Omics in human diseases

Index

- Omics data and Biological databases
- NGS methods
- NGS data analysis
- **Prediction and interpretation of pathogenic variants**
- Protein-protein interaction networks

Course organization 2022/2023

Frontal lecture/ guided practical activity

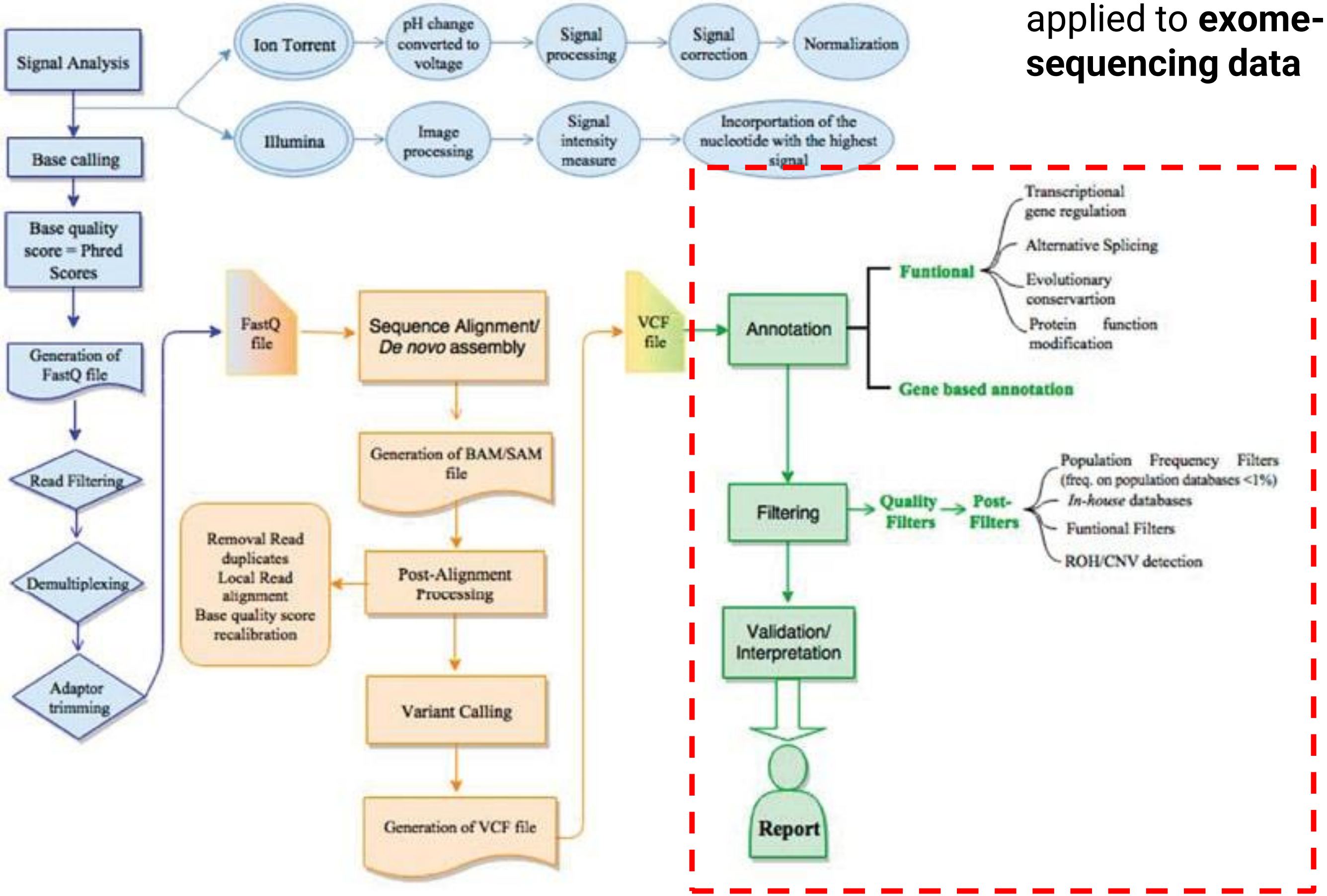
How to pass the exam: multiple choice quiz (50%) + results from practical activities (50%) + Bonus points, e.g. summary of previous lecture (up to 10%)

Mail: **emanuela.leonardi@unipd.it**

BioComputing

NGS analysis workflow

The same data analysis tools used for WGS can be applied to **exome-sequencing data**


Signal Analysis
Ion Torrent
pH change converted to voltage
Signal processing
Signal correction
Normalization
Illumina
Image processing
Signal intensity measure
Incorporation of the nucleotide with the highest signal
Base calling
Base quality score = Phred Scores
Generation of FastQ file
Read Filtering
Demultiplexing
Adaptor trimming
FastQ file
Sequence Alignment/ De novo assembly
Generation of BAM/SAM file
Post-Alignment Processing
Removal Read duplicates Local Read alignment Base quality score recalibration
Variant Calling
Generation of VCF file
VCF file
Annotation
Funtional
Transcriptional gene regulation
Alternative Splicing
Evolutionary conservartion
Protein function modification
Gene based annotation
Filtering
Quality Filters
Post-Filters
Population Frequency Filters (freq. on population databases <1%)
In-house databases
Funtional Filters
ROH/CNV detection
Validation/ Interpretation
Report

Variant filtering: criteria

- Genes list
- Sequencing parameters (filter artifactuals)
- Variant types (exonic – intronic)
- Variant class (frameshift - synonymous)
- Population frequency
- **Pathogenicity prediction score**
- **Conservation score**
- Clinical interpretation of genetic variants by the ACMG/AMP 2015 guideline
- Presence in mutation databases

(PMID: 28118812;
https://www.nature.com/articles/s41525-021-00227-3)

Variant filtering: pathogenicity scores

Computational methods to infer variant pathogenicity

- performance of current methods: 90% of pathogenicity predictions are correct, such methods **identified only 10%–20% of pathogenic variants**
- Current guidelines for clinical variant interpretation recommend that all computational methods be (at best) treated as "**weak evidence"**.
- MaveDB: "variant effect maps" , fewer than 1% of the 4,000 human disease-associated proteins

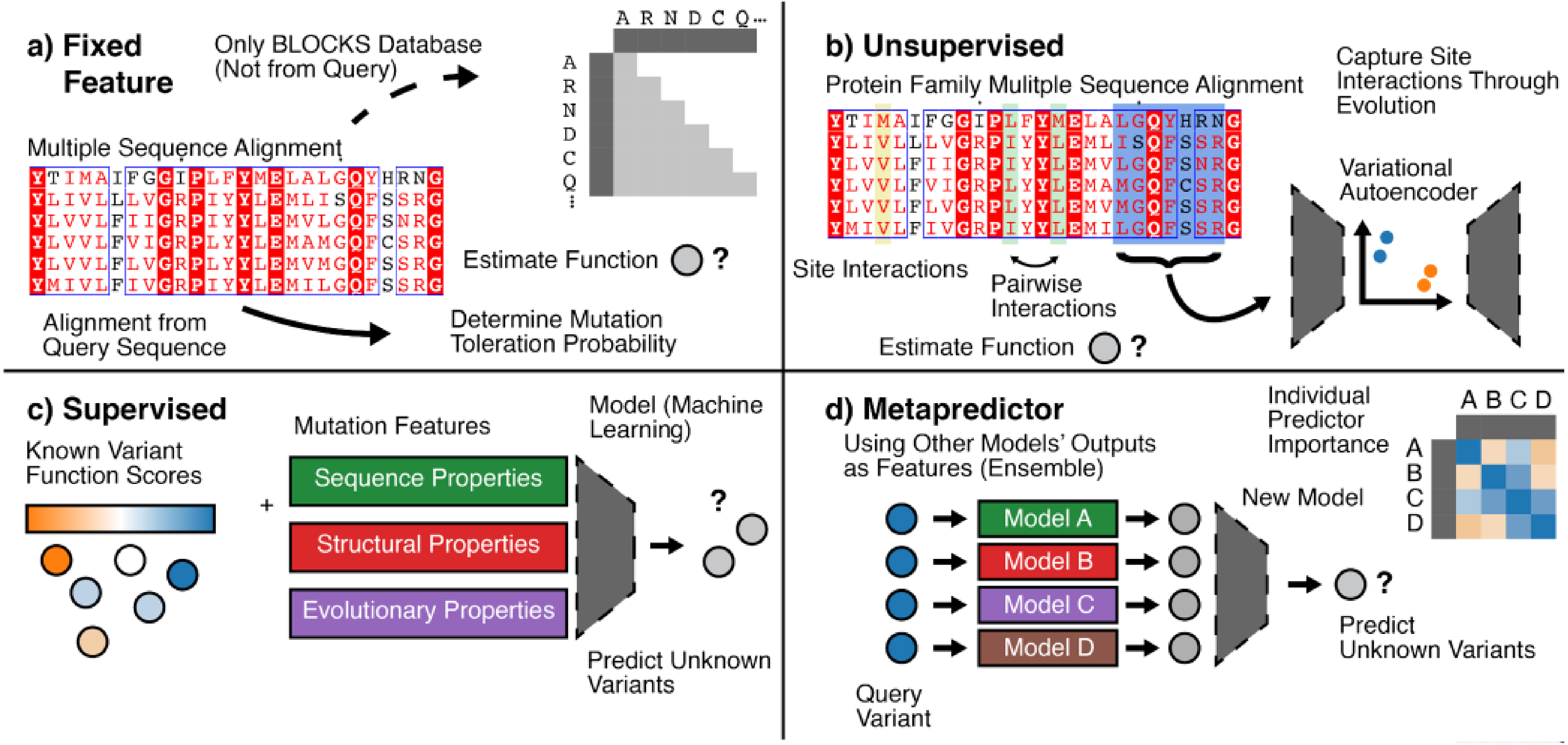
Variant filtering: Annovar pathogenicity scores

Score (dbtype)	Categorical Prediction
SIFT (sift)	D: Deleterious (sift<=0.05); T: tolerated (sift>0.05)
PolyPhen 2 HDIV (pp2 hdiv)	D: Probably damaging (>=0.957), P: possibly damaging (0.453<=pp2_hdiv<=0.956); B: benign (pp2_hdiv<=0.452)
PolyPhen 2 HVar (pp2 hvar)	D: Probably damaging (>=0.909), P: possibly damaging (0.447<=pp2_hdiv<=0.909); B: benign (pp2_hdiv<=0.446)
LRT (lrt)	D: Deleterious; N: Neutral; U: Unknown
MutationTaster (mt)	A" ("disease_causing_automatic"); "D" ("disease_causing"); "N" ("polymorphism"); "P" ("polymorphism_automatic"
MutationAssessor (ma)	H: high; M: medium; L: low; N: neutral. H/M means functional and L/N means non-functional
FATHMM (fathmm)	D: Deleterious; T: Tolerated
PROVEAN pred	
MetaSVM (metasvm)	D: Deleterious; T: Tolerated
MetaLR (metalr)	D: Deleterious; T: Tolerated
M-CAP pred	
CADD	D >25
fathmm-MKL_coding_pred	D: Deleterious; N: Neutral
GERP++ (gerp++)	higher scores are more deleterious (>3)
PhyloP (phylop)	higher scores are more deleterious
SiPhy (siphy)	higher scores are more deleterious
DANN	
Eigen	

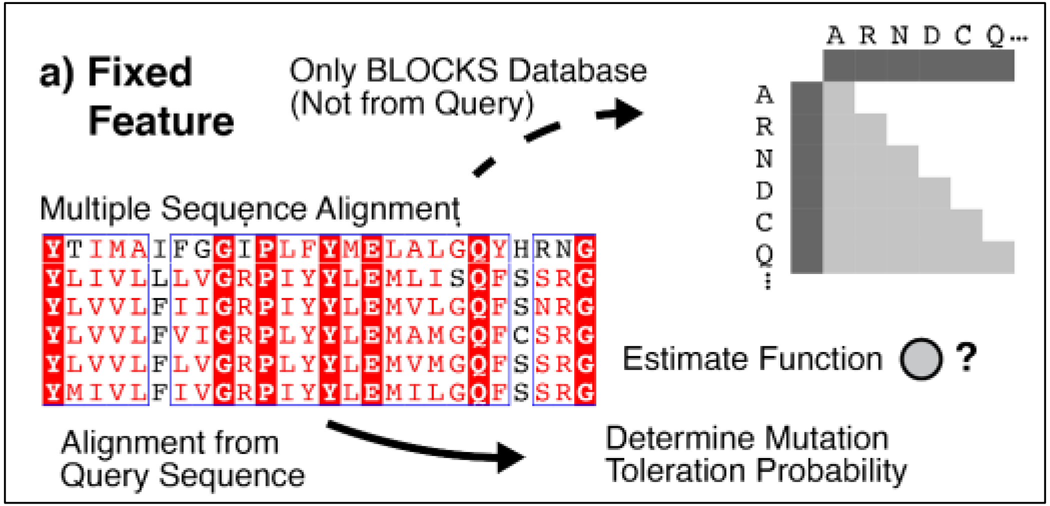
Variant filtering: pathogenicity scores

Approach	Training set	Conservation analysis	Structural attributes	Annotations
Randome forest; Alignment score; Bayesan classification; SVM (Support Vector Machine); Machine learning	HGMD, Swiss Prot; Protein Mutant Database	PSIC, position-specific independent counts; PFAM; PSI-BLAST; Sequence environment, sequence profiles	Predicted attributes; Homologue mapping	Swiss-Prot; Pfam domain; GO

Variant filtering: Variant effect prediction (VEP) tools


a) Fixed Feature
Only BLOCKS Database (Not from Query)
A R N D C Q ...
A
R
N
D
C
Q
⋮
Multiple Sequence Alignment
YTIMAIFGGIPLFYMELALGQYHRNG
YLIVLLLVGRPIYYLEMLISQFSSRG
YLVVLFIIGRPIYYLEMVLGQFSNRG
YLVVLFVIGRPLYYLEMAMGQFCSRG
YLVVLFLVGRPLYYLEMVMGQFSSRG
YMIVLFIVGRPIYYLEMILGQFSSRG
Estimate Function ?
Alignment from Query Sequence
Determine Mutation Toleration Probability
b) Unsupervised
Protein Family Mulitple Sequence Alignment
Capture Site Interactions Through Evolution
YTIMAIFGGIPLFYMELALGQYHRNG
YLIVLLLVGRPIYYLEMLISQFSSRG
YLVVLFIIGRPIYYLEMVLGQFSNRG
YLVVLFVIGRPLYYLEMAMGQFCSRG
YLVVLFLVGRPLYYLEMVMGQFSSRG
YMIVLFIVGRPIYYLEMILGQFSSRG
Variational Autoencoder
Site Interactions
Pairwise Interactions
Estimate Function ?
c) Supervised
Known Variant Function Scores
+
Mutation Features
Sequence Properties
Structural Properties
Evolutionary Properties
Model (Machine Learning)
?
Predict Unknown Variants
d) Metapredictor
Using Other Models' Outputs as Features (Ensemble)
Individual Predictor Importance
A B C D
A
B
C
D
New Model
Model A
Model B
Model C
Model D
?
Predict Unknown Variants
Query Variant

Variant effect prediction (VEP) tools


a) Fixed Feature
Only BLOCKS Database (Not from Query)
A R N D C Q...
A
R
N
D
C
Q
Multiple Sequence Alignment
YTIMAIFGGIPLFYMELALGQYHRNG
YLIVLLLVGRPIYYLEMLISQFSSRG
YLVVLFIIGRPIYYLEMVLGQFSNRG
YLVVLFVIGRPLYYLEMAMGQFCSRG
YLVVLFLVGRPLYYLEMVMGQFSSRG
YMIVLFIVGRPIYYLEMILGQFSSRG
Estimate Function ?
Alignment from Query Sequence
Determine Mutation Toleration Probability

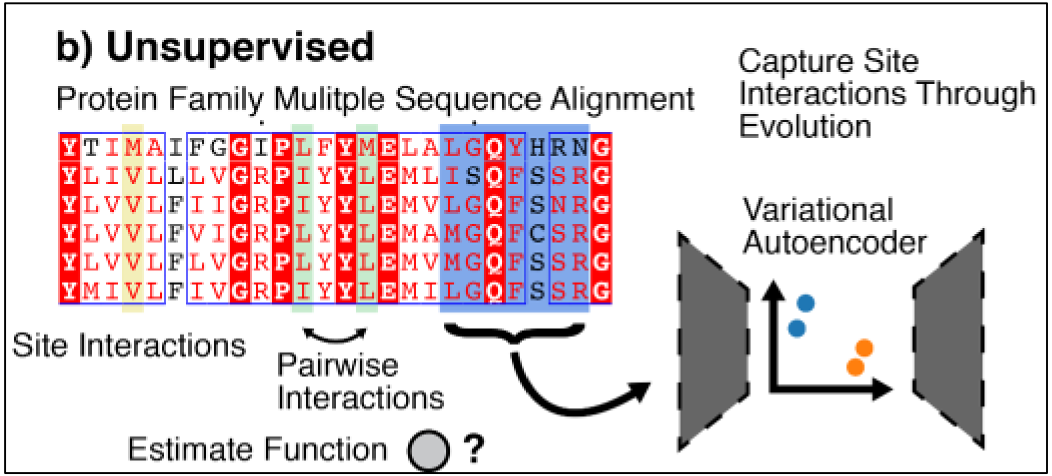
- Evaluate **mutation toleration** at a given site
- Direct statistics from given inputs and
- features
- based on averaged quantities about amino acid frequency and amino acid properties (size, charge, and hydrophobicity)

Statistical methods:

BLOSUM62, a block substitution scoring Matrix, used as the default scoring matrix for many multiple sequence alignment (MSA) methods.
SIFT (and SIFT 4G) is the commonly used static feature method

Sorting Intolerant From Tolerant

Variant effect prediction (VEP) tools


b) Unsupervised
Protein Family Mulitple Sequence Alignment
YTIMAIFGGIPLFYMELALGQYHRNG
YLIVLLLVGRPIYYLEMLISQFSSRG
YLVVLFIIGRPIYYLEMVLGQFSNRG
YLVVLFVIGRPLYYLEMAMGQFCSRG
YLVVLFLVGRPLYYLEMVMGQFSSRG
YMIVLFIVGRPIYYLEMILGQFSSRG
Site Interactions
Pairwise Interactions
Estimate Function ?
Capture Site Interactions Through Evolution
Variational Autoencoder

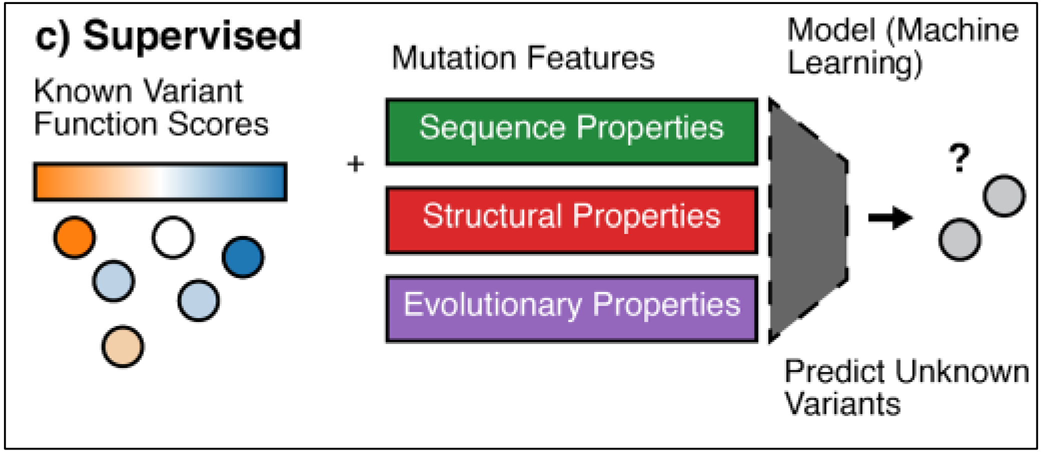
- **do not fit experimental data**
- capture the frequencies and dependencies among amino acid residues given **evolutionary pressure**
- dependencies among residues **(epistasis**) affect molecular function

EVmutation is not a deep learning-based (DL) method but rather a statistical model capturing dependencies across pairs of residues

DeepSequence, variational autoencoder framework to model the evolutionary fitness landscape, from a protein family's evolutionary history

While unsupervised methods readily generalize across protein space, they neglect to learn from the many labeled mutagenesis **datasets** appearing in the literature

Variant effect prediction (VEP) tools

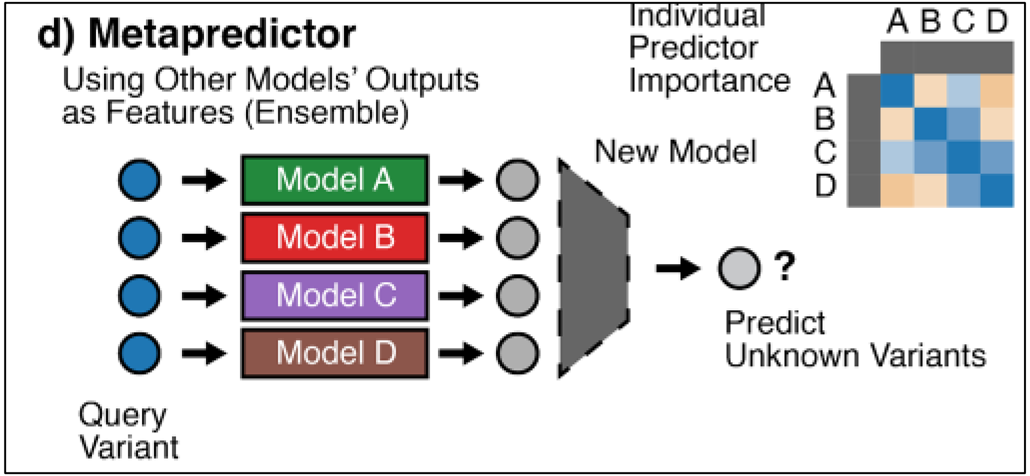


- Use large-scale datasets
- Rely on fitting models to experimental
- measurements of the fitness landscape often is derived from Deep mutational scan (DMS) experiments

SNPs&GO a support vector machine (SVM) ML model was used to categorize mutations as disease-causing or not
Envision a random forest (RF) regressor on DMS datasets of eight proteins

using datasets for assessing model performance, caution should be taken as model architectures, parameters, and hyperparameters may over-represent training set protein families, and real-world prediction accuracy may be overstated

Variant effect prediction (VEP) tools

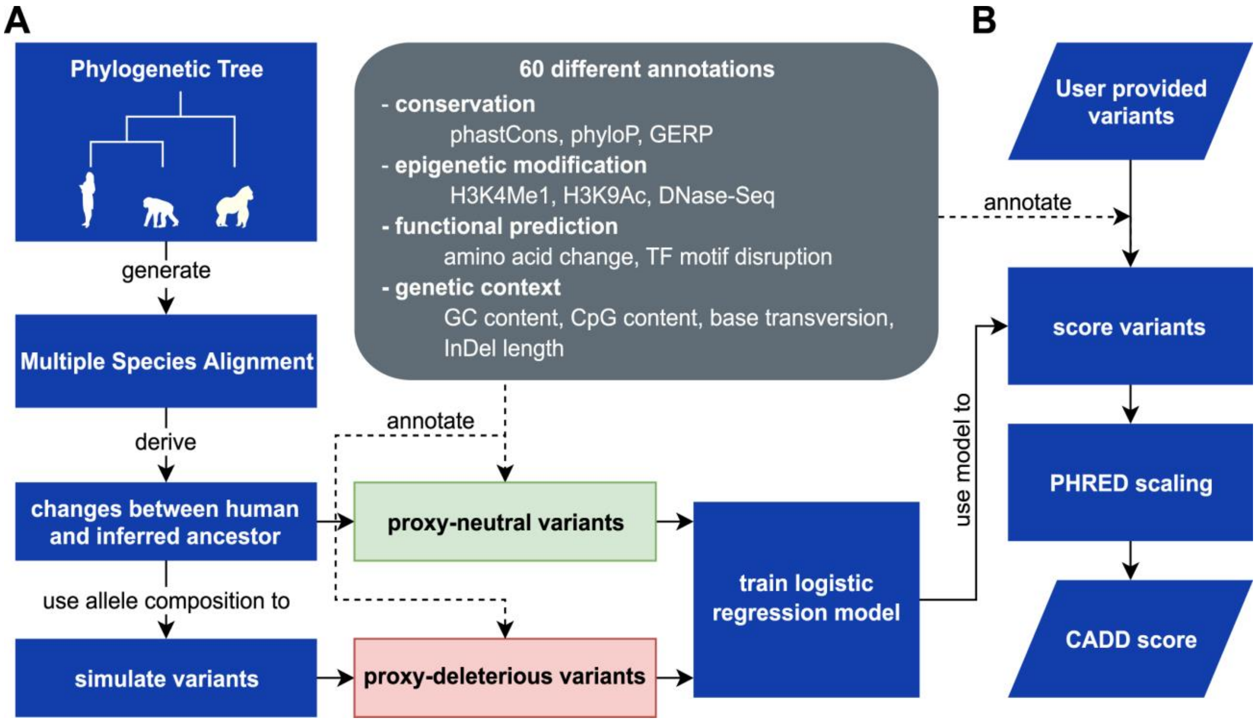


- Leverage the predictive power of ensembling models together to improve performance
- The outputs of other VEP models are used as inputs and typically trained similarly to supervised learning methods

REVEL (Rare Exome Variant Ensemble Learner) combines 13 other VEP tools as features (MutPred, FATHMM v2.3, VEST 3.0, PolyPhen-2, SIFT, PROVEAN, MutationAssessor, MutationTaster, LRT, GERP++, SiPhy, phyloP, and phastCons)

- is an RF ML method trained on rare neutral and disease variants
- The features of greatest significance in the developed RF were the FATHMM and VEST models.
- FATHMM employs a hidden Markov modeling method to analyze MSAs, which is unique among the other techniques

Combined Annotation-Dependent Depletion (CADD)

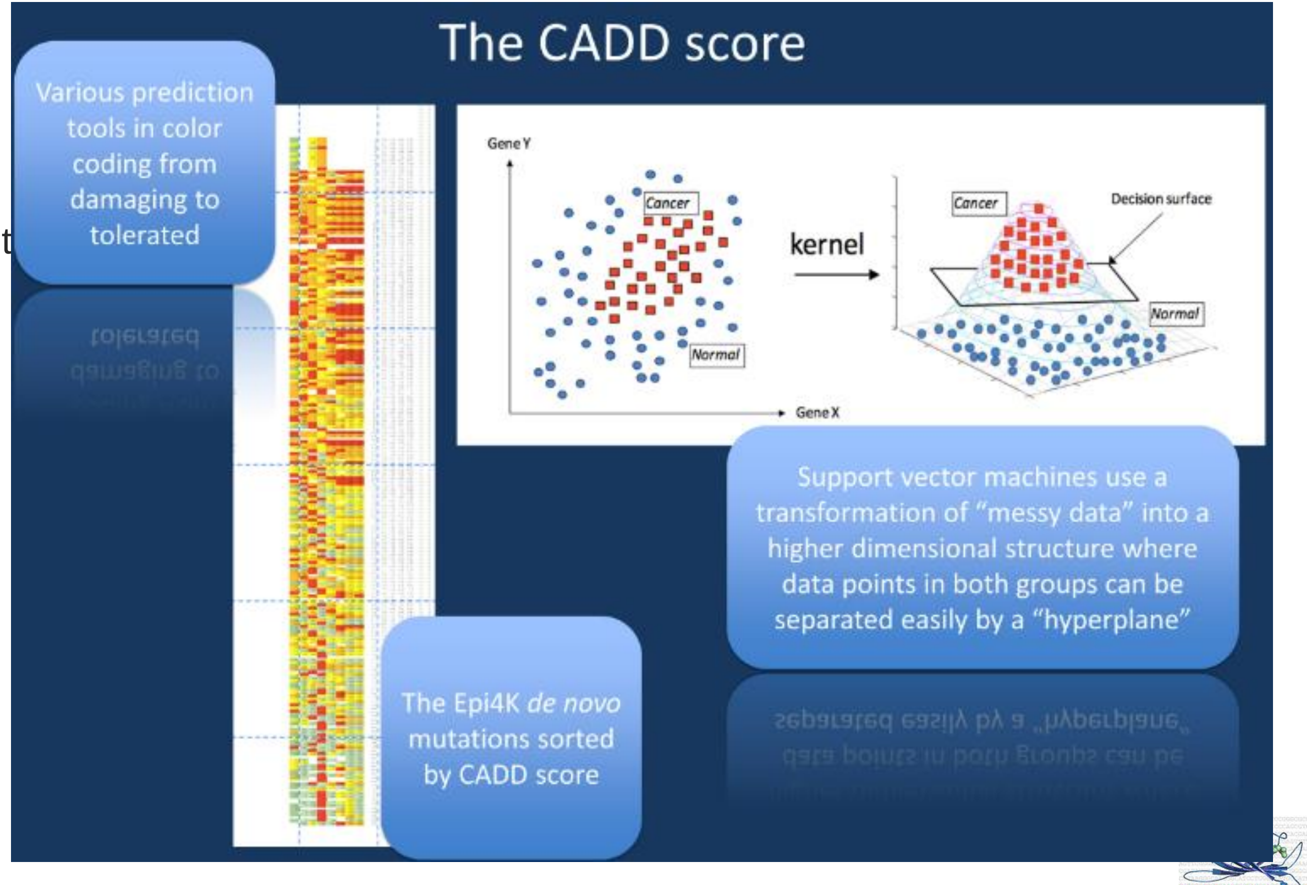

A
Phylogenetic Tree
generate
Multiple Species Alignment
derive
changes between human and inferred ancestor
use allele composition to
simulate variants
60 different annotations
- conservation
phastCons, phyloP, GERP
- epigenetic modification
H3K4Me1, H3K9Ac, DNase-Seq
- functional prediction
amino acid change, TF motif disruption
- genetic context
GC content, CpG content, base transversion, InDel length
annotate
proxy-neutral variants
proxy-deleterious variants
train logistic regression model
use model to
B
User provided variants
annotate
score variants
PHRED scaling
CADD score

Using more than 60 diverse annotations, a machine learning model is trained to classify variants as *proxy-neutral* versus *proxy-deleterious*

Combined Annotation-Dependent Depletion (CADD)

Score that ranks genetic variants, including single nucleotide variants (SNVs) and short inserts and deletions (InDels), throughout the human genome reference assembly

Reference genome SNVs at the 10th-% of CADD scores are assigned to CADD-10
top 1% to CADD-20,
top 0.1% to CADD-30



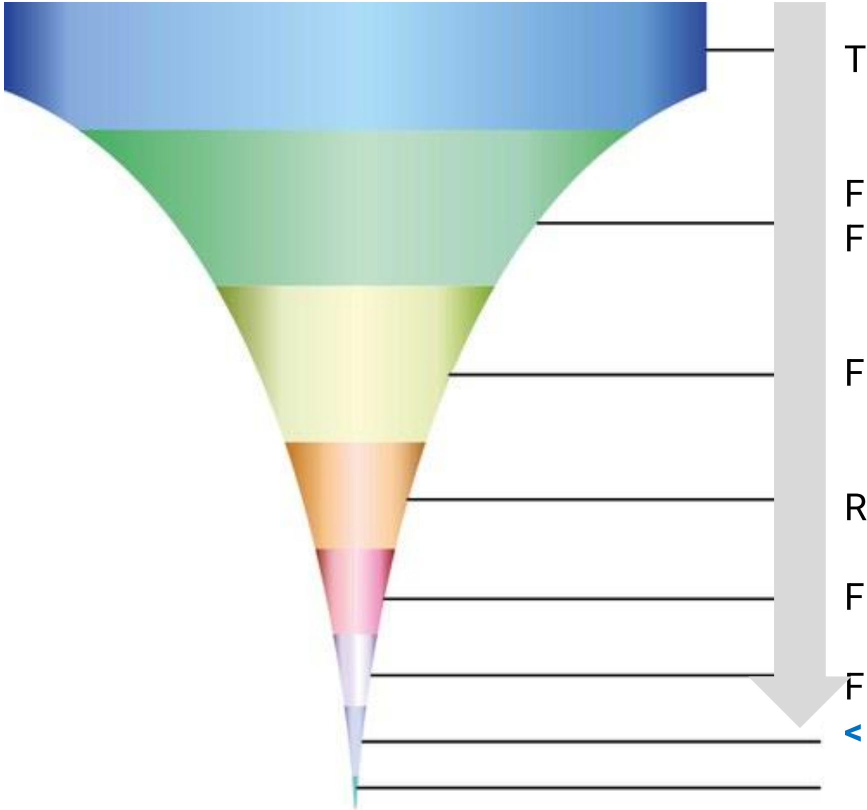
The CADD score

Various prediction tools in color coding from damaging to tolerated

Support vector machines use a transformation of “messy data” into a higher dimensional structure where data points in both groups can be separated easily by a “hyperplane”

The Epi4K *de novo* mutations sorted by CADD score

Variant filtering



Total Number of variants

Filtered by Internal Lab database
Filtered low quality variants

Filtered by gene list (Not for gene panel)

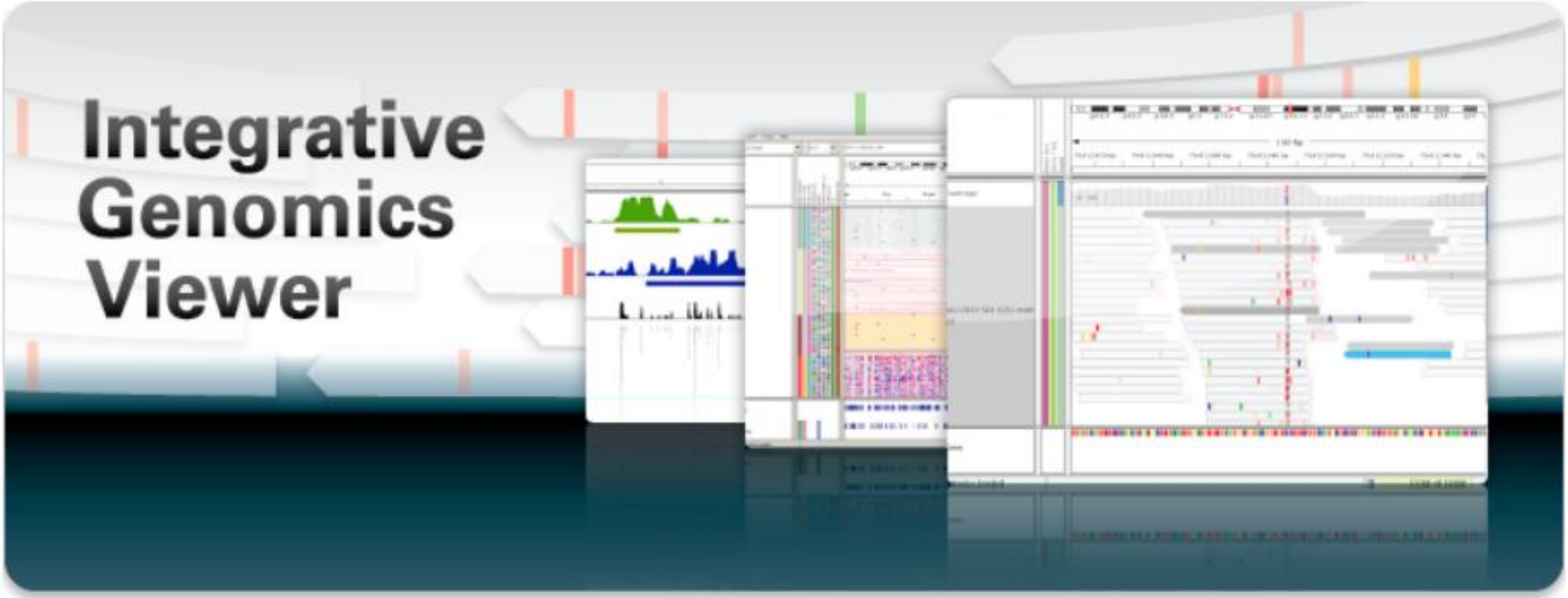
Removed intronic and synonimous variants

Filtered by population frequency

Filtered by pathogenicity score

< 10 variants to check

Variant Filtering: visual inspection



Variant Filtering: Integrative Genome Viewer (IGV)

Visual inspection can greatly increase the confidence in calls, reduce the risk of false positives, and help characterize complex events

A number of **IGV features** have been developed specifically to aid this manual review step:

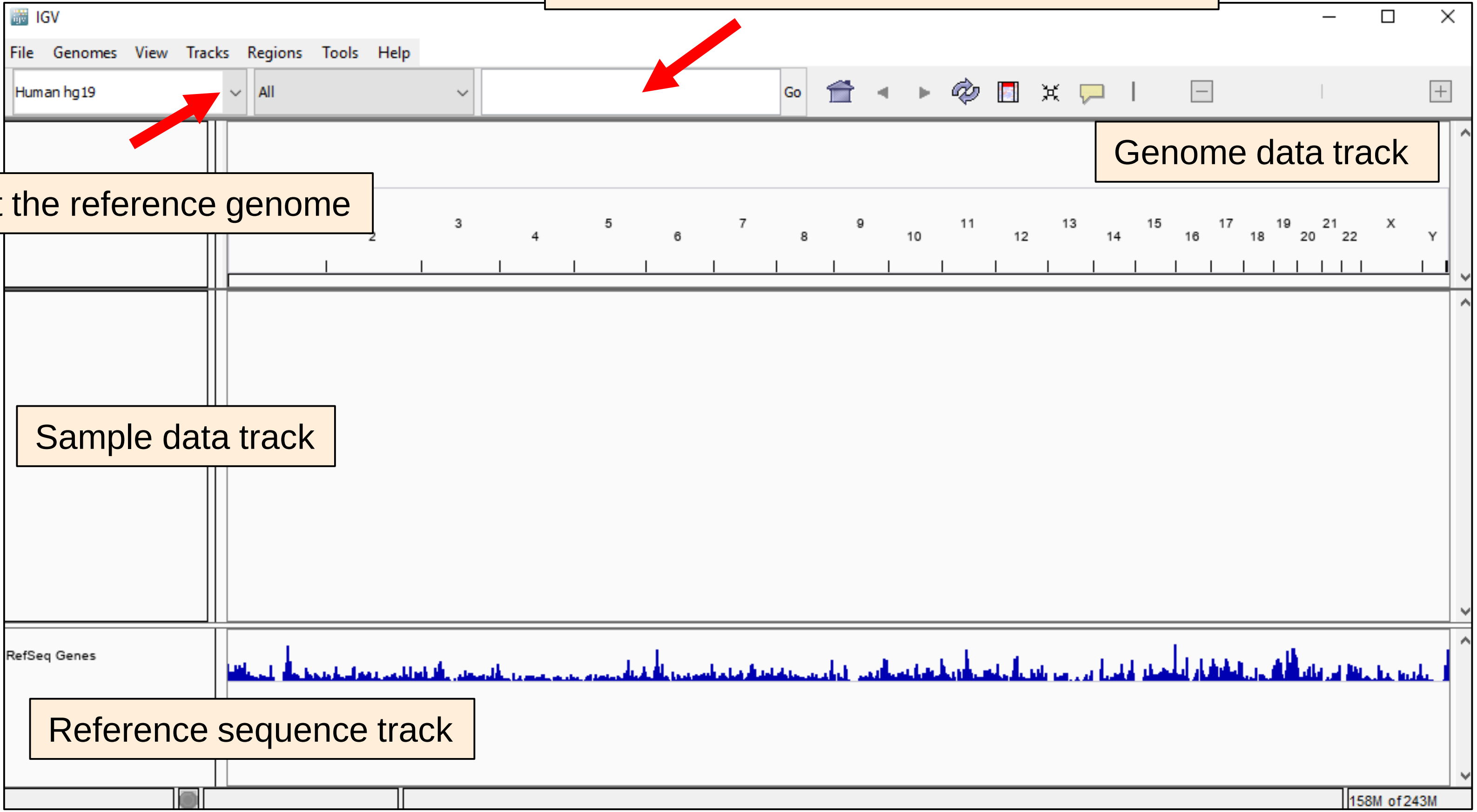
- highlighting mismatched bases in individual reads in color to aid detection of unusual patterns and **mismapped alignment**
- highlighting ambiguously mapped reads (**mapping quality**), indicative of high reference sequence homology, as such regions are known to produce many false positives
- shading of mismatched bases by **read base quality**, as clusters of bases of low quality can be indicative of sequencing errors
- **sorting, grouping, and coloring alignments** by alignment, sequencing, and platform metadata, which can be useful for detecting **systematic errors** upstream of read alignment

Variant Filtering: Integrative Genome Viewer (IGV)

- User Interface
- Download Reference Genome (Human hg19 for the gene panel sequencing)
- Viewing Sequencing Data with IGV
  - Alignment (BAM, SAM)
  - Variants (VCF)

https://youtu.be/E_G8z_2gTYM

IGV: user interface



Search by gene or chromosome position

Genome data track

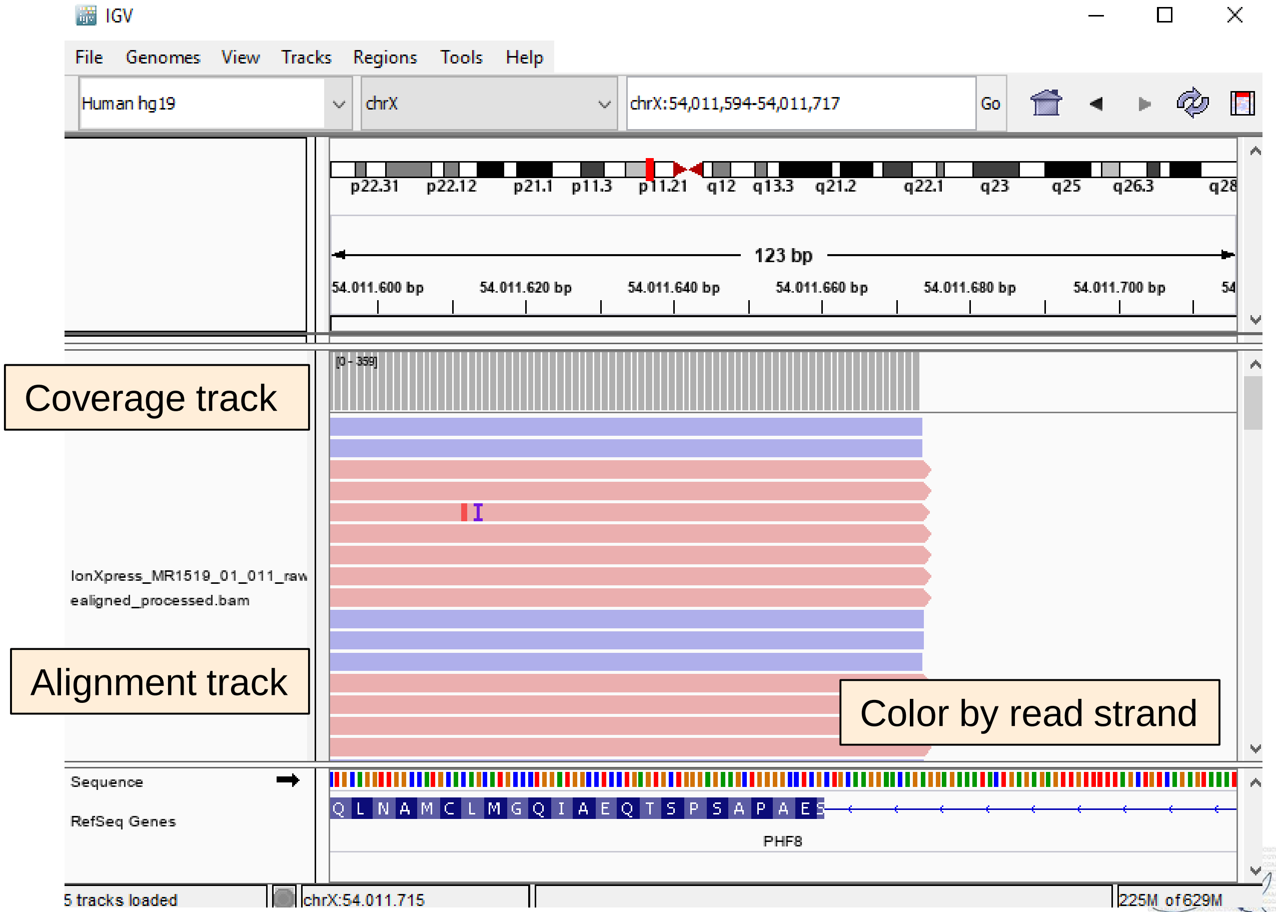
Select the reference genome

Sample data track

Reference sequence track

IGV: viewing alignment

- Upload a **BAM** or SAM file
- The main data file must include the *.bam*
- The index file should have the same filename but with the *.bai*
- When loading by file, IGV automatically searches for the index file within the same directory as the data file



Human hg19

chrX

chrX:54,011,594-54,011,717

IonXpress_MR1519_01_011_rawealigned_processed.bam

5 tracks loaded

chrX:54.011.715

225M of 629M

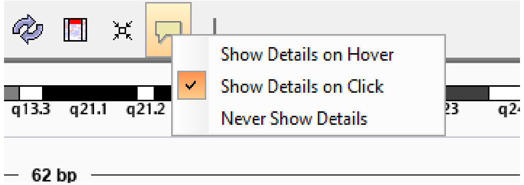
Coverage track

Alignment track

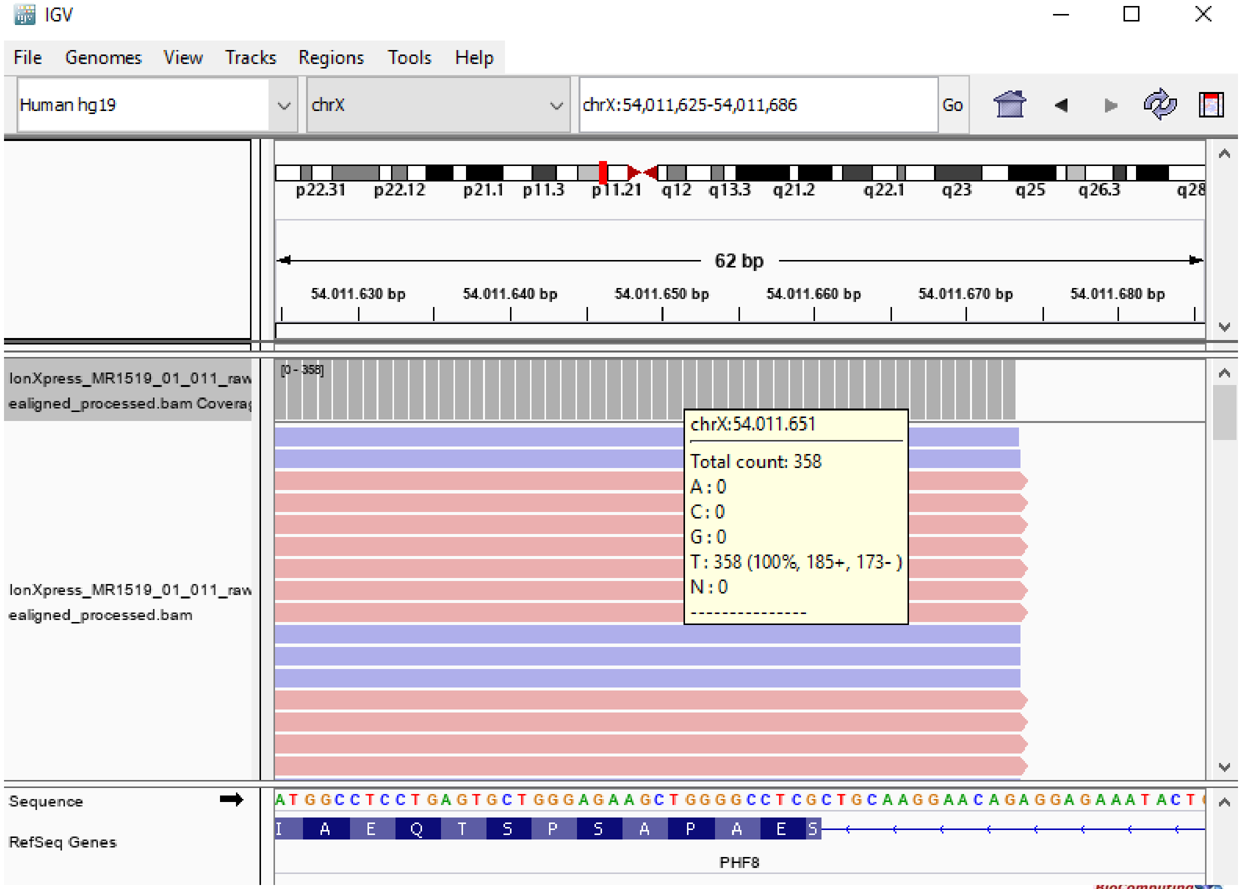
Color by read strand

IGV: viewing alignment

Coloring and sorting alignments
1. Use **Pop-up menù**



2. From menù
View/Preferences



Human hg19

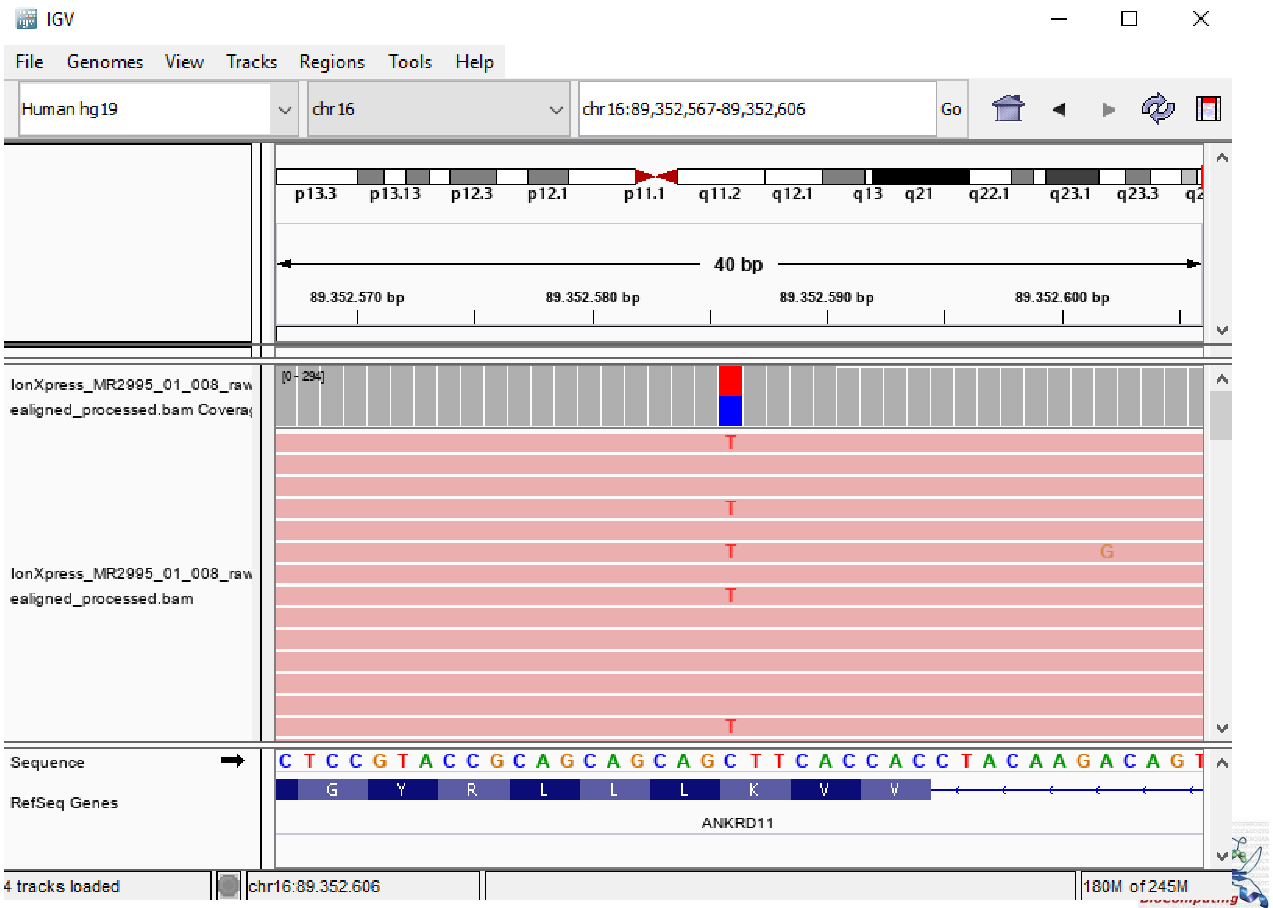
chrX

chrX:54,011,625-54,011,686

```
chrX:54.011.651
Total count: 358
A : 0
C : 0
G : 0
T : 358 (100%, 185+, 173- )
N : 0
---------------
```

IGV: viewing variant

- Upload a **VCF** file
- VCF data files must be indexed for viewing in IGV
- Search for a variant by chr position
- Use pop-up menù or change preferences to change color code



Human hg19

chr16

chr16:89,352,567-89,352,606

IonXpress_MR2995_01_008_rawealigned_processed.bam Coverag

IonXpress_MR2995_01_008_rawealigned_processed.bam

4 tracks loaded

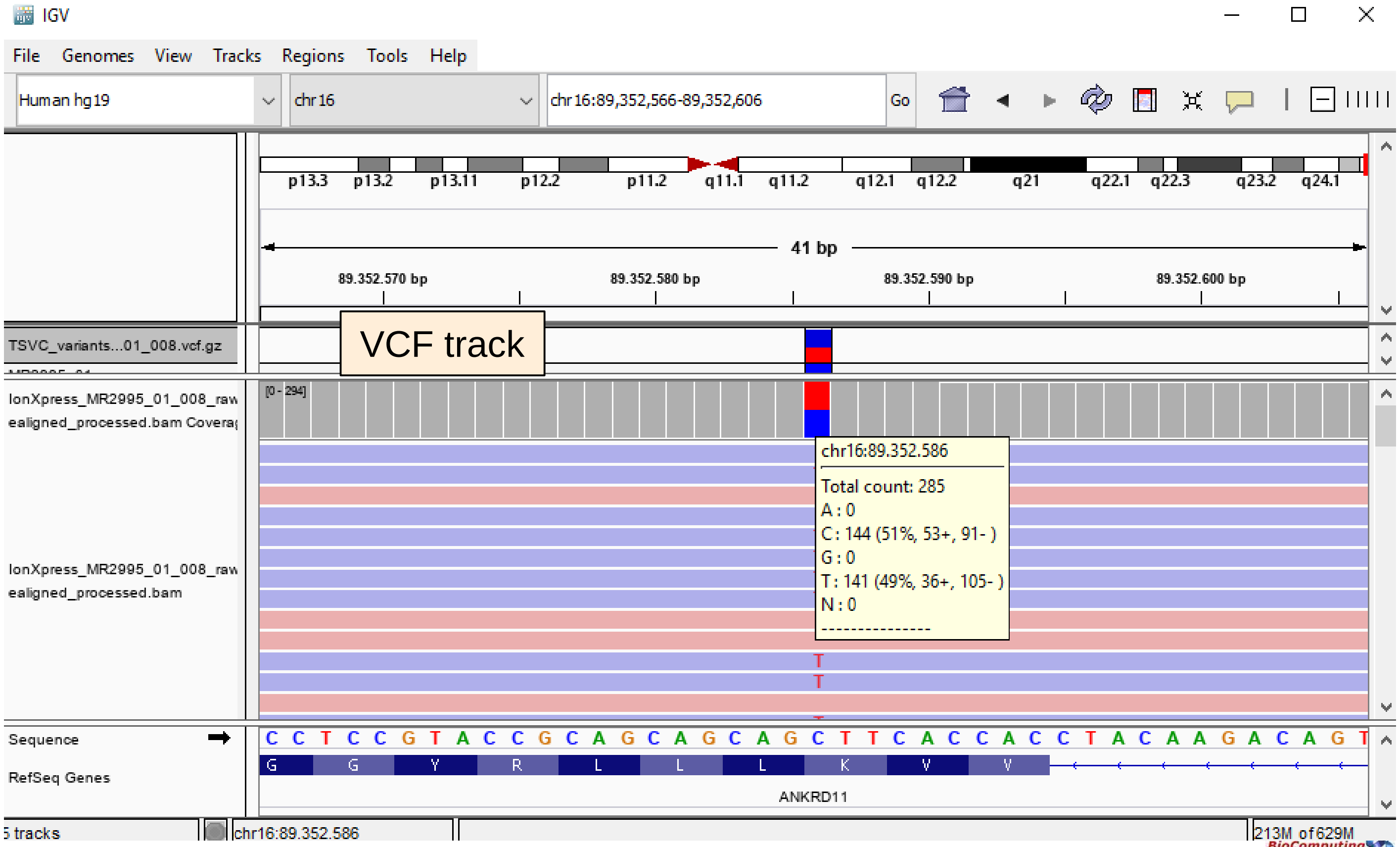
chr16:89.352.606

180M of 245M

IGV: viewing variant (nucleotide substitution)

chr16:89352586:C:T

ANKRD11



chr16:89.352.586

Total count: 285

A : 0

C : 144 (51%, 53+, 91-)

G : 0

T : 141 (49%, 36+, 105-)

N : 0

IGV: viewing variant (nt substitution)

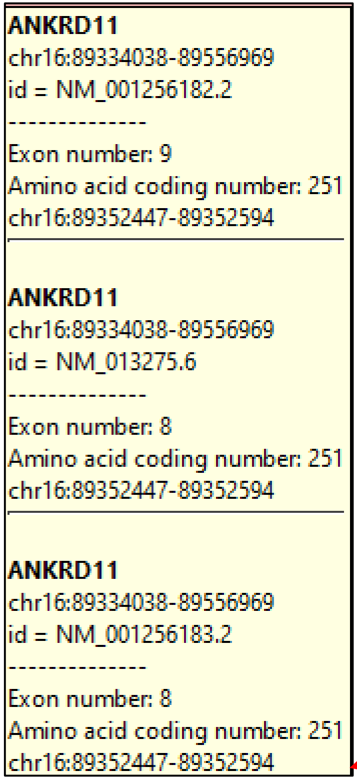
chr16:89352586:C:T

ANKRD11

NM_013275:exon8

c.753G>A p.K251K

Synonymous variant



ANKRD11
chr16:89334038-89556969
id = NM_001256182.2

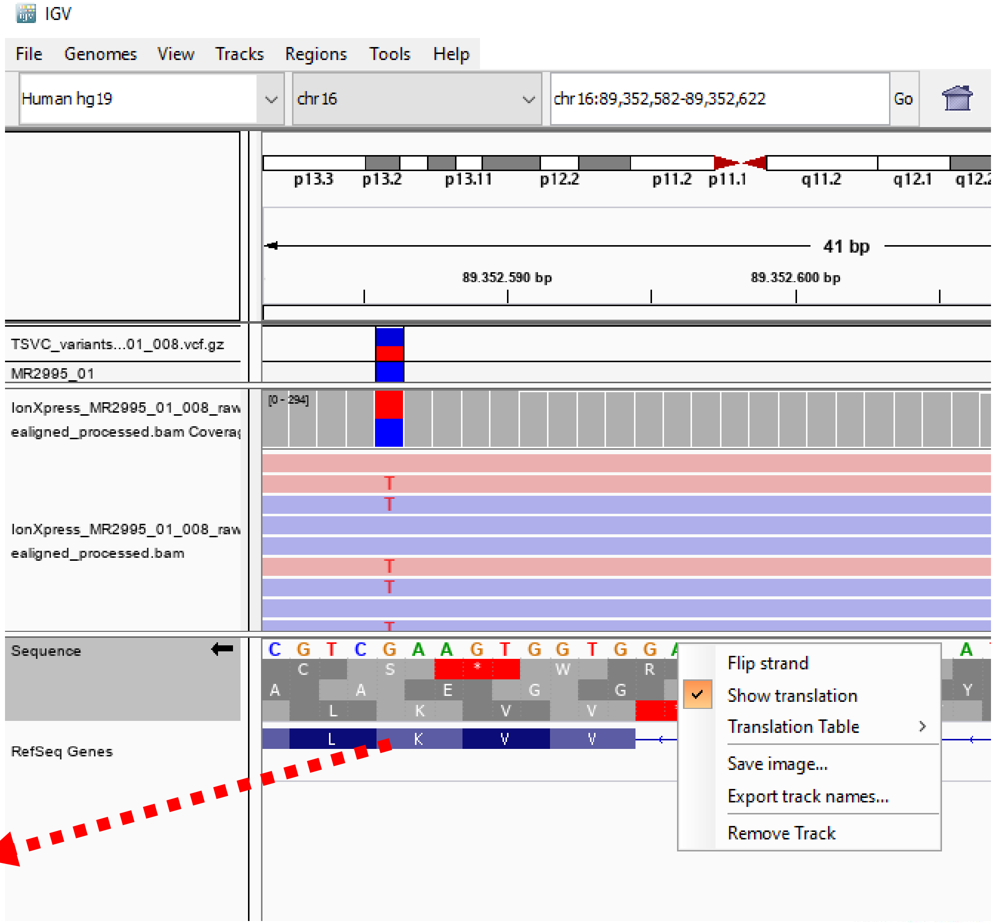
Exon number: 9
Amino acid coding number: 251
chr16:89352447-89352594

ANKRD11
chr16:89334038-89556969
id = NM_013275.6

Exon number: 8
Amino acid coding number: 251
chr16:89352447-89352594

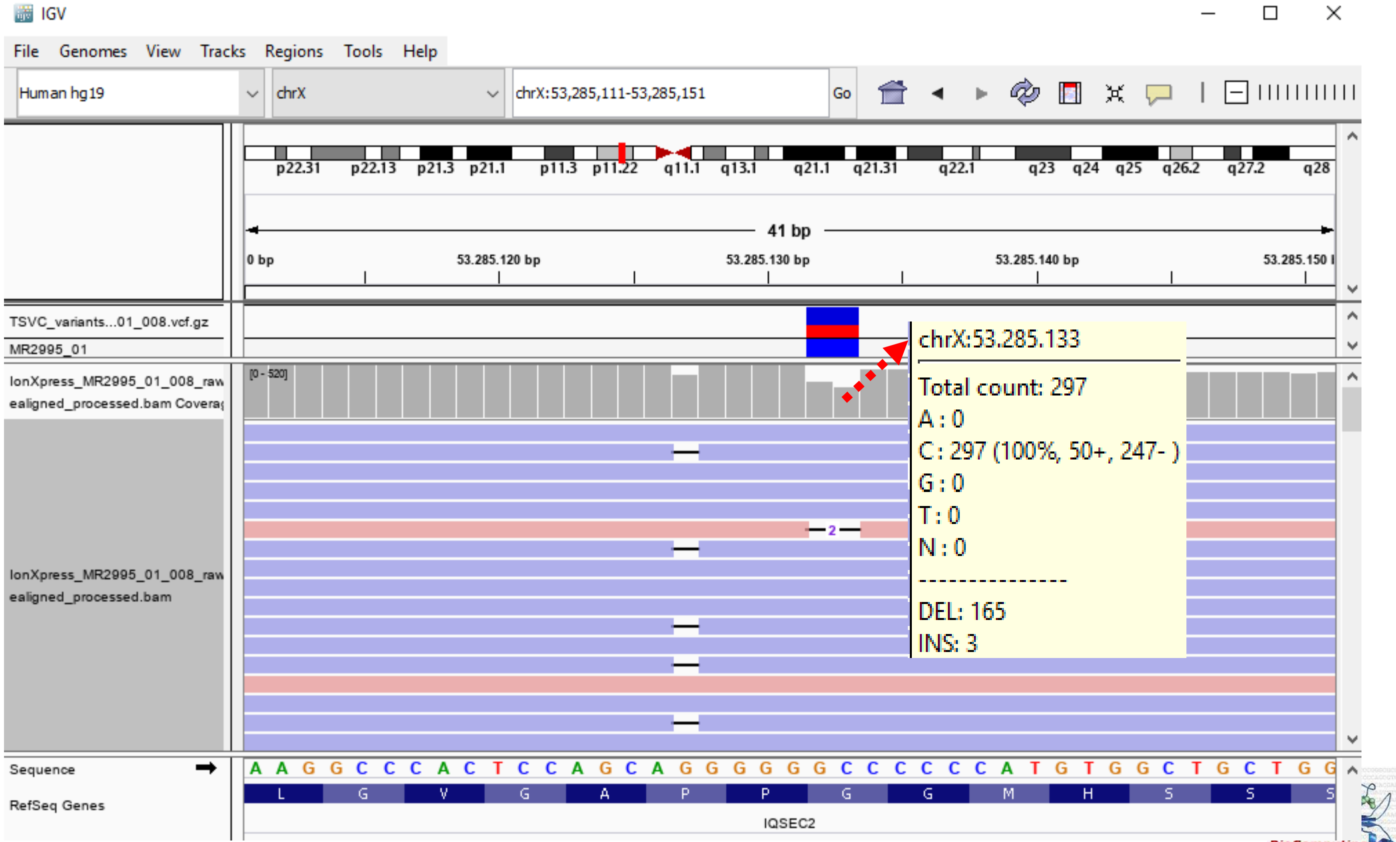
ANKRD11
chr16:89334038-89556969
id = NM_001256183.2

Exon number: 8
Amino acid coding number: 251
chr16:89352447-89352594



IGV: viewing variant (nt deletion)

chrX:53285126:AGGGGGGC:AGGGGG,AGGGGGC


IGV
File Genomes View Tracks Regions Tools Help
Human hg19
chrX
chrX:53,285,111-53,285,151
Go
p22.31 p22.13 p21.3 p21.1 p11.3 p11.22 q11.1 q13.1 q21.1 q21.31 q22.1 q23 q24 q25 q26.2 q27.2 q28
41 bp
0 bp
53.285.120 bp
53.285.130 bp
53.285.140 bp
53.285.150 bp
TSVC_variants...01_008.vcf.gz
MR2995_01
IonXpress_MR2995_01_008_rawealigned_processed.bam Coverag
[0 - 520]
IonXpress_MR2995_01_008_rawealigned_processed.bam
2
Sequence
A A G G C C C A C T C C A G C A G G G G G G C C C C C C A T G T G G C T G C T G G
RefSeq Genes
L G V G A P P G G M H S S S
IQSEC2

chrX:53.285.133

Total count: 297
A : 0
C : 297 (100%, 50+, 247-)
G : 0
T : 0
N : 0

DEL: 165
INS: 3

IGV: viewing variant (nt deletion)

chrX:53285126:AGGGGGGC:AGGGGG,AGGGGGC

GQ: 557, DP: 461, AF: 0.426304,0

IQSEC2:NM015075:exon3:c.239delC:p.P80fs

IQSEC2:NM015075:exon3:c.233-234del:p.G78fs

Annowar annotation from dbsnp: rs782460038

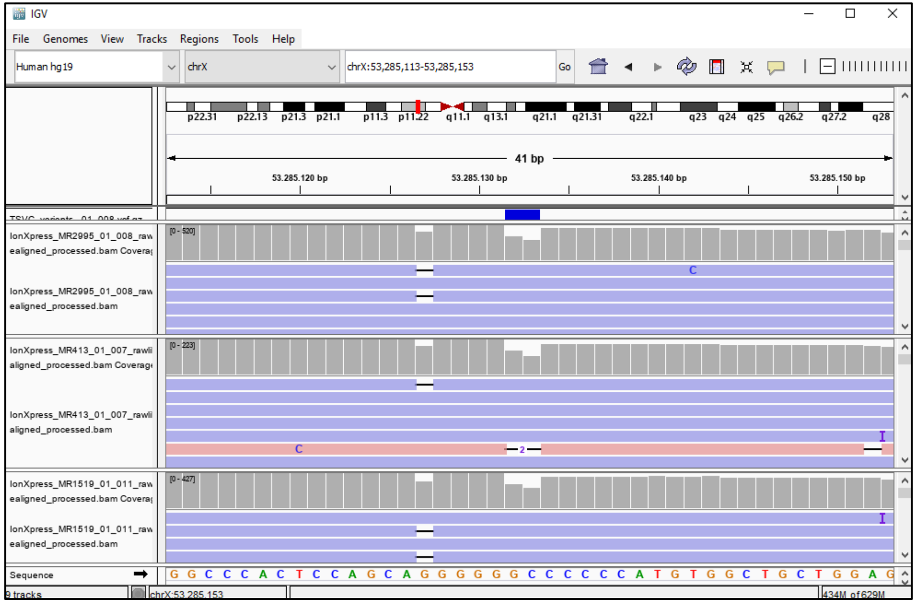
NM_015075.2:c.239del

Frequency!!!
delG=0.00000 (0/44894, ExAC)

Is in a polymeric region

Is present in with other VCFs

Probably error sequencing



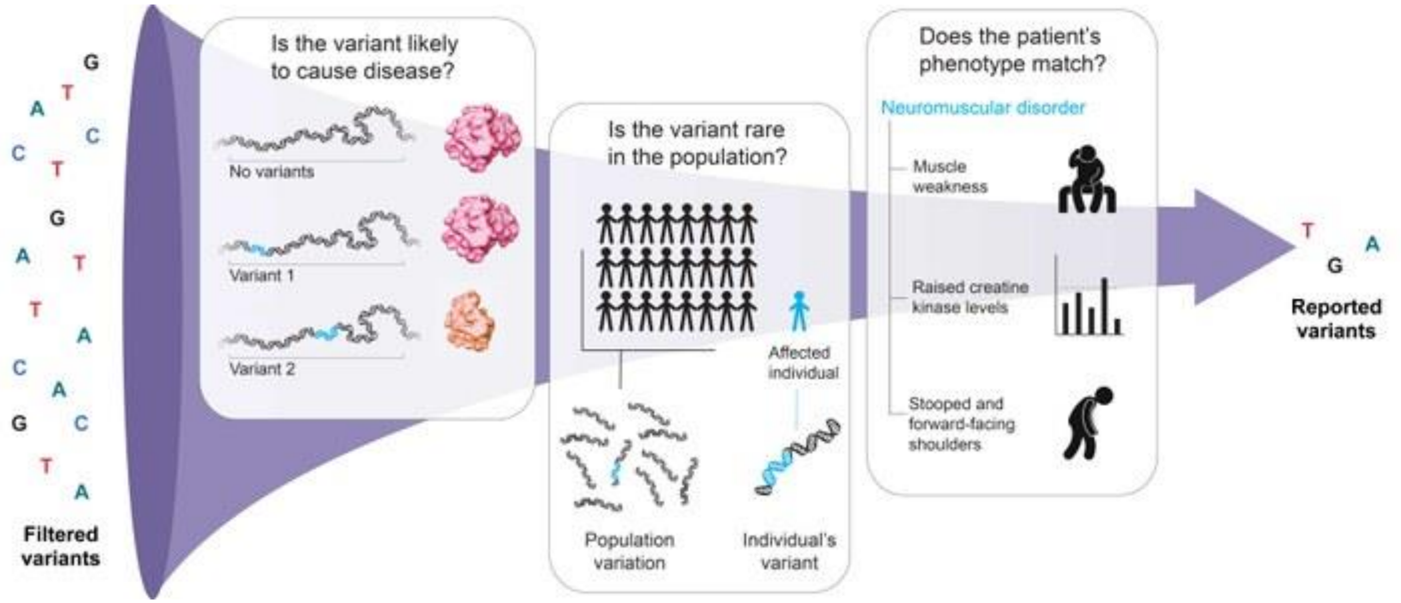
Human hg19

chrX

chrX:53,285,113-53,285,153

IGV
File
Genomes
View
Tracks
Regions
Tools
Help
Go
p22.31
p22.13
p21.3
p21.1
p11.3
p11.22
q11.1
q13.1
q21.1
q21.31
q22.1
q23
q24
q25
q26.2
q27.2
q28
41 bp
53.285.120 bp
53.285.130 bp
53.285.140 bp
53.285.150 bp
TSVC_variants_01_008.vcf.gz
IonXpress_MR2995_01_008_raw ealigned_processed.bam Coverag
[0 - 520]
C
IonXpress_MR2995_01_008_raw ealigned_processed.bam
IonXpress_MR413_01_007_rawli aligned_processed.bam Coverage
[0 - 223]
IonXpress_MR413_01_007_rawli aligned_processed.bam
C
2
I
IonXpress_MR1519_01_011_raw ealigned_processed.bam Coverag
[0 - 427]
IonXpress_MR1519_01_011_raw ealigned_processed.bam
I
Sequence
G G C C C A C T C C A G C A G G G G G G C C C C C C A T G T G G C T G C T G G A G
9 tracks
chrX:53,285,153
434M of 629M

Variant interpretation: criteria


G
T
A
C
C
T
G
A
T
T
A
C
A
G
C
T
A
Filtered variants
Is the variant likely to cause disease?
No variants
Variant 1
Variant 2
Is the variant rare in the population?
Affected individual
Population variation
Individual's variant
Does the patient's phenotype match?
Neuromuscular disorder
Muscle weakness
Raised creatine kinase levels
Stooped and forward-facing shoulders
T
A
G
Reported variants