Best-practice guidelines for reporting diagnostic NGS variants

Fran Smith

Clinical Scientist

Molecular Pathology, King's College Hospital

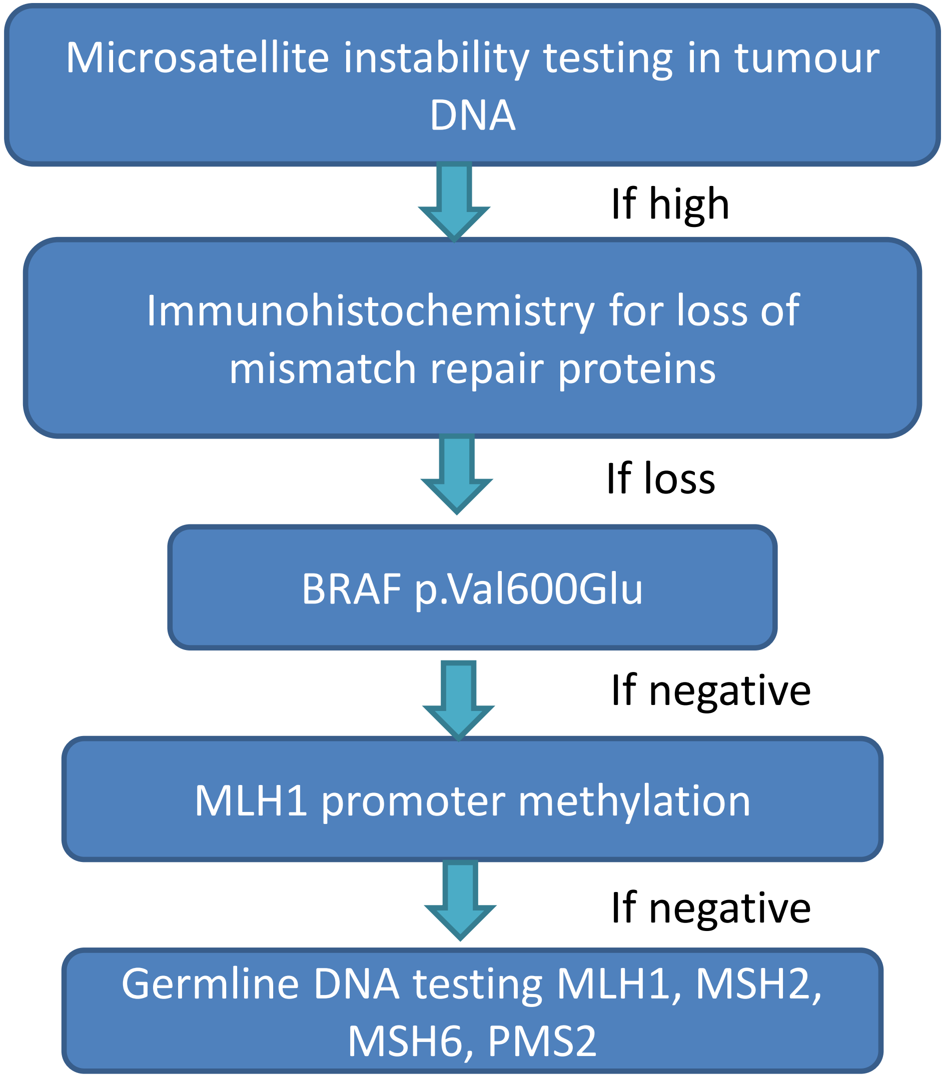
Overview:

- Background
  - Whole genome sequencing in context of clinical testing
- What we detect
- Variant classification systems and guidelines
  - ACGS
  - ACMG
- Terminology
- Functional effects
  - What significance do these variants have?
- Evidence we can use to interpret variants
  - Effect on protein
  - Splice site prediction tools
  - Missense variants
  - Databases
  - Frequency
  - Functional assays
  - Segregation
- Variant confirmation and validation
  - What it means to be ‘diagnostic’

Context

- Genetics in diagnosis of disease has been very expensive
  - Pre 2012 £600 for one or two genes
  - Can now order a whole exome for about the same price
- Genetic testing was only undertaken when the clinician was fairly sure of the diagnosis
- Now we can sequence all the genes and relate to phenotype after

Example: Lynch syndrome testing pathway



Microsatellite instability testing in tumour DNA

If high

Immunohistochemistry for loss of mismatch repair proteins

If loss

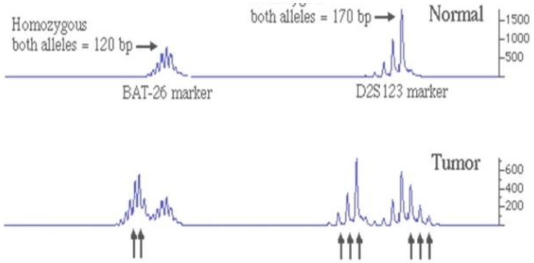
BRAF p.Val600Glu

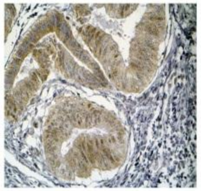
If negative

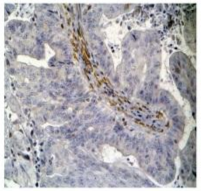
MLH1 promoter methylation

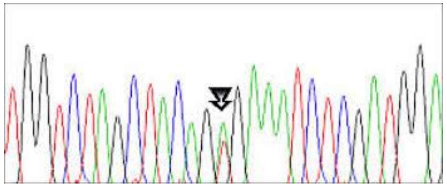
If negative

Germline DNA testing MLH1, MSH2, MSH6, PMS2




hMLH1: positive staining


hMLH1: negative staining



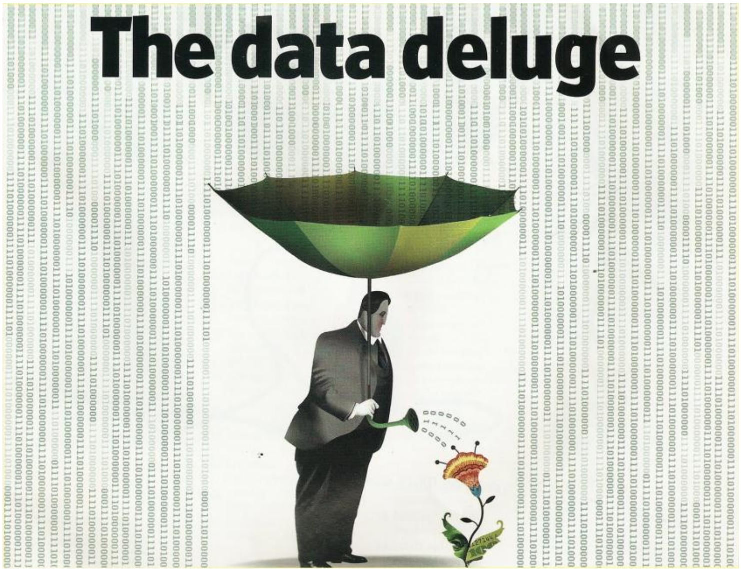
Whole genome sequencing (WGS)

- Only ~1 % of the genome codes for proteins
- Many other functional elements
- Numbers of variants expected from 1 genome:

Type of variant	Average number in European	Average number in African
SNP	3.53 million	4.31 million
Indel	546k	625k
Large deletion	939	1.1k
CNV	157	170
Non-synonymous	10.2k	12.2k
Synonymous	11.2k	13.8k
Intronic	1.68 million	2.06 million

Data from 1000 genomes project

Out of all these variants…which ones cause disease???

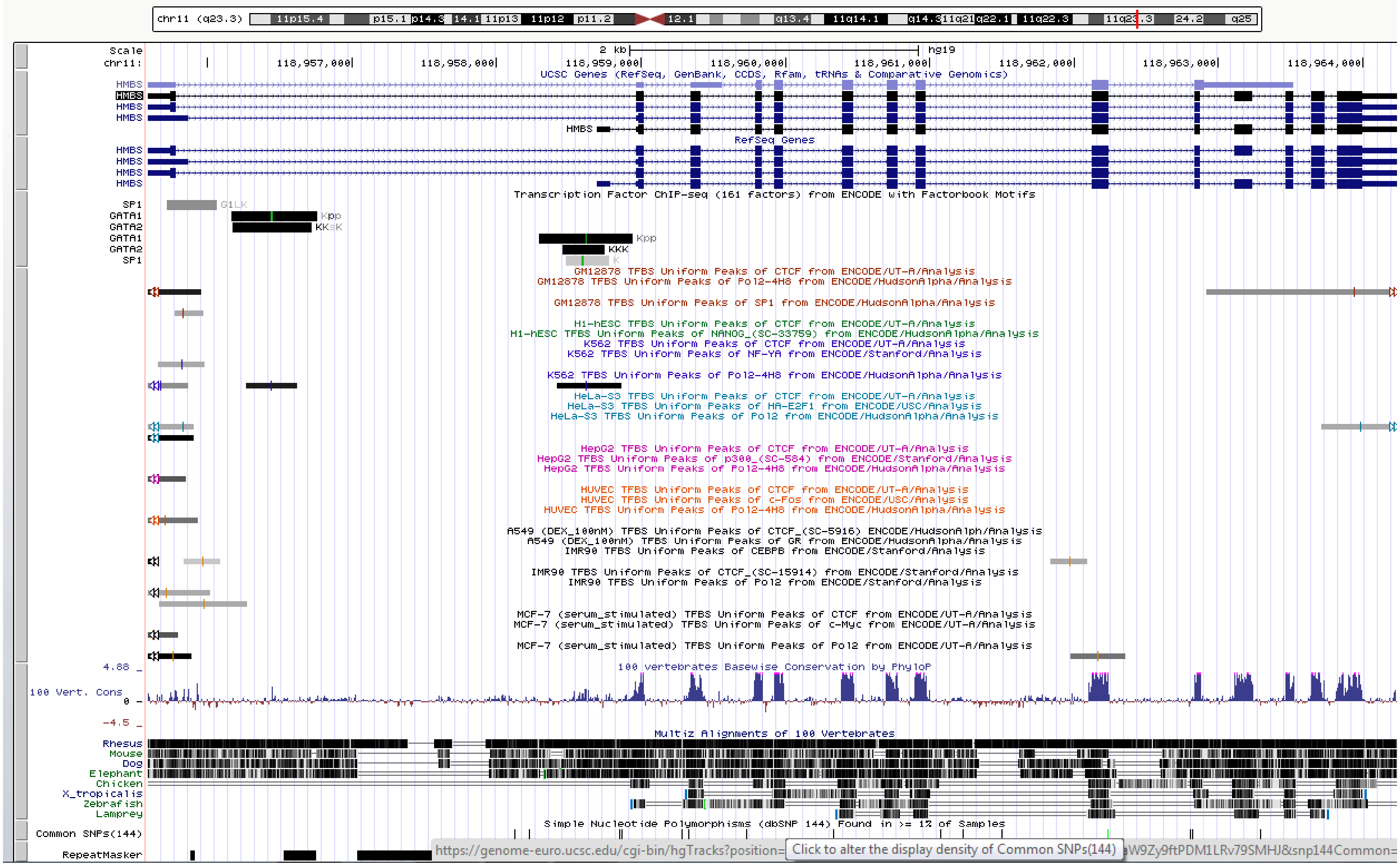


- We must name them systematically (HGVS)
- We must classify them using a standard evidence based approach

HGVS nomenclature

- Must be called against a reference sequence
  - RefSeq
  - LRG – Locus Reference Genomic
  - Genome build
- g. is for genomic locations c. is for coding DNA sequence
- p. is for protein
- m. is for mitochondrial DNA
- Coding nomenclature is where c.1 is the first base of the translation initiation codon

Transcripts


chr11 (q23.3)
11p15.4
p15.1
p14.3
14.1
11p13
11p12
p11.2
12.1
q13.4
11q14.1
q14.3
11q21
q22.1
11q22.3
11q23.3
24.2
q25
Scale
chr11:
2 kb
hg19
118,957,000
118,958,000
118,959,000
118,960,000
118,961,000
118,962,000
118,963,000
118,964,000
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
HMBS
HMBS
HMBS
HMBS
HMBS
RefSeq Genes
HMBS
HMBS
HMBS
HMBS
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
SP1
GATA1
GATA2
GATA1
GATA2
SP1
G1LK
Kpp
KKsK
Kpp
KKK
K
GM12878 TFBS Uniform Peaks of CTCF from ENCODE/UT-A/Analysis
GM12878 TFBS Uniform Peaks of Pol2-4H8 from ENCODE/HudsonAlpha/Analysis
GM12878 TFBS Uniform Peaks of SP1 from ENCODE/HudsonAlpha/Analysis
H1-hESC TFBS Uniform Peaks of CTCF from ENCODE/UT-A/Analysis
H1-hESC TFBS Uniform Peaks of NANOG_(SC-33759) from ENCODE/HudsonAlpha/Analysis
K562 TFBS Uniform Peaks of CTCF from ENCODE/UT-A/Analysis
K562 TFBS Uniform Peaks of NF-YA from ENCODE/Stanford/Analysis
K562 TFBS Uniform Peaks of Pol2-4H8 from ENCODE/HudsonAlpha/Analysis
HeLa-S3 TFBS Uniform Peaks of CTCF from ENCODE/UT-A/Analysis
HeLa-S3 TFBS Uniform Peaks of HA-E2F1 from ENCODE/USC/Analysis
HeLa-S3 TFBS Uniform Peaks of Pol2 from ENCODE/HudsonAlpha/Analysis
HepG2 TFBS Uniform Peaks of CTCF from ENCODE/UT-A/Analysis
HepG2 TFBS Uniform Peaks of p300_(SC-584) from ENCODE/Stanford/Analysis
HepG2 TFBS Uniform Peaks of Pol2-4H8 from ENCODE/HudsonAlpha/Analysis
HUVEC TFBS Uniform Peaks of CTCF from ENCODE/UT-A/Analysis
HUVEC TFBS Uniform Peaks of c-Fos from ENCODE/USC/Analysis
HUVEC TFBS Uniform Peaks of Pol2-4H8 from ENCODE/HudsonAlpha/Analysis
A549 (DEX_100nM) TFBS Uniform Peaks of CTCF_(SC-5916) ENCODE/HudsonAlph/Analysis
A549 (DEX_100nM) TFBS Uniform Peaks of GR from ENCODE/HudsonAlpha/Analysis
IMR90 TFBS Uniform Peaks of CEBPB from ENCODE/Stanford/Analysis
IMR90 TFBS Uniform Peaks of CTCF_(SC-15914) from ENCODE/Stanford/Analysis
IMR90 TFBS Uniform Peaks of Pol2 from ENCODE/Stanford/Analysis
MCF-7 (serum_stimulated) TFBS Uniform Peaks of CTCF from ENCODE/UT-A/Analysis
MCF-7 (serum_stimulated) TFBS Uniform Peaks of c-Myc from ENCODE/UT-A/Analysis
MCF-7 (serum_stimulated) TFBS Uniform Peaks of Pol2 from ENCODE/UT-A/Analysis
4.88 _
100 Vert. Cons
0 -
-4.5 _
100 vertebrates Basewise Conservation by PhyloP
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
Simple Nucleotide Polymorphisms (dbSNP 144) Found in >= 1% of Samples
Common SNPs(144)
RepeatMasker
https://genome-euro.ucsc.edu/cgi-bin/hgTracks?position=
Click to alter the display density of Common SNPs(144)
aW9Zy9ftPDM1LRv79SMHJ&snp144Common=

Classification – UK & US system

Class	Description	Report wording
1	Clearly not pathogenic Benign	Not reported
2	Unlikely to be pathogenic Likely benign	Diagnosis not confirmed
3	Uncertain significance	Diagnosis not confirmed or excluded
4	Likely to be pathogenic	Consistent with diagnosis
5	Certainly pathogenic	Confirms diagnosis

Terms used to describe sequence variants

- **Pathogenic** - contributes mechanistically to disease, but is not necessarily fully penetrant
- **Damaging** - alters the normal levels or biochemical function of a gene or gene product – not necessarily causative
- **Deleterious** - reduces the reproductive fitness of carriers, targeted by purifying natural selection
- **Benign** – not harmful in effect

Key considerations

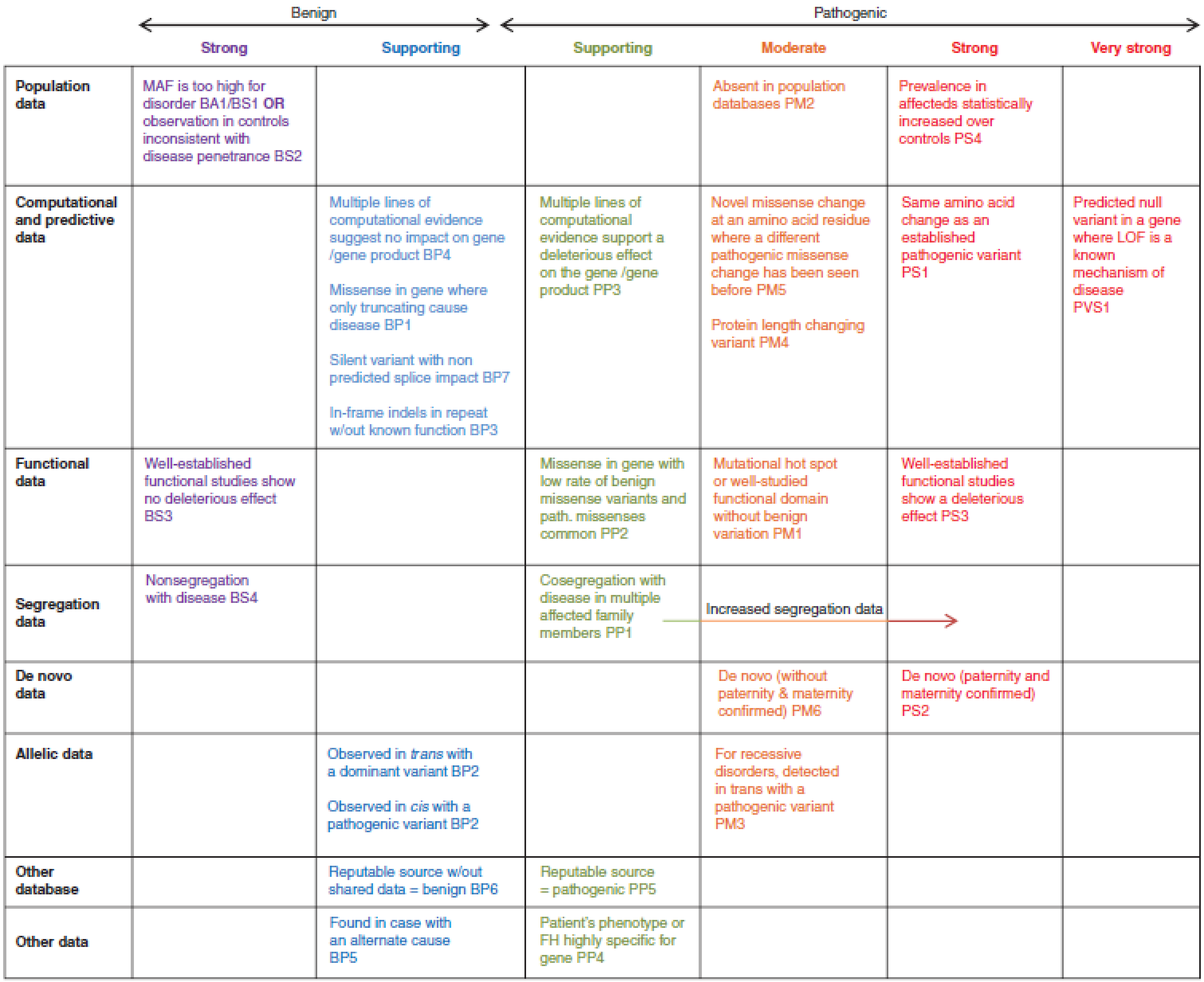
- What is the expected inheritance pattern?
  - Autosomal dominant
  - Autosomal recessive
  - X-linked
  - De novo
- What is the mechanism of pathogenicity?
  - Loss of function
  - Gain of function
  - Haploinsufficiency
- What is the frequency of the disease?
  - Can you rule out common variants (>1%)?

Evidence base for classification

- ACMG (American College of Medical Genetics) have very useful descriptive classification system:

Pathogenic		Example
PVS1	Very Strong	Nonsense variant
PS1-6	Strong	Well established functional assay
PM1-6	Moderate	In mutational hotspot
PP1-5	Supporting	Cosegregates with disease

Benign		Example
BA1	Stand-alone	Allele frequency >5%
BS1-4	Strong	Lack of effect on functional assay
BP1-7	Supporting	Observed in *trans* with pathogenic variant



	Benign		Pathogenic			
	Strong	Supporting	Supporting	Moderate	Strong	Very strong
Population data	MAF is too high for disorder BA1/BS1 OR observation in controls inconsistent with disease penetrance BS2			Absent in population databases PM2	Prevalence in affecteds statistically increased over controls PS4	
Computational and predictive data		Multiple lines of computational evidence suggest no impact on gene /gene product BP4 Missense in gene where only truncating cause disease BP1 Silent variant with non predicted splice impact BP7 In-frame indels in repeat w/out known function BP3	Multiple lines of computational evidence support a deleterious effect on the gene /gene product PP3	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before PM5 Protein length changing variant PM4	Same amino acid change as an established pathogenic variant PS1	Predicted null variant in a gene where LOF is a known mechanism of disease PVS1
Functional data	Well-established functional studies show no deleterious effect BS3		Missense in gene with low rate of benign missense variants and path. missenses common PP2	Mutational hot spot or well-studied functional domain without benign variation PM1	Well-established functional studies show a deleterious effect PS3	
Segregation data	Nonsegregation with disease BS4		Cosegregation with disease in multiple affected family members PP1	Increased segregation data →		
De novo data				De novo (without paternity & maternity confirmed) PM6	De novo (paternity and maternity confirmed) PS2	
Allelic data		Observed in *trans* with a dominant variant BP2 Observed in *cis* with a pathogenic variant BP2		For recessive disorders, detected in trans with a pathogenic variant PM3		
Other database		Reputable source w/out shared data = benign BP6	Reputable source = pathogenic PP5			
Other data		Found in case with an alternate cause BP5	Patient's phenotype or FH highly specific for gene PP4			

Figure 1 Evidence framework. This chart organizes each of the criteria by the type of evidence as well as the strength of the criteria for a benign (left side) or pathogenic (right side) assertion. Evidence code descriptions can be found in **Tables 3** and **4**. BS, benign strong; BP, benign supporting; FH, family history; LOF, loss of function; MAF, minor allele frequency; path., pathogenic; PM, pathogenic moderate; PP, pathogenic supporting; PS, pathogenic strong; PVS, pathogenic very strong.

What effect does the variant have?

- Is it in a region known to be functional?
  - Protein coding regions
  - Flanking splice site
  - Promoter
- However we know from ENCODE that there are many more regions of the genome that are functional

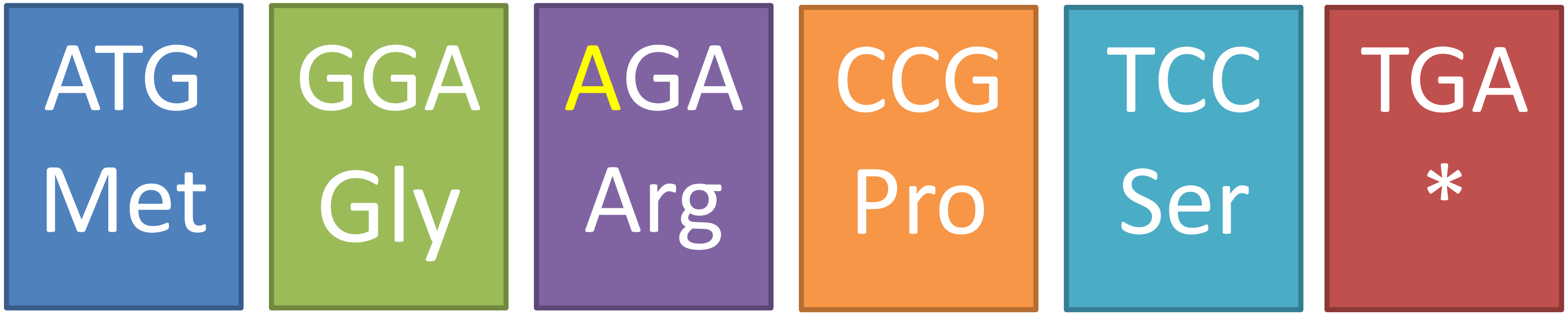
In a known functional region = more likely to be pathogenic

Types of variant that will severely affect the protein …*there are always exceptions*

- Nonsense
  - Premature stop codons – nonsense mediated decay (NMD)
- Frameshift
  - Translated protein scrambled – often results in NMD
- Canonical splice site
  - Causes exon skipping and loss of functional domains or frameshift

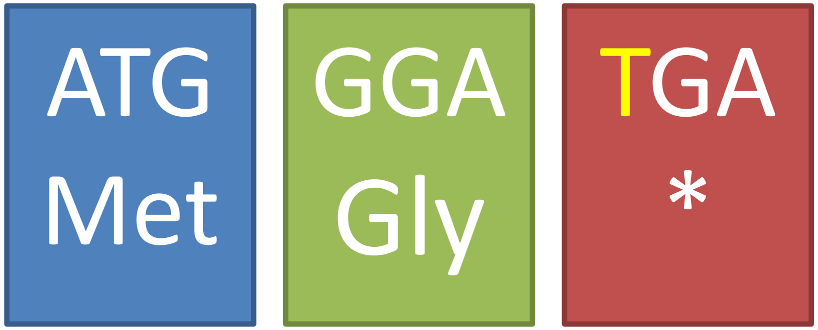
Nonsense

Normal sequence



DNA sequence	ATG	GGA	AGA	CCG	TCC	TGA
Protein sequence	Met	Gly	Arg	Pro	Ser	*

Mutated sequence c.7A>T

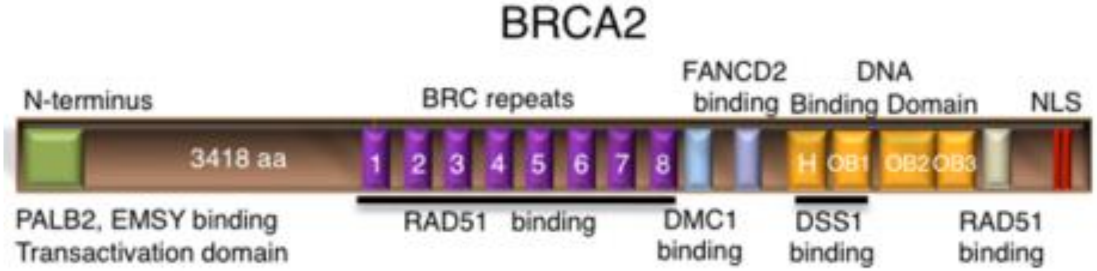


DNA sequence	ATG	GGA	TGA
Protein sequence	Met	Gly	*

Loss of part of functional part of protein = likely to be pathogenic

Exception to the rule…

- A truncating variant in the last exon
- BRCA2: c.9976A > T (p.Lys3326*)
- Protein is still functional



- Common in European population
- Was mistakenly assigned as a pathogenic variant
- Always be aware of the context of your mutation within the gene

Frameshift

Scrambled reading frame = likely to be pathogenic

Normal sequence

DNA sequence	ATG	GGA	AGA	CCG	TCC	TGA
Protein sequence	Met	Gly	Arg	Pro	Ser	*

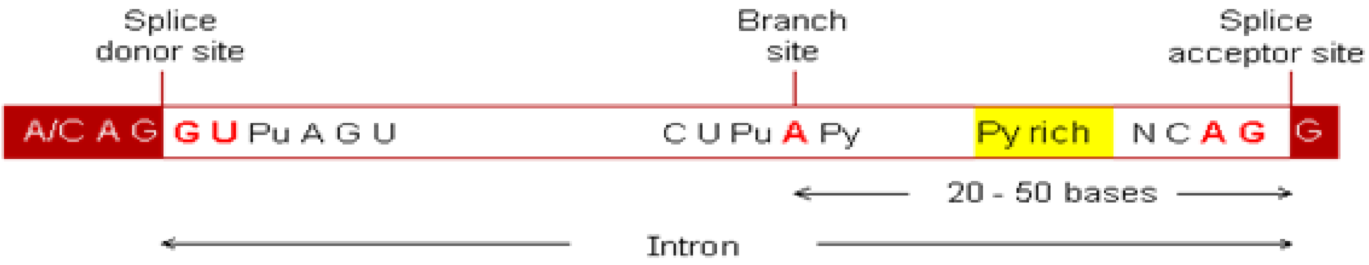
Mutated sequence c.7delT

DNA sequence	ATG	GGA	GAC	CGT	CCT	GA…
Protein sequence	Met	Gly	Asp	Arg	Pro	…..

Splice site

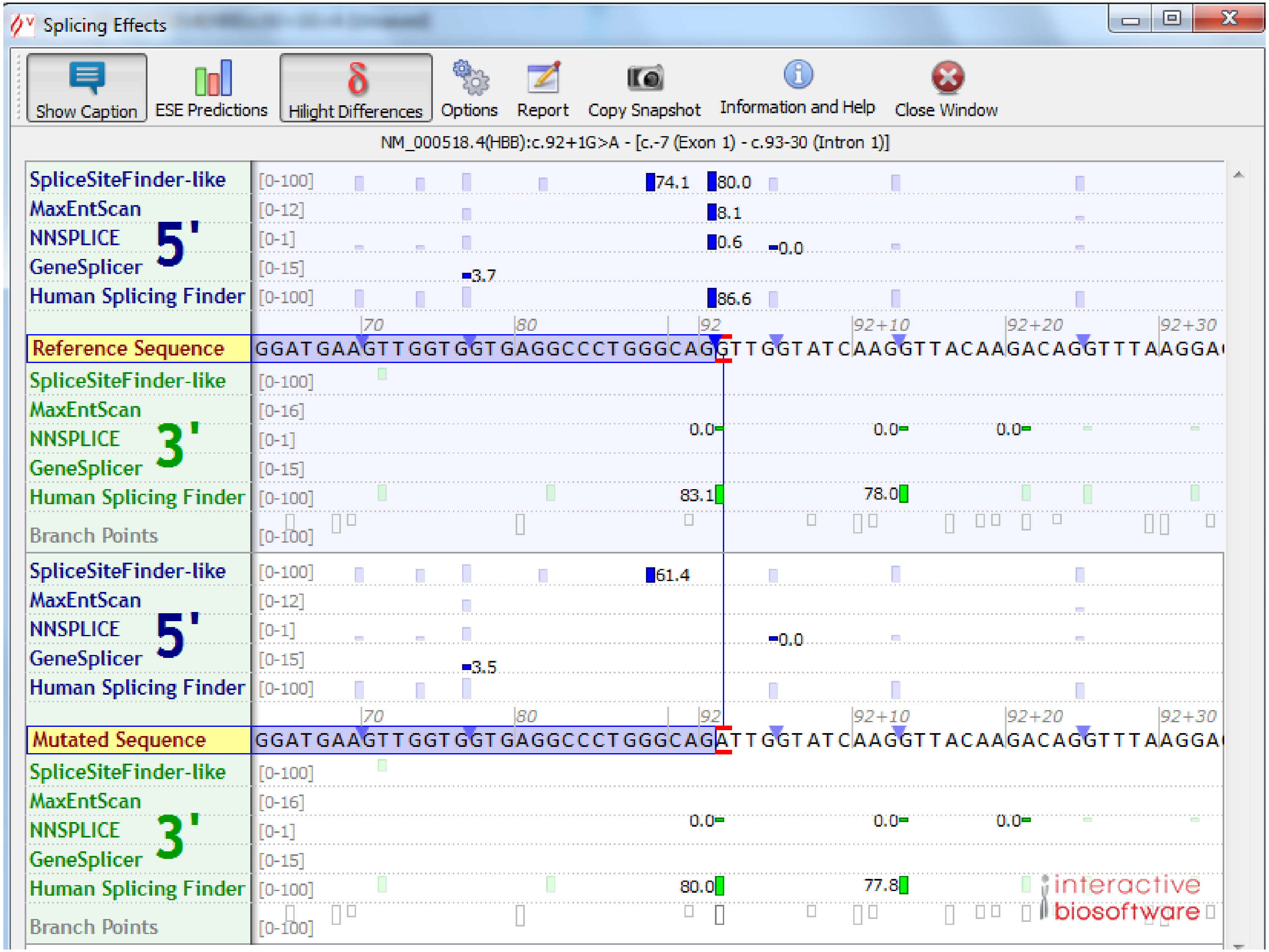
Affects splicing = likely to be pathogenic

- Splicing is all about recognition of sequence motifs by the spliceosome:



- Variants in the consensus splice site can lead to exon skipping or incorporation of intronic sequence into the mRNA transcript
- Cryptic splice sites can also be activated

Splice site prediction software



NM_000518.4(HBB):c.92+1G>A - [c.-7 (Exon 1) - c.93-30 (Intron 1)]

Reference Sequence: GGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGGTTACAAGACAGGTTTAAGGA

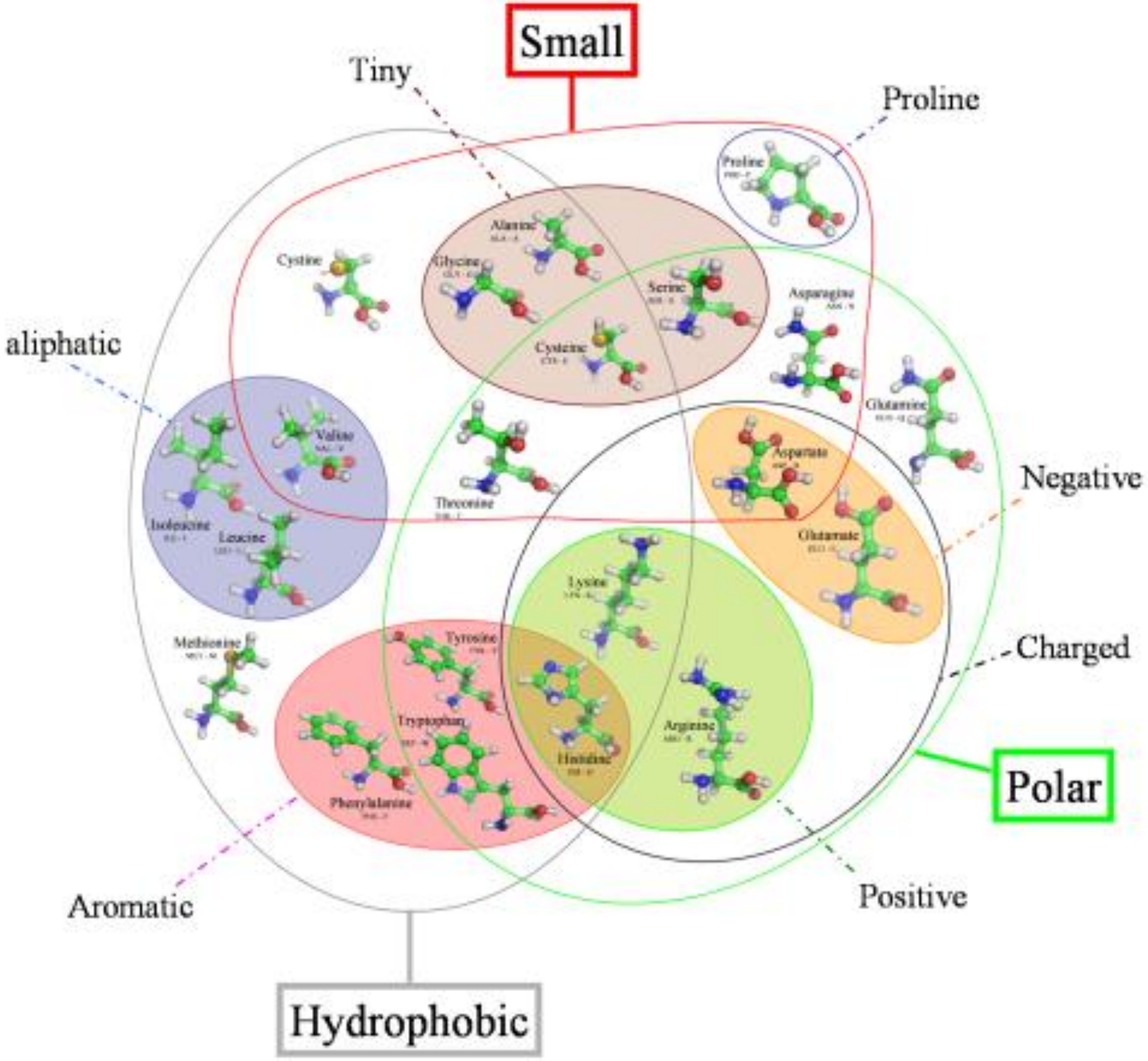
Mutated Sequence: GGATGAAGTTGGTGGTGAGGCCCTGGGCAGATTGGTATCAAGGTTACAAGACAGGTTTAAGGA

Subtle effects on protein - missense

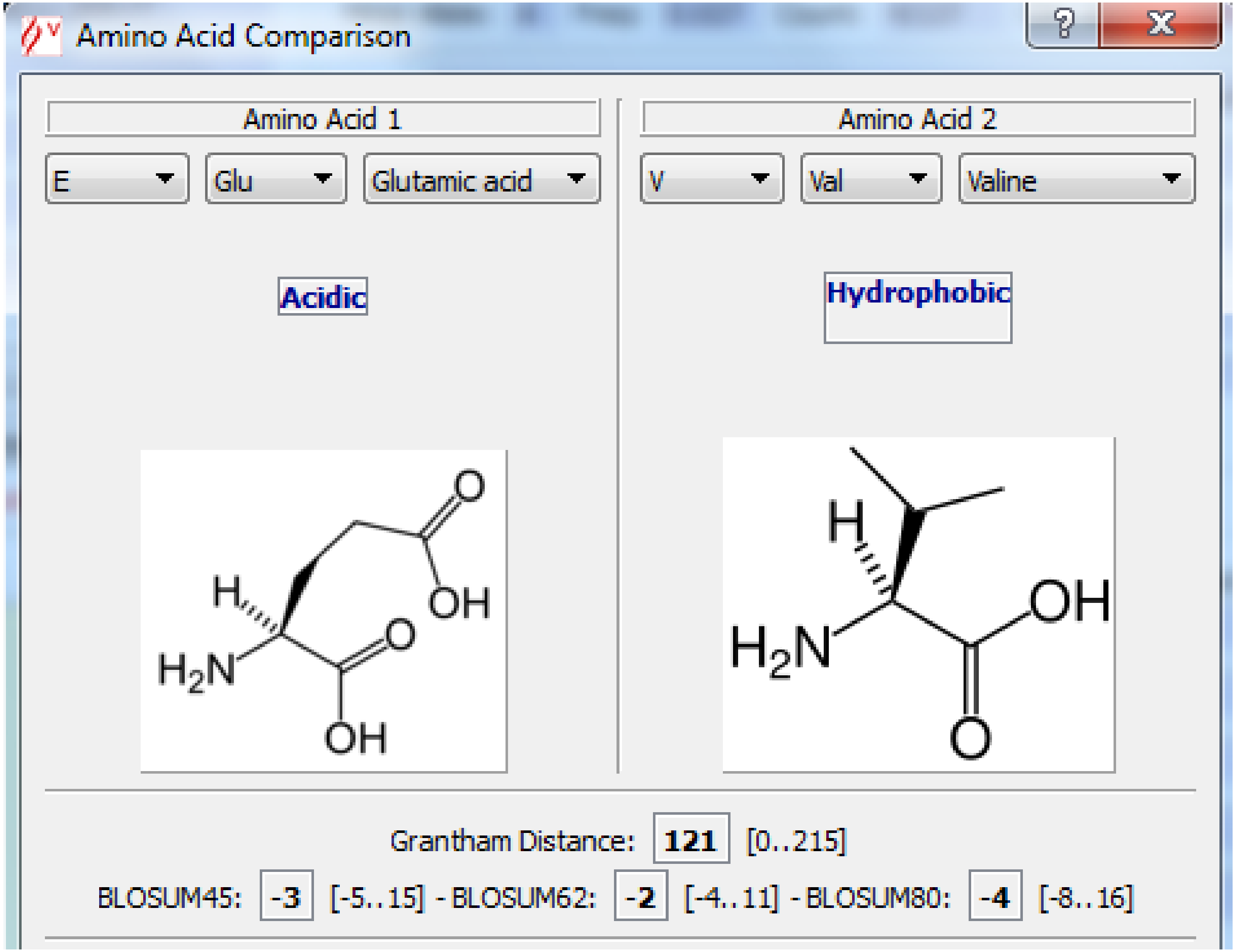
- Substituting one amino acid for another
- Can affect functional part of protein
- Adding or removing cysteine alters potential for forming disulphide bridges
- Grantham distance: a measure of physiochemical difference

Large physiochemical difference = more likely to be pathogenic

Amino acid properties



Grantham distance



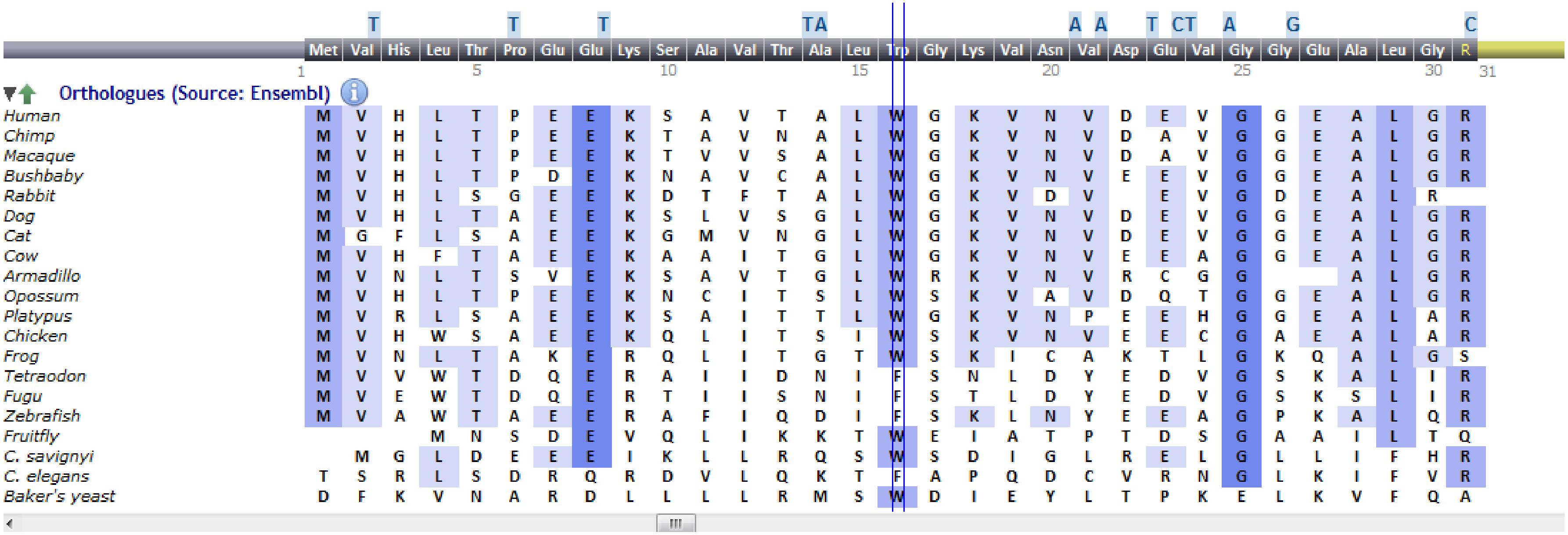
E | Glu | Glutamic acid

V | Val | Valine

Grantham Distance: 121 [0..215]

BLOSUM45: -3 [-5..15] - BLOSUM62: -2 [-4..11] - BLOSUM80: -4 [-8..16]

Multiple sequence alignments



Orthologues (Source: Ensembl)

		T				T		T						TA						A	A		T	CT	A	G					C
	Met	Val	His	Leu	Thr	Pro	Glu	Glu	Lys	Ser	Ala	Val	Thr	Ala	Leu	Trp	Gly	Lys	Val	Asn	Val	Asp	Glu	Val	Gly	Gly	Glu	Ala	Leu	Gly	R
	1				5					10					15					20					25					30	31
Human	M	V	H	L	T	P	E	E	K	S	A	V	T	A	L	W	G	K	V	N	V	D	E	V	G	G	E	A	L	G	R
Chimp	M	V	H	L	T	P	E	E	K	T	A	V	N	A	L	W	G	K	V	N	V	D	A	V	G	G	E	A	L	G	R
Macaque	M	V	H	L	T	P	E	E	K	T	V	V	S	A	L	W	G	K	V	N	V	D	A	V	G	G	E	A	L	G	R
Bushbaby	M	V	H	L	T	P	D	E	K	N	A	V	C	A	L	W	G	K	V	N	V	E	E	V	G	G	E	A	L	G	R
Rabbit	M	V	H	L	S	G	E	E	K	D	T	F	T	A	L	W	G	K	V	D	V		E	V	G	D	E	A	L	R	
Dog	M	V	H	L	T	A	E	E	K	S	L	V	S	G	L	W	G	K	V	N	V	D	E	V	G	G	E	A	L	G	R
Cat	M	G	F	L	S	A	E	E	K	G	M	V	N	G	L	W	G	K	V	N	V	D	E	V	G	G	E	A	L	G	R
Cow	M	V	H	F	T	A	E	E	K	A	A	I	T	G	L	W	G	K	V	N	V	E	E	A	G	G	E	A	L	G	R
Armadillo	M	V	N	L	T	S	V	E	K	S	A	V	T	G	L	W	R	K	V	N	V	R	C	G	G			A	L	G	R
Opossum	M	V	H	L	T	P	E	E	K	N	C	I	T	S	L	W	S	K	V	A	V	D	Q	T	G	G	E	A	L	G	R
Platypus	M	V	R	L	S	A	E	E	K	S	A	I	T	T	L	W	G	K	V	N	P	E	E	H	G	G	E	A	L	A	R
Chicken	M	V	H	W	S	A	E	E	K	Q	L	I	T	S	I	W	S	K	V	N	V	E	E	C	G	A	E	A	L	A	R
Frog	M	V	N	L	T	A	K	E	R	Q	L	I	T	G	T	W	S	K	I	C	A	K	T	L	G	K	Q	A	L	G	S
Tetraodon	M	V	V	W	T	D	Q	E	R	A	I	I	D	N	I	F	S	N	L	D	Y	E	D	V	G	S	K	A	L	I	R
Fugu	M	V	E	W	T	D	Q	E	R	T	I	I	S	N	I	F	S	T	L	D	Y	E	D	V	G	S	K	S	L	I	R
Zebrafish	M	V	A	W	T	A	E	E	R	A	F	I	Q	D	I	F	S	K	L	N	Y	E	E	A	G	P	K	A	L	Q	R
Fruitfly				M	N	S	D	E	V	Q	L	I	K	K	T	W	E	I	A	T	P	T	D	S	G	A	A	I	L	T	Q
C. savignyi		M	G	L	D	E	E	E	I	K	L	L	R	Q	S	W	S	D	I	G	L	R	E	L	G	L	L	I	F	H	R
C. elegans	T	S	R	L	S	D	R	Q	R	D	V	L	Q	K	T	F	A	P	Q	D	C	V	R	N	G	L	K	I	F	V	R
Baker's yeast	D	F	K	V	N	A	R	D	L	L	L	L	R	M	S	W	D	I	E	Y	L	T	P	K	E	L	K	V	F	Q	A

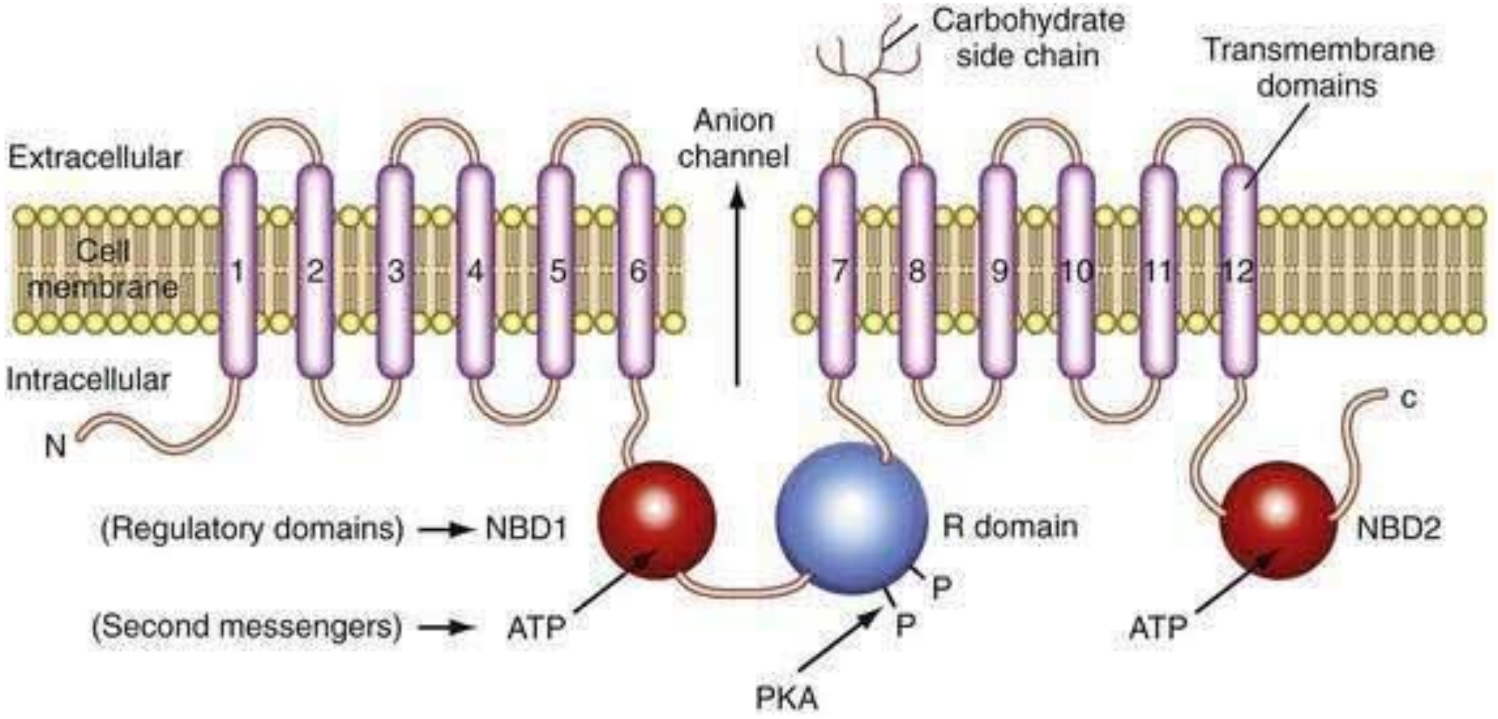
- Assessment of evolutionary conservation

More conserved = more important for function

Not conserved = less likely to be pathogenic

Protein domains/mutational hotspots

CFTR protein structure



In important protein domain = more likely to be pathogenic

Missense prediction tools

- In silico prediction
- SIFT – Sorting intolerant from tolerant
- PolyPhen
- AlignGVGD
- MutationTaster
- All are based upon physiochemical differences, position in the protein and multiple sequence alignments or a combination
- Use with CAUTION
- Never use in isolation

Very difficult to tell anything from missense prediction tools!

HBB c.20A>T; p.Glu7Val
HbS Sickle

- One example exemplifies why you must consider ALL THE DATA
- Missense
- 4.85% in African population (ExAC)
- Weakly conserved nucleotide (phyloP: 0.04 [-14.1;6.4])
- Moderately conserved amino acid (considering 20 species)
- Moderate physicochemical difference between Glu and Val (Grantham dist.: 121 [0-215])
- This variant is in protein domain: Globin, subset
- Align GVGD: C0 (GV: 164.97 - GD: 0.00)
- SIFT: Tolerated (score: 0.1, median: 2.21)
- MutationTaster: polymorphism (p-value: 1)

Databases

Reported as pathogenic = more likely to be pathogenic

- Leiden Open Variation database LOVD
- Human Gene Mutation Database HGMD
- ClinVar
- Online Mendelian Inheritance in Man OMIM
- DECIPHER
- Many smaller disease specific databases
- Considerations
  - Is it curated/updated?
  - HGVS nomenclature and transcript
  - Validated data
  - Source and independence of the observations listed
  - Always reassess their data

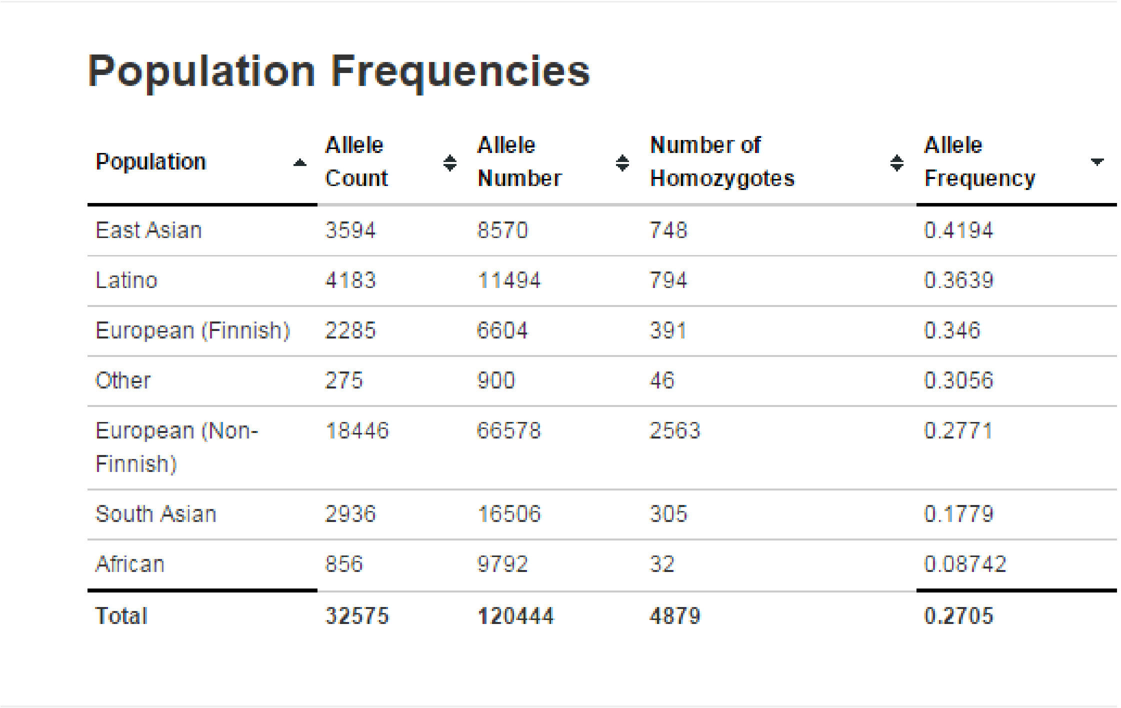
Variant frequency

Higher frequency = less likely to be pathogenic

- Probably one of the most useful ways of assigning classes 1 and 2 (not pathogenic)
- Consider the incidence of the disease vs variant frequency and inheritance
- Several large databases that hold variant frequency data
  - ExAC – Exome Aggregation consortium
  - dbSNP
  - EVS – Exome variant server
- According to ACMG: >5% freq is stand-alone support for classification as benign

Example:

- You find a variant at 36% in the Latino population in ExAC



Population Frequencies

Population ▲	Allele Count ⬍	Allele Number ⬍	Number of Homozygotes ⬍	Allele Frequency ▼
East Asian	3594	8570	748	0.4194
Latino	4183	11494	794	0.3639
European (Finnish)	2285	6604	391	0.346
Other	275	900	46	0.3056
European (Non-Finnish)	18446	66578	2563	0.2771
South Asian	2936	16506	305	0.1779
African	856	9792	32	0.08742
Total	**32575**	**120444**	**4879**	**0.2705**

- What are your considerations?

Functional assays

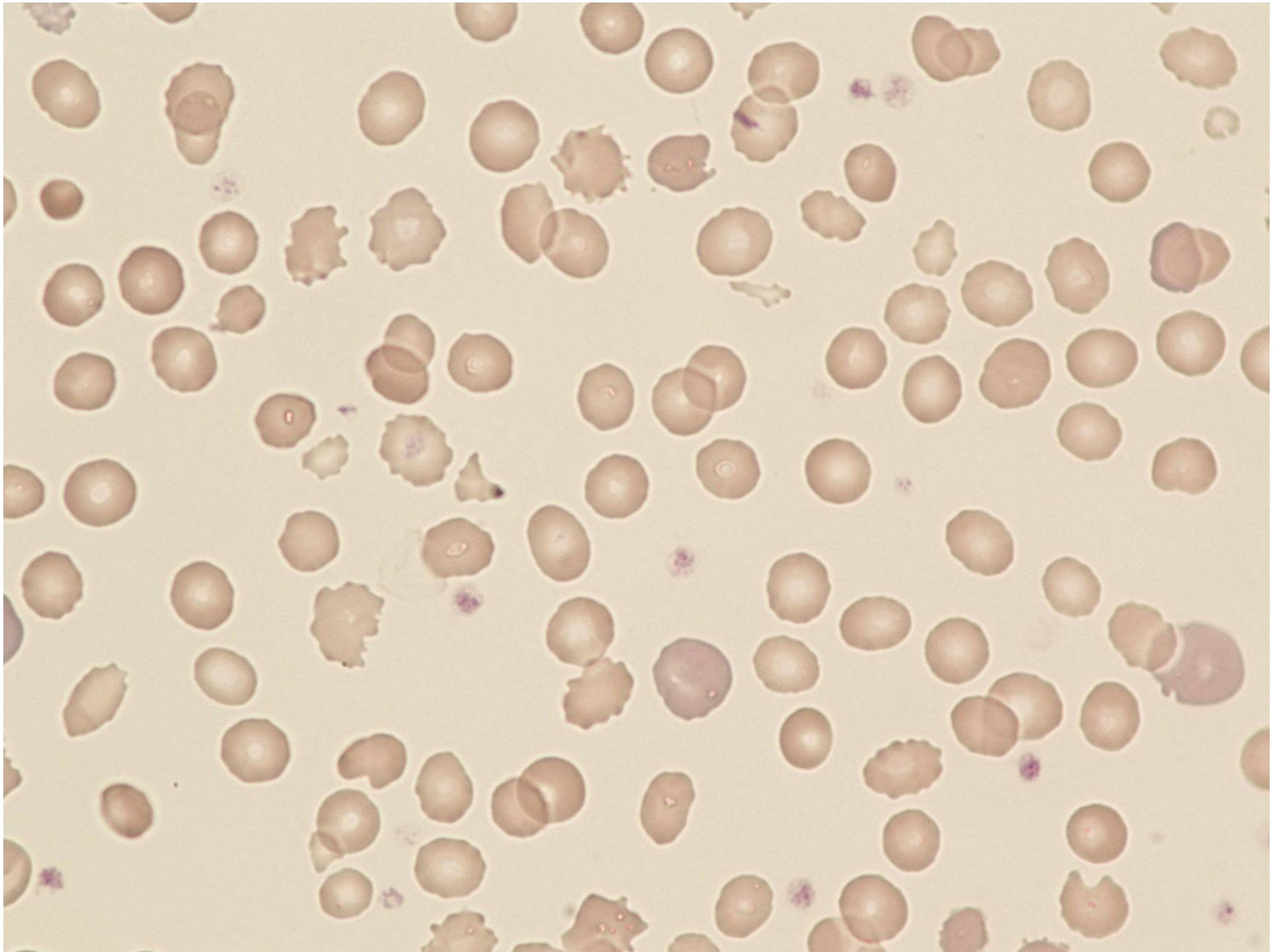
Effect seen functionally = more likely to be pathogenic

- Can be very useful in looking at the *in vivo* effect of a variant
- Many metabolic enzymes have robust functional assays
  - Pyruvate kinase
- mRNA sequencing very good at predicting effects of splicing variants

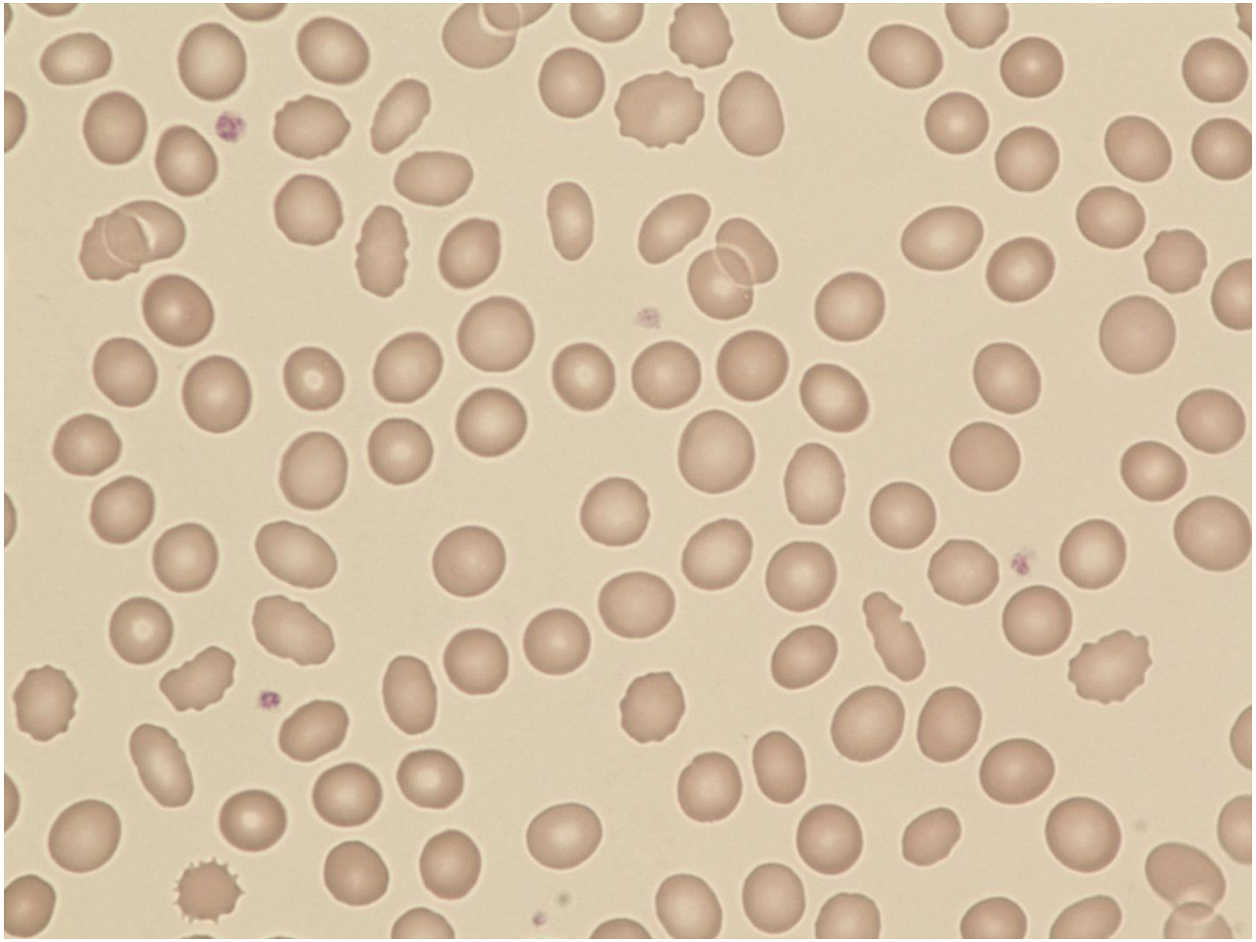
Example:

- You find a PKLR class 3 missense in *trans* with a class 5 in a patient with haemolytic anaemia
- PKLR activity can be assayed on a fresh EDTA sample
- PKLR assay shows the patient to be deficient
- This data can help to re-classify the class 3 variant as pathogenic

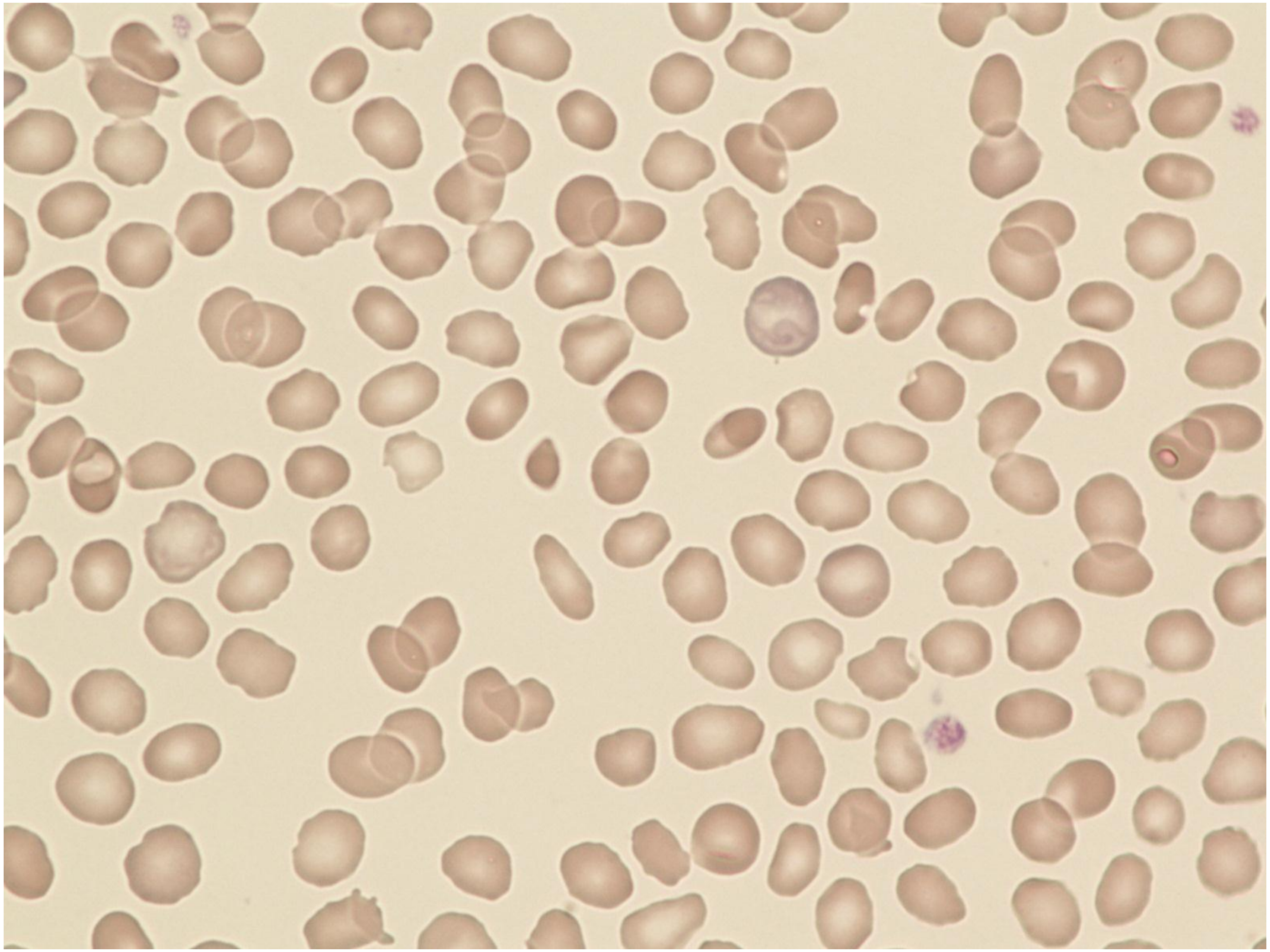
- Blood film: You are looking at the functional effect



- Mother:
  - Homozygous *SPTA1:* c.[5572C>G; 6531-12C>T]; p.[Leu1858Val;?] Low expression allele



- Father:
    - Heterozygous *SPTA1:* c.83G>A; p.Arg28His

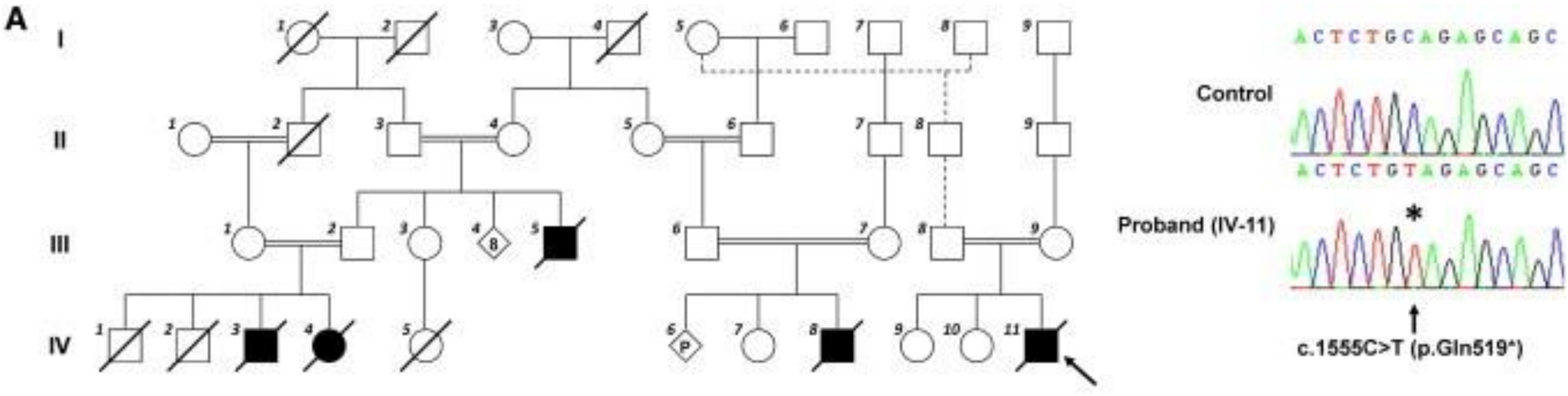


Segregation

Segregates with disease in family = more likely to be pathogenic

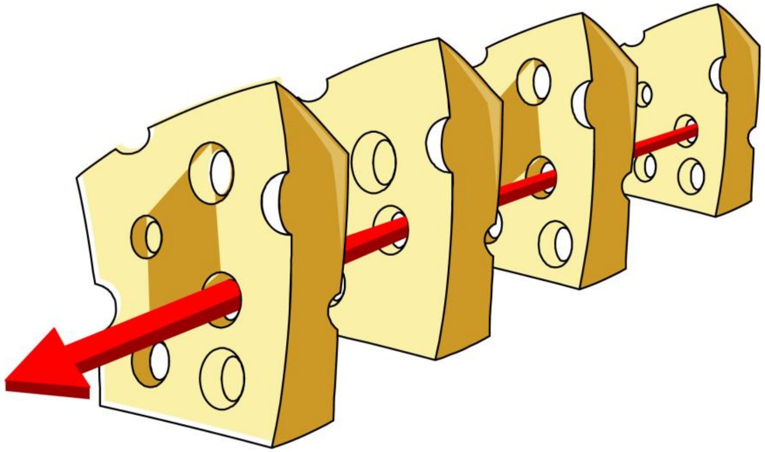
Found in unaffected = less likely to be pathogenic

- Is it in *cis/trans* with other variants in the patient?
- Testing of other affected or unaffected family members



Diagnostic validation of results

- What does it mean to be ‘diagnostic’?
  - ISO15189 accreditation for quality standards
  - The cheese model
- We confirm a variant by Sanger sequencing
- Reassess pathogenicity in relation to phenotype
- Report



What should be included on the clinical report?

- Results
  - HGVS nomenclature
  - Gene name, cDNA, protein
  - Disease
  - Inheritance
- Interpretation
  - Evidence supporting variant classification
  - Does it explain the patient’s phenotype?
  - Supplementary testing

Report…

- Methodology
  - Laboratory and analysis tools used
  - Limitations
- Risk to offspring
- Referral for genetic counselling
- Incidental findings??

Incidental findings

- Much debate on if we should report pathogenic variants that are not associated with the condition that the patient has been referred for
  - Eg Cystic Fibrosis carrier status
  - Eg Late onset dominant disorders such as BRCA
- For 100k the patient will decide at consent

Examples

- Hereditary spastic paraplegia referral
- Heterozygous *REEP1* variant c.471del p.(Thr158fs)
- What questions will you ask?

- Is this gene known to be associated with the condition?
  - Yes
- Mode of inheritance?
  - Autosomal dominant
- How big is the gene?
  - 7 exons
- Where does the variant lie in the gene
  - Penultimate exon