Variant data analysis and prioritization using HGVA

TrainMALTA

Cambridge, UK

2nd June 2017

Marta Bleda Latorre

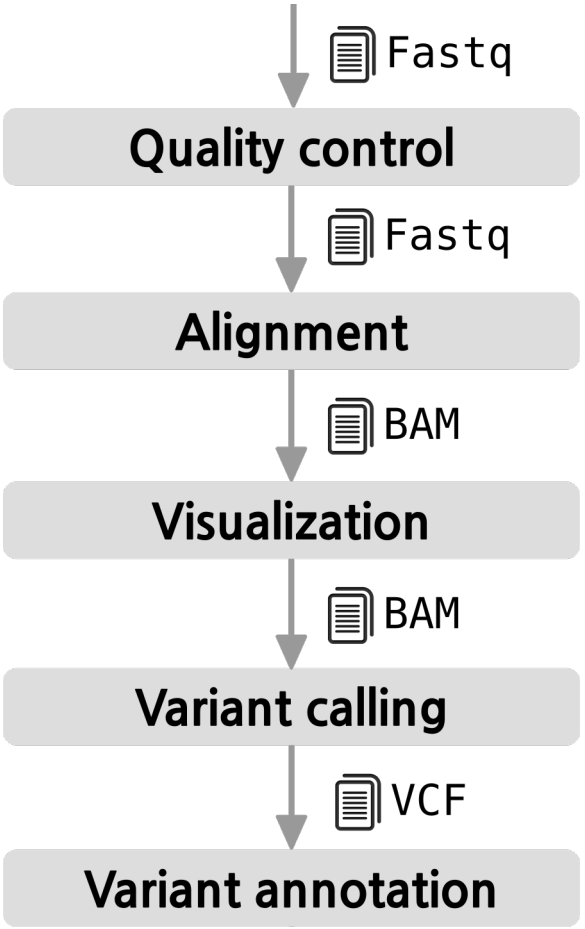
mb2033@cam.ac.uk

Research Associate at the Department of Medicine

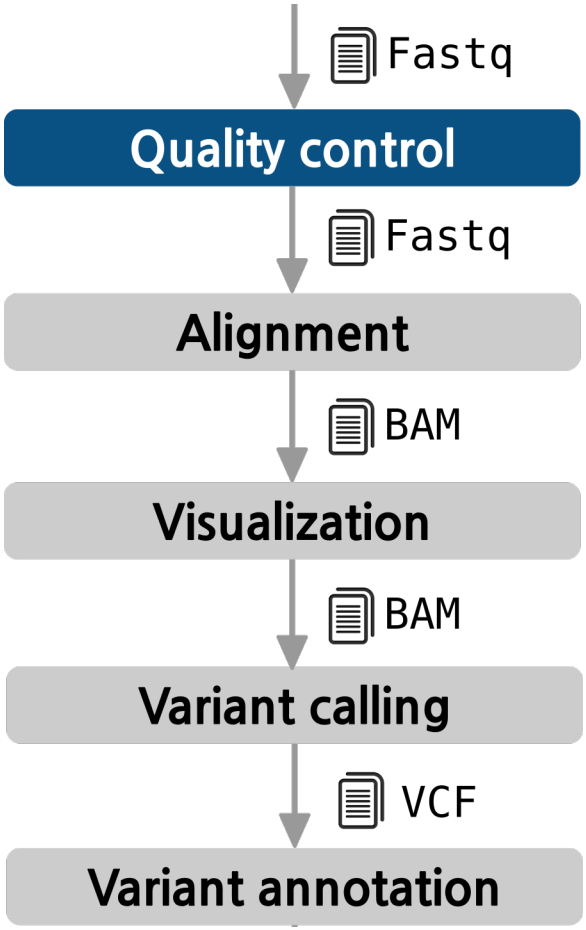
University of Cambridge

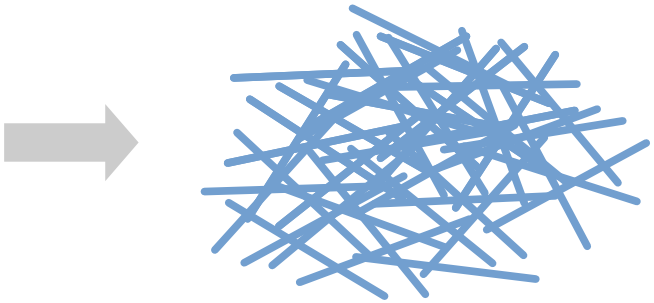
Cambridge, UK

The pipeline

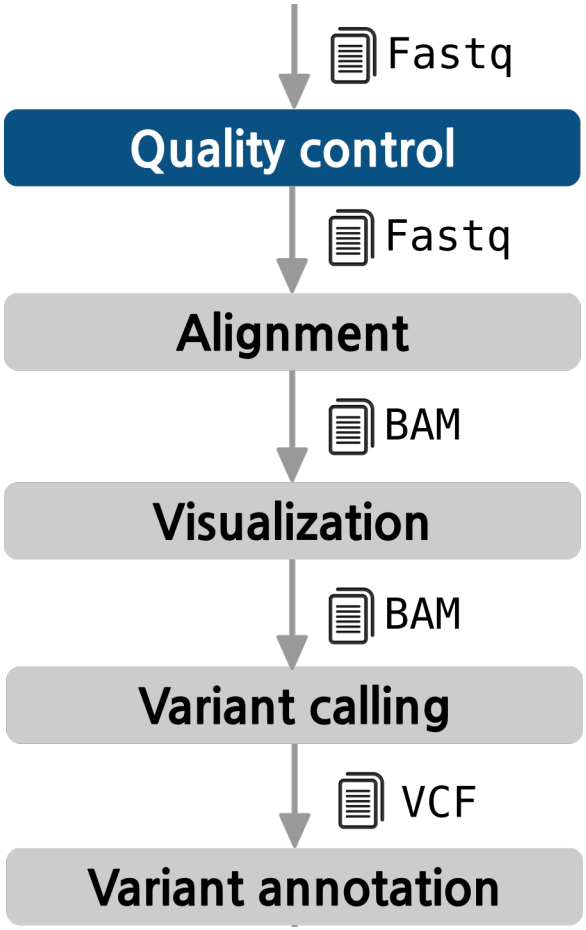


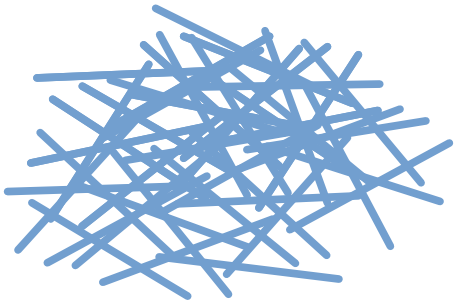
The pipeline

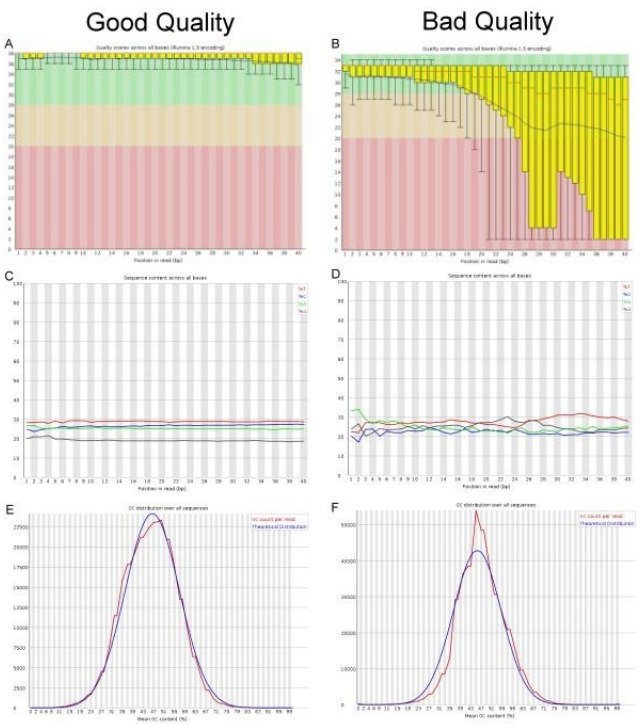


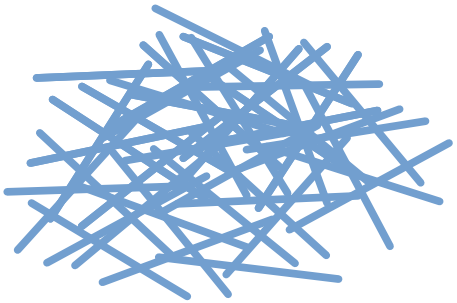


The pipeline

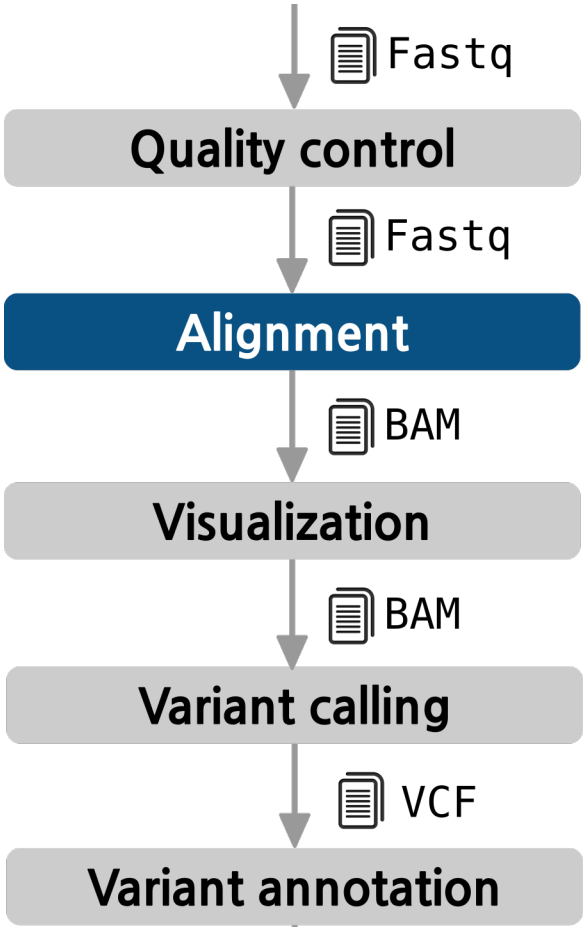


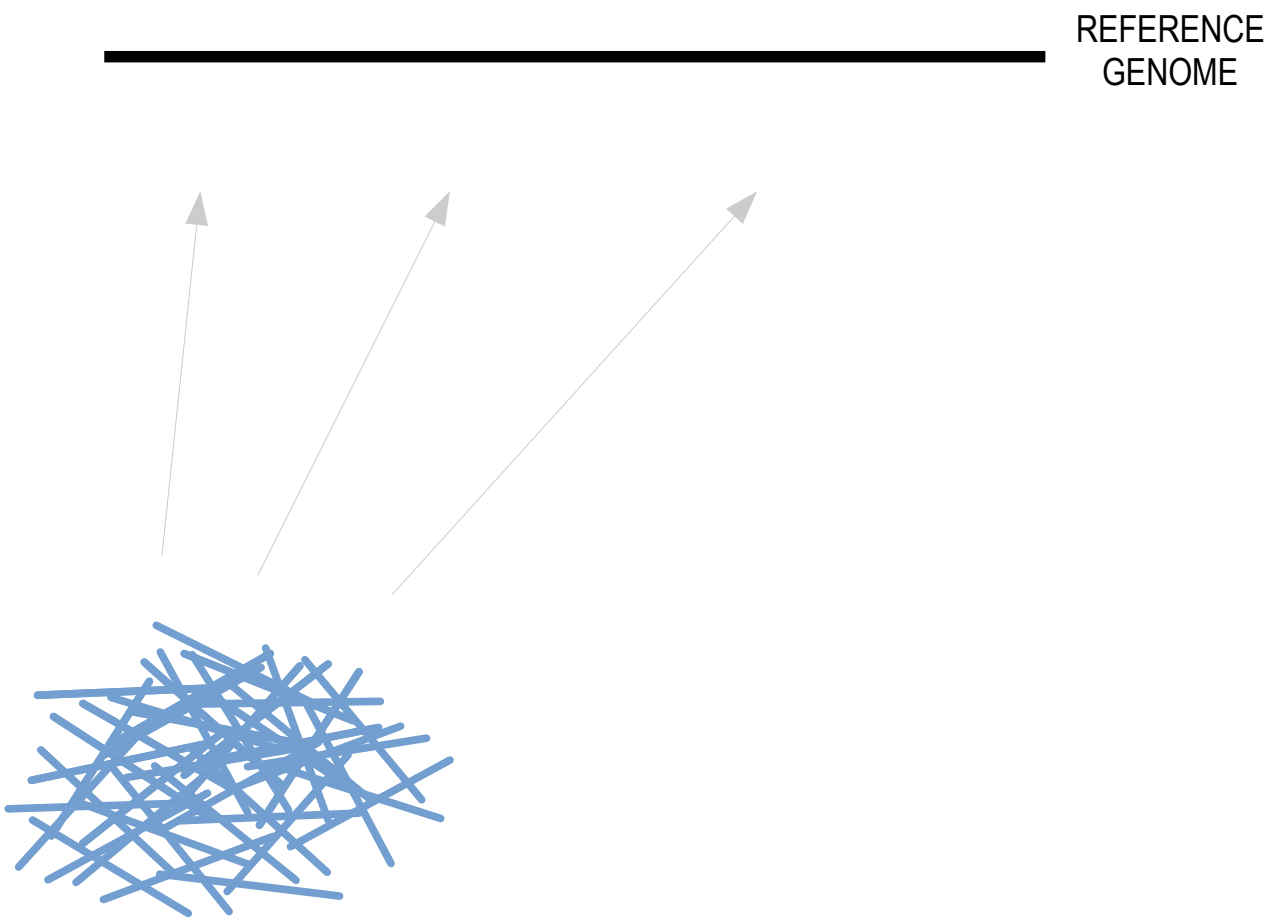




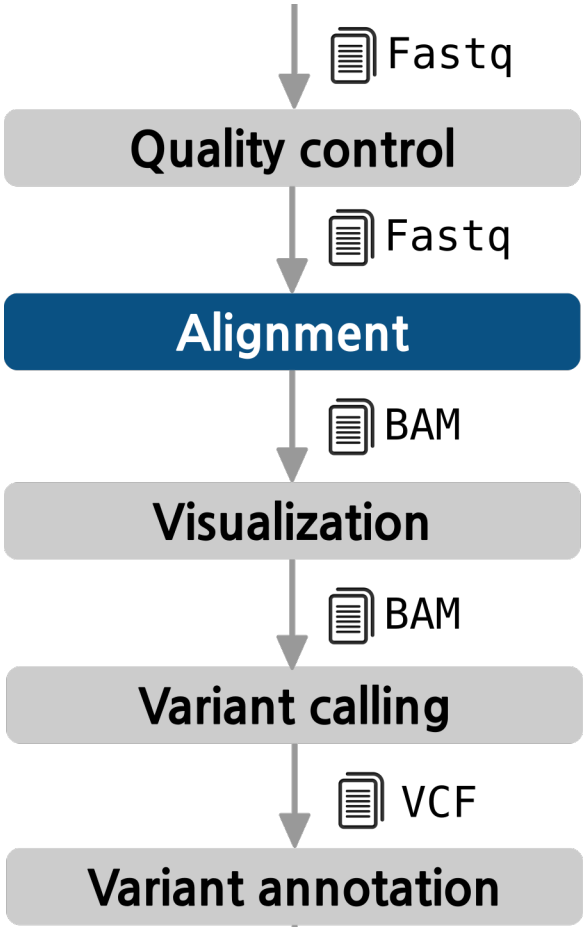


The pipeline

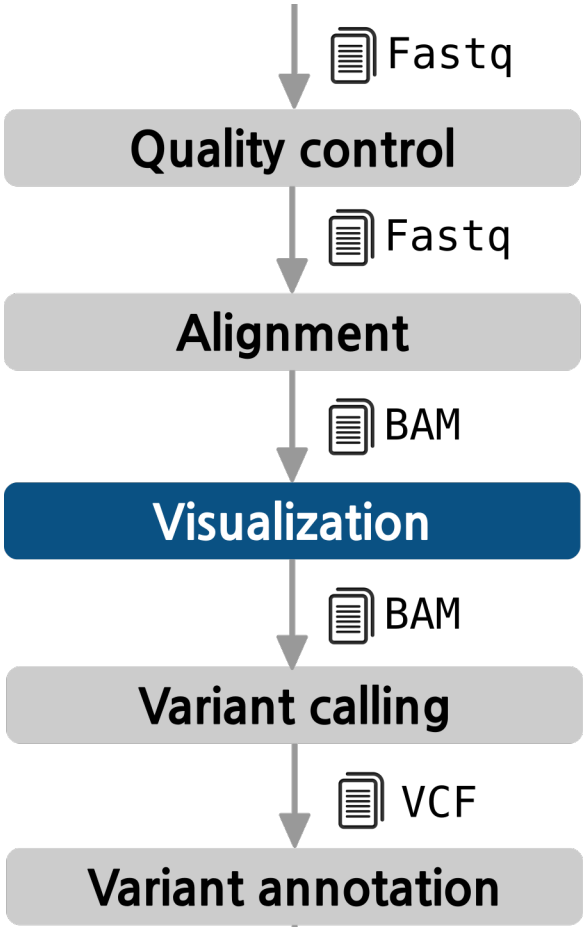


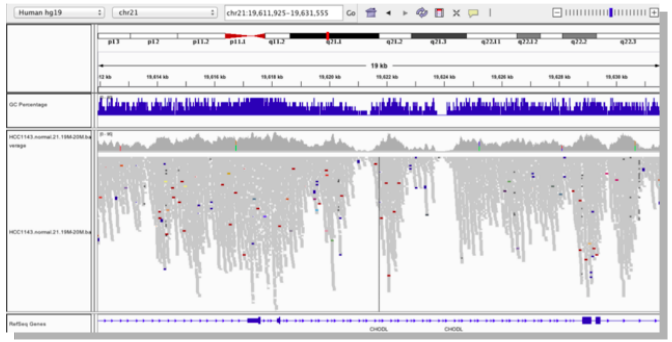


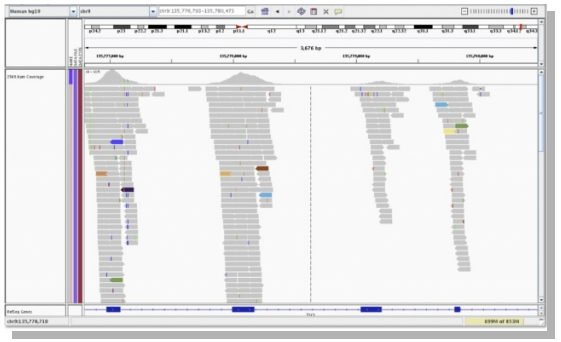
The pipeline

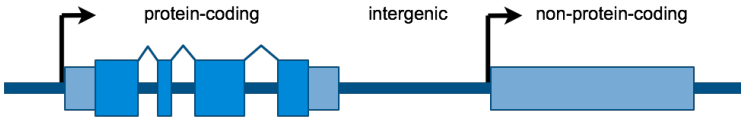


The pipeline





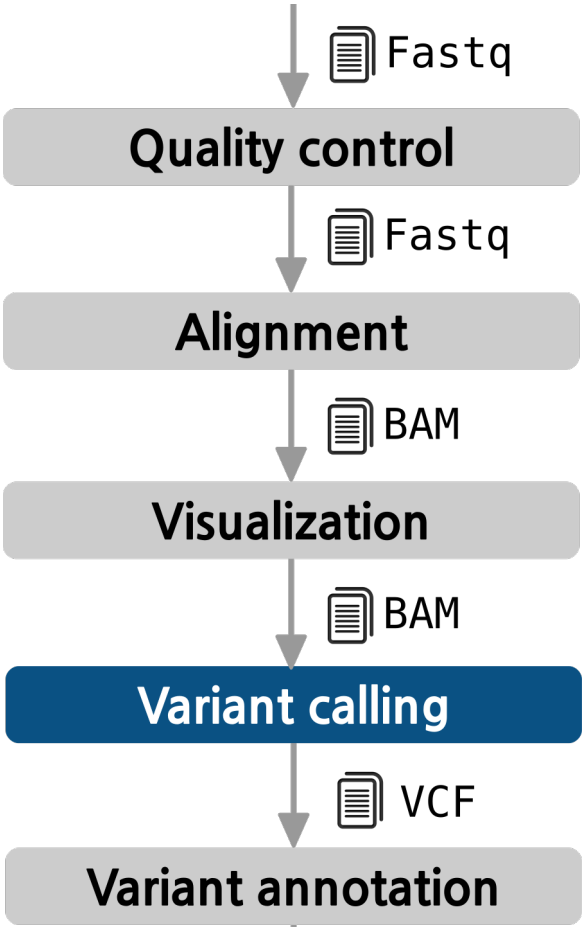


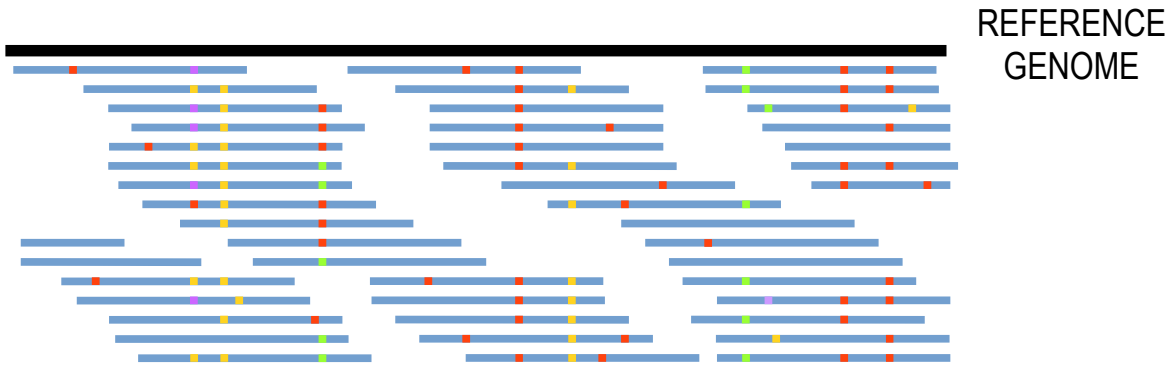


Regions covered by WES (64 MB, 2%)

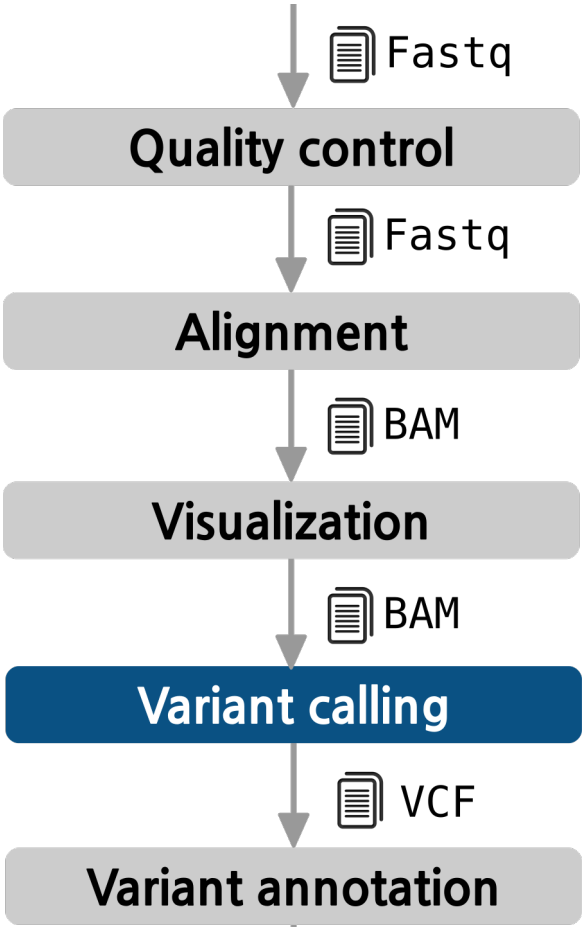
Regions covered by WGS (~3000 MB, 98%)

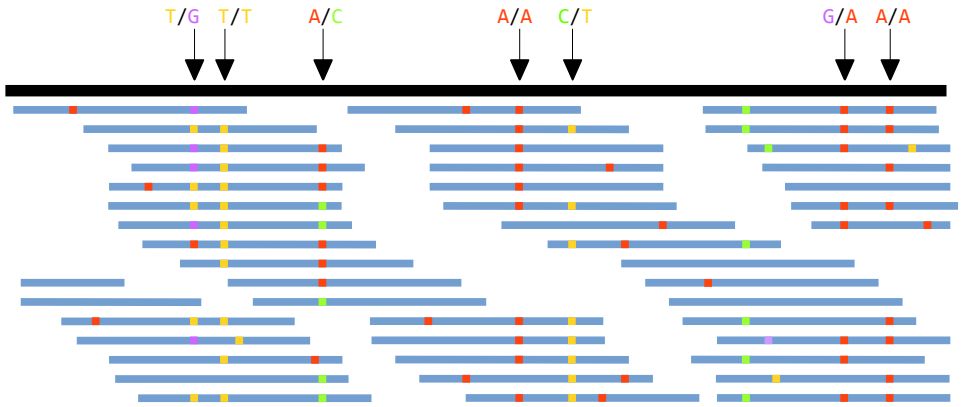
The pipeline





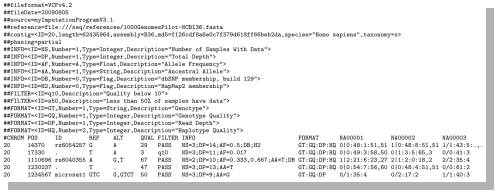
The pipeline

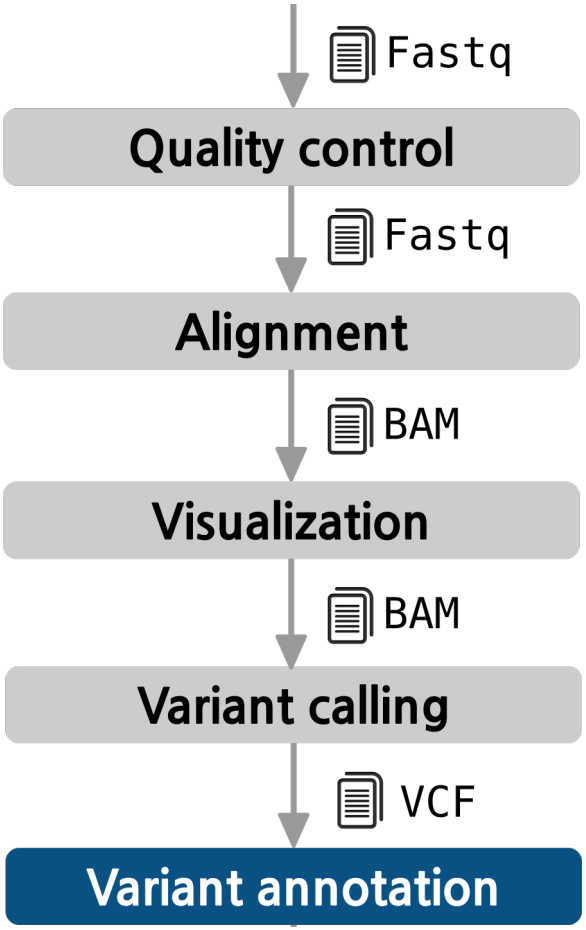


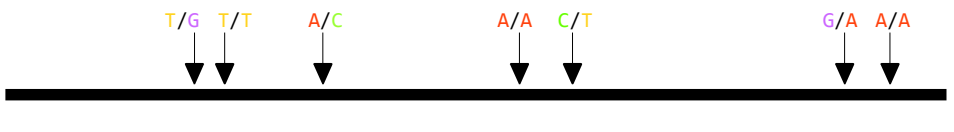



```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number 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        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
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|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

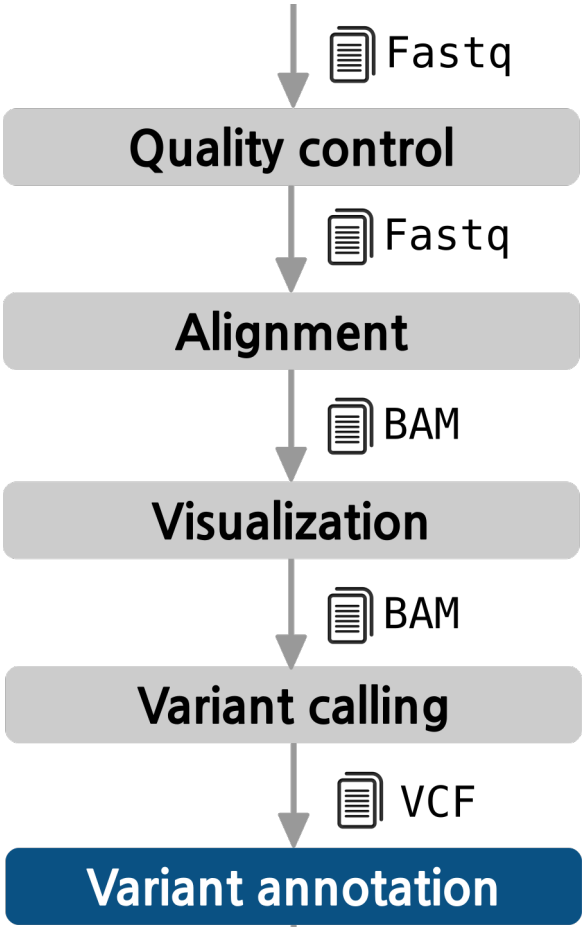
The pipeline

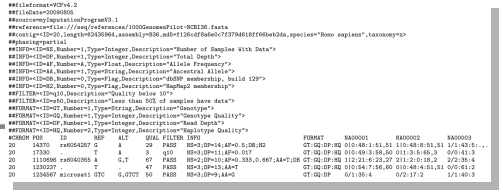


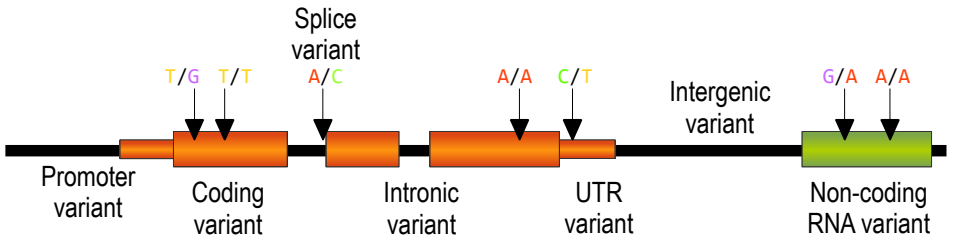




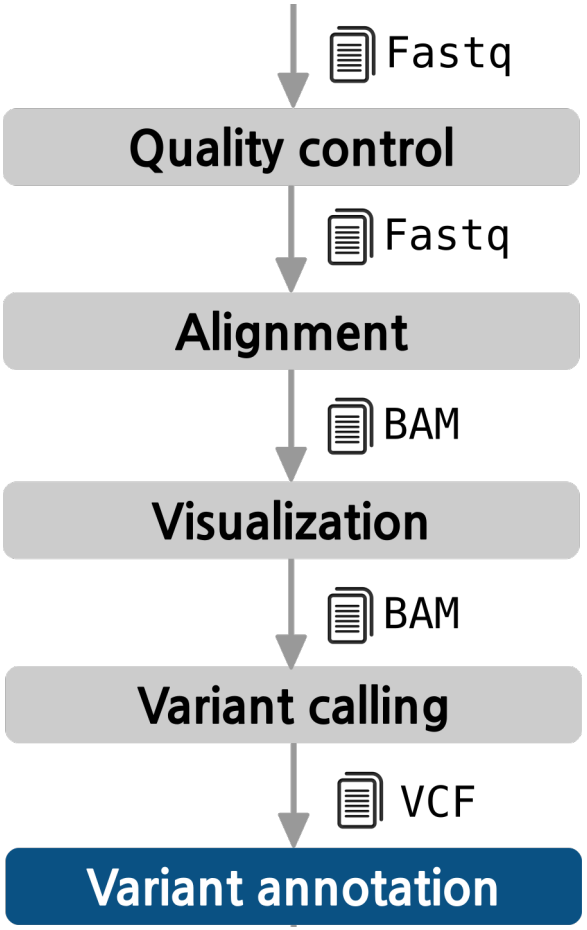
The pipeline

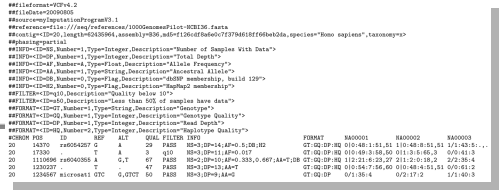


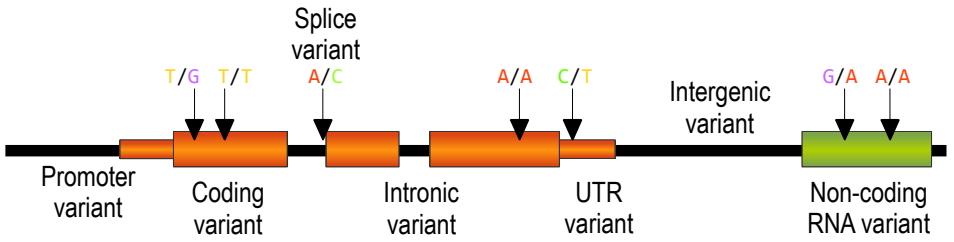




The pipeline







Challenges

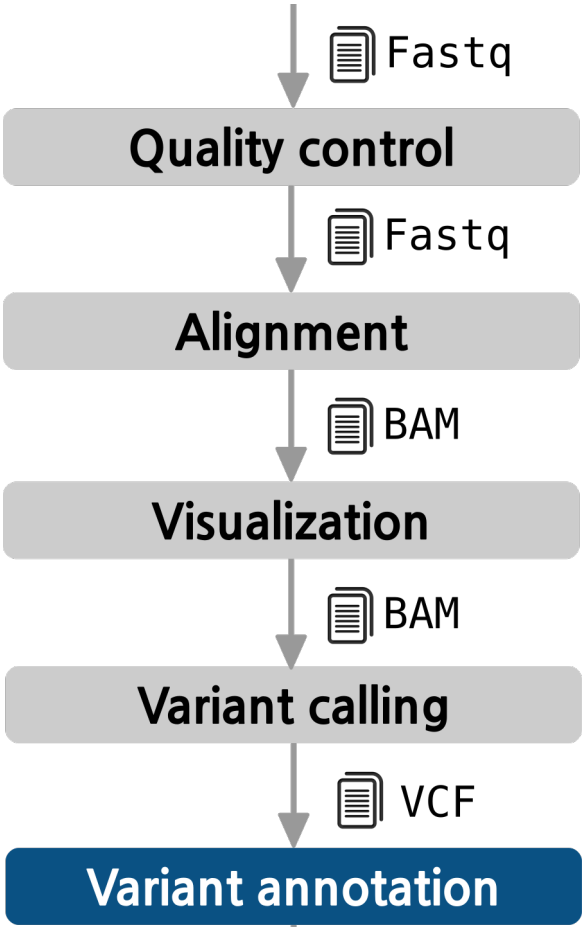
- An individual **exome** carries between **25,000** and **50,000** variants
- A **whole genome** can carry **3.5 million** variants on average
- After annotating there will be **hundreds** of **deleterious** variants

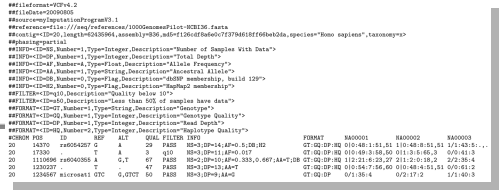
1000 Genomes Project Consortium. *A map of human genome variation from population-scale sequencing.* **Nature**. 2010 Oct 28;467(7319):1061-73. PubMed PMID: 20981092

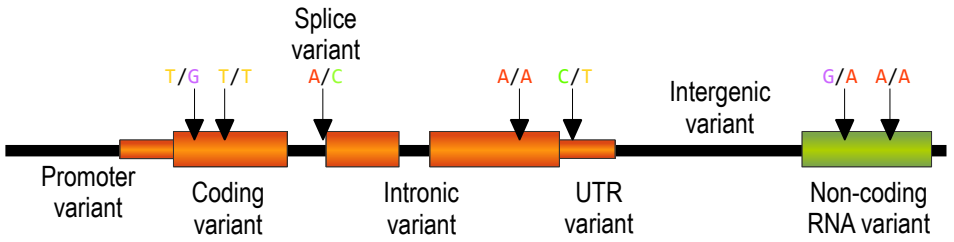
On average, every healthy person is found to carry:

- **~11,000 synonymous** variants
- **~11,000 non-synonymous** variants
- **250 to 300 loss-of-function** variants in annotated genes
- **50 to 100** variants previously implicated in **inherited disorders**

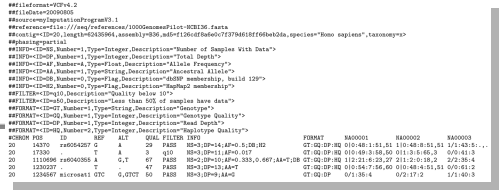
The pipeline

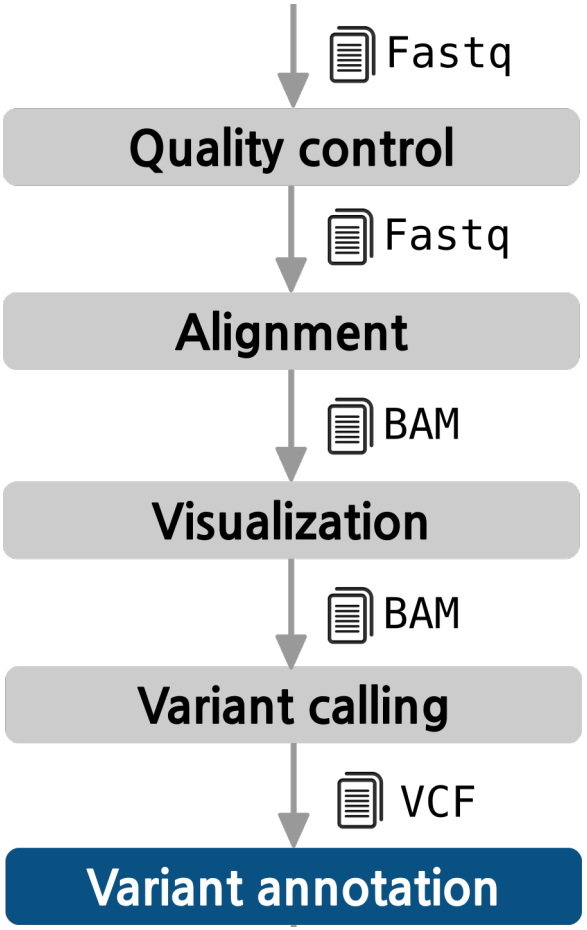


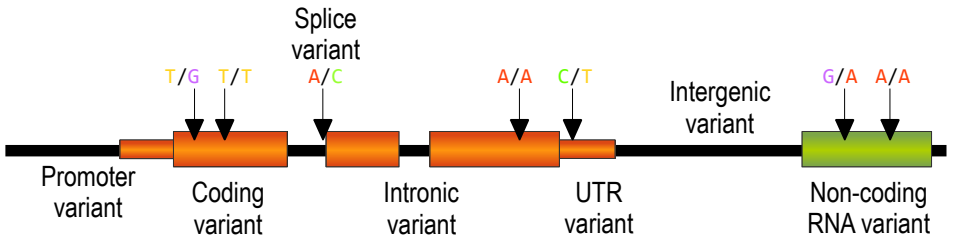




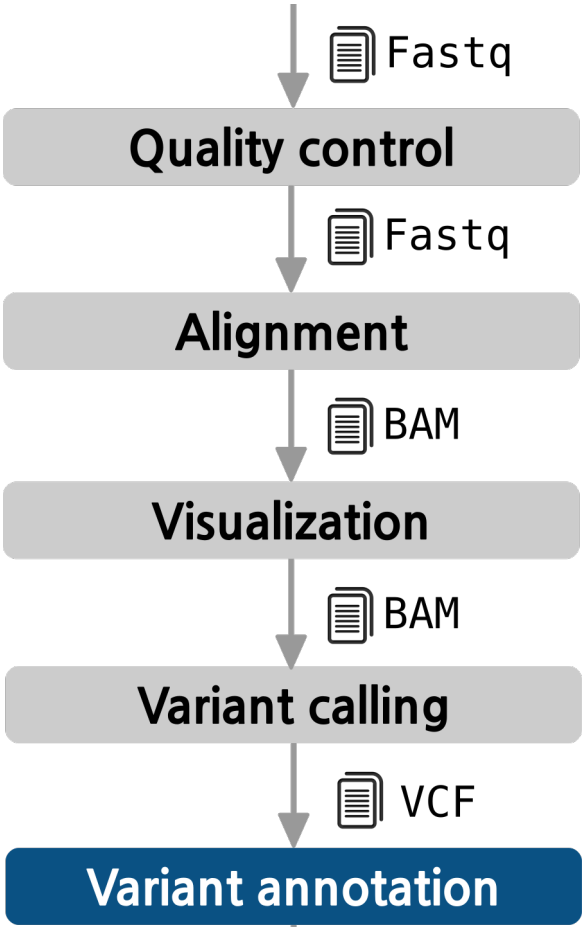
The pipeline

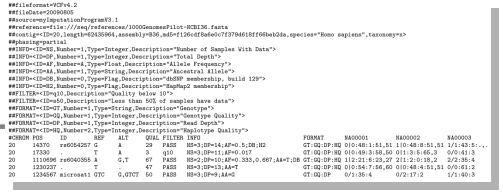


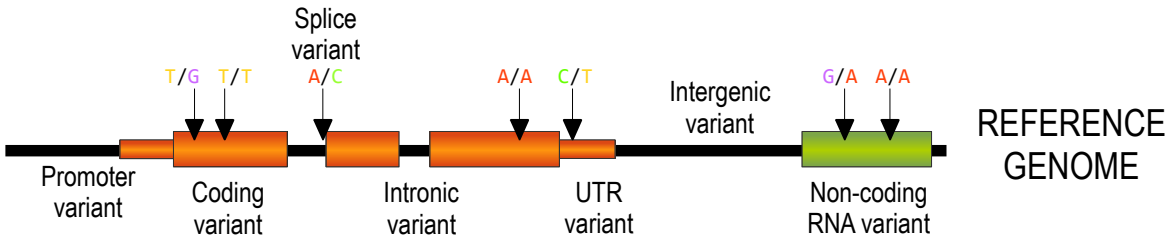




The pipeline

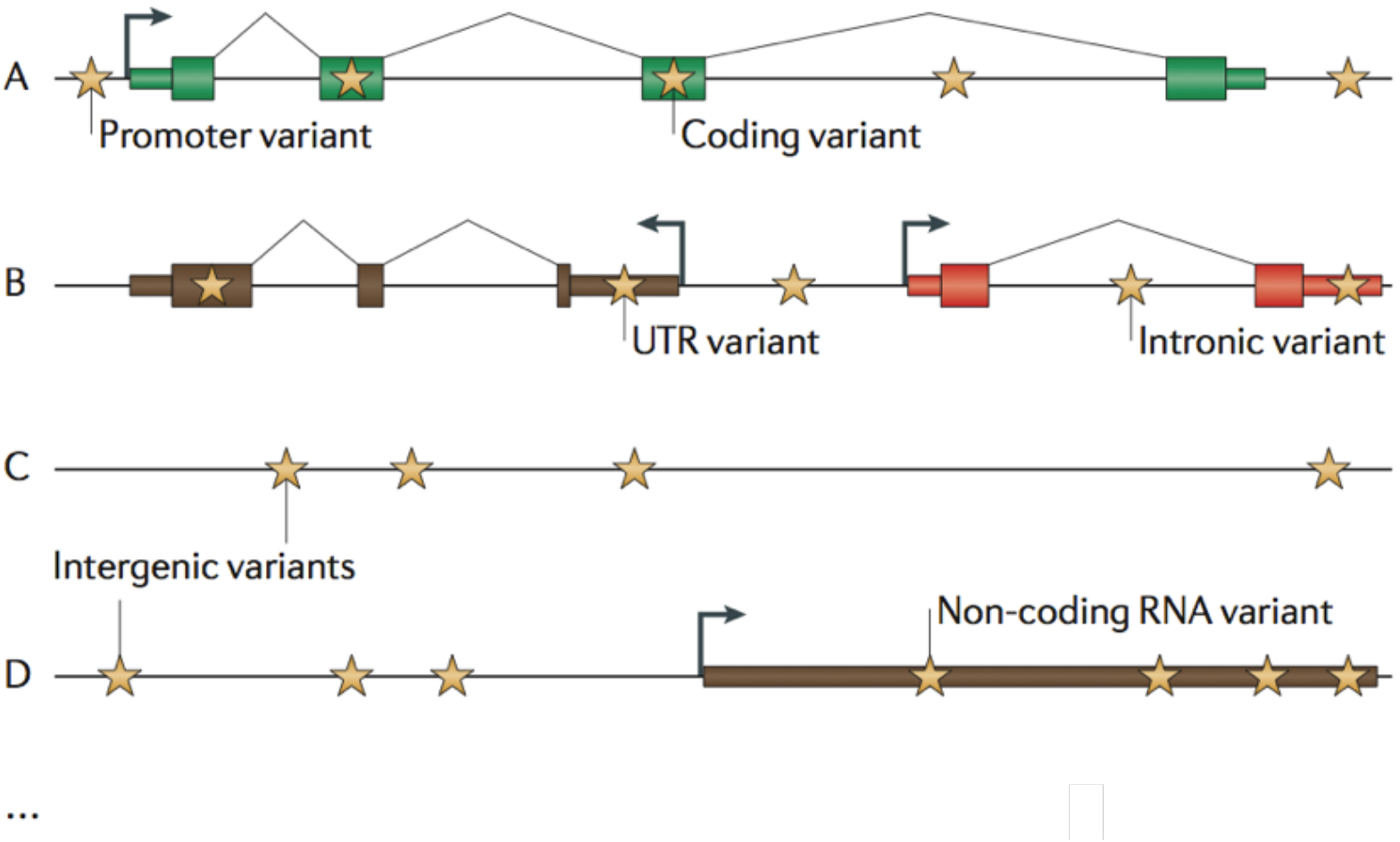






Variant annotation

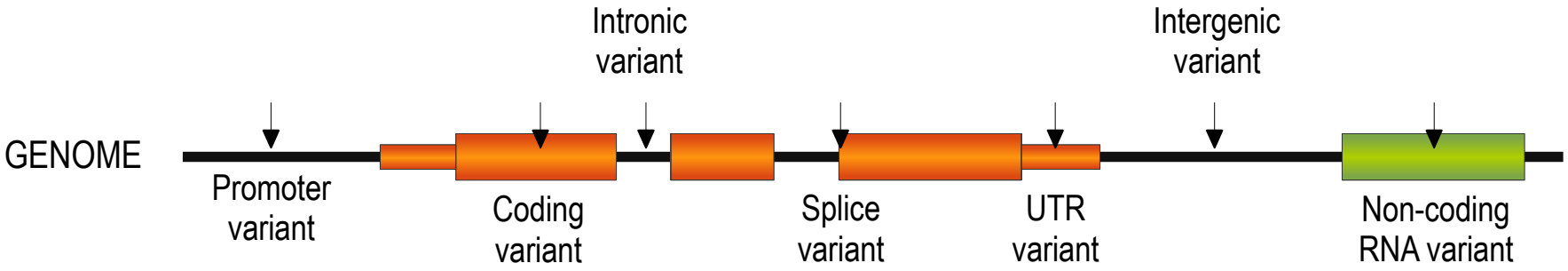
Consequence types



Cooper *et al.*, 2011

Variant annotation

External databases



UniProt

e!Ensembl

OMIM

KEGG

Pfam

NCBI

TargetScan Prediction of microRNA targets

ENCODE

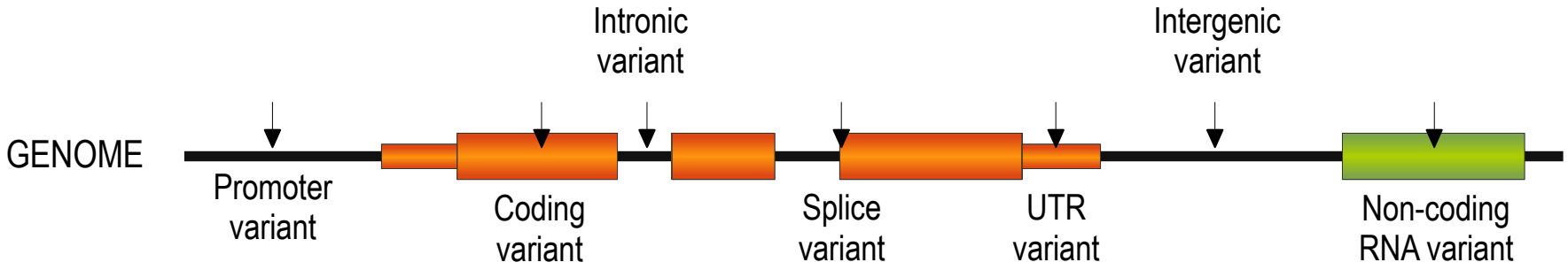
COSMIC

miRTarBase

miRBase

Variant annotation

Reference datasets



Population allele frequency is one of the most powerful discriminators of genetic variant pathogenicity

AF from large-scale genomic datasets can be used for efficient **filtering of candidate disease-causing variants**

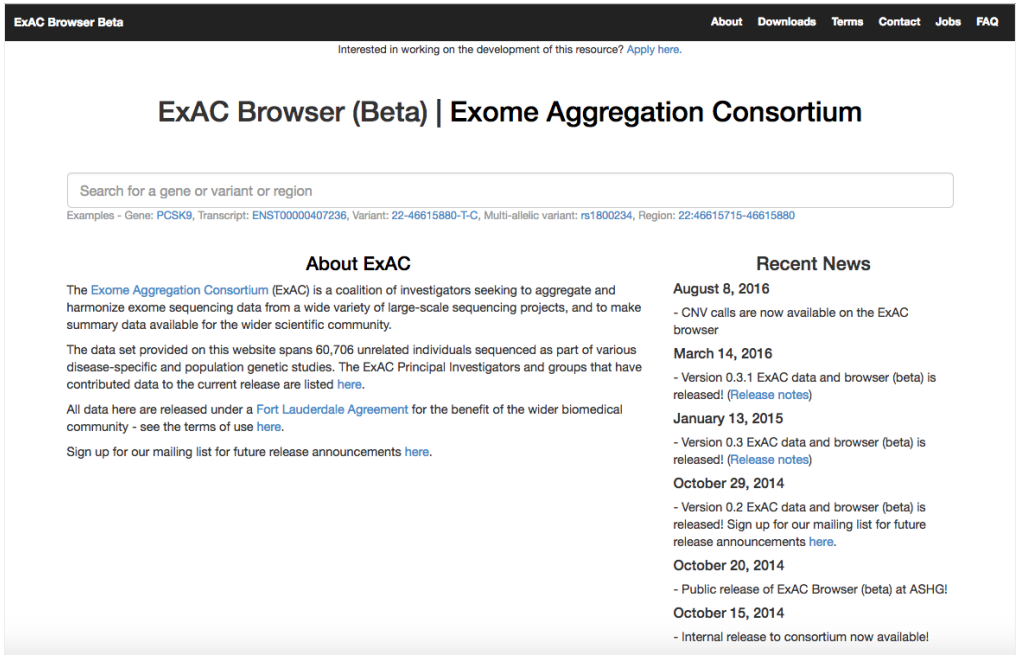
Common variants are unlikely to be pathogenic

- **1000 Genomes** (2,504 samples)
- **Exome Sequencing Project (ESP)** (6,503 samples)
- **ExAC** (60,706 samples)
- **GnomAD** (138,632 samples)

Variant annotation

Reference datasets: ExAC

Exome Aggregation Consortium (ExAC)



http://exac.broadinstitute.org/

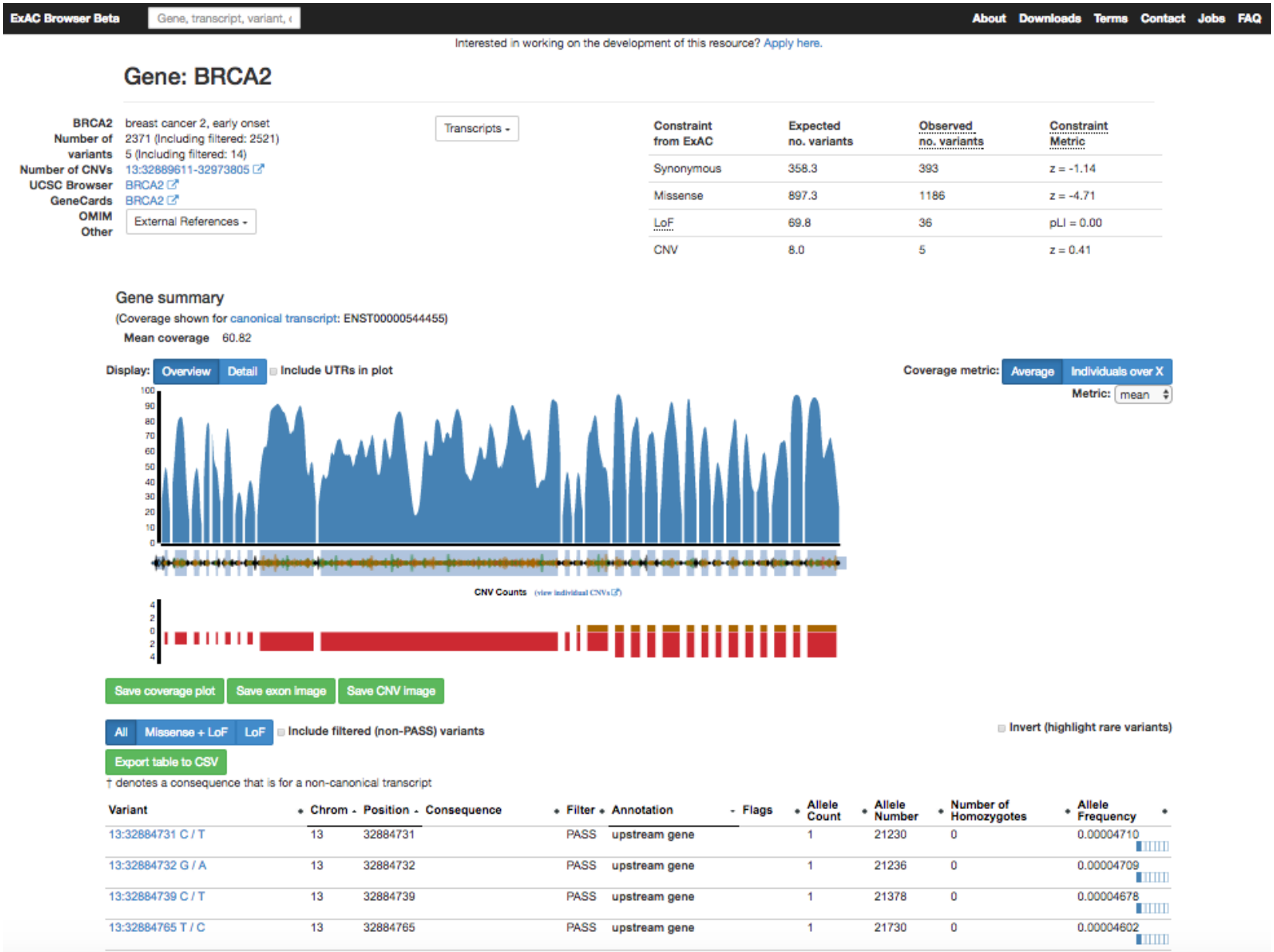
Contributing projects

- 1000 Genomes
- Bulgarian Trios
- Finland-United States Investigation of NIDDM Genetics (FUSION)
- GoT2D
- Inflammatory Bowel Disease
- METabolic Syndrome In Men (METSIM)
- Jackson Heart Study
- Myocardial Infarction Genetics Consortium:
- Italian Atherosclerosis, Thrombosis, and Vascular Biology Working Group
- Ottawa Genomics Heart Study
- Pakistan Risk of Myocardial Infarction Study (PROMIS)
- Precocious Coronary Artery Disease Study (PROCARDIS)
- Registre Gironi del COR (REGICOR)
- NHLBI-GO Exome Sequencing Project (ESP), incl. 96 PAH cases
- National Institute of Mental Health (NIMH) Controls
- SIGMA-T2D
- Sequencing in Suomi (SISu)
- Swedish Schizophrenia & Bipolar Studies
- T2D-GENES
- Schizophrenia Trios from Taiwan
- The Cancer Genome Atlas (TCGA)
- Tourette Syndrome Association International Consortium for Genomics (TSAICG)

- Aggregation of high-quality exome (protein-coding region) sequence data for **60,706 individuals** of diverse ethnicities
- 7.4M variants: **one variant every 8 base pairs** within exons
- Allows calculation of objective **metrics of pathogenicity** for sequence variants

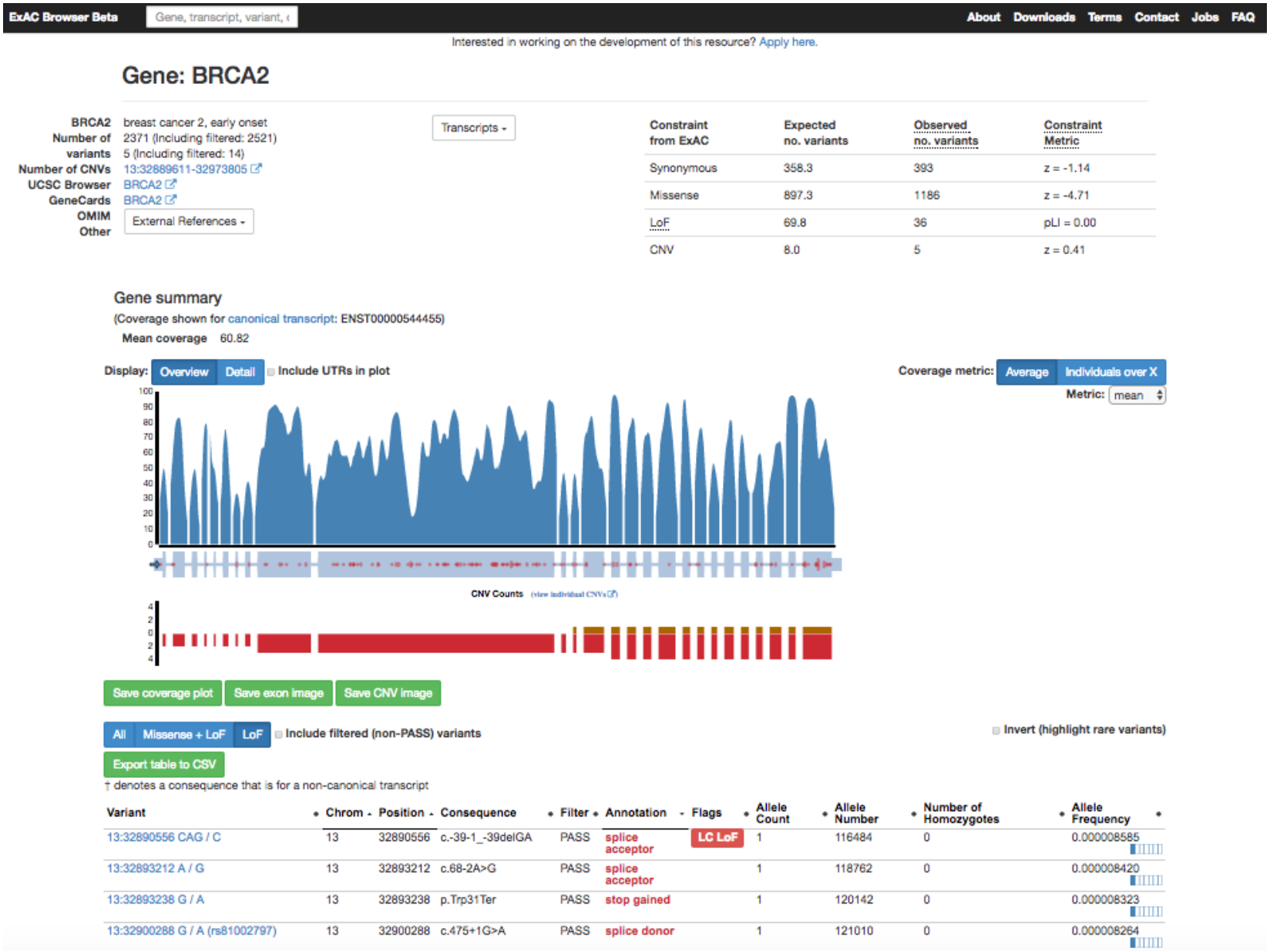
Variant annotation

Reference datasets: ExAC

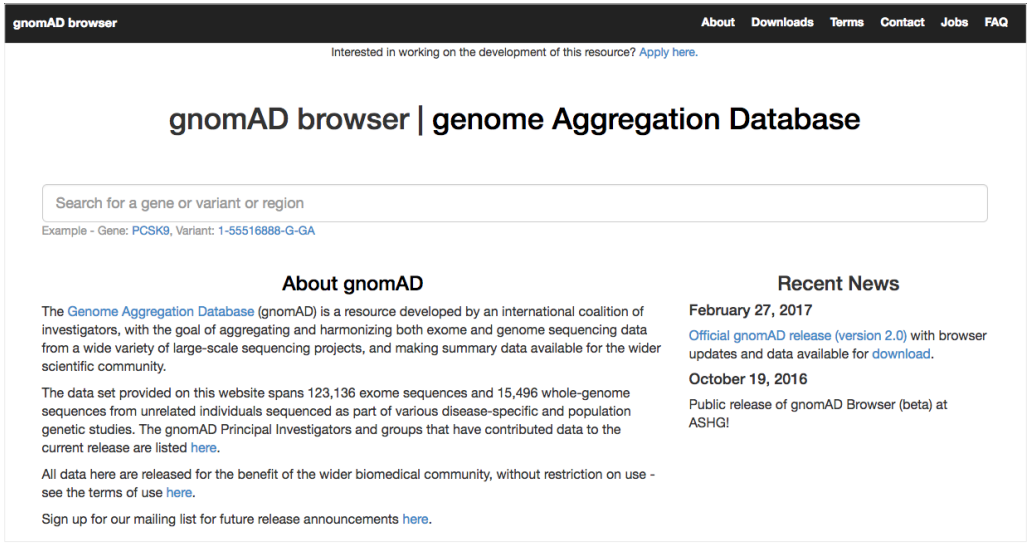


Variant annotation

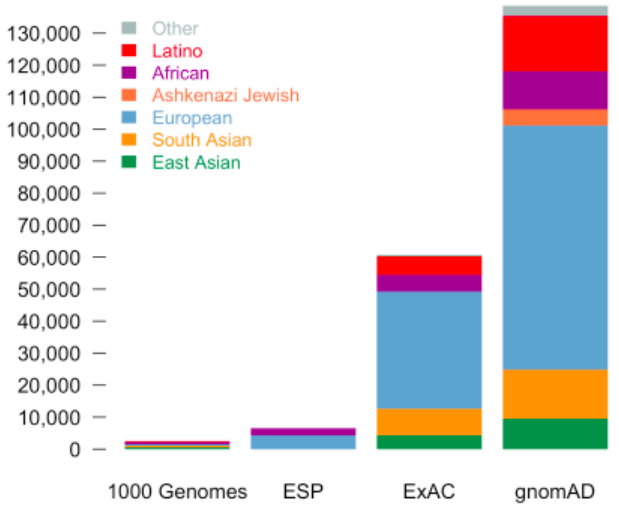
Reference datasets: ExAC



The genome Aggregation Database (gnomAD)



http://gnomad.broadinstitute.org/



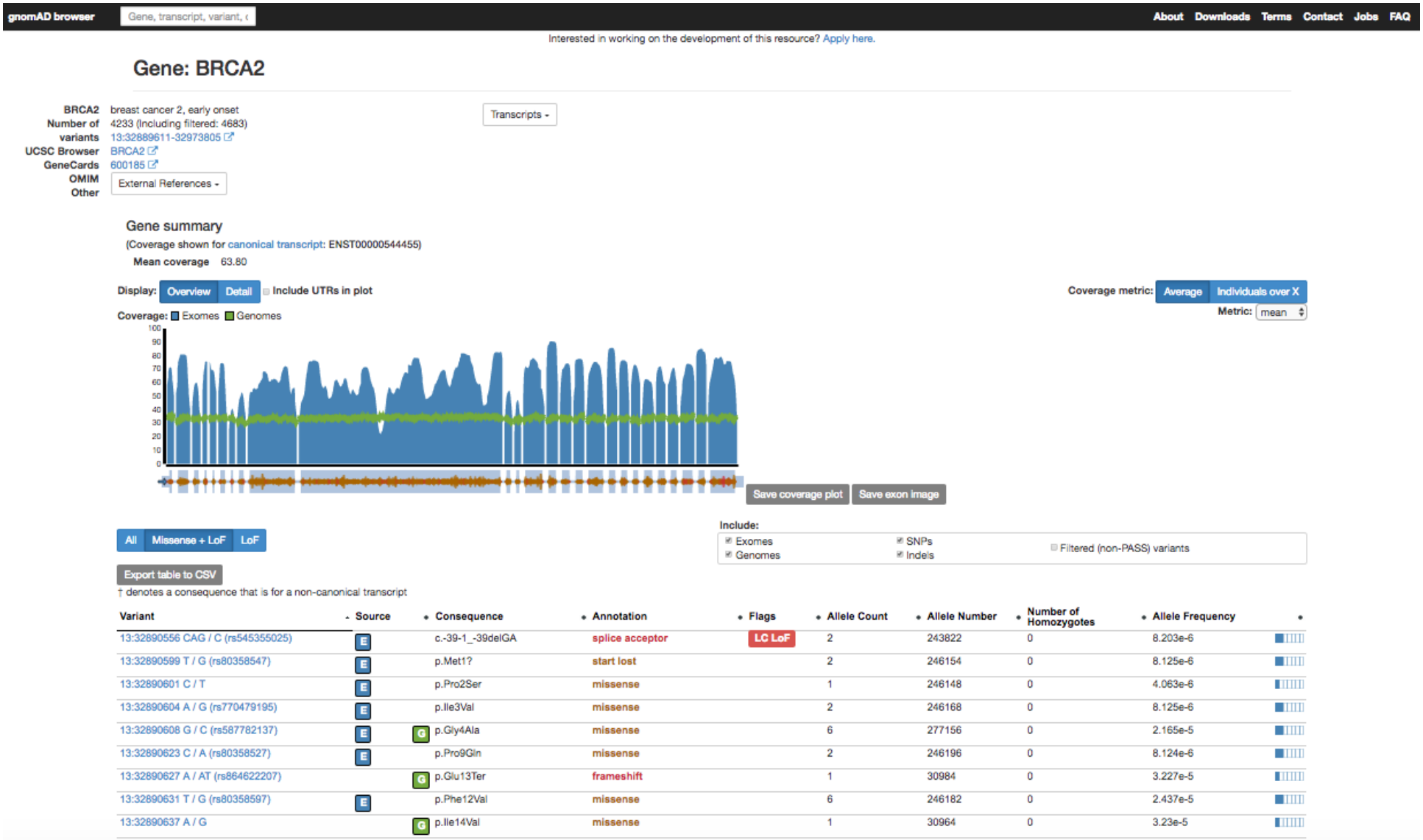
POPULATION	DESCRIPTION	GENOMES	EXOMES	TOTAL
AFR	African/African American	4,368	7,652	12,020
AMR	Admixed American	419	16,791	17,210
ASJ	Ashkenazi Jewish	151	4,925	5,076
EAS	East Asian	811	8,624	9,435
FIN	Finnish	1,747	11,150	12,897
NFE	Non-Finnish European	7,509	55,860	63,369
SAS	South Asian	0	15,391	15,391
OTH	Other (population not assigned)	491	2,743	3,234
	Total	15,496	123,136	138,632

- Released on February 2017
- Two callsets:
  - **123,136 exomes**
  - **15,496 whole genomes**
- Exomes and genomes called separately but analyzed together

https://macarthurlab.org/2017/02/27/the-genome-aggregation-database-gnomad/

Variant annotation

Reference datasets: gnomAD

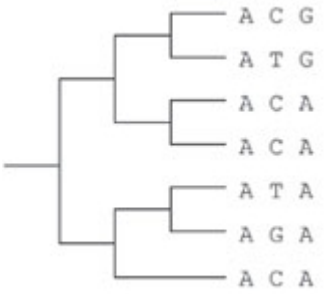


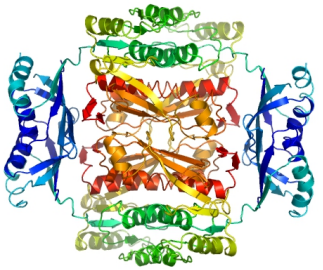
Variant	Source	Consequence	Annotation	Flags	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
13:32890556 CAG / C (rs545355025)	E	c.-39-1_-39delGA	splice acceptor	LC LoF	2	243822	0	8.203e-6
13:32890599 T / G (rs80358547)	E	p.Met1?	start lost		2	246154	0	8.125e-6
13:32890601 C / T	E	p.Pro2Ser	missense		1	246148	0	4.063e-6
13:32890604 A / G (rs770479195)	E	p.Ile3Val	missense		2	246168	0	8.125e-6
13:32890608 G / C (rs587782137)	E	G p.Gly4Ala	missense		6	277156	0	2.165e-5
13:32890623 C / A (rs80358527)	E	p.Pro9Gln	missense		2	246196	0	8.124e-6
13:32890627 A / AT (rs864622207)		G p.Glu13Ter	frameshift		1	30984	0	3.227e-5
13:32890631 T / G (rs80358597)	E	p.Phe12Val	missense		6	246182	0	2.437e-5
13:32890637 A / G		G p.Ile14Val	missense		1	30964	0	3.23e-5

Variant annotation

Deleteriousness scores

- **SIFT**: functional prediction, protein sequence conservation among homologs. Score: 1 (tolerated) - 0 (deleterious)
- **PolyPhen**: functional prediction, protein sequence and structure features. Score: 0 (benign) - 1 (damaging)
- **CADD**: ensemble score, combines 63 distinct variant annotation features retrieved from Ensembl VEP, Encode, UCSC genome browser. Phred score (i.e. 30 = 99.9% accurate or 1 in 1000is incorrect)





DNA sequence conservation scores

- **GERP**: maximum likelihood evolutionary rate estimation, predicts sites under evolutionary constraints
- **PhyloP**: base-wise conservation score derived from Multiz alignment of 100 vertebrate species
- **PhastCons**: evolutionary conserved elements derived from Multiz alignment of 100 vertebrate species (phylogenetic hidden Markov model)

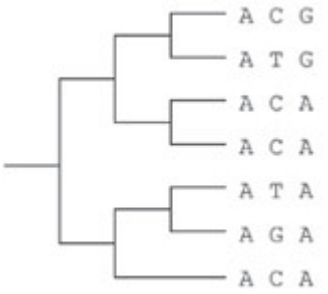
Variant annotation

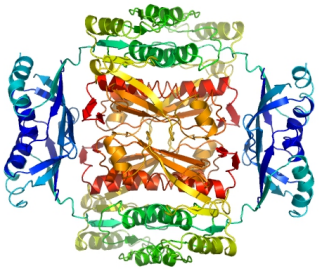
Deleteriousness scores

- **SIFT**: functional prediction, protein sequence conservation among homologs. Sc
- **PolyPhen**: functional cture features. Score: 0 (be damaging)
- **CADD**: ensemble sco tion features retrieved fro browser. Phred score is incorrect)

protein sequence and structure based prediction

Score base on various informative genome-wide annotations





DNA sequence conservation scores

- **GERP**: maximum likelihood evolutionary rate estimation, predicts sites under evolutionary constraints
- **PhyloP**: base-wise co alignment of 100 ver
- **PhastCons**: evolutionary conserved elements derived from Multiz alignment of 100 vertebrate species (phylogenetic hidden Markov model)

Measures of DNA coservation

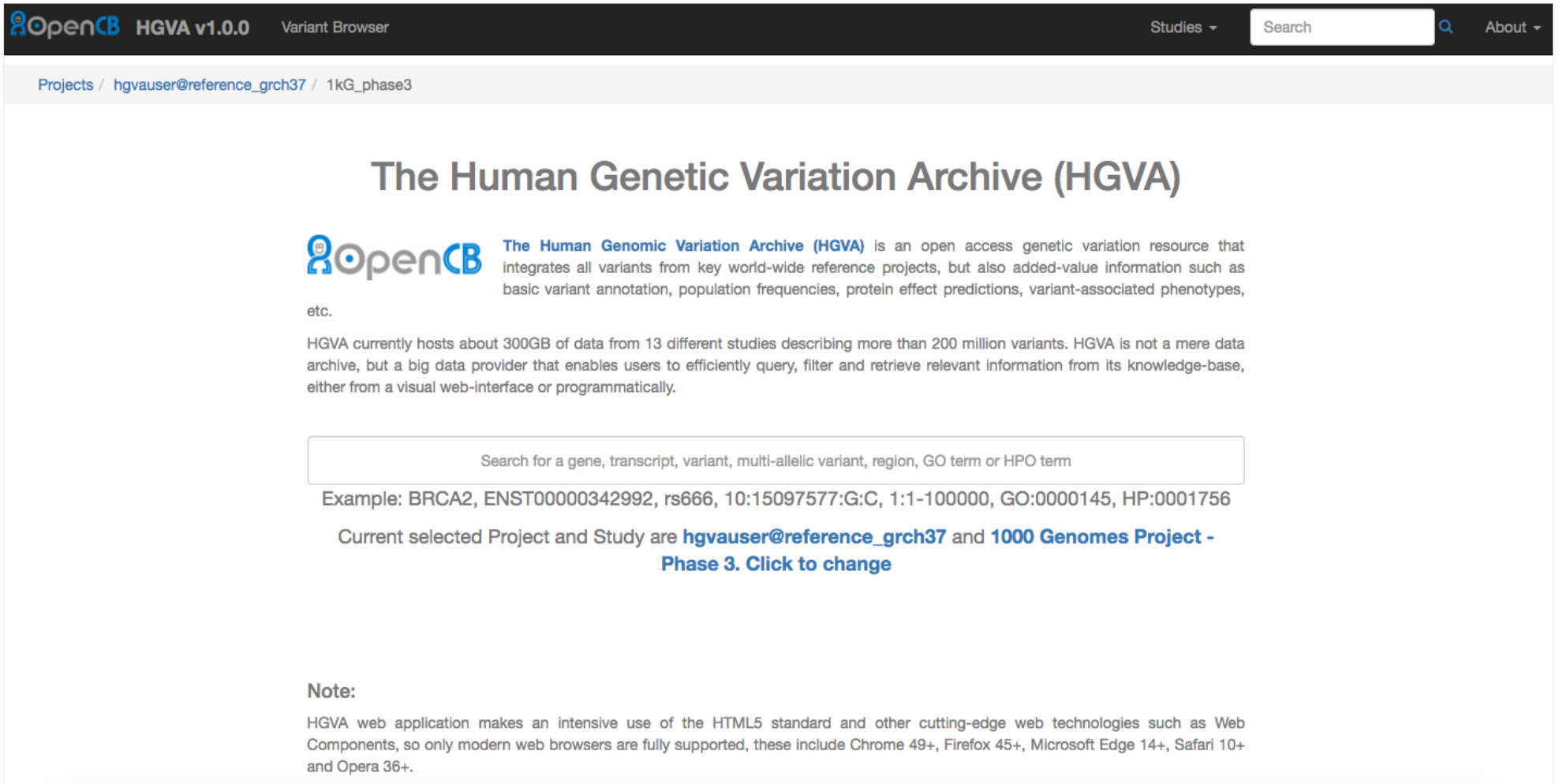
Variant annotation

Deleteriousness scores

Name	Category	Score used for analysis	Deleterious threshold	Information used
SIFT	Function prediction	1 − Score	>0.95	Protein sequence conservation among homologs
PolyPhen-2	Function prediction	Score	>0.5	Eight protein sequence features, three protein structure features
LRT	Function prediction	Score * 0.5 (if Omega ≥1) or 1 − Score * 0.5 (if Omega <1)	P	DNA sequence evolutionary model
MutationTaster	Function prediction	Score (if A or D) or 1 − Score (if N or P)	>0.5	DNA sequence conservation, splice site prediction, mRNA stability prediction and protein feature annotations
Mutation Assessor	Function prediction	(Score-Min)/(Max − Min)	>0.65	Sequence homology of protein families and sub-families within and between species
FATHMM	Function prediction	1 − (Score-Min)/(Max − Min)	≥0.45	Sequence homology
GERP++ RS	Conservation score	Score	>4.4	DNA sequence conservation
PhyloP	Conservation score	Score	>1.6	DNA sequence conservation
SiPhy	Conservation score	Score	>12.17	Inferred nucleotide substitution pattern per site
PON-P	Ensemble score	Score	P	Random forest methodology-based pipeline integrating five predictors
PANTHER	Function prediction	Score	P	Phylogenetic trees based on protein sequences
PhD-SNP	Function prediction	Score	P	SVM-based method using protein sequence and profile information
SNAP	Function prediction	Score	P	Neural network-based method using DNA sequence information as well as functional and structural annotations
SNPs&GO	Function prediction	Score	P	SVM-based method using information from protein sequence, protein sequence profile and protein function
MutPred	Function prediction	Score	>0.5	Protein sequence-based model using SIFT and a gain/loss of 14 different structural and functional properties
KGGSeq	Ensemble score	Score	P	Filtration and prioritization framework using information from three levels: genetic level, variant-gene level and knowledge level
CONDEL	Ensemble score	Score	>0.49	Weighted average of the normalized scores of five methods
CADD	Ensemble score	Score	>15	63 distinct variant annotation retrieved from Ensembl Variant Effect Predictor (VEP), data from the ENCODE project and information from UCSC genome browser tracks

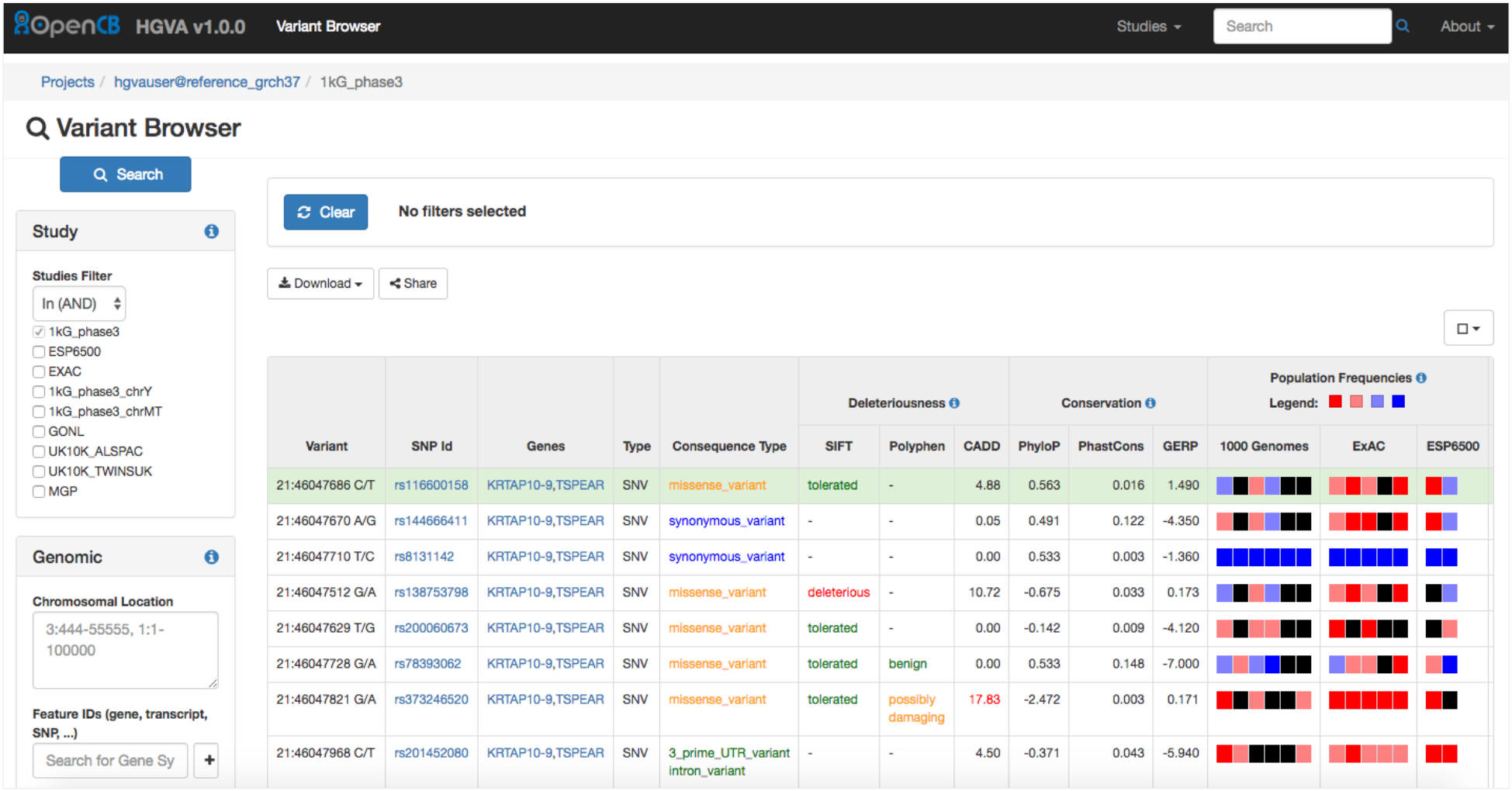
The Human Genomic Variation Archive (HGVA)

http://hgva.opencb.org/



The Human Genomic Variation Archive (HGVA)

http://hgva.opencb.org/



THANK YOU.