

Variant data analysis and prioritization using HGVA

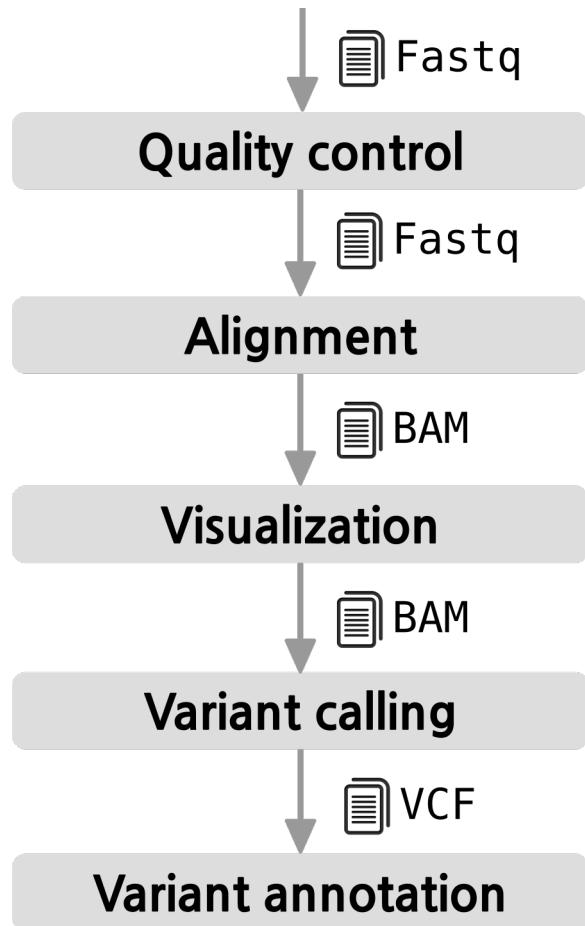
TrainMALTA
Cambridge, UK
2nd June 2017

Marta Bleda Latorre
mb2033@cam.ac.uk
Research Associate at the Department of Medicine
University of Cambridge
Cambridge, UK



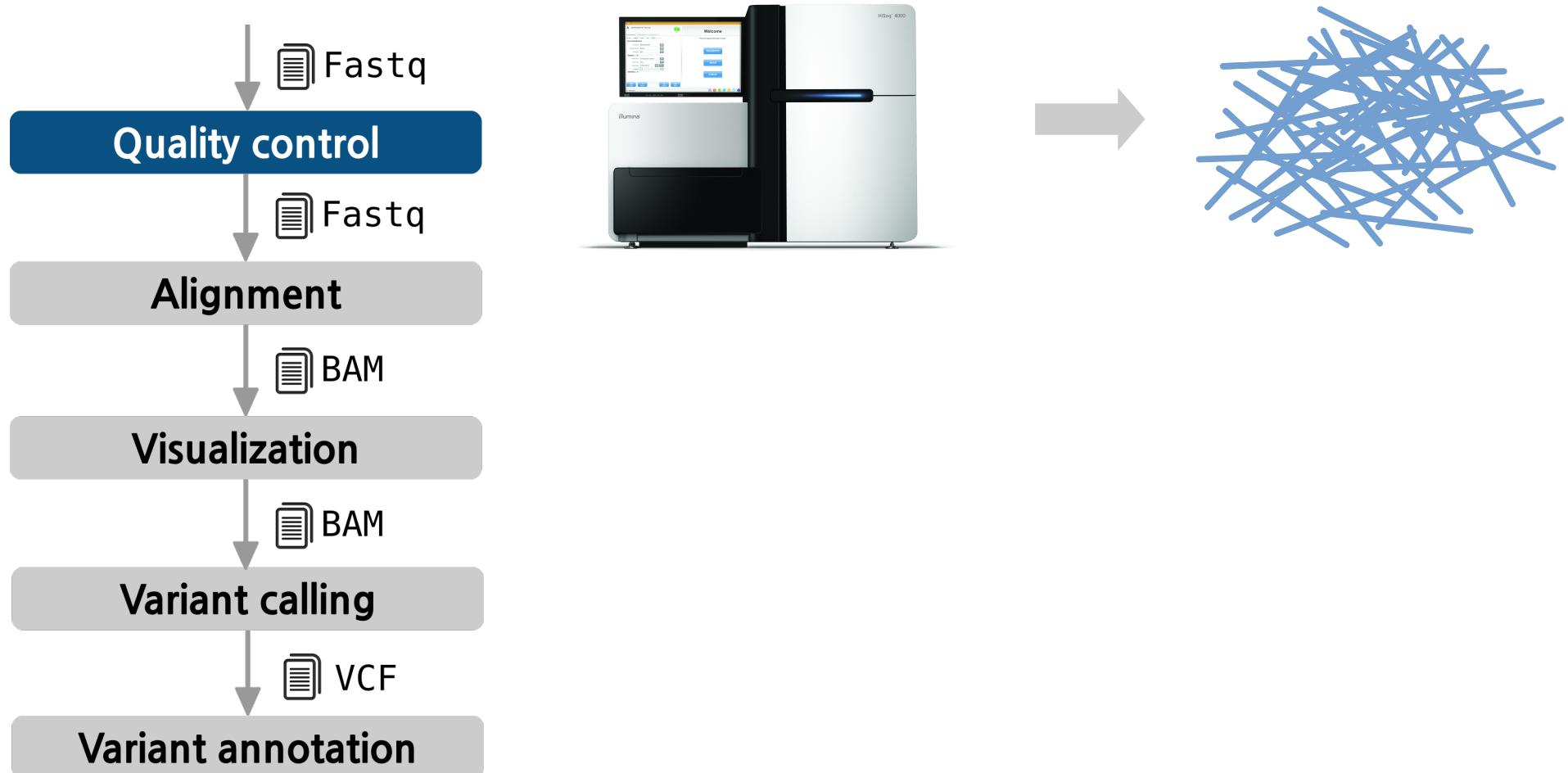
UNIVERSITY OF
CAMBRIDGE

The pipeline



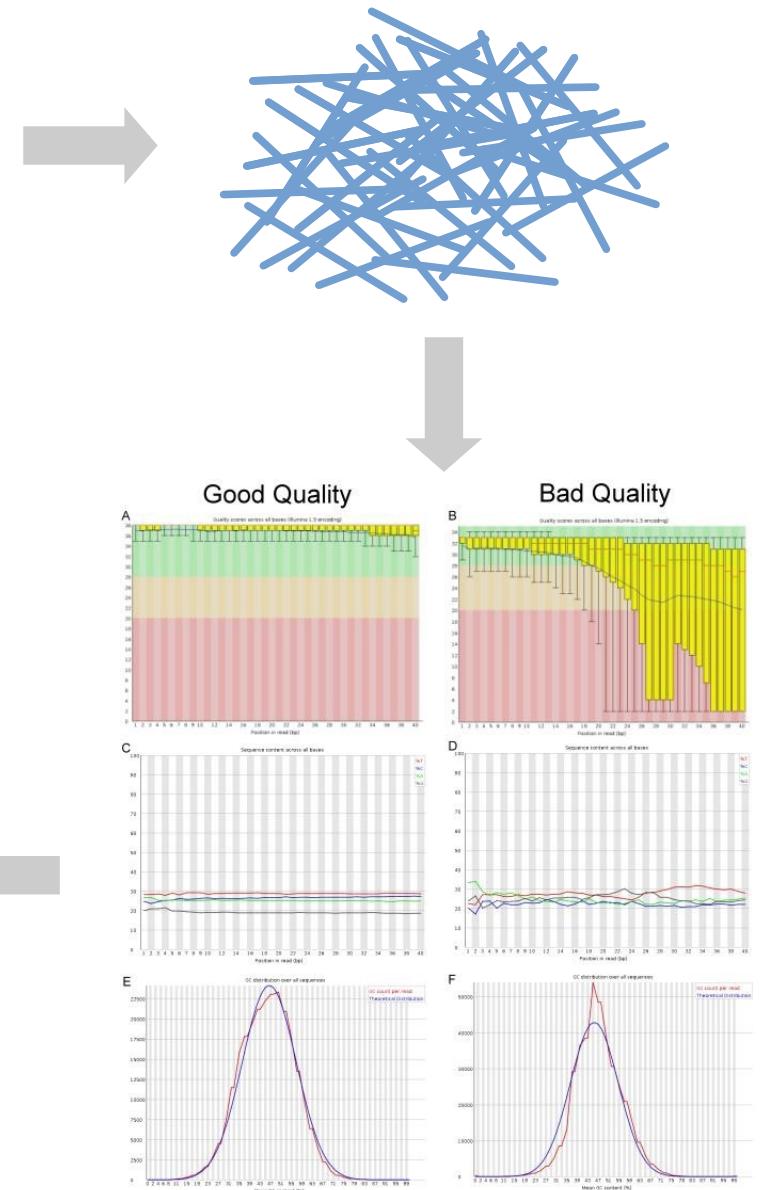
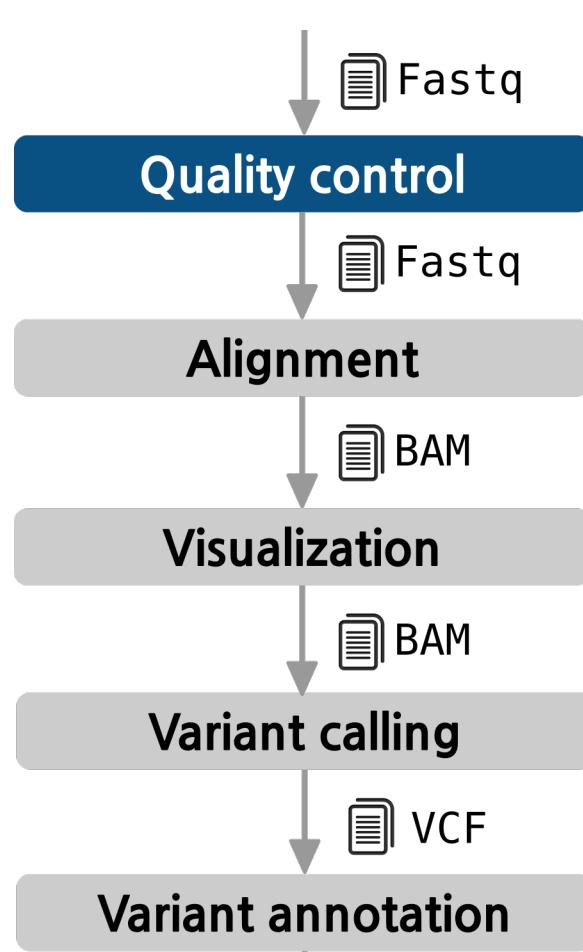
The pipeline

Fastq QC



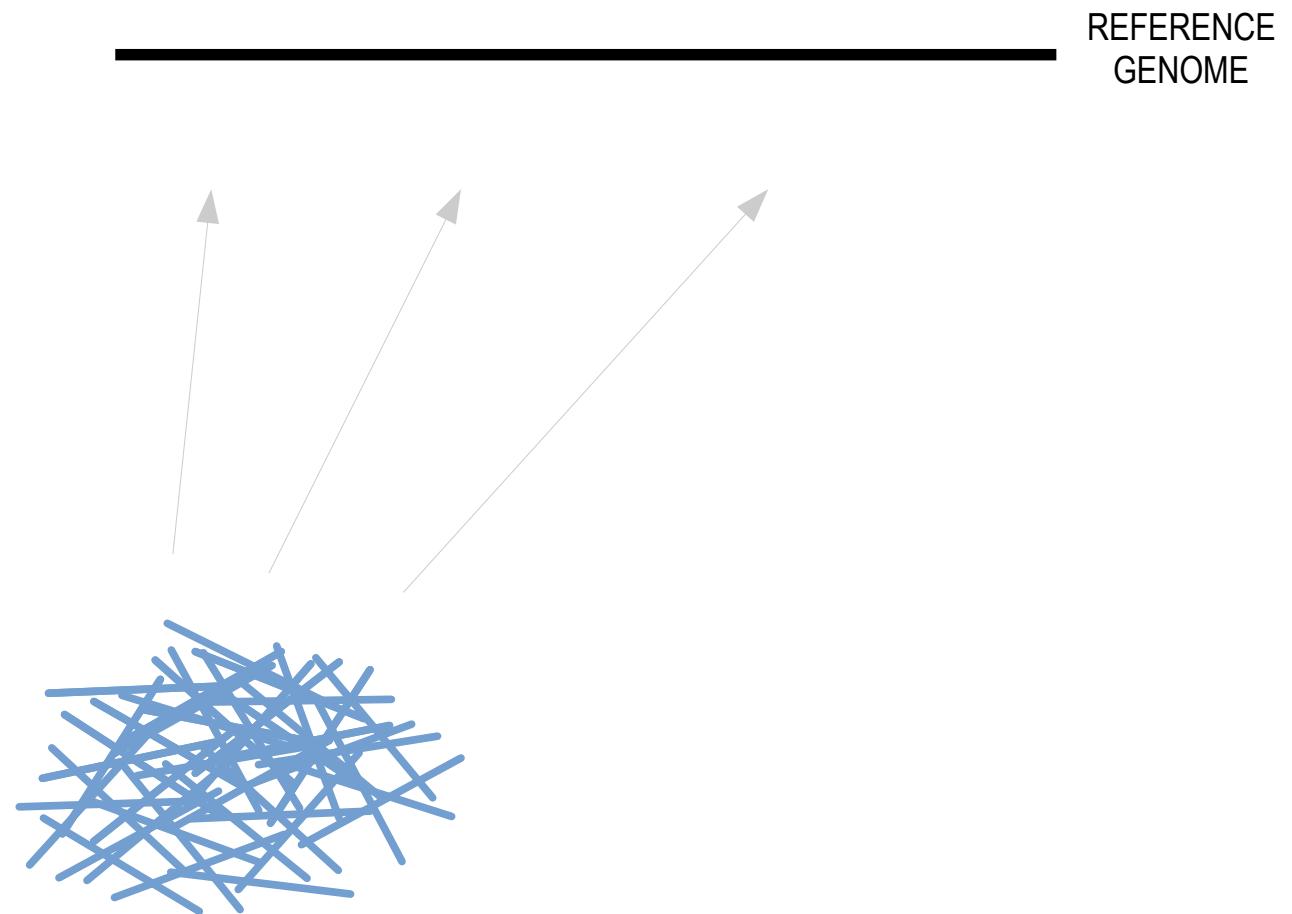
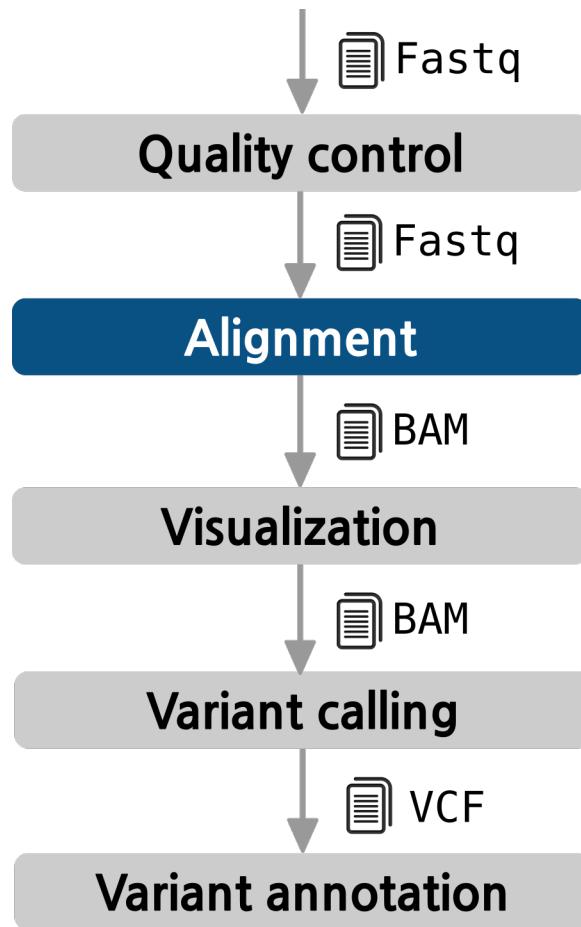
The pipeline

Fastq QC



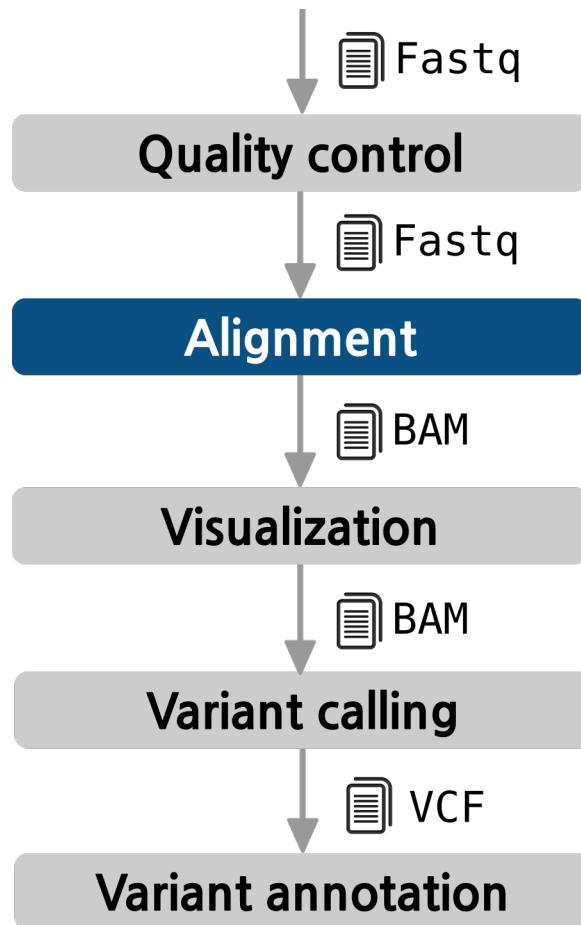
The pipeline

Alignment



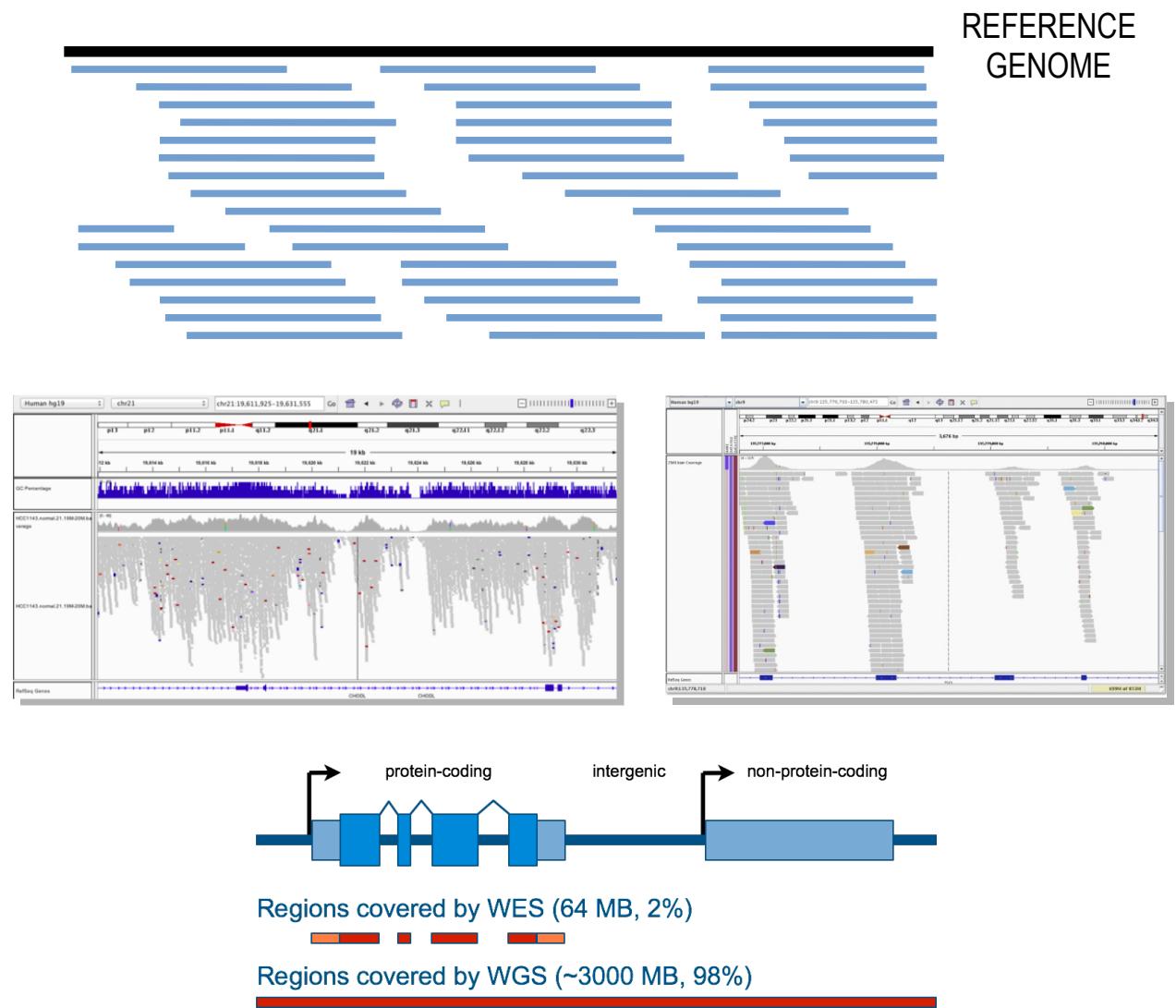
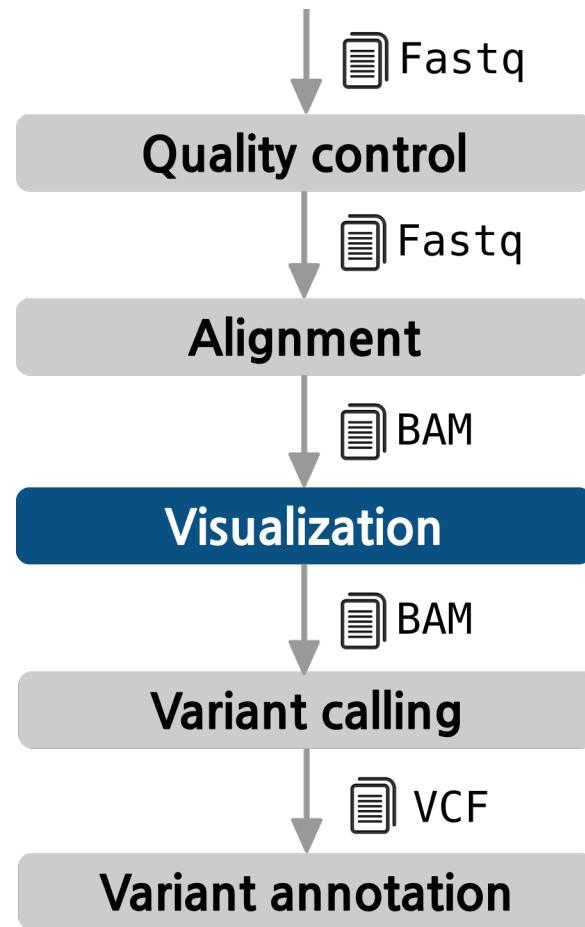
The pipeline

Alignment



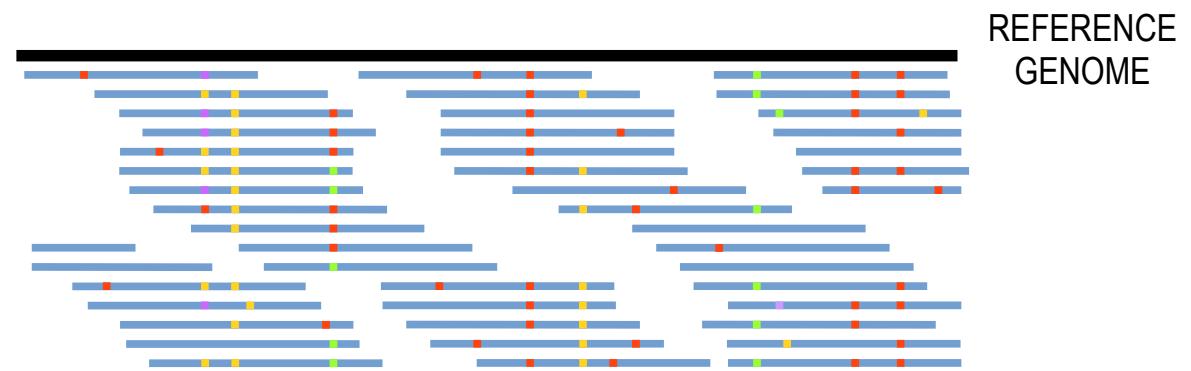
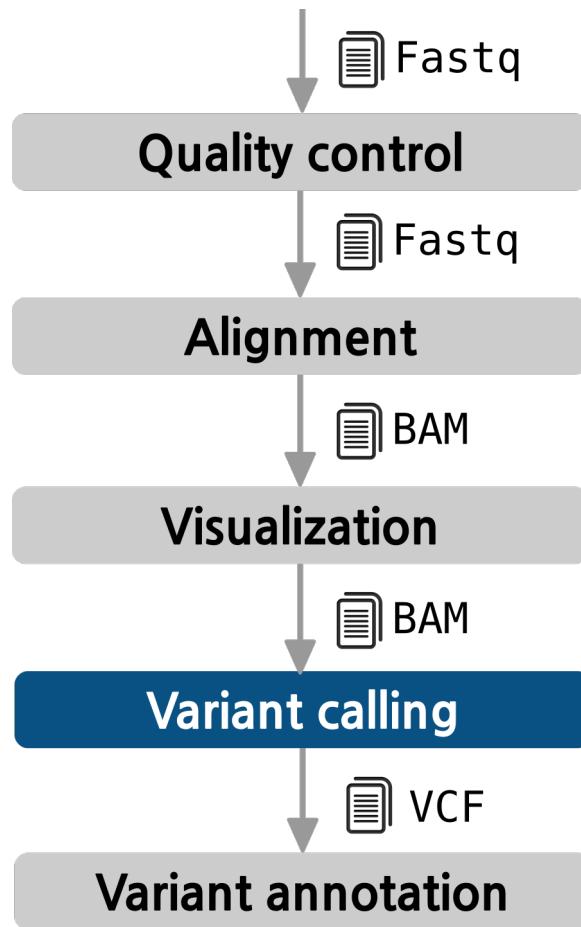
The pipeline

Visualization



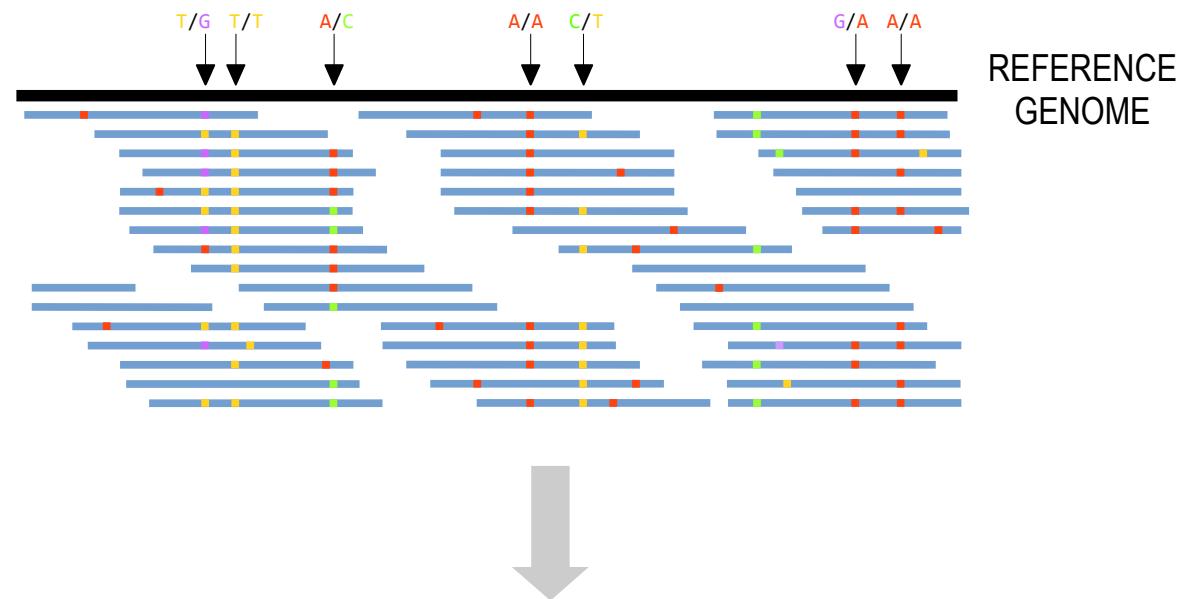
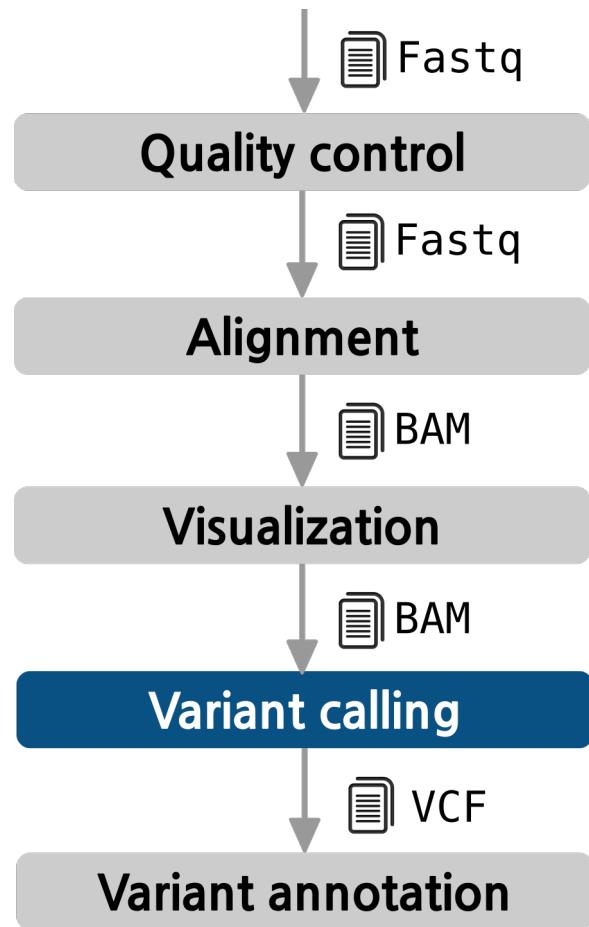
The pipeline

Variant Calling



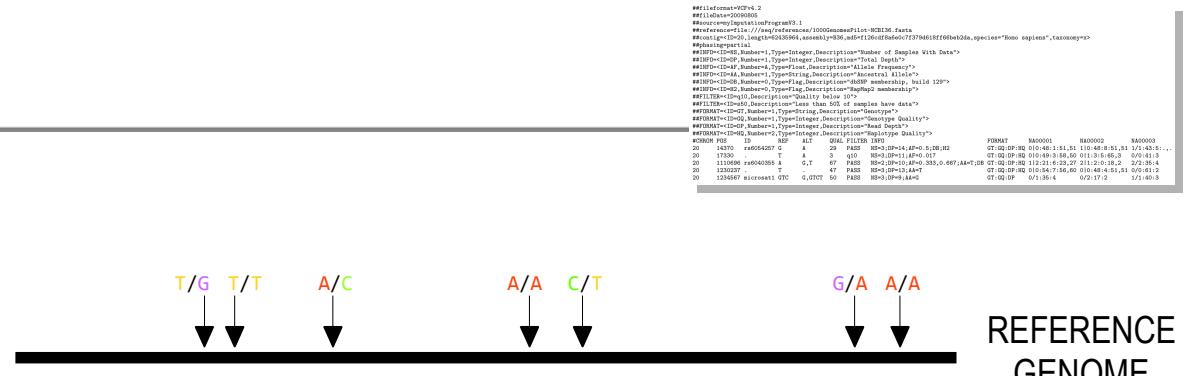
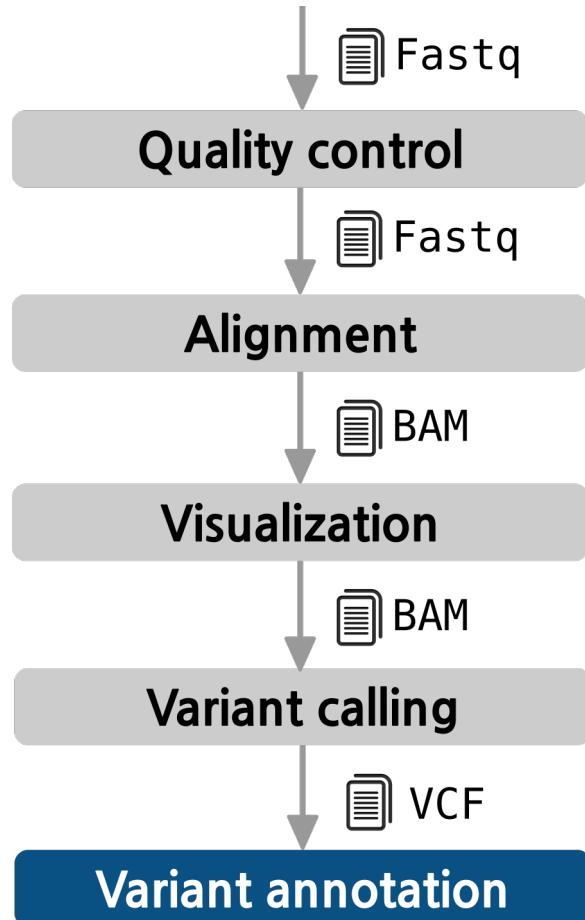
The pipeline

Variant Calling

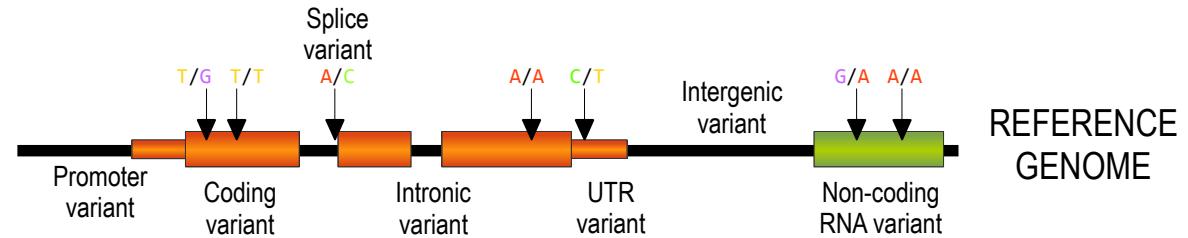
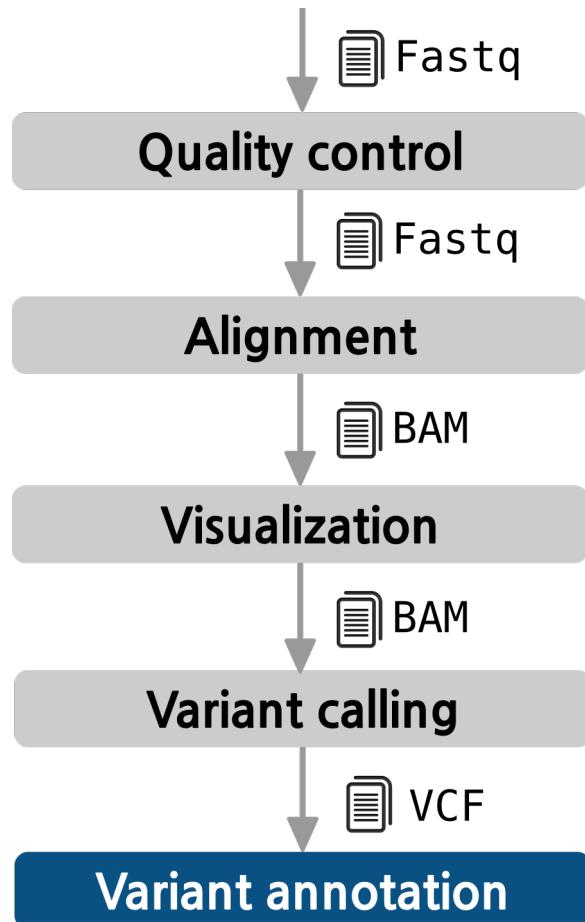


```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=1,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



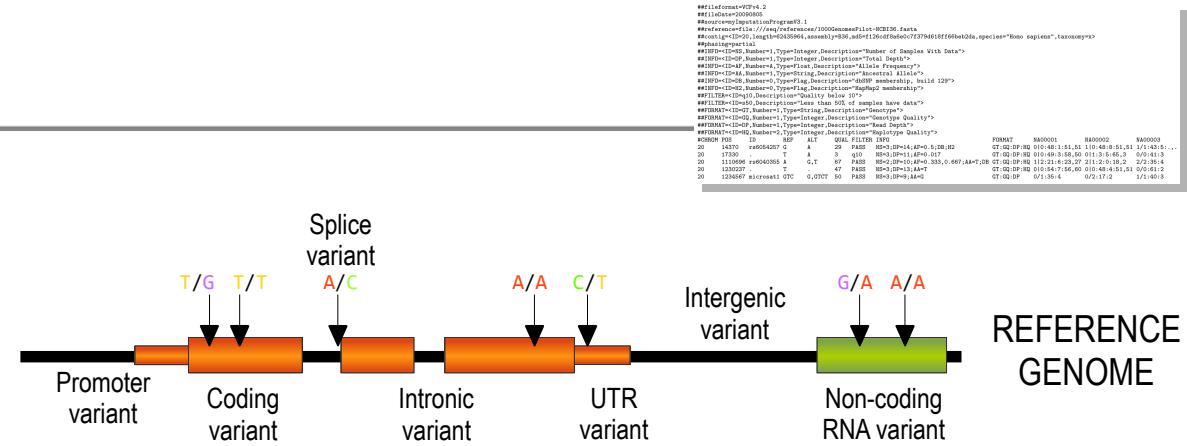
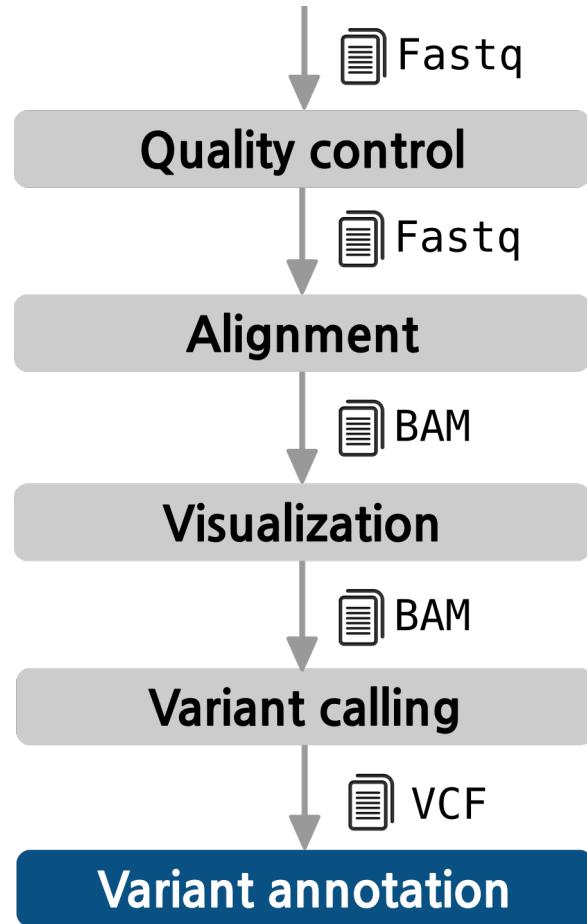
The pipeline



REFERENCE GENOME

Marta Bleda | Variant data analysis and prioritization using HGVA

The pipeline



Challenges

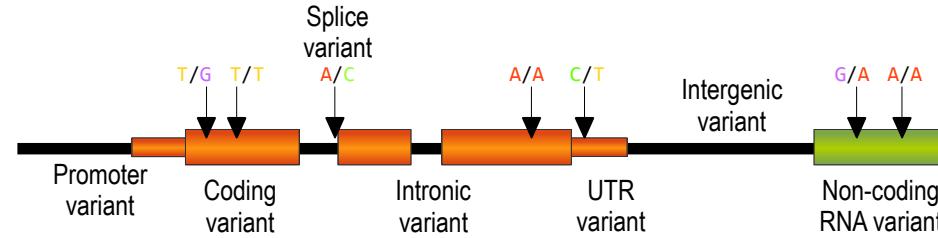
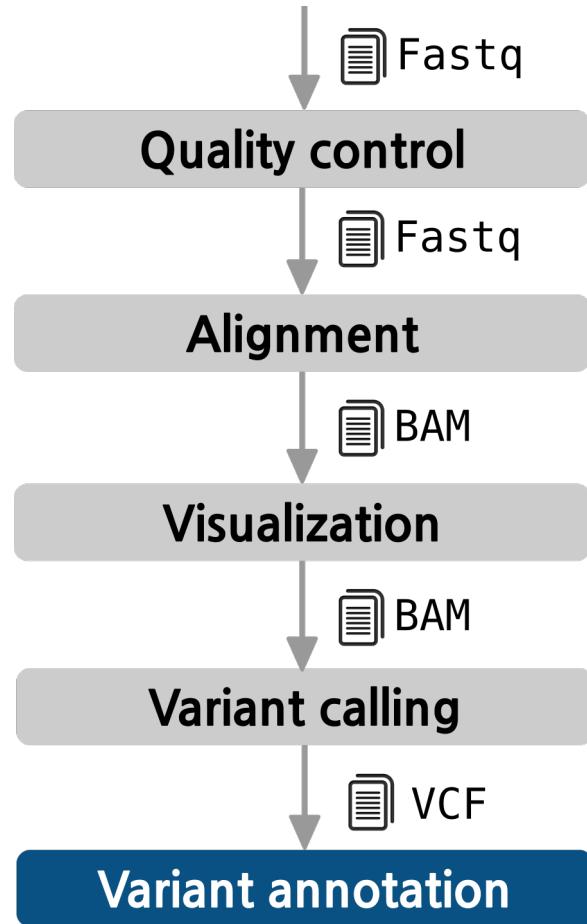
- An individual **exome** carries between 25,000 and 50,000 variants
 - A **whole genome** can carry 3.5 million variants on average
 - After annotating there will be **hundreds** of deleterious variants

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

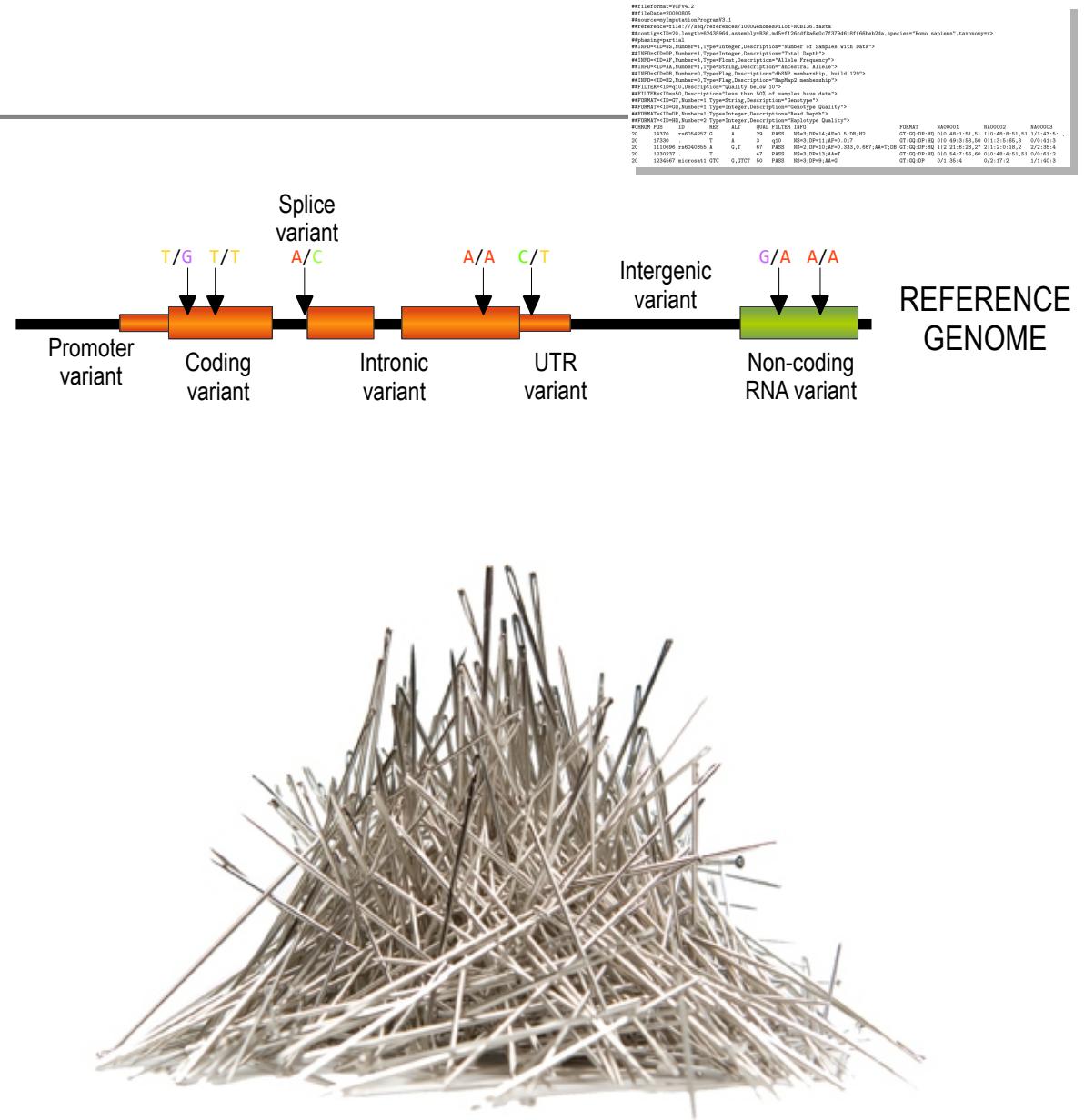
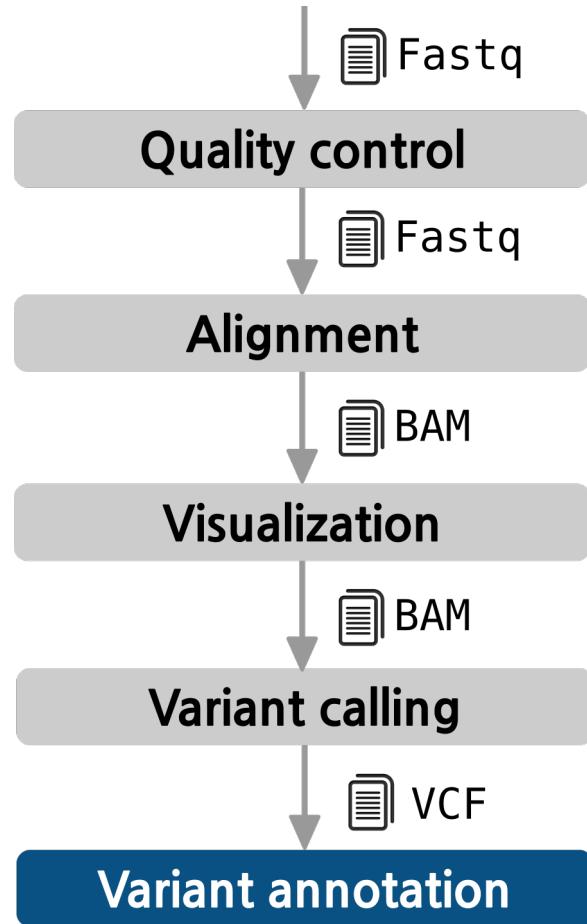


REFERENCE GENOME

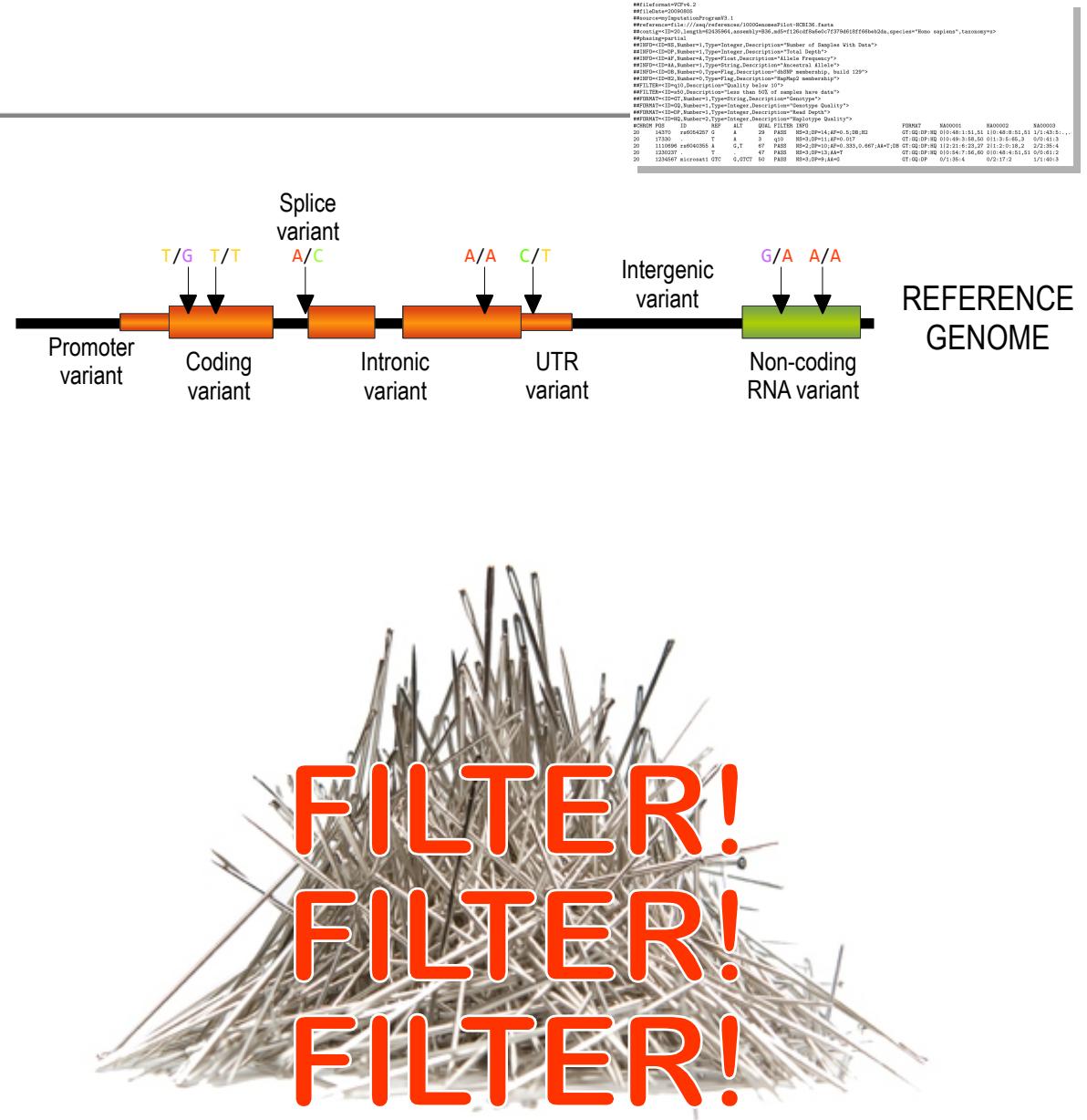
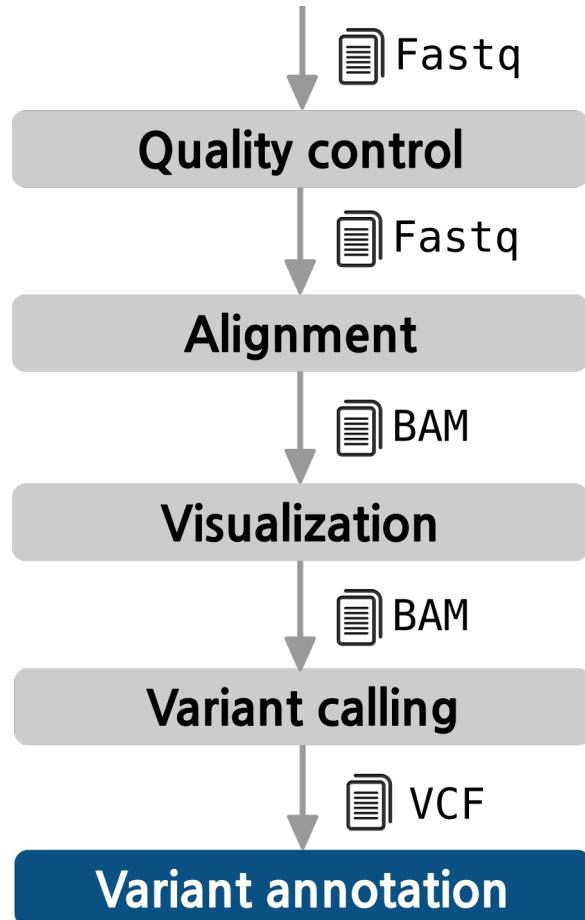


Marta Bleda | Variant data analysis and prioritization using HGVA

The pipeline

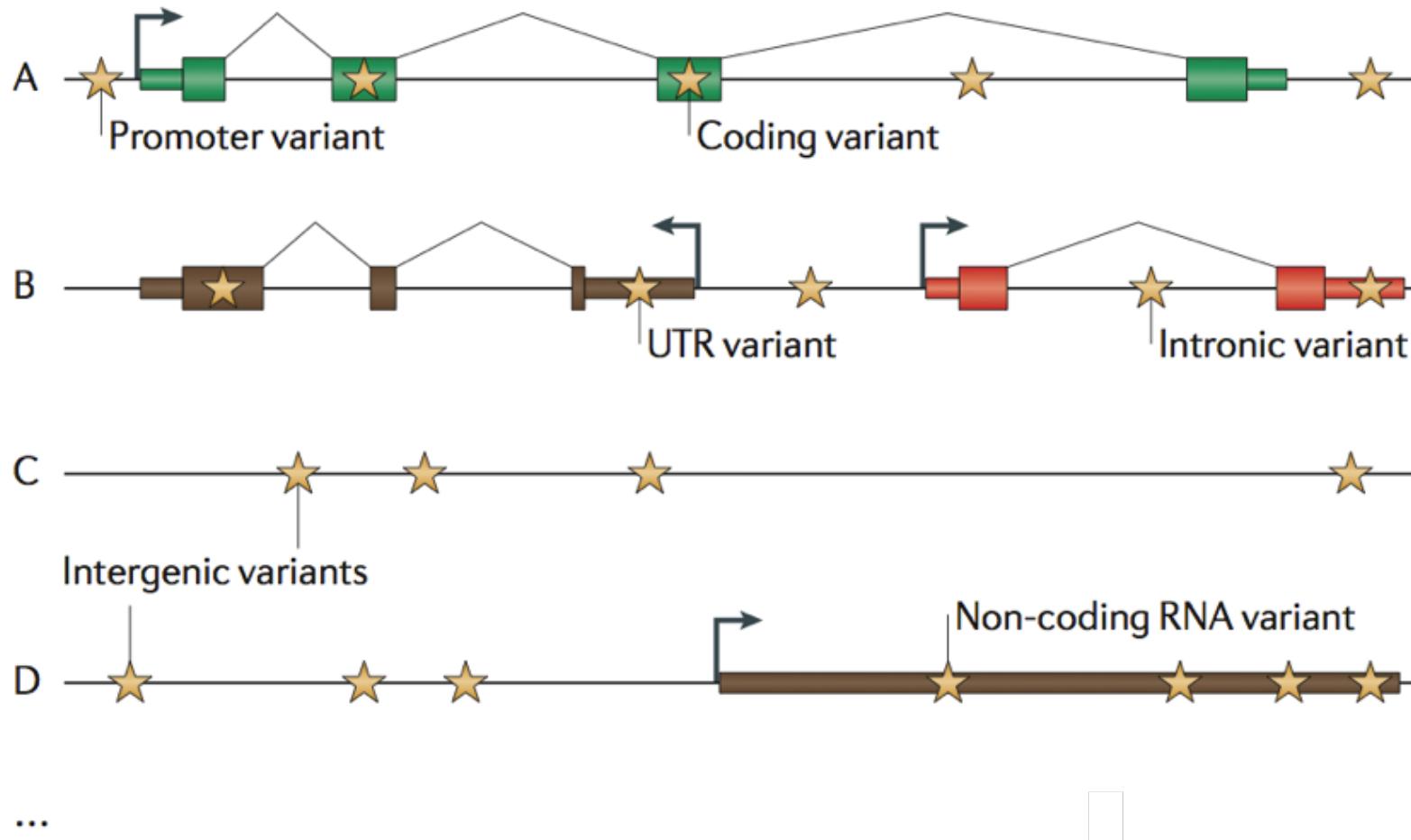


The pipeline



Variant annotation

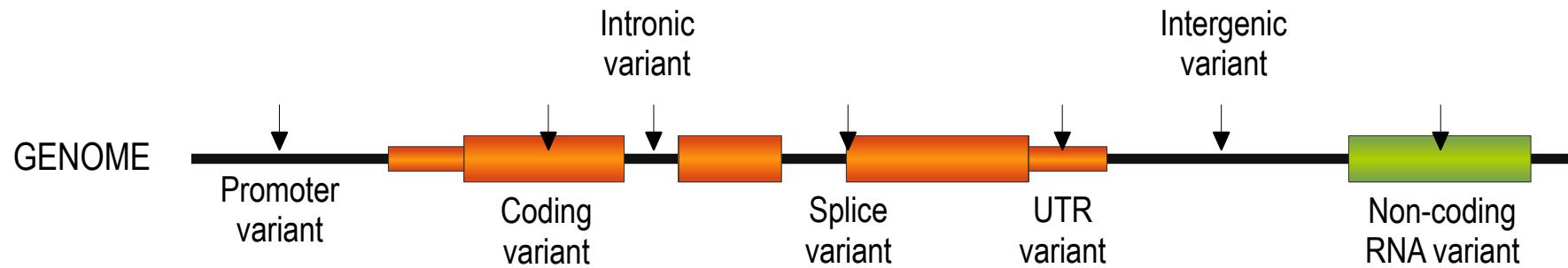
Consequence types



Cooper et al., 2011

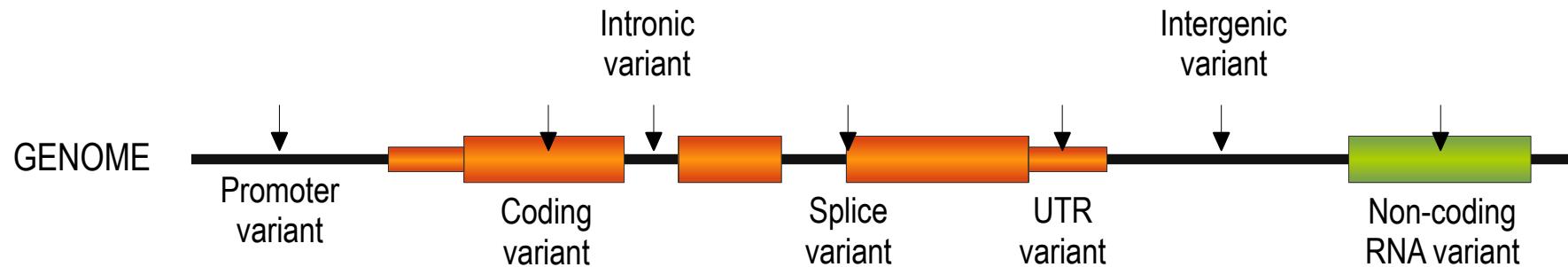
Variant annotation

External databases



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)

Exome Aggregation Consortium (ExAC)

ExAC Browser Beta

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

ExAC Browser (Beta) | Exome Aggregation Consortium

Search for a gene or variant or region

Examples - Gene: [PCSK9](#), Transcript: [ENST00000407236](#), Variant: [22-46615880-T-C](#), Multi-allelic variant: [rs1800234](#), Region: [22:46615715-46615880](#)

About ExAC

The Exome Aggregation Consortium (ExAC) is a coalition of investigators seeking to aggregate and harmonize exome sequencing data from a wide variety of large-scale sequencing projects, and to make summary data available for the wider scientific community.

The data set provided on this website spans 60,706 unrelated individuals sequenced as part of various disease-specific and population genetic studies. The ExAC Principal Investigators and groups that have contributed data to the current release are listed [here](#).

All data here are released under a [Fort Lauderdale Agreement](#) for the benefit of the wider biomedical community - see the terms of use [here](#).

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

Recent News

August 8, 2016
- CNV calls are now available on the ExAC browser

March 14, 2016
- Version 0.3.1 ExAC data and browser (beta) is released! ([Release notes](#))

January 13, 2015
- Version 0.3 ExAC data and browser (beta) is released! ([Release notes](#))

October 29, 2014
- Version 0.2 ExAC data and browser (beta) is released! Sign up for our mailing list for future release announcements [here](#).

October 20, 2014
- Public release of ExAC Browser (beta) at ASHG!

October 15, 2014
- Internal release to consortium now available!

<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-DO 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

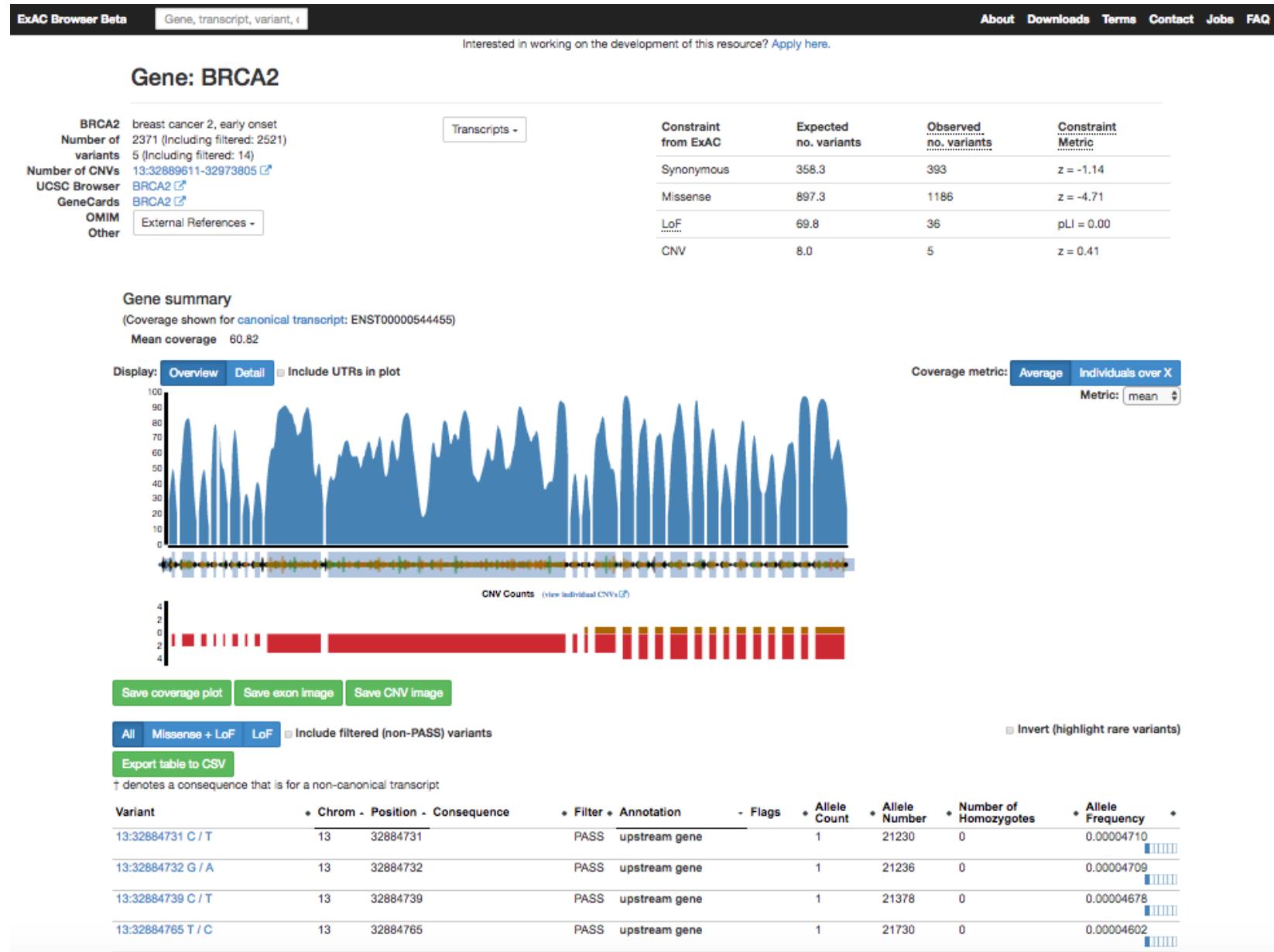
doi:10.1038/nature19057

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



Variant annotation

Reference datasets: ExAC

ExAC Browser Beta Gene, transcript, variant, ... Interested in working on the development of this resource? [Apply here.](#)

[About](#) [Downloads](#) [Terms](#) [Contact](#) [Jobs](#) [FAQ](#)

Gene: BRCA2

BRCA2 breast cancer 2, early onset
Number of variants 2371 (Including filtered: 2521)
5 (Including filtered: 14)
Number of CNVs 13:32889611-32973805 [View](#)
UCSC Browser [BRCA2](#)
GeneCards [BRCA2](#)
OMIM [External References](#)

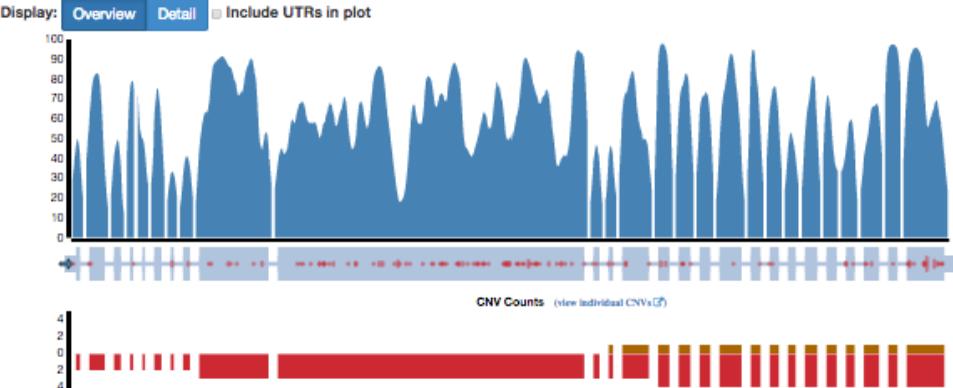
Transcripts	Constraint from ExAC	Expected no. variants	Observed no. variants	Constraint Metric
Transcripts	Synonymous	358.3	393	$z = -1.14$
	Missense	897.3	1186	$z = -4.71$
	LoF	69.8	36	pLI = 0.00
	CNV	8.0	5	$z = 0.41$

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 Invert (highlight rare variants)

[Export table to CSV](#)

† 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

About Downloads Terms Contact Jobs FAQ

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

gnomAD browser | genome Aggregation Database

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](#).

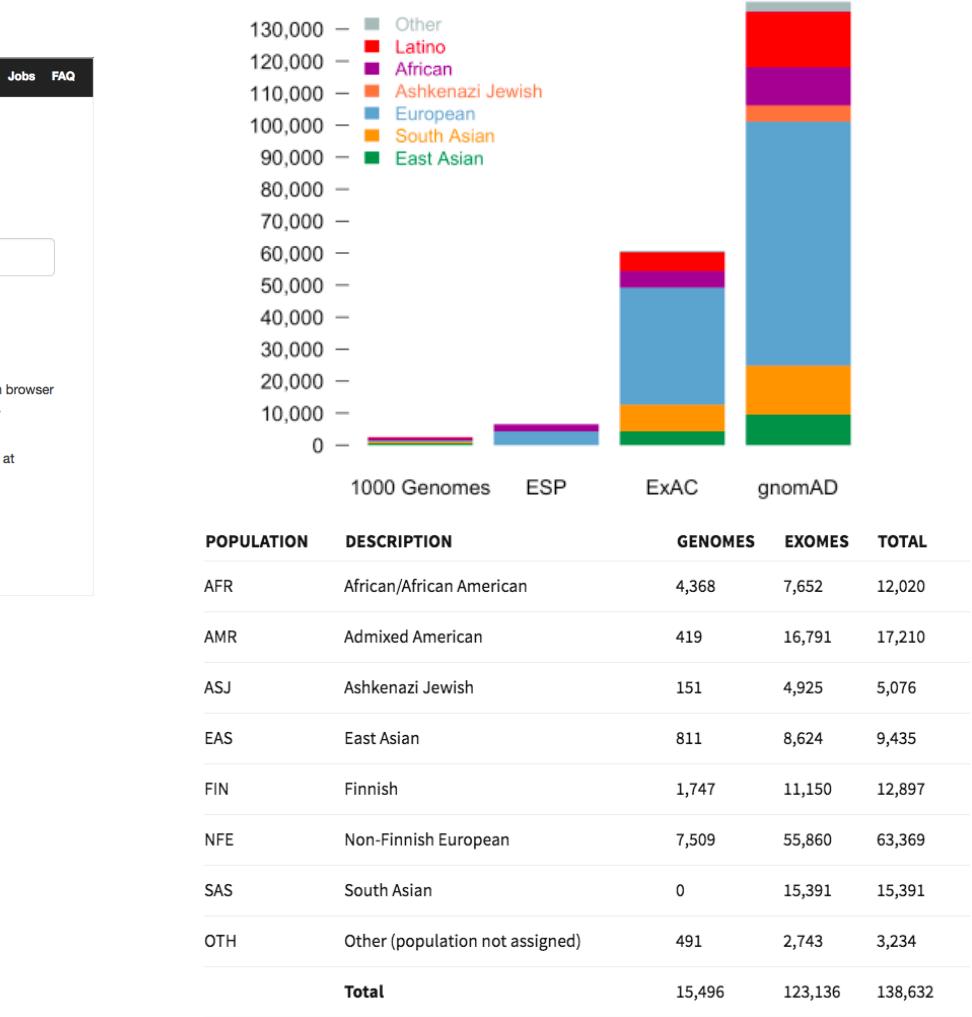
Recent News

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/>

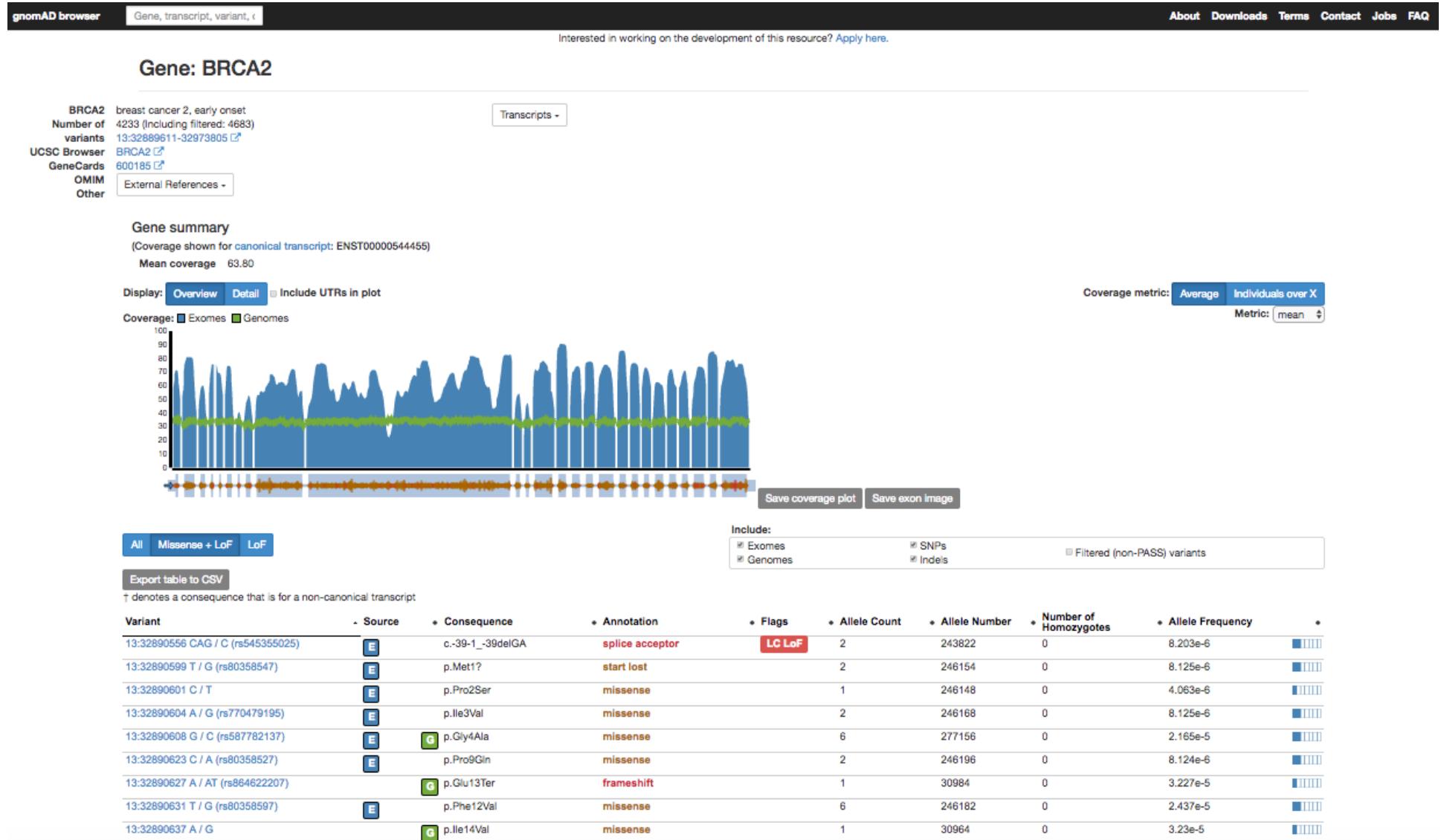
- Released on February 2017
- Two callsets:
 - 123,136 exomes
 - 15,496 whole genomes
- Exomes and genomes called separately but analyzed together



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

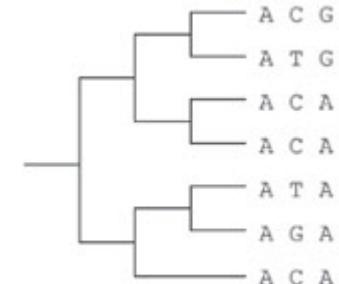
Variant annotation

Reference datasets: gnomAD



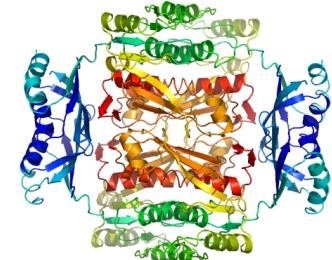
Deleteriousness scores

- **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)



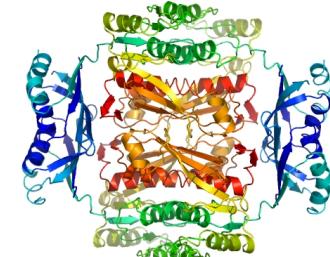
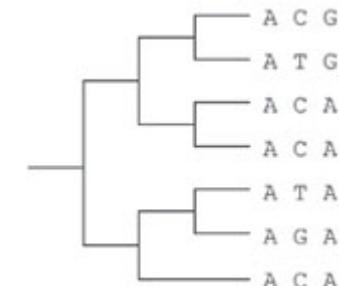
DNA sequence conservation scores

- **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 base on various informative genome-wide annotations
- **PolyPhen**: functional prediction based on protein sequence and structure features. Score: 0 (benign), 1 (damaging)
- **CADD**: ensemble score combining GERP, PolyPhen and SIFT features retrieved from browser. Phred score (incorrect)



DNA sequence conservation scores

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

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/>



OpenCB HGVA v1.0.0 Variant Browser Studies ▾ Search About ▾

Projects / hgvauser@reference_grch37 / 1kG_phase3

The Human Genetic Variation Archive (HGVA)

OpenCB 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.

Search for a gene, transcript, variant, multi-allelic variant, region, GO term or HPO term

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 ▾

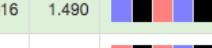
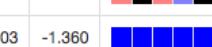
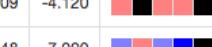
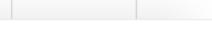
Projects / hgvauser@reference_grch37 / 1kG_phase3

Variant Browser

Search

Clear No filters selected

Download Share

Variant	SNP Id	Genes	Type	Consequence Type	Deleteriousness			Conservation			Population Frequencies		
					SIFT	PolypHEN	CADD	PhyloP	PhastCons	GERP	1000 Genomes	ExAC	ESP6500
21:46047686 C/T	rs116600158	KRTAP10-9,TSPEAR	SNV	missense_variant	tolerated	-	4.88	0.563	0.016	1.490			
21:46047670 A/G	rs144666411	KRTAP10-9,TSPEAR	SNV	synonymous_variant	-	-	0.05	0.491	0.122	-4.350			
21:46047710 T/C	rs8131142	KRTAP10-9,TSPEAR	SNV	synonymous_variant	-	-	0.00	0.533	0.003	-1.360			
21:46047512 G/A	rs138753798	KRTAP10-9,TSPEAR	SNV	missense_variant	deleterious	-	10.72	-0.675	0.033	0.173			
21:46047629 T/G	rs200060673	KRTAP10-9,TSPEAR	SNV	missense_variant	tolerated	-	0.00	-0.142	0.009	-4.120			
21:46047728 G/A	rs78393062	KRTAP10-9,TSPEAR	SNV	missense_variant	tolerated	benign	0.00	0.533	0.148	-7.000			
21:46047821 G/A	rs373246520	KRTAP10-9,TSPEAR	SNV	missense_variant	tolerated	possibly damaging	17.83	-2.472	0.003	0.171			
21:46047968 C/T	rs201452080	KRTAP10-9,TSPEAR	SNV	3_prime_UTR_variant intron_variant	-	-	4.50	-0.371	0.043	-5.940			

THANK YOU.