05. Functional annotation

Erik Dassi & Davide Bressan

Genomics Technologies Lab

Where are we ?

1. *QC*
2. *Alignment*
3. *Variant calling:* you have obtained a VCF-format file containing all the variants identified in your cell line
4. **Variant annotation**

Variant functional annotation

Provide functional information to allow understanding the properties, genic context, and potential functional effect of a variant

- In which loci does the variant fall? Is it a protein-coding gene ? Is it within a promoter, an UTR, or a coding exon ?

- What happens to the transcript/protein sequence if the variant is present ? Do we get a change in splicing ? A different aminoacid ?

- Is the variant already associated with the onset or progression of a disease (i.e. pathogenic) ?

Variant functional annotation

An (incomplete) list of tools:

- VEP (Variant Effect Predictor, Ensembl)

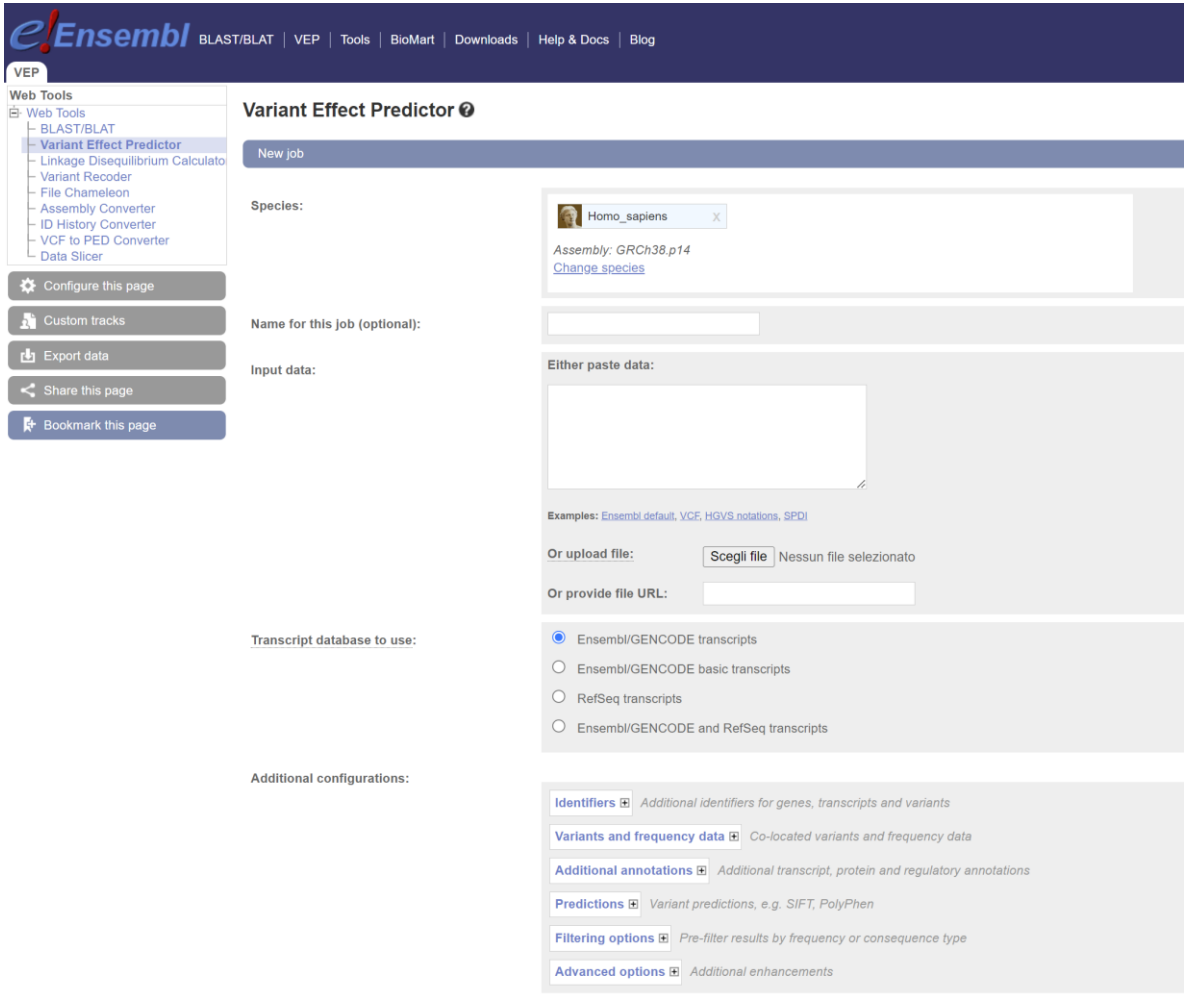
- VAI (Variant Annotation Integrator, UCSC)

- SnpEff

- ANNOVAR

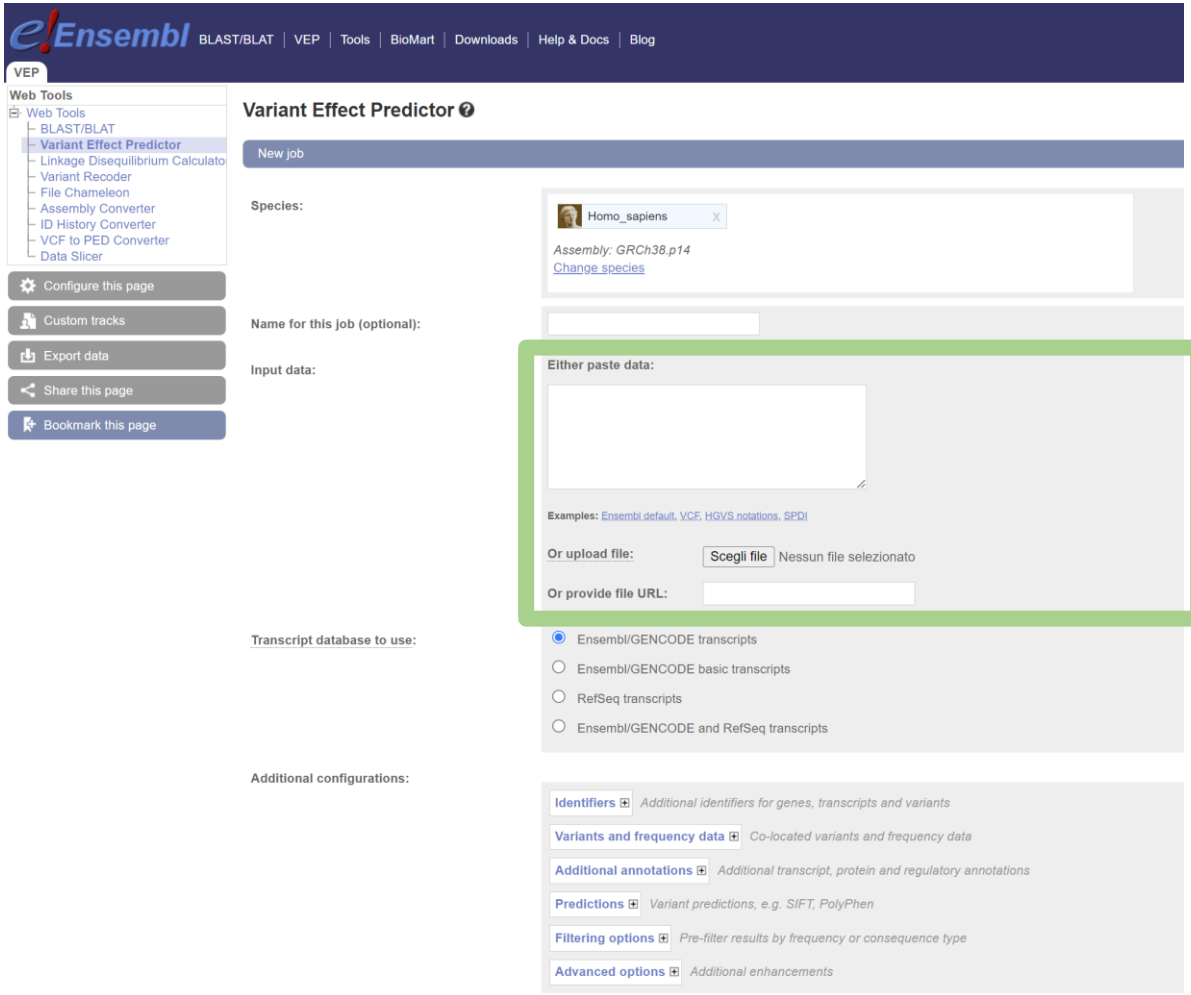
The Variant Effect Predictor (VEP)

https://www.ensembl.org/Tools/VEP



The Variant Effect Predictor (VEP)

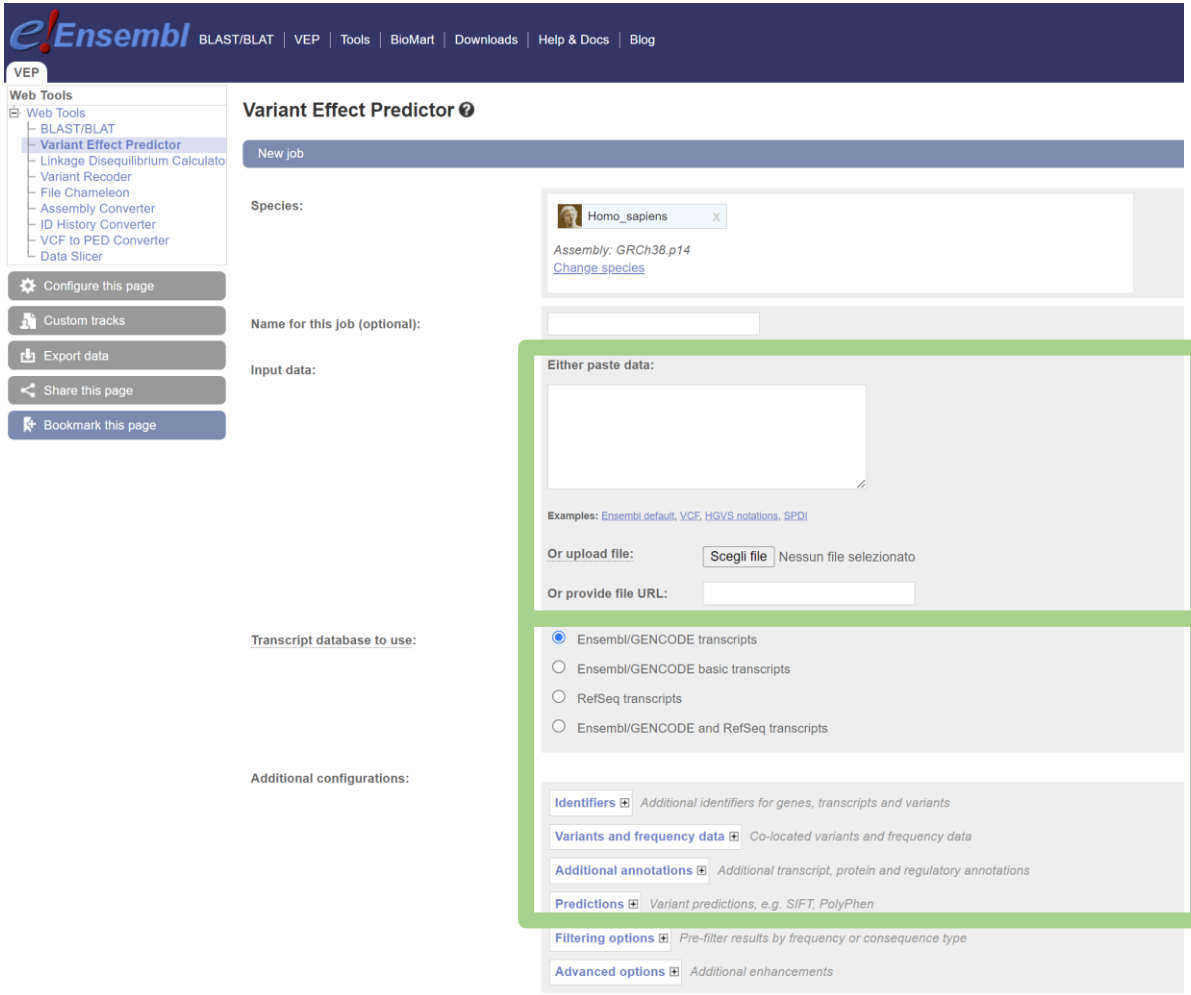
https://www.ensembl.org/Tools/VEP



Your data

The Variant Effect Predictor (VEP)

https://www.ensembl.org/Tools/VEP



Your data

Requested annotation

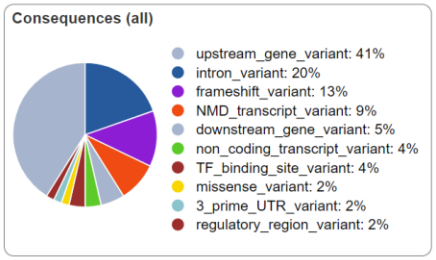
The Variant Effect Predictor (VEP)

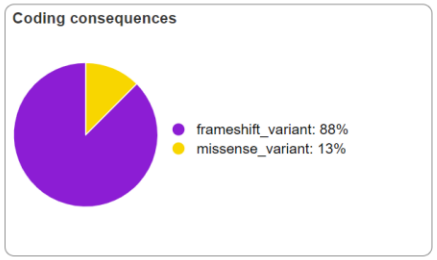
Variant Effect Predictor results

Job details

Summary statistics

Category	Count
Variants processed	3
Variants filtered out	0
Novel / existing variants	-
Overlapped genes	6
Overlapped transcripts	46
Overlapped regulatory features	1





Results preview

Navigation (per variant) — Page: 1 of 1 | Show: 1 All variants

Filters — Uploaded variant | is | defined | Add

Download — All: VCF VEP TXT — BioMart: Variants Genes

New job

Show/hide columns (13 hidden)

Uploaded variant	Location	Allele	Consequence	Symbol	Gene	Feature type	Feature	Biotype	Exon
1_65568_A/C	1:65568-65568	C	downstream_gene_variant	OR4G11P	ENSG00000240361	Transcript	ENST00000492842.2	transcribed_unprocessed_pseudogene	-
1_65568_A/C	1:65568-65568	C	missense_variant	OR4F5	ENSG00000186092	Transcript	ENST00000641515.2	protein_coding	2/3
1_65568_A/C	1:65568-65568	C	downstream_gene_variant	-	ENSG00000290826	Transcript	ENST00000642116.1	lncRNA	-
2_265023_C/T	2:265023-265023	T	intron_variant	ACP1	ENSG00000143727	Transcript	ENST00000272065.10	protein_coding	-
2_265023_C/T	2:265023-265023	T	intron_variant	ACP1	ENSG00000143727	Transcript	ENST00000272067.11	protein_coding	-
2_265023_C/T	2:265023-265023	T	upstream_gene_variant	SH3YL1	ENSG00000035115	Transcript	ENST00000356150.10	protein_coding	-
2_265023_C/T	2:265023-265023	T	upstream_gene_variant	SH3YL1	ENSG00000035115	Transcript	ENST00000402632.5	protein_coding	-
2_265023_C/T	2:265023-265023	T	upstream_gene_variant	SH3YL1	ENSG00000035115	Transcript	ENST00000403657.5	protein_coding	-
2_265023_C/T	2:265023-265023	T	upstream_gene_variant	SH3YL1	ENSG00000035115	Transcript	ENST00000403658.5	protein_coding	-

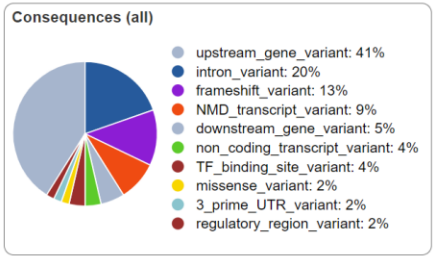
The Variant Effect Predictor (VEP)

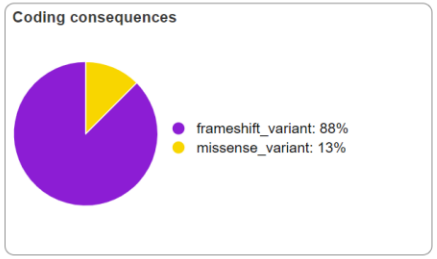
Variant Effect Predictor results

Job details

Summary statistics

Category	Count
Variants processed	3
Variants filtered out	0
Novel / existing variants	-
Overlapped genes	6
Overlapped transcripts	46
Overlapped regulatory features	1





Results preview

Navigation (per variant) — Page: 1 of 1 | Show: 1 All variants

Filters — Uploaded variant | is | defined | Add

Download — All: VCF VEP TXT — BioMart: Variants Genes

New job

Show/hide columns (13 hidden)

Uploaded variant	Location	Allele	Consequence	Symbol	Gene	Feature type	Feature	Biotype	Exon
1_65568_A/C	1:65568-65568	C	downstream_gene_variant	OR4G11P	ENSG00000240361	Transcript	ENST00000492842.2	transcribed_unprocessed_pseudogene	-
1_65568_A/C	1:65568-65568	C	missense_variant	OR4F5	ENSG00000186092	Transcript	ENST00000641515.2	protein_coding	2/3
1_65568_A/C	1:65568-65568	C	downstream_gene_variant	-	ENSG00000290826	Transcript	ENST00000642116.1	lncRNA	-
2_265023_C/T	2:265023-265023	T	intron_variant	ACP1	ENSG00000143727	Transcript	ENST00000272065.10	protein_coding	-
2_265023_C/T	2:265023-265023	T	intron_variant	ACP1	ENSG00000143727	Transcript	ENST00000272067.11	protein_coding	-
2_265023_C/T	2:265023-265023	T	upstream_gene_variant	SH3YL1	ENSG00000035115	Transcript	ENST00000356150.10	protein_coding	-
2_265023_C/T	2:265023-265023	T	upstream_gene_variant	SH3YL1	ENSG00000035115	Transcript	ENST00000402632.5	protein_coding	-
2_265023_C/T	2:265023-265023	T	upstream_gene_variant	SH3YL1	ENSG00000035115	Transcript	ENST00000403657.5	protein_coding	-
2_265023_C/T	2:265023-265023	T	upstream_gene_variant	SH3YL1	ENSG00000035115	Transcript	ENST00000403658.5	protein_coding	-

Consequence

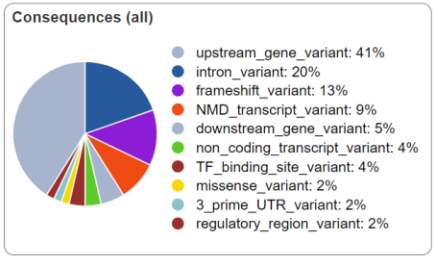
The Variant Effect Predictor (VEP)

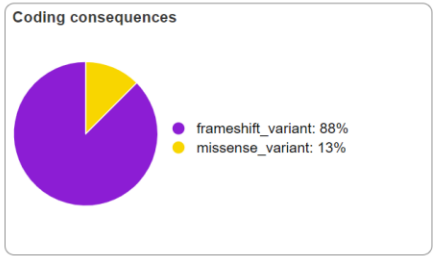
Variant Effect Predictor results

Job details

Summary statistics

Category	Count
Variants processed	3
Variants filtered out	0
Novel / existing variants	-
Overlapped genes	6
Overlapped transcripts	46
Overlapped regulatory features	1





Results preview

Navigation (per variant) Page: 1 of 1 | Show: 1 All variants

Filters: Uploaded variant | is | defined | Add

Download All: VCF VEP TXT BioMart: Variants Genes

New job

Show/hide columns (13 hidden)

Uploaded variant	Location	Allele	Consequence	Symbol	Gene	Feature type	Feature	Biotype	Exon
1_65568_A/C	1:65568-65568	C	downstream_gene_variant	OR4G11P	ENSG00000240361	Transcript	ENST00000492842.2	transcribed_unpr…essed_pseudogene	-
1_65568_A/C	1:65568-65568	C	missense_variant	OR4F5	ENSG00000186092	Transcript	ENST00000641515.2	protein_coding	2/3
1_65568_A/C	1:65568-65568	C	downstream_gene_variant	-	ENSG00000290826	Transcript	ENST00000642116.1	lncRNA	-
2_265023_C/T	2:265023-265023	T	intron_variant	ACP1	ENSG00000143727	Transcript	ENST00000272065.10	protein_coding	-
2_265023_C/T	2:265023-265023	T	intron_variant	ACP1	ENSG00000143727	Transcript	ENST00000272067.11	protein_coding	-
2_265023_C/T	2:265023-265023	T	upstream_gene_variant	SH3YL1	ENSG00000035115	Transcript	ENST00000356150.10	protein_coding	-
2_265023_C/T	2:265023-265023	T	upstream_gene_variant	SH3YL1	ENSG00000035115	Transcript	ENST00000402632.5	protein_coding	-
2_265023_C/T	2:265023-265023	T	upstream_gene_variant	SH3YL1	ENSG00000035115	Transcript	ENST00000403657.5	protein_coding	-
2_265023_C/T	2:265023-265023	T	upstream_gene_variant	SH3YL1	ENSG00000035115	Transcript	ENST00000403658.5	protein_coding	-

Consequence → **Gene / transcript annotation**

The Variant Effect Predictor (VEP)

APPRIS	SIFT	PolyPhen
-	-	-
P1	0.06	0
-	-	-
P3	-	-
A1	-	-
P1	-	-
-	-	-
-	-	-
-	-	-

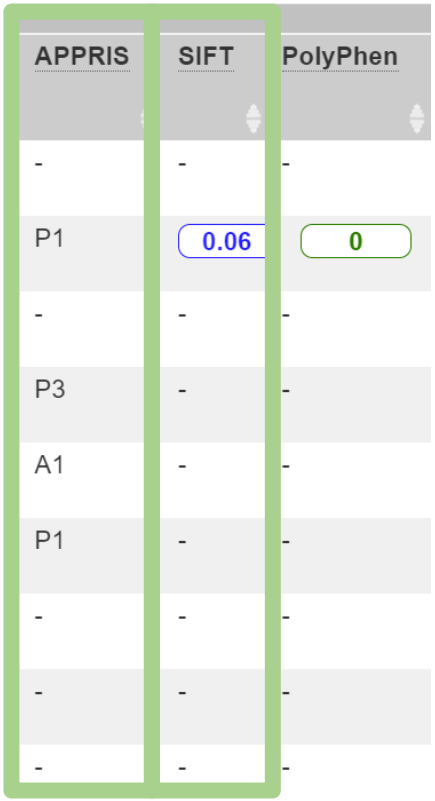
The Variant Effect Predictor (VEP)

APPRIS:
identifies functionally important transcripts (Primary 1 -> 5, Alternative 1 -> 2)

APPRIS	SIFT	PolyPhen
-	-	-
P1	0.06	0
-	-	-
P3	-	-
A1	-	-
P1	-	-
-	-	-
-	-	-
-	-	-

The Variant Effect Predictor (VEP)

APPRIS:
identifies functionally important transcripts (Primary 1 -> 5, Alternative 1 -> 2)



APPRIS	SIFT	PolyPhen
-	-	-
P1	0.06	0
-	-	-
P3	-	-
A1	-	-
P1	-	-
-	-	-
-	-	-
-	-	-

SIFT:
predicts whether an aminoacid substitution affects protein function by homology
[0-1, the lower the worse]

The Variant Effect Predictor (VEP)

APPRIS	SIFT	PolyPhen
-	-	-
P1	0.06	0
-	-	-
P3	-	-
A1	-	-
P1	-	-
-	-	-
-	-	-
-	-	-

APPRIS:
identifies functionally important transcripts (Primary 1 -> 5, Alternative 1 -> 2)

PolyPhen:
Predicts the impact of substitution based on sequence, phylogenetic and structural information [0-1, the lower the worse]

SIFT:
predicts whether an aminoacid substitution affects protein function by homology [0-1, the lower the worse]

Variant databases

An (incomplete) list of databases:

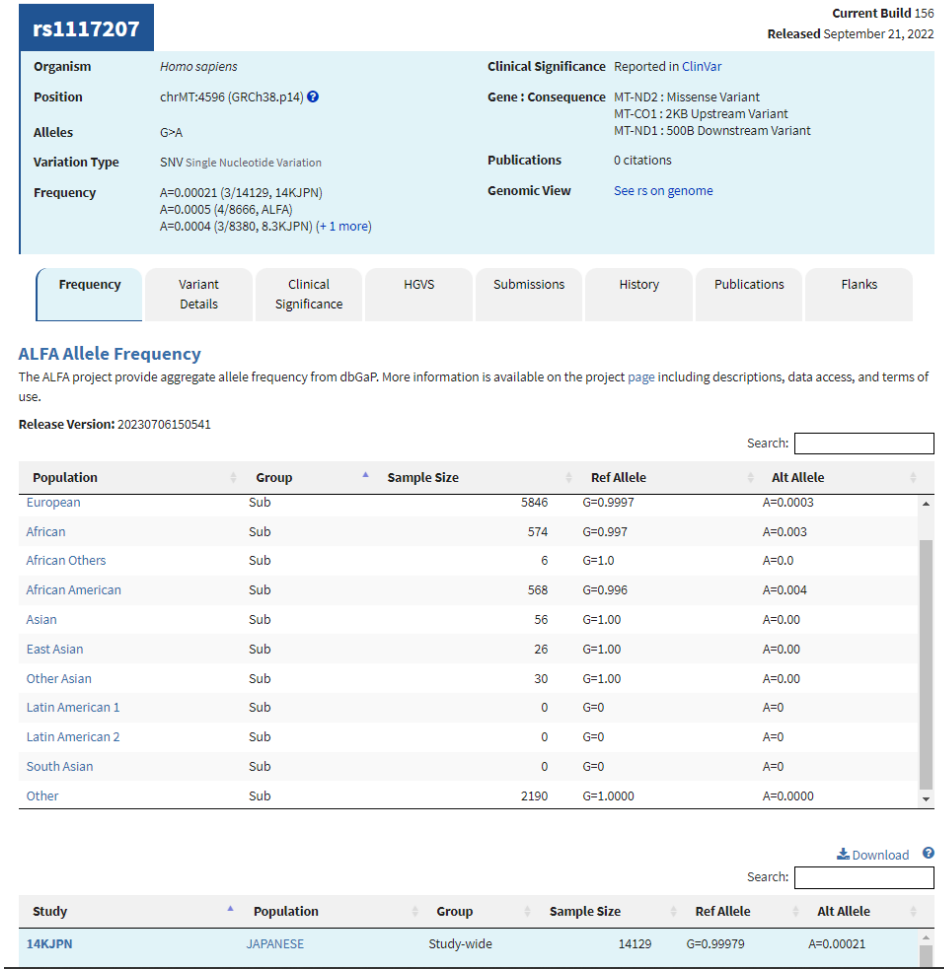
- dbSNP (NCBI)

- ClinVar (variants of health relevance)

- OMIM (Online Mendelian Inheritance in Man)

dbSNP

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



rs1117207

Current Build 156
Released September 21, 2022

Organism	*Homo sapiens*	**Clinical Significance**	Reported in ClinVar
Position	chrMT:4596 (GRCh38.p14)	**Gene : Consequence**	MT-ND2 : Missense Variant MT-CO1 : 2KB Upstream Variant MT-ND1 : 500B Downstream Variant
Alleles	G>A		
Variation Type	SNV Single Nucleotide Variation	**Publications**	0 citations
Frequency	A=0.00021 (3/14129, 14KJPN) A=0.0005 (4/8666, ALFA) A=0.0004 (3/8380, 8.3KJPN) (+ 1 more)	**Genomic View**	See rs on genome

Frequency | Variant Details | Clinical Significance | HGVS | Submissions | History | Publications | Flanks

ALFA Allele Frequency

The ALFA project provide aggregate allele frequency from dbGaP. More information is available on the project page including descriptions, data access, and terms of use.

Release Version: 20230706150541

Search:

Population	Group	Sample Size	Ref Allele	Alt Allele
European	Sub	5846	G=0.9997	A=0.0003
African	Sub	574	G=0.997	A=0.003
African Others	Sub	6	G=1.0	A=0.0
African American	Sub	568	G=0.996	A=0.004
Asian	Sub	56	G=1.00	A=0.00
East Asian	Sub	26	G=1.00	A=0.00
Other Asian	Sub	30	G=1.00	A=0.00
Latin American 1	Sub	0	G=0	A=0
Latin American 2	Sub	0	G=0	A=0
South Asian	Sub	0	G=0	A=0
Other	Sub	2190	G=1.0000	A=0.0000

Download

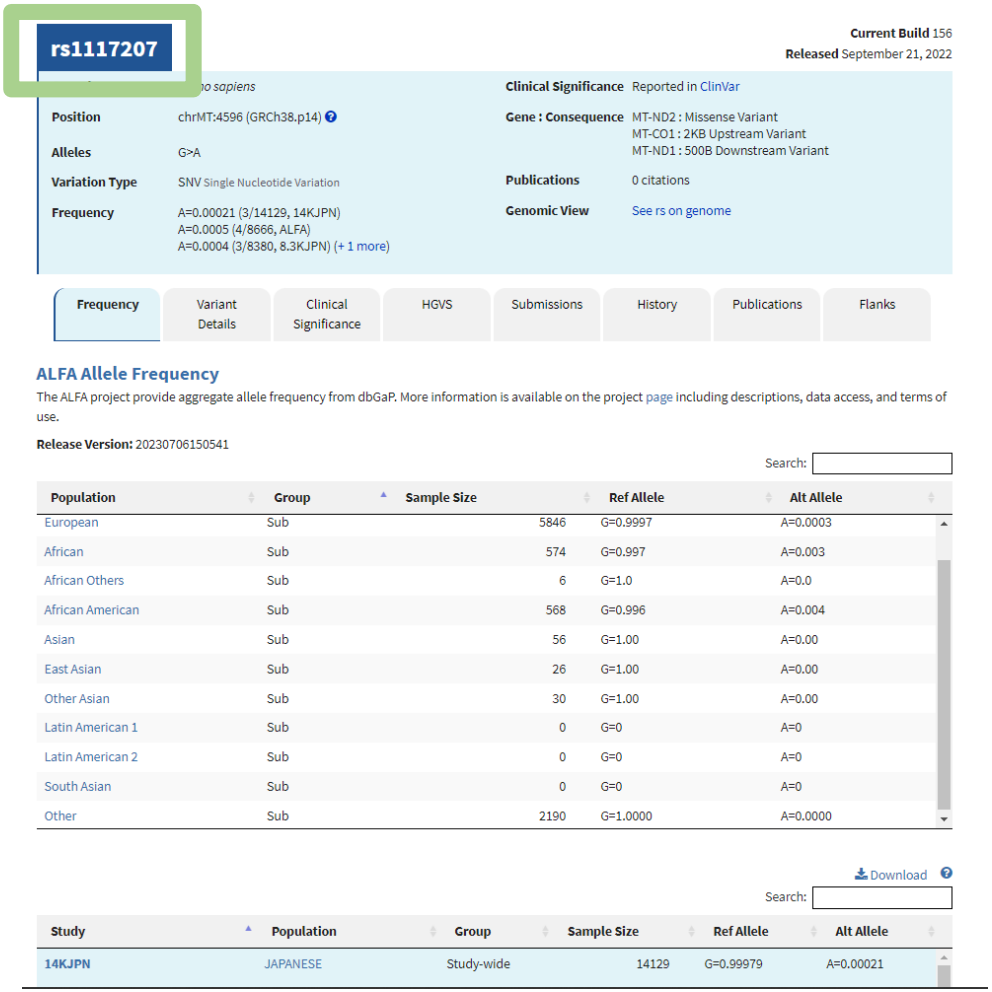
Search:

Study	Population	Group	Sample Size	Ref Allele	Alt Allele
14KJPN	JAPANESE	Study-wide	14129	G=0.99979	A=0.00021

dbSNP

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

Variant ID

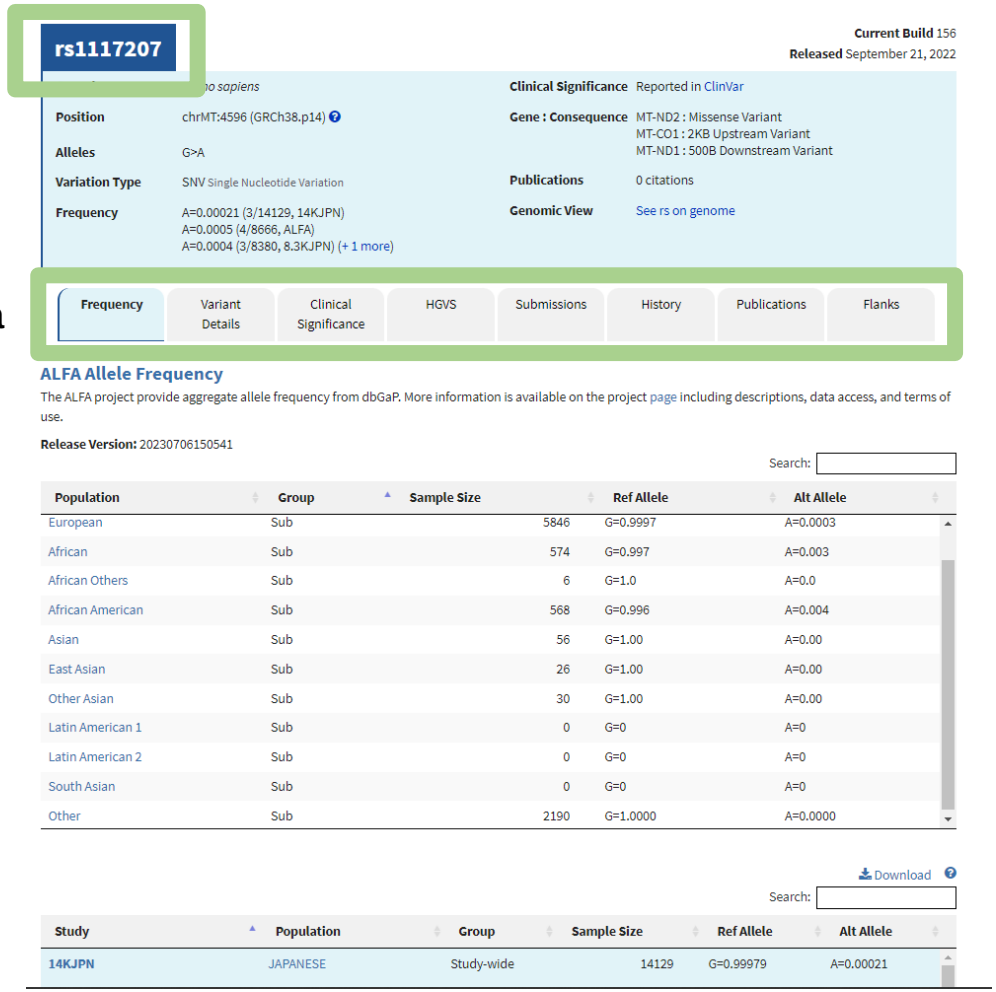


dbSNP

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

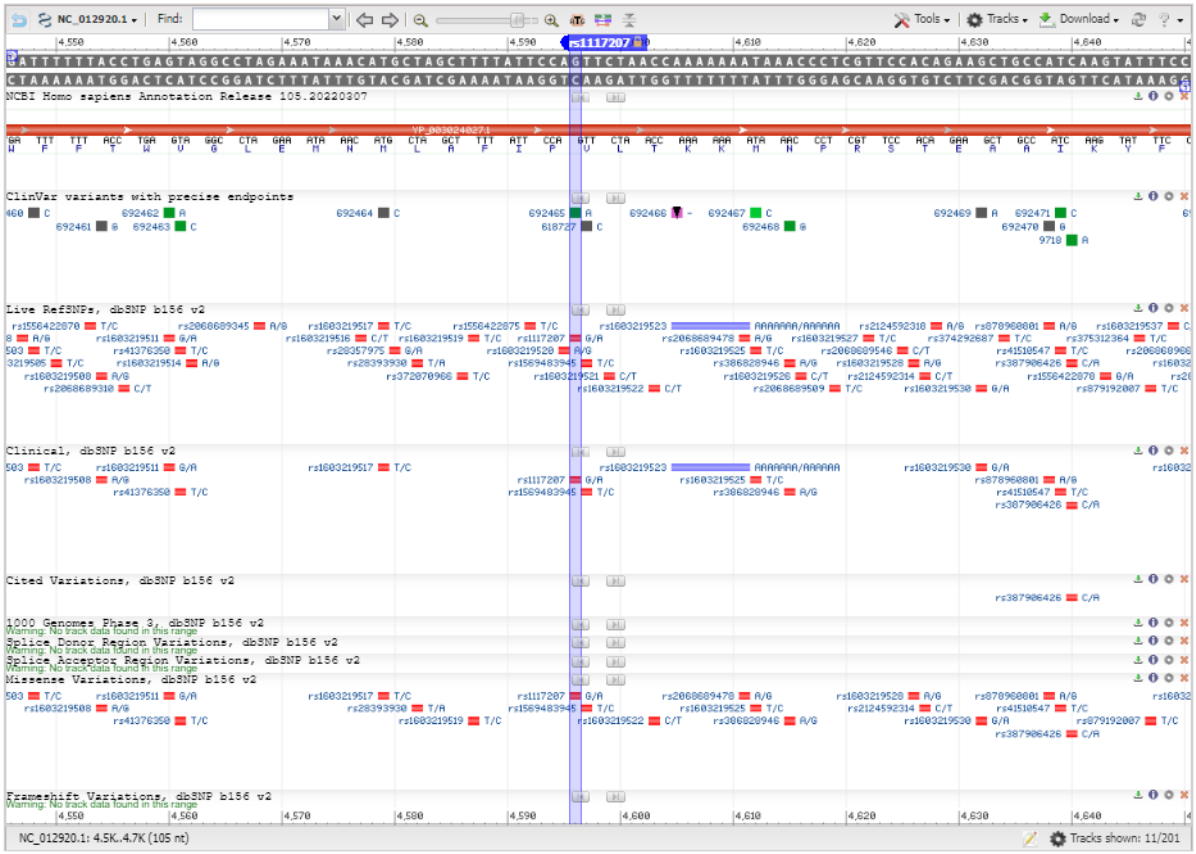
Variant ID

Annotation and data



dbSNP

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



What should you do?

1. Run VEP on your VCF file
2. Explore the results, also checking with dbSNP / OMIM
3. Download the resulting annotated VCF for later use
4. Upload it to the Google Drive folder:

https://drive.google.com/drive/folders/1UfVf1y1gLiV0XOwdDasJHnRZss5-Mte_?usp=sharing

Questions ?