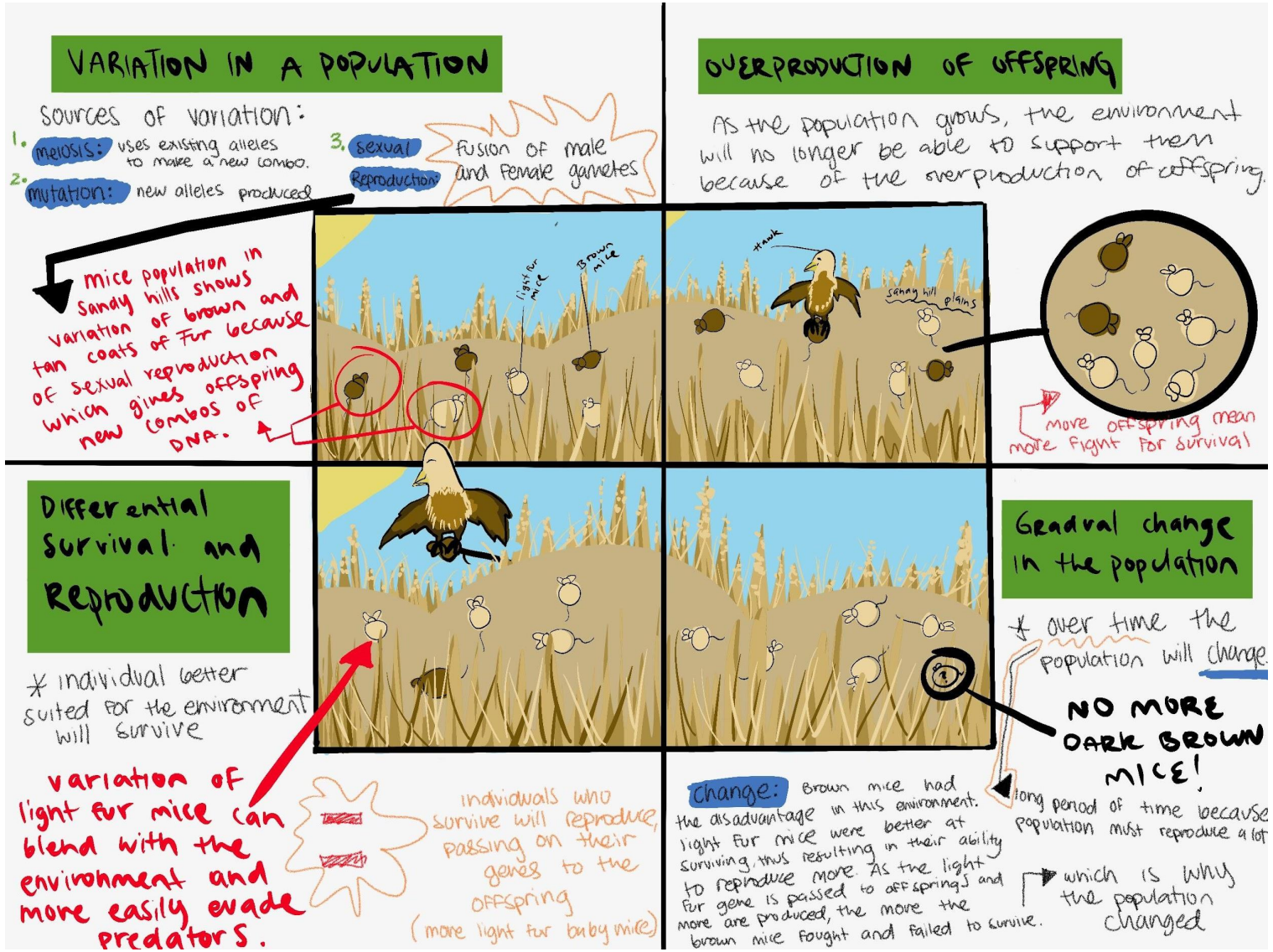
Variant Annotation

SM Adadey

Annotation

- An extra information associated with a particular point[1]
- A note of explanation or comment added to a text or diagram[2]



wikimedia

[1]*Wikipedia,* [2]*Oxford Languages*

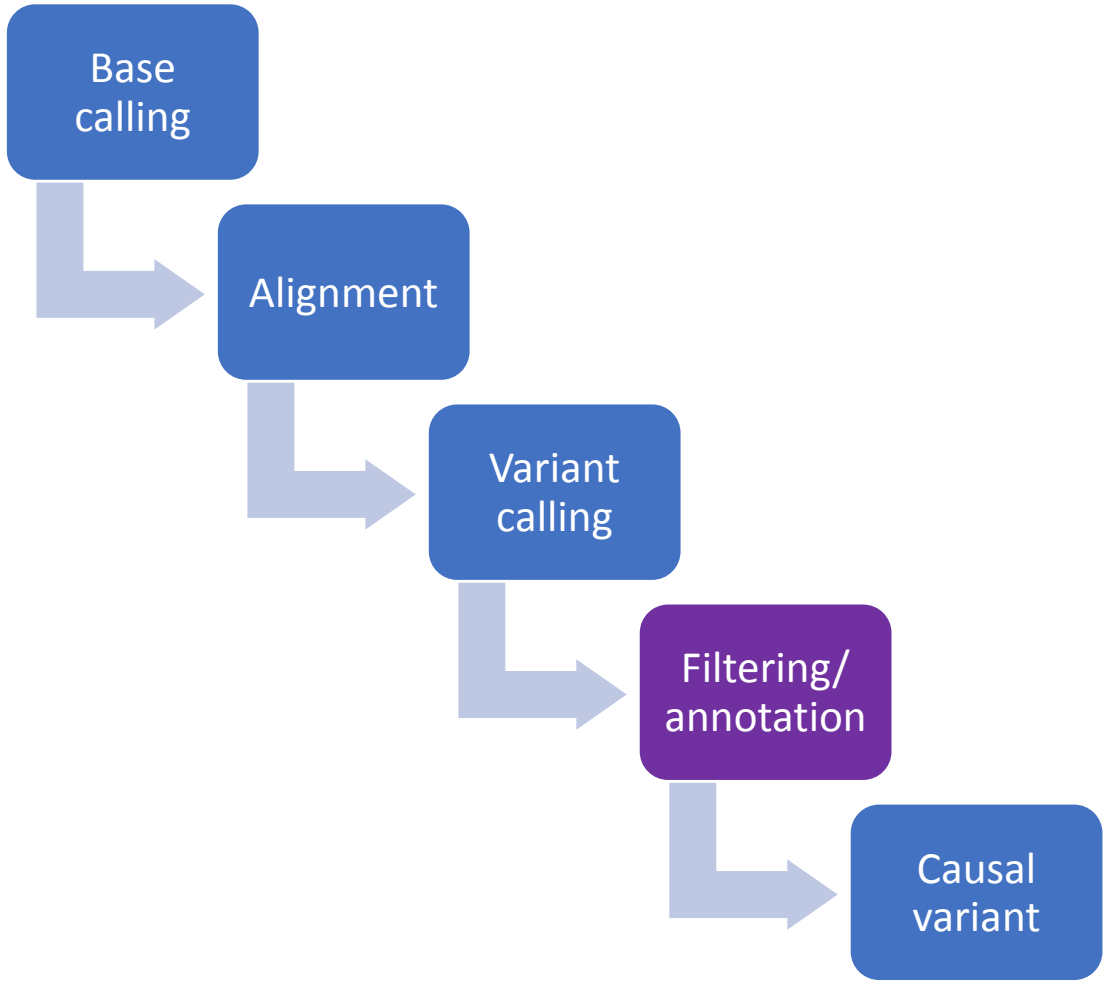
Outline

Variant annotation

- Describe **types of variants**: short vs structural, somatic vs germline
- Describe **variant consequences** and why these are important
- Describe **algorithms** such as SIFT & PolyPhen, CADD
- G2P gene lists

Computational resources and tools that are useful to determine the functional part of variants

- VEP in the Ensembl browser
- Using VEP with the G2P plugin for variant prioritisation
- VEP exercise



What must I do with this vcf?

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4
```

What must I do with the vcf?

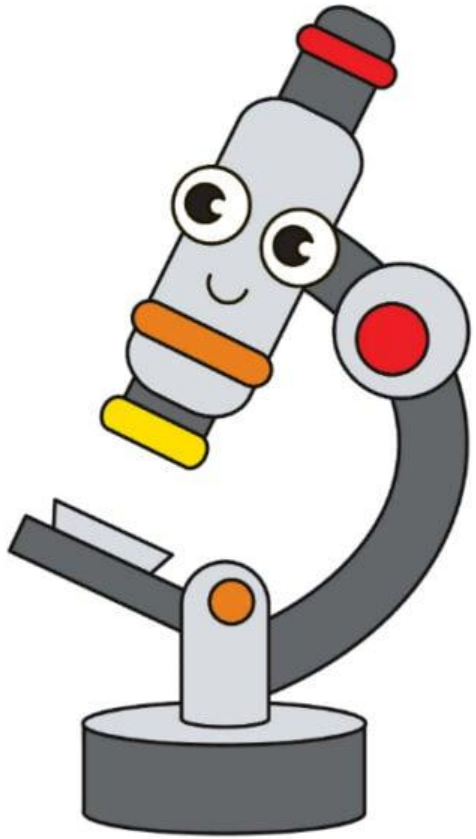
Annotation?

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",
##phasing=partial
```

- Assigning functional information to DNA variants

```
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4
```

Annotation



#VCF/CI	VCF/POS	VCF/ID	VCF/REF	VCF/ALT	VCF/QU/	VCF/FILT	CAVA/GI	CAVA/TRANSCRIF	CAVA/CSN	IMPACT	:AVA/LO	CAVA/CI	CADD_p	MetaSVM	Mutation	PROVEA	VEST4_s	BIOR::GM	BIOR::GM	BIOR::GM	BI
1	6E+07	.	A	ACTG	35	PASS	PCSK9	NM_174936.3	c.63_65dupGCT_p.Leu23dup	2	Ex1	IF	.	.	.	.	.	.	.	.	.
2	5E+07	.	G	A	35	PASS	MSH6	NM_000179.2	c.3556+146G>A	3	In6/7	INT	.	.	.	.	.	.	.	.	.
2	5E+07	.	C	G	35	PASS	MSH6	NM_000179.2	c.1186C>G_p.Leu396Val	2	Ex4	NSY	21.7	T	D	N	0.32	1556	276930	0.0056	.
13	3E+07	.	T	A,TA	35	PASS	BRCA2	NM_000059.3	c.68-4dupA	3	In2/3	SS	.	.	.	.	.	787	274652	0.0029	.
13	3E+07	.	T	A,TA	35	PASS	BRCA2	NM_000059.3	c.68-4dupA	3	In2/3	SS	.	.	.	.	.	787	274652	0.0029	.
19	1E+07	.	A	G	35	PASS	LDLR	NM_000527.4	c.1510A>G_p.Lys504Glu	2	Ex10	NSY	21.8	T	D	N	0.347	5	246252	2E-05	.
2	2E+07	.	C	T	35	PASS	APOB	NM_000384.2	c.5066G>A_p.Arg1689His	2	Ex26	NSY	25.9	T	D	D	0.718	396	276304	0.0014	.
7	6E+06	.	T	C	35	PASS	PMS2	NM_000535.6	c.2324A>G_p.Asn775Ser	2	Ex14	NSY	23.2	T	D	D	0.391	88	275058	0.0003	.
17	4E+07	.	A	C	35	PASS	BRCA1	NM_007294.3	c.736T>G_p.Leu246Val	2	Ex10	NSY	11.52	D	D	N	0.358	82	276040	0.0003	.
19	1E+07	.	G	A	35	PASS	LDLR	NM_000527.4	c.313+1G>A	1	In3/4	ESS	24.8	.	D	.	.	7	246208	3E-05	.
7	6E+06	.	T	A	35	PASS	PMS2	NM_000535.6	c.251-2A>T	1	In3/4	ESS	27	.	D	.	.	1	239164	4E-06	.
3	4E+07	.	AG	CC	35	PASS	MLH1	NM_000249.3	c.1410-2_1410-1delinsCC	1	In12/13	ESS	.	.	.	.	.	.	.	.	.
3	4E+07	.	AG	CC	35	PASS	MLH1	NM_000249.3	c.1410-2_1410-1delinsCC	1	In12/13	ESS	.	.	.	.	.	.	.	.	.
2	5E+07	.	A	AAC	35	PASS	MSH6	NM_000179.2	c.843_844insAC	1	Ex4	FS	.	.	.	.	.	1	245844	4E-06	.
7	6E+06	.	CTT	C	35	PASS	PMS2	NM_000535.6	c.1312_1313delAA	1	Ex11	FS	.	.	.	.	.	.	.	.	.
13	3E+07	.	AGCAAG	A	35	PASS	BRCA2	NM_000059.3	c.6024_6035delinsTGCTGTT	1	Ex11	FS	.	.	.	.	.	.	.	.	.
2	2E+07	.	C	A	35	PASS	APOB	NM_000384.2	c.409G>T_p.Glu137X	1	Ex5	SG	34	.	A	.	0.796	.	.	.	.
2	2E+07	.	C	A	35	PASS	APOB	NM_000384.2	c.409G>T_p.Glu137X	1	Ex5	SG	34	.	A	.	0.796	.	.	.	.
2	2E+07	.	C	A	35	PASS	APOB	NM_000384.2	c.409G>T_p.Glu137X	1	Ex5	SG	34	.	A	.	0.796	.	.	.	.
19	1E+07	.	C	A	35	PASS	LDLR	NM_000527.4	c.2546C>A_p.Ser849X	1	Ex17	SG	42	.	D	.	0.71	.	.	.	.
1	6E+07	.	G	GGAGGA	35	PASS	PCSK9	NM_174936.3	c.101_106dup6_p.Glu34_Asp3	2	Ex1	IF	.	.	.	.	.	19	175140	0.0001	.
2	2E+07	.	CTCA	C	35	PASS	APOB	NM_000384.2	c.6633_6641delTGA_p.Asp2	2	Ex26	IF	.	.	.	.	.	1266	245070	0.0052	.
2	2E+07	.	CTCA	C	35	PASS	APOB	NM_000384.2	c.6633_6641delTGA_p.Asp2	2	Ex26	IF	.	.	.	.	.	1266	245070	0.0052	.
7	6E+06	.	A	G	35	PASS	PMS2	NM_000535.6	c.2T>C_p.Met1?	2	Ex1	IM	22.7	D	D	N	0.953	1	245322	4E-06	.
17	4E+07	.	G	GCCT	35	PASS	BRCA1	NM_007294.3	c.2_3insAGG_p.Met1?	2	Ex2	IM	.	.	.	.	.	.	.	.	.
17	4E+07	.	TT	T	35	PASS	BRCA1	NM_007294.3	c.1delA_p.Met1?	1	Ex2	FS	.	.	.	.	.	.	.	.	.
17	4E+07	.	C	G	35	PASS	BRCA1	NM_007294.3	c.5591G>C_p.X1864SerextX3	2	Ex23	SL	11.18	.	N	.	0.431	.	.	.	.
13	3E+07	.	A	AT	35	PASS	BRCA2	NM_000059.3	c.10256_10257insT	1	Ex27	FS	.	.	.	.	.	.	.	.	.
7	6E+06	.	CTGA	C	35	PASS	MSH2	NM_000251.2	c.2583_2585delGAA_p.Gln86	2	Ex15	IF	.	.	.	.	.	.	.	.	.
2	5E+07	.	A	G	35	PASS	MSH6	NM_000179.2	c.3649A>G_p.Arg1217Gly	2	Ex8	EE	33	D	D	D	0.937	1	30978	3E-05	.
2	5E+07	.	A	G	35	PASS	MSH6	NM_000179.2	:.3649A>G_p.Arg1217Gly	2	Ex8	EE	33	D	D	D	0.937	1	30978	3E-05	.
19	1E+07	.	C	T	35	PASS	LDLR	NM_000527.4	c.2388C>T_p.	2	Ex16	EE	.	.	.	.	.	11	277170	4E-05	.
2	5E+07	.	G	GGGG	35	PASS	MSH6	NM_000179.2	c.3802_3803insGGG_p.Met1	2	Ex9	EE	.	.	.	.	.	.	.	.	.
2	5E+07	.	TTGG	T	35	PASS	MSH6	NM_000179.2	c.3170_3172delTGG_p.Leu105	2	Ex4	EE	.	.	.	.	.	.	.	.	.
19	1E+07	.	C	T	35	PASS	LDLR	NM_000527.4	c.1920C>T_p.	3	Ex13	SY	.	.	.	.	.	1179	277244	0.0043	.
19	1E+07	.	T	C	35	PASS	LDLR	NM_000					.	.	.	.	.	1525	276844	0.0055	.
2	5E+07	.	A	AT,ATT	35	PASS	MSH6	NM_000					.	.	.	.	.	.	.	.	.
2	5E+07	.	AT	A	35	PASS	MSH6	NM_000					.	.	.	.	.	.	.	.	.
19	1E+07	.	C	T	35	PASS	LDLR	NM_000					.	.	.	.	.	189	30954	0.0061	.
17	4E+07	.	C	CACA	35	PASS	BRCA1	NM_007294.3	c.5154-2331_5154-2330dupTT	3	In16/19	INT	.	.	.	.	.	.	.	.	.
2	5E+07	.	CTT	C,CT	35	PASS	MSH6	NM_000179.2	c.4002-10delT	3	In3/10	INT	.	.	.	.	.	.	.	.	.
17	4E+07	.	G	A	35	PASS	BRCA1	NM_007294.3	c.-20+11C>T	3	In1/2	5PU	.	.	.	.	.	69	156718	0.0004	.
17	4E+07	.	C	CAT	35	PASS	BRCA1	NM_007294.3	c.-19-22_-19-21dupAT	3	In1/2	5PU	.	.	.	.	.	.	.	.	.
17	4E+07	.	GTTTTT	G	35	PASS	BRCA1	NM_007294.3	c.-20+521_-20+525del5	3	In1/2	5PU	.	.	.	.	.	35	30914	0.0031	0
17	4E+07	.	C	T	35	PASS	BRCA1	NM_007294.3	c.+1332G>A	3	3UTR	3PU	.	.	.	.	.	1496	159758	0.0094	.
19	1E+07	.	TTA	T	35	PASS	LDLR	NM_000527.4	c.+2210_+2211delTA	3	3UTR	3PU	.	.	.	.	.	.	.	.	.
2	2E+07	.	C	A	35	PASS	APOB	NM_000384.2	c.409G>T_p.Glu137X	1	Ex5	SG	34	.	A	.	0.796	.	.	.	.
19	1E+07	.	G	A	35	PASS	LDLR	NM_000527.4	c.1359-1G>A	1	In9/10	ESS	24.8	.	D	.	.	.	.	.	.
2	2E+07	.	A	C	35	PASS	APOB	NM_000384.2	c.4503T>G_p.Tyr1501X	1	Ex26	SG	27.5	.	A	.	0.862	.	.	.	.
7	6E+06	.	G	A	35	PASS	PMS2	NM_000535.6	c.730C>T_p.Gln244X	1	Ex7	SG	43	.	A	.	0.865	.	.	.	.
7	6E+06	.	G	A	35	PASS	PMS2	NM_000535.6	c.730C>T_p.Gln244X	1	Ex7	SG	43	.	A	.	0.865	.	.	.	.
19	1E+07	.	T	G	35	PASS	LDLR	NM_000527.4	c.1942T>G_p.Ser648Ala	2	Ex13	NSY	21.2	D	N	N	0.401	4	277228	1E-05	.

+/23.9kb	Deletion	0	0	0	0
+/44.5kb/	Substituti	5	2	2	0
+/44.5kb/	Substituti	12	2	0	0
+/23.9kb	Insertion	3	1	0	0
+/23.9kb	Deletion	2	2	0	0
+/44.5kb/	Substituti	0	1	0	0
-/81.2kb/	Insertion	1	0	0	0
+/23.9kb	Deletion	0	0	1	0
-/81.2kb/	Substituti	2	2	1	0
-/81.2kb/	Insertion	1	1	0	0
-/81.2kb/	Deletion	0	1	0	0
-/81.2kb/	Substituti	1	0	0	0
+/44.5kb/	Deletion	0	1	0	0
		0	0	0	0
		0	0	0	3
		0	0	0	0
		0	0	0	0
		0	0	0	0
+/44.5kb/	Substituti	0	0	0	1

Gene IDs

Transcripts

Variants

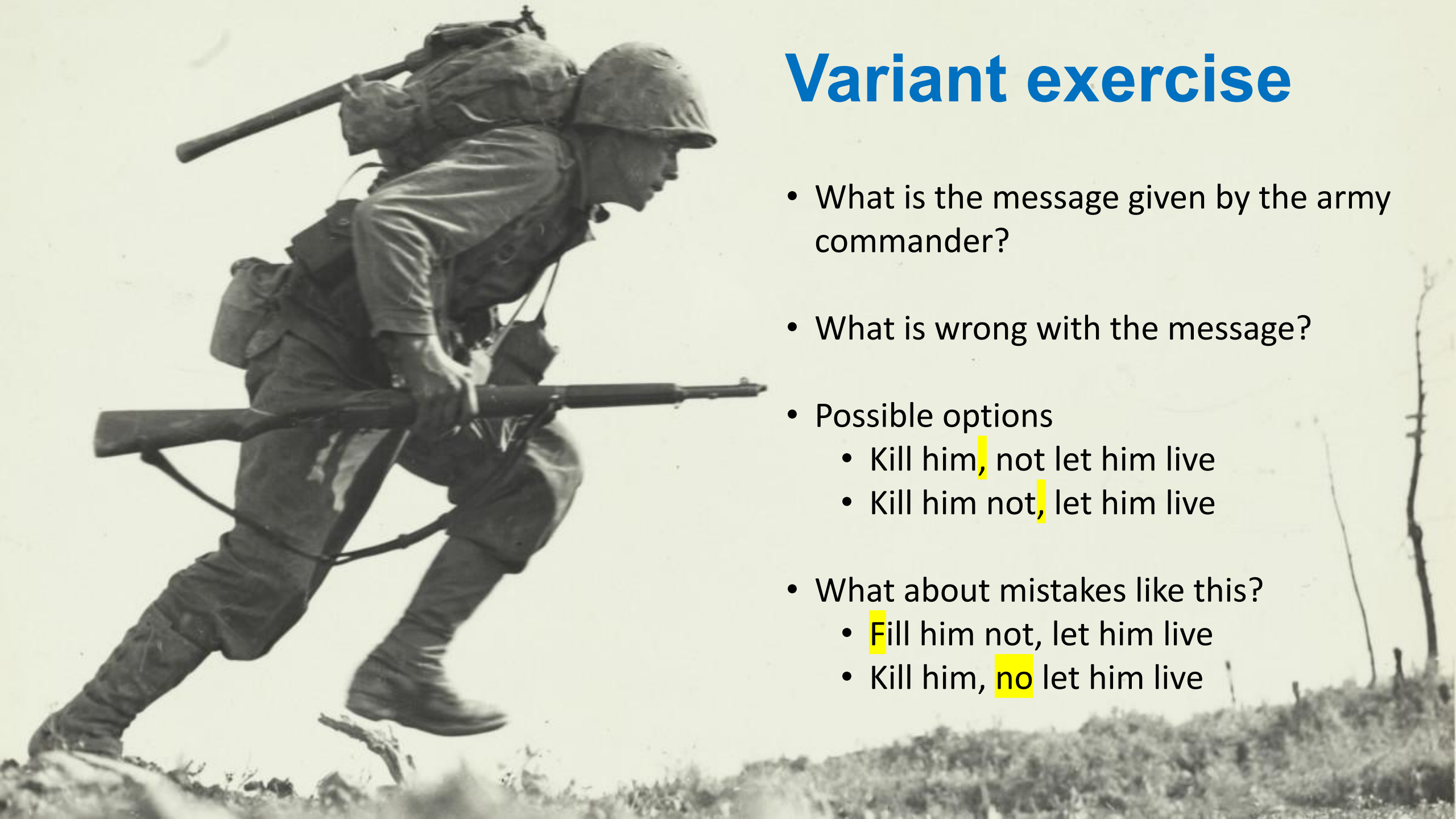
In silico predictions

Allele frequencies

Variant exercise

What kind of message is this from the commander?

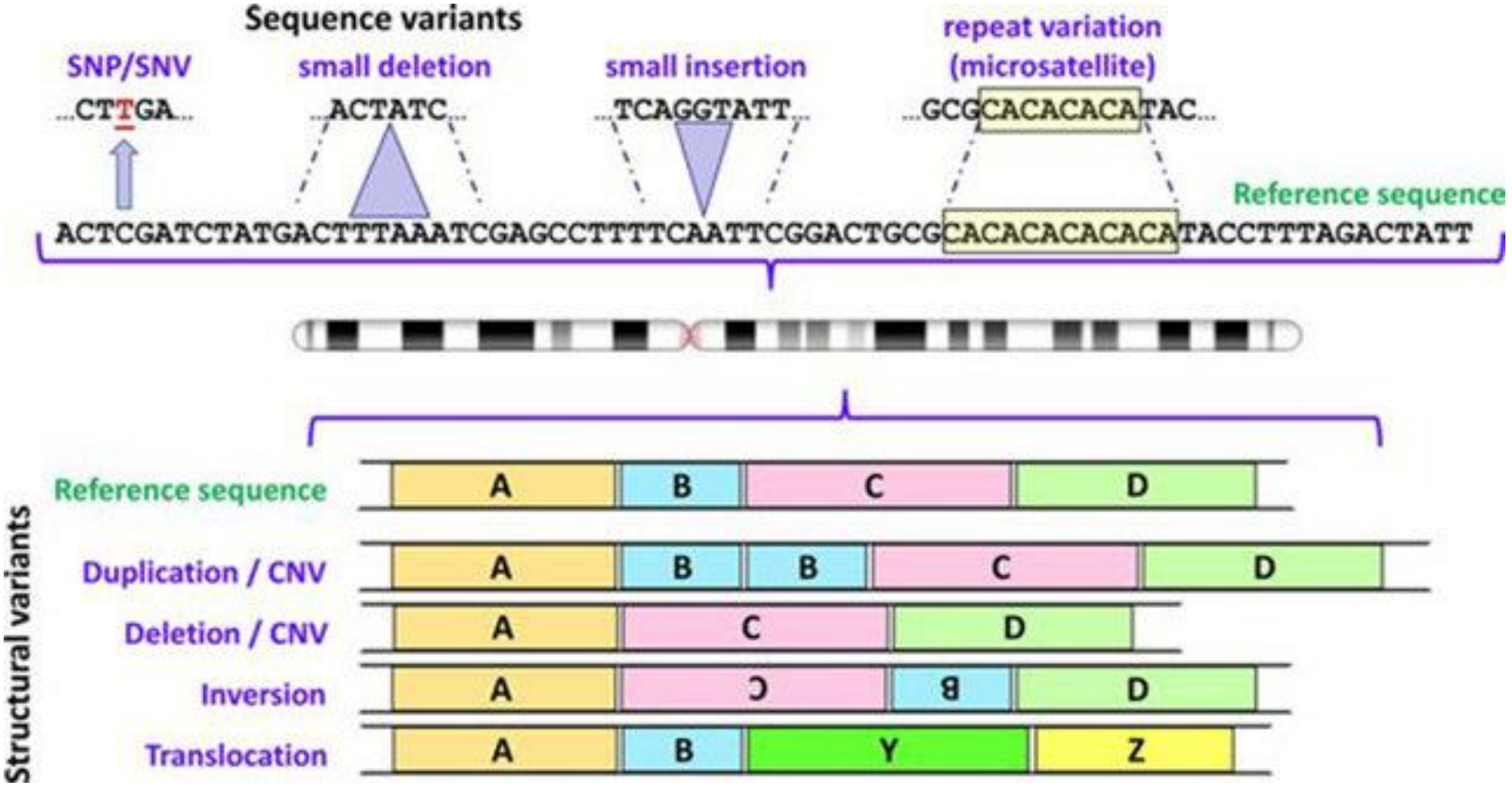
"Kill him not let him live"



Variant exercise

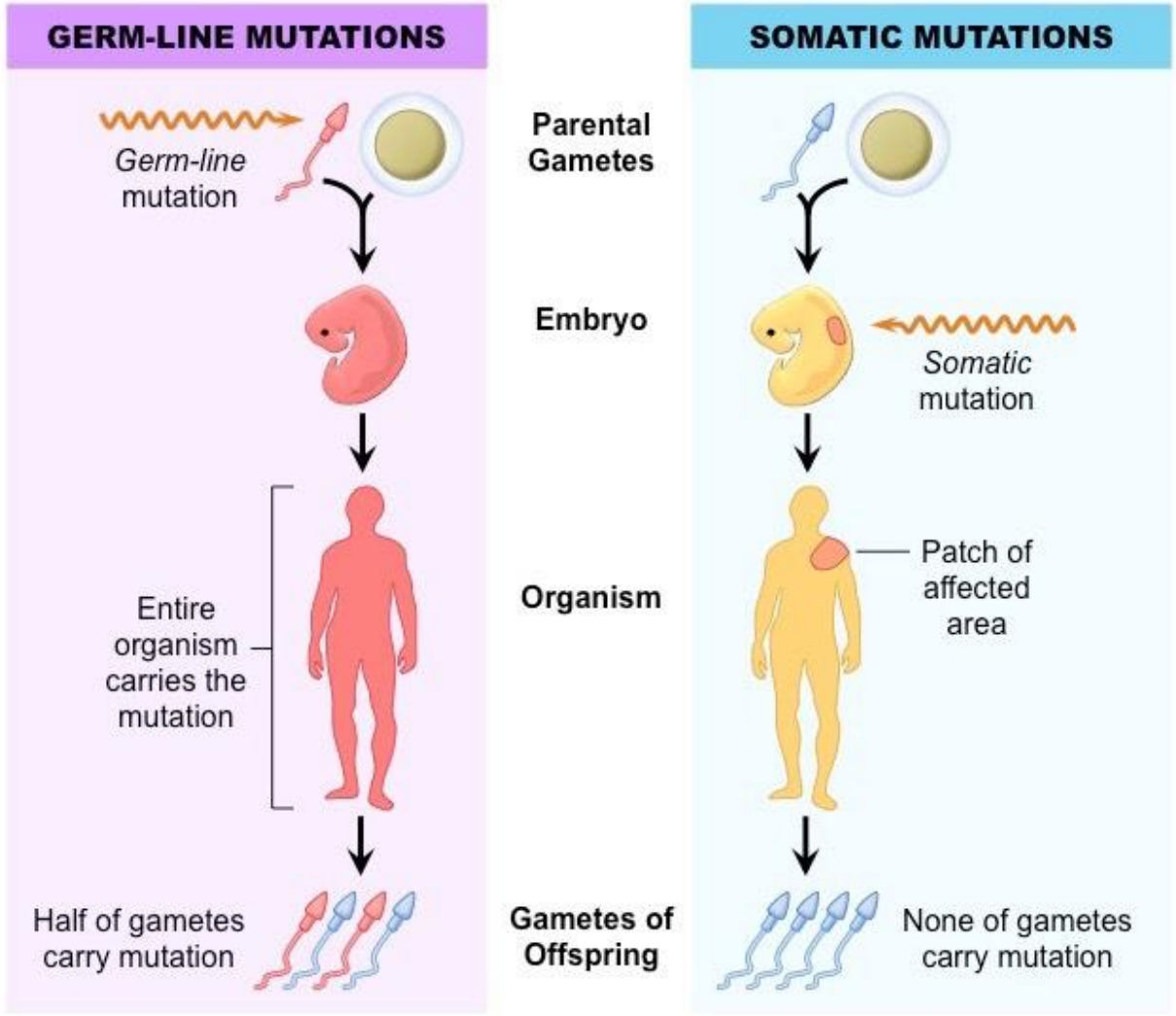
- What is the message given by the army commander?
- What is wrong with the message?
- Possible options
  - Kill him, not let him live
  - Kill him not, let him live
- What about mistakes like this?
  - Fill him not, let him live
  - Kill him, no let him live

Variant types (short and structural)

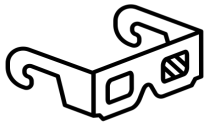


DOI: 10.1042/EBC20170053

Variant types (somatic and germline)



https://ib.bioninja.com.au/standard-level/topic-3-genetics/33-meiosis/somatic-vs-germline-mutatio.html



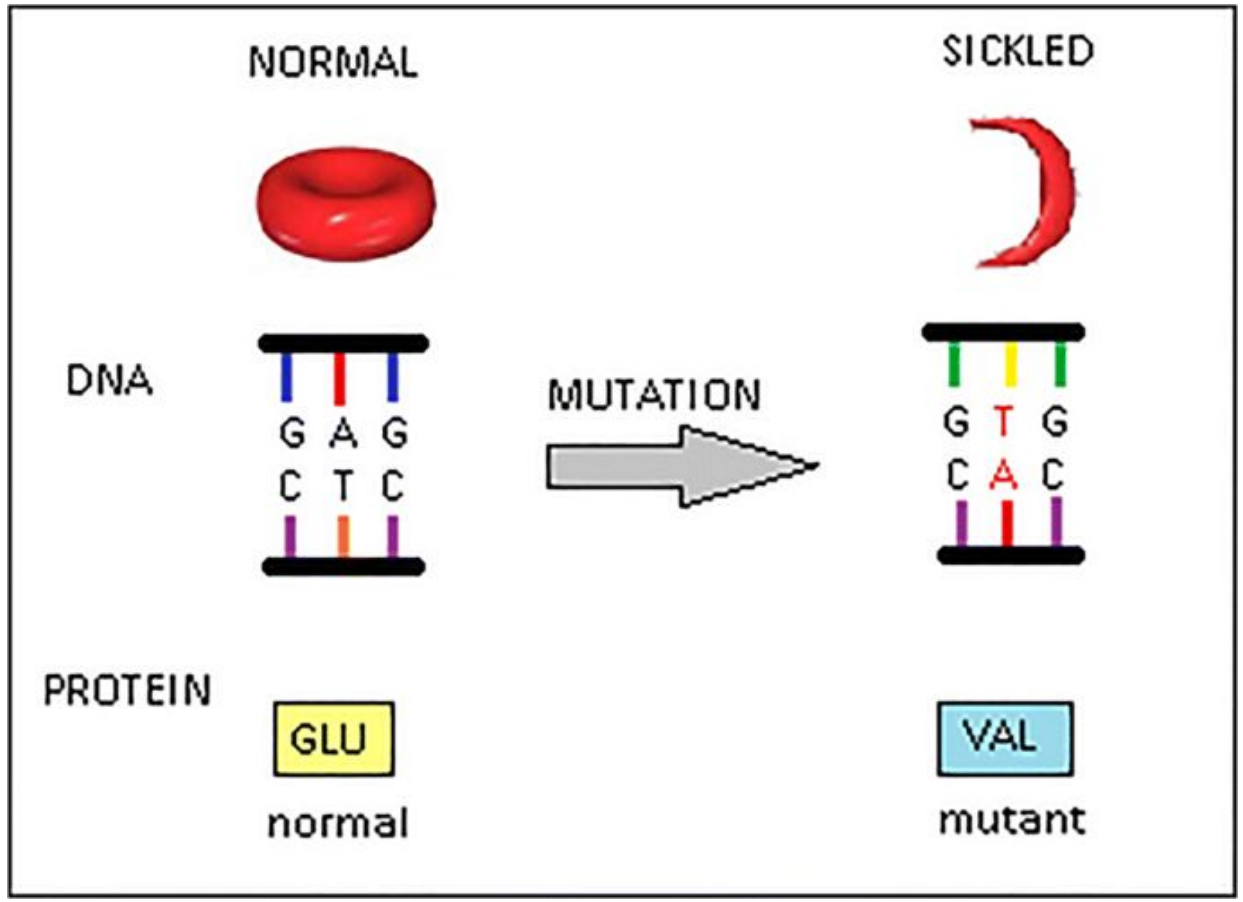
Variant consequences

- Essential role of DNA
  - Preservation and transmission of genetic information

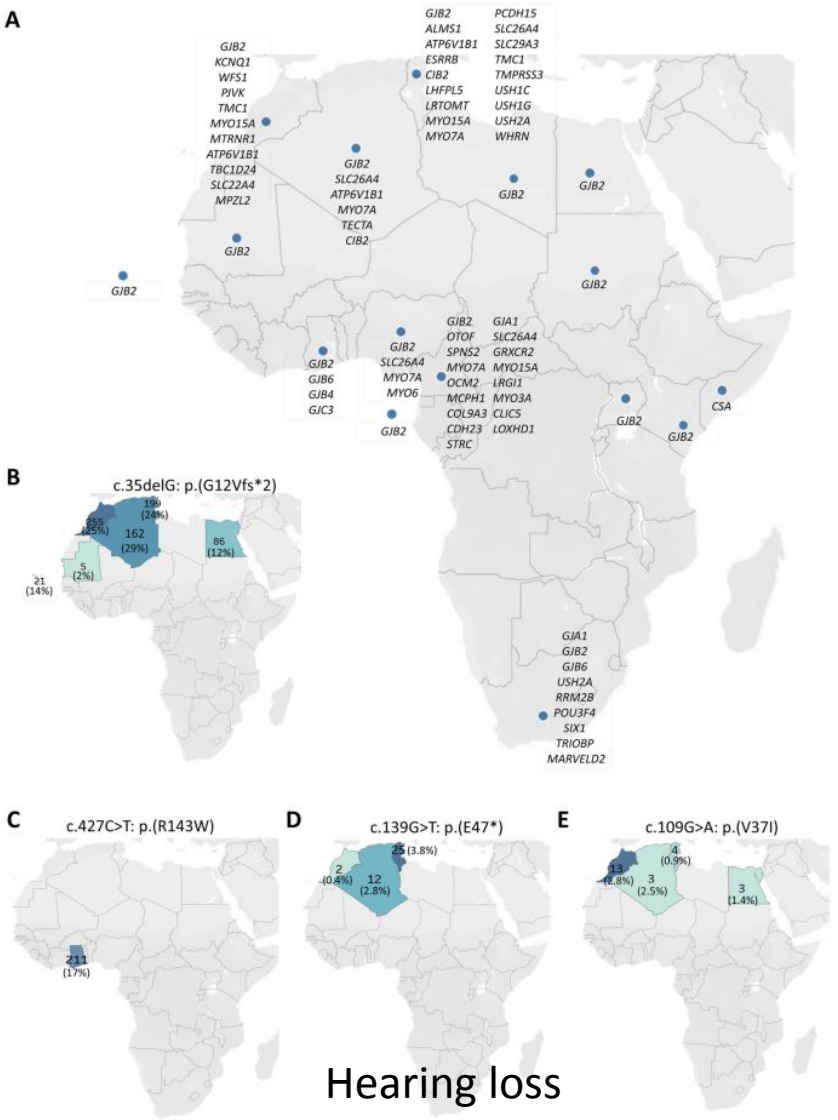
- Effect of variation
  - May cause no change (but may help in population genetics)
  - Alter gene activity or protein function
    - Introduce different traits in an organism

  - Advantageous
    - Helps the individual survive and reproduce, the genetic variation
    - More likely to be passed to the next generation (natural selection).

Examples of SNV diseases



Sickle cell disease



Hearing loss

Examples of structural variation diseases

Type of Disease phenotype	Disease	Structural Variant	Reference
Rare (sporadic) disease	Williams-Beuren Syndrome	Deletion of *ELN* + others	65
	Velo-Cardio-Facial Syndrome	Deletion of *TBX1* + others	66
	Autism	Deletion in 16p11.2	48-50
Rare (Mendelian) disease	Haemophilia A	Inversion disrupting *F8*	67
	Charcot-Marie-Tooth type 1A	Duplication of *PMP22*	36
	Juvenile Nephronophthisis	Deletion of *NPHP1*	37
Common Disease	Psoriasis	Multiallelic CNV of Beta-defensins	43
	Systemic Lupus Erythematosus	Multiallelic CNV of *C4*	44
	Malaria susceptibility	Deletion of alpha-globin	40
	HIV susceptibility	Multi-allelic CNV of *CCL3L1*	41
Pharmacogenetic	Codeine metabolism	Multi-allelic CNV of *CYP2D6*	68
	Carcinogen metabolism	Deletion of *GSTM1*	69

doi: 10.1016/j.tig.2008.03.001

Tools for Pathogenicity Scores

- CADD
- DANN
- DEOGEN2
- EIGEN
- EIGEN PC
- FATHMM
- FATHMM-MKL
- FATHMM-XF
- LIST-S2
- LRT
- M-CAP
- Mutation assessor
- MutationTaster
- MVP
- Polyphen2 HDIV
- Polyphen2 HVAR
- PrimateAI
- PROVEAN
- SIFT

Combined Annotation Dependent Depletion (CADD) Score

- Uses predictions from different annotations of genetic variation
  - Missense
  - Intronic
  - Stop-Gain
  - Insertions
  - Deletions

- Uses Support Vector Machine (SVM) training algorithm based on 63 annotations
  - Conservation metrics
  - Regulatory information
  - Transcript information
  - Protein-level scores)

Sorts Intolerant From Tolerant (SIFT)

- *In silico* prediction tool for nonsynonymous variants
- Based on sequence homology derived from closely related sequences collected through PSI-BLAST
- Values
  - Range 0 to 1
  - Values less than 0.05 usually considered intolerant.

Some useful variant clinical interpretation tools

https://varsome.com/

WGLab/**InterVar**

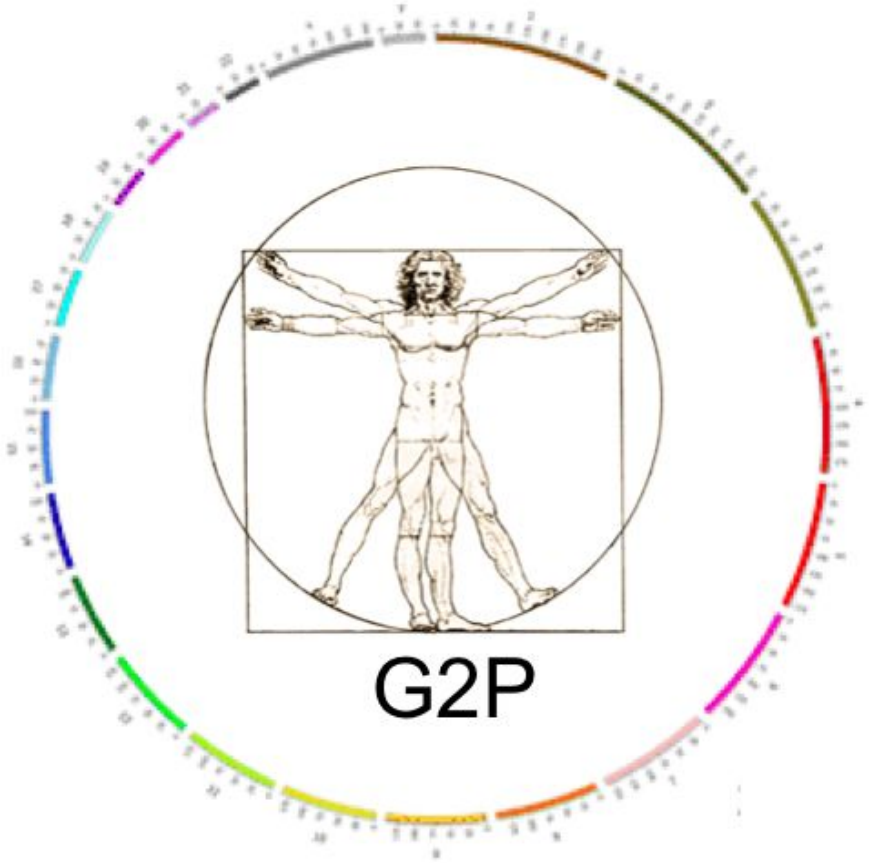
A bioinformatics software tool for clinical interpretation of genetic variants by the 2015 ACMG-AMP guideline

WGLAB

2 Contributors | 44 Issues | 150 Stars | 81 Forks

https://wintervar.wglab.org/

Some useful gene to phenotype tools



OMIM

The Most Authentic Catalogue
of
Human Genes and Associated DiseaseS

https://www.ncbi.nlm.nih.gov/omim

https://omim.org/

https://www.ebi.ac.uk/gene2phenotype

https://www.omim.org/

Browsing Genes and Genomes with Ensembl

EMBL-EBI