

Variant data analysis and prioritization using HGVA

TrainMALTA

Cambridge, UK

2nd June 2017

Marta Bleda Latorre

mb2033@cam.ac.uk

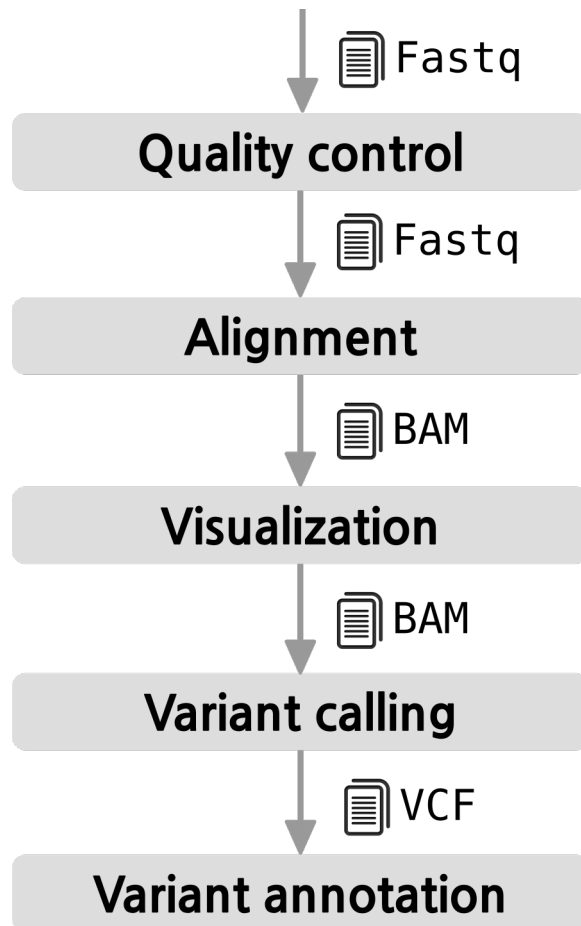
Research Associate at the Department of Medicine

University of Cambridge

Cambridge, UK

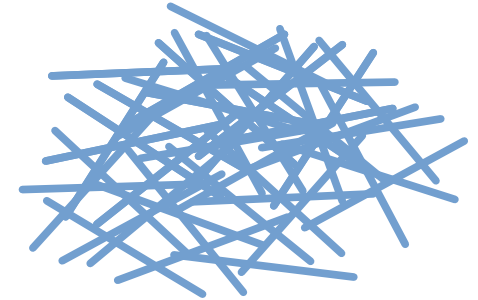
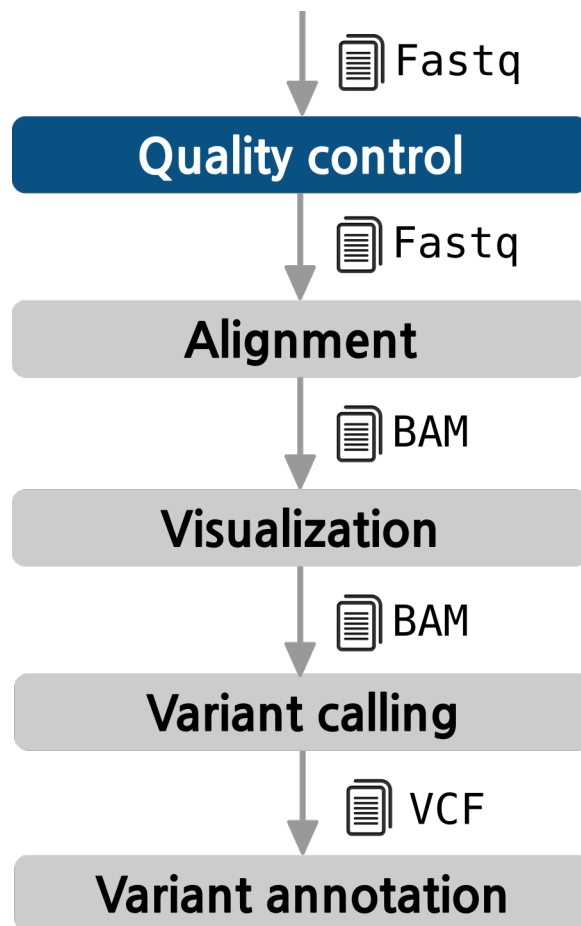


The pipeline



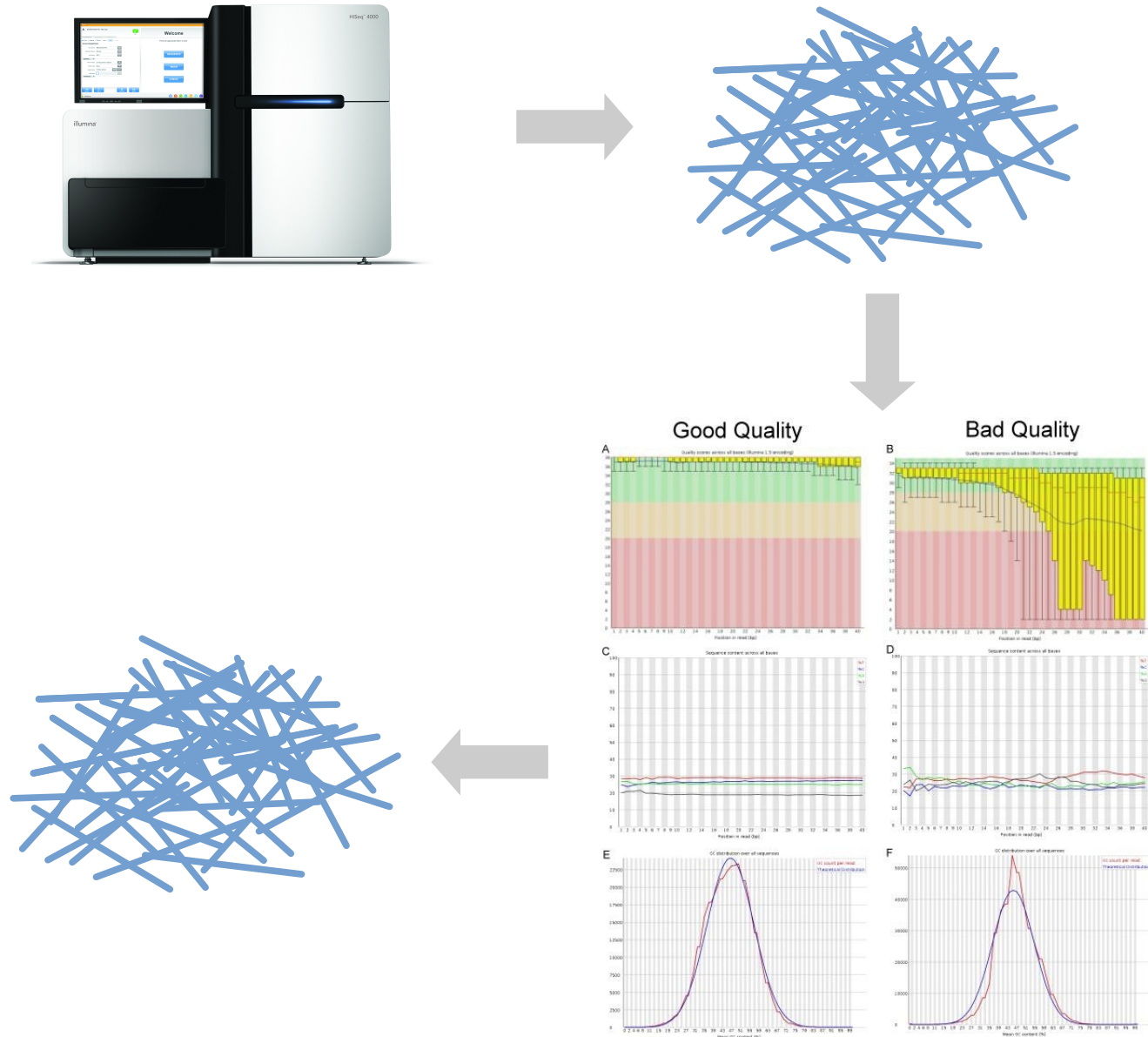
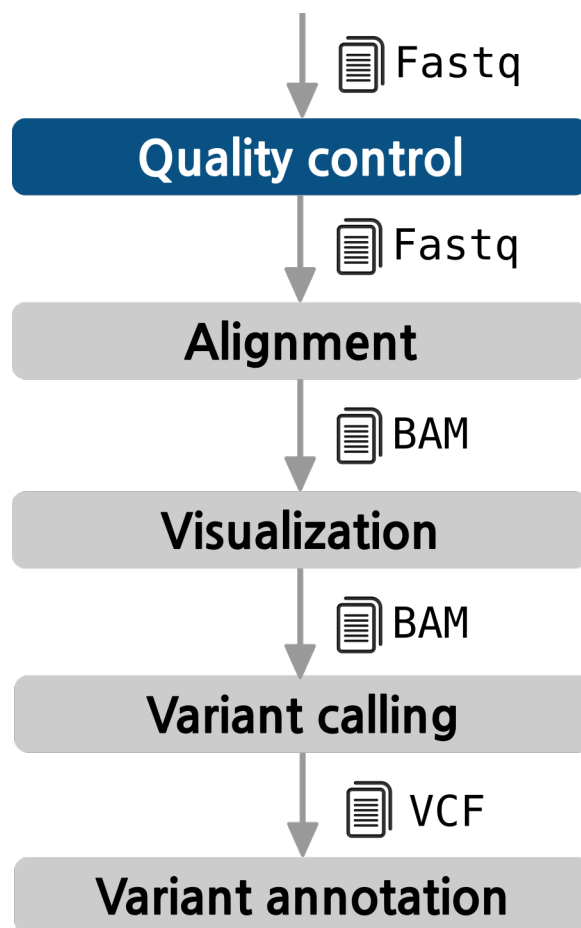
The pipeline

Fastq QC



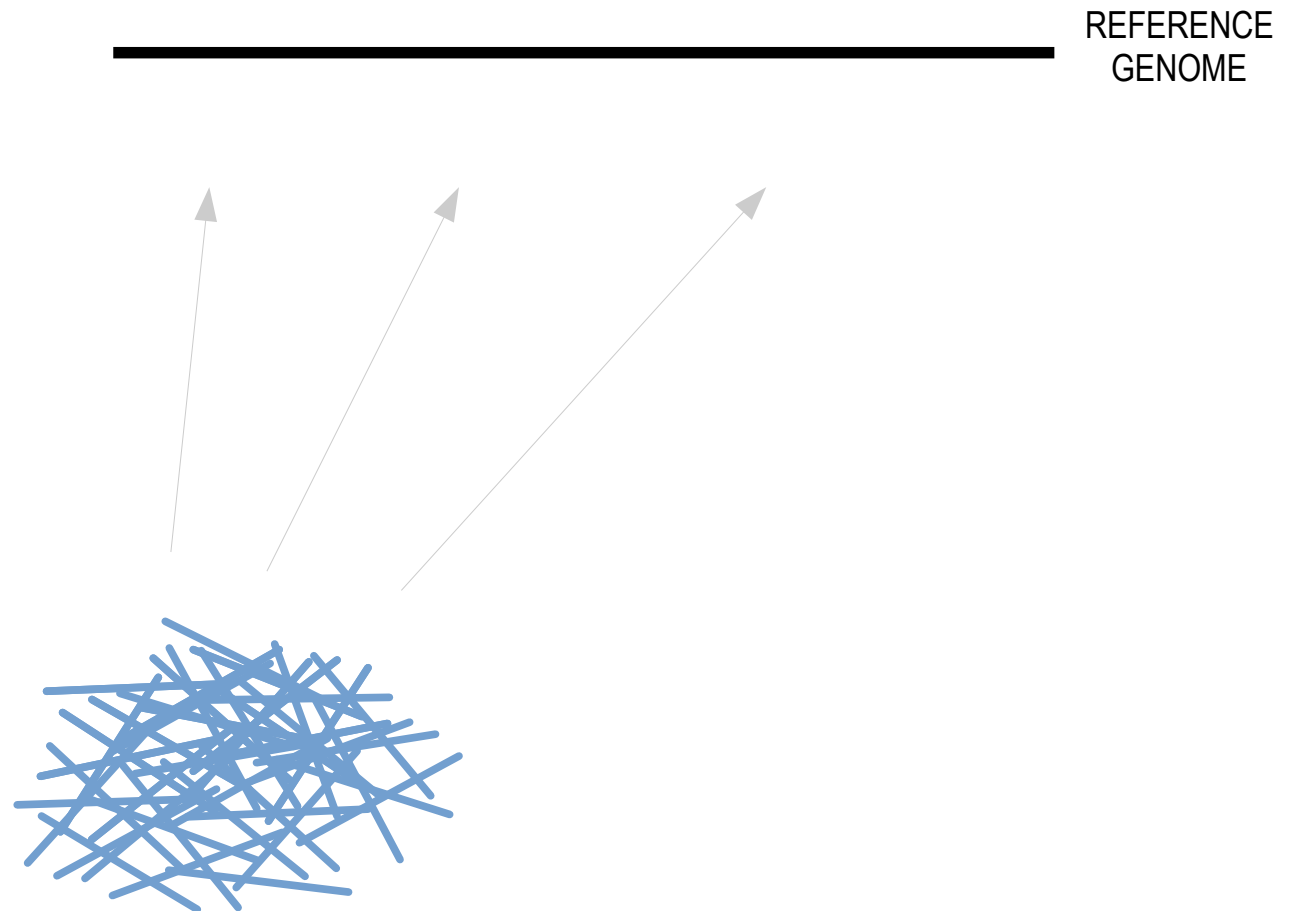
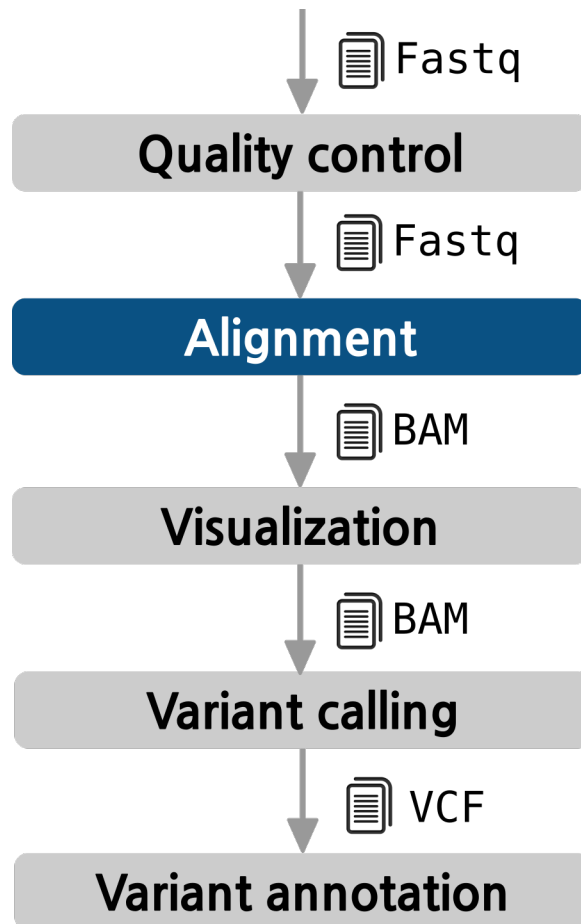
The pipeline

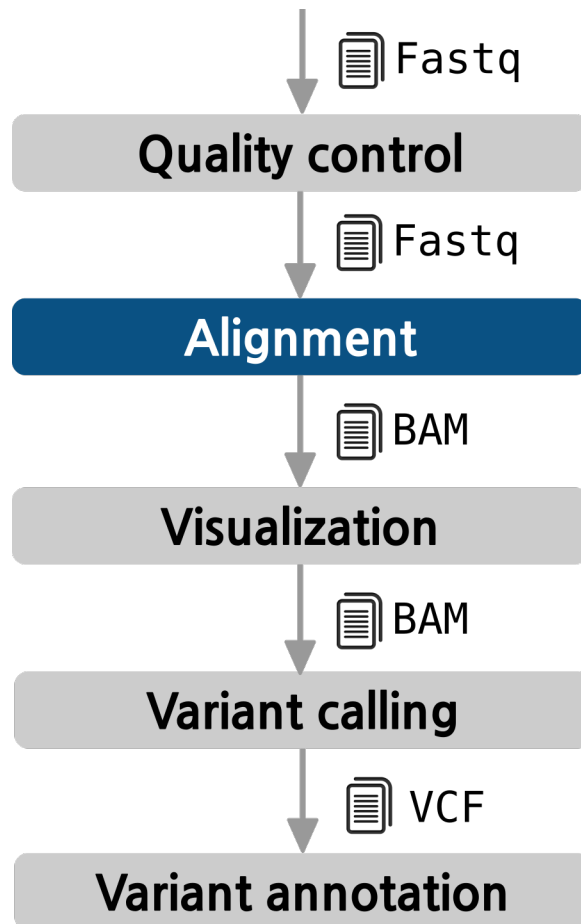
Fastq QC



The pipeline

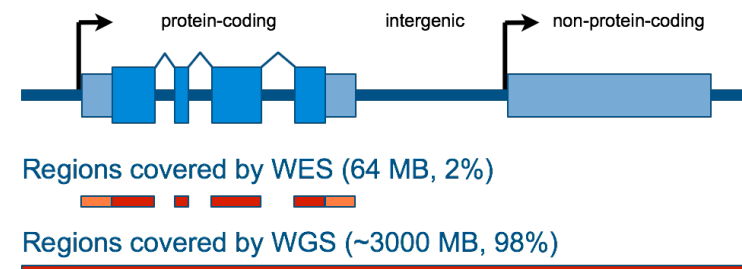
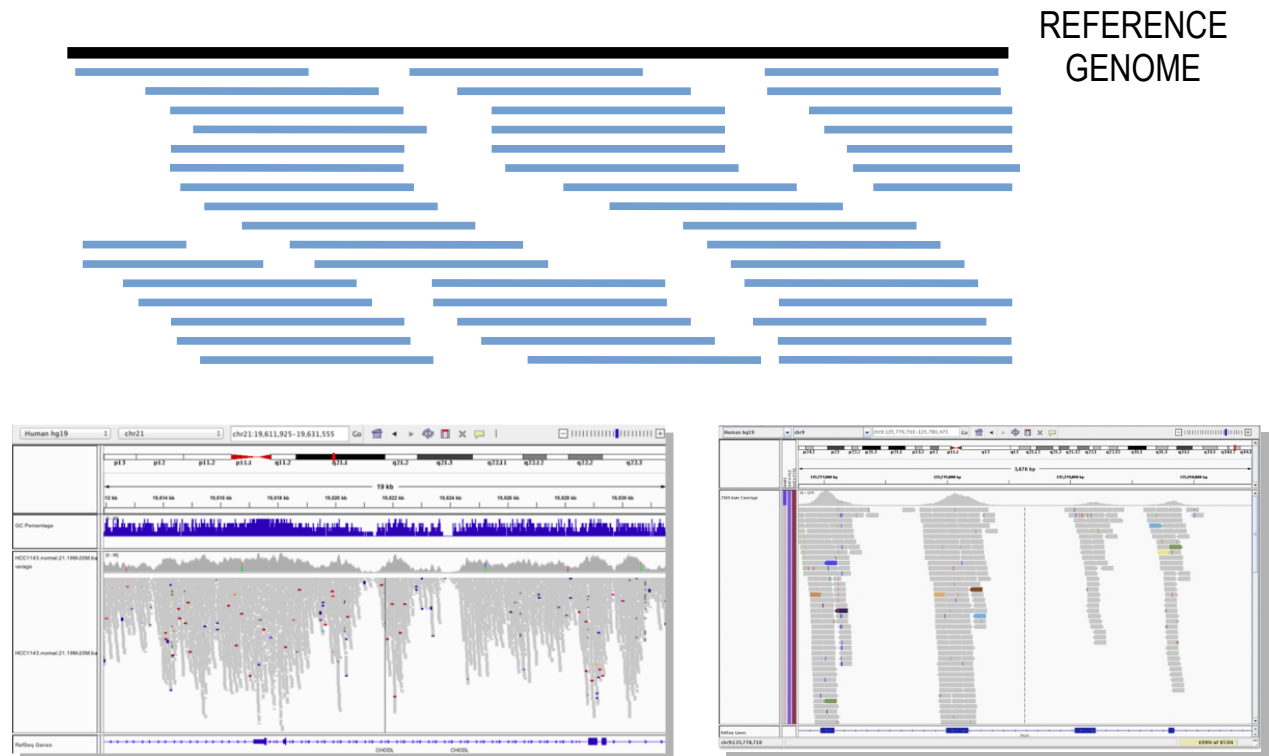
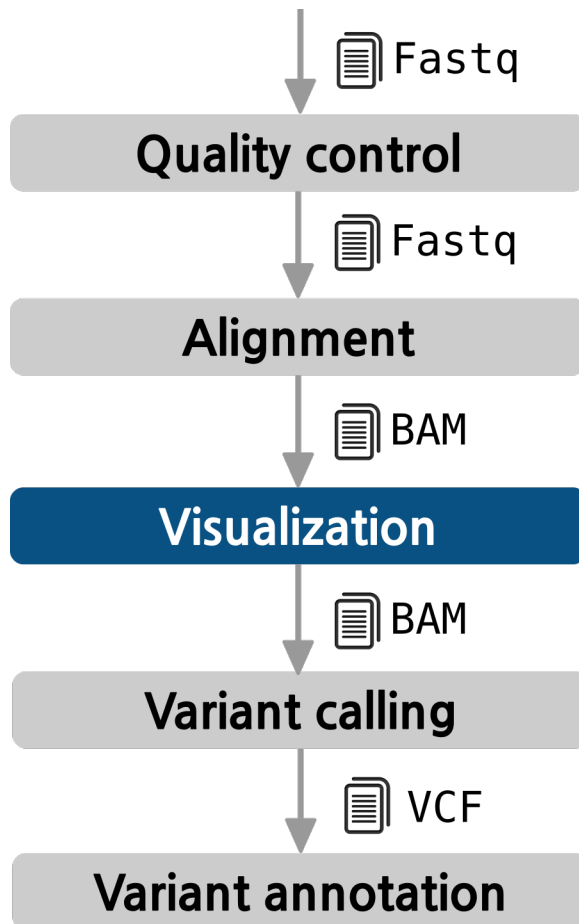
Alignment

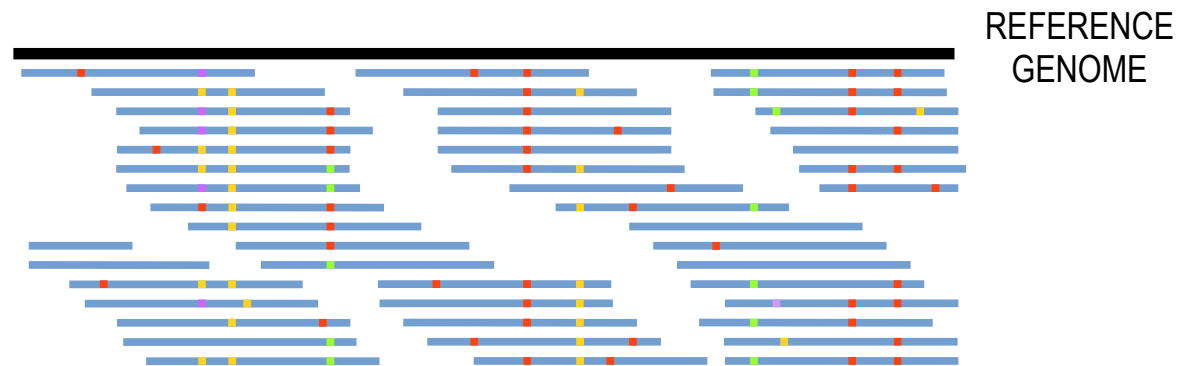
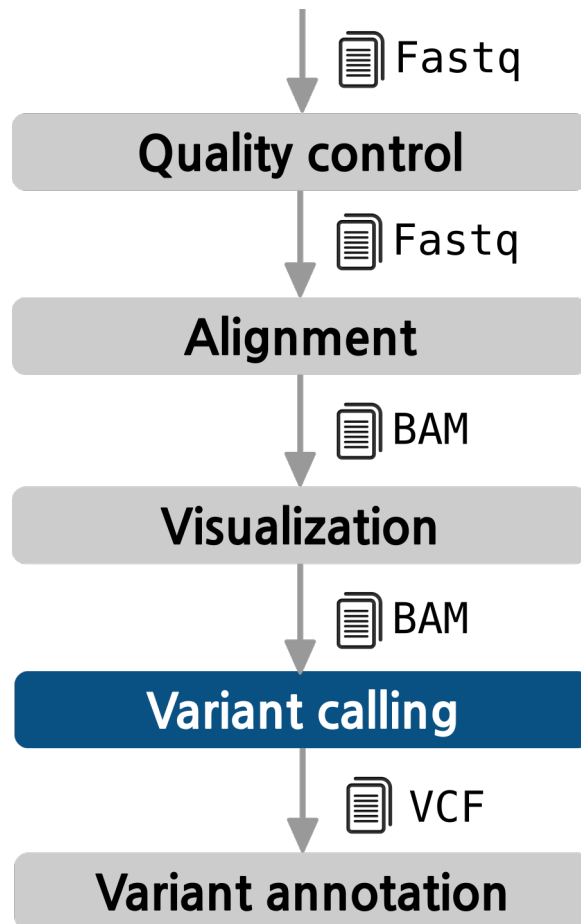




The pipeline

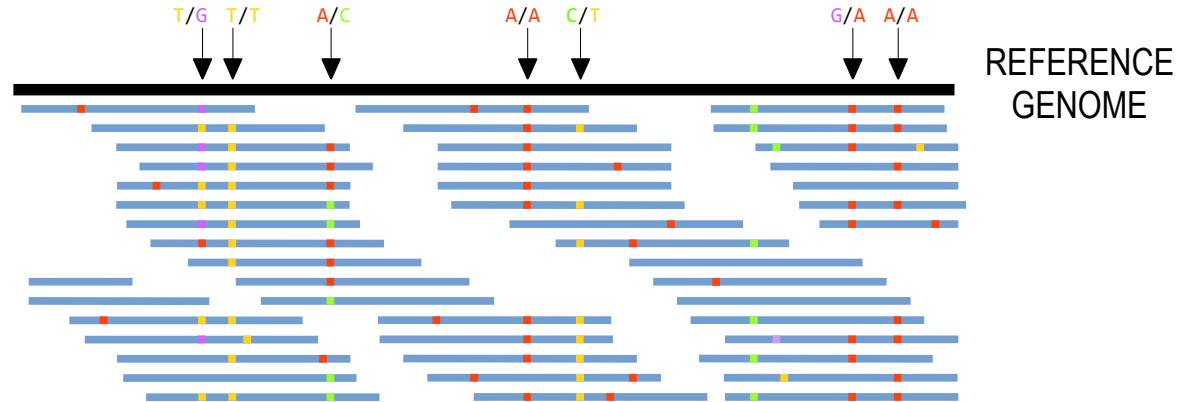
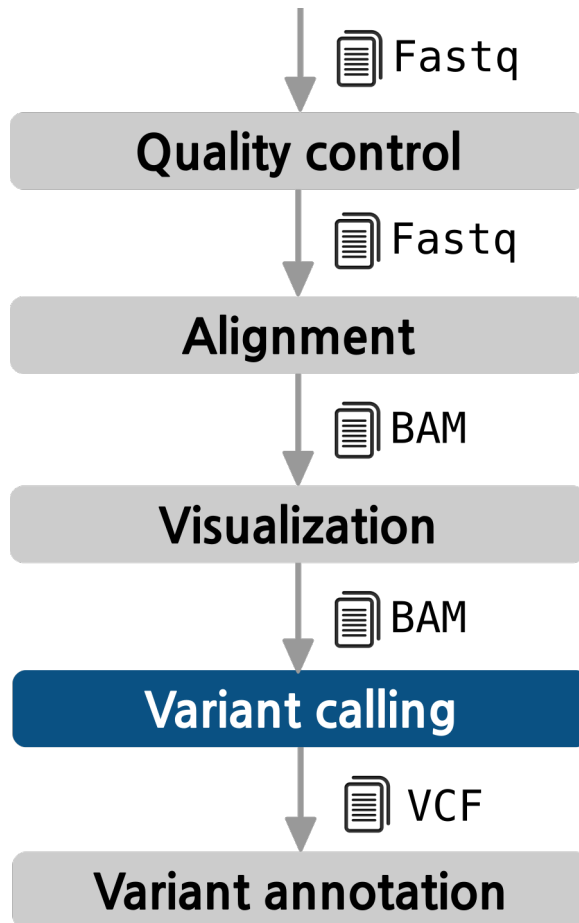
Visualization





The pipeline

Variant Calling



```
##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 NA000001 NA000002 NA000003
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 diagram illustrates the variant calling workflow. It begins with a **Fastq** file input to the **Quality control** step. The output is another **Fastq** file, which is then aligned to a reference genome in the **Alignment** step, resulting in a **BAM** file. This **BAM** file is used for **Visualization**, also resulting in a **BAM** file. The **BAM** file is then processed by **Variant calling**, which outputs a **VCF** file. Finally, the **VCF** file is used for **Variant annotation**.

To the right of the workflow, a **REFERENCE GENOME** is shown as a horizontal line. Arrows point to specific positions on the genome, indicating variant calls with their genotypes: **T/G**, **T/T**, **A/C**, **A/A**, **C/T**, **G/A**, and **A/A**.

[illegible]

The diagram illustrates the variant calling pipeline and the types of variants identified in a genomic track.

Variant Calling Pipeline:

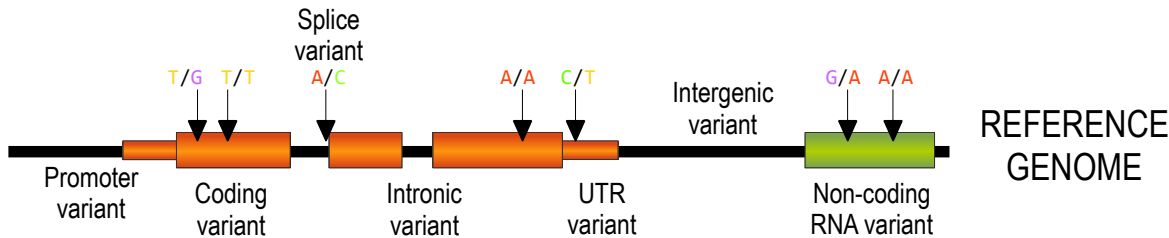
- Quality control** (Input: Fastq)
- Alignment** (Input: Fastq)
- Visualization** (Input: BAM)
- Variant calling** (Input: BAM)
- Variant annotation** (Input: VCF)

Genomic Track Visualization:

The track shows a reference genome with various variant types identified:

- Promoter variant**
- Coding variant** (T/G, T/T)
- Splice variant** (A/C)
- Intronic variant**
- UTR variant** (A/A, C/T)
- Intergenic variant**
- Non-coding RNA variant** (G/A, A/A)

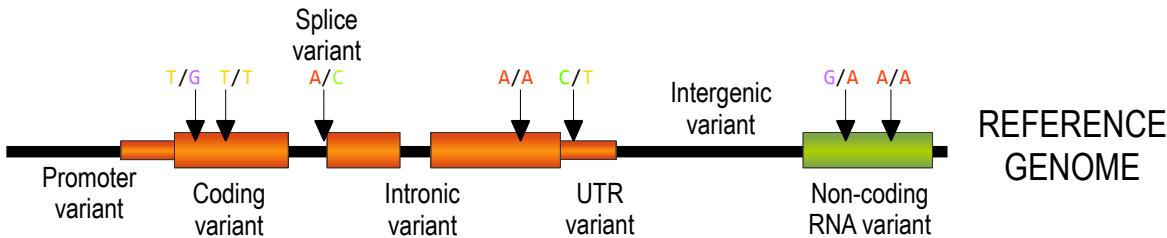
REFERENCE GENOME

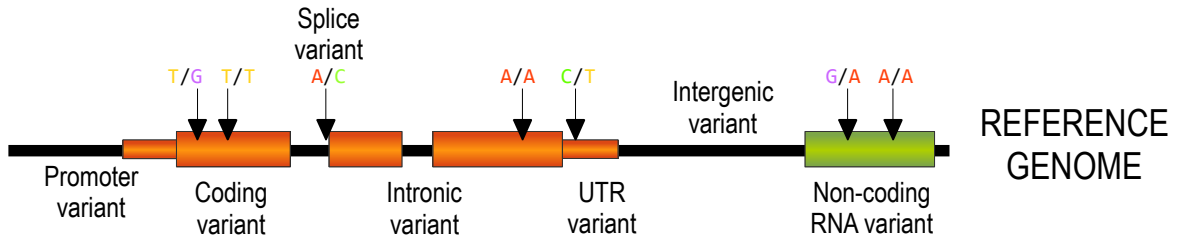
[illegible]

- 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

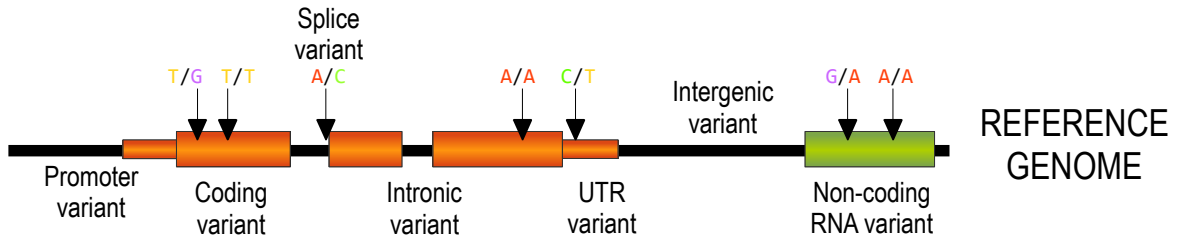
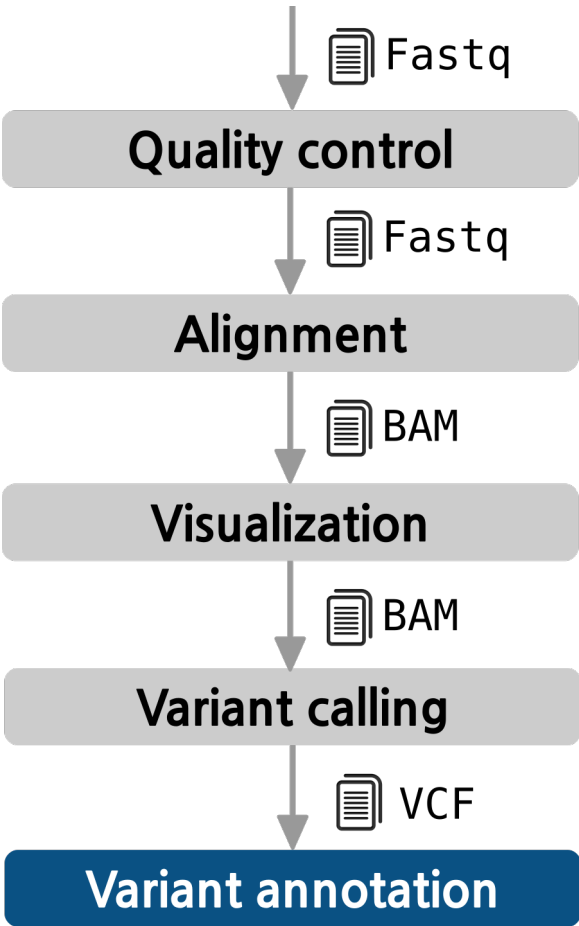
On average, every healthy person is found to carry:

- Marta Bleda | Variant data analysis and prioritization using HGVA

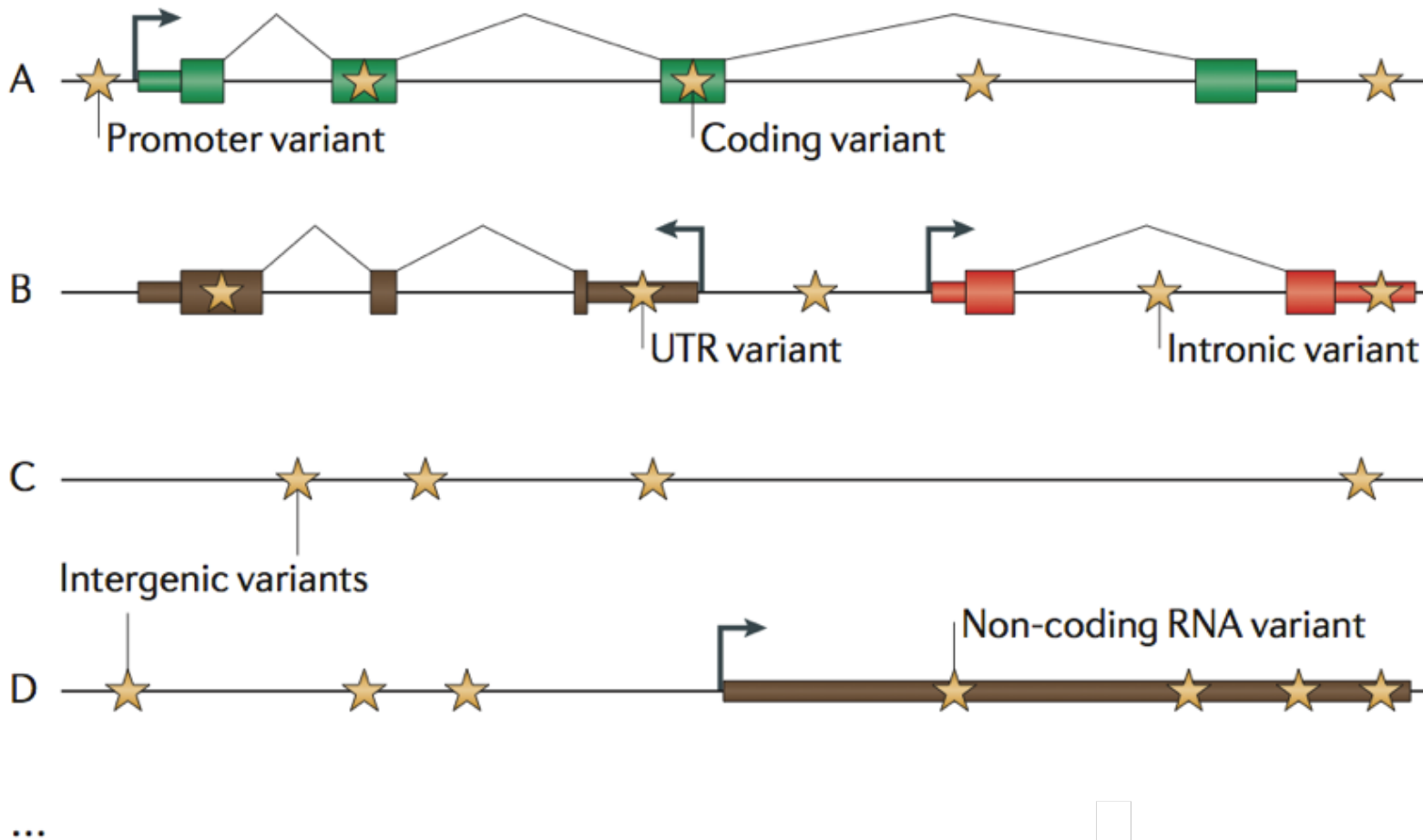
[illegible]

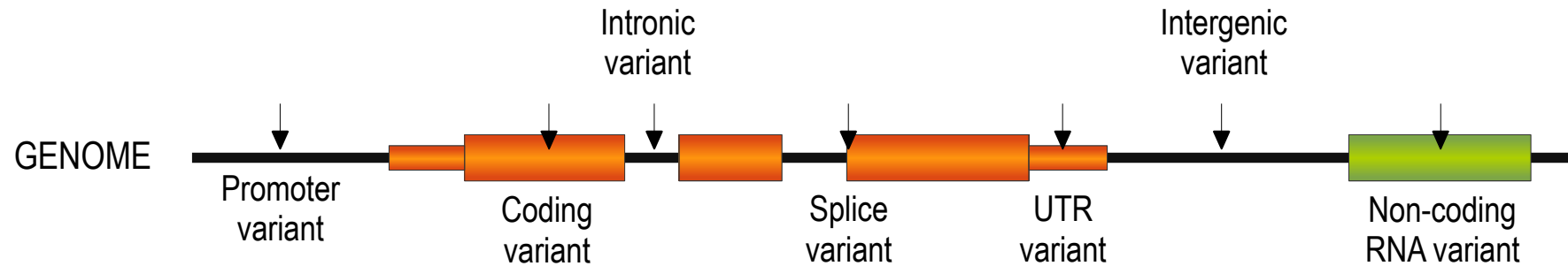
[illegible]

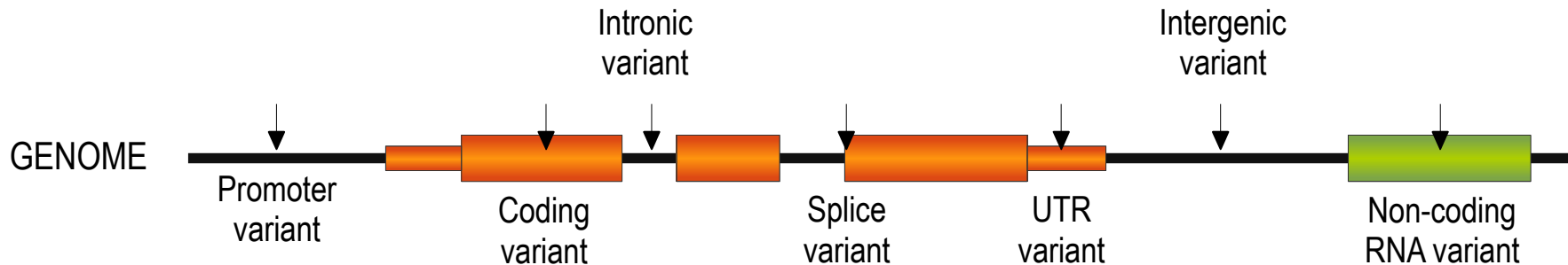
The pipeline



FILTER!
FILTER!
FILTER!







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)

Exome Aggregation Consortium (ExAC)

The screenshot shows the ExAC Browser Beta website. At the top, there's a navigation bar with links: About, Downloads, Terms, Contact, Jobs, and FAQ. Below the navigation bar, there's a search bar with the placeholder text "Search for a gene or variant or region". Below the search bar, there's a line of example text: "Examples - Gene: PCSK9, Transcript: ENST00000407236, Variant: 22-46615880-T-C, Multi-allelic variant: rs1800234, Region: 22:46615715-46615880". The main content area is divided into two columns. The left column is titled "About ExAC" and contains text about the consortium's mission and data availability. The right column is titled "Recent News" and lists several updates, including the release of CNV calls, version 0.3.1, and version 0.2.

<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)

ARTICLE

OPEN

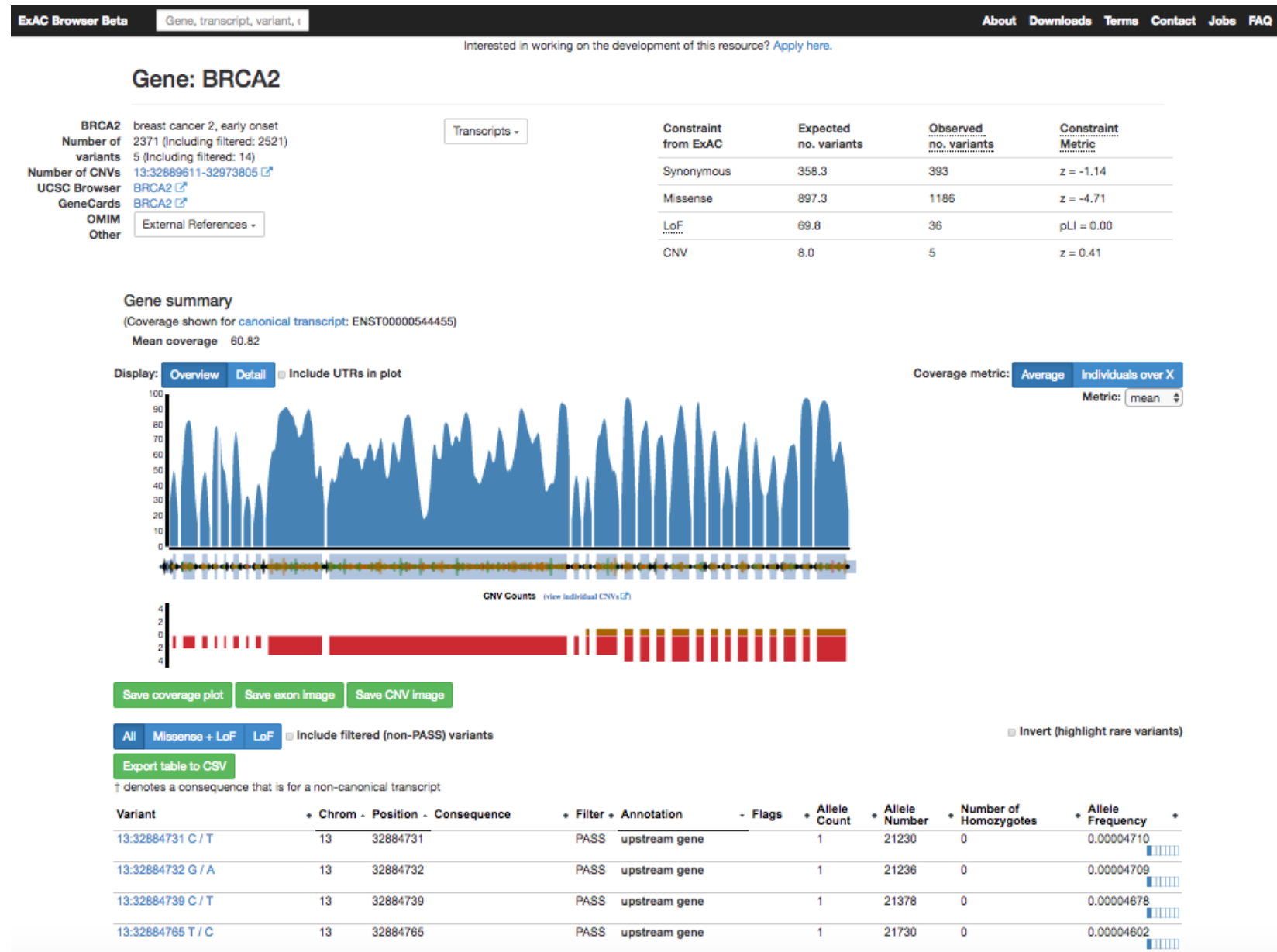
doi:10.1038/nature19057

Analysis of protein-coding genetic variation in 60,706 humans

- 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



ExAC Browser Beta

Gene, transcript, variant, c

About Downloads Terms Contact Jobs FAQ

Interested in working on the development of this resource? [Apply here.](#)

Gene: BRCA2

BRCA2

breast cancer 2, early onset

Number of variants 2371 (including filtered: 2521)

Number of CNVs 5 (including filtered: 14)

UCSC Browser [BRCA2](#)

GeneCards [BRCA2](#)

OMIM [BRCA2](#)

Other [External References](#)

Transcripts

Constraint from ExAC	Expected no. variants	Observed no. variants	Constraint Metric
Synonymous	358.3	393	$z = -1.14$
Missense	897.3	1186	$z = -4.71$
LoF	69.8	36	$pLI = 0.00$
CNV	8.0	5	$z = 0.41$

Gene summary

(Coverage shown for canonical transcript: ENST00000544455)

Mean coverage 60.82

Display: Overview Detail

☐ Include UTRs in plot

Coverage metric: Average Individuals over X

Metric: mean

Save coverage plot Save exon image Save CNV image

All Missense + LoF LoF

☐ Include filtered (non-PASS) variants

Export table to CSV

☐ Invert (highlight rare variants)

† denotes a consequence that is for a non-canonical transcript

Variant	Chrom	Position	Consequence	Filter	Annotation	Flags	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
13:32890556 CAG / C	13	32890556	c.-39-1_39delGA	PASS	splice acceptor	LC LoF	1	116484	0	0.000008585
13:32893212 A / G	13	32893212	c.68-2A>G	PASS	splice acceptor		1	118762	0	0.000008420
13:32893238 G / A	13	32893238	p.Trp31Ter	PASS	stop gained		1	120142	0	0.000008323
13:32900288 G / A (rs81002797)	13	32900288	c.475+1G>A	PASS	splice donor		1	121010	0	0.000008264

The genome Aggregation Database (gnomAD)

gnomAD browser | genome Aggregation Database

Interested in working on the development of this resource? [Apply here.](#)

Search for a gene or variant or region

Example - Gene: PCSK9, Variant: 1-55516888-G-GA

About gnomAD

The [Genome Aggregation Database](#) (gnomAD) is a resource developed by an international coalition of investigators, with the goal of aggregating and harmonizing both exome and genome sequencing data from a wide variety of large-scale sequencing projects, and making summary data available for the wider scientific community.

The data set provided on this website spans 123,136 exome sequences and 15,496 whole-genome sequences from unrelated individuals sequenced as part of various disease-specific and population genetic studies. The gnomAD Principal Investigators and groups that have contributed data to the current release are listed [here](#).

All data here are released for the benefit of the wider biomedical community, without restriction on use - see the terms of use [here](#).

Sign up for our mailing list for future release announcements [here](#).

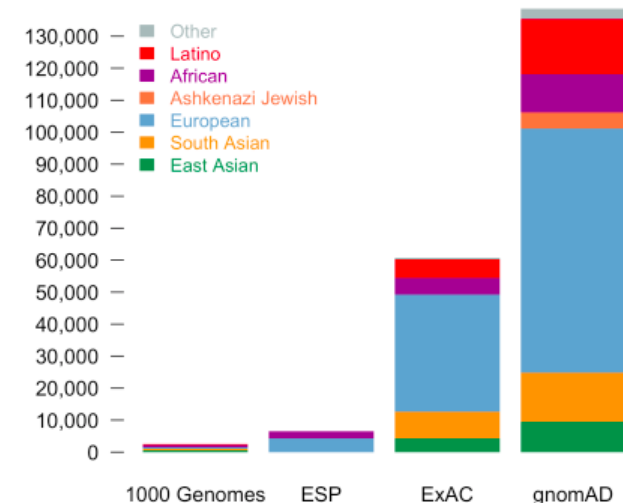
Recent News

February 27, 2017
[Official gnomAD release \(version 2.0\)](#) with browser updates and data available for [download](#).

October 19, 2016
Public release of gnomAD Browser (beta) at ASHG!

<http://gnomad.broadinstitute.org/>

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

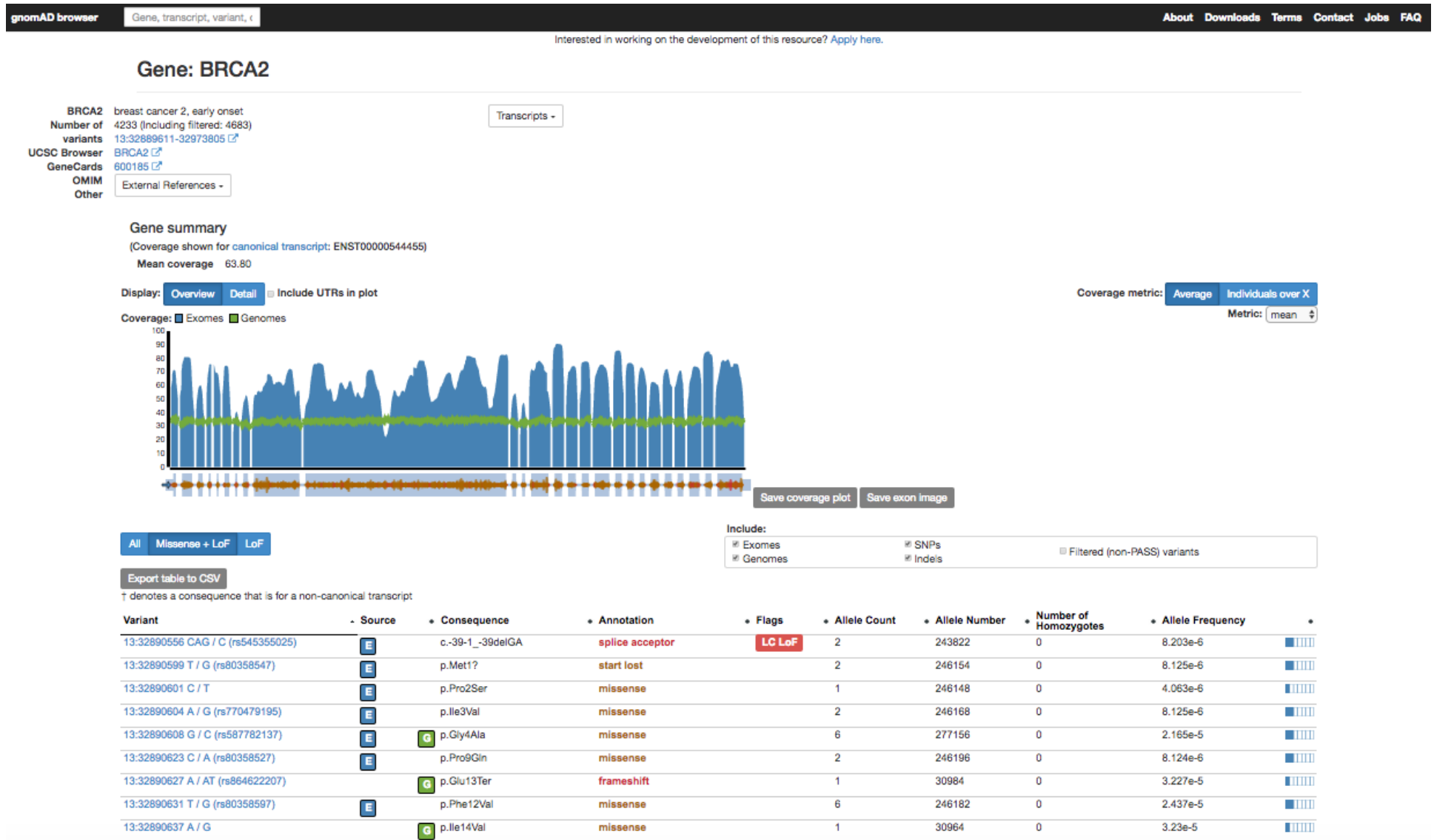


POPULATION	DESCRIPTION	GENOMES	EXOMES	TOTAL
AFR	African/African American	4,368	7,652	12,020
AMR	Admixed American	419	16,791	17,210
ASJ	Ashkenazi Jewish	151	4,925	5,076
EAS	East Asian	811	8,624	9,435
FIN	Finnish	1,747	11,150	12,897
NFE	Non-Finnish European	7,509	55,860	63,369
SAS	South Asian	0	15,391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

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

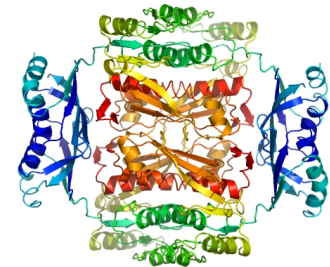
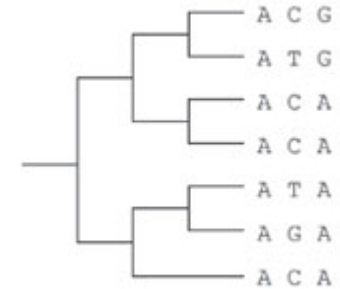
Variant annotation

Reference datasets: gnomAD



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 1000 is incorrect)

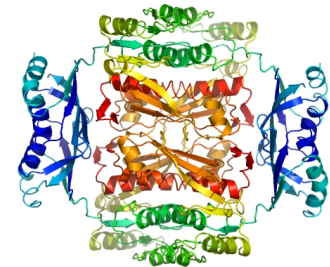
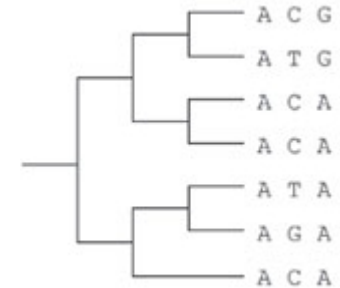


DNA sequence conservation scores

- **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)

Deleteriousness scores

- **SIFT**: functional prediction, protein sequence conservation among homologs. Score: 0 (benign) - 1 (damaging)
- **PolyPhen**: functional prediction based on protein sequence and structure features. Score: 0 (benign) - 1 (damaging)
- **CADD**: ensemble score combining 62 distinct variant annotation features retrieved from Ensembl browser. Phred score (0 is incorrect)



DNA sequence conservation scores

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

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/>




UNIVERSITY OF
CAMBRIDGE



 **HGVA v1.0.0** Variant Browser Studies ▾ Q About ▾

Projects / [hgvauser@reference_grch37](#) / [1kG_phase3](#)

The Human Genetic Variation Archive (HGVA)



The **Human Genomic Variation Archive (HGVA)** is an open access genetic variation resource that integrates all variants from key world-wide reference projects, but also added-value information such as basic variant annotation, population frequencies, protein effect predictions, variant-associated phenotypes, etc.

HGVA currently hosts about 300GB of data from 13 different studies describing more than 200 million variants. HGVA is not a mere data archive, but a big data provider that enables users to efficiently query, filter and retrieve relevant information from its knowledge-base, either from a visual web-interface or programmatically.

Example: BRCA2, ENST00000342992, rs666, 10:15097577:G:C, 1:1-100000, GO:0000145, HP:0001756

Current selected Project and Study are [hgvauser@reference_grch37](#) and **1000 Genomes Project - Phase 3**. [Click to change](#)

Note:

HGVA web application makes an intensive use of the HTML5 standard and other cutting-edge web technologies such as Web Components, so only modern web browsers are fully supported, these include Chrome 49+, Firefox 45+, Microsoft Edge 14+, Safari 10+ and Opera 36+.

The Human Genomic Variation Archive (HGVA)

<http://hgva.opencb.org/>

OpenCB

HGVA v1.0.0

Variant Browser

Studies

Search

About

Projects / hgvauser@reference_grch37 / 1kG_phase3

Variant Browser

Search

Clear

No filters selected

Download

Share

Study

Studies Filter

In (AND)

☒ 1kG_phase3

☐ ESP6500

☐ EXAC

☐ 1kG_phase3_chrY

☐ 1kG_phase3_chrMT

☐ GONL

☐ UK10K_ALSPAC

☐ UK10K_TWINSUK

☐ MGP

Genomic

Chromosomal Location

3:444-55555, 1:1-100000

Feature IDs (gene, transcript, SNP, ...)

Search for Gene Sy

+

Download

Share

Variant

SNP Id

Genes

Type

Consequence Type

Deleteriousness

SIFT

Polyphen

CADD

Conservation

PhyloP

PhastCons

GERP

Population Frequencies

Legend: Red Blue Black

1000 Genomes

ExAC

ESP6500

21:46047686 C/T	rs116600158	KRTAP10-9,TSPEAR	SNV	missense_variant	tolerated	-	4.88	0.563	0.016	1.490	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>
21:46047670 A/G	rs144666411	KRTAP10-9,TSPEAR	SNV	synonymous_variant	-	-	0.05	0.491	0.122	-4.350	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>
21:46047710 T/C	rs8131142	KRTAP10-9,TSPEAR	SNV	synonymous_variant	-	-	0.00	0.533	0.003	-1.360	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>
21:46047512 G/A	rs138753798	KRTAP10-9,TSPEAR	SNV	missense_variant	deleterious	-	10.72	-0.675	0.033	0.173	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>
21:46047629 T/G	rs200060673	KRTAP10-9,TSPEAR	SNV	missense_variant	tolerated	-	0.00	-0.142	0.009	-4.120	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>
21:46047728 G/A	rs78393062	KRTAP10-9,TSPEAR	SNV	missense_variant	tolerated	benign	0.00	0.533	0.148	-7.000	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>
21:46047821 G/A	rs373246520	KRTAP10-9,TSPEAR	SNV	missense_variant	tolerated	possibly damaging	17.83	-2.472	0.003	0.171	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>
21:46047968 C/T	rs201452080	KRTAP10-9,TSPEAR	SNV	3_prime_UTR_variant intron_variant	-	-	4.50	-0.371	0.043	-5.940	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>

THANK YOU.