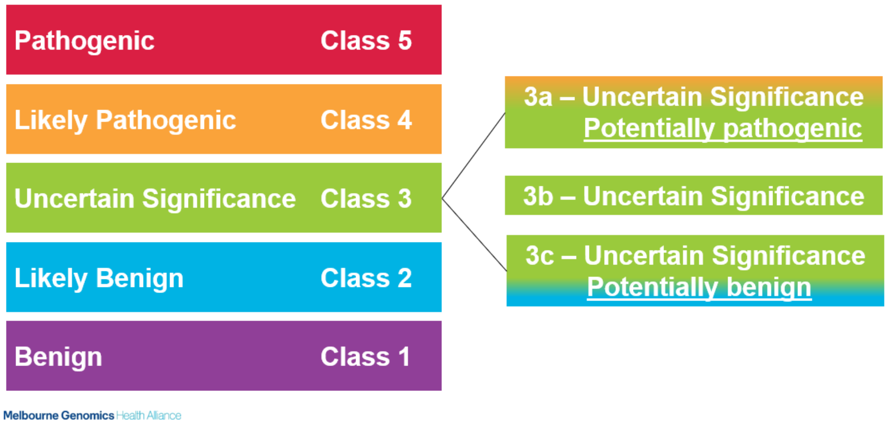

Pathogenic Class 5
Likely Pathogenic Class 4
Uncertain Significance Class 3
Likely Benign Class 2
Benign Class 1
3a – Uncertain Significance Potentially pathogenic
3b – Uncertain Significance
3c – Uncertain Significance Potentially benign
Melbourne Genomics Health Alliance

*Figure 34 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 based on the reported variant interpretation*

| Type of Variant | Class | Implications for patients |
|---|---|---|
| **Pathogenic** | 5 | Cause for condition identified<br>Can be used to direct management<br>Can be used for family planning<br>Can be used for predictive testing |
| **Likely pathogenic** | 4 | Cause for condition likely been identified<br>May be used to direct management<br>May be used for family planning<br>May be used for predictive testing in other family members |
| **Uncertain significance*** | 3 | Cause for condition still unclear<br>Cannot be used to direct management<br>Cannot be used for family planning<br>Cannot be used for predictive testing in other family members<br>*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 |

[7] Tavtigian et al, Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. Genet Med. 2018 Sep;20(9):1054-1060. doi: 10.1038/gim.2017.210. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6336098/

# 10 Analysing cancer variants

We saw earlier (see Figure 4) that variants arising in mature somatic cells can cause cancer. These somatic variants, including SNVs, CNVs, gene fusions and chromosomal rearrangements, are not heritable. This contrasts with heritable germline variants associated with cancer susceptibility syndromes.

Cancers are typically initiated by a single variant (an acquired or inherited cancer susceptibility gene) plus a 'second hit' [8] (another alteration in the other allele of the affected gene). Tumours also accumulate more mutations over time. Variant analysis of solid tumours identifies both somatic and germline variants relevant to the cancer. Blood cancers can occur by accumulation of both inherited and acquired variants, so testing of haematological cancers involves both somatic and germline variant analysis.

Solid tumour variant analysis must account for the proportion of cancerous cells in a tumour sample, the total load of variants acquired as a tumour evolves - the tumour mutational burden (TMB) - and 'mutational signatures' - characteristic combinations of mutation types arising due to errors in DNA replication and repair, or the action of genotoxins. Genetic profiles reflect the evolution of the tumour and can potentially provide predictive biomarkers for therapy.

## Genomic testing in cancer

- **Tissue quality:** Appropriate tumour tissue sampling and preparation is essential to obtain high quality DNA required for NGS. Tumour purity, assessed by histopathology, is required for variant analysis and interpreting the variant allele frequency (VAF). Solid tumour testing may involve tumour-normal analysis, i.e. sequencing of both the tumour and a blood sample (non-cancer cells) to compare the cancer cell genome to the patient's own germline genome.
- **Variants detected:** Cancer genomic test reports may include somatic SNVs, CNVs, gene fusions, germline variants. The frequency of the variant allele is an important factor in identifying somatic versus germline variants. Genomic sequencing can also determine mutational signatures.
- **Genomic analysis**: NGS for cancer genome testing typically uses multigene panels established for known cancer variants and types of cancer, including comprehensive cancer panels and small targeted panels. Other tests such as MLPA detect CNVs in targeted cancer-related genes. Fusions are detected by FISH and RNA sequencing (the latter is currently available only a research setting). Sequencing for blood cancers employs panels for both acquired and inherited variants.

Genomic analysis of solid tumours investigates the variant(s) in the tumour cells, and for some cancers additional genomic alterations such as microsatellite instability (MSI), plus investigation of the therapeutic implications of the identified variants, such as drug resistance or sensitivity.

Curation of cancer variants follows a similar path to that of germline variant interpretation (Chapter 9), with additional classification related to the **diagnostic, therapeutic,** and **prognostic** outcomes.

## Classifying cancer variants

Information sources used in cancer variant curation and classification include:

- Cancer specific tools and databases, e.g. The Cancer Genome Atlas (TCGA), OncoKB, cBioPortal, NIH National Cancer Institute GDC Data Portal, Jackson Lab JAX-CKB, CIViC and COSMIC, St Jude's PECAN (paediatric cancer) (see Appendix 3).

[8] The 'two hit' hypothesis: see https://www.nature.com/scitable/topicpage/tumor-suppressor-ts-genes-and-the-two-887/; Chernoff 2021 The two-hit theory hits 50, Mol Cell Biol v.32(22) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8694077/

- Classification guidelines for additional classification based on clinical utility have been developed by the Association for Molecular Pathology (AMP)[9] and European Society for Medical Oncology (ESMO). Evidence is grouped into tiers to guide evidence-based clinical action (the AMP guidelines are summarised in Table 5). Guidelines are updated as new evidence becomes available.

*Table 5 Summary of AMP Classification tiers for cancer variants*[7]

| 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.<br>No clear evidence of cancer association | Variant at significant frequency in general population and databases; no published evidence of cancer association |

[9] 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

# 11 Genomic test reports

Reports from different laboratories vary in style, layout and amount of variant curation detail. Some laboratories sub-classify VUSs, some report all VUSs thus requiring the clinician to make a judgement about their value, further investigation or actionability. Some labs list all the genes in the panels used, and so on. Here we summarise the main elements generally covered in a genomic test report, with examples.

## Elements of a genomic test report

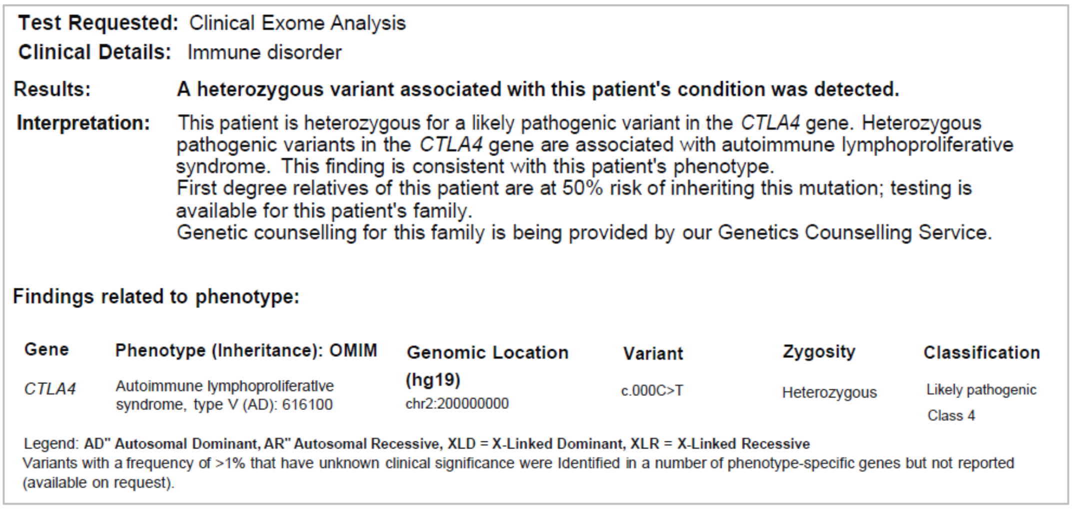
- Test requested
- Reason for referral
- Test type, panels and gene lists used
- Result summary
- Interpretation of the findings
- Findings related to phenotype: e.g. a table listing gene(s), variant(s), zygosity, classification, inheritance pattern
- Variant description: detailed description of curation results, with reference to data sources and publications
- Methods: laboratory & technical details; includes prioritised gene lists for exome/genome analysis

*For a brief guide to variant nomenclature used in genomic test reports, see page 43.*

## Germline genomic test report - example

### A clinical exome for a young adult with an immunological condition

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 altered for deidentification).



**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<br>Class 4 |

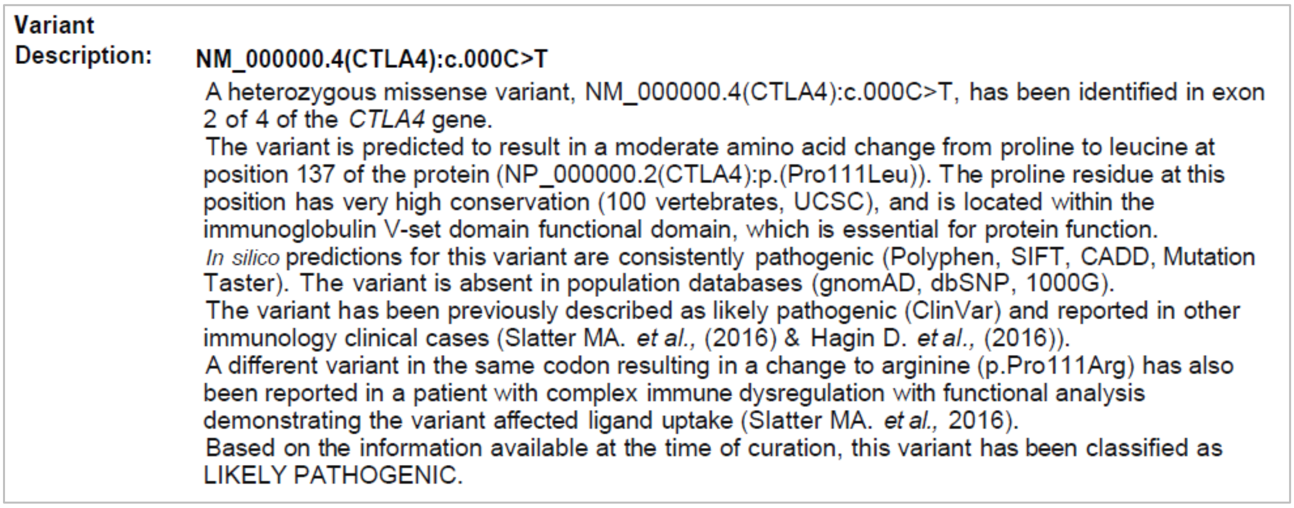
Legend: AD" **Autosomal Dominant, AR" Autosomal Recessive, XLD = X-Linked Dominant, XLR = X-Linked Recessive**
Variants with a frequency of >1% that have unknown clinical significance were Identified in a number of phenotype-specific genes but not reported (available on request).

*Figure 35 Exome test report – result summary and interpretation (genomic coordinates are altered)*

This report includes a summary of the results, including zygosity (heterozygous, the patient has one copy of the variant) and whether a variant associated with, or causing, the condition is found (Figure 35). The interpretation paragraph includes the name of the gene (*CTLA4*), the variant classification (likely pathogenic), previous association of this gene and/or the specific variant with the patient's condition (associated with autoimmune lymphoproliferative disorders), or other disorders. If parents 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 was done so inheritance is unknown (which is typically stated on the report).

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.000C>T') 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.

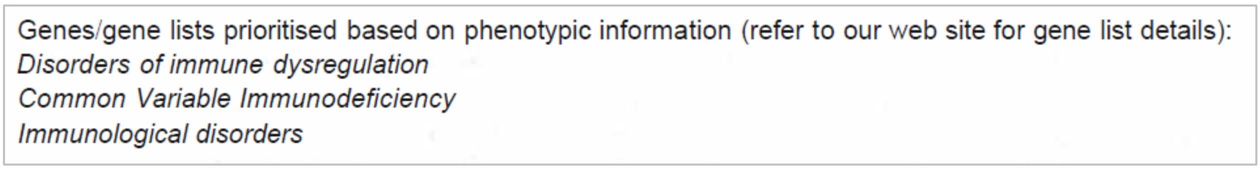


**Variant Description:** **NM_000000.4(CTLA4):c.000C>T**

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 36  Exome test report – variant description*

The variant description 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.



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 37  Exome test report - gene lists*

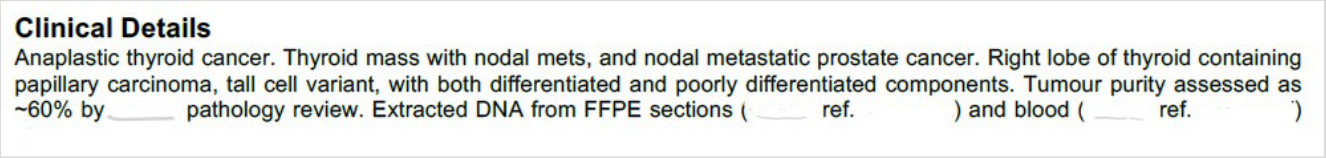
Genomic reports also provide technical information about limitations of the tests, coverage of the sequence, information not reported, and the gene lists (Figure 37) used in analysis. In this case, gene lists covering a wide range of immunological disorders. Different gene lists may cover some of the same genes, but selected because some gene lists might have been more recently updated.

The amount and depth of technical information in reports differs with the lab. Some labs report several VUSs, not all laboratories sub-classify VUSs, and they may not offer recommendations for clinical action. Contact your local friendly clinical geneticist for further guidance with genomic testing.

## Cancer genomic test report - example

### A comprehensive cancer panel test for adult thyroid cancer

View the following extracts from a comprehensive cancer panel report for a man with anaplastic thyroid cancer. The report summarises the clinical details of the patient's cancer and tumour purity (Figure 38); the quality of a cancer sample in terms of DNA preservation and proportion of tumour cells is important for genomic test analysis.

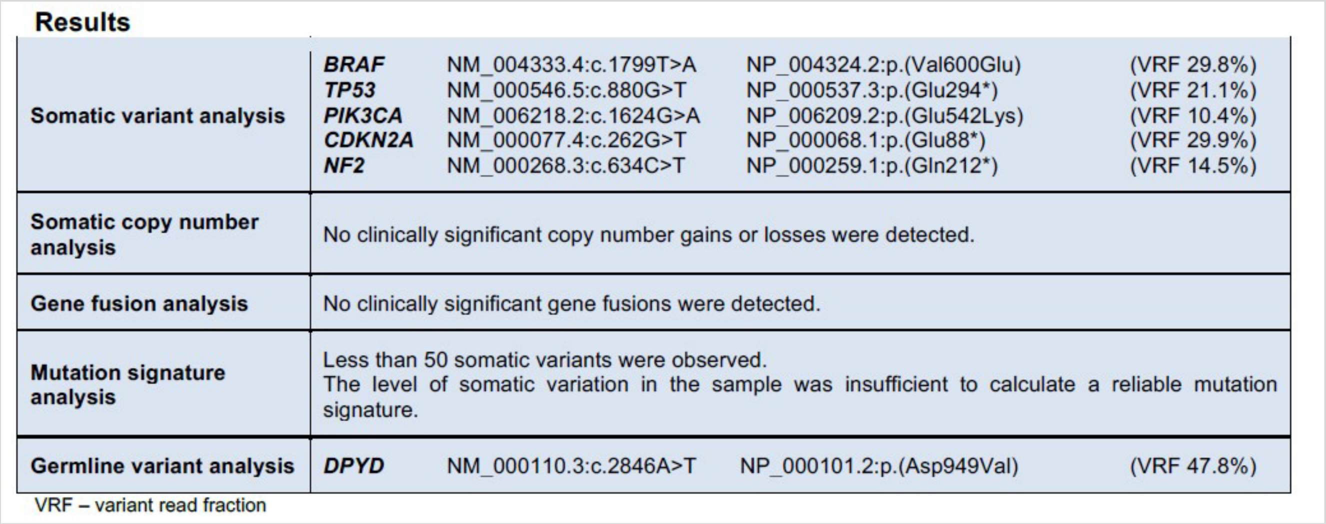


**Clinical Details**

Anaplastic thyroid cancer. Thyroid mass with nodal mets, and nodal metastatic prostate cancer. Right lobe of thyroid containing papillary carcinoma, tall cell variant, with both differentiated and poorly differentiated components. Tumour purity assessed as ~60% by______ pathology review. Extracted DNA from FFPE sections (______ ref. ) and blood ( ______ ref. )

*Figure 38 Cancer report: Clinical details*

The variants relevant to the cancer are listed, with five somatic and one germline variants (SNVs) detected in this sample, and no fusions or CNVs (Figure 39). Elsewhere in the report several somatic VUSs are listed.



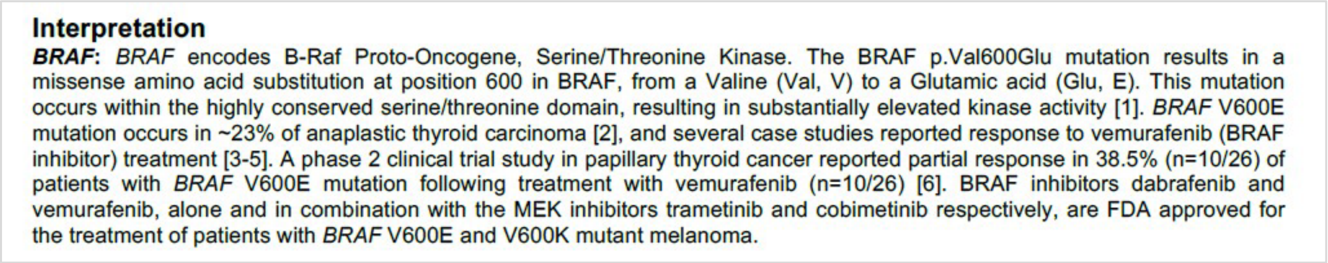
**Results**

| | | | | |
|---|---|---|---|---|
| **Somatic variant analysis** | ***BRAF***<br>***TP53***<br>***PIK3CA***<br>***CDKN2A***<br>***NF2*** | NM_004333.4:c.1799T>A<br>NM_000546.5:c.880G>T<br>NM_006218.2:c.1624G>A<br>NM_000077.4:c.262G>T<br>NM_000268.3:c.634C>T | NP_004324.2:p.(Val600Glu)<br>NP_000537.3:p.(Glu294*)<br>NP_006209.2:p.(Glu542Lys)<br>NP_000068.1:p.(Glu88*)<br>NP_000259.1:p.(Gln212*) | (VRF 29.8%)<br>(VRF 21.1%)<br>(VRF 10.4%)<br>(VRF 29.9%)<br>(VRF 14.5%) |
| **Somatic copy number analysis** | No clinically significant copy number gains or losses were detected. | | | |
| **Gene fusion analysis** | No clinically significant gene fusions were detected. | | | |
| **Mutation signature analysis** | Less than 50 somatic variants were observed.<br>The level of somatic variation in the sample was insufficient to calculate a reliable mutation signature. | | | |
| **Germline variant analysis** | ***DPYD*** | NM_000110.3:c.2846A>T | NP_000101.2:p.(Asp949Val) | (VRF 47.8%) |

VRF – variant read fraction

*Figure 39 Cancer report: Result – variants identified.*

Each of the clinically relevant variants is described, with the example of BRAF shown (Figure 40). The VUSs are not described in detail.

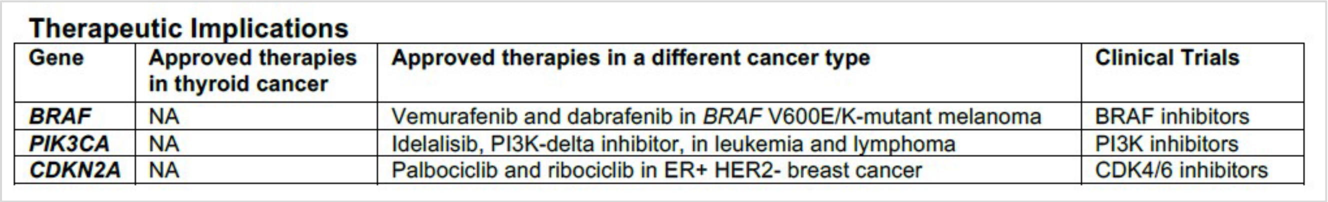


**Interpretation**

***BRAF***: *BRAF* encodes B-Raf Proto-Oncogene, Serine/Threonine Kinase. The BRAF p.Val600Glu mutation results in a missense amino acid substitution at position 600 in BRAF, from a Valine (Val, V) to a Glutamic acid (Glu, E). This mutation occurs within the highly conserved serine/threonine domain, resulting in substantially elevated kinase activity [1]. *BRAF* V600E mutation occurs in ~23% of anaplastic thyroid carcinoma [2], and several case studies reported response to vemurafenib (BRAF inhibitor) treatment [3-5]. A phase 2 clinical trial study in papillary thyroid cancer reported partial response in 38.5% (n=10/26) of patients with *BRAF* V600E mutation following treatment with vemurafenib (n=10/26) [6]. BRAF inhibitors dabrafenib and vemurafenib, alone and in combination with the MEK inhibitors trametinib and cobimetinib respectively, are FDA approved for the treatment of patients with *BRAF* V600E and V600K mutant melanoma.

*Figure 40 Cancer report: Variant interpretation/description of information used in the classification.*

Somatic variants with possible implications for access to approved therapies. In this case, at the time of analysis and reporting, there were no gene/variant specific therapies approved for thyroid cancer, but off-label therapeutics were possible (Figure 41).

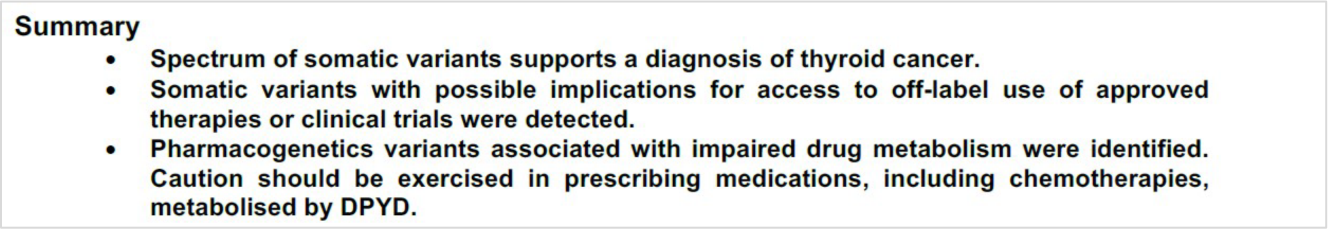


**Therapeutic Implications**

| Gene | Approved therapies in thyroid cancer | Approved therapies in a different cancer type | Clinical Trials |
|---|---|---|---|
| ***BRAF*** | NA | Vemurafenib and dabrafenib in *BRAF* V600E/K-mutant melanoma | BRAF inhibitors |
| ***PIK3CA*** | NA | Idelalisib, PI3K-delta inhibitor, in leukemia and lymphoma | PI3K inhibitors |
| ***CDKN2A*** | NA | Palbociclib and ribociclib in ER+ HER2- breast cancer | CDK4/6 inhibitors |

*Figure 41 Cancer report: Therapeutic implications of the results; (potentially) actionable variants*

The germline variant identified in this tumour was associated with impaired metabolism of cancer therapeutics. The overall summary (Figure 42):



**Summary**

- **Spectrum of somatic variants supports a diagnosis of thyroid cancer.**
- **Somatic variants with possible implications for access to off-label use of approved therapies or clinical trials were detected.**
- **Pharmacogenetics variants associated with impaired drug metabolism were identified. Caution should be exercised in prescribing medications, including chemotherapies, metabolised by DPYD.**

*Figure 42 Cancer report: Summary*

## Nomenclature [10]

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___'<br>• The accession number for the coding/mRNA sequence is recorded as 'NM___' | **Examples *CFTR* [11] gene**<br>• e.g. NG_016465.4<br>• e.g. NM_000492.3 |
| **Naming the location and type of variant:** | • The genomic DNA sequence is recorded as 'g.'<br>• The coding DNA (or mRNA) sequence is recorded as 'c.'<br>• The protein sequence is recorded as 'p.'<br>• The mitochondrial DNA sequence is recorded as 'm.' | **Example: *CFTR* most common deletion**<br>• g.98809_98811delCTT<br>• c.1521_1523delCTT<br>• 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 |

[10] **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/

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

# Appendix 1 Glossary

| 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.<br>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. |
| Chimera; chimeric | An individual having cells with genetically distinct cells |
| 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<br>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;<br>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. |
| Coverage | In DNA sequencing: the proportion of DNA 'read' during sequencing (breadth) and the number of reads per region of DNA (depth) |
| 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 | Deletions and insertions in close proximity on a DNA strand that produces a new variant. |
| DNA | Deoxyribonuleic acid. Genetic material of life on earth. Built from 4 deoxyribonucleotides– 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 | |
| Epigenetics | Heritable DNA modification that alters gene expression without changing the DNA sequence or genetic code. Commonly methyl-groups (methylation) and acetyl-groups (acetylation) 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 | |
| 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 | |
| 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. |
| 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 | |
| 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. |
| Heterogeneity | Genetic heterogeneity: When a germline condition can be caused by one of several different genes; In cancer, genetic heterogeneity is the variation in range of mutations found in the cells in a tumour sample.<br>Cellular heterogeneity: In cancer, the variation in cell types (normal vs tumour cells) within a tumour. |
| 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. |
| Histone | Protein complex in the cell nucleus around which long strands of chromosomal DNA coil. Provides structural support for chromosomes. |
| Homology | In relation to 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. |
| HRD - Homologous recombination deficiency | Homologous recombination is a cellular mechanism to repair breaks and crosslinks in DNA. Mutations in genes in this pathway can lead to deficiency in the mechanism and predispose to cancer. |
| I-J | |
| *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 insertion or deletion. Collective term for insertions and deletions. |
| 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 | |
| Karyotype | Arrangement of chromosomes showing the number and structure of the set of chromosomes in a species or individual. |
| Liquid biopsy | A blood sample used for detecting circulation biomarkers, such as cell free DNA (cfDNA) or circulating tumour DNA (ctDNA). |
| LOH - Loss of heterozygosity | Loss of allelic variation in regions of the genome (consequently, regions of homozygosity). In germline conditions, can result from uniparental disomy. In cancer, a common event in development of cancer and often seen in tumour cells. |

| 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* |
| Microsatellite | Regions of the genome containing repeated sequences of nucleotides; 1-6 bases typically repeated up to 50 times. Also called short tandem repeats (STRs).<br>Also see MSI – microsatellite instability. |
| Missense | Genetic variant (nucleotide substitution) causing a change in an amino acid in the resulting protein. Also called non-synonymous. |
| MMR - Mismatch repair | Mismatch repair is a cellular mechanism to repair incorrect bases inserted during DNA replication. Some cancers are caused by mismatch repair deficiency (MMRD). |
| Monogenic | Condition or phenotype caused by a variant in one gene |
| Mosaic; mosaicism | Variant: A variant present only in some cells of the individual.<br>Germline or gonadal mosaicism: some eggs or sperm carry a variant that is not present in the other body cells/tissues. |
| 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. |
| MSI – Microsatellite instability | A change in the number of repeats within microsatellites (regions of the genome with short repeated sequences); can be caused by impaired mismatch repair (MMR). |
| 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. |
| N-O | |
| 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.<br>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). |
| Oncogene | A proto-oncogene activated by mutation, causing abnormal cell growth. |
| 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 | |
| Panel | see ‘multigene panel test’ |
| PanelApp | Publicly available knowledgebase and source of gene panels for genomic sequence analysis. See PanelApp Australia https://panelapp.agha.umccr.org/ |
| 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. |
| Proto-oncogene | A gene involved in cell growth that can cause normal cells to become cancer cells when activated by mutation. |
| Pseudogene | An inactive version of a gene; originating as a functional protein-coding gene but altered by mutations through evolution. |
| 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 | Ribonucleic acid. Composed of 4 ribonucleotide bases: adenine (A), cytosine (C) guanine (G) and uracil (U). Different types have different roles in cells: messenger (mRNA), transfer (tRNA), ribosomal (rRNA), long non-coding RNA (lncRNA) and microRNAs. |
| RNA processing | Modification of the primary transcript, including splicing, addition of 5’CAP and 3’ poly-A tail to produce mature mRNA. |
| 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. |
| Signature (mutational) | Tumour mutational signatures are characteristic patterns of DNA alteration or base changes for a given mutation aetiology. |
| 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 | |
| TMB - Tumour mutational burden | The number of different mutations in a tumour cell. |
| 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. |
| TSG - Tumour suppressor gene | A gene that controls cell growth. Mutations that cause loss of function can cause uncontrolled cell growth and cancer. |
| Tumour purity | The proportion of cancerous cells in a tumour/tumour biopsy. |
| 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 | |
| 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. |

# Appendix 2 Resources and databases – Germline Genomics

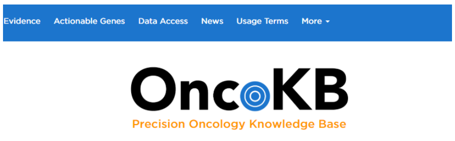
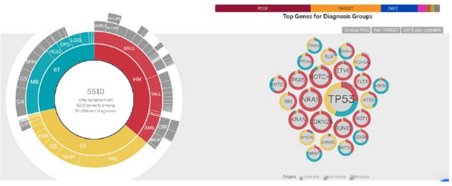
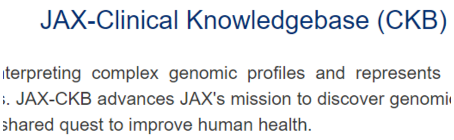
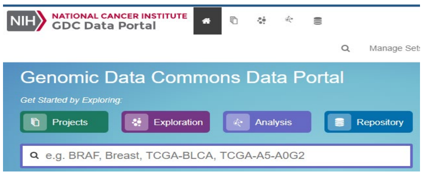
These web resources can be useful for clinicians to access information on genetic conditions and understand the resources used for variant classification.

Some of these sites are referenced in genomic test reports. Most are open access but many require some experience to navigate effectively.

| Resource, Source, URL, | Type of information |
|---|---|
| 1 Genetic Conditions and general genetic information | |
| **MedlinePlus**<br>National Library of Medicine (NLM) USA<br>https://medlineplus.gov/genetics/ | Genetic conditions. Search by Condition, Gene or Chromosome. Links to readable explanations on human genetics. |
| **Gene Reviews**<br>https://www.ncbi.nlm.nih.gov/books/NBK1116/ | International resource for clinicians, provides clinically relevant and medically actionable information for many inherited conditions. |
| **EviQ**<br>Cancer Institute NSW<br>https://www.eviq.org.au/ | Information on inherited cancer susceptibility. |
| **Unique**<br>Rare Chromosome Disorder Support Group, UK<br>https://www.rarechromo.org/disorder-guides | Guides and databases of chromosomal disorders. Genotype database has information/ links to cytogenetics, array or sequencing data/results. |
| 2 Databases of human genomic variants and tools for variant interpretation | |
| Including relationships between variants and phenotypes, common variants – polymorphisms (SNPs) and disease-related variants. | |
| **ClinVar**<br>NCBI<br>www.ncbi.nlm.nih.gov/clinvar/ | ClinVar aggregates information about genomic variation and its relationship to human health. |

| | | |
|---|---|---|
| **OMIM**<br>https://www.omim.org/ | OMIM<br>50 YEARS OMIM<br>Human Genetics Knowledge for the World | An Online Catalog of Human Genes and Genetic Disorders |
| **UCSC** Genomics Institute<br>http://genome.ucsc.edu | UCSC Genome Browser | A Genome Browser. A web-based tool, displays portion of a genome at any scale, accompanied by aligned annotation "tracks". |
| **gnomAD**<br>Broad Institute<br>http://gnomad.broadinstitute.org | gnomAD<br>genome aggregation database | The Genome Aggregation Database (gnomAD). A population database of exome & genome sequences from around the world. |
| **DECIPHER** (Protein)<br>Wellcome Sanger Institute<br>https://www.deciphergenomics.org/about/overview | DECIPHER<br>GRCh38 | An interactive web-based database which incorporates a suite of tools designed to aid the interpretation of genomic variants. |

# Appendix 3 Resources and databases – Somatic genomics

| Database URL, Authors or Institute | | Purpose |
|---|---|---|
| **OncoKB**<br>http://oncokb.org/#/<br>Memorial Sloane Kettering Cancer Centre | <br>Evidence Actionable Genes Data Access News Usage Terms More<br>OncoKB<br>Precision Oncology Knowledge Base | Databases of variants and the related biological effect, prevalence, prognostic information and treatment implications. Gene curation |
| **St Jude PeCan data portal**<br>https://pecan.stjude.cloud/<br>St. Jude Children's Research Hospital | <br>Top Genes for Diagnosis Groups | PeCan - interactive visualisations of **paediatric cancer mutations** from St. Jude Children's Research Hospital and collaborators projects |
| **JAX-CKB**<br>https://ckb.jax.org/<br>The Jackson Laboratory, JAX Genomic Medicine, USA | <br>JAX-Clinical Knowledgebase (CKB)<br>ıterpreting complex genomic profiles and represents<br>;. JAX-CKB advances JAX's mission to discover genomi<br>shared quest to improve human health. | A dynamic digital resource for interpreting complex cancer genomic profiles in the context of protein impact, therapies, and clinical trials. |
| **COSMIC**<br>https://cancer.sanger.ac.uk/cosmic<br>Wellcome Sanger Institute, UK | <br>COSMIC<br>Catalogue Of Somatic Mutations In Cancer<br>Projects ▼ Data ▼ Tools ▼ News ▼ | Catalogue of somatic mutations in cancer |
| **CIViC**<br>https://civic.genome.wustl.edu/home<br>NIH-NCI, USA<br>Washington University, St Louis, School of Medicine | <br>CIViC<br>Clinical Interpretations of Variants in Cancer | Clinical Interpretation of variants in cancer. Discussion forum for interpretation of peer-reviewed publications on clinical relevance of variants or biomarker alterations. |
| **cBioportal**<br>http://www.cbioportal.org/index.do | <br>cBioPortal<br>FOR CANCER GENOMICS | Tools to visualise and analyse cancer genomic data sets |
| **GDC portal**<br>https://portal.gdc.cancer.gov/<br>Genomic Data Commons National Cancer Institute (NIH, USA). | <br>NIH NATIONAL CANCER INSTITUTE GDC Data Portal<br>Manage Sets<br>Genomic Data Commons Data Portal<br>Get Started by Exploring:<br>Projects Exploration Analysis Repository<br>e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-A0G2 | **Genomic Data Commons** data portal – projects; exploration; analysis; repository |
| **My cancer genome**<br>https://www.mycancergenome.org/<br>Vanderbilt-Ingram Cancer Center | <br>My Cancer Genome ®<br>Genetically Informed Cancer Medicine | Clinical impact of molecular biomarkers in cancer-related genes, proteins, biomarkers and anticancer therapies. |

**Melbourne Genomics**
Health Alliance

**melbournegenomics.org.au**

C/O WEHI
1G Royal Parade, Parkville, VIC 3052
Tel: +61 3 9936 6499 enquiries@melbournegenomics.org.au
**Email: education@melbournegenomics.org.au**

Alliance members

Supported by