NO CONFLICTS TO DISCLOSE

BERLIN
INSTITUTE
OF HEALTH
Charité & Max Delbrück Center

ESHG 2017 – W05

Defining “mutation” or “polymorphism” using prediction tools

Organizer: Malte Spielmann (University of Washington)

Presenters: Martin Kircher (BIH) & Dominik Seelow (Charité)

Copenhagen, 2017-05-28

BERLIN
INSTITUTE
OF HEALTH
Charité & Max Delbrück Center

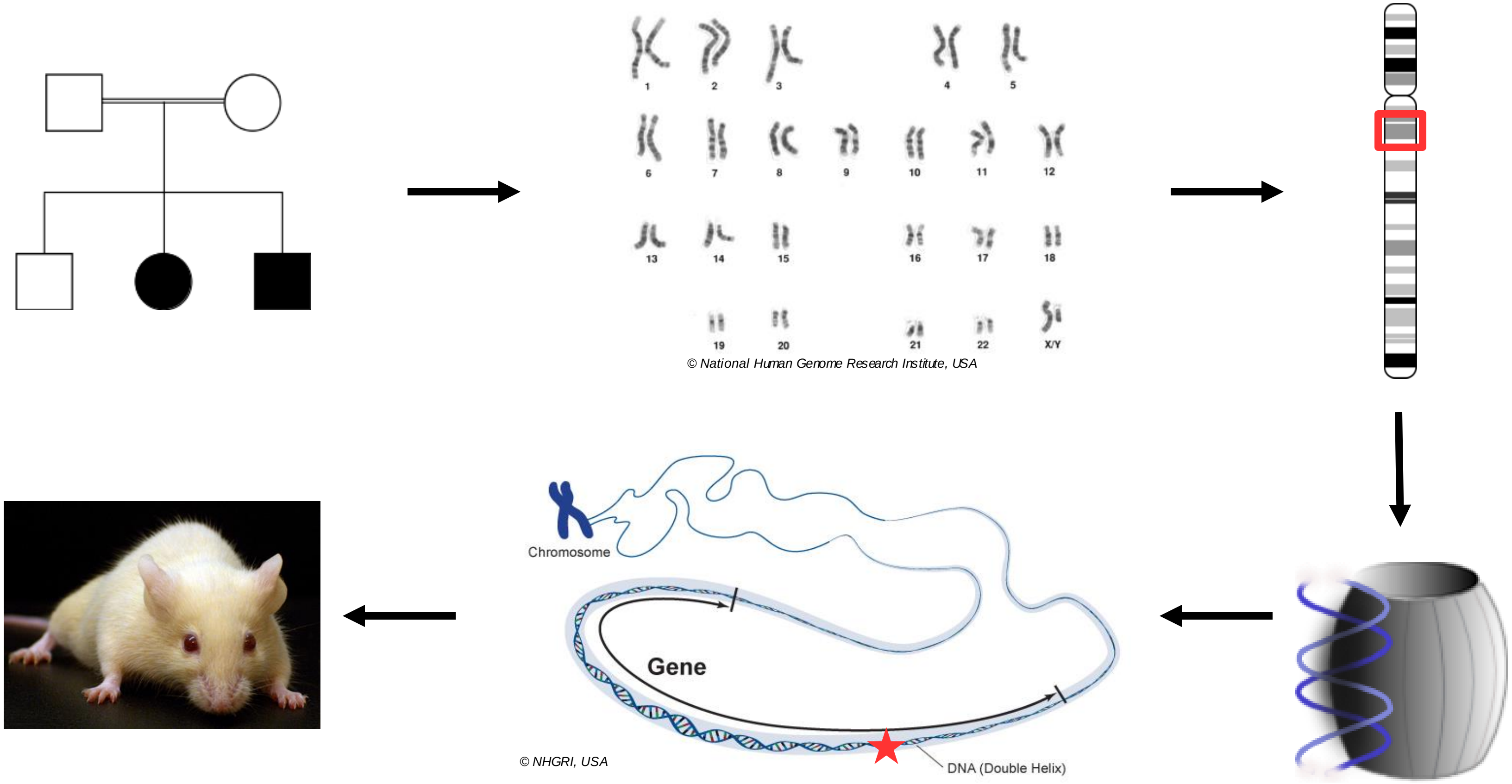
CONTENT

1. Welcome and opening remarks
2. Variants and polymorphisms in the context of disease
3. Annotation of variants
4. Considerations of variant filtering

(short break)

5. Assessment of variants
6. Challenges of interpreting non-coding variants
7. Questions and participant feedback

Discovery of disease mutations (past)



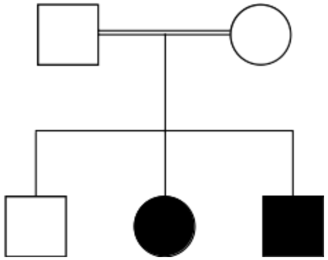
© National Human Genome Research Institute, USA

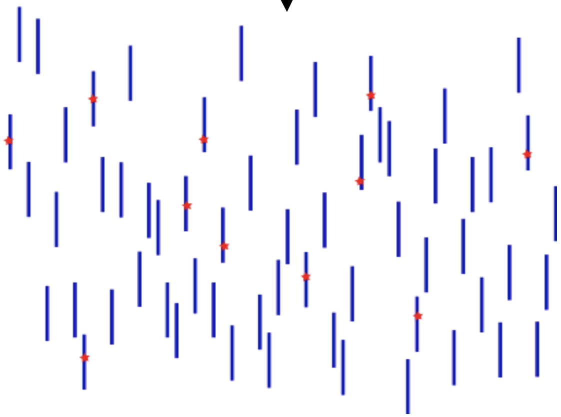
© NHGRI, USA

sequencing single candidate genes ► single variants ► confirmation

Discovery of disease mutations (present)

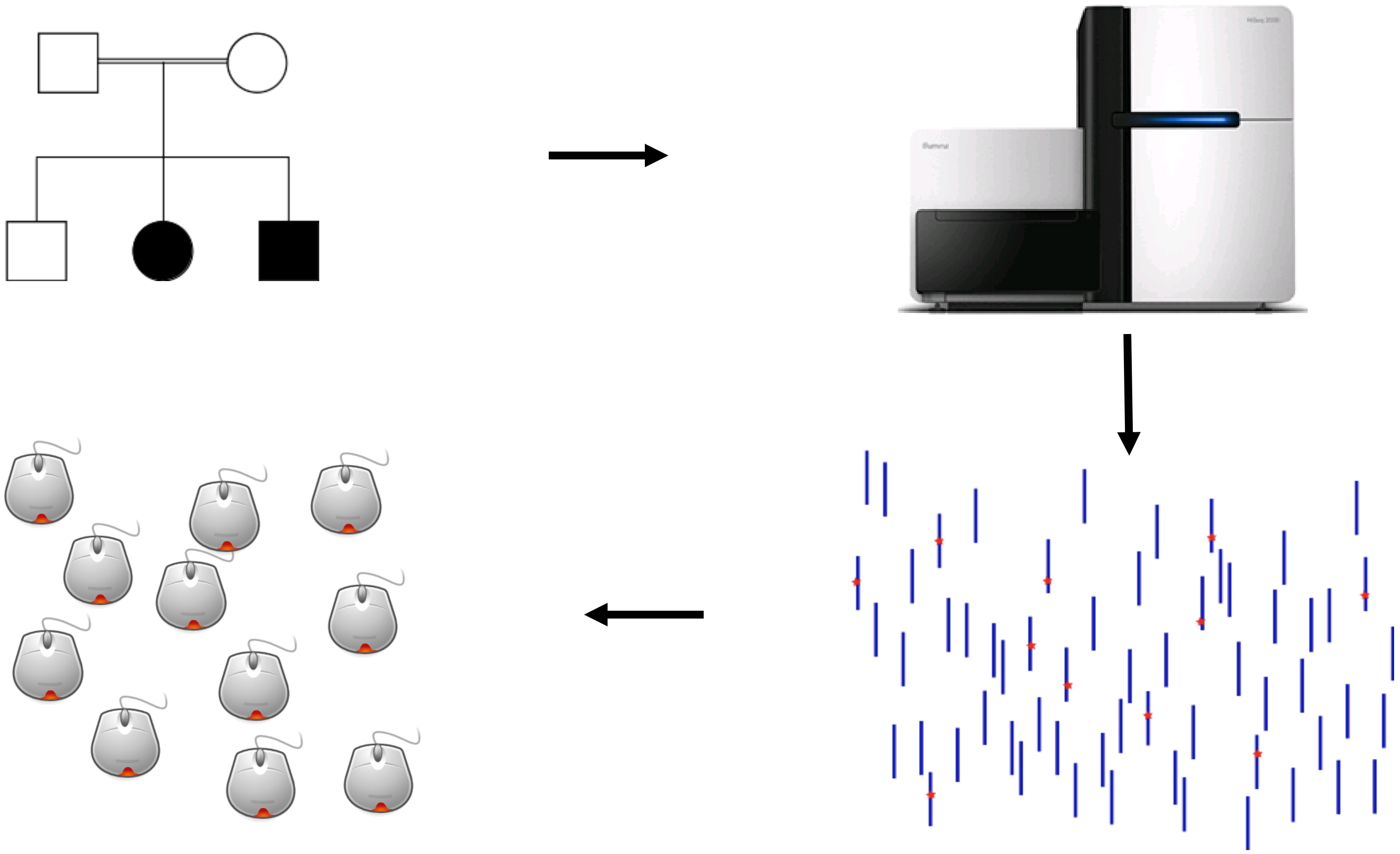
BERLIN INSTITUTE OF HEALTH
Charité & Max Delbrück Center





sequencing all genes ▶ 10,000+ variants ▶ ?

Discovery of disease mutations (present)



sequencing all genes ▶ 10,000+ variants ▶ bioinformatics

VARIANT AND POLYMORPHISMS IN THE CONTEXT OF DISEASE

1. Genetic models of disease
2. Reference-based analysis & Variant Call Format (VCF)
3. Which variants to trust?
4. Visualizing alignment files with IGV

Variant, mutation and polymorphism

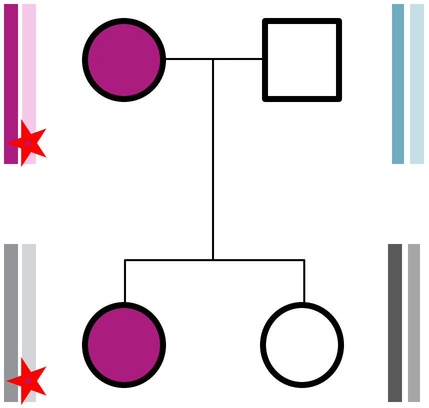
- **Variant:**
  Sequence difference identified in a comparison to a reference.
  Can be used with modifiers: e.g. pathogenic, likely pathogenic, uncertain significance, likely benign, or benign
- **Mutation:**
  Variant identified in a paired sequencing effort (e.g. cancer vs. normal, somatic vs. germline, parents vs. offspring)
  Earlier: rare sequence change; potentially damaging
- **Polymorphism:**
  Variant identified across multiple unrelated individuals
  Earlier: DNA variant occurring with 1% or higher frequency in a population; considered neutral

doi: 10.1186/s12920-015-0115-z

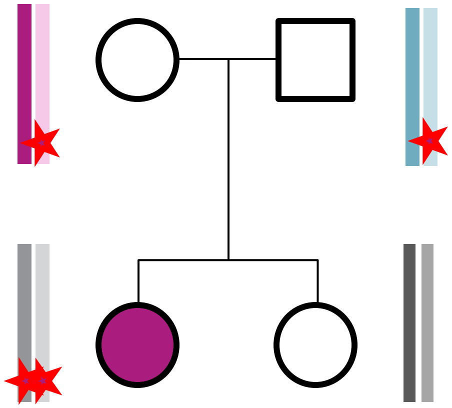
Genetic models of disease (1)

- Dominant / recessive
- Homozygous / heterozygous / compound

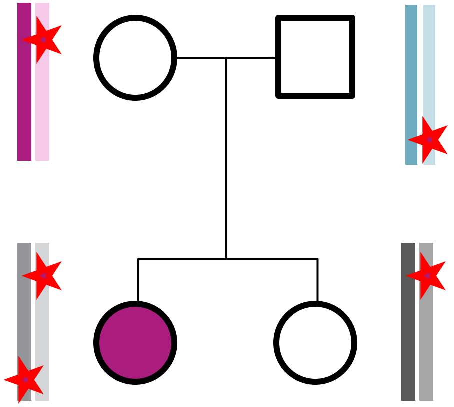
Dominant



Homozygous recessive



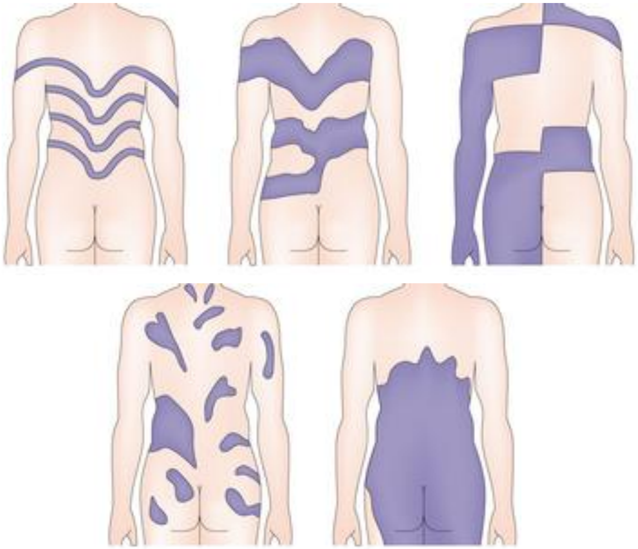
Compound* heterozygote recessive

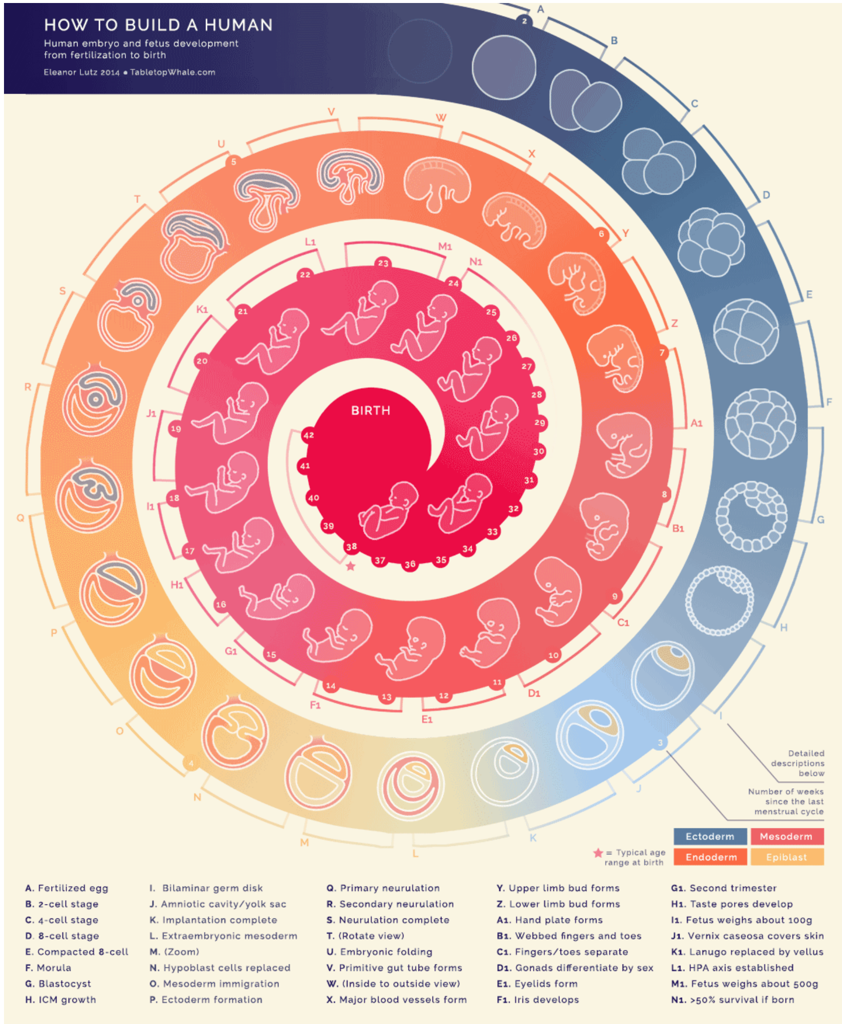


* require phase information

Genetic models of disease (2)

- Inherited vs. *De Novo*
- Somatic vs. Germline
- Mosaicism / Cancer




HOW TO BUILD A HUMAN
Human embryo and fetus development from fertilization to birth
Eleanor Lutz 2014 • TabletopWhale.com
BIRTH
Detailed descriptions below
Number of weeks since the last menstrual cycle
Ectoderm
Mesoderm
Endoderm
Epiblast
★ = Typical age range at birth

A. Fertilized egg
B. 2-cell stage
C. 4-cell stage
D. 8-cell stage
E. Compacted 8-cell
F. Morula
G. Blastocyst
H. ICM growth

I. Bilaminar germ disk
J. Amniotic cavity/yolk sac
K. Implantation complete
L. Extraembryonic mesoderm
M. (Zoom)
N. Hypoblast cells replaced
O. Mesoderm immigration
P. Ectoderm formation

Q. Primary neurulation
R. Secondary neurulation
S. Neurulation complete
T. (Rotate view)
U. Embryonic folding
V. Primitive gut tube forms
W. (Inside to outside view)
X. Major blood vessels form

Y. Upper limb bud forms
Z. Lower limb bud forms
A1. Hand plate forms
B1. Webbed fingers and toes
C1. Fingers/toes separate
D1. Gonads differentiate by sex
E1. Eyelids form
F1. Iris develops

G1. Second trimester
H1. Taste pores develop
I1. Fetus weighs about 100g
J1. Vernix caseosa covers skin
K1. Lanugo replaced by vellus
L1. HPA axis established
M1. Fetus weighs about 500g
N1. >50% survival if born

Left: doi: 10.1038/nrg3424
Right: http://tabletopwhale.com/2014/12/16/how-to-build-a-human.html

Whole genome vs. exome sequencing

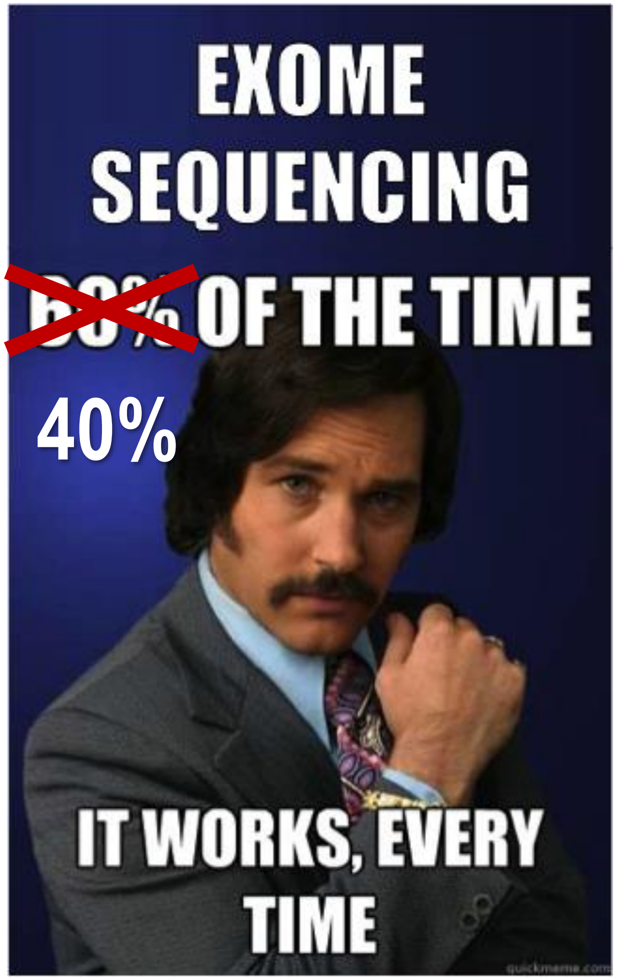
Exome sequencing (~ €350):

- ~25,000 - 50,000 variants
  mostly within annotated genes
- 1,000 - 2,000 'rare' variants

Whole genome (~ €1000):

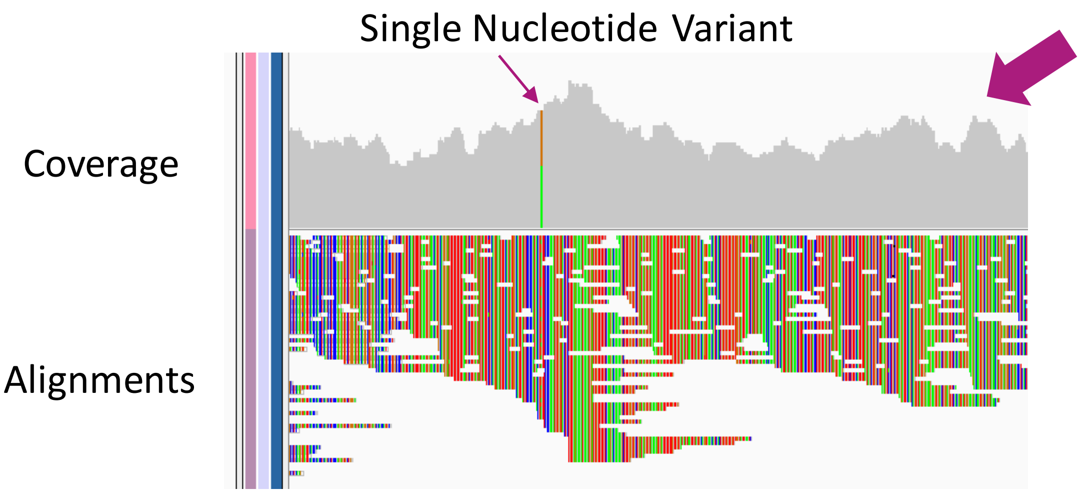
- 1 - 3 million variants
  mostly outside of annotated genes
- 150,000 - 500,000 'rare' variants

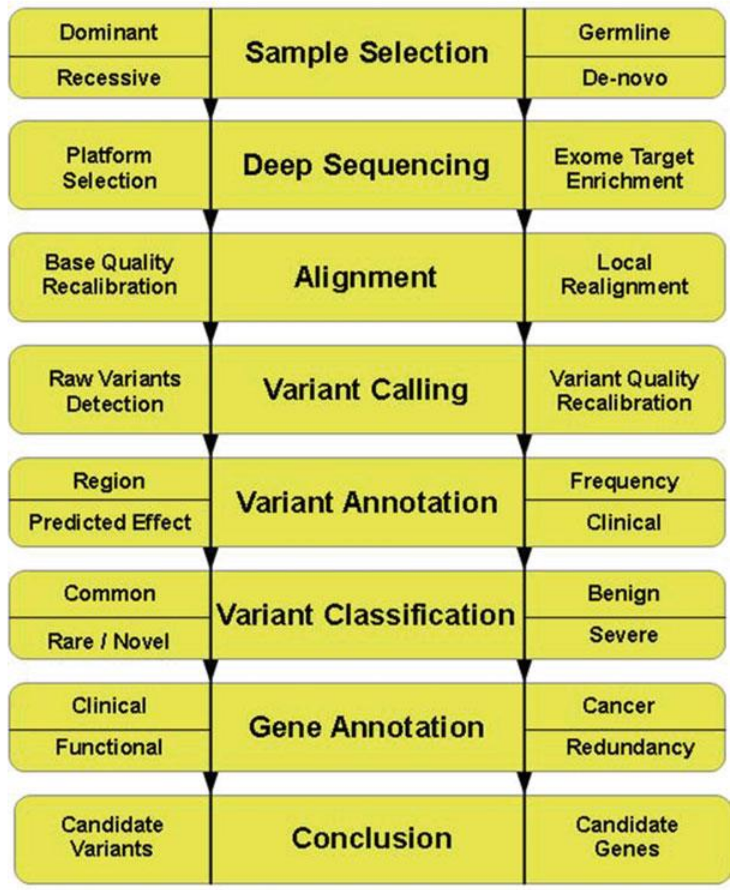
➢ **Prices without variant interpretation!**



Reference-based variant analysis

Due to complexity of assembling and annotating genomes, best practice workflows involve alignments to a reference genome




Dominant | Sample Selection | Germline
Recessive | | De-novo

Platform Selection | Deep Sequencing | Exome Target Enrichment

Base Quality Recalibration | Alignment | Local Realignment

Raw Variants Detection | Variant Calling | Variant Quality Recalibration

Region | Variant Annotation | Frequency
Predicted Effect | | Clinical

Common | Variant Classification | Benign
Rare / Novel | | Severe

Clinical | Gene Annotation | Cancer
Functional | | Redundancy

Candidate Variants | Conclusion | Candidate Genes

DOI 10.1007/978-1-62703-514-9_8

Variant Call Format (VCF)

BERLIN INSTITUTE OF HEALTH
Charité & Max Delbrück Center

- Tab-separated text format for storing variant information (typically SNPs, indels; but also structural variants)
- Development and specification driven by 1000 Genomes project
- Generated by many variant caller / genotyper packages
- Input for most downstream tools (e.g. Gemini, SeattleSeq, Variant Effect Predictor, SNPeff, AnnoVar, CADD)
- Official format specification: http://samtools.github.io/hts-specs/VCFv4.2.pdf

Variant Call Format: header

```
##fileformat=VCFv4.0
##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=.,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
20      14370 rs6054257  G    A   29     PASS     NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:D
20      17330 .          T    A   3      q10      NS=3;DP=11;AF=0.017      GT:GQ:D
```

Variant Call Format: variant lines

```
#CHROM  POS    ID         REF  ALT  QUAL  FILTER  INFO                    FORMAT
20      14370  rs6054257  G    A    29    PASS    NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:D
20      17330  .          T    A    3     q10     NS=3;DP=11;AF=0.017     GT:GQ:D
```

CHROM	chromosome
POS	position (1st base having position 1, positions are sorted numerically, in increasing order)
ID	semi-colon separated list of unique identifiers or '.'
REF	reference base(s): A,C,G,T,N
ALT	comma separated list of alternate non-reference alleles
QUAL	phred-scaled quality score
FILTER	filter that position passes

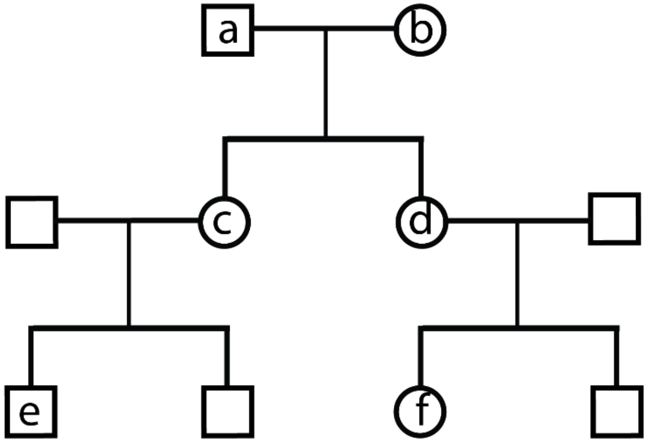
```
TER   INFO                    FORMAT       NA00001
S     NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ  0|0:48:1:51,51
      NS=3;DP=11;AF=0.017     GT:GQ:DP:HQ  0|1:3:5:65,3
```

INFO	additional information as a semicolon-separated series of short keys with optional values in the format: <key>=<data>[,data]
FORMAT	data to be provided for each of the samples
ACTUAL_SAMPLES	information in the order of FORMAT

Always quality control samples first

- Sequencing: quality scores
- Sample quality: molecule length, DNA damage, PCR replicates
- Sample purity: environmental / sample contamination
- Completeness of coverage / fraction bases uncovered
- Sex check: Alignments in Y unique regions? X chromosome heterozygosity?
- Relatedness/kinship estimates?
- Agreement with inheritance model?

Relatedness:
a-b ~ 0
a-c/d ~ 1/2
a-e/f ~ 1/4, e-d ~ 1/4
e-f ~ 1/8



Measures of relatedness

- Tools like bcftools, PLINK, KING allow to test sex and relatedness

Relationship	R	Kinship
identical twins	1.0000	0.5000
parent-offspring	0.5000	0.2500
full siblings	0.5000	0.2500
grandparent-grandchild	0.2500	0.1250
half siblings	0.2500	0.1250
aunt/uncle-nephew/niece	0.2500	0.1250
double first cousins	0.2500	0.1250
great grandparent-great grandchild	0.1250	0.0625
first cousins	0.1250	0.0625
second cousins	0.0313	0.0157
third cousins	0.0078	0.0039
fourth cousins	0.0020	0.0010

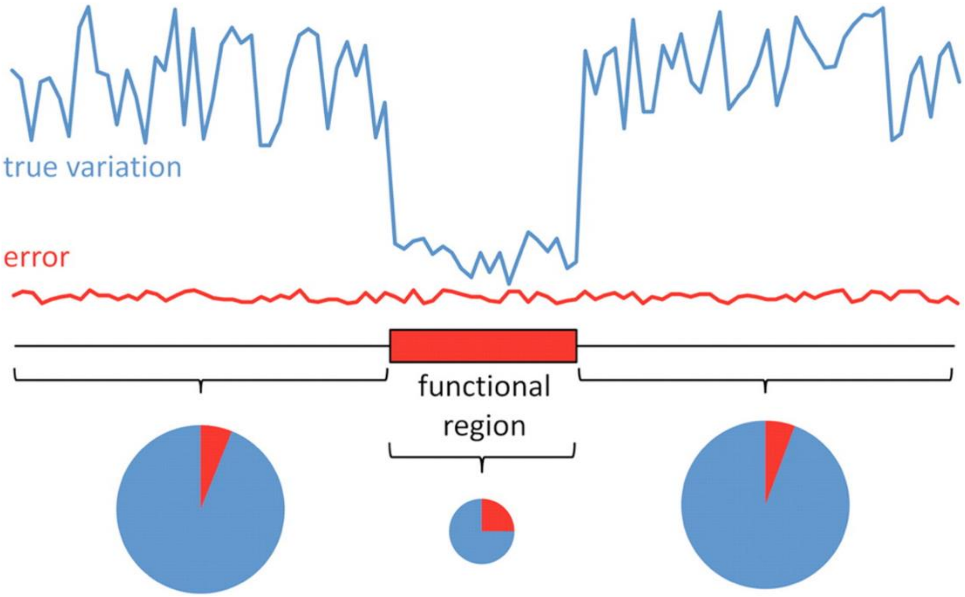
Which variants to trust?

BERLIN INSTITUTE OF HEALTH
Charité & Max Delbrück Center

- **Targeted vs shotgun sequencing**
  - Targeted sequencing with larger variation in coverage
  - Check targeted regions are covered at a minimum depth
  - Candidate variants: always check genotype quality, allele balance, strand balance, sequencing depth
- **Systematic errors, long variants and structural variants**
  - Collect/ask for list of commonly observed variants
  - Note that intermediate-sized (~30-100bp) InDels are the most difficult to call from short-read technologies
  - Check for overlap with known structural variants/segmental duplications

Sanity checks for individual variants

- Variants with high impact functional annotation are enriched for false positives
- Check overlap with segmental duplications or repetitive elements
- Study frequency vs. database frequency
  - Common allele in study absent from public database, or rare variant in study at high-frequency in database
  - Hardy-Weinberg equilibrium for homozygote and heterozygote carriers ($p^2 + 2pq + q^2 = 1$)



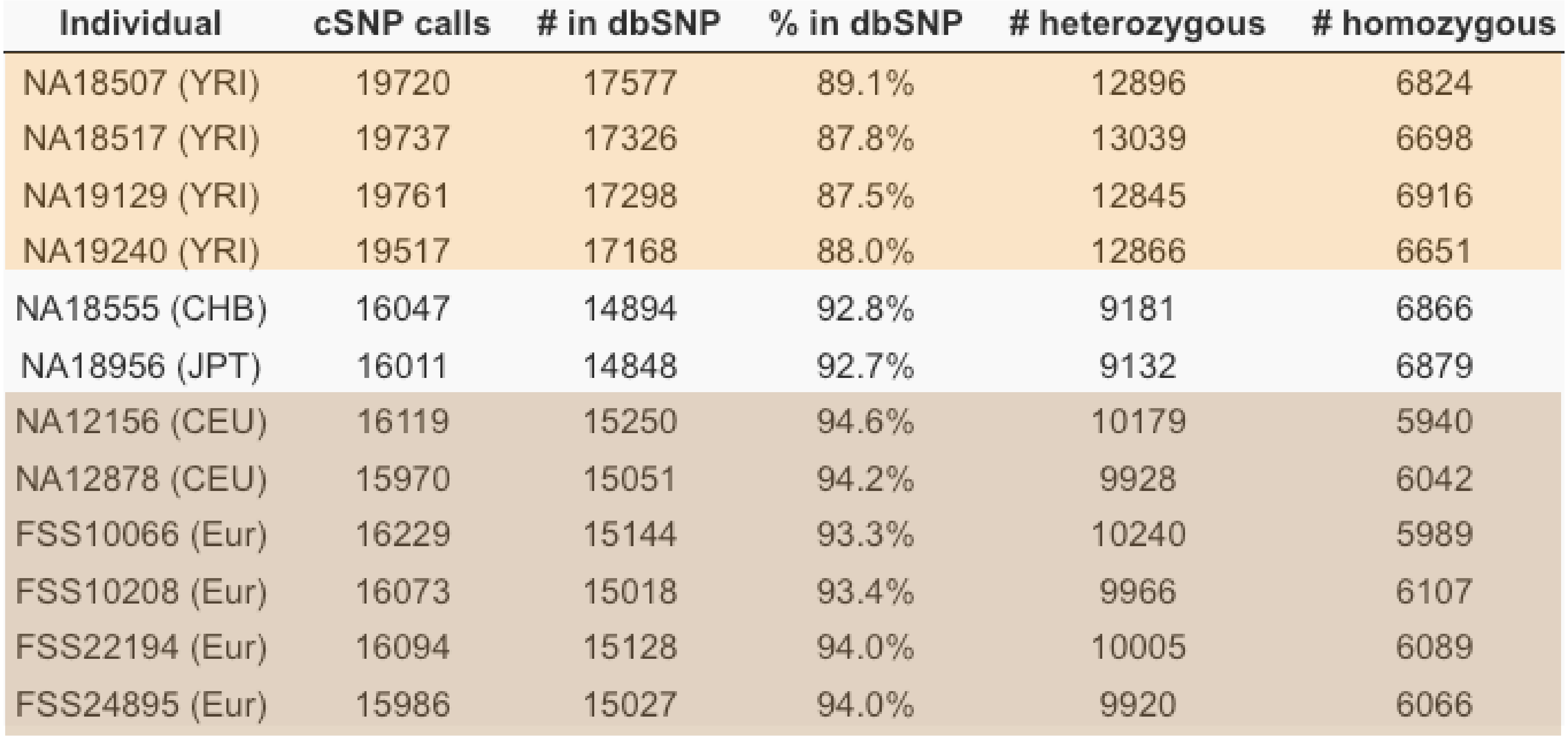
MacArthur and Tyler-Smith 2010 Hum Mol Gen

Overall quality: Known allele frequency spectrum

BERLIN INSTITUTE OF HEALTH
Charité & Max Delbrück Center

- Vast majority of alleles in any one sample should be common and present in databases
- Most variants in a large sample of people are rare
- Rare/novel variants are overwhelmingly heterozygous
- Number of stop codons, typically ~100 per genome (most are common variants)
- Transition-to-transversion ratio for mammals:
  - Transitions about 2x more frequent than transversions
  - Within coding exons, the ratio is closer to 3:1, as transitions are less likely to change amino acids, random errors yield a ratio of 1:2

Public database coverage

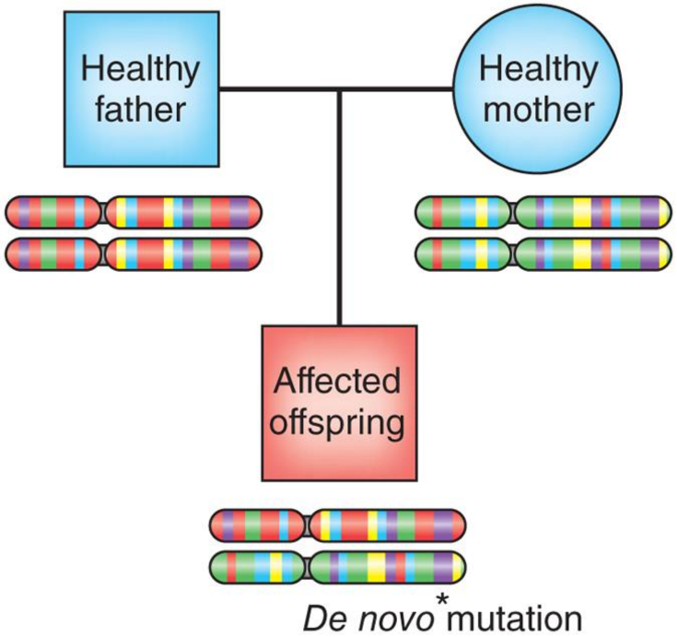


Individual	cSNP calls	# in dbSNP	% in dbSNP	# heterozygous	# homozygous
NA18507 (YRI)	19720	17577	89.1%	12896	6824
NA18517 (YRI)	19737	17326	87.8%	13039	6698
NA19129 (YRI)	19761	17298	87.5%	12845	6916
NA19240 (YRI)	19517	17168	88.0%	12866	6651
NA18555 (CHB)	16047	14894	92.8%	9181	6866
NA18956 (JPT)	16011	14848	92.7%	9132	6879
NA12156 (CEU)	16119	15250	94.6%	10179	5940
NA12878 (CEU)	15970	15051	94.2%	9928	6042
FSS10066 (Eur)	16229	15144	93.3%	10240	5989
FSS10208 (Eur)	16073	15018	93.4%	9966	6107
FSS22194 (Eur)	16094	15128	94.0%	10005	6089
FSS24895 (Eur)	15986	15027	94.0%	9920	6066

→ More variants identified in exomes from African than in European ancestry, larger proportion of European variants covered in public databases

De Novo Mutations and Errors

- **Assuming:**
  - Mutation rate of 2.5 x 10^{-8}
  - 20 Mbp of captured exome
  - Calling false positive rate (false heterozygote) of 1 x 10^{-6} (specificity of 99.9999%, Q60)
- **We expect:**
  - ~0.5 actual *de novo* non-synonymous variants per proband, and 20 false positives, i.e. FDR = 97.6%
  - Not considering false negative variants in parents…



Systematic errors & likely false positives

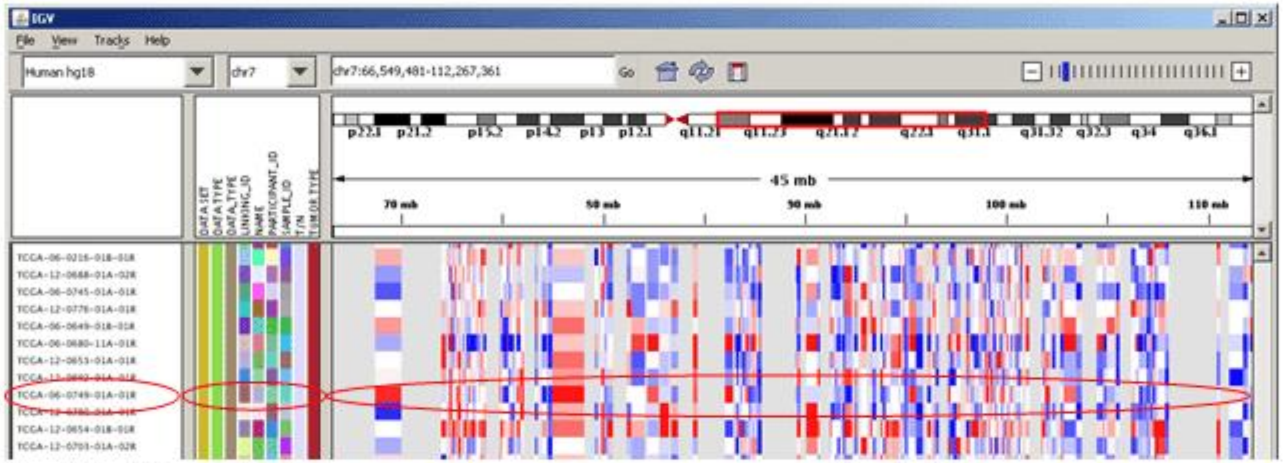
- Verify your variants using a different technology before follow up
- Unless isolated population, unrelated cases frequently have different mutations
- Is gene a likely false positive?
  - Large genes: *TTN, USH2A*
  - Lots of paralogs/part of gene family: *MUC*, ANK**
  - Don't rule out if phenotype makes sense! E.g.
    - *TTN:* dilated cardiomyopathy and muscular dystrophy
    - *MUC1*: medullary cystic, kidney disease
    - *KRT**: ichthyosis, keratoderma, keratosis

Checking underlying alignment files

BERLIN INSTITUTE OF HEALTH
Charité & Max Delbrück Center

- Integrative Genomics Viewer (IGV)
  - Java-based genome-browser, download/documentation: http://www.broadinstitute.org/igv/
  - Support for diverse data files, e.g. sorted .sam, .bam, .aligned, .psl, .pslx, and .bed, and multiple tracks

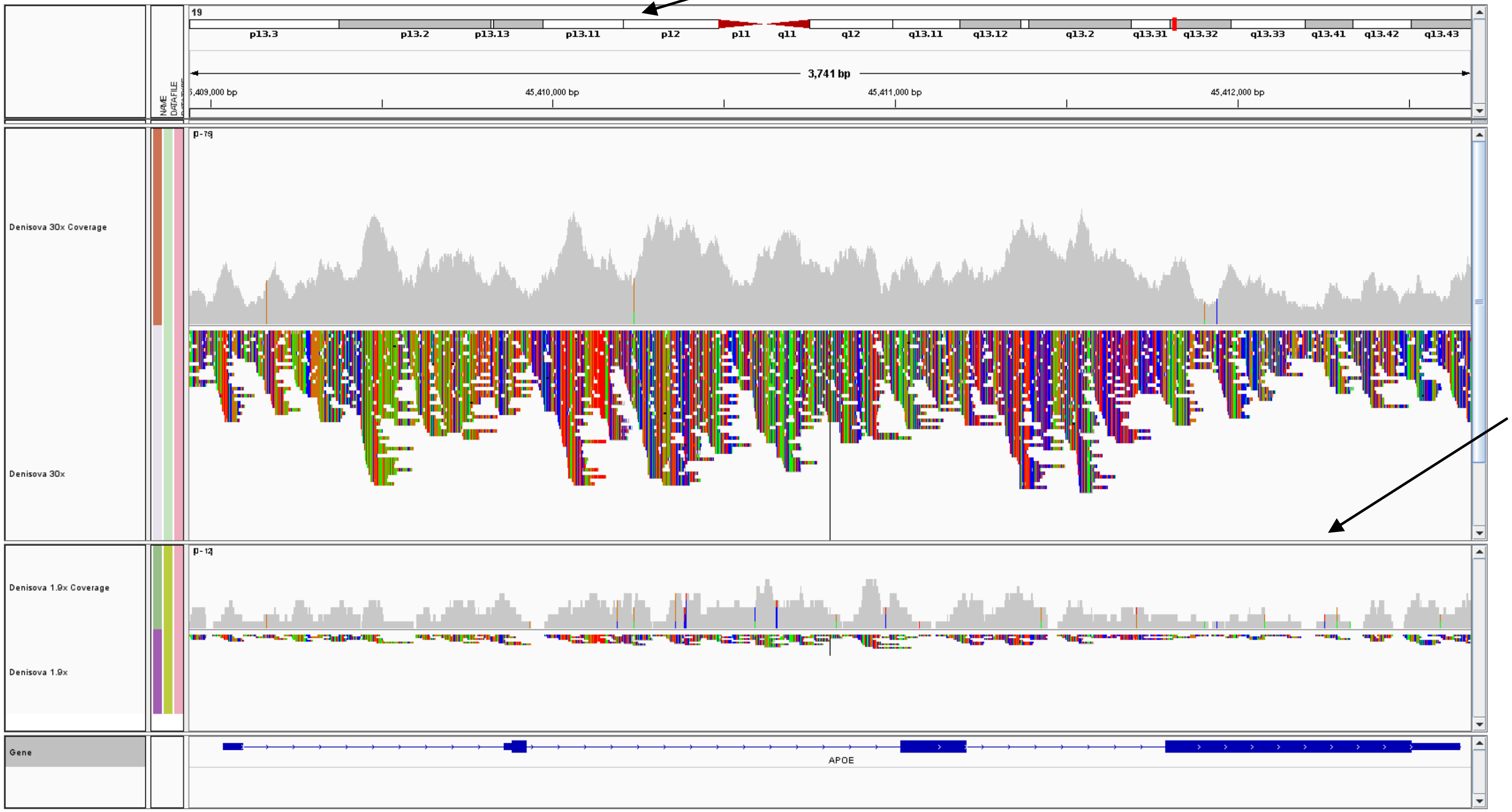
J.T. Robinson et al.
Nature Biotechnology
29, 24–26 (2011)



Track identifier Attributes Track data

Integrative Genomics Viewer (IGV)

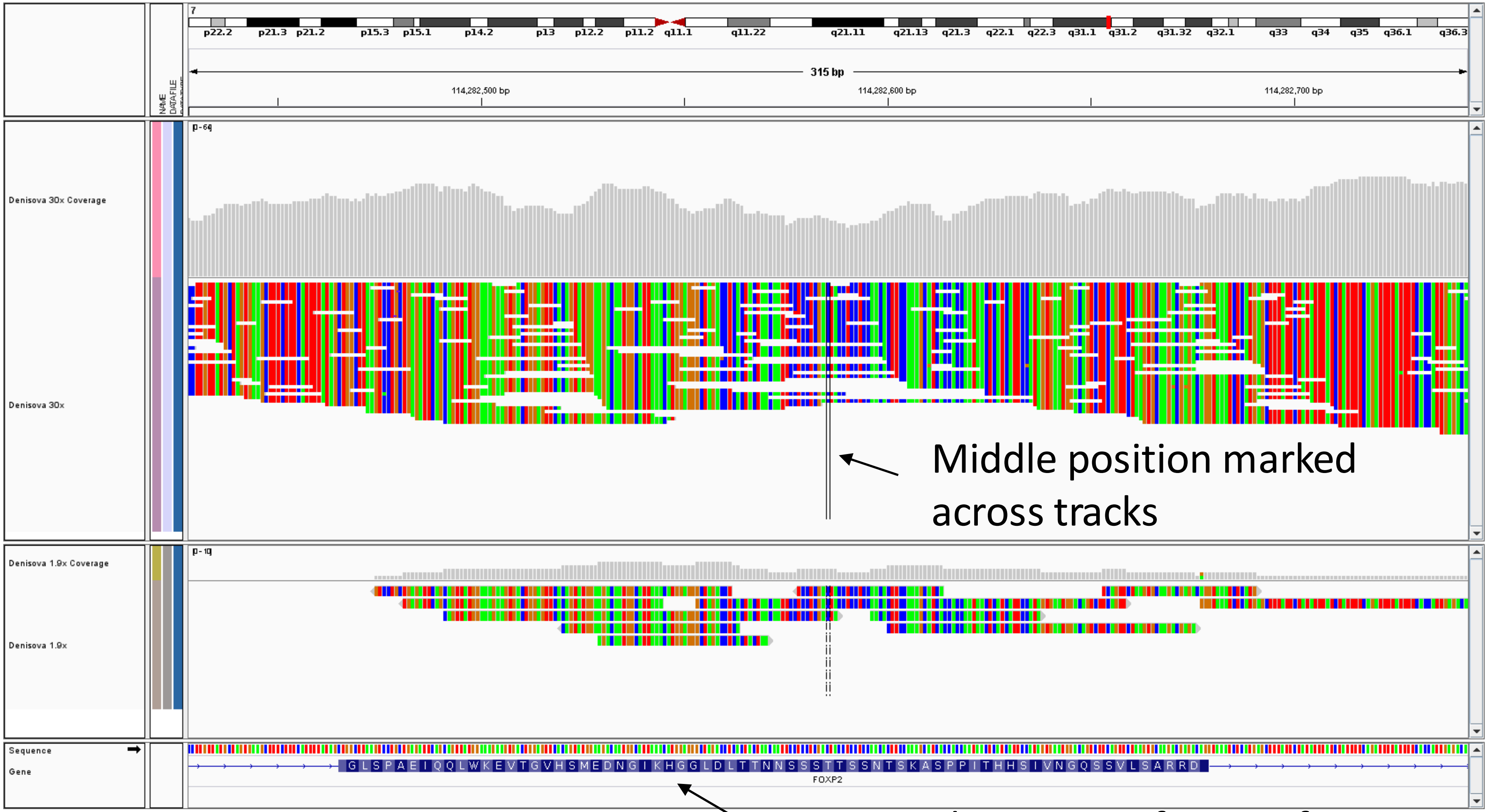
Chromosome view for easier navigation



Compare samples in multiple BAM files

Gene/ transcript annotation

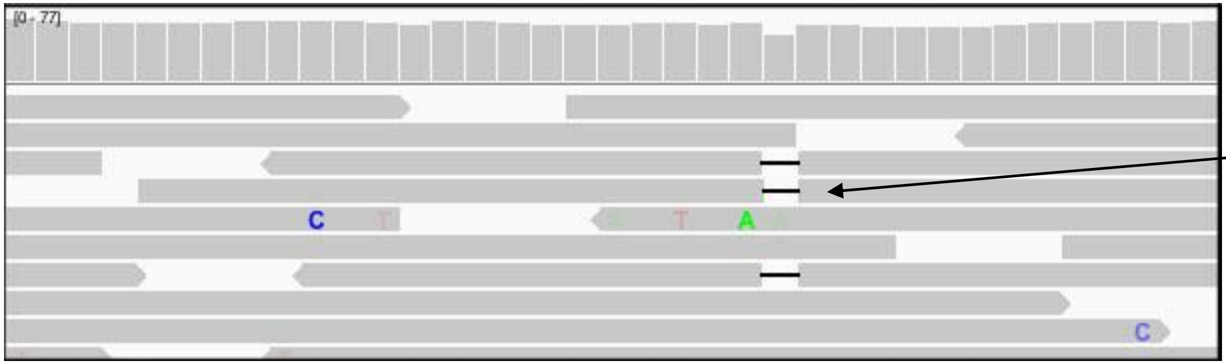
A more detailed view


7
p22.2 p21.3 p21.2 p15.3 p15.1 p14.2 p13 p12.2 p11.2 q11.1 q11.22 q21.11 q21.13 q21.3 q22.1 q22.3 q31.1 q31.2 q31.32 q32.1 q33 q34 q35 q36.1 q36.3
315 bp
114,282,500 bp
114,282,600 bp
114,282,700 bp
NAME
DATA FILE
[0 - 64]
Denisova 30x Coverage
Denisova 30x
[0 - 10]
Denisova 1.9x Coverage
Denisova 1.9x
Sequence
Gene
G L S P A E I Q Q L W K E V T G V H S M E D N G I K H G G L D L T T N N S S S T T S S N T S K A S P P I T H H S I V N G Q S S V L S A R R D
FOXP2

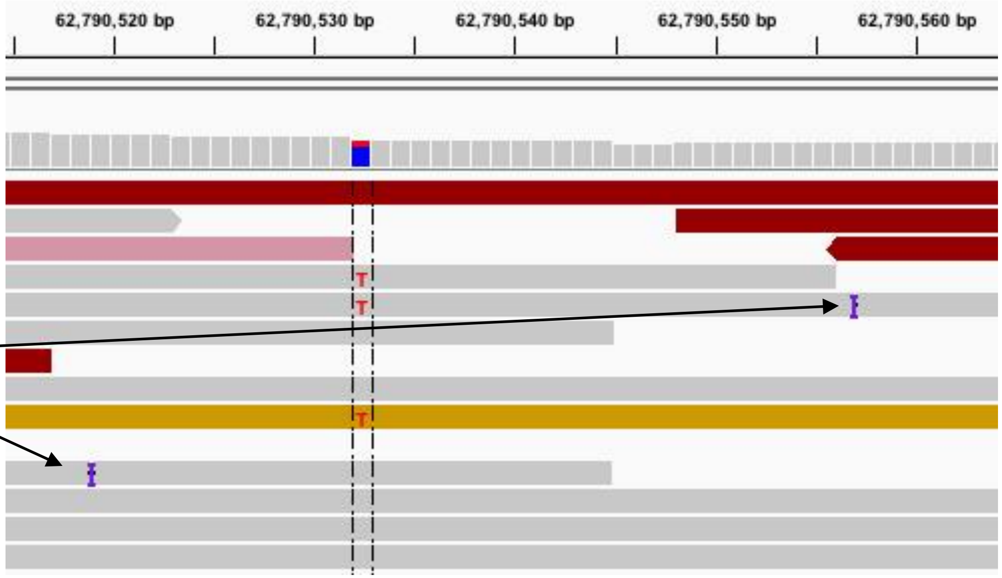
Middle position marked across tracks

Amino acid sequence from RefSeq

Insertions/deletions



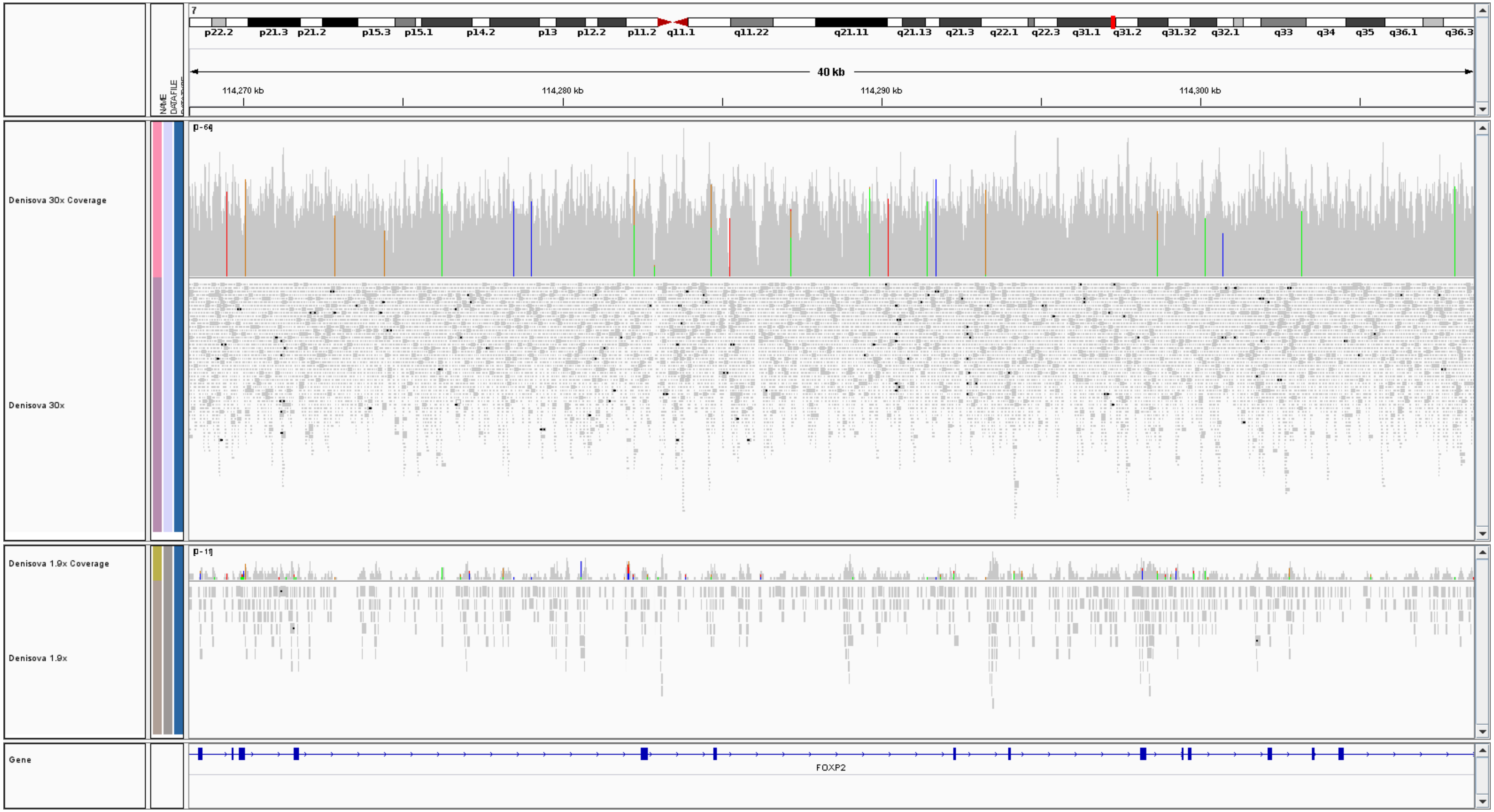
Deletion in reads



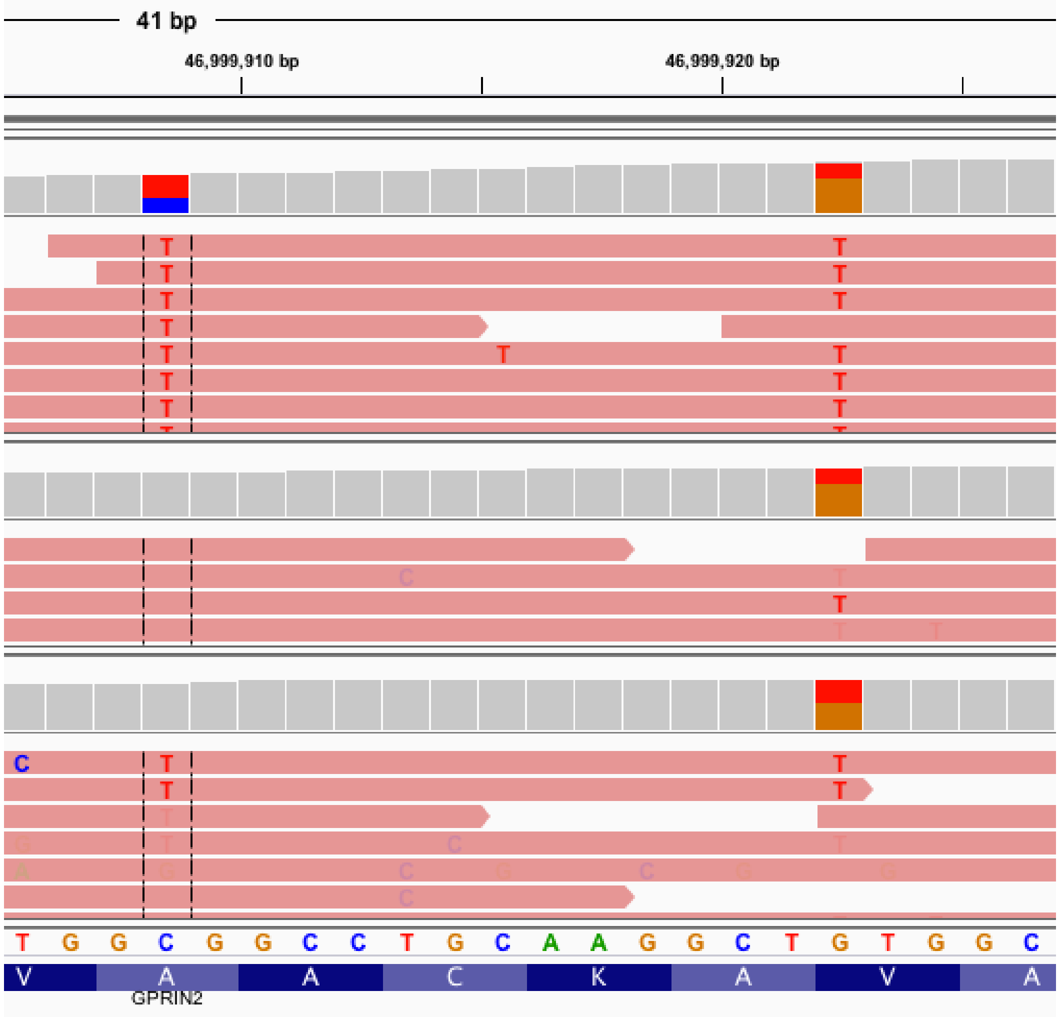
Insertion in reads

http://www.broadinstitute.org/igv/book/export/html/6

Viewing limit ~40kb



Allelic balance?



Total count: 15
A : 0
C : 6 (40%, 1+, 5-)
G : 0
T : 9 (60%, 9+, 0-)
N : 0

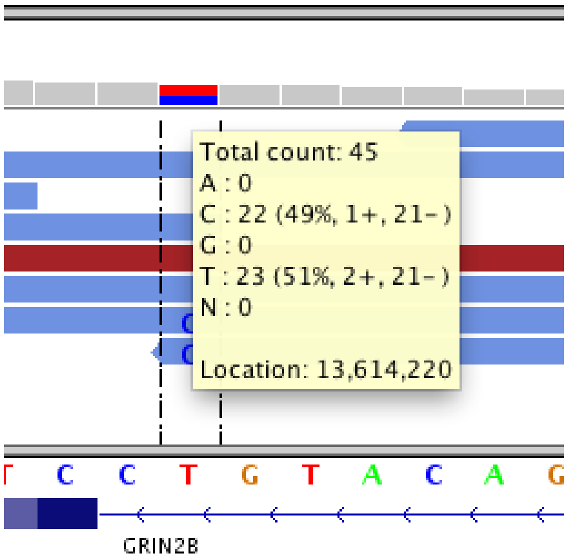
Proband

Father

Total count: 26
A : 0
C : 21 (81%, 19+, 2-)
G : 1 (4%, 1+, 0-)
T : 4 (15%, 4+, 0-)
N : 0

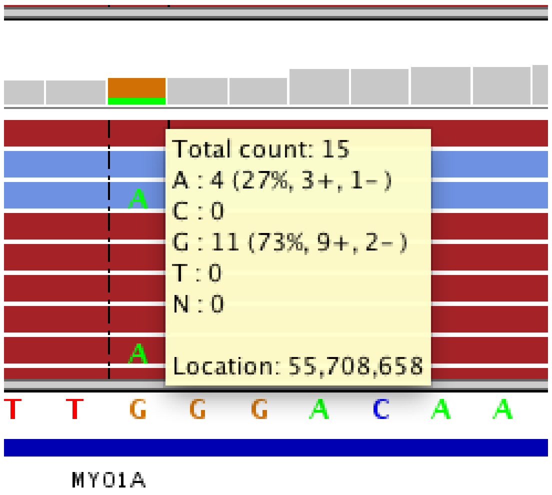
Mother

Strand balance?



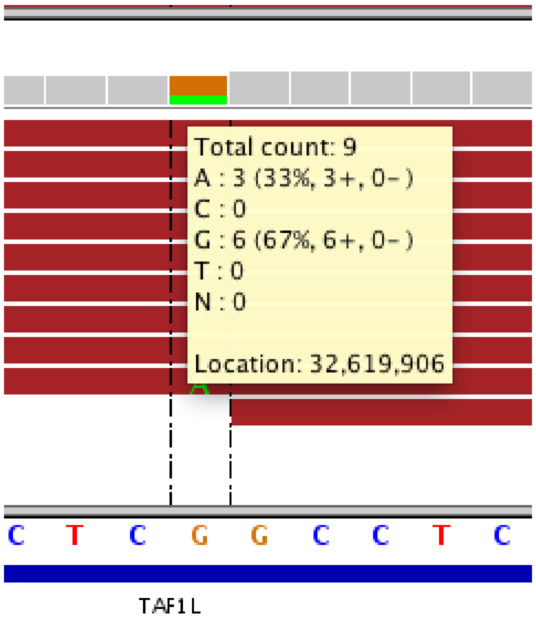
Total count: 45
A : 0
C : 22 (49%, 1+, 21-)
G : 0
T : 23 (51%, 2+, 21-)
N : 0

Location: 13,614,220



Total count: 15
A : 4 (27%, 3+, 1-)
C : 0
G : 11 (73%, 9+, 2-)
T : 0
N : 0

Location: 55,708,658

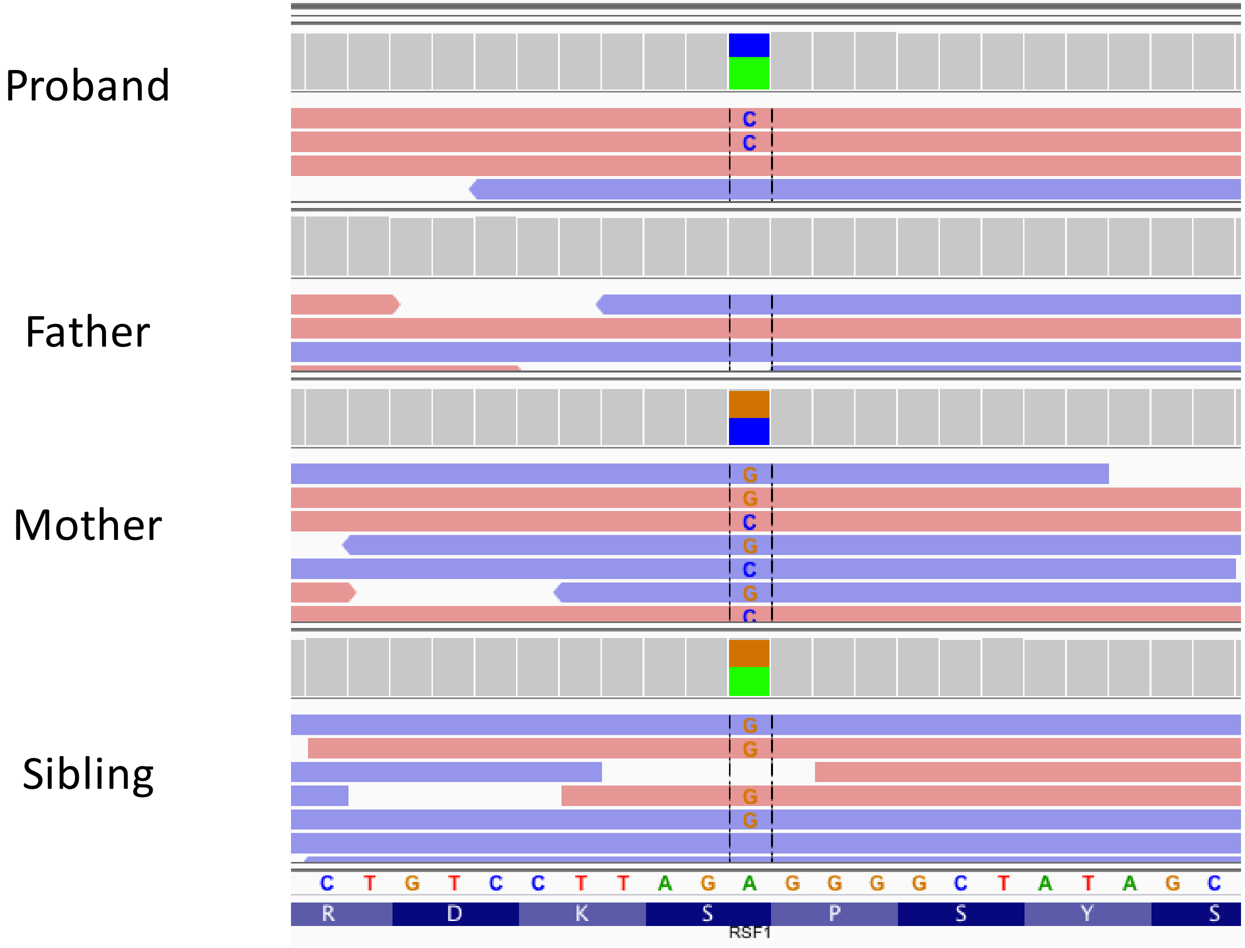


Total count: 9
A : 3 (33%, 3+, 0-)
C : 0
G : 6 (67%, 6+, 0-)
T : 0
N : 0

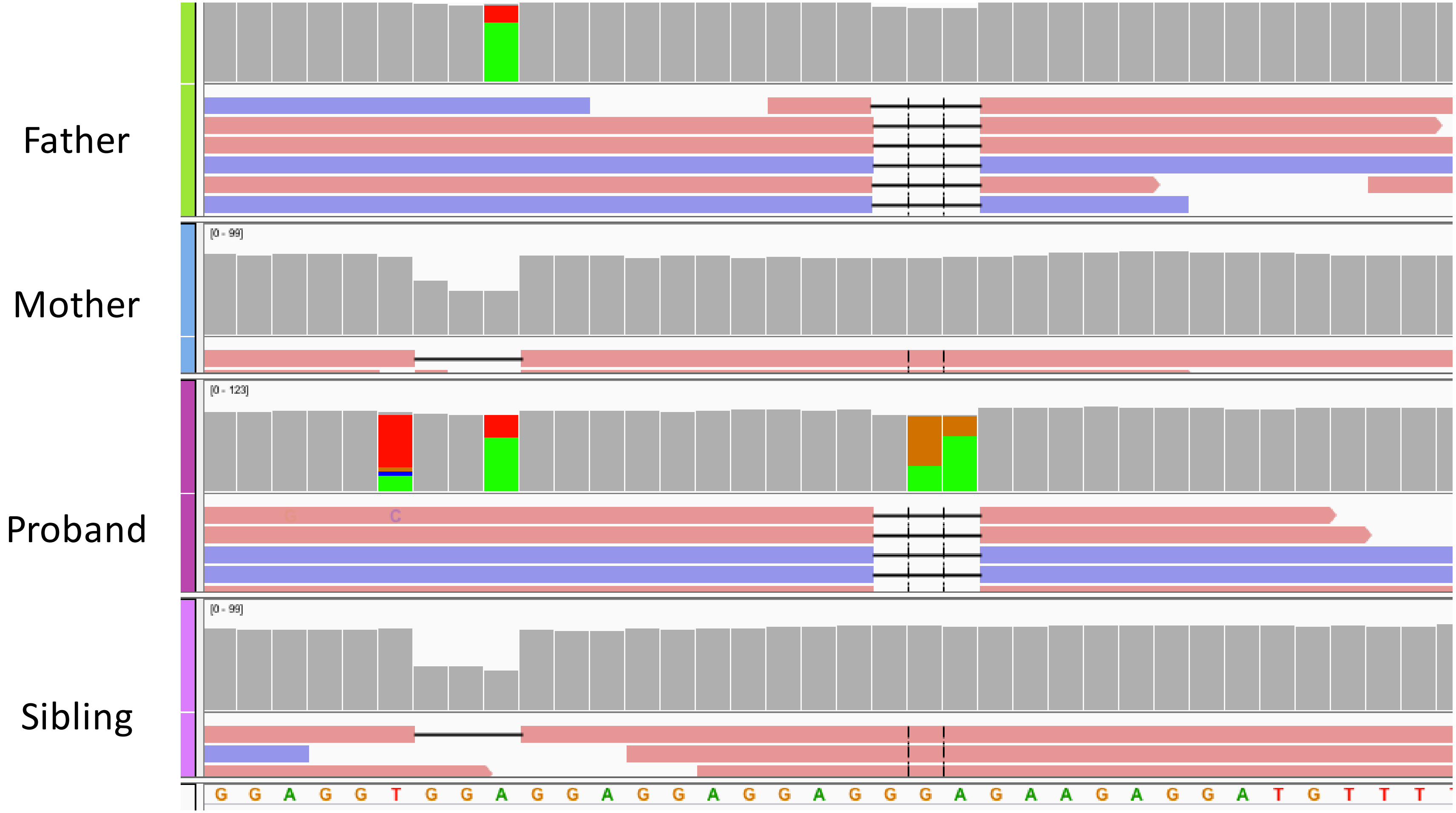
Location: 32,619,906

Note: You only expect even sampling from both strands if both strands can make it into your sequencing library and sequencing reaction

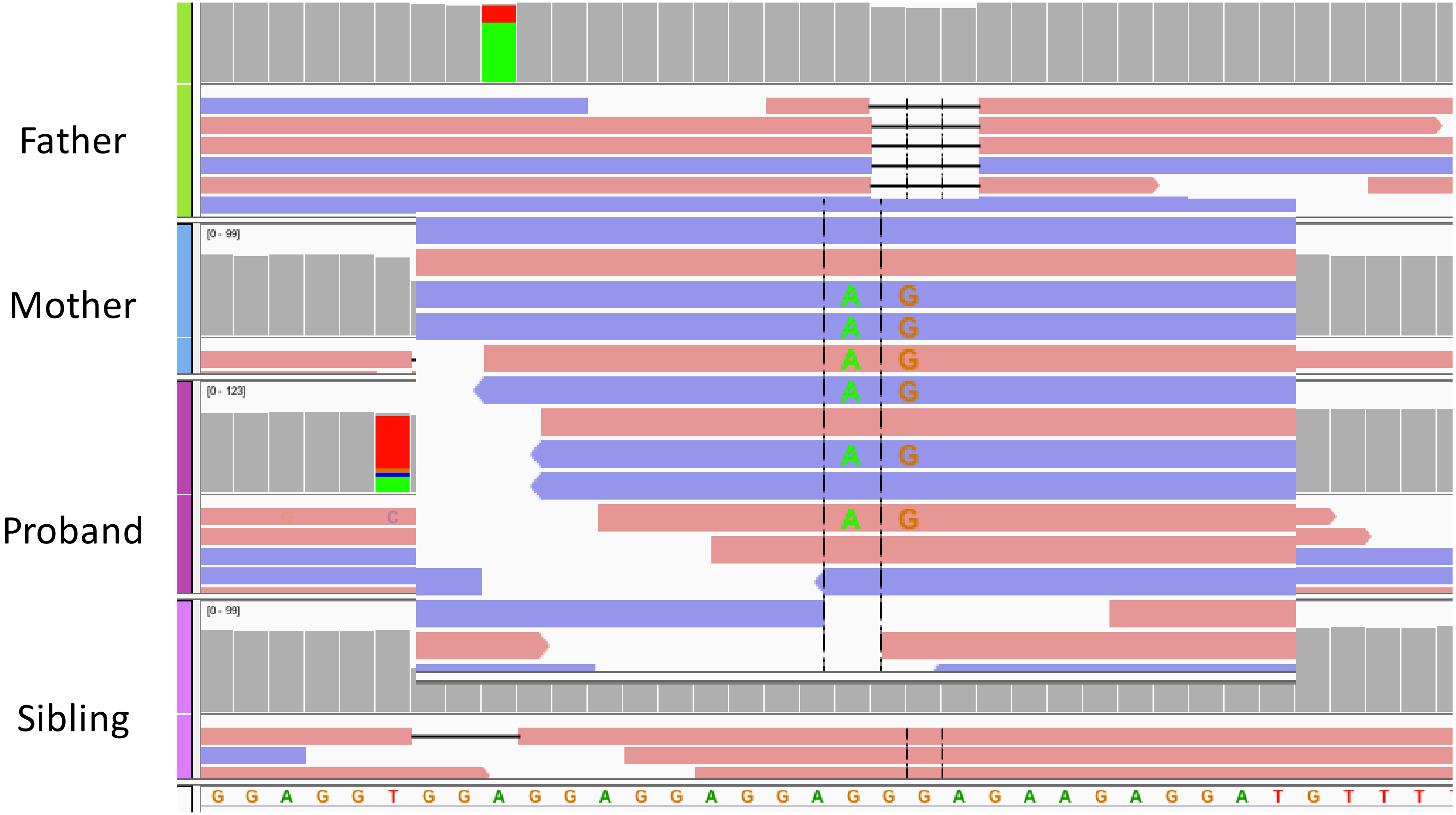
Tri-allelic sites



Low complexity regions

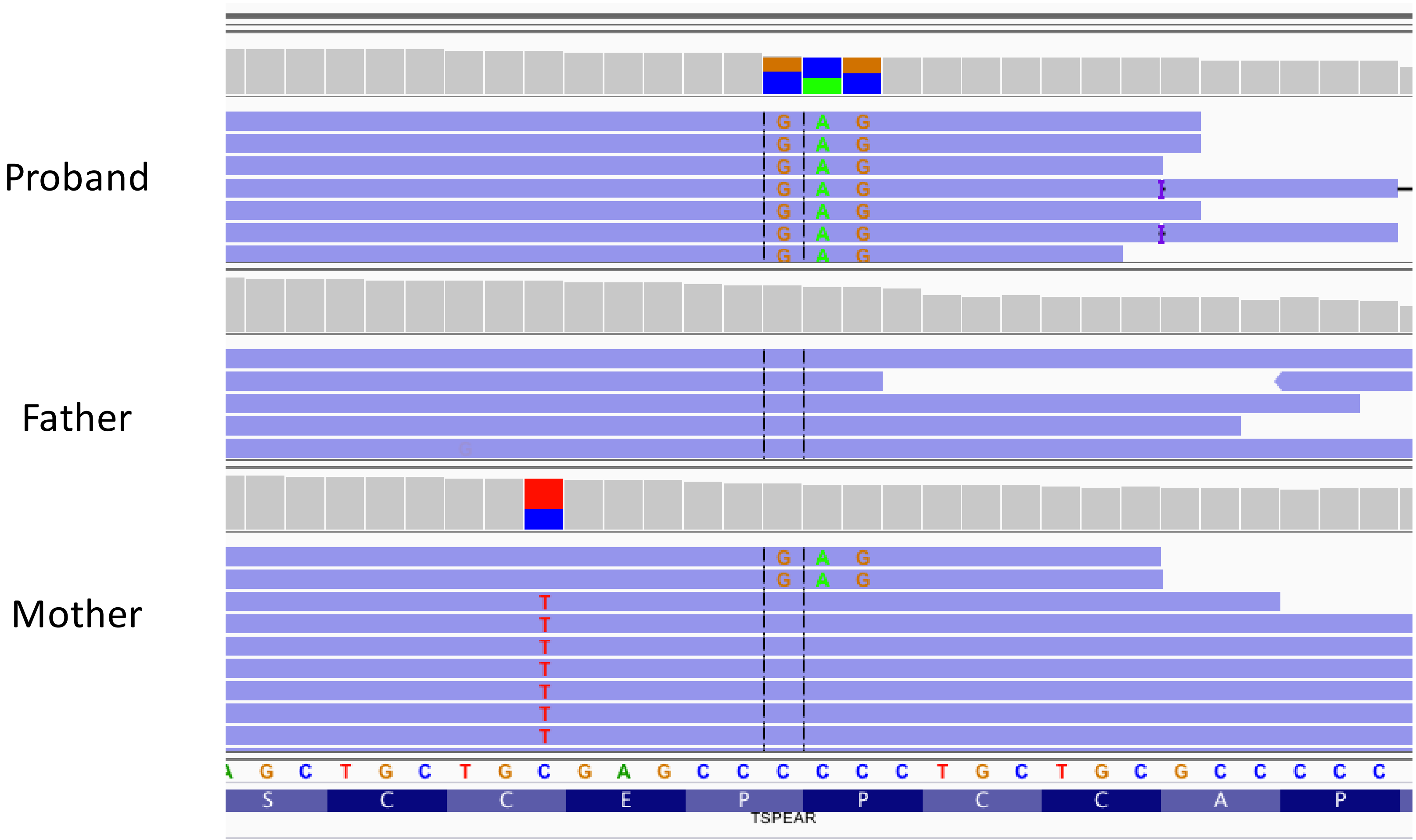

Father
Mother
[0 - 99]
Proband
[0 - 123]
Sibling
[0 - 99]
G G A G G T G G A G G A G G A G G A G G G A G A A G A G G A T G T T T

Low complexity regions

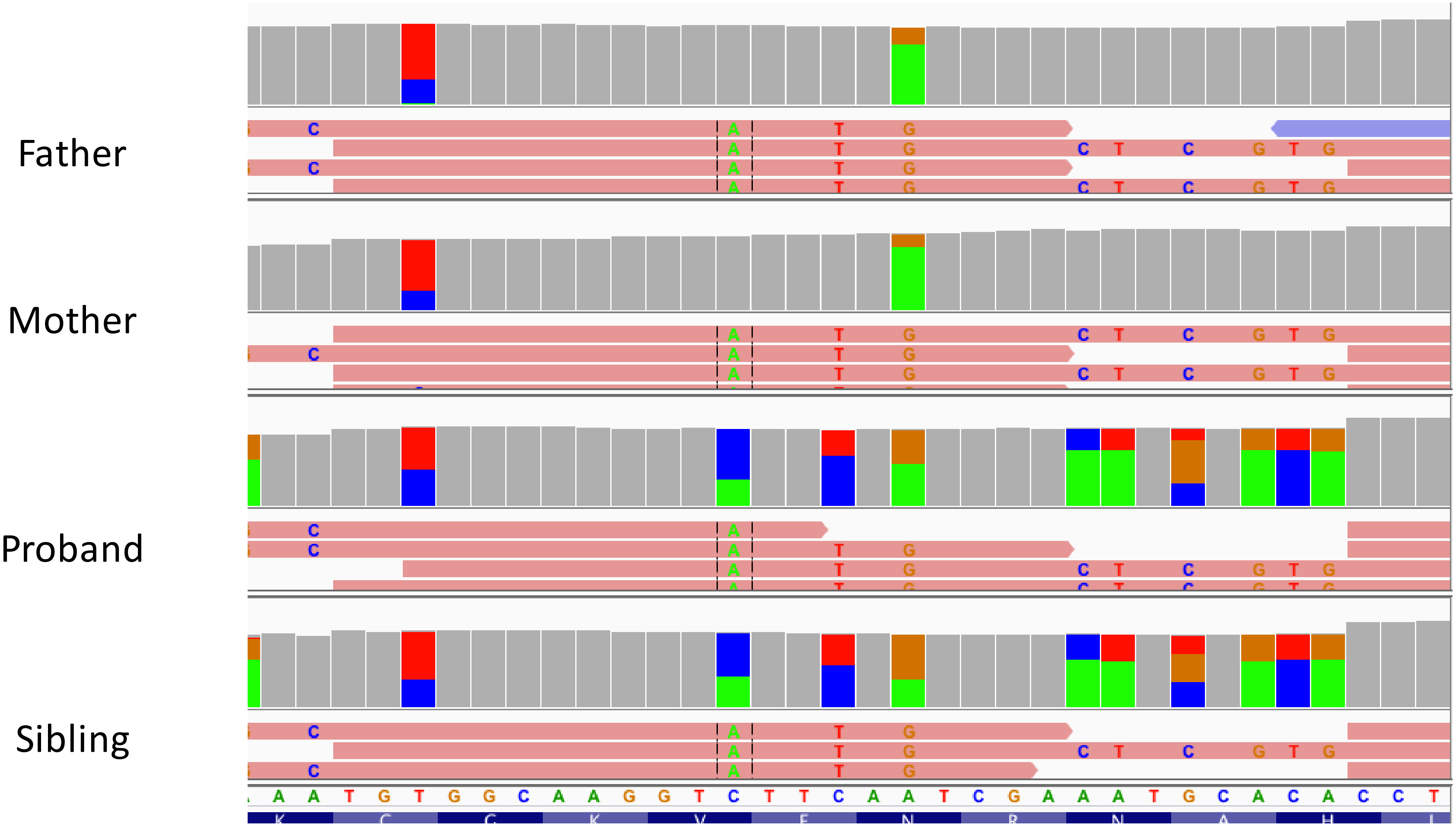


Low complexity regions (2)

BERLIN INSTITUTE OF HEALTH
Charité & Max Delbrück Center



Segmental duplication / assembly issue


Father
Mother
Proband
Sibling
A A T G T G G C A A G G T C T T C A A T C G A A A T G C A C A C C T
K C G K V F N R N A H I

ANNOTATION OF VARIANTS

1. Gene model sources and genome builds
2. Transcript models and predicted variant effects
3. HGVS - usage and validation
4. Variant sources, databases and underlying evidence
5. Variant beacons

Human genome builds

- GRCh37 / hg19 was released in 02/2009 and is still widely used
  - Ensembl and UCSC differ in mitochondrial genome sequence
- GRCh38 / hg38 first released 12/2013
  - Extended patch system (now p10, 01/2017)
  - Patches and alternate haplotypes complicate alignment and other algorithms, causing very slow adaptation

GRCh38 updates:

- > 100 assembly gaps closed or reduced
- MT: Cambridge Reference Sequence (rCRS)
- 261 alternate loci
- Centromere model integrated
- 150 Mb increase in non-N bases

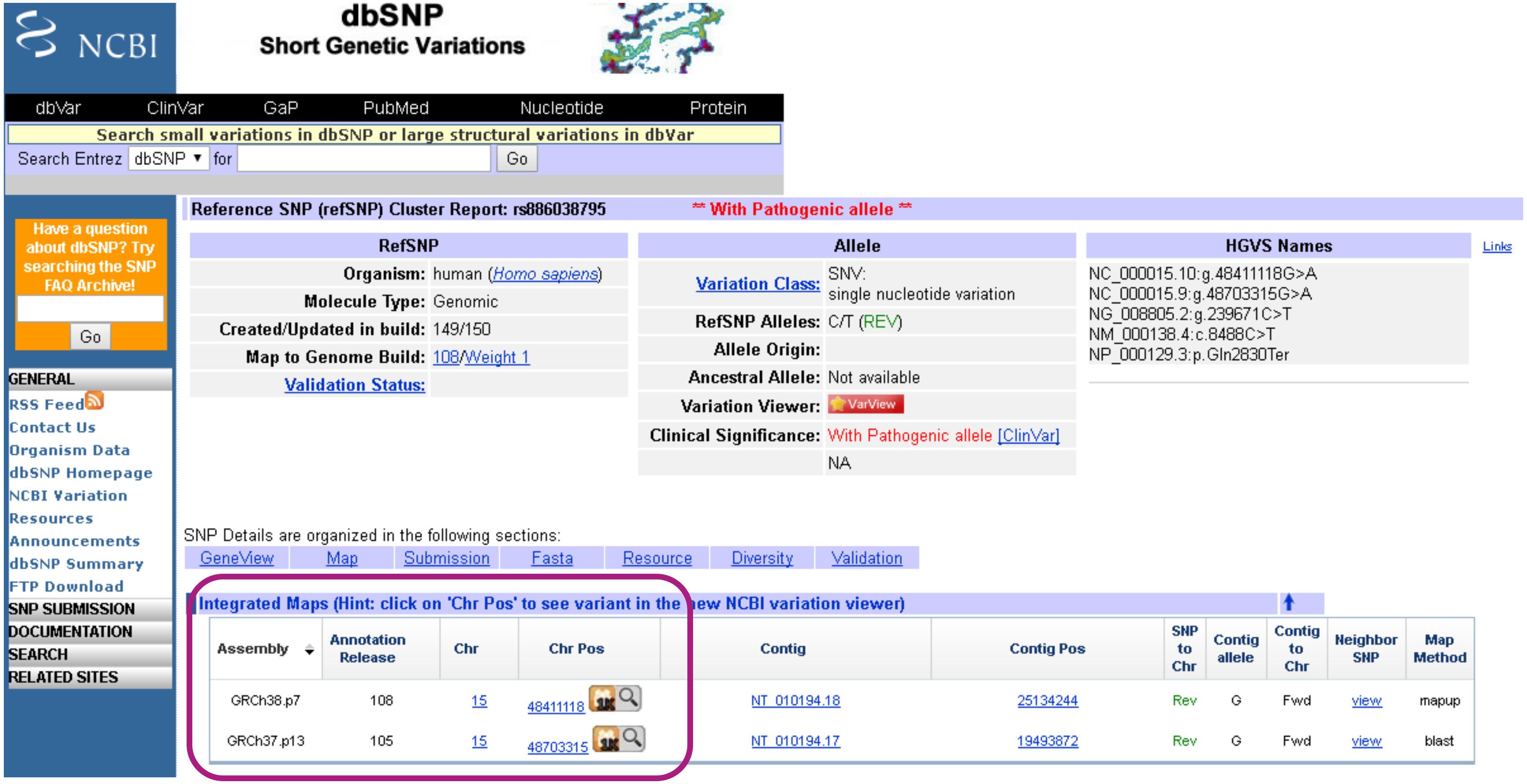
Coordinate conversions:

- http://www.ensembl.org/Homo_sapiens/Tools/AssemblyConverter
- https://www.ncbi.nlm.nih.gov/genome/tools/remap
- http://genome.ucsc.edu/cgi-bin/hgLiftOver

Human genome builds (2)

BERLIN INSTITUTE OF HEALTH
Charité & Max Delbrück Center

- *Make sure to check for the appropriate genome build before providing coordinates to any tools!*



dbSNP
Short Genetic Variations

Search small variations in dbSNP or large structural variations in dbVar

Reference SNP (refSNP) Cluster Report: rs886038795 ** With Pathogenic allele **

RefSNP	
Organism:	human (*Homo sapiens*)
Molecule Type:	Genomic
Created/Updated in build:	149/150
Map to Genome Build:	108/Weight 1
Validation Status:	

Allele	
Variation Class:	SNV: single nucleotide variation
RefSNP Alleles:	C/T (REV)
Allele Origin:	
Ancestral Allele:	Not available
Variation Viewer:	VarView
Clinical Significance:	With Pathogenic allele [ClinVar]
	NA

HGVS Names

NC_000015.10:g.48411118G>A
NC_000015.9:g.48703315G>A
NG_008805.2:g.239671C>T
NM_000138.4:c.8488C>T
NP_000129.3:p.Gln2830Ter

SNP Details are organized in the following sections:

Integrated Maps (Hint: click on 'Chr Pos' to see variant in the new NCBI variation viewer)

Assembly	Annotation Release	Chr	Chr Pos	Contig	Contig Pos	SNP to Chr	Contig allele	Contig to Chr	Neighbor SNP	Map Method
GRCh38.p7	108	15	48411118	NT_010194.18	25134244	Rev	G	Fwd	view	mapup
GRCh37.p13	105	15	48703315	NT_010194.17	19493872	Rev	G	Fwd	view	blast

Gene model sources

- NCBI (RefSeq), Ensembl (GENCODE), UCSC (knownGenes) distribute independent gene/transcript annotation sets
- Ensembl provides most comprehensive set
- Collaborative consensus coding sequence (CCDS) curates and revises a joined gene/transcript set

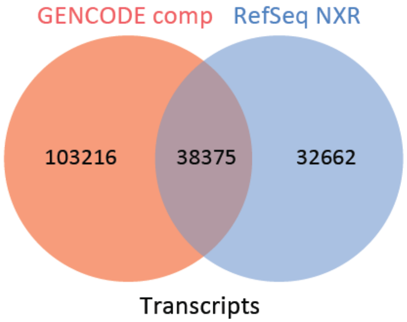
Annotating ~80 million variants in the WGS500 project (doi: 10.1186/gm543)

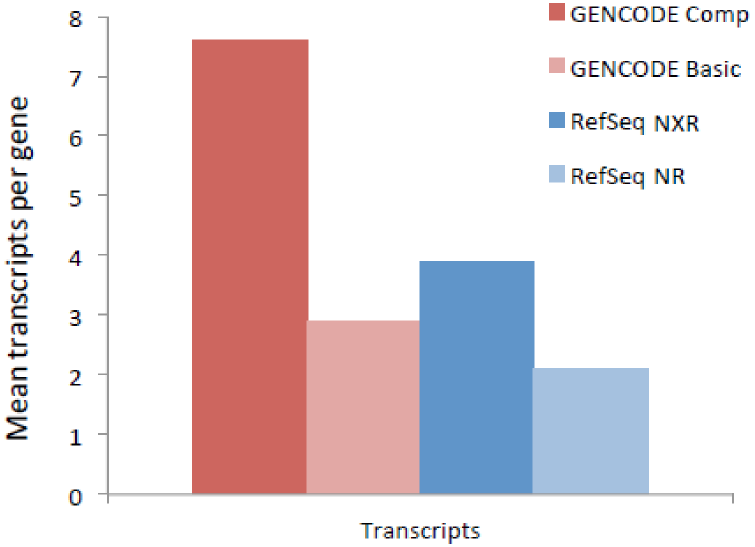
	REF+ENS	RefSeq	Ensembl	Match	Overall match [%]
Stopgain (SNV)	15,835	14,183	14,960	13,308	84.04
Frameshift insertion	6,980	5,298	6,495	4,813	68.95
Frameshift deletion	7,491	4,547	7,380	4,436	59.22
Stoploss (SNV)	946	503	906	463	48.94
Splicing	47,878	14,154	45,839	12,115	25.30
Nonsynonymous (SNV)	321,669	291,898	315,592	285,821	88.86

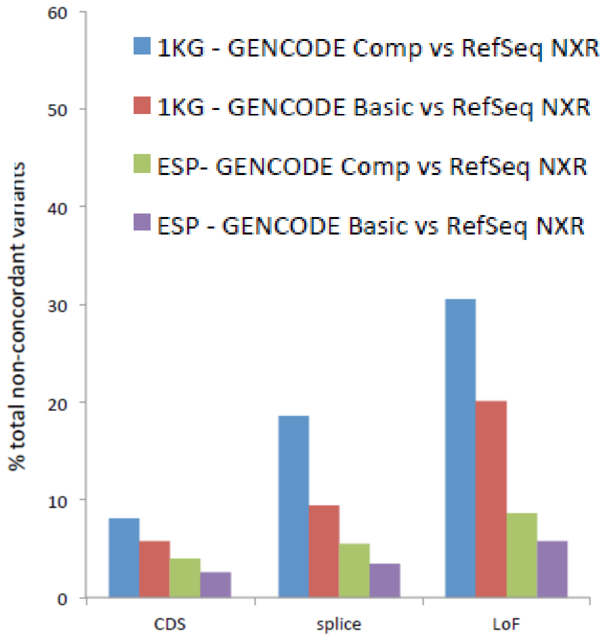
Transcript models and predicted variant effects

BERLIN INSTITUTE OF HEALTH
Charité & Max Delbrück Center

- Annotation sources differ significantly on transcript level
- Be inclusive to not miss a potentially damaging variant
- Never assume that annotations are perfect, if in doubt validate predicted transcript effect





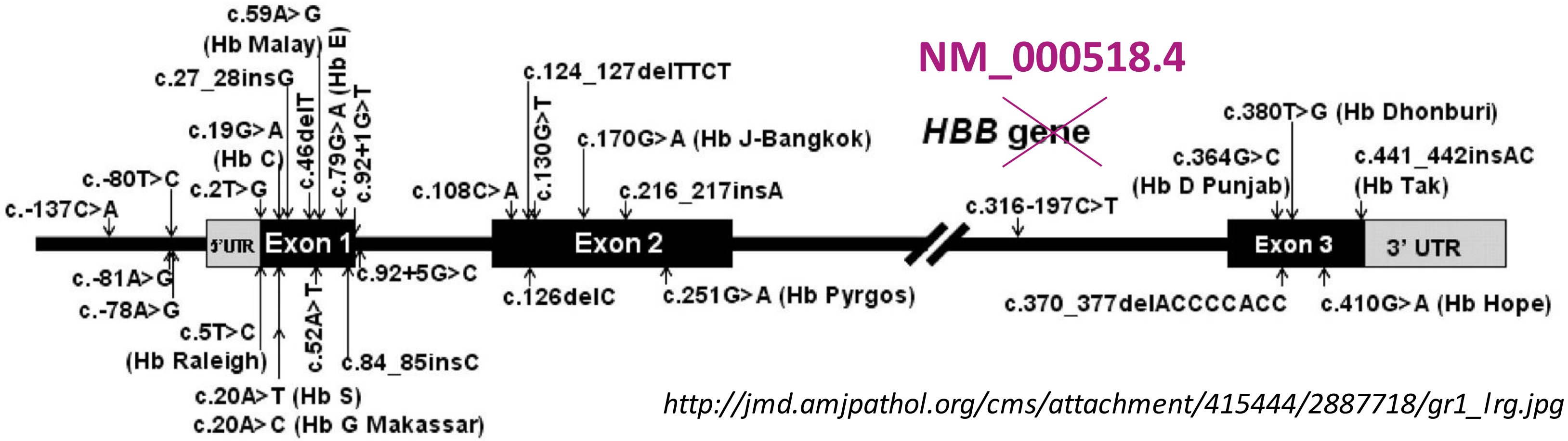


Modified from Frankish A et al. BMC Genomics 2015

HGVS - usage and validation

BERLIN INSTITUTE OF HEALTH
Charité & Max Delbrück Center

- Sequence Variant Nomenclature (http://varnomen.hgvs.org/)
- Frequently used in medical publications, unfortunately with large variation/deviations from standard
- Mostly impossible to computationally process
- If you *must* use it, run validation and conversion tools: mutalyzer.nl / VEP



http://jmd.amjpathol.org/cms/attachment/415444/2887718/gr1_lrg.jpg

Variant databases

- Many sources for variants around coding sequences: ESP, ExAC
  and genome-wide:
  1000 Genomes, UK10K, gnomAD, Haplotype Reference Consortium (HRC), Genomics England
- General variant repository for small and large studies: dbSNP
- Structural variants: dbVar, DGV

1000 Genomes

Exome Aggregation Consortium (ExAC)

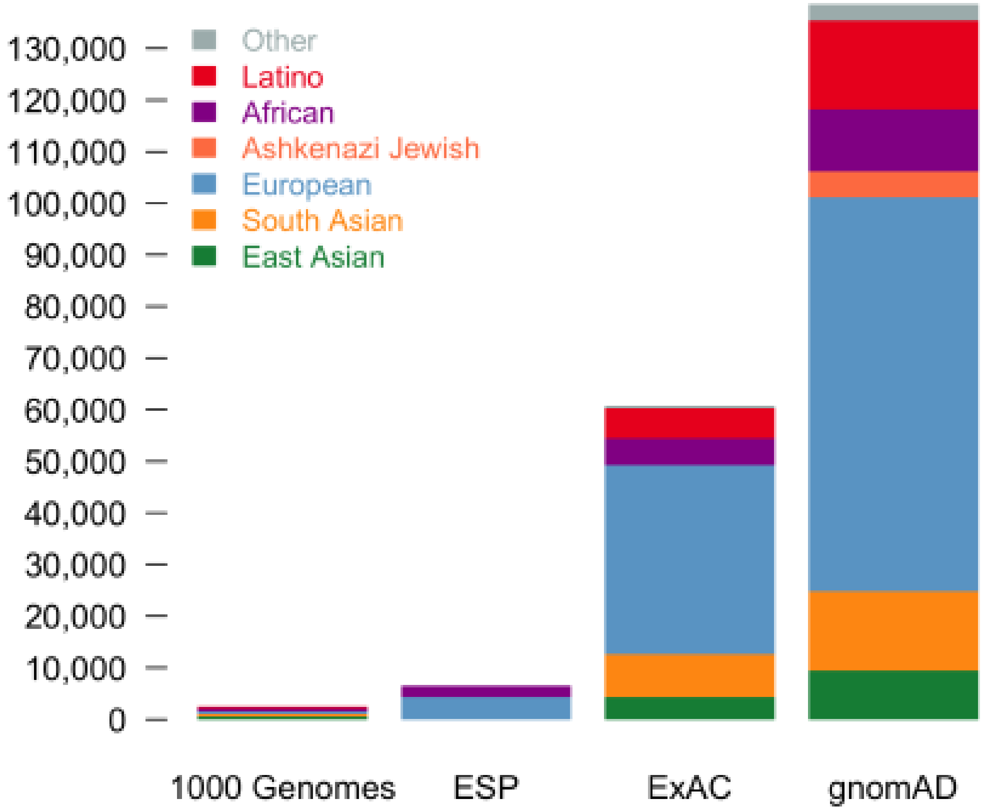
Genome Aggregation Database (gnomAD)

dbSNP
Short Genetic Variations

Genomics
england

Human populations in variant databases

- Variant data bases are biased towards individuals of European ancestry
- Frequencies summaries only available for some larger populations



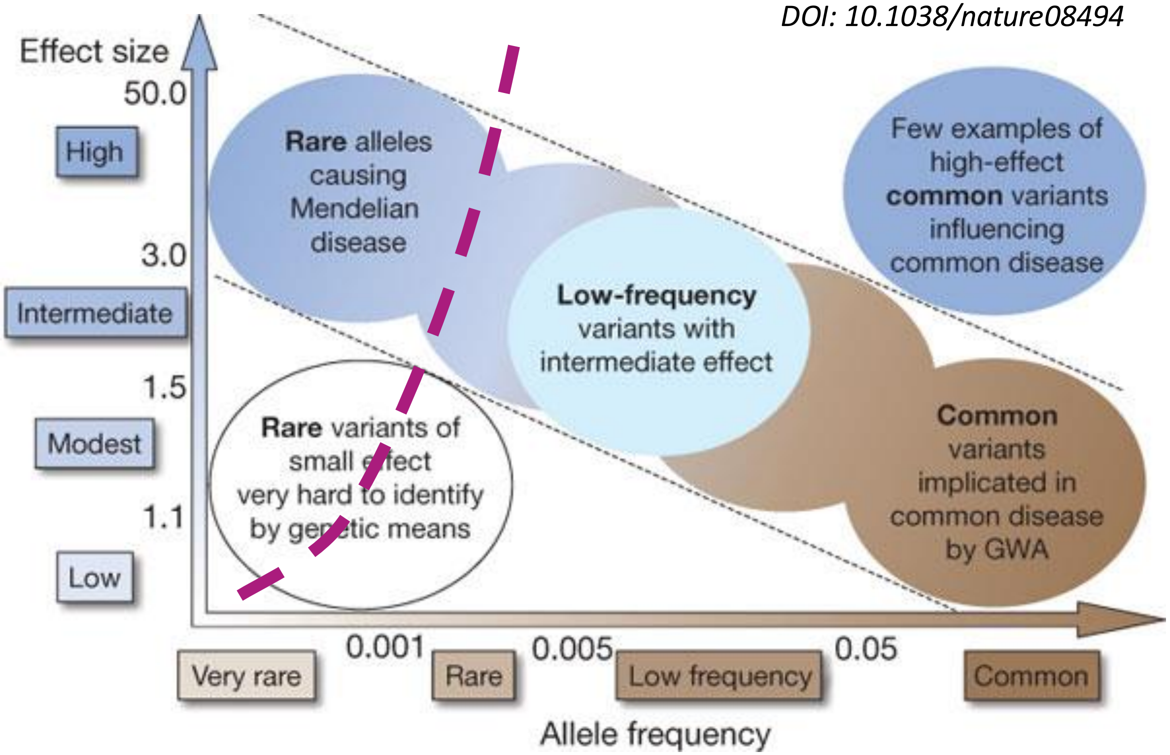
https://macarthurlab.org/2017/02/27/the-genome-aggregation-database-gnomad/

Databases include disease variants

BERLIN INSTITUTE OF HEALTH
Charité & Max Delbrück Center

- 1000 Genome project and others recruited “healthy individuals” – does not mean that disease alleles are absent!
- gnomAD excludes individuals with severe pediatric diseases
- Late-onset and less severe disease alleles likely present

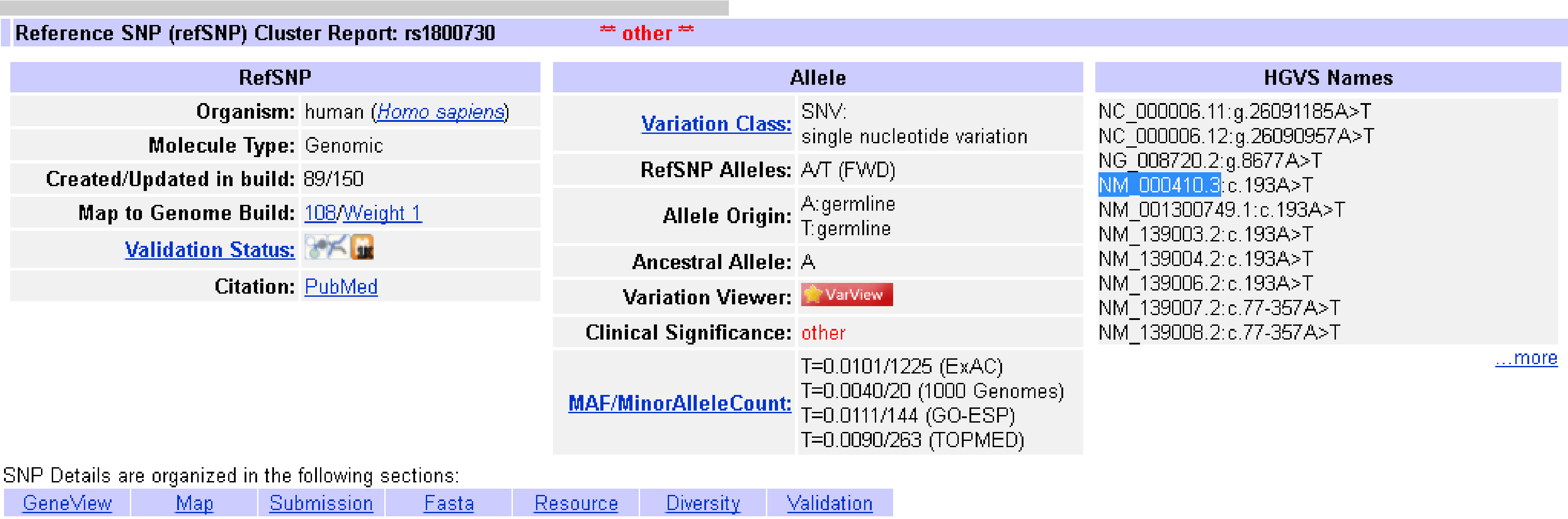
DOI: 10.1038/nature08494



Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Few examples of high-effect common variants influencing common disease
Low-frequency variants with intermediate effect
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
0.001
0.005
0.05
Very rare
Rare
Low frequency
Common
Allele frequency

dbSNP is not your database of choice

- > 325M rsIDs, only 130M with frequency information
- rsIDs reference a loci + allele length, *not* an allele; issues when frequency and genotype information are linked
- Somatic as well as germline, disease variants as well common



Reference SNP (refSNP) Cluster Report: rs1800730 ** other **

RefSNP	
Organism:	human (*Homo sapiens*)
Molecule Type:	Genomic
Created/Updated in build:	89/150
Map to Genome Build:	108/Weight 1
Validation Status:	
Citation:	PubMed

Allele	
Variation Class:	SNV: single nucleotide variation
RefSNP Alleles:	A/T (FWD)
Allele Origin:	A:germline T:germline
Ancestral Allele:	A
Variation Viewer:	VarView
Clinical Significance:	other

MAF/MinorAlleleCount:	T=0.0101/1225 (ExAC)
	T=0.0040/20 (1000 Genomes)
	T=0.0111/144 (GO-ESP)
	T=0.0090/263 (TOPMED)

HGVS Names
NC_000006.11:g.26091185A>T
NC_000006.12:g.26090957A>T
NG_008720.2:g.8677A>T
NM_000410.3:c.193A>T
NM_001300749.1:c.193A>T
NM_139003.2:c.193A>T
NM_139004.2:c.193A>T
NM_139006.2:c.193A>T
NM_139007.2:c.77-357A>T
NM_139008.2:c.77-357A>T
...more

SNP Details are organized in the following sections:

Clinical variant sources

NCBI ClinVar

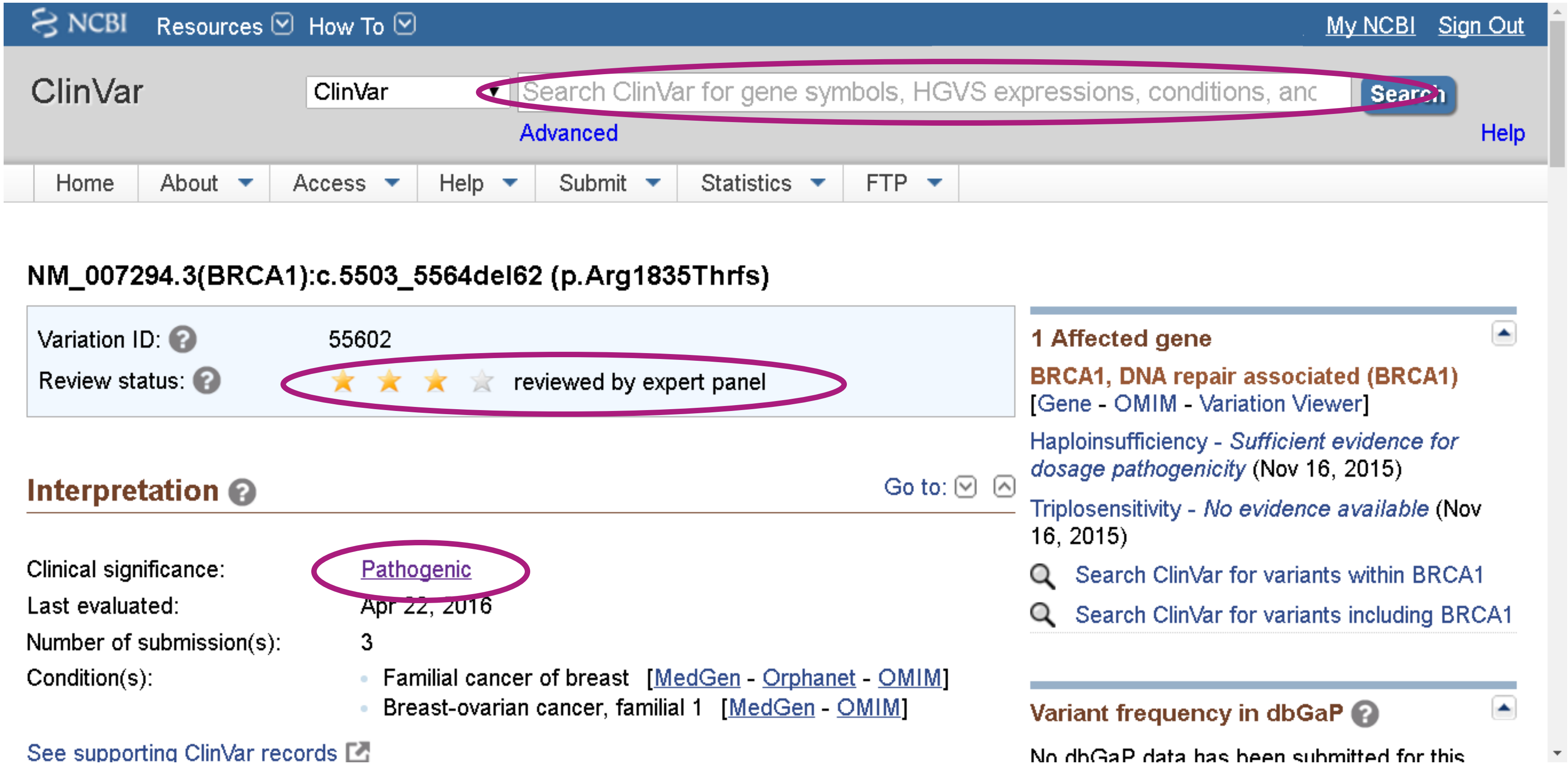
- Public domain, free
- >261k variants: 39k ‘pathogenic’ and 55k ‘benign’
- Clinical labs major submitters
- Goal: present agreement or conflict in clinical significance assignment
- Linking underlying evidence

Human Gene Mutation Database

- Commercial (Qiagen)
- Curated: inherited disease
- >203k mutations (2017.1)
- GWAS and associated variants
- Reference published evidence
- Free academic version with fewer variants (2 year delay)

Clinical variant sources: ClinVar

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



NM_007294.3(BRCA1):c.5503_5564del62 (p.Arg1835Thrfs)

Variation ID:	55602
Review status:	★★★☆ reviewed by expert panel

Interpretation

Go to:

Clinical significance:	Pathogenic
Last evaluated:	Apr 22, 2016
Number of submission(s):	3
Condition(s):	Familial cancer of breast [MedGen - Orphanet - OMIM] Breast-ovarian cancer, familial 1 [MedGen - OMIM]

See supporting ClinVar records

1 Affected gene

BRCA1, DNA repair associated (BRCA1)
[Gene - OMIM - Variation Viewer]

Haploinsufficiency - *Sufficient evidence for dosage pathogenicity* (Nov 16, 2015)

Triplosensitivity - *No evidence available* (Nov 16, 2015)

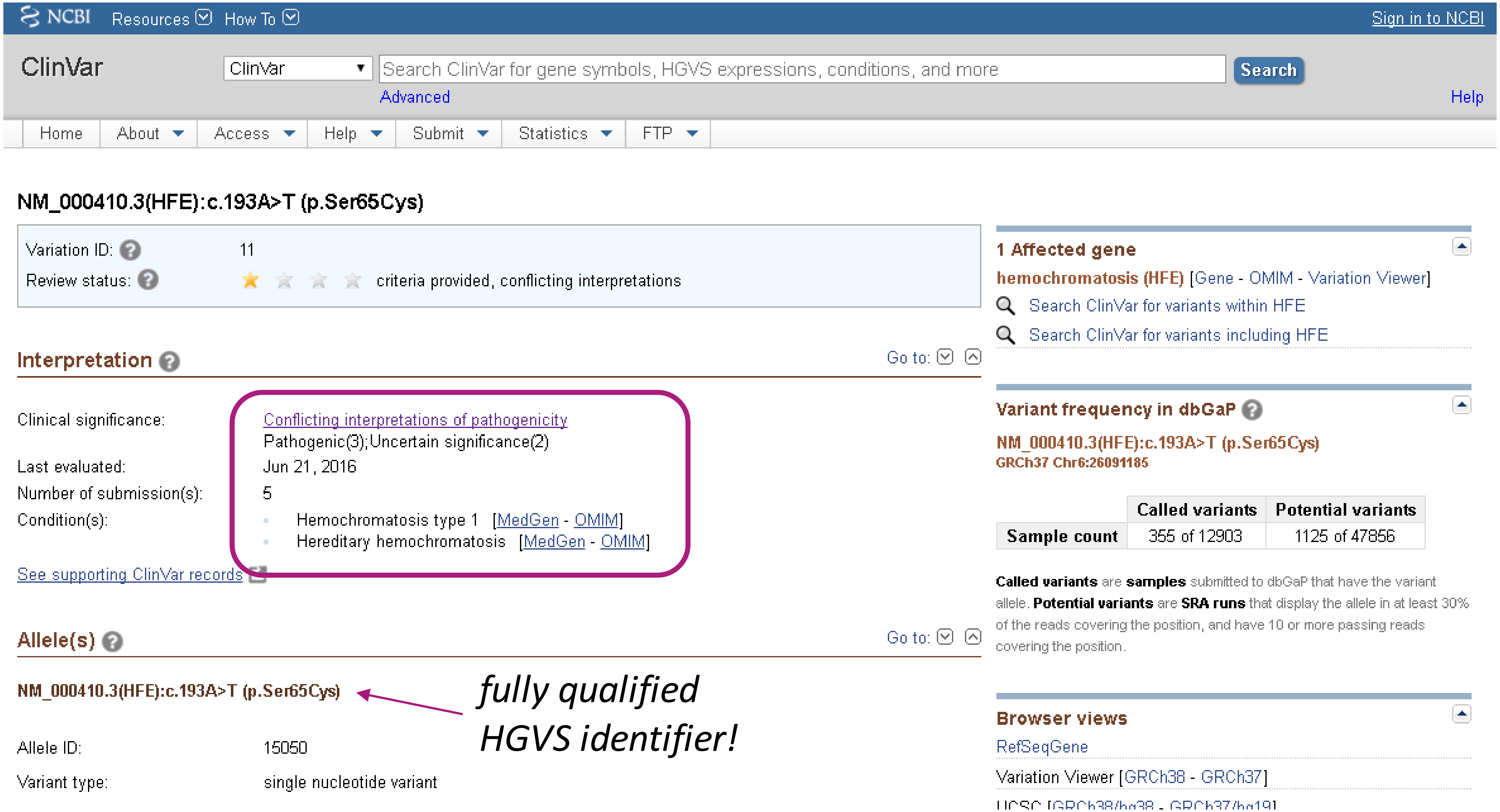
Search ClinVar for variants within BRCA1

Search ClinVar for variants including BRCA1

Variant frequency in dbGaP

No dbGaP data has been submitted for this

Clinical variant sources: ClinVar (2)



NM_000410.3(HFE):c.193A>T (p.Ser65Cys)

Variation ID:	11
Review status:	★☆☆☆ criteria provided, conflicting interpretations

Interpretation

Go to:

Clinical significance:	Conflicting interpretations of pathogenicity Pathogenic(3);Uncertain significance(2)
Last evaluated:	Jun 21, 2016
Number of submission(s):	5
Condition(s):	Hemochromatosis type 1 [MedGen - OMIM] Hereditary hemochromatosis [MedGen - OMIM]

See supporting ClinVar records

Allele(s)

Go to:

NM_000410.3(HFE):c.193A>T (p.Ser65Cys)

fully qualified HGVS identifier!

Allele ID:	15050
Variant type:	single nucleotide variant

1 Affected gene

hemochromatosis (HFE) [Gene - OMIM - Variation Viewer]

Search ClinVar for variants within HFE

Search ClinVar for variants including HFE

Variant frequency in dbGaP

NM_000410.3(HFE):c.193A>T (p.Ser65Cys)
GRCh37 Chr6:26091185

	Called variants	Potential variants
Sample count	355 of 12903	1125 of 47856

Called variants are **samples** submitted to dbGaP that have the variant allele. **Potential variants** are **SRA runs** that display the allele in at least 30% of the reads covering the position, and have 10 or more passing reads covering the position.

Browser views

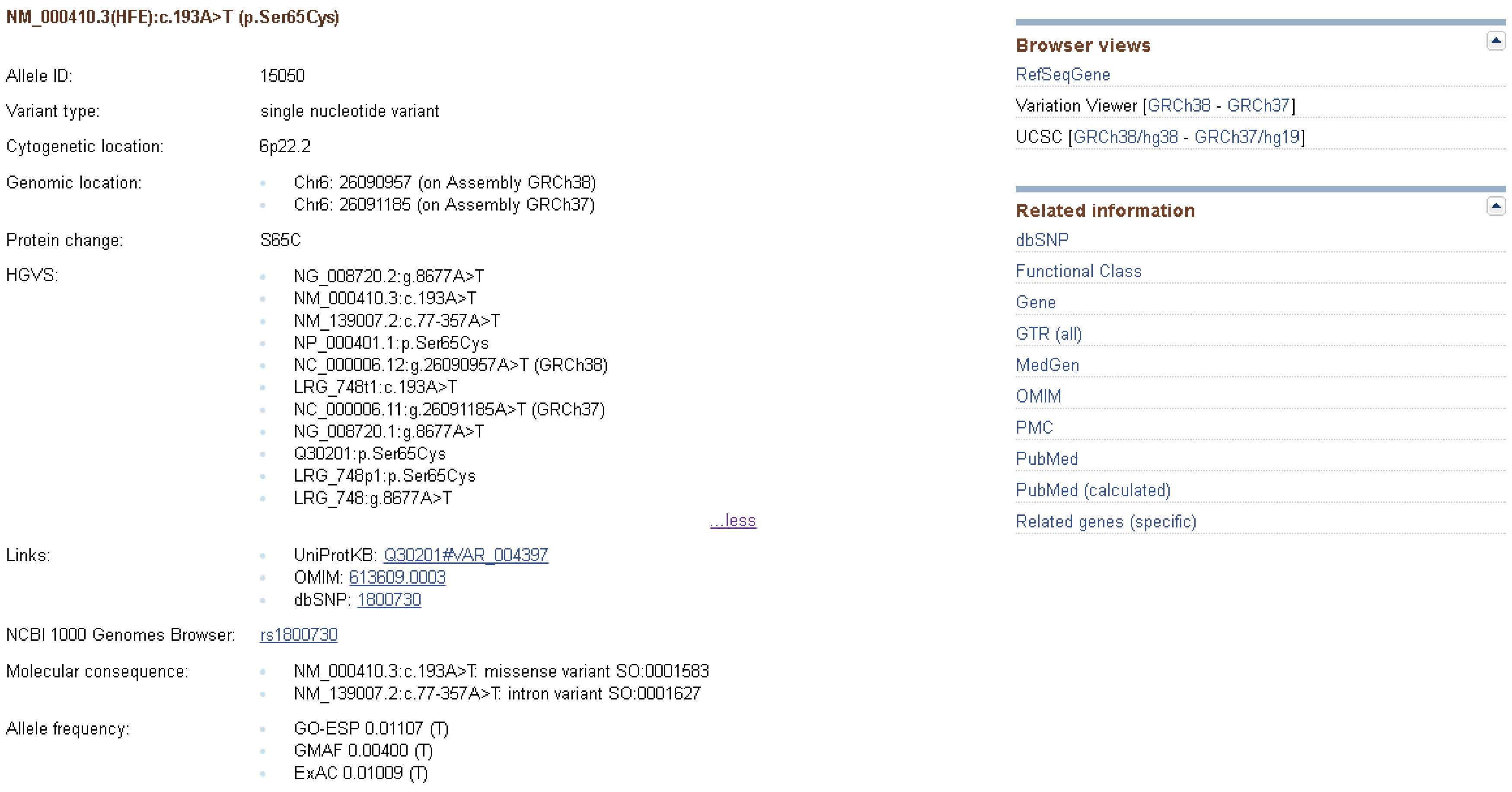
RefSeqGene

Variation Viewer [GRCh38 - GRCh37]

UCSC [GRCh38/hg38 - GRCh37/hg19]

Clinical variant sources: ClinVar (2)

BERLIN INSTITUTE OF HEALTH
Charité & Max Delbrück Center



NM_000410.3(HFE):c.193A>T (p.Ser65Cys)

Allele ID:	15050
Variant type:	single nucleotide variant
Cytogenetic location:	6p22.2
Genomic location:	Chr6: 26090957 (on Assembly GRCh38) Chr6: 26091185 (on Assembly GRCh37)
Protein change:	S65C
HGVS:	NG_008720.2:g.8677A>T NM_000410.3:c.193A>T NM_139007.2:c.77-357A>T NP_000401.1:p.Ser65Cys NC_000006.12:g.26090957A>T (GRCh38) LRG_748t1:c.193A>T NC_000006.11:g.26091185A>T (GRCh37) NG_008720.1:g.8677A>T Q30201:p.Ser65Cys LRG_748p1:p.Ser65Cys LRG_748:g.8677A>T ...less
Links:	UniProtKB: Q30201#VAR_004397 OMIM: 613609.0003 dbSNP: 1800730
NCBI 1000 Genomes Browser:	rs1800730
Molecular consequence:	NM_000410.3:c.193A>T: missense variant SO:0001583 NM_139007.2:c.77-357A>T: intron variant SO:0001627
Allele frequency:	GO-ESP 0.01107 (T) GMAF 0.00400 (T) ExAC 0.01009 (T)

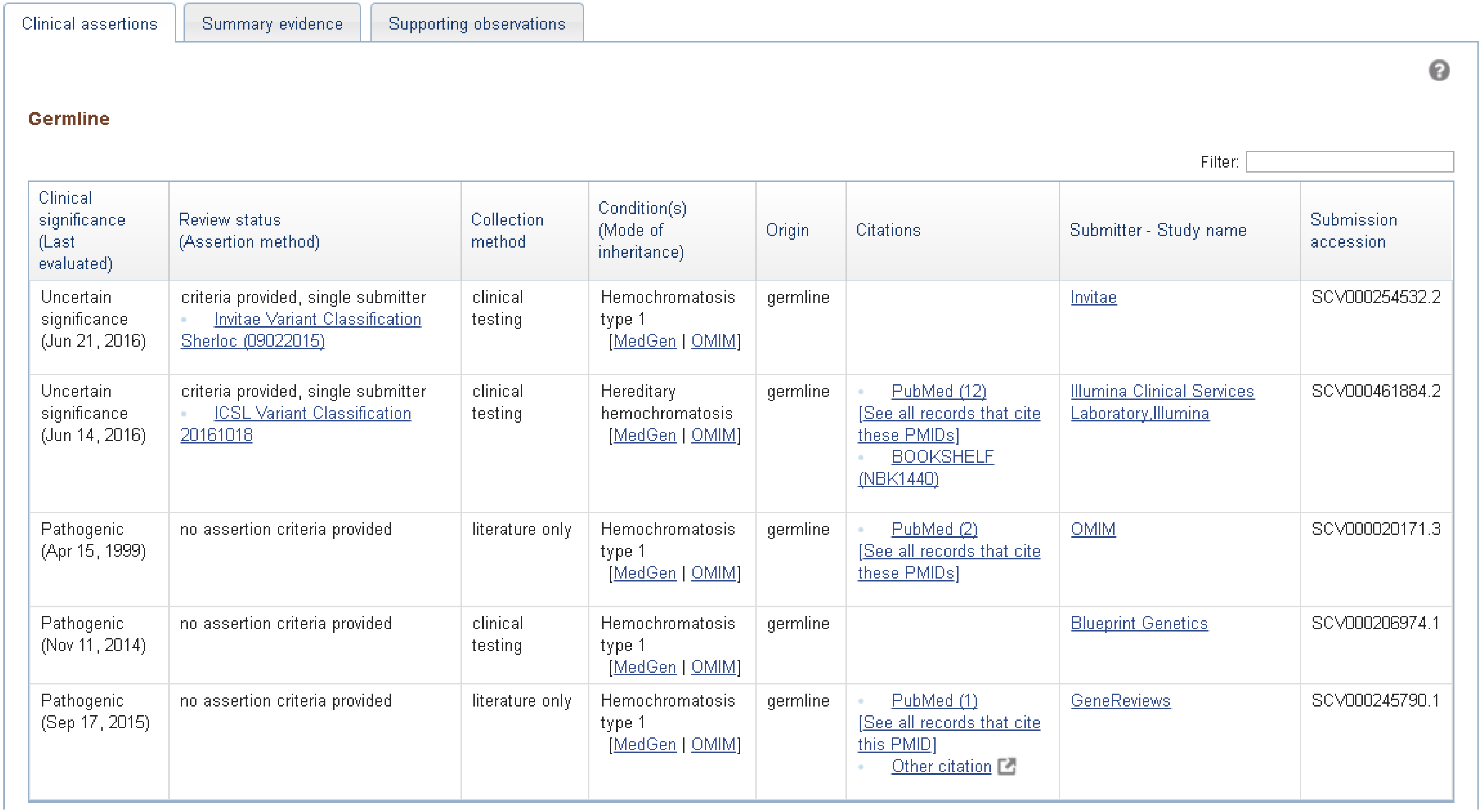
Browser views

- RefSeqGene
- Variation Viewer [GRCh38 - GRCh37]
- UCSC [GRCh38/hg38 - GRCh37/hg19]

Related information

- dbSNP
- Functional Class
- Gene
- GTR (all)
- MedGen
- OMIM
- PMC
- PubMed
- PubMed (calculated)
- Related genes (specific)

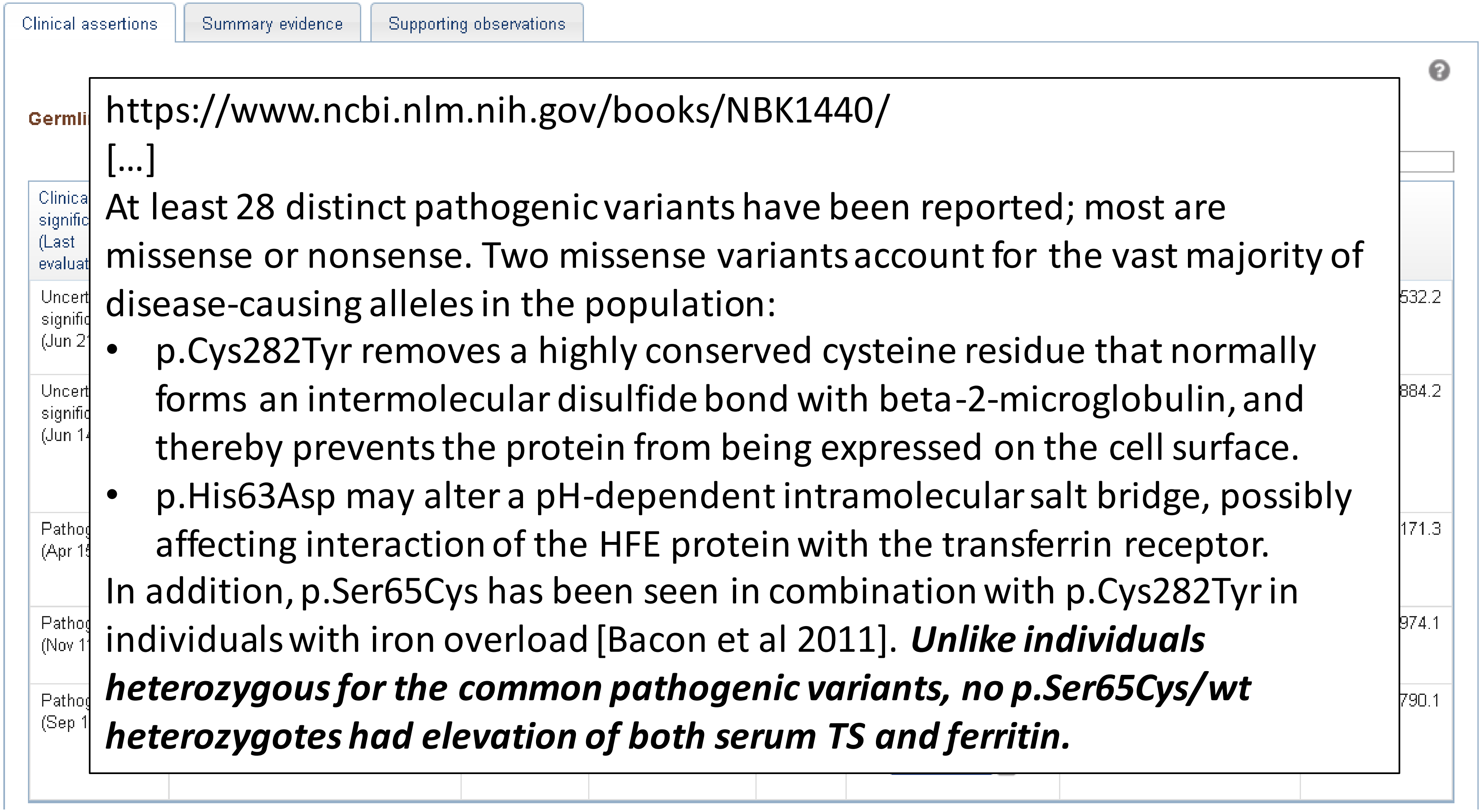
Clinical variant sources: ClinVar (2)



Germline

Clinical significance (Last evaluated)	Review status (Assertion method)	Collection method	Condition(s) (Mode of inheritance)	Origin	Citations	Submitter - Study name	Submission accession
Uncertain significance (Jun 21, 2016)	criteria provided, single submitter • Invitae Variant Classification Sherloc (09022015)	clinical testing	Hemochromatosis type 1 [MedGen | OMIM]	germline		Invitae	SCV000254532.2
Uncertain significance (Jun 14, 2016)	criteria provided, single submitter • ICSL Variant Classification 20161018	clinical testing	Hereditary hemochromatosis [MedGen | OMIM]	germline	• PubMed (12) [See all records that cite these PMIDs] • BOOKSHELF (NBK1440)	Illumina Clinical Services Laboratory,Illumina	SCV000461884.2
Pathogenic (Apr 15, 1999)	no assertion criteria provided	literature only	Hemochromatosis type 1 [MedGen | OMIM]	germline	• PubMed (2) [See all records that cite these PMIDs]	OMIM	SCV000020171.3
Pathogenic (Nov 11, 2014)	no assertion criteria provided	clinical testing	Hemochromatosis type 1 [MedGen | OMIM]	germline		Blueprint Genetics	SCV000206974.1
Pathogenic (Sep 17, 2015)	no assertion criteria provided	literature only	Hemochromatosis type 1 [MedGen | OMIM]	germline	• PubMed (1) [See all records that cite this PMID] • Other citation	GeneReviews	SCV000245790.1

Clinical variant sources: ClinVar (2)



https://www.ncbi.nlm.nih.gov/books/NBK1440/

[…]

At least 28 distinct pathogenic variants have been reported; most are missense or nonsense. Two missense variants account for the vast majority of disease-causing alleles in the population:

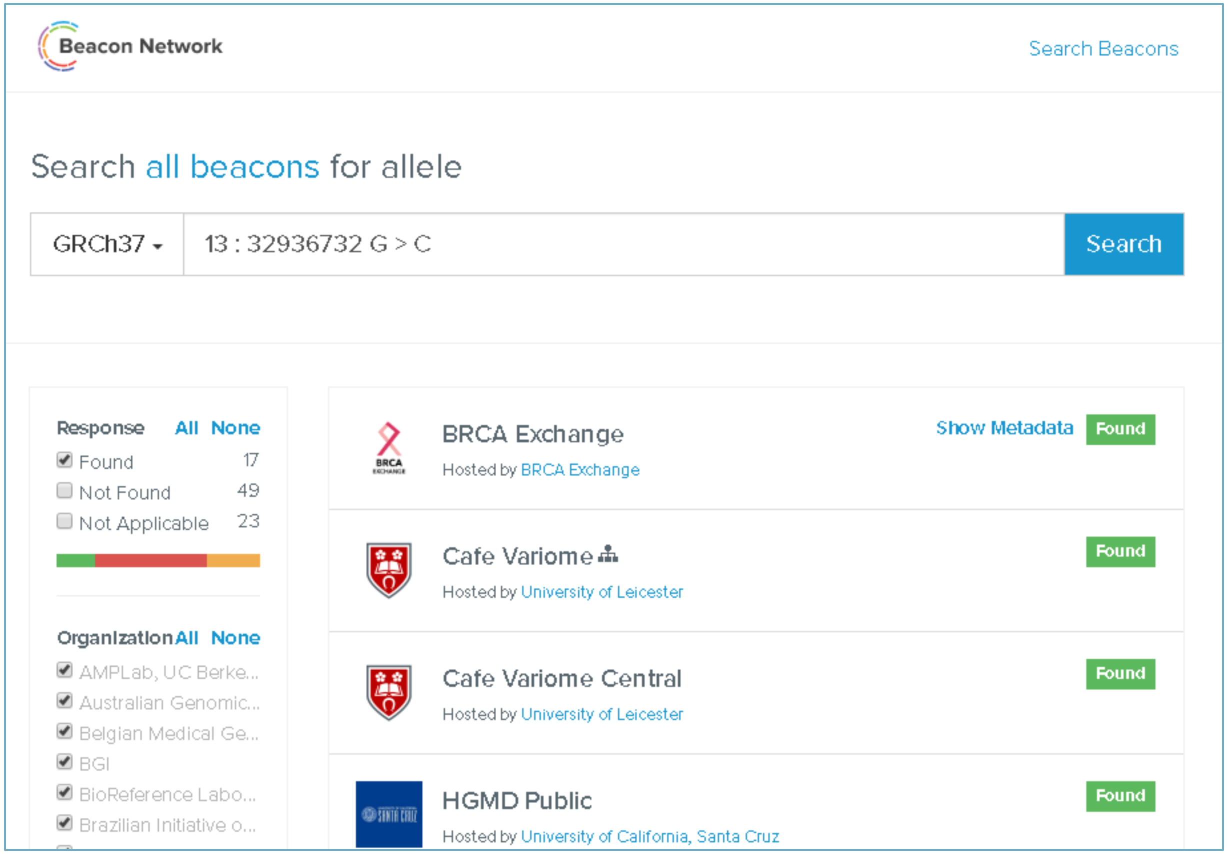
- p.Cys282Tyr removes a highly conserved cysteine residue that normally forms an intermolecular disulfide bond with beta-2-microglobulin, and thereby prevents the protein from being expressed on the cell surface.
- p.His63Asp may alter a pH-dependent intramolecular salt bridge, possibly affecting interaction of the HFE protein with the transferrin receptor.

In addition, p.Ser65Cys has been seen in combination with p.Cys282Tyr in individuals with iron overload [Bacon et al 2011]. ***Unlike individuals heterozygous for the common pathogenic variants, no p.Ser65Cys/wt heterozygotes had elevation of both serum TS and ferritin.***

Variant beacons: beacon-network.org

- Web services trying to balance desire of sharing genomic data with need for data protection – only one question:

  Does a specific variant exist in your database?



Search all beacons for allele

GRCh37 ▾ | 13 : 32936732 G > C

Response	All None
☑ Found	17
□ Not Found	49
□ Not Applicable	23

Organization All None

☑ AMPLab, UC Berke...
☑ Australian Genomic...
☑ Belgian Medical Ge...
☑ BGI
☑ BioReference Labo...
☑ Brazilian Initiative o...

Beacon	Hosted by		Response
BRCA Exchange	BRCA Exchange	Show Metadata	Found
Cafe Variome	University of Leicester		Found
Cafe Variome Central	University of Leicester		Found
HGMD Public	University of California, Santa Cruz		Found

https://beacon-network.org//#/search?pos=32936732&chrom=13&allele=C&ref=G&rs=GRCh37

VARIANT FILTERING

Mendelian disorders

- rare variant
- severe effect
- early onset / high penetrance

Search for known disease mutations

Caveats

- ClinVar is not comprehensive
- HGMD is expensive
- many wrong entries in both (revealed by ExAC etc.)
- do not include novel mutations
- **the phenotype should match yours!**

Exclude harmless polymorphisms

1000 Genomes Project:

- 2,500 genomes
- no severe Mendelian disorders

ExAC:

- 60,000 exomes
- no severe Mendelian disorders

gnomAD:

- 120,000 exomes + 15,000 genomes

1000 Genomes

Exome Aggregation Consortium (ExAC)

Genome Aggregation Database (gnomAD)

dbSNP
Short Genetic Variations

Exclude harmless polymorphisms

Caveats:

- dbSNP contains disease mutations

➔ **do not filter for dbSNP IDs!**

- ‘private’ or population-specific variants are not covered
- gnomAD is not limited to ‘healthy’ individuals

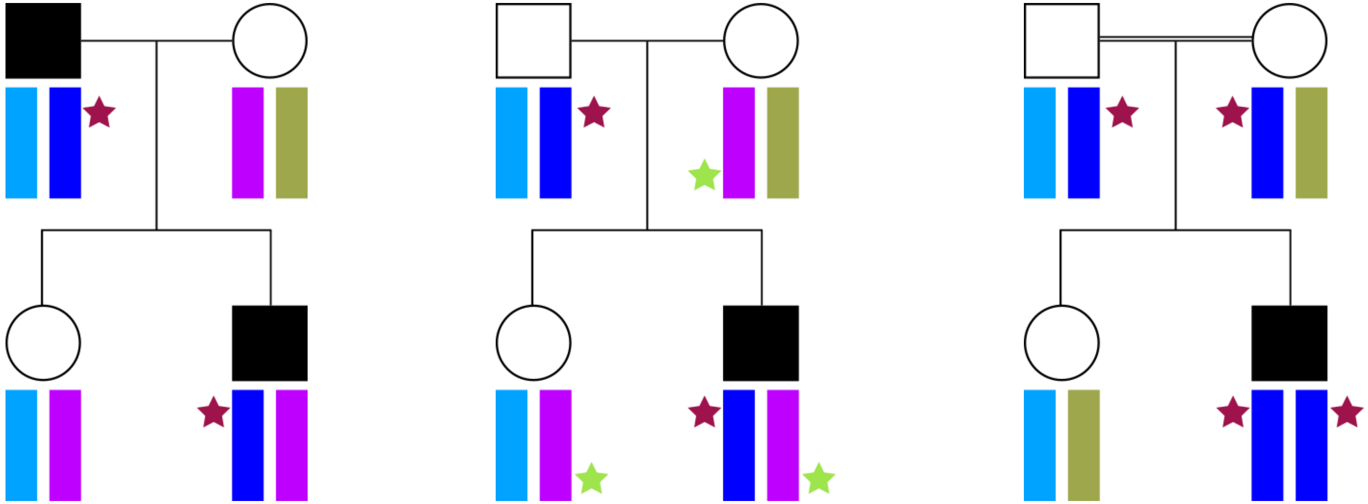
1000 Genomes

Exome Aggregation Consortium (ExAC)

Genome Aggregation Database (gnomAD)

dbSNP
Short Genetic Variations

Consider inheritance

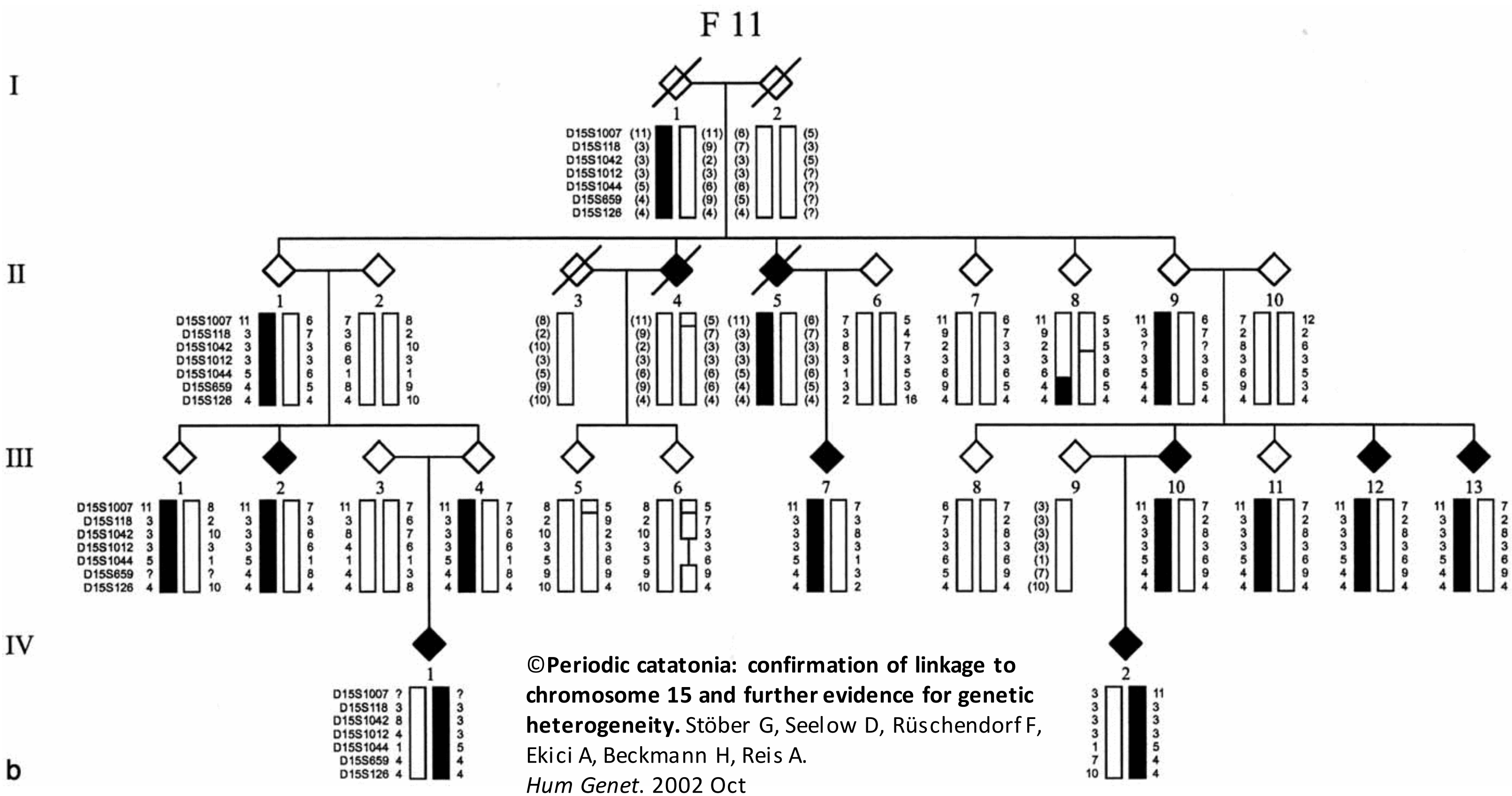


- fully penetrant ‘dominant alleles’ should not be present in healthy indivuals
- ‘common’ disease mutations occur in heterozygous state

➔ adapt filtering strategy to MOI

➔ **recessive disorders: filter for homozygosity, not for allele frequencies**

Consider incomplete penetrance



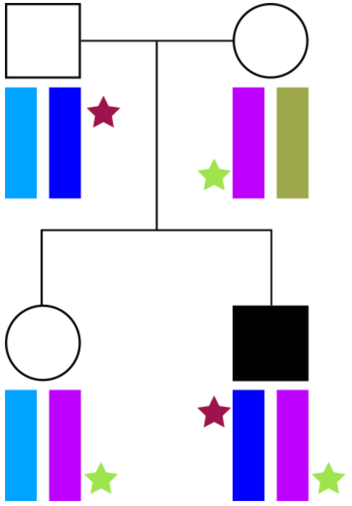
©Periodic catatonia: confirmation of linkage to chromosome 15 and further evidence for genetic heterogeneity. Stöber G, Seelow D, Rüschendorf F, Ekici A, Beckmann H, Reis A. *Hum Genet*. 2002 Oct

- allele carriers may be healthy

➔ **allow higher allele frequencies in healthy individuals**

Consider compound heterozygosity

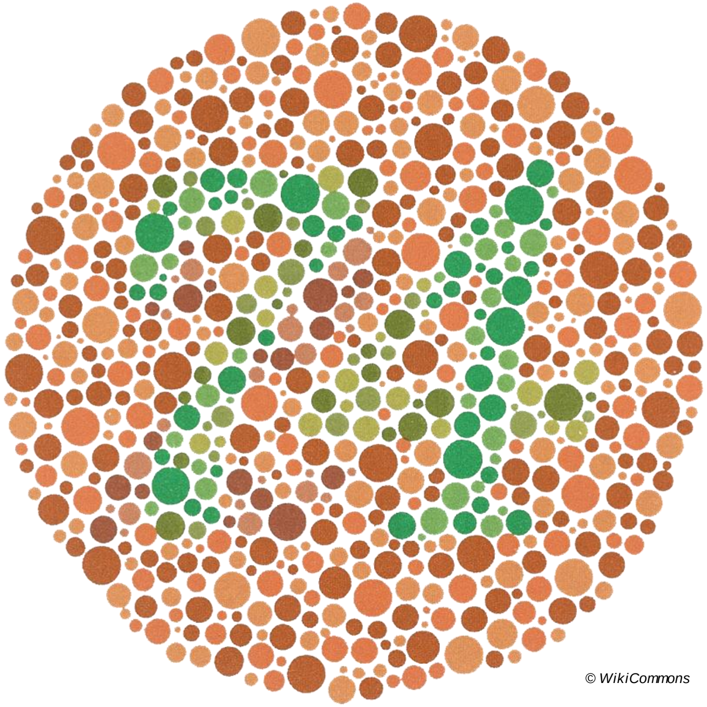
BERLIN INSTITUTE OF HEALTH
Charité & Max Delbrück Center



- patients do not have to be homozygous

➔ **do not exclude heterozygous genotypes**

Consider the disease frequency



© WikiCommons

- alleles causing ‘common’ recessive monogenic diseases are not rare

➔ **do not filter too strictly (or only for homozygosity)**

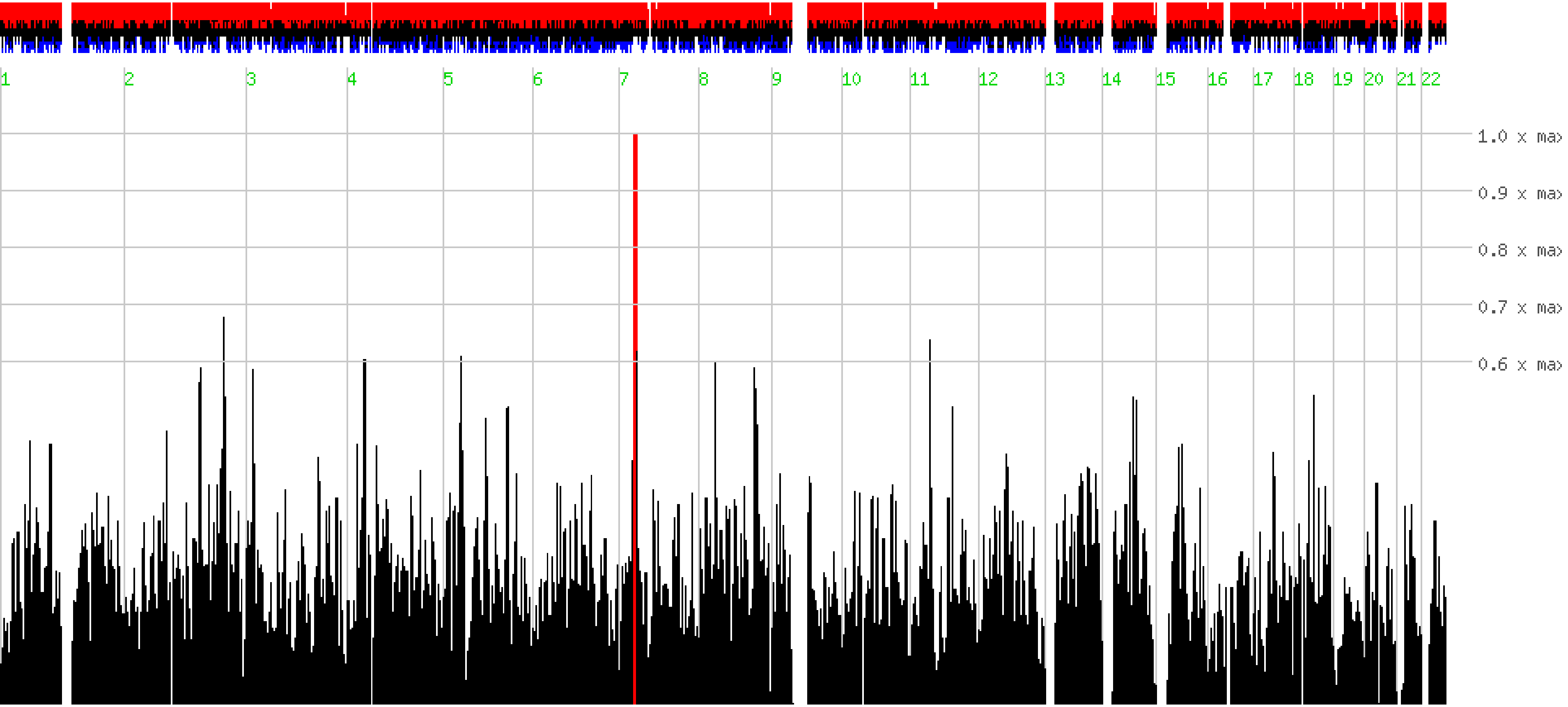
➔ **use different thresholds, depending on disease frequency**

Create an in-house database

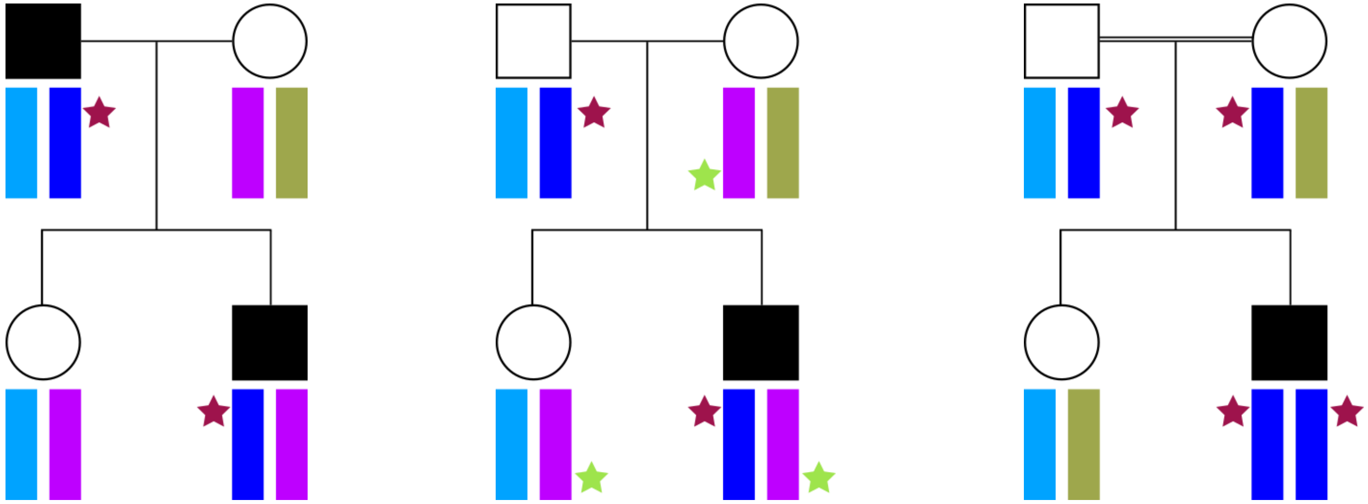
Filter against your own data!

- removes population- (or family-) specific variants
- reveals alignment artefacts

Limit to disease loci



Check for suitable genotypes



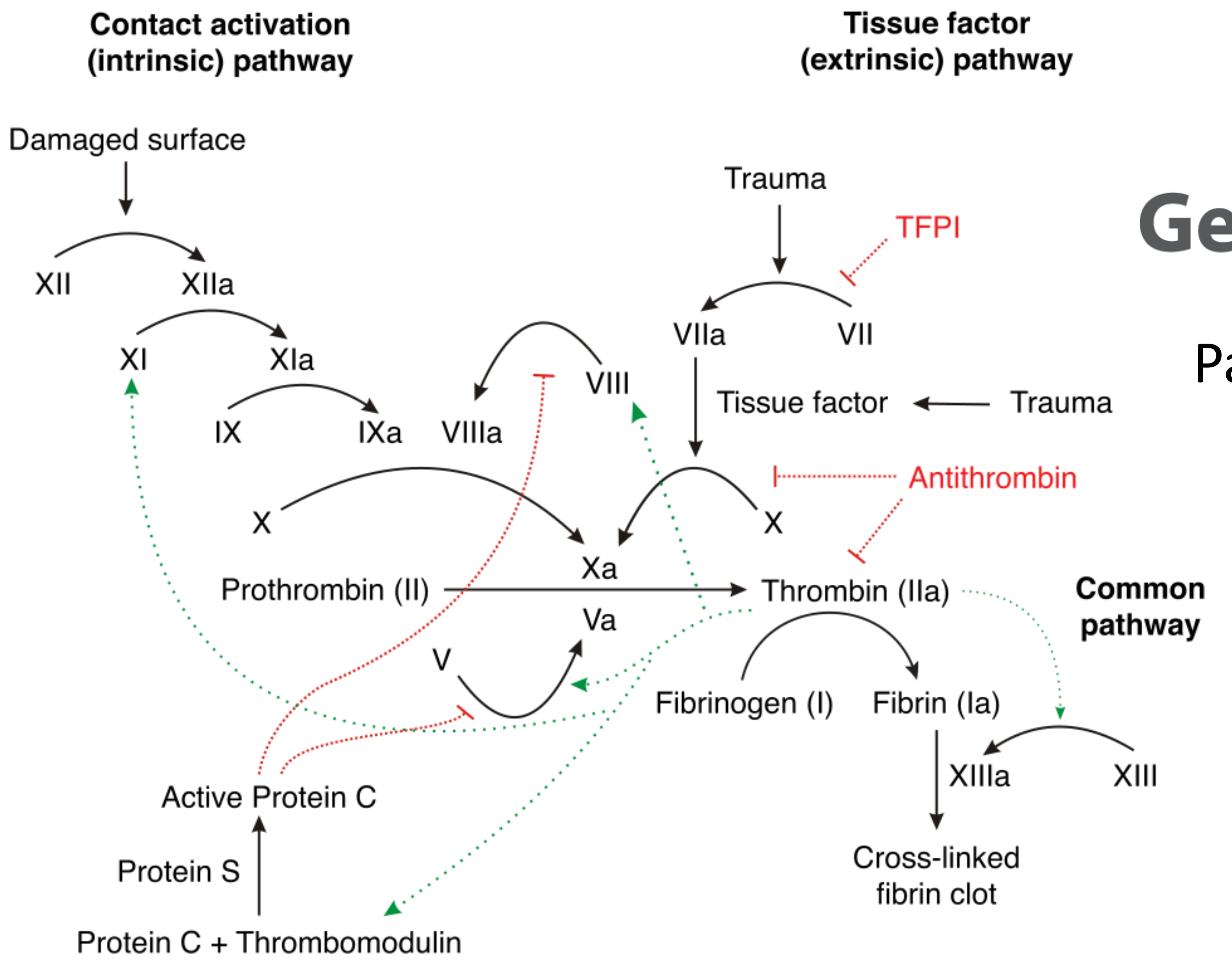
GEMINI

A FRAMEWORK FOR MINING GENOME VARIATION

VAAST

Limit to candidate genes / gene panels

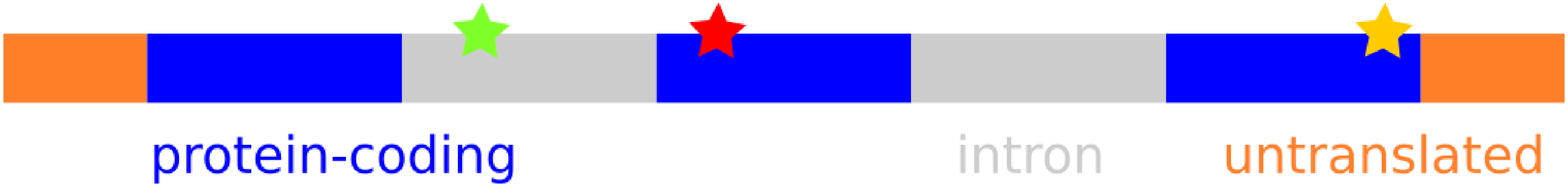
BERLIN INSTITUTE OF HEALTH
Charité & Max Delbrück Center

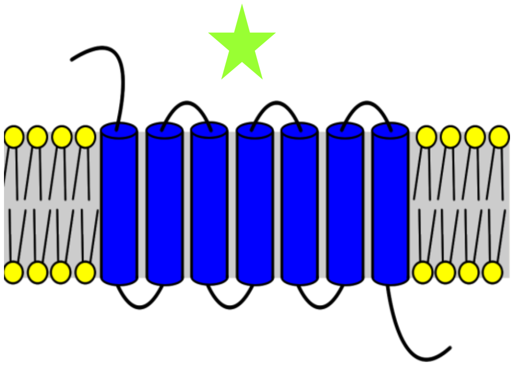


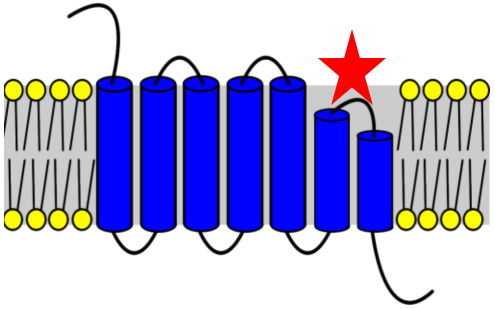
Genomics england PanelApp

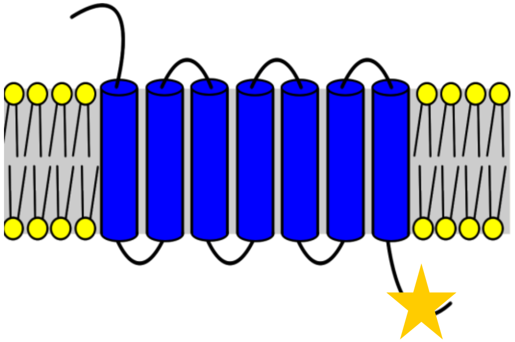
©Joe D (WikiCommons)

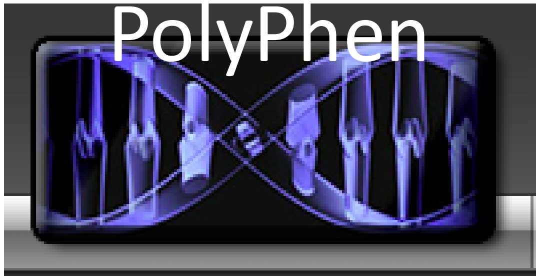
Limit to 'damaging' variants





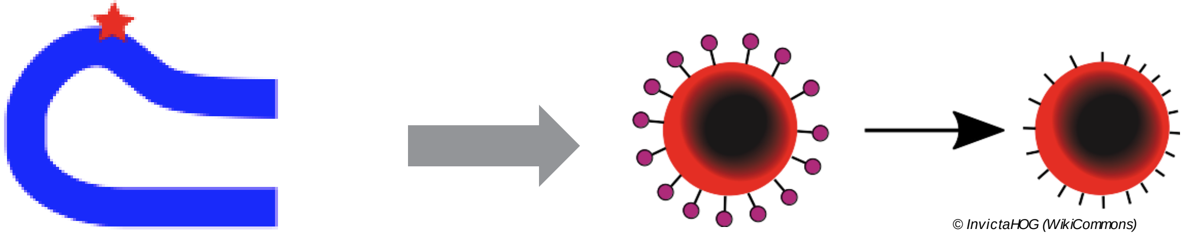






Consider the disease / phenotype

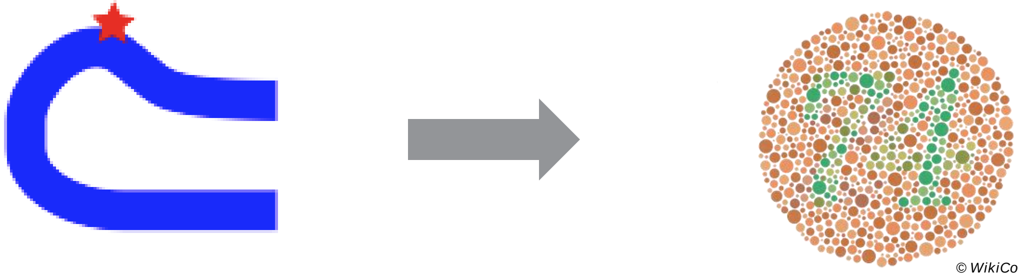
ABO gene



© InvictaHOG (WikiCommons)

may change blood type

OPN1LW gene

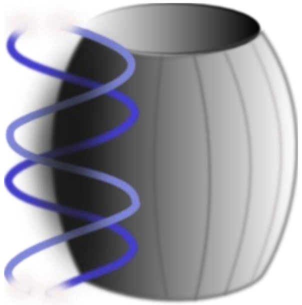


© WikiCommons

may cause colour blindness

Gene prioritisation tools

GeneCards®
HUMAN GENE DATABASE

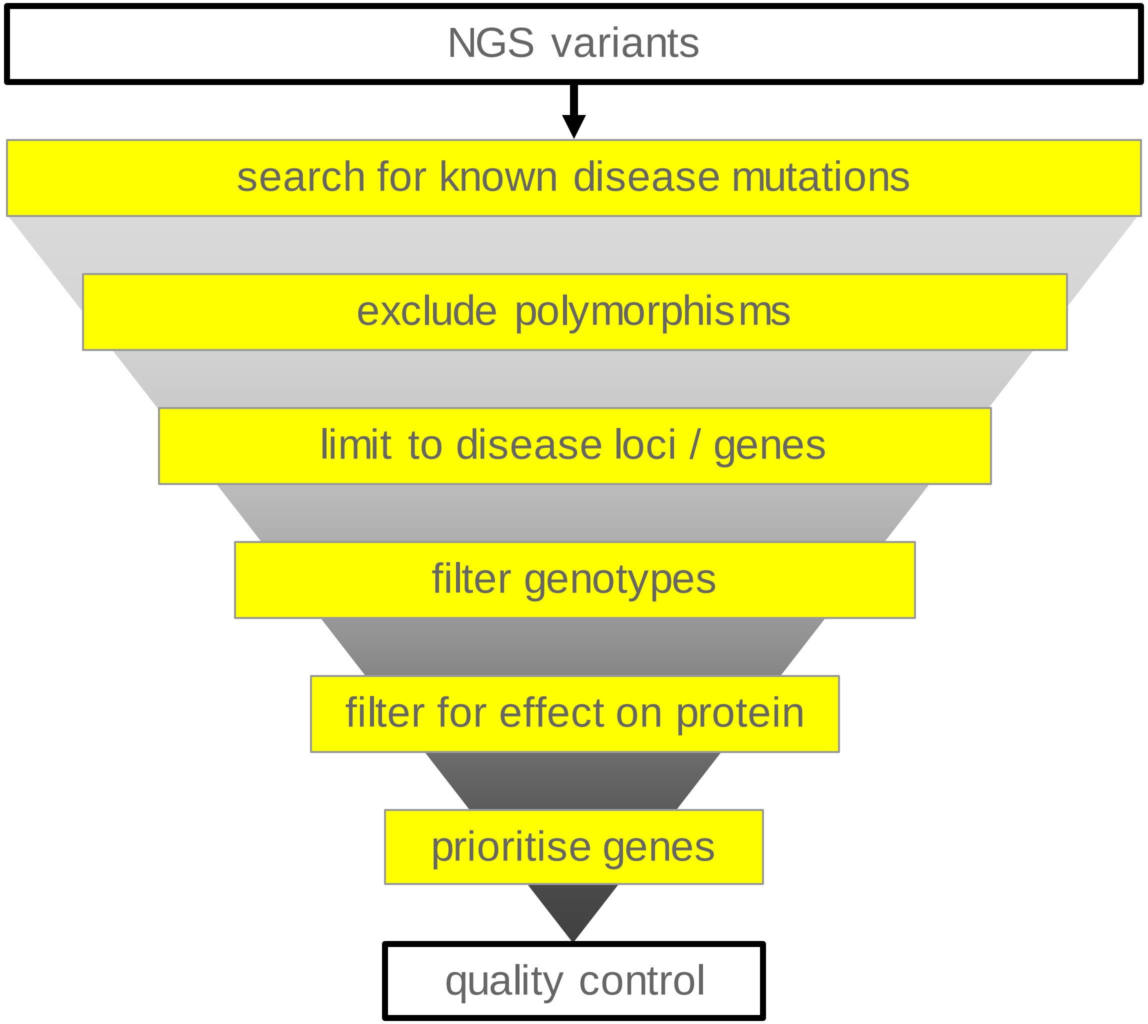

GeneDistiller

ENDEAVOUR

Phenomizer

GeneWalker

Variant filtering in a nutshell



NGS variants

search for known disease mutations

exclude polymorphisms

limit to disease loci / genes

filter genotypes

filter for effect on protein

prioritise genes

quality control

Quality control: inspect your variant

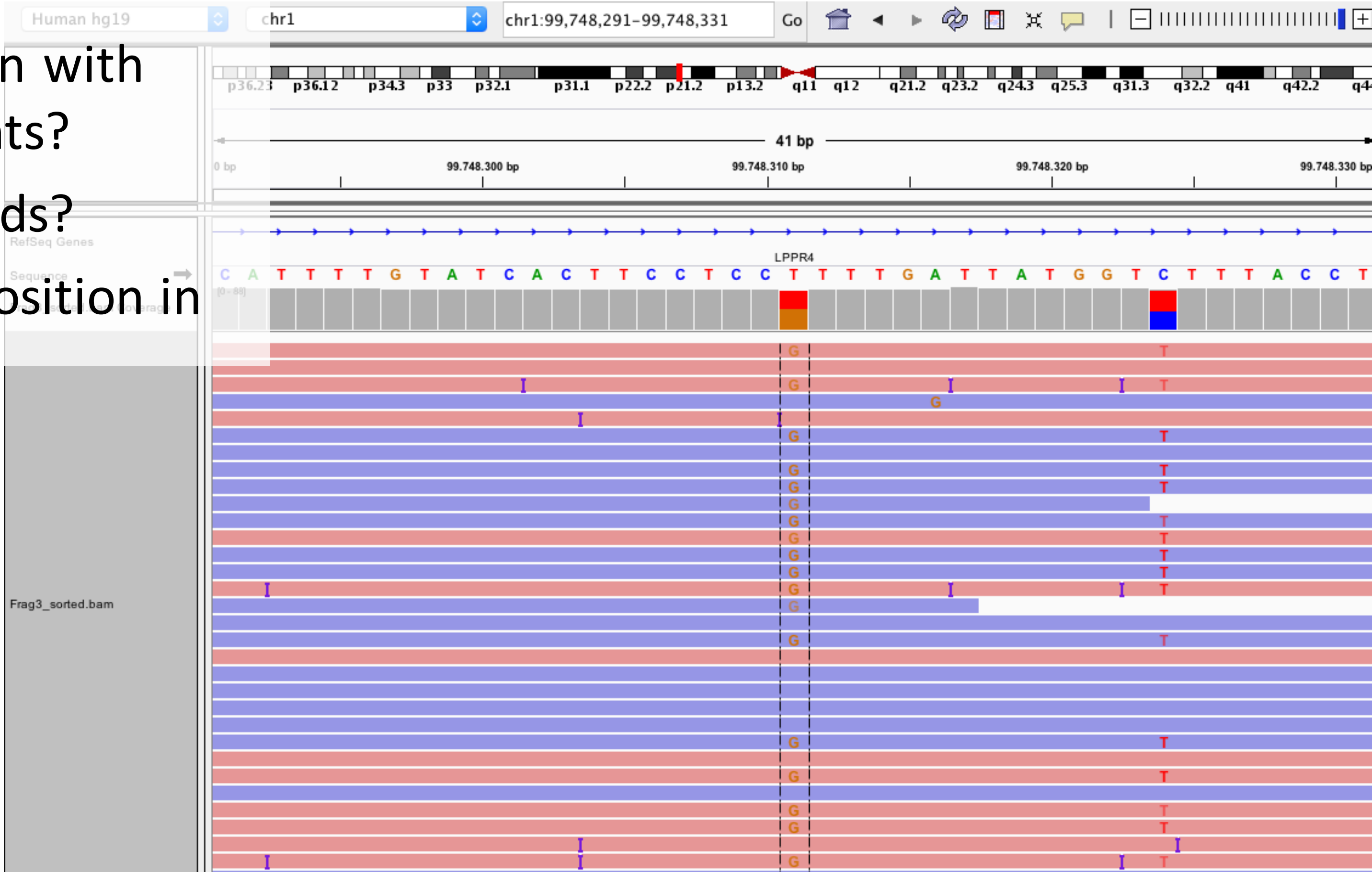
- is it sufficiently covered?
- is it (frequently) found in polymorphism databases?
- is it reported in ClinVar / HGMD?
- do you see it in the parents?
- do you see it in other samples?

Test for co-segregation

- reveals incompatibilities with the pedigree
- inevitable for suspected compound heterozygosity

Quality control: IGV (Broad Institute)

- sufficient coverage?
- variant on both strands?
- co-segregation with nearby variants?
- on both strands?
- at different position in the reads?



After the break

- Assessment of variants within protein-coding genes
- A use case for the identification of disease mutations
- Predicting the effect of non-coding variants

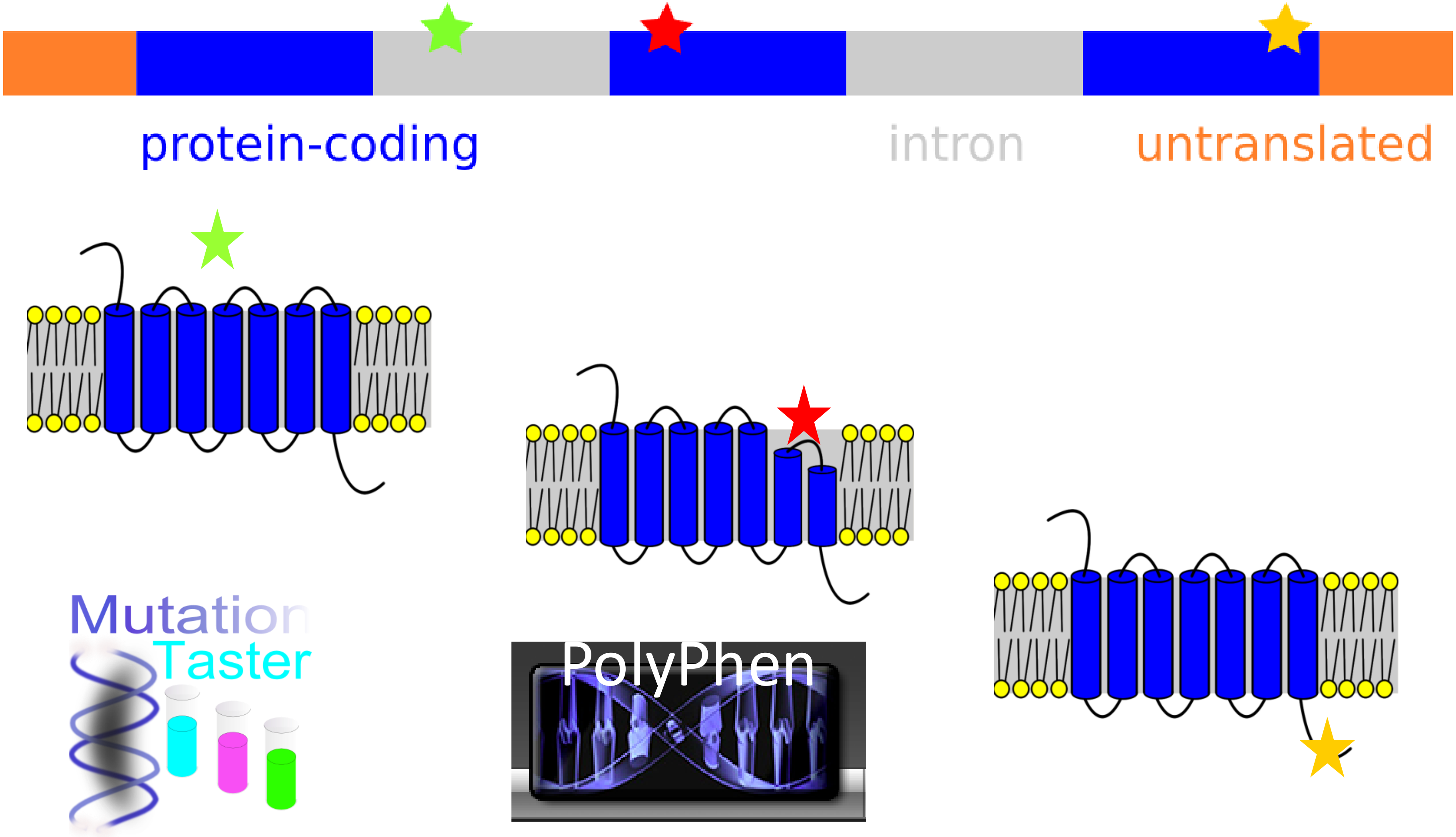
SHORT BREAK

VARIANT ASSESSMENT

Mendelian disorders

- rare variant
- severe effect
- early onset / high penetrance

Consequences of variants



Non-coding variants within genes

splice site

- loss/gain of exons -> affects the CDS
- frameshift -> affects the CDS
- transcript lost/misregulated

promoter / TSS

- gene/transcript lost/misregulated

UTRs

- polyadq signal lost
- miRNA binding sites changed
- transcript misregulated

Coding variants within genes

may affect

- splicing
- functional domains
- structure
- activity

can cause

- premature termination codon -> NMD
- frameshift
- loss/gain/substitution of amino acids

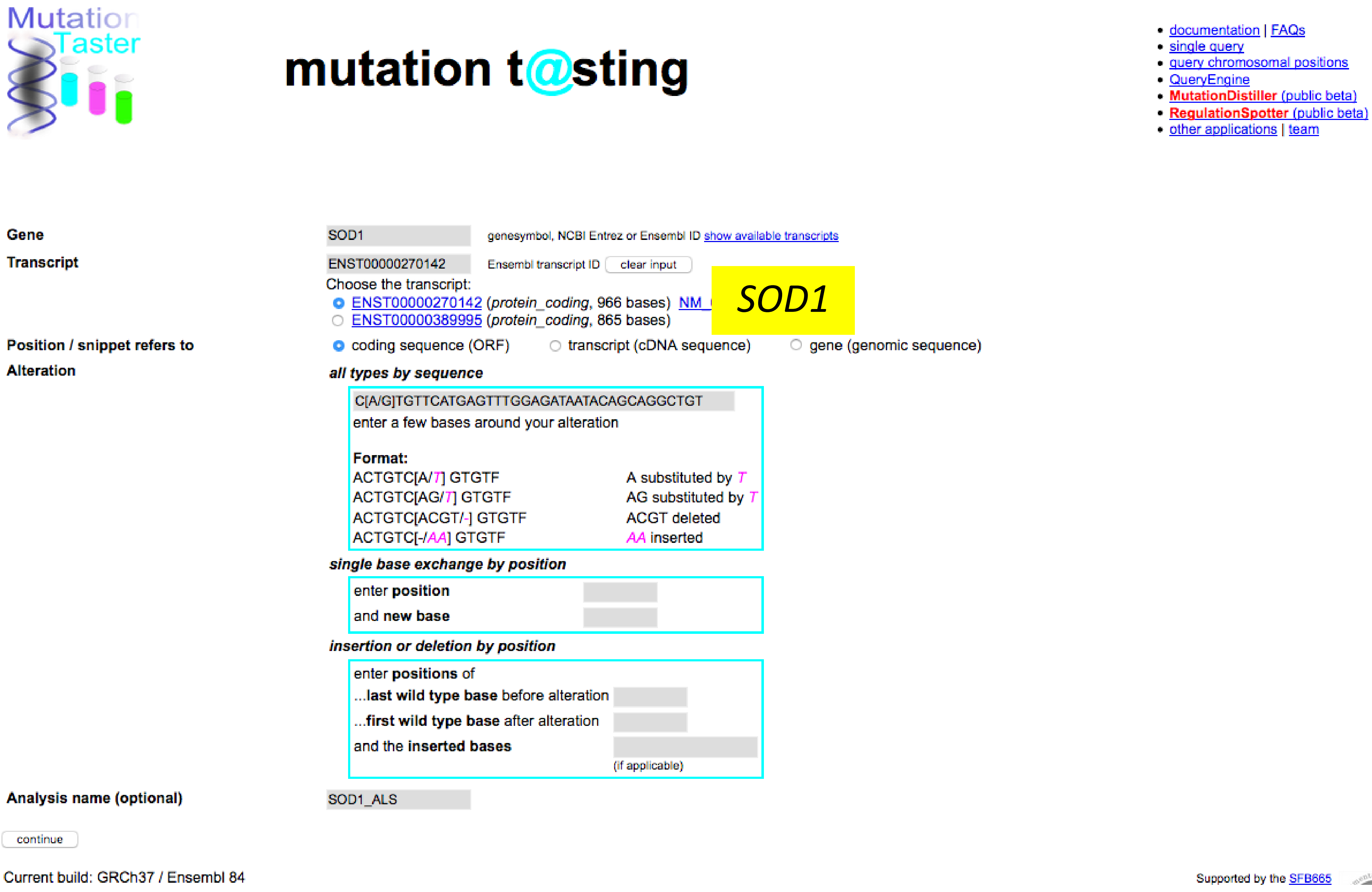
not limited to *missense/nonsense* variants

Predicting the disease-causing effect of DNA variants with MutationTaster

BERLIN INSTITUTE OF HEALTH
Charité & Max Delbrück Center

Jana Marie Schwarz

MutationTaster evaluates disease-causing potential of sequence alterations.
Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. *Nat Methods*. 2010
MutationTaster2: mutation prediction for the deep-sequencing age.
Schwarz JM, Cooper DN, Schuelke M, Seelow D. *Nat Methods*. 2014



mutation t@sting

- documentation | FAQs
- single query
- query chromosomal positions
- QueryEngine
- MutationDistiller (public beta)
- RegulationSpotter (public beta)
- other applications | team

Gene	SOD1 genesymbol, NCBI Entrez or Ensembl ID show available transcripts
Transcript	ENST00000270142 Ensembl transcript ID clear input Choose the transcript: ● ENST00000270142 (*protein_coding*, 966 bases) NM_ ○ ENST00000389995 (*protein_coding*, 865 bases)
Position / snippet refers to	● coding sequence (ORF) ○ transcript (cDNA sequence) ○ gene (genomic sequence)
Alteration	***all types by sequence***

C[A/G]TGTTCATGAGTTTGGAGATAATACAGCAGGCTGT

enter a few bases around your alteration

Format:

ACTGTC[A/*T*] GTGTF	A substituted by *T*
ACTGTC[AG/*T*] GTGTF	AG substituted by *T*
ACTGTC[ACGT/-] GTGTF	ACGT deleted
ACTGTC[-/*AA*] GTGTF	*AA* inserted

single base exchange by position

enter **position**

and new **base**

insertion or deletion by position

enter **positions** of

...**last wild type base** before alteration

...**first wild type base** after alteration

and the **inserted bases**

(if applicable)

Analysis name (optional) SOD1_ALS

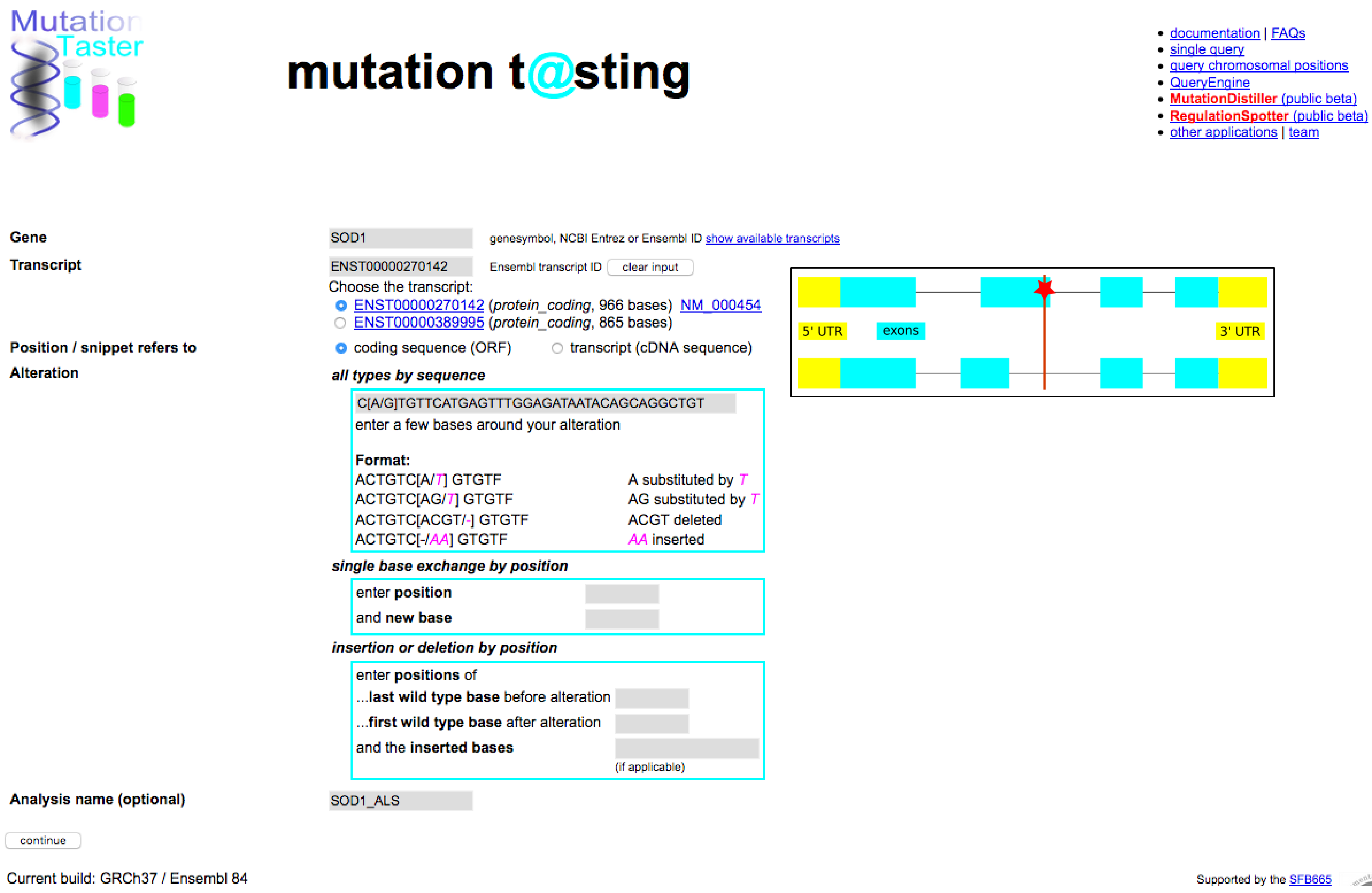
continue

Current build: GRCh37 / Ensembl 84

Supported by the SFB665

SOD1

http://www.mutationtaster.org/



mutation t@sting

- documentation | FAQs
- single query
- query chromosomal positions
- QueryEngine
- MutationDistiller (public beta)
- RegulationSpotter (public beta)
- other applications | team

Gene SOD1 genesymbol, NCBI Entrez or Ensembl ID show available transcripts

Transcript ENST00000270142 Ensembl transcript ID clear input

Choose the transcript:

● ENST00000270142 (*protein_coding*, 966 bases) NM_000454
○ ENST00000389995 (*protein_coding*, 865 bases)

Position / snippet refers to ● coding sequence (ORF) ○ transcript (cDNA sequence)

Alteration

all types by sequence

C[A/G]TGTTCATGAGTTTGGAGATAATACAGCAGGCTGT

enter a few bases around your alteration

Format:

ACTGTC[A/*T*] GTGTF	A substituted by *T*
ACTGTC[AG/*T*] GTGTF	AG substituted by *T*
ACTGTC[ACGT/-] GTGTF	ACGT deleted
ACTGTC[-/*AA*] GTGTF	*AA* inserted

single base exchange by position

enter **position**

and **new base**

insertion or deletion by position

enter **positions** of

...**last wild type base** before alteration

...**first wild type base** after alteration

and the **inserted bases**

(if applicable)

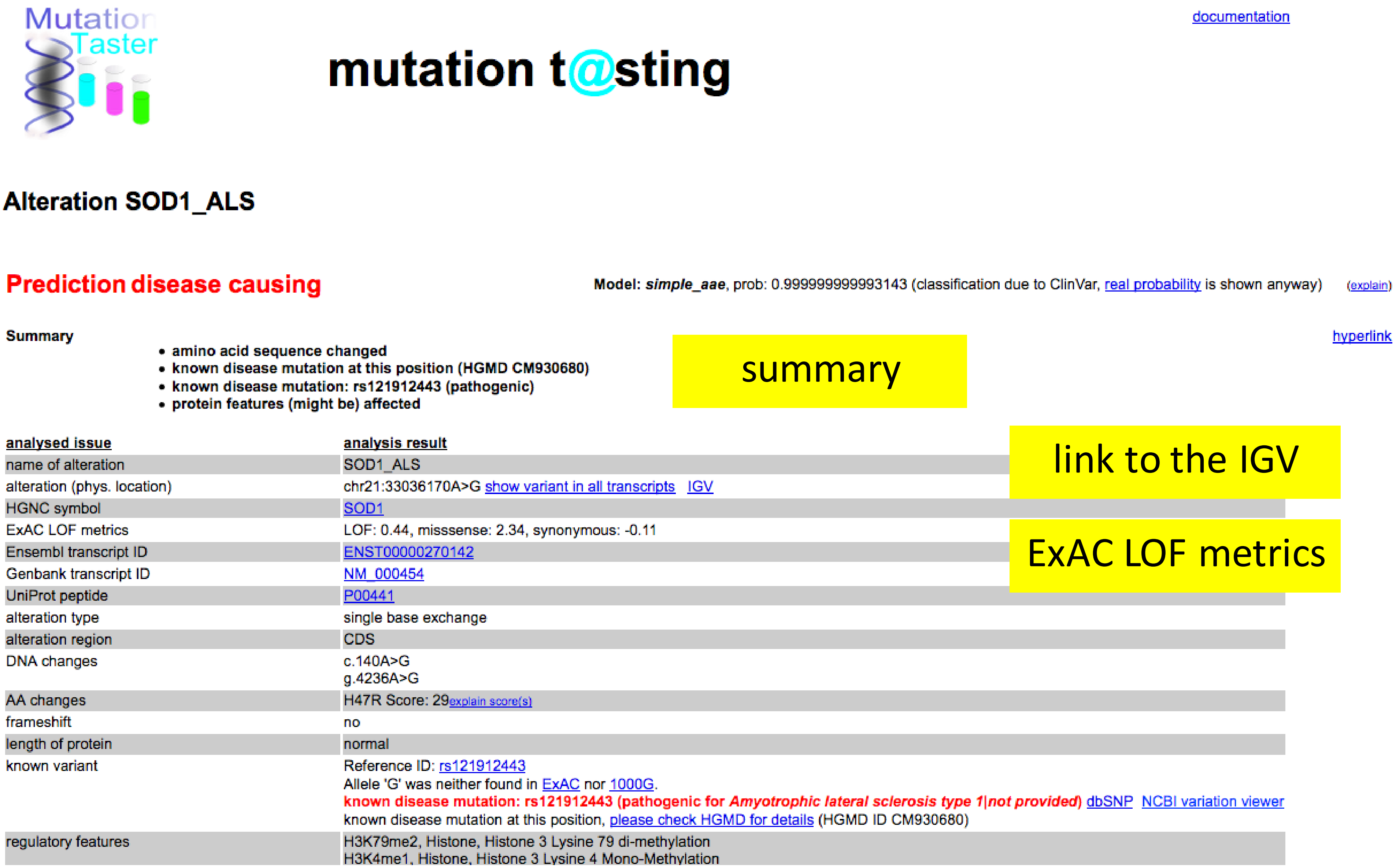
Analysis name (optional) SOD1_ALS

continue

Current build: GRCh37 / Ensembl 84

Supported by the SFB665

http://www.mutationtaster.org/



documentation

mutation t@sting

Alteration SOD1_ALS

Prediction disease causing

Model: ***simple_aae***, prob: 0.999999999993143 (classification due to ClinVar, real probability is shown anyway) (explain)

Summary

hyperlink

- **amino acid sequence changed**
- **known disease mutation at this position (HGMD CM930680)**
- **known disease mutation: rs121912443 (pathogenic)**
- **protein features (might be) affected**

summary

analysed issue	**analysis result**
name of alteration	SOD1_ALS
alteration (phys. location)	chr21:33036170A>G show variant in all transcripts IGV
HGNC symbol	SOD1
ExAC LOF metrics	LOF: 0.44, misssense: 2.34, synonymous: -0.11
Ensembl transcript ID	ENST00000270142
Genbank transcript ID	NM_000454
UniProt peptide	P00441
alteration type	single base exchange
alteration region	CDS
DNA changes	c.140A>G g.4236A>G
AA changes	H47R Score: 29explain score(s)
frameshift	no
length of protein	normal
known variant	Reference ID: rs121912443 Allele 'G' was neither found in ExAC nor 1000G. **known disease mutation: rs121912443 (pathogenic for *Amyotrophic lateral sclerosis type 1|not provided*)** dbSNP NCBI variation viewer known disease mutation at this position, please check HGMD for details (HGMD ID CM930680)
regulatory features	H3K79me2, Histone, Histone 3 Lysine 79 di-methylation H3K4me1, Histone, Histone 3 Lysine 4 Mono-Methylation

link to the IGV

ExAC LOF metrics

Essentiality or intolerance scores for genes

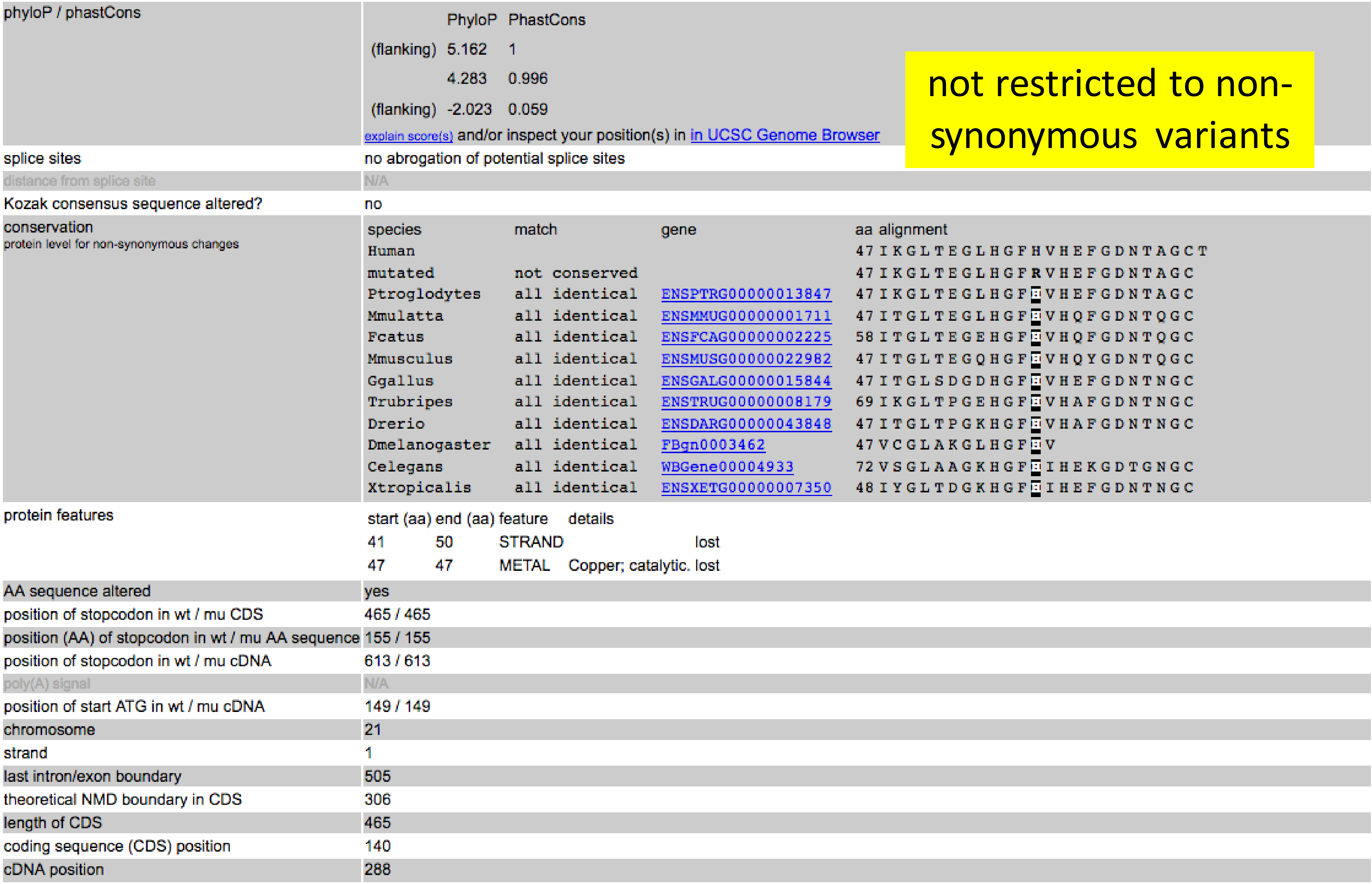
ExAC LoF / pLI

- intolerance to loss-of-function variants
- negative: gene seems to be tolerant to mutations
- positive: mutations more likely to cause disease

    (ALS1 can be AR **and** AD!)

subRVIS (Residual Variation Intolerance Score)

- including protein domains



not restricted to non-synonymous variants

phyloP / phastCons		PhyloP	PhastCons
	(flanking)	5.162	1
		4.283	0.996
	(flanking)	-2.023	0.059

explain score(s) and/or inspect your position(s) in in UCSC Genome Browser

splice sites	no abrogation of potential splice sites
distance from splice site	N/A
Kozak consensus sequence altered?	no

conservation
protein level for non-synonymous changes

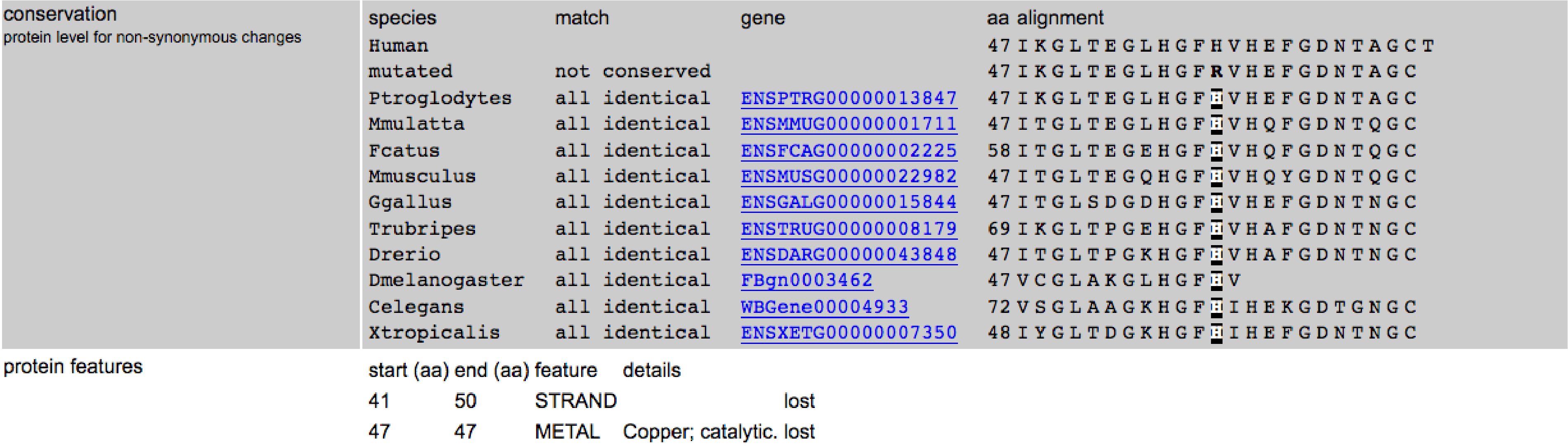
species	match	gene	aa alignment
Human			47 IKGLTEGLHGFHVHEFGDNTAGCT
mutated	not conserved		47 IKGLTEGLHGFRVHEFGDNTAGC
Ptroglodytes	all identical	ENSPTRG00000013847	47 IKGLTEGLHGFHVHEFGDNTAGC
Mmulatta	all identical	ENSMMUG00000001711	47 ITGLTEGLHGFHVHQFGDNTQGC
Fcatus	all identical	ENSFCAG00000002225	58 ITGLTEGEHGFHVHQFGDNTQGC
Mmusculus	all identical	ENSMUSG00000022982	47 ITGLTEGQHGFHVHQYGDNTQGC
Ggallus	all identical	ENSGALG00000015844	47 ITGLSDGDHGFHVHEFGDNTNGC
Trubripes	all identical	ENSTRUG00000008179	69 IKGLTPGEHGFHVHAFGDNTNGC
Drerio	all identical	ENSDARG00000043848	47 ITGLTPGKHGFHVHAFGDNTNGC
Dmelanogaster	all identical	FBgn0003462	47 VCGLAKGLHGFHV
Celegans	all identical	WBGene00004933	72 VSGLAAGKHGFHIHEKGDTGNGC
Xtropicalis	all identical	ENSXETG00000007350	48 IYGLTDGKHGFHIHEFGDNTNGC

protein features

start (aa)	end (aa)	feature	details	
41	50	STRAND		lost
47	47	METAL	Copper; catalytic.	lost

AA sequence altered	yes
position of stopcodon in wt / mu CDS	465 / 465
position (AA) of stopcodon in wt / mu AA sequence	155 / 155
position of stopcodon in wt / mu cDNA	613 / 613
poly(A) signal	N/A
position of start ATG in wt / mu cDNA	149 / 149
chromosome	21
strand	1
last intron/exon boundary	505
theoretical NMD boundary in CDS	306
length of CDS	465
coding sequence (CDS) position	140
cDNA position	288

Protein domains & conservation



conservation

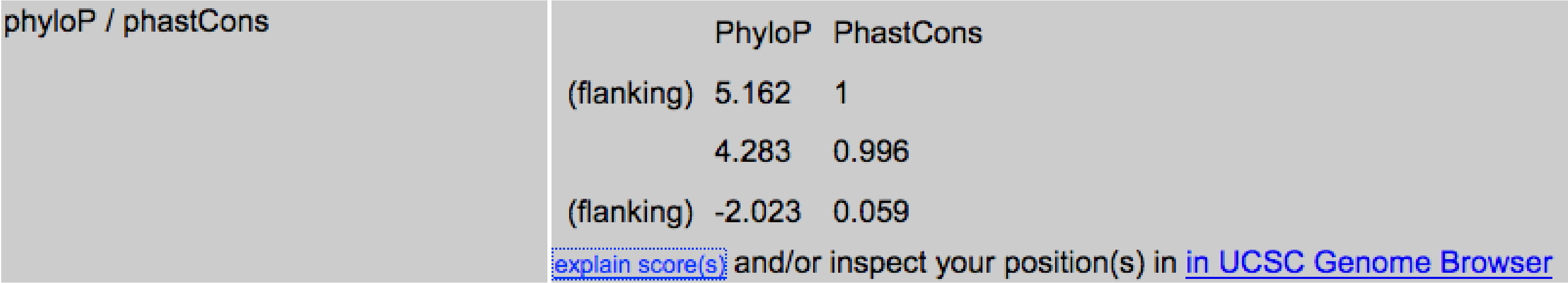
protein level for non-synonymous changes

species	match	gene	aa alignment
Human			47 I K G L T E G L H G F H V H E F G D N T A G C T
mutated	not conserved		47 I K G L T E G L H G F **R** V H E F G D N T A G C
Ptroglodytes	all identical	ENSPTRG00000013847	47 I K G L T E G L H G F H V H E F G D N T A G C
Mmulatta	all identical	ENSMMUG00000001711	47 I T G L T E G L H G F H V H Q F G D N T Q G C
Fcatus	all identical	ENSFCAG00000002225	58 I T G L T E G E H G F H V H Q F G D N T Q G C
Mmusculus	all identical	ENSMUSG00000022982	47 I T G L T E G Q H G F H V H Q Y G D N T Q G C
Ggallus	all identical	ENSGALG00000015844	47 I T G L S D G D H G F H V H E F G D N T N G C
Trubripes	all identical	ENSTRUG00000008179	69 I K G L T P G E H G F H V H A F G D N T N G C
Drerio	all identical	ENSDARG00000043848	47 I T G L T P G K H G F H V H A F G D N T N G C
Dmelanogaster	all identical	FBgn0003462	47 V C G L A K G L H G F H V
Celegans	all identical	WBGene00004933	72 V S G L A A G K H G F H I H E K G D T G N G C
Xtropicalis	all identical	ENSXETG00000007350	48 I Y G L T D G K H G F H I H E F G D N T N G C

protein features

start (aa)	end (aa)	feature	details	
41	50	STRAND		lost
47	47	METAL	Copper; catalytic.	lost

Phylogenetic conservation



phyloP / phastCons

	PhyloP	PhastCons
(flanking)	5.162	1
	4.283	0.996
(flanking)	-2.023	0.059

explain score(s) and/or inspect your position(s) in in UCSC Genome Browser

GERP (genomic evolutionary rate profiling)

- conservation of bases in different species

PhastCons

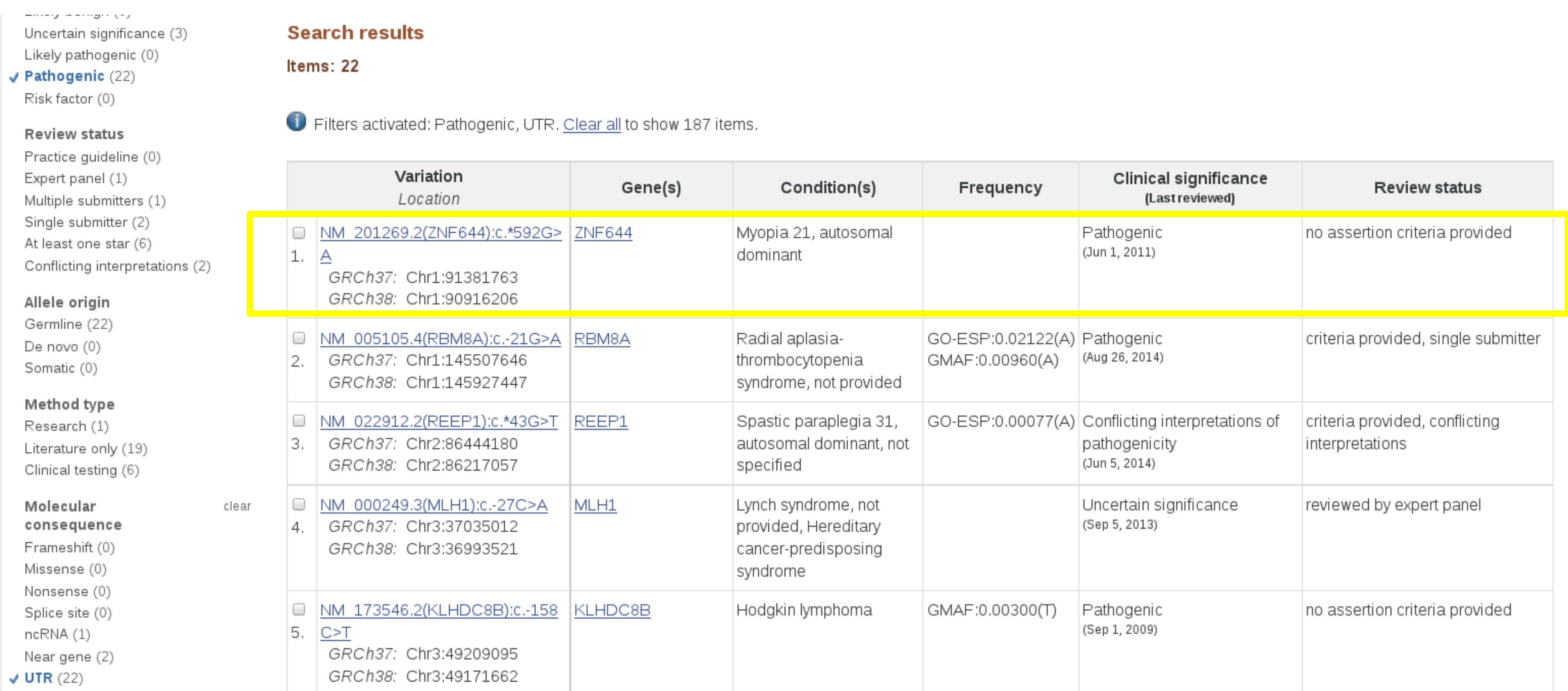
- multibase elements

phyloP

- ‘detection of lineage-specific conservation or acceleration’

(more in the non-coding part)

A non-coding example from ClinVar



- Uncertain significance (3)
- Likely pathogenic (0)
- ✓ **Pathogenic** (22)
- Risk factor (0)

Review status

- Practice guideline (0)
- Expert panel (1)
- Multiple submitters (1)
- Single submitter (2)
- At least one star (6)
- Conflicting interpretations (2)

Allele origin

- Germline (22)
- De novo (0)
- Somatic (0)

Method type

- Research (1)
- Literature only (19)
- Clinical testing (6)

Molecular consequence clear

- Frameshift (0)
- Missense (0)
- Nonsense (0)
- Splice site (0)
- ncRNA (1)
- Near gene (2)
- ✓ **UTR** (22)

Search results

Items: 22

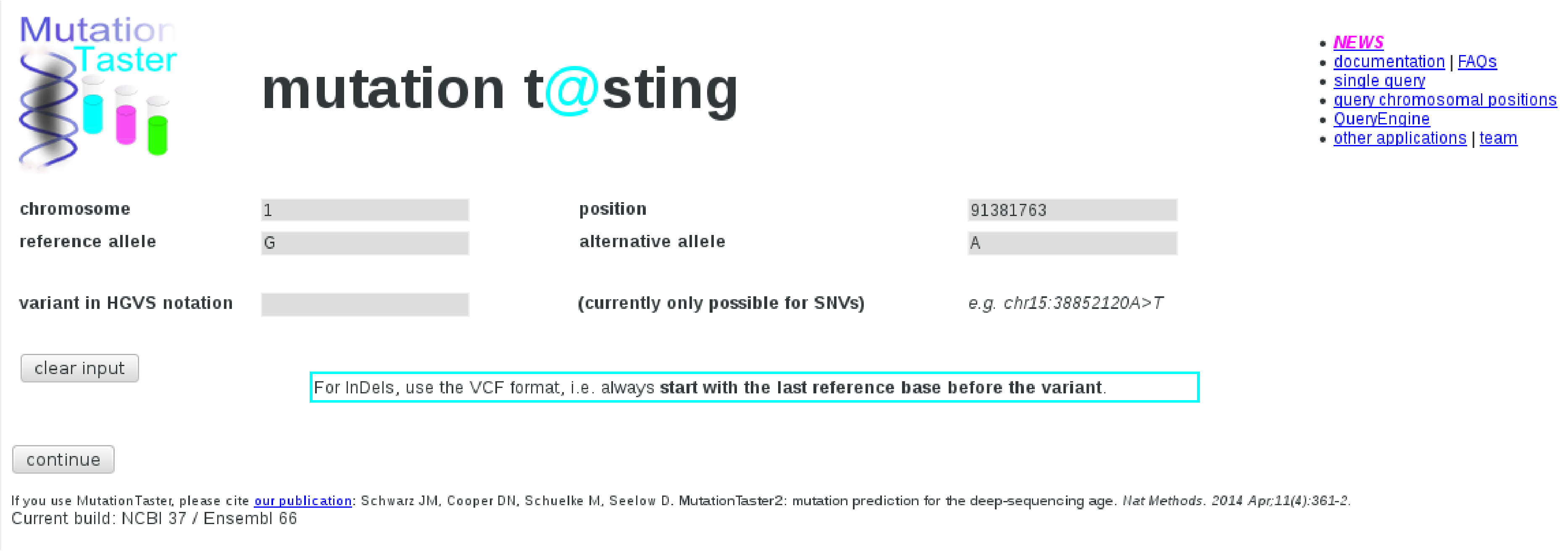
Filters activated: Pathogenic, UTR. Clear all to show 187 items.

	Variation *Location*	Gene(s)	Condition(s)	Frequency	Clinical significance (Last reviewed)	Review status
☐ 1.	NM_201269.2(ZNF644):c.*592G>A *GRCh37:* Chr1:91381763 *GRCh38:* Chr1:90916206	ZNF644	Myopia 21, autosomal dominant		Pathogenic (Jun 1, 2011)	no assertion criteria provided

☐ 2.	NM_005105.4(RBM8A):c.-21G>A
	GRCh37: Chr1:145507646
	GRCh38: Chr1:145927447
		RBM8A	Radial aplasia-thrombocytopenia syndrome, not provided	GO-ESP:0.02122(A)
				GMAF:0.00960(A)
					Pathogenic
					(Aug 26, 2014)
						criteria provided, single submitter

☐ 3.	NM_022912.2(REEP1):c.*43G>T *GRCh37:* Chr2:86444180 *GRCh38:* Chr2:86217057	REEP1	Spastic paraplegia 31, autosomal dominant, not specified	GO-ESP:0.00077(A)	Conflicting interpretations of pathogenicity (Jun 5, 2014)	criteria provided, conflicting interpretations
☐ 4.	NM_000249.3(MLH1):c.-27C>A *GRCh37:* Chr3:37035012 *GRCh38:* Chr3:36993521	MLH1	Lynch syndrome, not provided, Hereditary cancer-predisposing syndrome		Uncertain significance (Sep 5, 2013)	reviewed by expert panel
☐ 5.	NM_173546.2(KLHDC8B):c.-158C>T *GRCh37:* Chr3:49209095 *GRCh38:* Chr3:49171662	KLHDC8B	Hodgkin lymphoma	GMAF:0.00300(T)	Pathogenic (Sep 1, 2009)	no assertion criteria provided

How does it taste?



mutation t@sting

- *NEWS*
- documentation | FAQs
- single query
- query chromosomal positions
- QueryEngine
- other applications | team

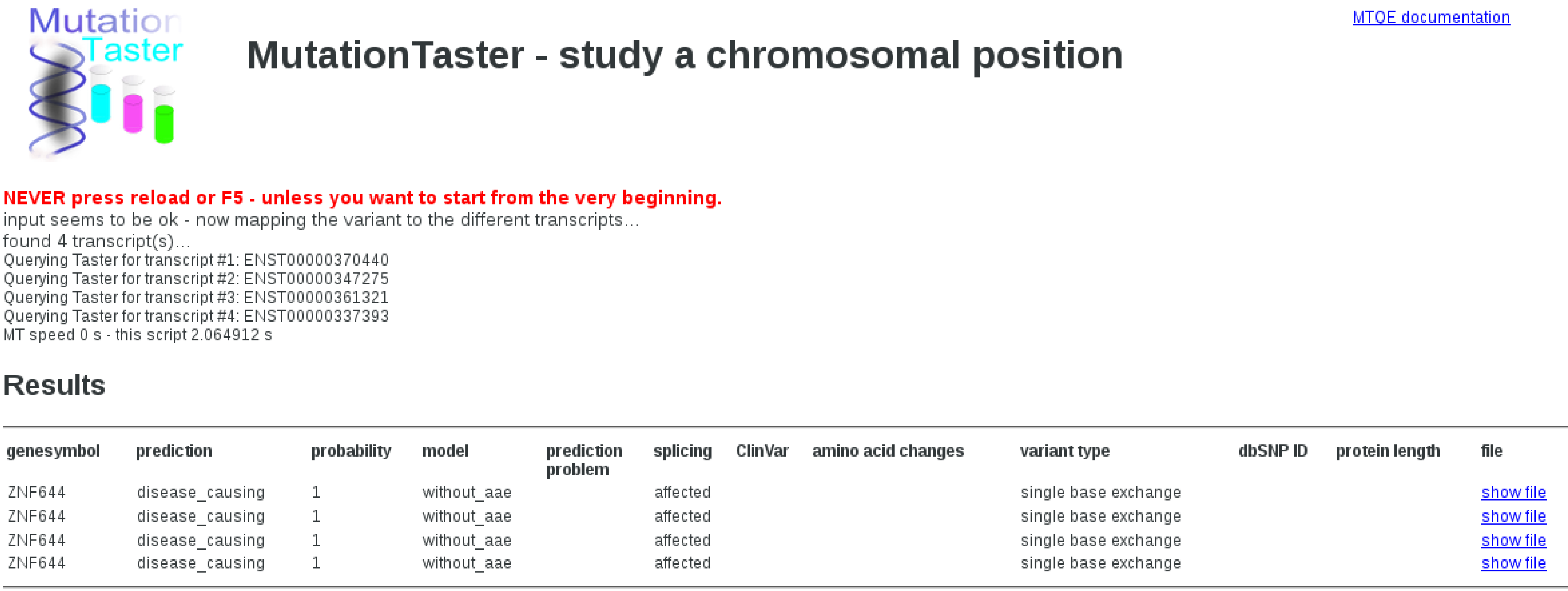
chromosome	1	position	91381763
reference allele	G	alternative allele	A
variant in HGVS notation		(currently only possible for SNVs)	*e.g. chr15:38852120A>T*

For InDels, use the VCF format, i.e. always **start with the last reference base before the variant**.

If you use MutationTaster, please cite our publication: Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods. 2014 Apr;11(4):361-2.*
Current build: NCBI 37 / Ensembl 66

Pretty bittersweet.

BERLIN
INSTITUTE
OF HEALTH
Charité & Max Delbrück Center



MutationTaster - study a chromosomal position

NEVER press reload or F5 - unless you want to start from the very beginning.

```
input seems to be ok - now mapping the variant to the different transcripts...
found 4 transcript(s)...
Querying Taster for transcript #1: ENST00000370440
Querying Taster for transcript #2: ENST00000347275
Querying Taster for transcript #3: ENST00000361321
Querying Taster for transcript #4: ENST00000337393
MT speed 0 s - this script 2.064912 s
```

Results

genesymbol	prediction	probability	model	prediction problem	splicing	ClinVar	amino acid changes	variant type	dbSNP ID	protein length	file
ZNF644	disease_causing	1	without_aae		affected			single base exchange			show file
ZNF644	disease_causing	1	without_aae		affected			single base exchange			show file
ZNF644	disease_causing	1	without_aae		affected			single base exchange			show file
ZNF644	disease_causing	1	without_aae		affected			single base exchange			show file

documentation

mutation t@sting

Prediction	disease causing	Model: *without_aae*, prob: 1 (explain)
Summary	- splice site changes	hyperlink

analysed issue	analysis result
name of alteration	no title
alteration (phys. location)	chr1:91381763C>A show variant in all transcripts IGV
HGNC symbol	ZNF644
Ensembl transcript ID	ENST00000361321
Genbank transcript ID	N/A
UniProt peptide	N/A
alteration type	single base exchange
alteration region	3'UTR
DNA changes	cDNA.1232G>T g.106067G>T
AA changes	N/A
position(s) of altered AA if AA alteration in CDS	N/A
frameshift	N/A
known variant	Variant was neither found in ExAC nor 1000G. Search ExAC.
regulatory features	H3K9me1, Histone, Histone 3 Lysine 9 mono-methylation H3K36me3, Histone, Histone 3 Lysine 36 Tri-Methylation H4K20me1, Histone, Histone 4 Lysine 20 mono-methylation
phyloP / phastCons	PhyloP PhastCons (flanking) 3.515 1 4.214 1 (flanking) 4.214 1 explain score(s) and/or inspect your position(s) in in UCSC Genome Browser
splice sites	splice site change occurs after stopcodon (at aa 302) splice site change occurs after stopcodon (at aa 305) splice site change occurs after stopcodon (at aa 306)

effect	gDNA position	score	detection sequence	exon-intron border
Acc marginally increased	106063	wt: 0.4310 / mu: 0.4311 (marginal change - not scored)	wt: CGGTTTTTTTTATACTAAAAAGTGGAGGGAGATTTGTTTAA mu: CGGTTTTTTTTATACTAAAAAGTGTAGGGAGATTTGTTTAA	aaaa|GTGG
Acc marginally increased	106058	wt: 0.2456 / mu: 0.2516 (marginal change - not scored)	wt: TGGAACGGTTTTTTTTATACTAAAAAGTGGAGGGAGATTTG mu: TGGAACGGTTTTTTTTATACTAAAAAGTGTAGGGAGATTTG	tact|AAAA
Donor increased	106072	wt: 0.36 / mu: 0.56	wt: GAGGGAGATTTGTTT mu: TAGGGAGATTTGTTT	GGGA|gatt
Donor increased	106060	wt: 0.22 / mu: 0.85	wt: TACTAAAAAGTGGAG mu: TACTAAAAAGTGTAG	CTAA|aaag
Donor marginally increased	106058	wt: 0.9832 / mu: 0.9913 (marginal change - not scored)	wt: TATACTAAAAAGTGG mu: TATACTAAAAAGTGT	TACT|aaaa
Donor gained	106070	0.31	mu: TGTAGGGAGATTTGT	TAGG|gaga

distance from splice site	22
Kozak consensus sequence altered?	N/A
conservation protein level for non-synonymous changes	N/A
protein features	N/A

How?

Once upon in my inbox

Subject: **New RX pharmacy**

WE NOW have online pharmacy take a look
......ablepharmacy.com
Payments are every Thursday like clockwork, no delays or arrays
Our "Low Price Pharmacy Store" design sports a professional array of pharmaceuticals.
This is definatly our top converting website.
Other product: enlargement pills
very popular sextoy

msg me with a valid email for an account

Once upon in my inbox

Subject: **New RX pharmacy**

WE NOW have online **pharmacy** take a look
......ablepharmacy.com
Payments are every Thursday like clockwork, no delays or arrays
Our "Low Price Pharmacy Store" design sports a professional array of pharmaceuticals.
This is definatly our top converting website.
Other product: **enlargement pills**
very popular **sextoy**

msg me with a valid email for an account

Mozilla Thunderbird uses a Bayes classifier

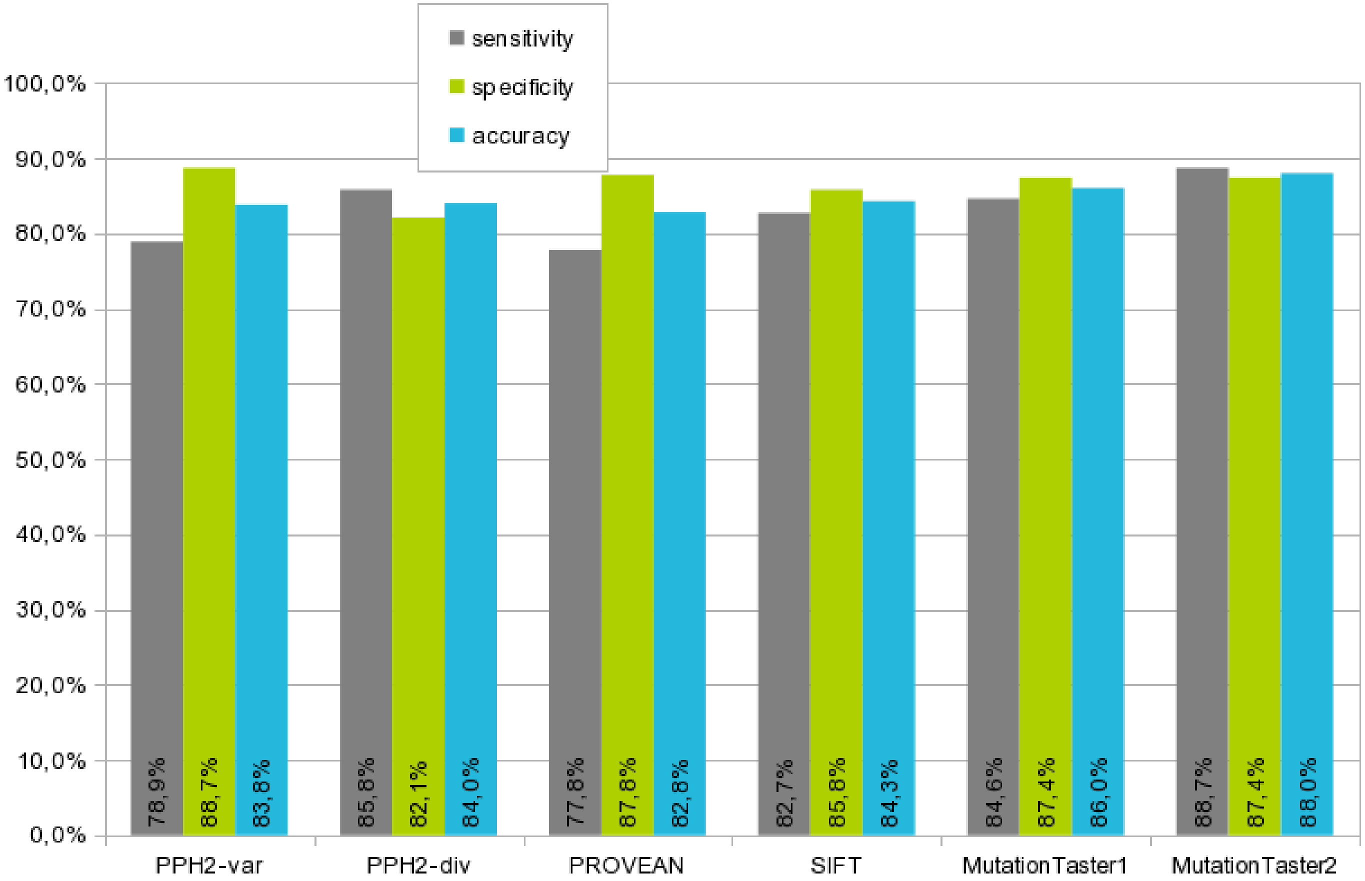
term	spam	ham
pharmacy	++	o
enlargement	++	o
pills	++	o
sextoy	++	--
website	+	+
abstract	-	++
MutationTaster	-	++

...and so does MutationTaster

BERLIN INSTITUTE OF HEALTH
Charité & Max Delbrück Center

Test result	**mutations**	**polymorph.**
abrogation of a splice site	49.5%	0.06%
loss of a transmembrane domain	7.3%	4.5%
loss of a disulfid bridge	2.9%	0.1%
trained with	DM from HGMD© Pro	20+ persons homozygous in 1000G

Comparison of different tools



PolyPhen

J. Craig Venter Institute SIFT

J. Craig Venter Institute PROVEAN

2 x 1,100 non-synonymous variants

Do not rely on predictions - include background knowledge

	MutationTaster2	PPH	SIFT	PROVEAN
all predictions				
FP	6	376	295	331
TN	2771	776	2482	2446
FPR	0.2%	32.6%	10.6%	11.9%
1152 variants predicted by all tools				
FP	6	376	274	290
TN	1146	776	878	862
FPR	0.5%	32.6%	23.8%	25.2%

exome of a healthy individual
all homozygous non-synonymous variants

Combination scores

- integrate different prediction tools
- integrate further data (may overlap!)
- also used/created by the all-in-one tools
- often only for non-synonymous variants!

examples

CADD
Combined Annotation Dependent Depletion

CONDEL
CONsensus DELeteriousness

dbNSFP

Database for functional prediction and annotation of all potential non-synonymous single-nucleotide variants

- non-synonymous variants
- splice site variants
- pre-computer values from many prediction tools
- many pre-computed combination scores
- allele frequencies

- no InDels!
- no web interface

VARIANT PRIORITISATION

BERLIN INSTITUTE OF HEALTH

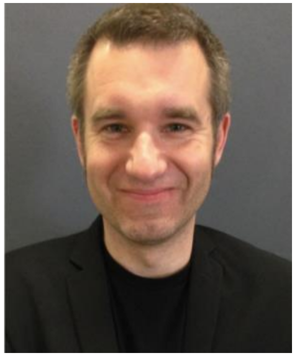
Charité & Max Delbrück Center

How can you interpret 10,000+ variants? Bioinformatics!

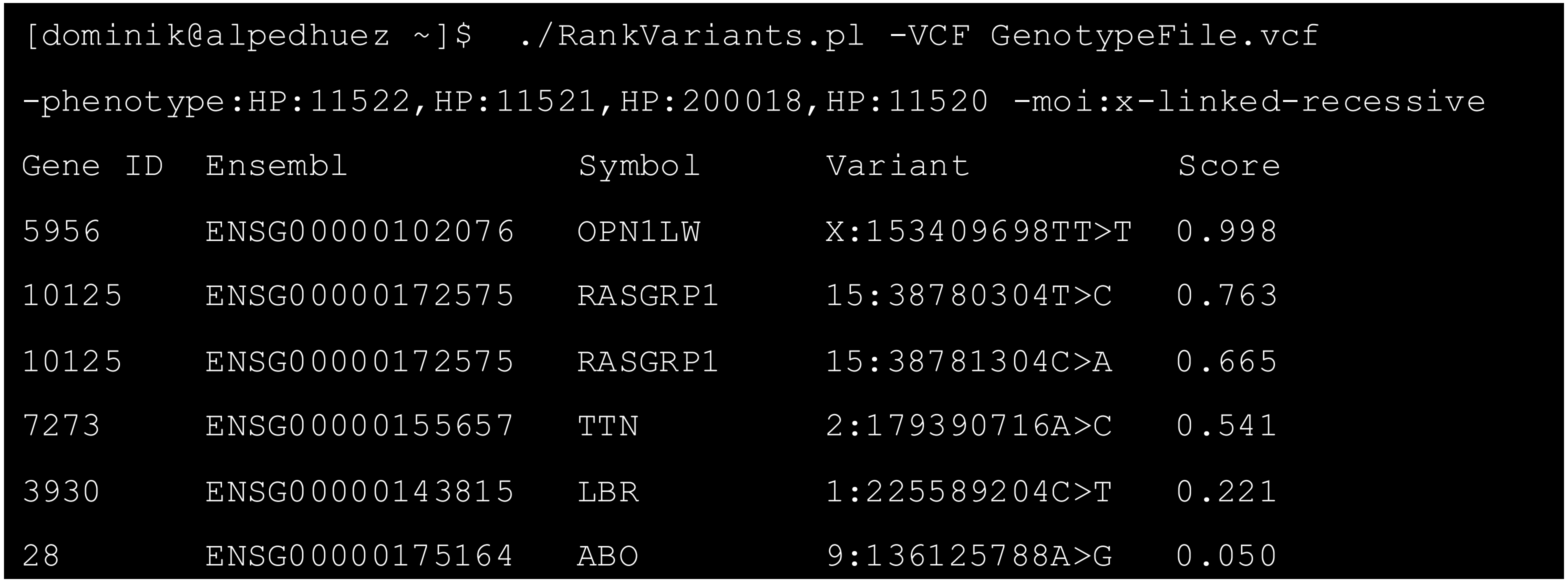
© Justin Stephens/Corbis Outline

140,000 variants!
Protanopia?
Deuteranopia?
Protanomaly?
Deuteranomaly?

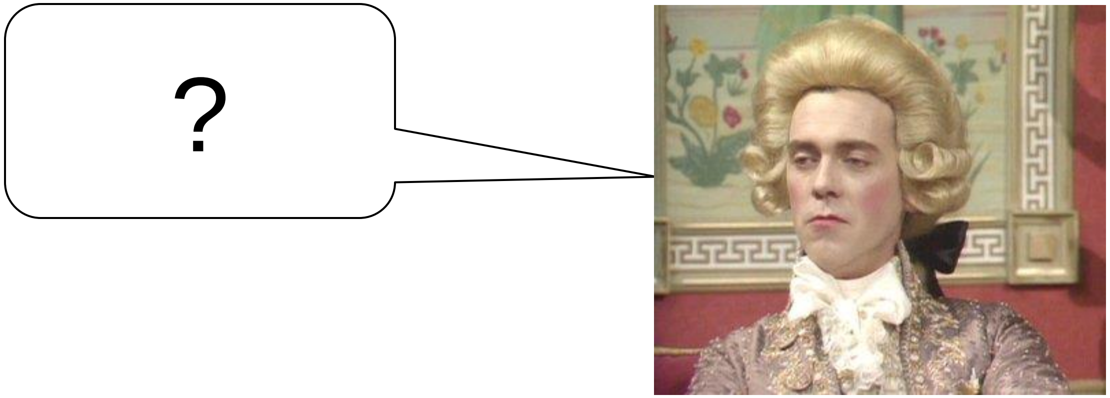
?



A world apart.



```
[dominik@alpedhuez ~]$  ./RankVariants.pl -VCF GenotypeFile.vcf
-phenotype:HP:11522,HP:11521,HP:200018,HP:11520 -moi:x-linked-recessive
Gene ID  Ensembl           Symbol      Variant          Score
5956     ENSG00000102076   OPN1LW      X:153409698TT>T  0.998
10125    ENSG00000172575   RASGRP1     15:38780304T>C   0.763
10125    ENSG00000172575   RASGRP1     15:38781304C>A   0.665
7273     ENSG00000155657   TTN         2:179390716A>C   0.541
3930     ENSG00000143815   LBR         1:225589204C>T   0.221
28       ENSG00000175164   ABO         9:136125788A>G   0.050
```



© BBC/Blackadder 3

Enough for a mouse model?

© Markus Schuelke (Charité)

Software should adapt to the user!

© *Wilson Afonso*
(WikiCommons)

(Some) all-in-one tools

eXtasy
Variant Prioritization

PhenIX

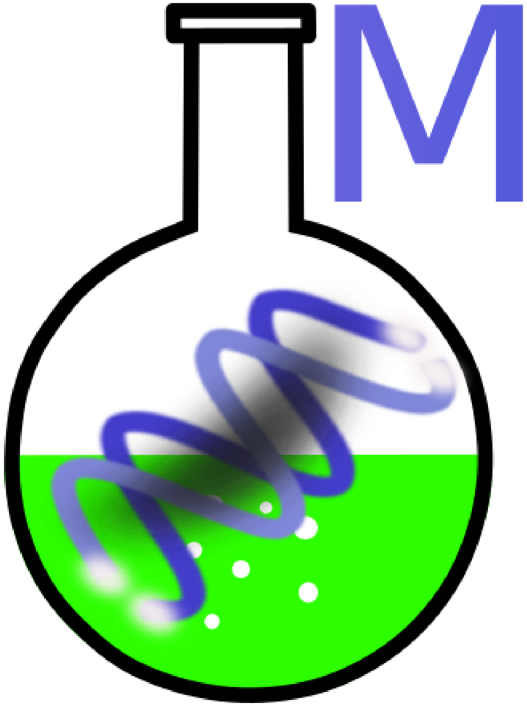
PHEVOR
PHENOTYPE DRIVEN VARIANT ONTOLOGICAL RE-RANKING TOOL

Exomiser

Phenolyzer

Mutation Distiller

Finding disease mutations with

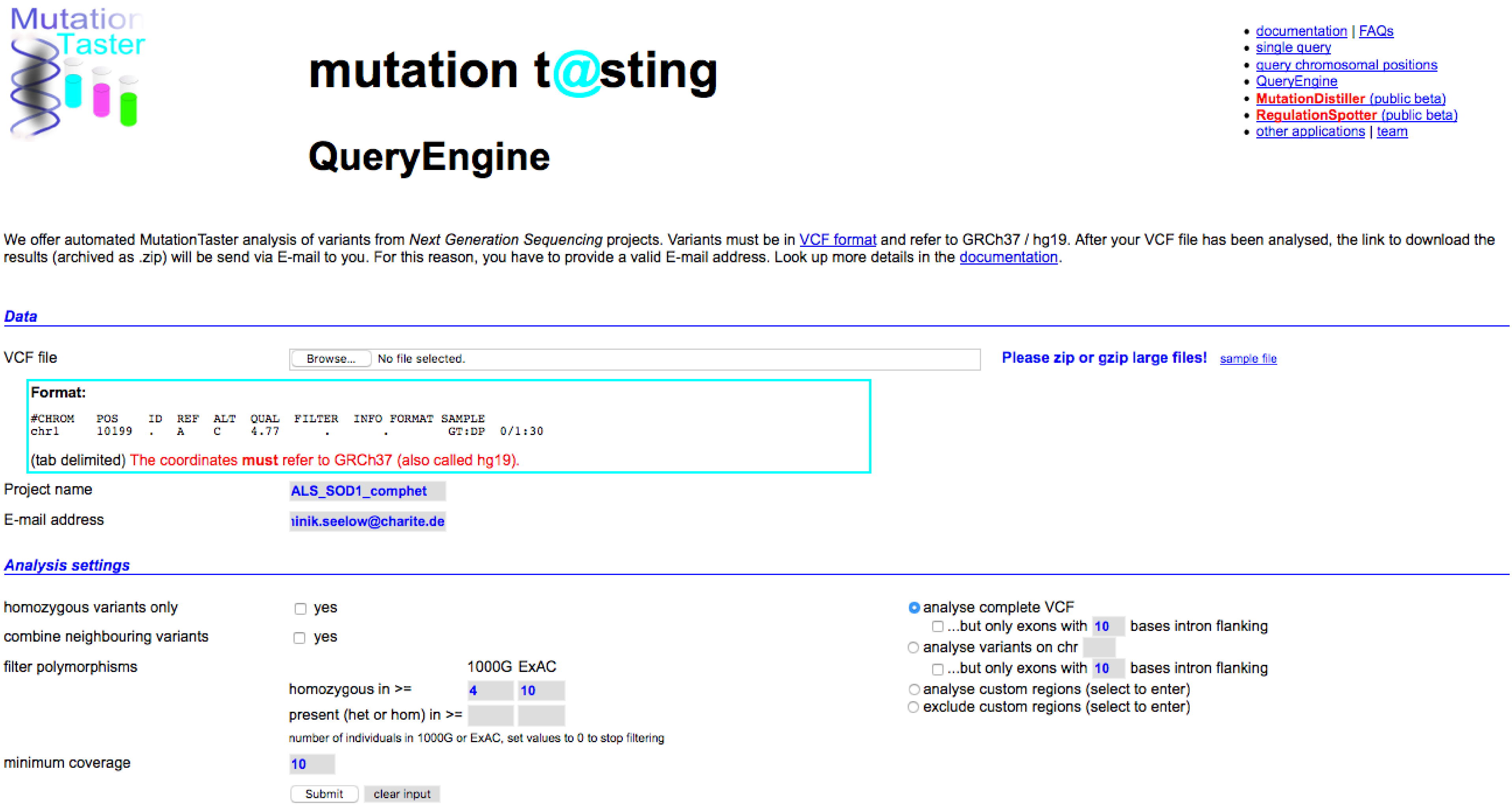


Mutation Distiller

Daniela
Hombach

Healthy exome plus two heteroz. SOD1 mutations (causing recessive ALS)

BERLIN
INSTITUTE
OF HEALTH
Charité & Max Delbrück Center



mutation t@sting

QueryEngine

- documentation | FAQs
- single query
- query chromosomal positions
- QueryEngine
- **MutationDistiller** (public beta)
- **RegulationSpotter** (public beta)
- other applications | team

We offer automated MutationTaster analysis of variants from *Next Generation Sequencing* projects. Variants must be in VCF format and refer to GRCh37 / hg19. After your VCF file has been analysed, the link to download the results (archived as .zip) will be send via E-mail to you. For this reason, you have to provide a valid E-mail address. Look up more details in the documentation.

Data

VCF file — Browse... No file selected. **Please zip or gzip large files!** sample file

Format:

```
#CHROM  POS    ID  REF  ALT  QUAL  FILTER  INFO  FORMAT  SAMPLE
chr1    10199  .   A    C    4.77    .       .     GT:DP   0/1:30
```

(tab delimited) The coordinates **must** refer to GRCh37 (also called hg19).

Project name: ALS_SOD1_comphet

E-mail address: ıinik.seelow@charite.de

Analysis settings

homozygous variants only — □ yes

combine neighbouring variants — □ yes

filter polymorphisms

	1000G	ExAC
homozygous in >=	4	10
present (het or hom) in >=		

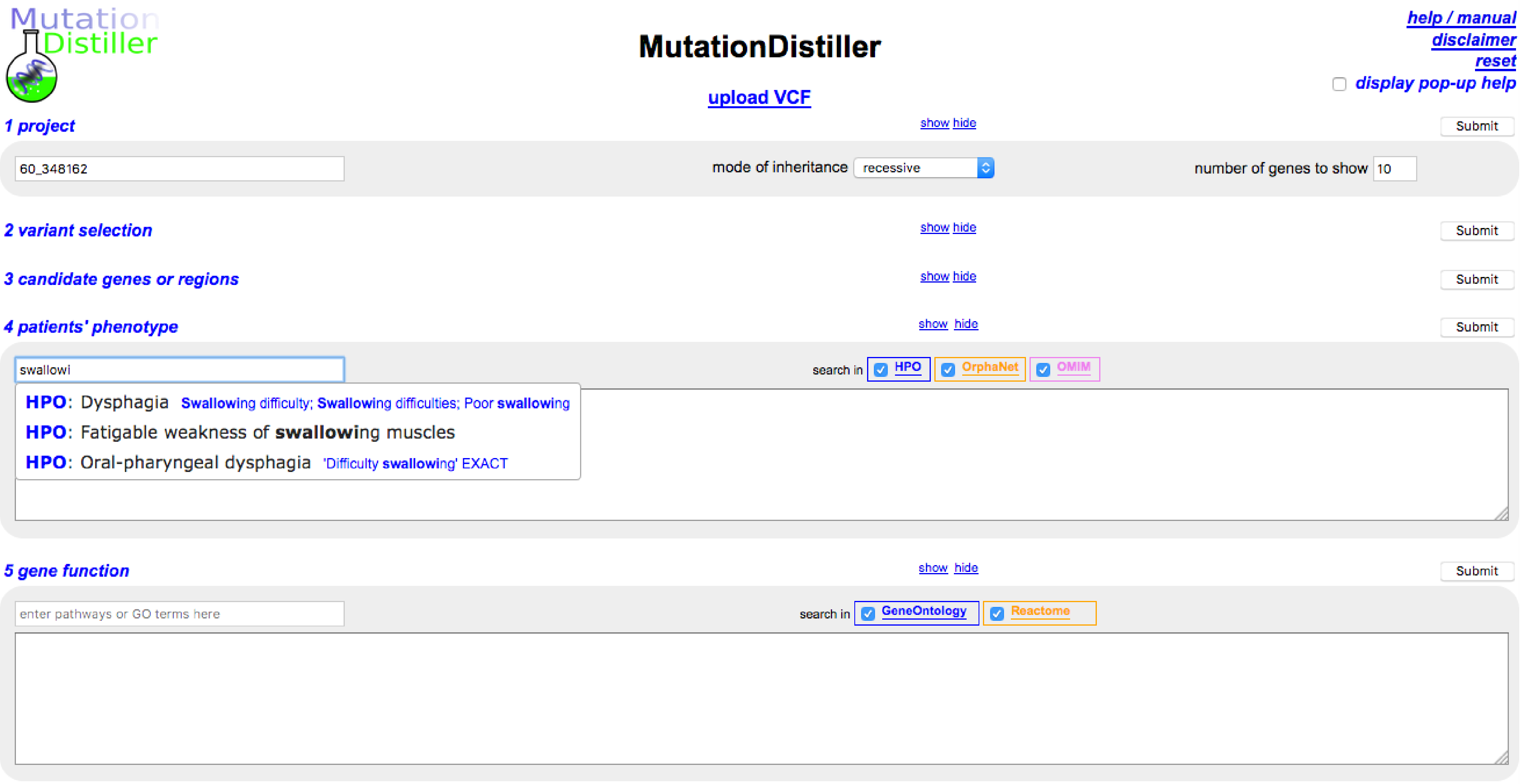
number of individuals in 1000G or ExAC, set values to 0 to stop filtering

minimum coverage: 10

● analyse complete VCF
□ ...but only exons with 10 bases intron flanking
○ analyse variants on chr
□ ...but only exons with 10 bases intron flanking
○ analyse custom regions (select to enter)
○ exclude custom regions (select to enter)

Healthy exome plus two heteroz. SOD1 mutations (causing recessive ALS)

BERLIN INSTITUTE OF HEALTH
Charité & Max Delbrück Center



MutationDistiller

MutationDistiller

help / manual
disclaimer
reset
□ display pop-up help

upload VCF

1 project show hide Submit

60_348162

mode of inheritance: recessive

number of genes to show: 10

2 variant selection show hide Submit

3 candidate genes or regions show hide Submit

4 patients' phenotype show hide Submit

swallowi

search in ☑ HPO ☑ OrphaNet ☑ OMIM

HPO: Dysphagia Swallowing difficulty; Swallowing difficulties; Poor swallowing
HPO: Fatigable weakness of **swallowi**ng muscles
HPO: Oral-pharyngeal dysphagia 'Difficulty swallowing' EXACT

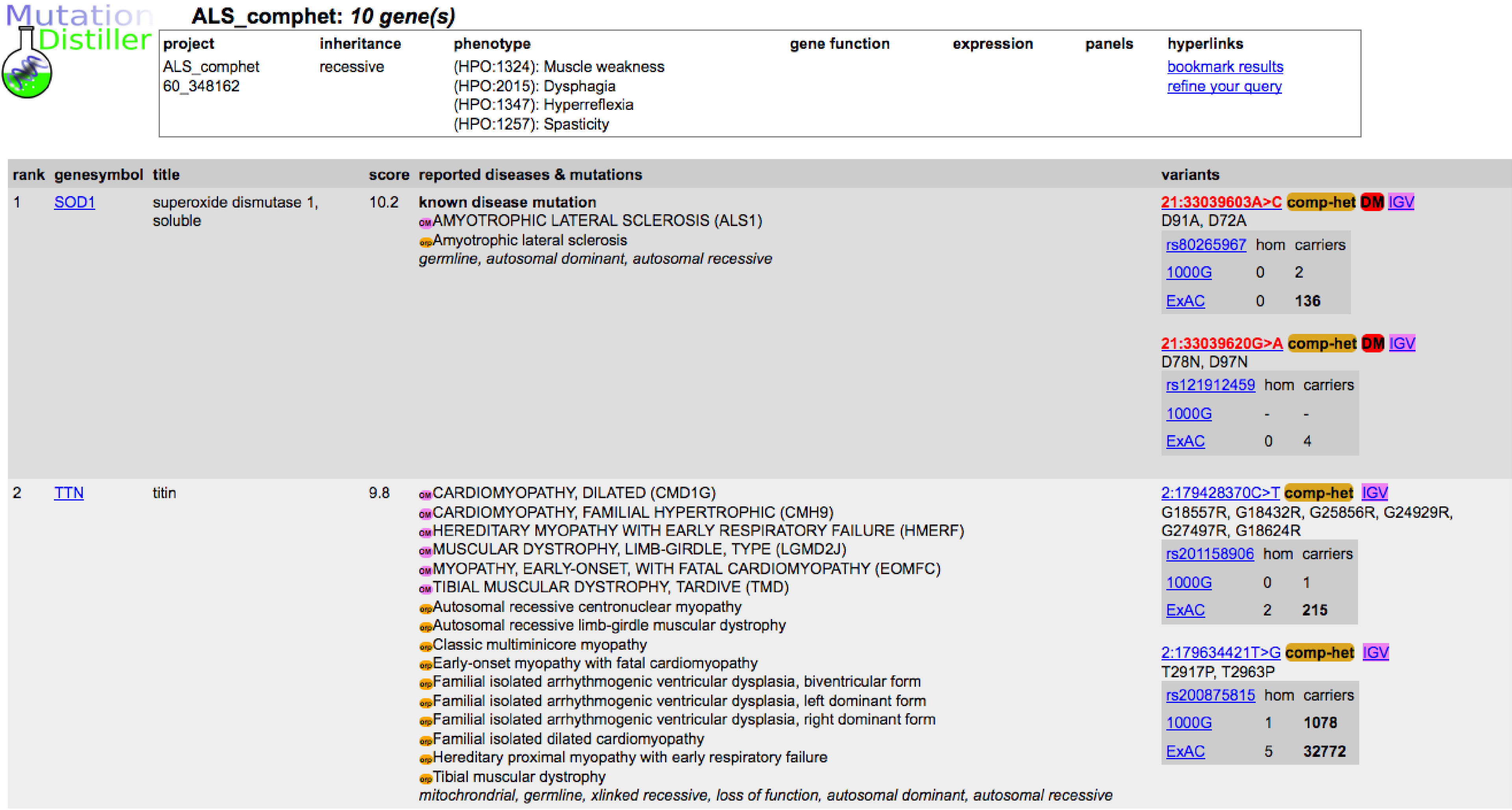
5 gene function show hide Submit

enter pathways or GO terms here

search in ☑ GeneOntology ☑ Reactome

Results (overview)

BERLIN
INSTITUTE
OF HEALTH
Charité & Max Delbrück Center



Mutation Distiller

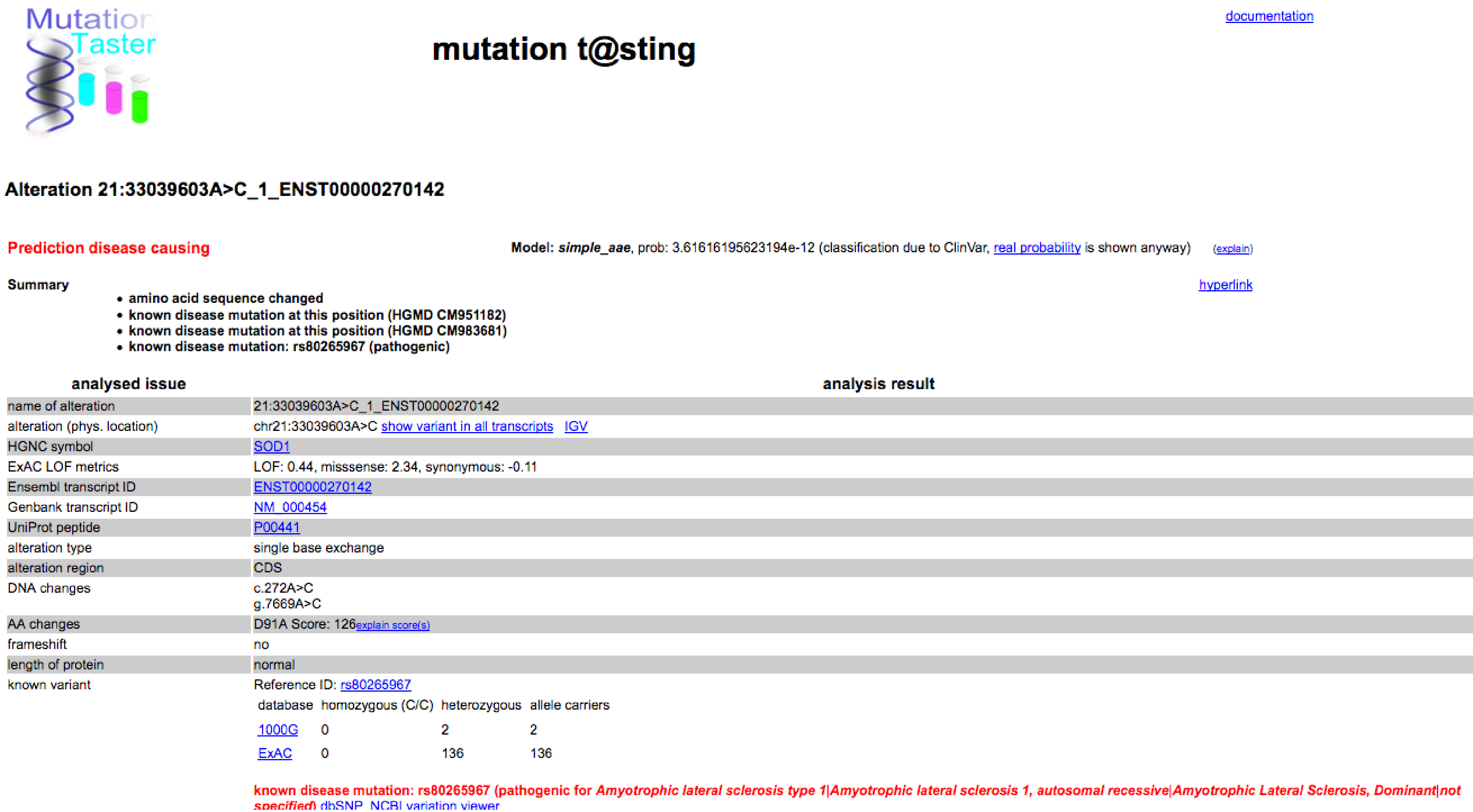
ALS_comphet: *10 gene(s)*

project	inheritance	phenotype	gene function	expression	panels	hyperlinks
ALS_comphet 60_348162	recessive	(HPO:1324): Muscle weakness (HPO:2015): Dysphagia (HPO:1347): Hyperreflexia (HPO:1257): Spasticity				bookmark results refine your query

rank	genesymbol	title	score	reported diseases & mutations	variants
1	SOD1	superoxide dismutase 1, soluble	10.2	**known disease mutation** OM AMYOTROPHIC LATERAL SCLEROSIS (ALS1) orp Amyotrophic lateral sclerosis *germline, autosomal dominant, autosomal recessive*	21:33039603A>C **comp-het** **DM** IGV D91A, D72A rs80265967 hom carriers 1000G 0 2 ExAC 0 **136** 21:33039620G>A **comp-het** **DM** IGV D78N, D97N rs121912459 hom carriers 1000G - - ExAC 0 4
2	TTN	titin	9.8	OM CARDIOMYOPATHY, DILATED (CMD1G) OM CARDIOMYOPATHY, FAMILIAL HYPERTROPHIC (CMH9) OM HEREDITARY MYOPATHY WITH EARLY RESPIRATORY FAILURE (HMERF) OM MUSCULAR DYSTROPHY, LIMB-GIRDLE, TYPE (LGMD2J) OM MYOPATHY, EARLY-ONSET, WITH FATAL CARDIOMYOPATHY (EOMFC) OM TIBIAL MUSCULAR DYSTROPHY, TARDIVE (TMD) orp Autosomal recessive centronuclear myopathy orp Autosomal recessive limb-girdle muscular dystrophy orp Classic multiminicore myopathy orp Early-onset myopathy with fatal cardiomyopathy orp Familial isolated arrhythmogenic ventricular dysplasia, biventricular form orp Familial isolated arrhythmogenic ventricular dysplasia, left dominant form orp Familial isolated arrhythmogenic ventricular dysplasia, right dominant form orp Familial isolated dilated cardiomyopathy orp Hereditary proximal myopathy with early respiratory failure orp Tibial muscular dystrophy *mitochrondrial, germline, xlinked recessive, loss of function, autosomal dominant, autosomal recessive*	2:179428370C>T **comp-het** IGV G18557R, G18432R, G25856R, G24929R, G27497R, G18624R rs201158906 hom carriers 1000G 0 1 ExAC 2 **215** 2:179634421T>G **comp-het** IGV T2917P, T2963P rs200875815 hom carriers 1000G 1 **1078** ExAC 5 **32772**

Inspect prediction details

BERLIN INSTITUTE OF HEALTH
Charité & Max Delbrück Center



documentation

mutation t@sting

Alteration 21:33039603A>C_1_ENST00000270142

Prediction disease causing

Model: ***simple_aae***, prob: 3.61616195623194e-12 (classification due to ClinVar, real probability is shown anyway) (explain)

Summary

hyperlink

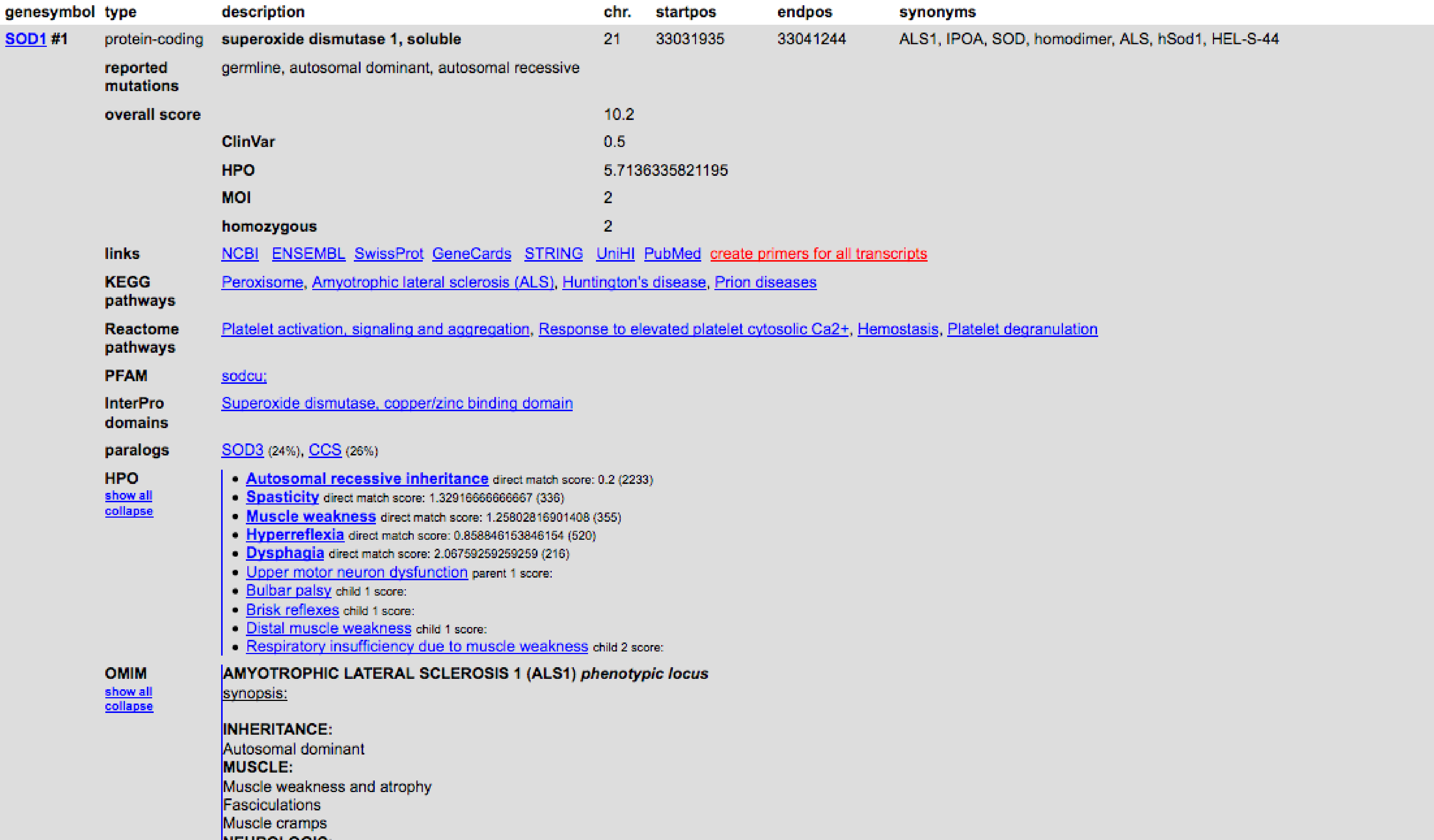
- **amino acid sequence changed**
- **known disease mutation at this position (HGMD CM951182)**
- **known disease mutation at this position (HGMD CM983681)**
- **known disease mutation: rs80265967 (pathogenic)**

analysed issue	analysis result
name of alteration	21:33039603A>C_1_ENST00000270142
alteration (phys. location)	chr21:33039603A>C show variant in all transcripts IGV
HGNC symbol	SOD1
ExAC LOF metrics	LOF: 0.44, misssense: 2.34, synonymous: -0.11
Ensembl transcript ID	ENST00000270142
Genbank transcript ID	NM_000454
UniProt peptide	P00441
alteration type	single base exchange
alteration region	CDS
DNA changes	c.272A>C g.7669A>C
AA changes	D91A Score: 126explain score(s)
frameshift	no
length of protein	normal
known variant	Reference ID: rs80265967

database	homozygous (C/C)	heterozygous	allele carriers
1000G	0	2	2
ExAC	0	136	136

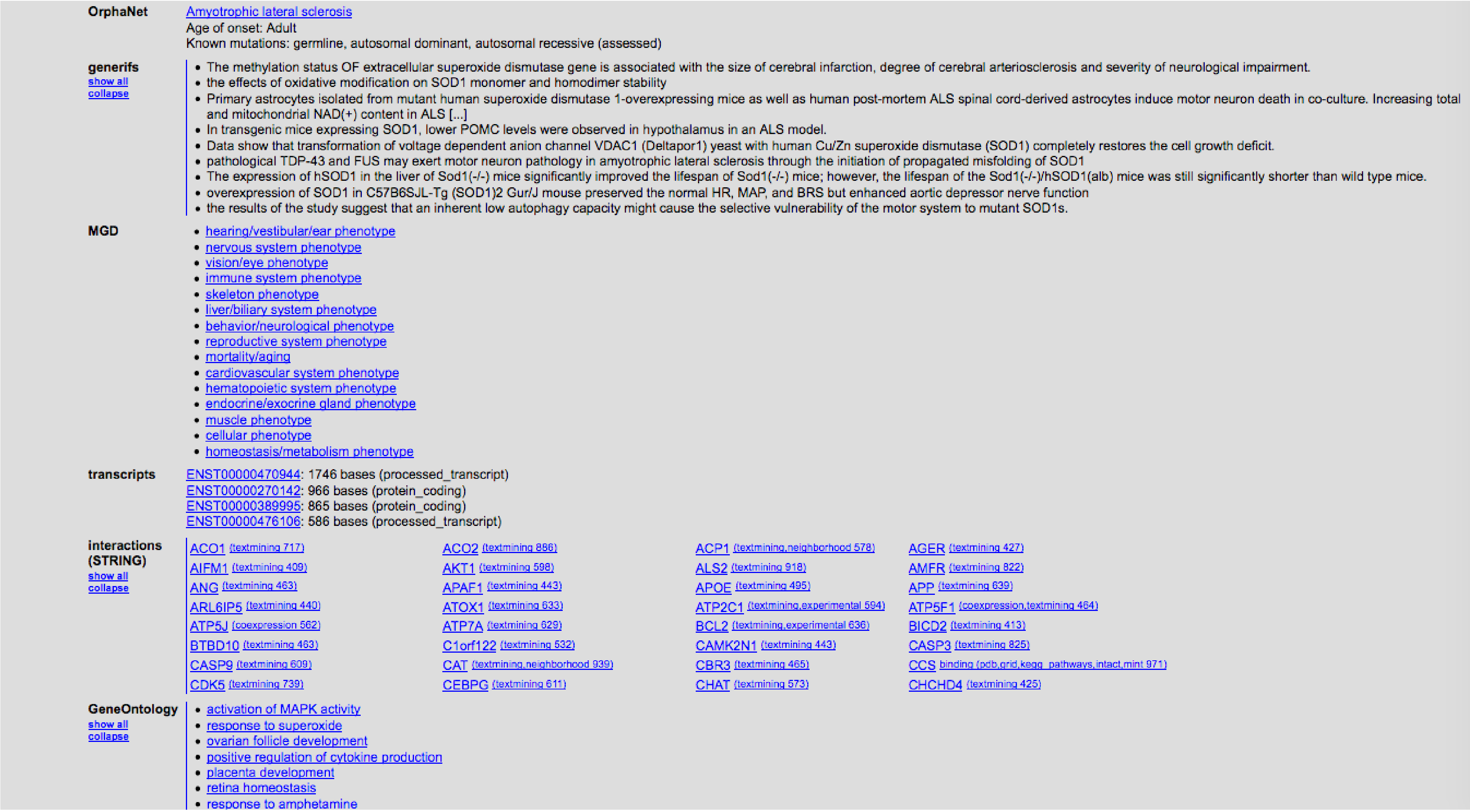
known disease mutation: rs80265967 (pathogenic for *Amyotrophic lateral sclerosis type 1|Amyotrophic lateral sclerosis 1, autosomal recessive|Amyotrophic Lateral Sclerosis, Dominant|not specified*) dbSNP NCBI variation viewer

Gene information included!



genesymbol	type	description	chr.	startpos	endpos	synonyms
SOD1 #1	protein-coding	**superoxide dismutase 1, soluble**	21	33031935	33041244	ALS1, IPOA, SOD, homodimer, ALS, hSod1, HEL-S-44
	reported mutations	germline, autosomal dominant, autosomal recessive				
	overall score		10.2			
		ClinVar	0.5			
		HPO	5.7136335821195			
		MOI	2			
		homozygous	2			
	links	NCBI ENSEMBL SwissProt GeneCards STRING UniHI PubMed create primers for all transcripts				
	KEGG pathways	Peroxisome, Amyotrophic lateral sclerosis (ALS), Huntington's disease, Prion diseases				
	Reactome pathways	Platelet activation, signaling and aggregation, Response to elevated platelet cytosolic Ca2+, Hemostasis, Platelet degranulation				
	PFAM	sodcu;				
	InterPro domains	Superoxide dismutase, copper/zinc binding domain				
	paralogs	SOD3 (24%), CCS (26%)				
	HPO show all collapse	• **Autosomal recessive inheritance** direct match score: 0.2 (2233) • **Spasticity** direct match score: 1.32916666666667 (336) • **Muscle weakness** direct match score: 1.25802816901408 (355) • **Hyperreflexia** direct match score: 0.858846153846154 (520) • **Dysphagia** direct match score: 2.06759259259259 (216) • Upper motor neuron dysfunction parent 1 score: • Bulbar palsy child 1 score: • Brisk reflexes child 1 score: • Distal muscle weakness child 1 score: • Respiratory insufficiency due to muscle weakness child 2 score:				
	OMIM show all collapse	**AMYOTROPHIC LATERAL SCLEROSIS 1 (ALS1)** ***phenotypic locus*** synopsis: **INHERITANCE:** Autosomal dominant **MUSCLE:** Muscle weakness and atrophy Fasciculations Muscle cramps **NEUROLOGIC:**				

Gene information included!



OrphaNet
Amyotrophic lateral sclerosis
Age of onset: Adult
Known mutations: germline, autosomal dominant, autosomal recessive (assessed)

generifs
show all
collapse

- The methylation status OF extracellular superoxide dismutase gene is associated with the size of cerebral infarction, degree of cerebral arteriosclerosis and severity of neurological impairment.
- the effects of oxidative modification on SOD1 monomer and homodimer stability
- Primary astrocytes isolated from mutant human superoxide dismutase 1-overexpressing mice as well as human post-mortem ALS spinal cord-derived astrocytes induce motor neuron death in co-culture. Increasing total and mitochondrial NAD(+) content in ALS [...]
- In transgenic mice expressing SOD1, lower POMC levels were observed in hypothalamus in an ALS model.
- Data show that transformation of voltage dependent anion channel VDAC1 (Deltapor1) yeast with human Cu/Zn superoxide dismutase (SOD1) completely restores the cell growth deficit.
- pathological TDP-43 and FUS may exert motor neuron pathology in amyotrophic lateral sclerosis through the initiation of propagated misfolding of SOD1
- The expression of hSOD1 in the liver of Sod1(-/-) mice significantly improved the lifespan of Sod1(-/-) mice; however, the lifespan of the Sod1(-/-)/hSOD1(alb) mice was still significantly shorter than wild type mice.
- overexpression of SOD1 in C57B6SJL-Tg (SOD1)2 Gur/J mouse preserved the normal HR, MAP, and BRS but enhanced aortic depressor nerve function
- the results of the study suggest that an inherent low autophagy capacity might cause the selective vulnerability of the motor system to mutant SOD1s.

MGD

- hearing/vestibular/ear phenotype
- nervous system phenotype
- vision/eye phenotype
- immune system phenotype
- skeleton phenotype
- liver/biliary system phenotype
- behavior/neurological phenotype
- reproductive system phenotype
- mortality/aging
- cardiovascular system phenotype
- hematopoietic system phenotype
- endocrine/exocrine gland phenotype
- muscle phenotype
- cellular phenotype
- homeostasis/metabolism phenotype

transcripts
ENST00000470944: 1746 bases (processed_transcript)
ENST00000270142: 966 bases (protein_coding)
ENST00000389995: 865 bases (protein_coding)
ENST00000476106: 586 bases (processed_transcript)

interactions (STRING)
show all
collapse

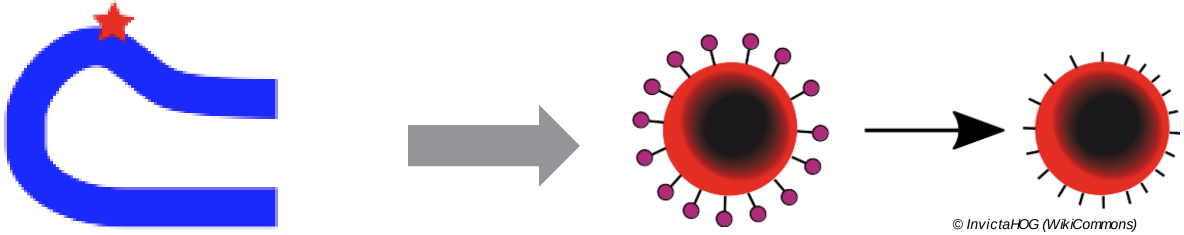
ACO1 (textmining 717)	ACO2 (textmining 886)	ACP1 (textmining,neighborhood 578)	AGER (textmining 427)
AIFM1 (textmining 409)	AKT1 (textmining 598)	ALS2 (textmining 918)	AMFR (textmining 822)
ANG (textmining 463)	APAF1 (textmining 443)	APOE (textmining 495)	APP (textmining 639)
ARL6IP5 (textmining 440)	ATOX1 (textmining 633)	ATP2C1 (textmining,experimental 594)	ATP5F1 (coexpression,textmining 464)
ATP5J (coexpression 562)	ATP7A (textmining 629)	BCL2 (textmining,experimental 636)	BICD2 (textmining 413)
BTBD10 (textmining 463)	C1orf122 (textmining 532)	CAMK2N1 (textmining 443)	CASP3 (textmining 825)
CASP9 (textmining 609)	CAT (textmining,neighborhood 939)	CBR3 (textmining 465)	CCS binding (pdb,grid,kegg_pathways,intact,mint 971)
CDK5 (textmining 739)	CEBPG (textmining 611)	CHAT (textmining 573)	CHCHD4 (textmining 425)

GeneOntology
show all
collapse

- activation of MAPK activity
- response to superoxide
- ovarian follicle development
- positive regulation of cytokine production
- placenta development
- retina homeostasis
- response to amphetamine
]

Consider the disease / phenotype

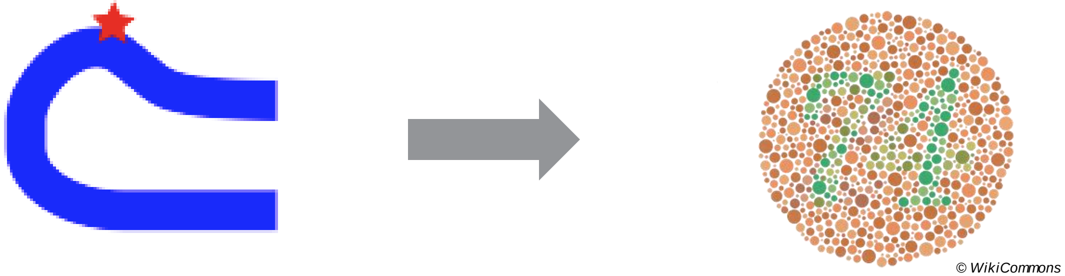
ABO gene



© InvictaHOG (WikiCommons)

may change
blood type

OPN1LW gene



© WikiCommons

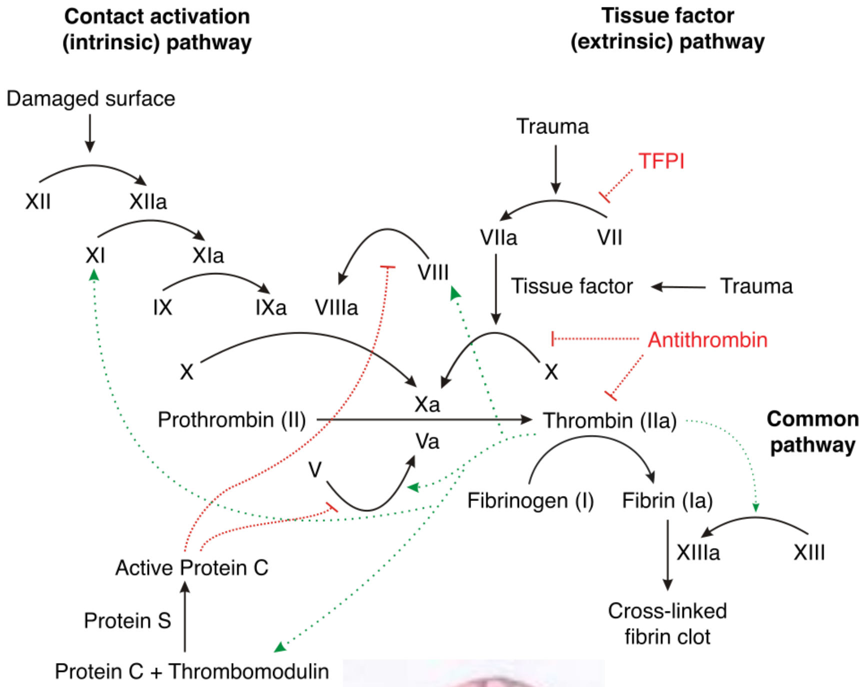
may cause
colour blindness

HPO

orphanet

MGI

Consider gene function & expression

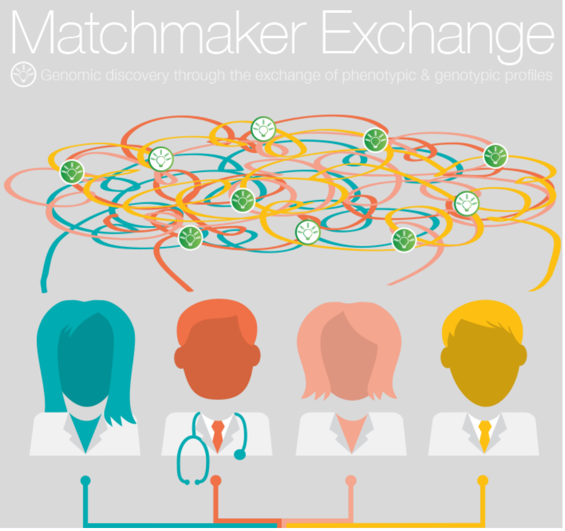


REACTOME

GENEONTOLOGY
Unifying Biology

ArrayExpress

Match-making


Matchmaker Exchange
Genomic discovery through the exchange of phenotypic & genotypic profiles

GeneMatcher

Cafe Variome

DECIPHER
GRCh37

PHENOME CENTRAL

- Did others report this variant?
- Find partners!
- Please share your variants of unknown significance!

General considerations

Use your brain!

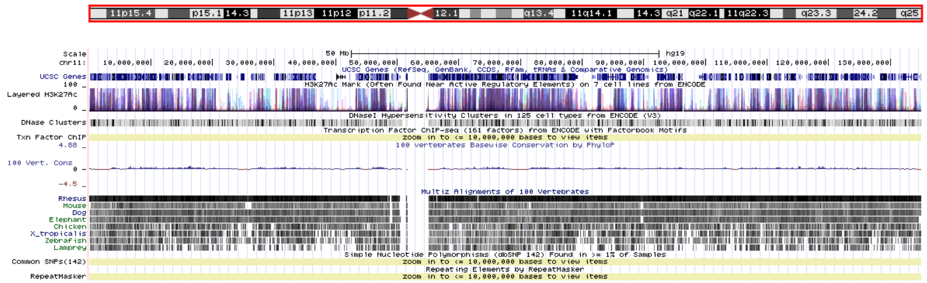
- don’t trust predictors blindly
- disease databases may be wrong
- think of reduced penetrance & compound heterozgyosity
- do not exclude synonymous variants
- check variant with IGV
- look up polymorphism databases
- consider phenotype & gene function
- consider gene expression
- **do segregation analysis!**

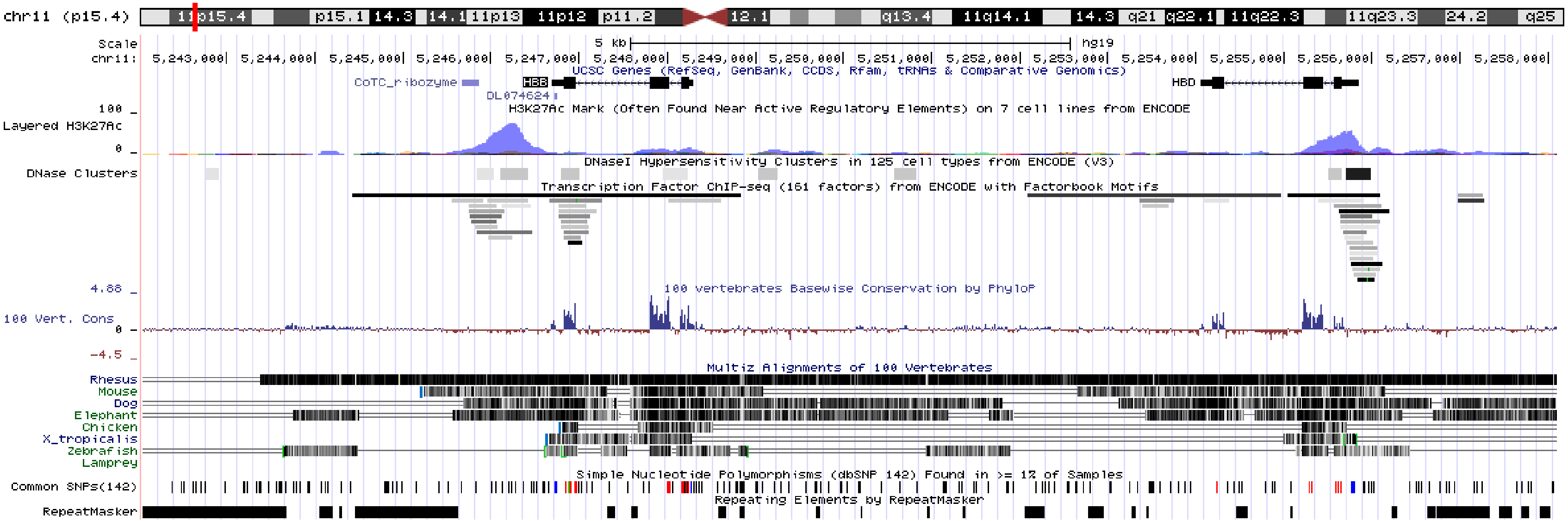
CHALLENGES OF INTERPRETING NON-CODING VARIANTS

1. No information vs information overload
2. Combined variant scores
3. Experimental assessment using reporter assays

No information vs information overload

Chromosome 11: 135 Mb of sequence
UCSC Genome Browser

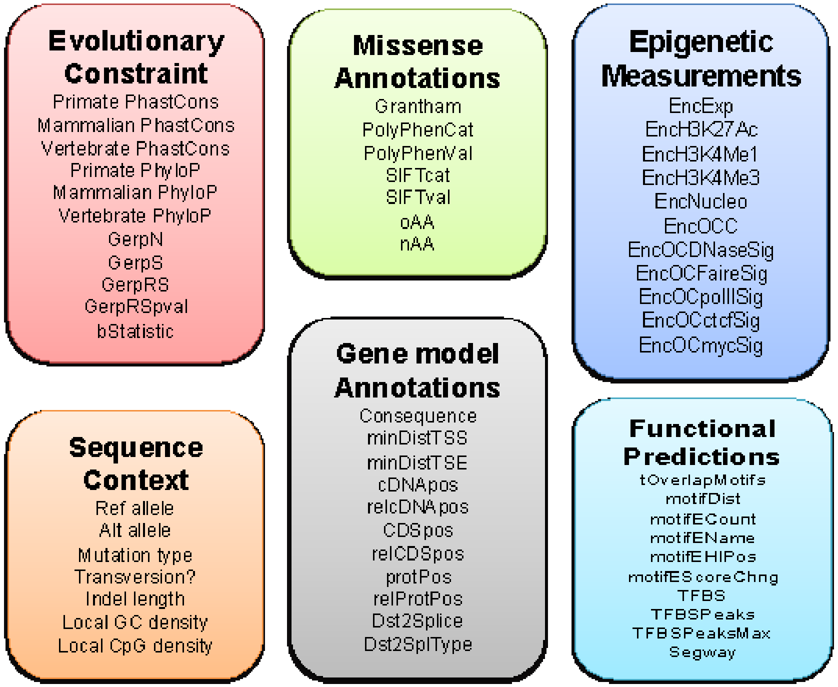

11p15.4
p15.1
14.3
11p13
11p12
p11.2
12.1
q13.4
11q14.1
14.3
q21
q22.1
11q22.3
q23.3
24.2
q25
Scale
chr11:
50 Mb
hg19
10,000,000
20,000,000
30,000,000
40,000,000
50,000,000
60,000,000
70,000,000
80,000,000
90,000,000
100,000,000
110,000,000
120,000,000
130,000,000
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
UCSC Genes
100
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
0
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
DNase Clusters
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
Txn Factor ChIP
zoom in to <= 10,000,000 bases to view items
4.88
100 vertebrates Basewise Conservation by PhyloP
100 Vert. Cons
0
-4.5
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples
Common SNPs(142)
zoom in to <= 10,000,000 bases to view items
Repeating Elements by RepeatMasker
RepeatMasker
zoom in to <= 10,000,000 bases to view items


chr11 (p15.4)
11p15.4
p15.1
14.3
14.1
11p13
11p12
p11.2
12.1
q13.4
11q14.1
14.3
q21
q22.1
11q22.3
11q23.3
24.2
q25
Scale
chr11:
5 kb
hg19
5,243,000
5,244,000
5,245,000
5,246,000
5,247,000
5,248,000
5,249,000
5,250,000
5,251,000
5,252,000
5,253,000
5,254,000
5,255,000
5,256,000
5,257,000
5,258,000
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
CoTC_ribozyme
HBB
DL074624
HBD
100
Layered H3K27Ac
0
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
DNase Clusters
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
4.88
100 vertebrates Basewise Conservation by PhyloP
100 Vert. Cons
0
-4.5
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples
Common SNPs(142)
Repeating Elements by RepeatMasker
RepeatMasker

Which annotation to use?

BERLIN INSTITUTE OF HEALTH
Charité & Max Delbrück Center

- Expanding panoply of partially correlated annotations
- Different scales, transformations – clustering, orthogonalization?
- Apply to variously overlapping subsets of genomic variants
- Most annotations are only defined in very specific contexts: power of domain-specific scores



Evolutionary Constraint
Primate PhastCons
Mammalian PhastCons
Vertebrate PhastCons
Primate PhyloP
Mammalian PhyloP
Vertebrate PhyloP
GerpN
GerpS
GerpRS
GerpRSpval
bStatistic

Missense Annotations
Grantham
PolyPhenCat
PolyPhenVal
SIFTcat
SIFTval
oAA
nAA

Epigenetic Measurements
EncExp
EncH3K27Ac
EncH3K4Me1
EncH3K4Me3
EncNucleo
EncOCC
EncOCDNaseSig
EncOCFaireSig
EncOCpolIISig
EncOCctcfSig
EncOCmycSig

Sequence Context
Ref allele
Alt allele
Mutation type
Transversion?
Indel length
Local GC density
Local CpG density

Gene model Annotations
Consequence
minDistTSS
minDistTSE
cDNApos
relcDNApos
CDSpos
relCDSpos
protPos
relProtPos
Dst2Splice
Dst2SplType

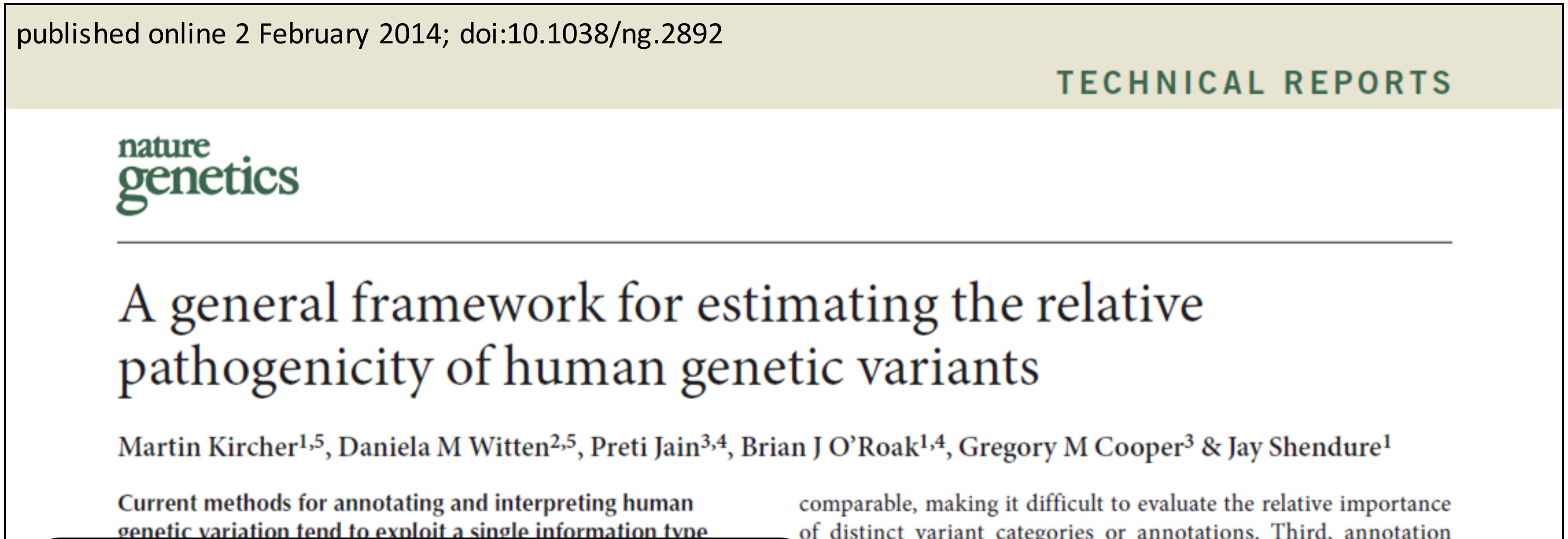
Functional Predictions
tOverlapMotifs
motifDist
motifECount
motifEName
motifEHIPos
motifEScoreChng
TFBS
TFBSPeaks
TFBSPeaksMax
Segway

Combined variant scores

- CADD/DANN: http://cadd.gs.washington.edu
- DeepSEA: http://deepsea.princeton.edu
- Eigen: http://www.columbia.edu/~ii2135/download.html
- FATHMM-MKL: http://fathmm.biocompute.org.uk/
- FunSeq2: http://funseq2.gersteinlab.org/
- GAWAVA: ftp://ftp.sanger.ac.uk/pub/resources/software/gwava/v1.0/VEP_plugin/
- ReMM: https://charite.github.io/software-remm-score.html
- LINSIGHT: http://compgen.cshl.edu/~yihuang/LINSIGHT/

…

Combined variant scores: CADD (1)



published online 2 February 2014; doi:10.1038/ng.2892

TECHNICAL REPORTS

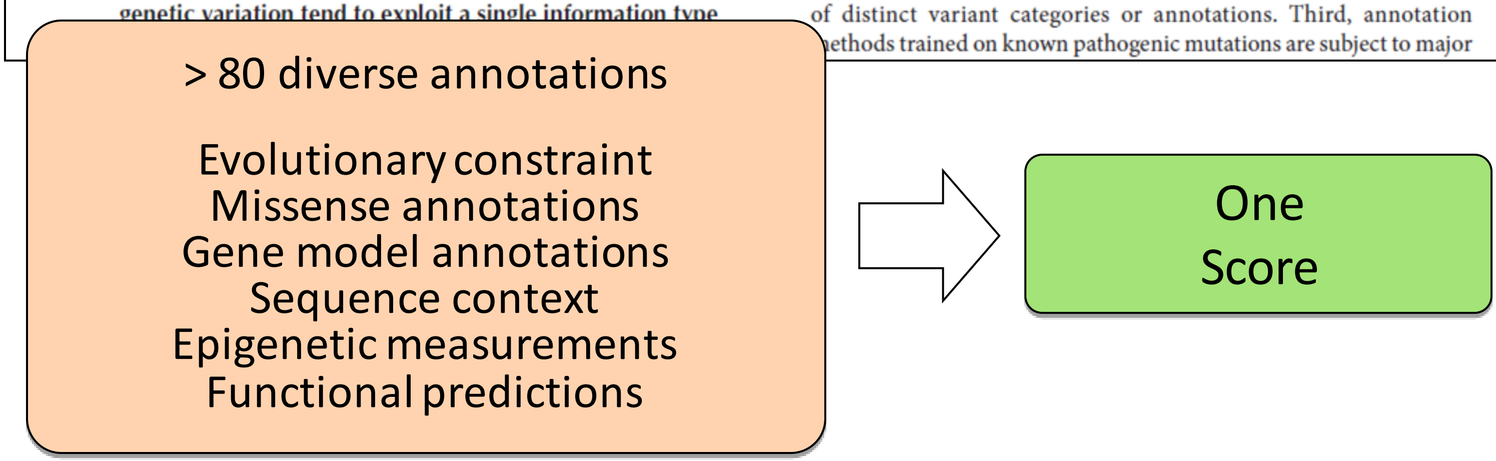
nature genetics

A general framework for estimating the relative pathogenicity of human genetic variants

Martin Kircher[1,5], Daniela M Witten[2,5], Preti Jain[3,4], Brian J O'Roak[1,4], Gregory M Cooper[3] & Jay Shendure[1]

Current methods for annotating and interpreting human genetic variation tend to exploit a single information type

comparable, making it difficult to evaluate the relative importance of distinct variant categories or annotations. Third, annotation methods trained on known pathogenic mutations are subject to major



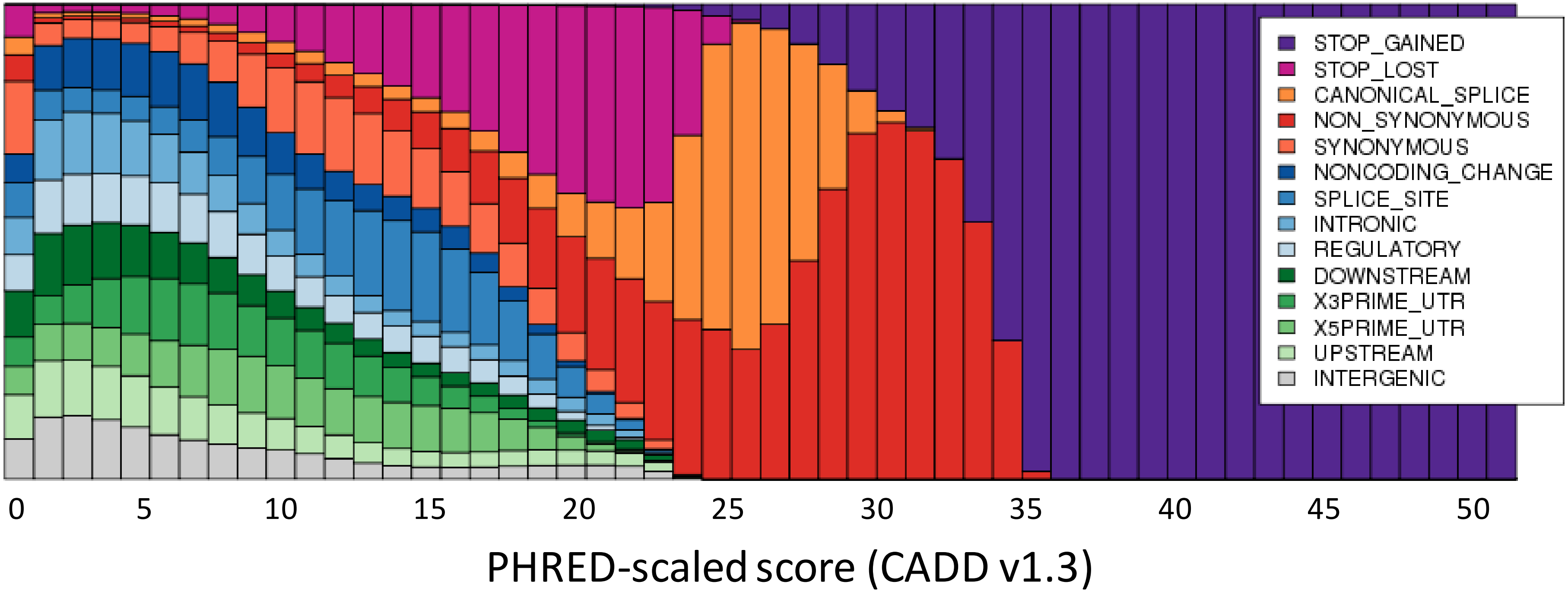
> 80 diverse annotations

Evolutionary constraint
Missense annotations
Gene model annotations
Sequence context
Epigenetic measurements
Functional predictions

One
Score

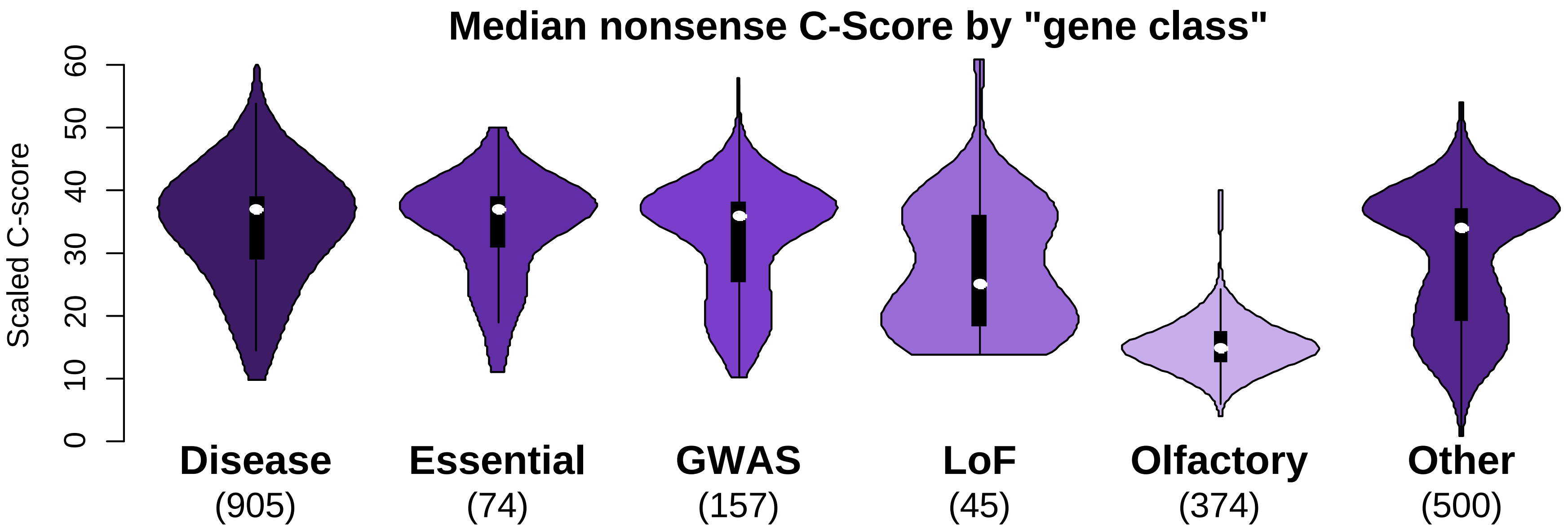
Combined variant scores: CADD (2)

- All variants are ranked relative to all nine billion possible substitutions in the human genome
- Median scores by categories are inline with common hierarchies



Combined variant scores: CADD (3)

- Scores provide resolution across and within functional categories

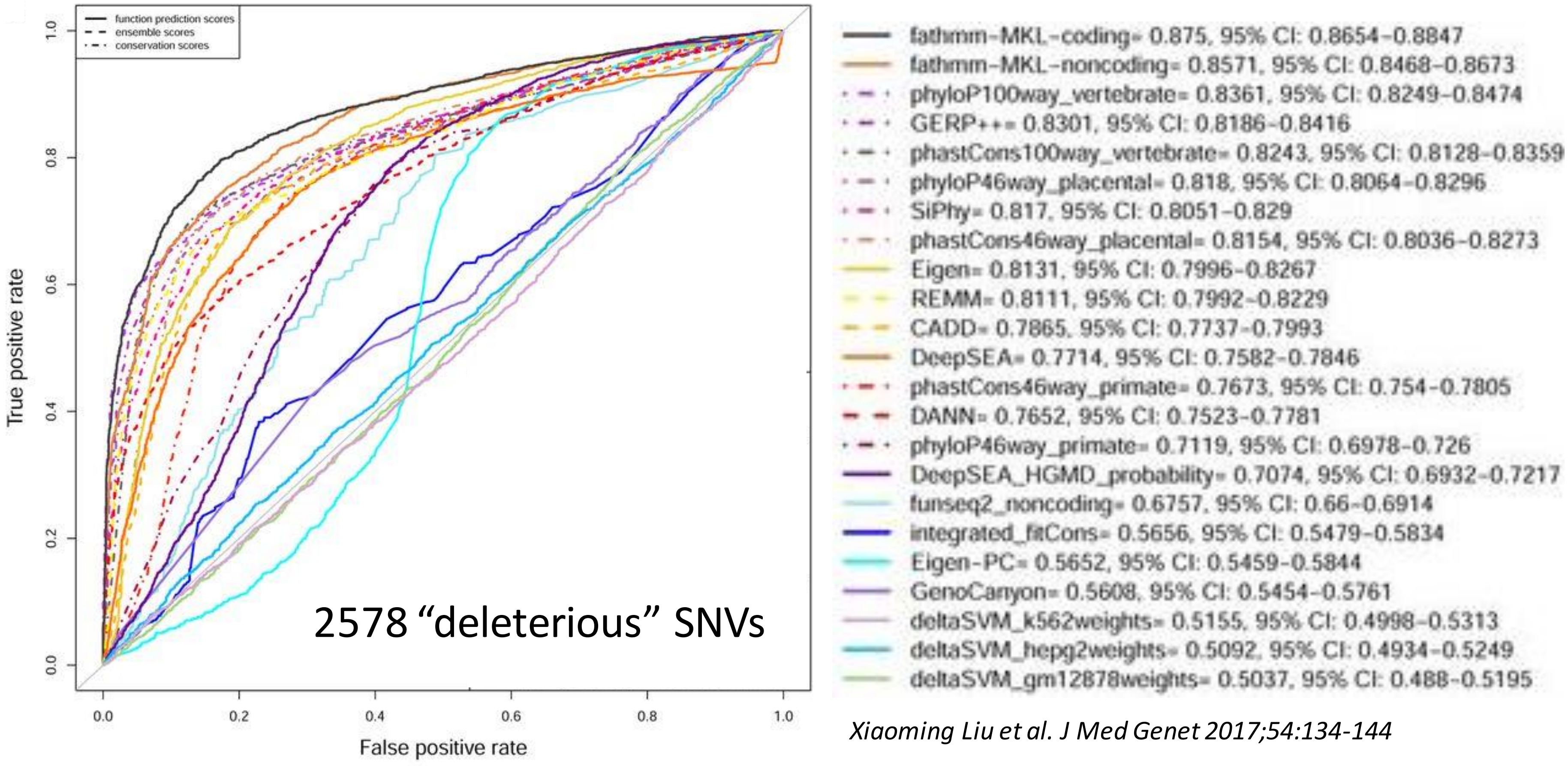


doi:10.1038/ng.2892

How well does it work for non-coding disease variants?

BERLIN INSTITUTE OF HEALTH
Charité & Max Delbrück Center

Performance on non-coding variants in HGMD database8 V.2015.4



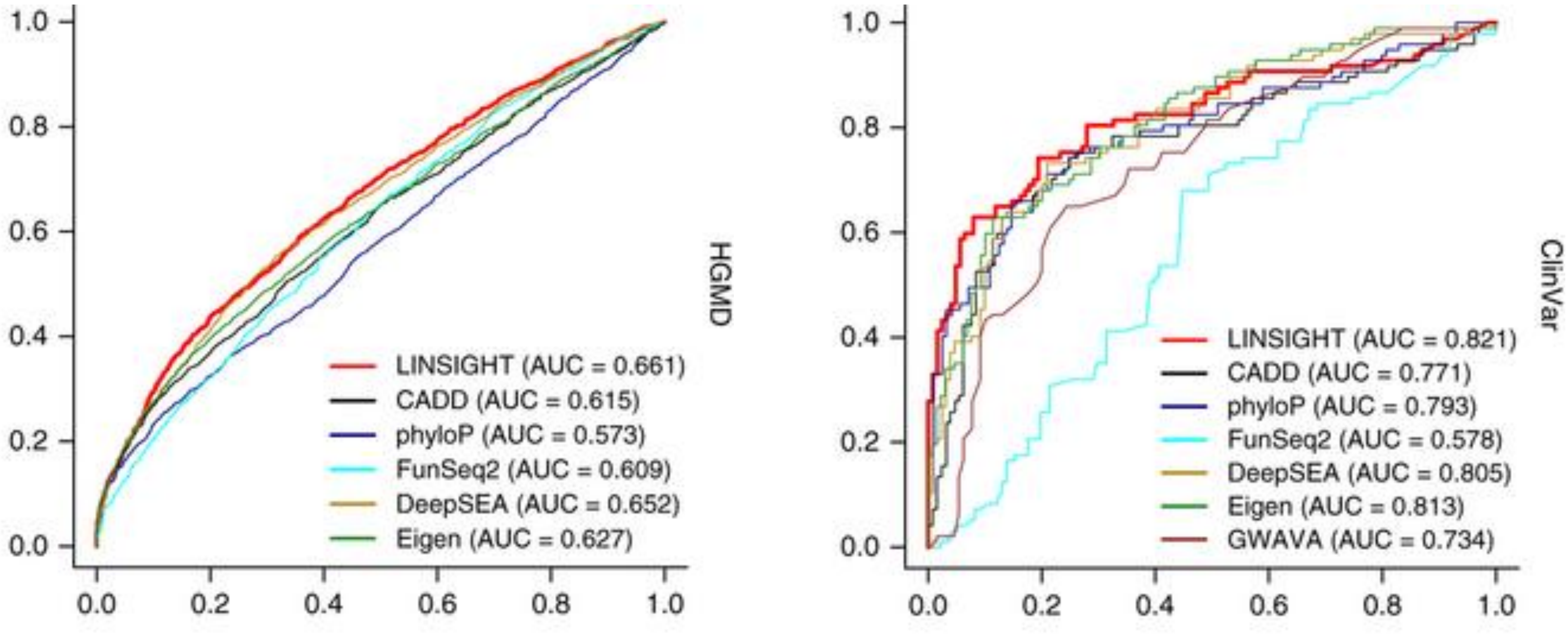
2578 “deleterious” SNVs

Xiaoming Liu et al. J Med Genet 2017;54:134-144

Other data set, similar problems, other results…

BERLIN INSTITUTE OF HEALTH
Charité & Max Delbrück Center

More rigorously matched benign set and comparison between non-overlapping HGMD (n=1495) and ClinVar (n=101) sets



Huang YF et al. Nature Genetics 49, 618–624 (2017), DOI: 10.1038/ng.3810

Few known high quality non-coding mutations?

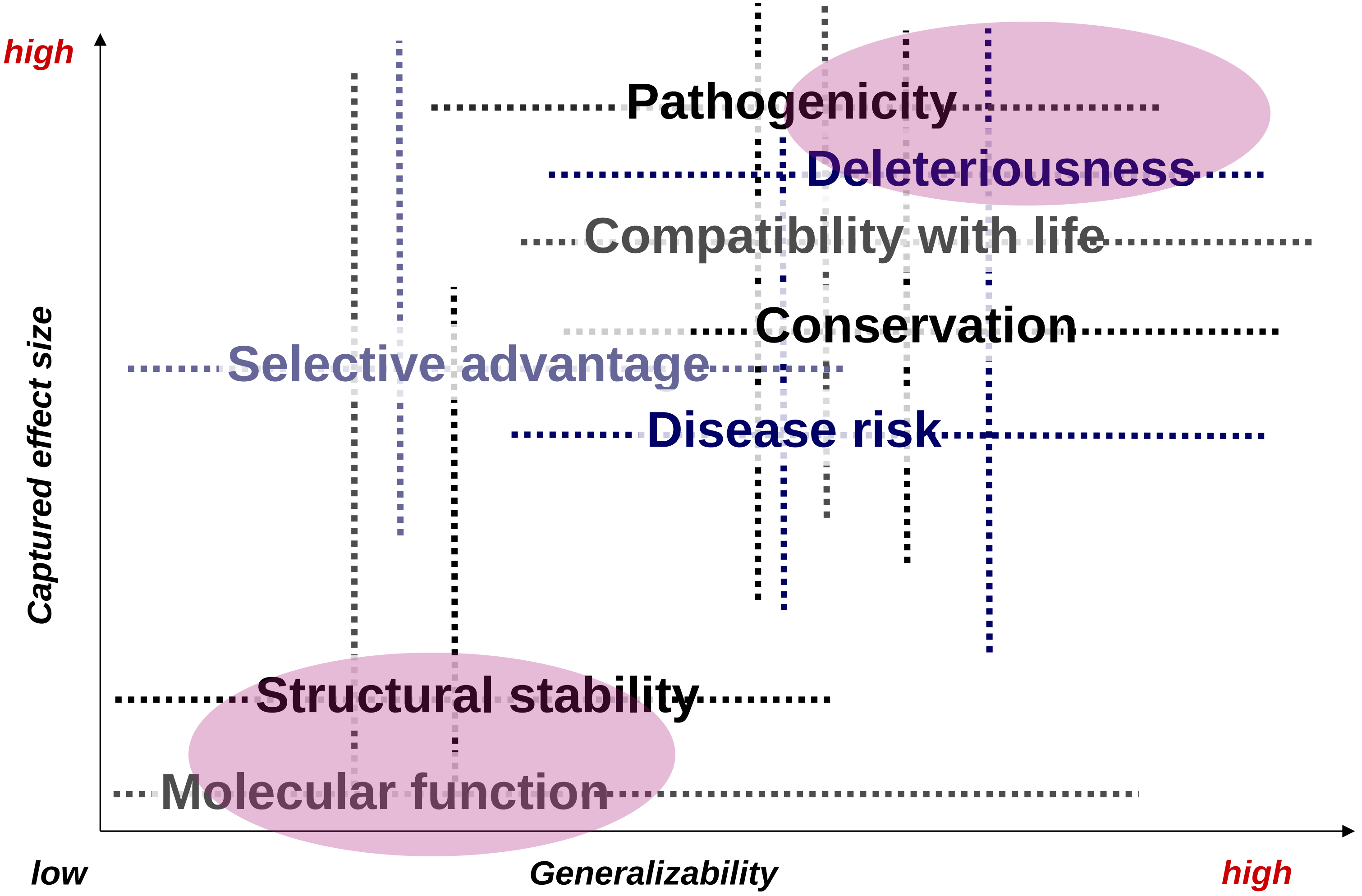
- Recent study used HGMD as well as literature research:

Category	Count
Enhancer	42
Promoter	142
5′ UTR	153
3′ UTR	43
Large non-coding RNA gene	65
MicroRNA gene	5
Imprinting control region	3
Total	***453***
Total single-nucleotide variants	***406***

Smedley D & Schubach M et al. AJHG 2016

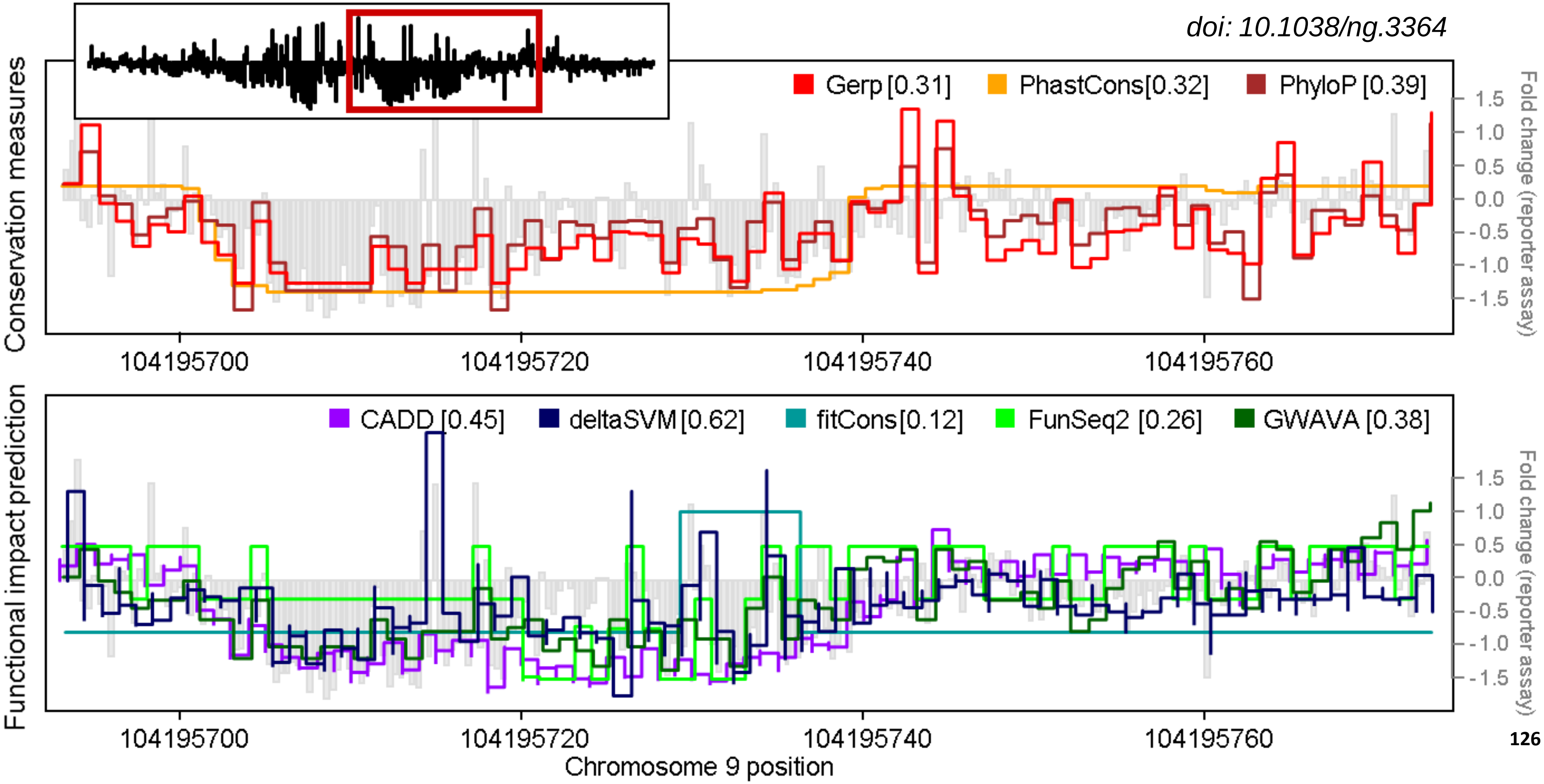
- Variants are clustered:
  - 142 promoter variants in 52 genes, 11 genes contribute 50%
  - 18 genes contribute 50% of all 338 promoter+UTR variants
  - 65 RNA gene mutations are in only 3 genes

Are we looking for the right effect size?



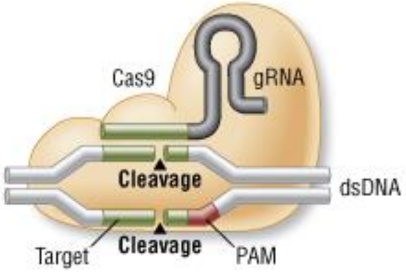
Expression effects & non-coding scores

Saturation mutagenesis of ALDOB enhancer (*Patwardhan et al, 2012*) correlated with measures of sequence conservation (*top*) and functional constraint/variant impact scores (*bottom*)

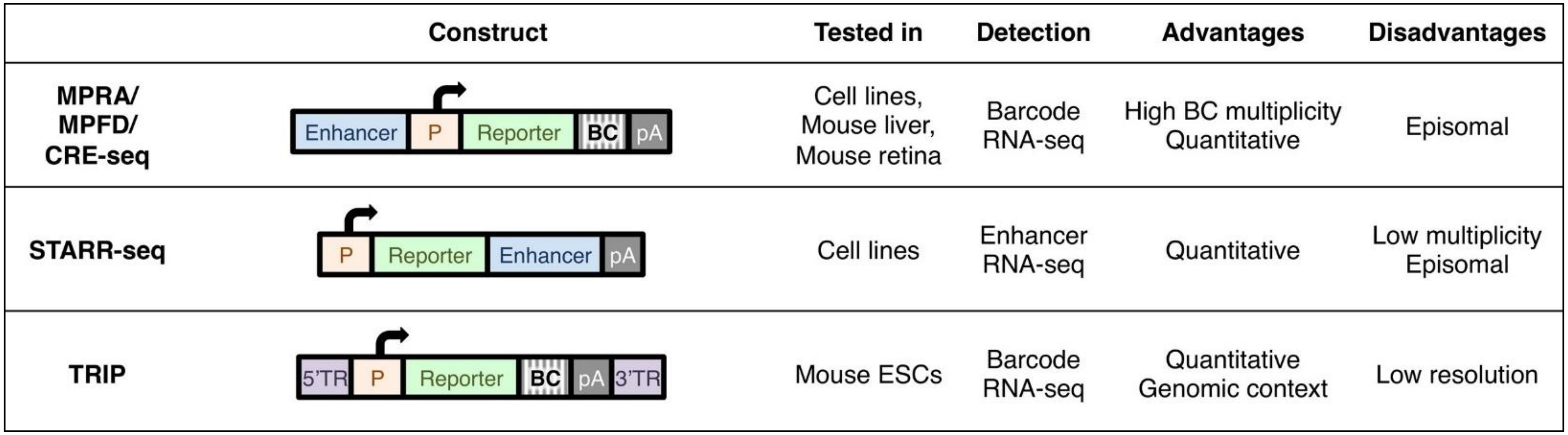


Can we obtain more non-coding variants from high-throughput assays?

- CRISPR/Cas9: mutation, deletion, activator/repressor screens, …
- MPRAs
  1. Dense read outs for mutations in select regions
  2. Test activity of regions (cataloging elements / learning rules)
  3. Large sets of readouts for genomically scattered mutations



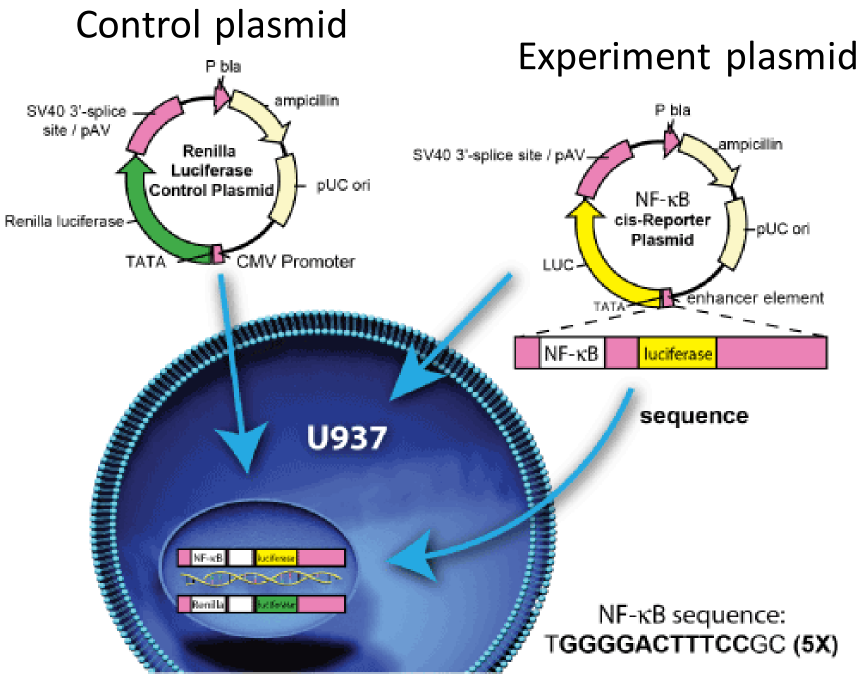
www.neb.com



	Construct	Tested in	Detection	Advantages	Disadvantages
MPRA/ MPFD/ CRE-seq	Enhancer | P | Reporter | BC | pA	Cell lines, Mouse liver, Mouse retina	Barcode RNA-seq	High BC multiplicity Quantitative	Episomal
STARR-seq	P | Reporter | Enhancer | pA	Cell lines	Enhancer RNA-seq	Quantitative	Low multiplicity Episomal
TRIP	5'TR | P | Reporter | BC | pA | 3'TR	Mouse ESCs	Barcode RNA-seq	Quantitative Genomic context	Low resolution

F. Inoue & N. Ahituv, Genomics 106(3), September 2015, Pages 159-164

Background: reporter assays

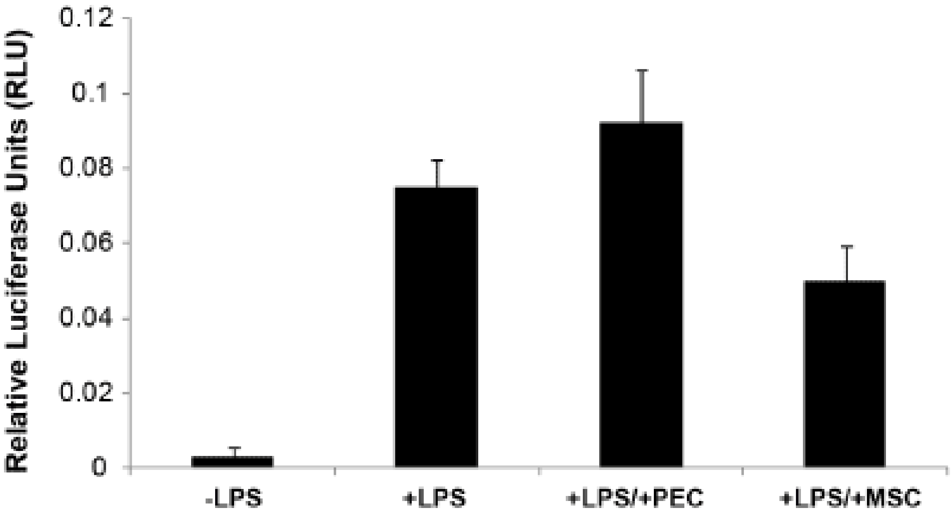


Control plasmid

Experiment plasmid

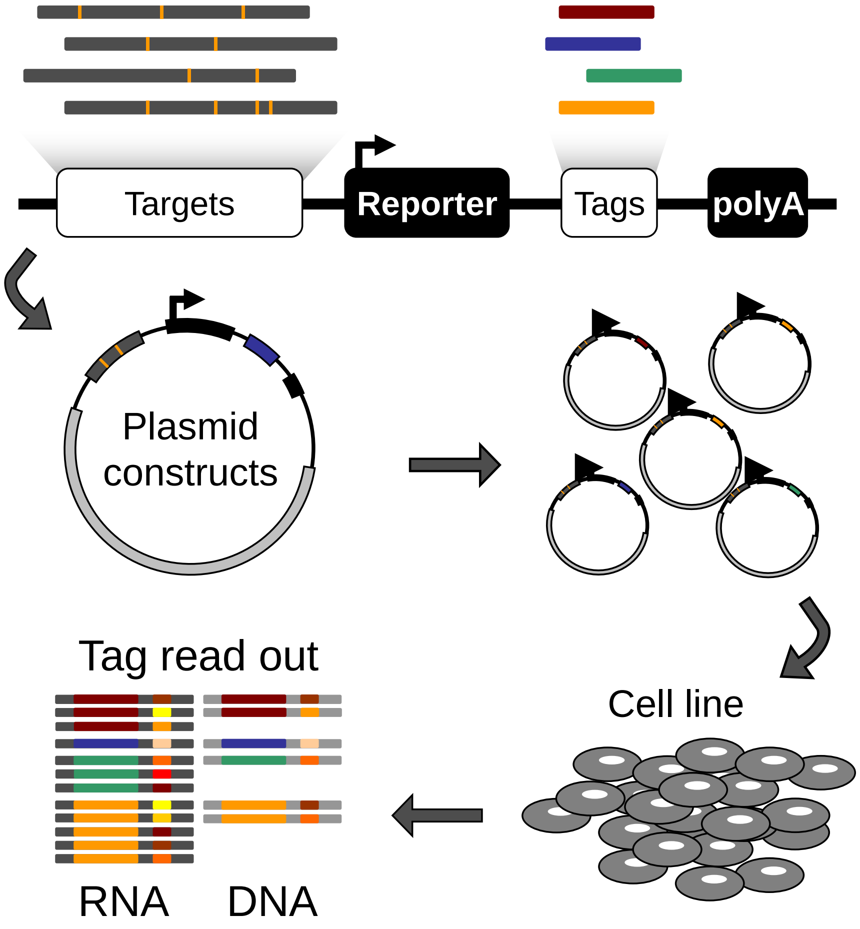
NF-κB sequence:
TGGGGACTTTCCGC (5X)

https://www.omicsonline.org/ articles-images/2157-7552-S3-001-g004.html



Massively parallel reporter assays (MPRA)

- Generate sequence variants
- Integrate plasmid or lenti library containing tag sequences
- Learn association between tags and sequence variants
- Express in cell line and collect RNA & DNA to read-out tags
- Analyze RNA/DNA ratio



TERT promoter

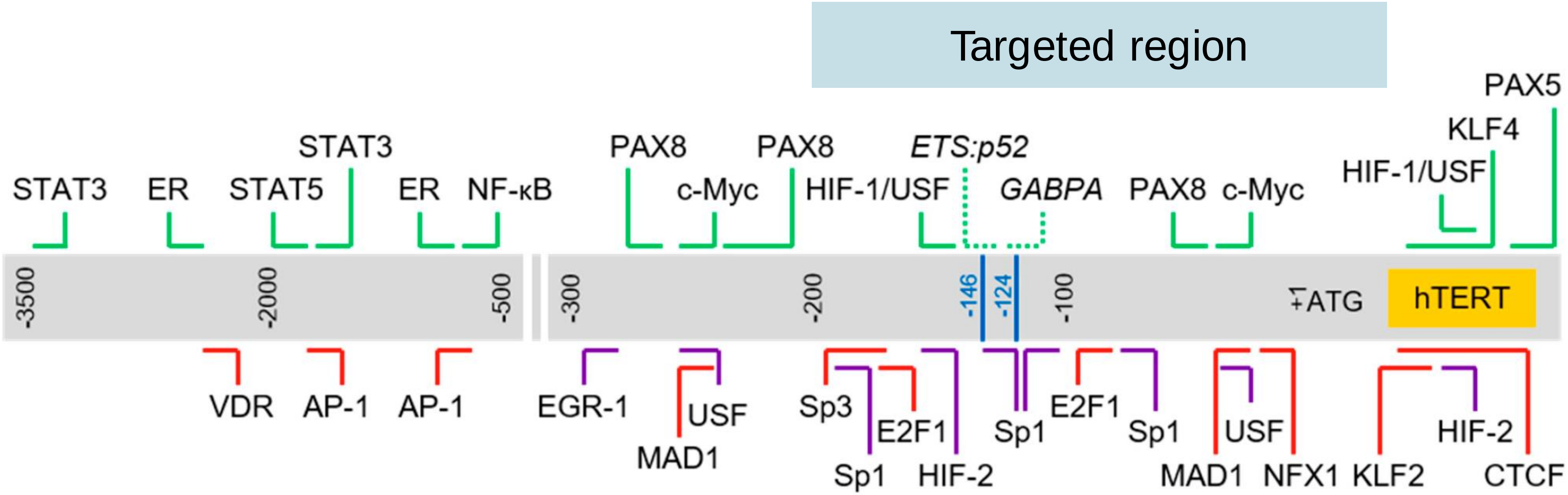


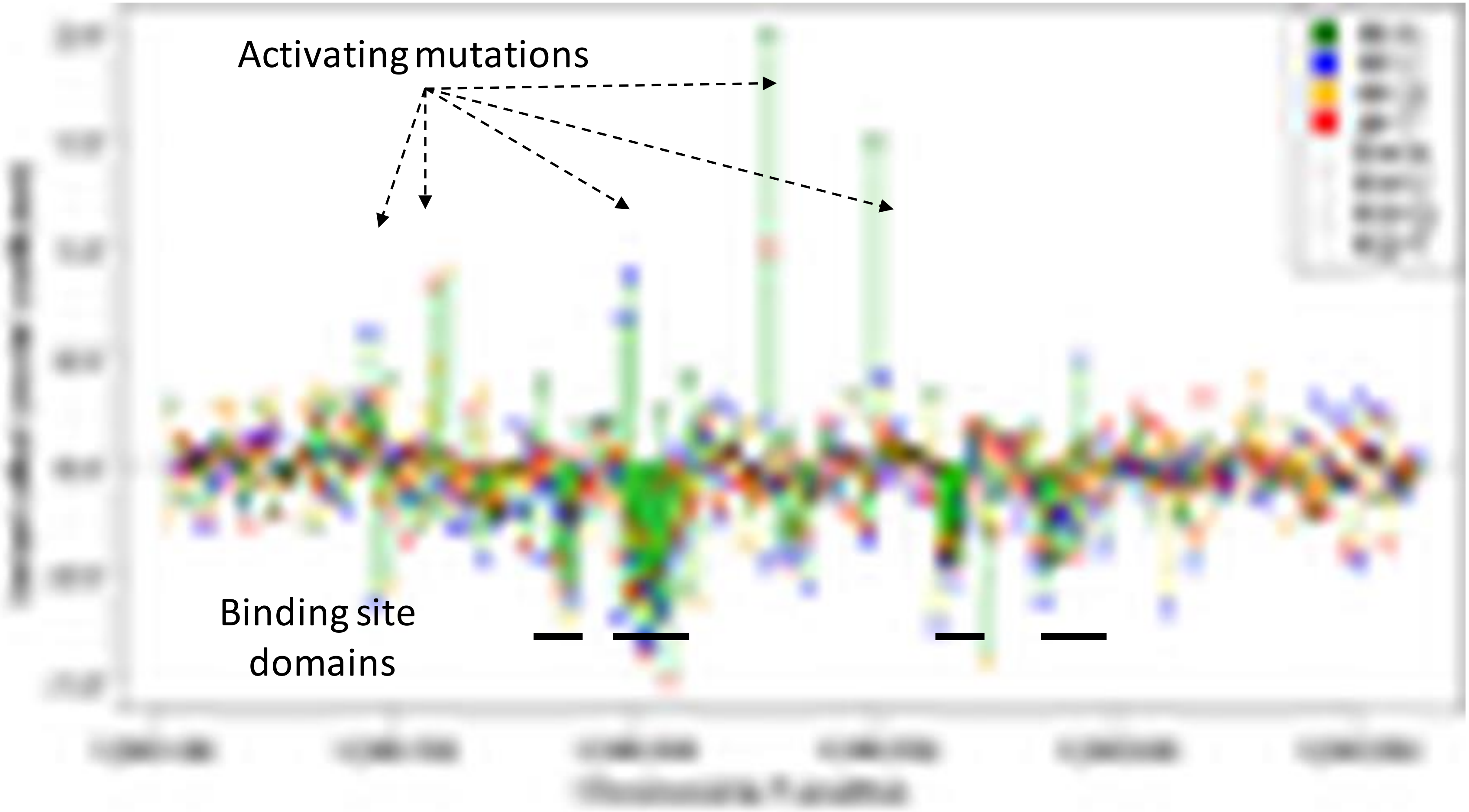
Figure 1. Schematic of transcription factor binding sites in human Telomerase Reverse Transcriptase (*hTERT*) promoter. Chromosomal sequence extending from 3.5 kb upstream and 150 bp downstream of *hTERT* translation start site (+1) is represented by the gray box. Horizontal lines above and below the box indicate approximate binding sites of respective transcription factors. Blue lines: hotspot promoter mutations ("-124" corresponds to C228T mutation; "-146" corresponds to C250T mutation); green: activator; red: repressor; purple: regulator with dual roles; dotted line: regulator bound to sites created by hotspot mutations.

Genes 2016, 7(8), 50; doi:10.3390/genes7080050

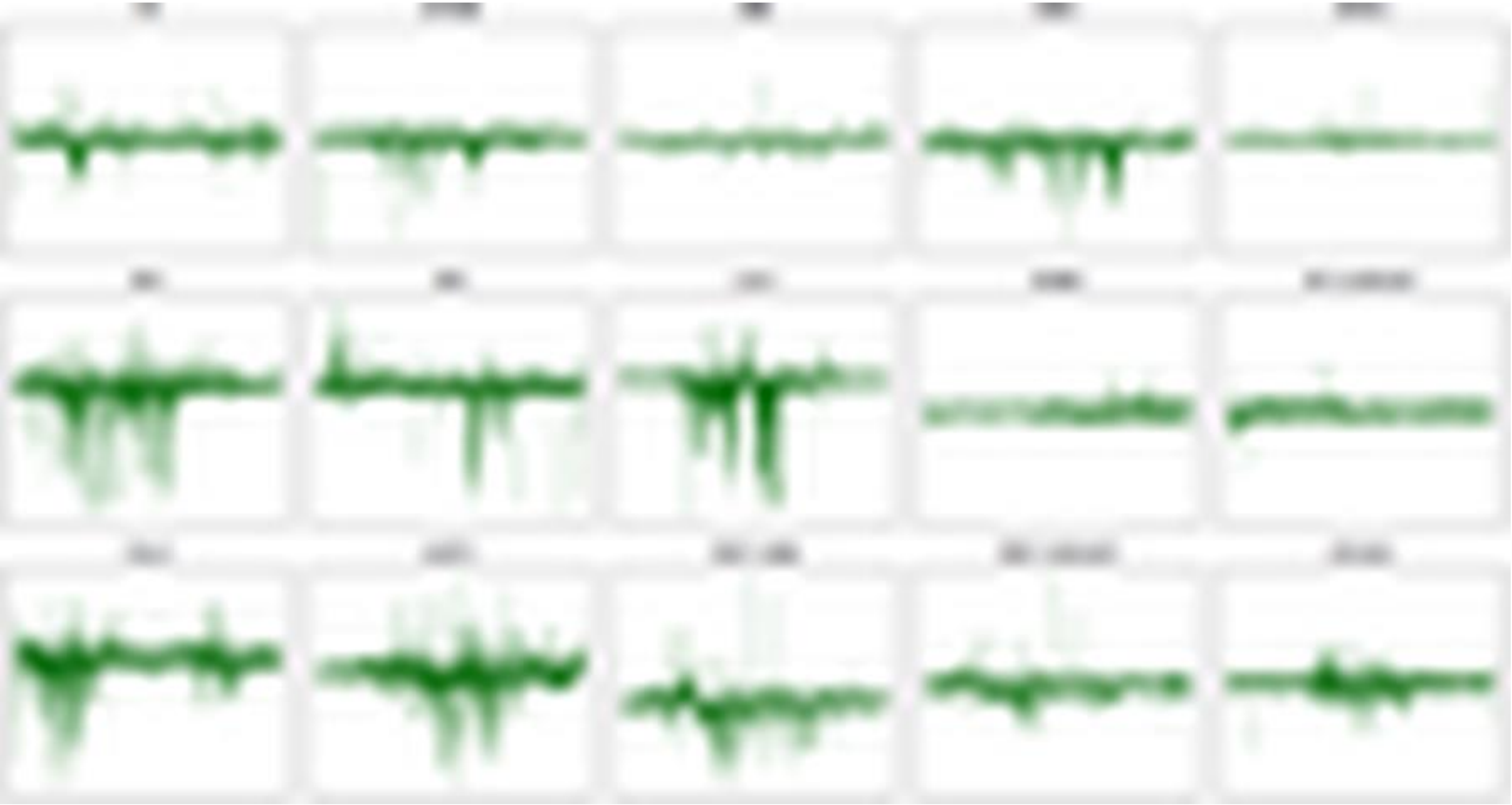
Saturation mutagenesis of TERT promoter in HEK293T

BERLIN INSTITUTE OF HEALTH
Charité & Max Delbrück Center

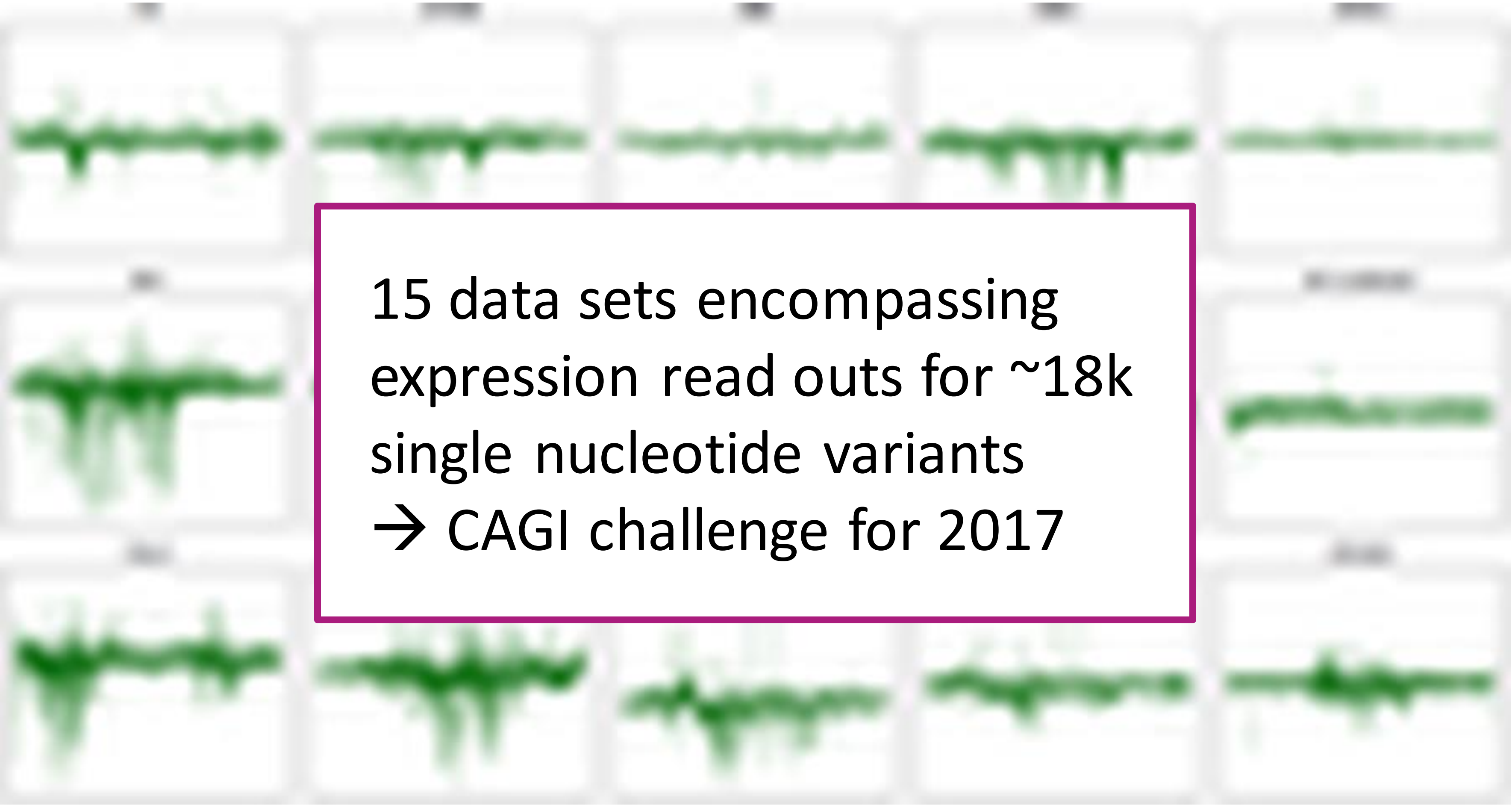
TERT (259bp)



New saturation mutagenesis data sets



New saturation mutagenesis data sets

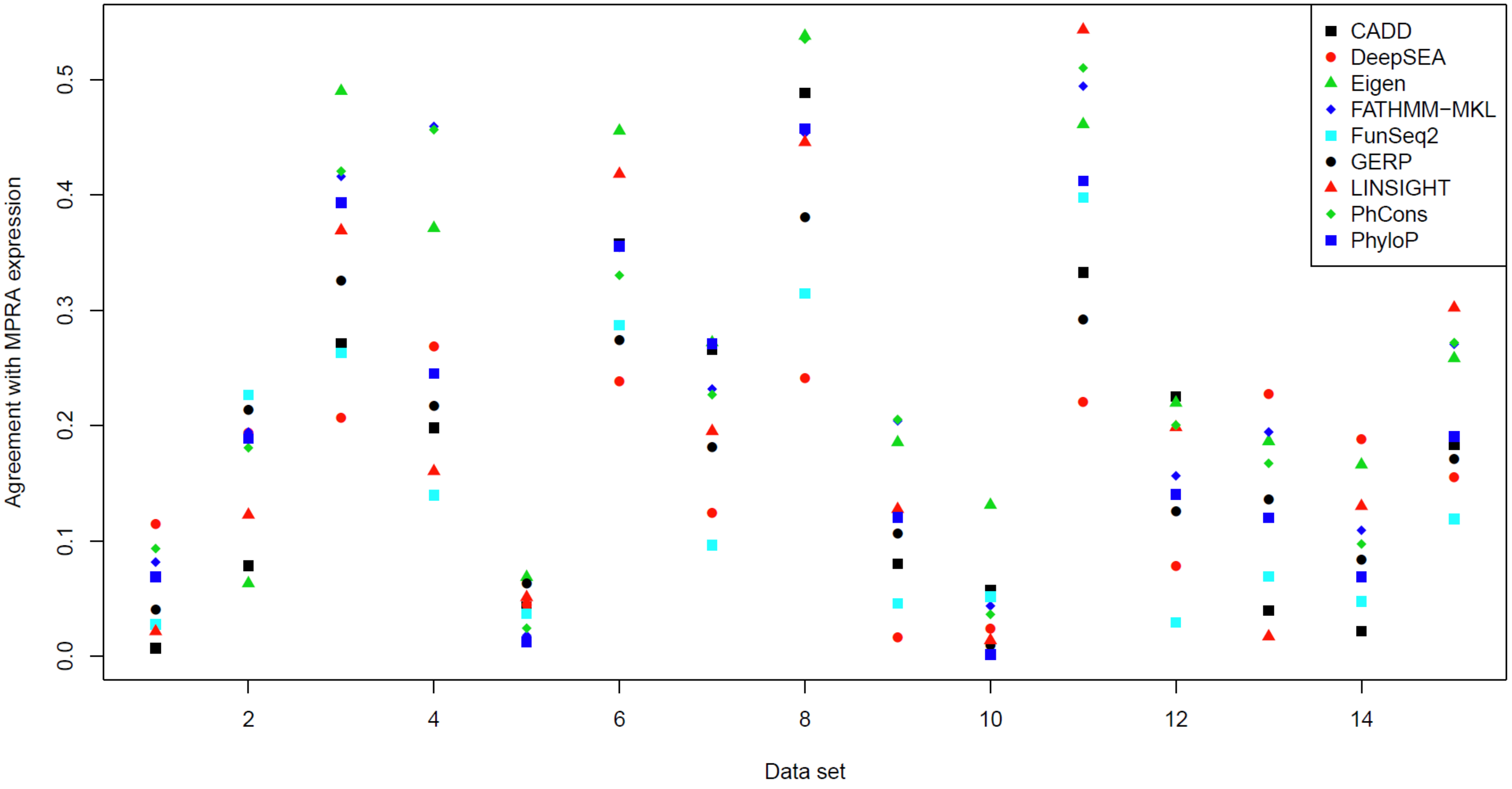


15 data sets encompassing expression read outs for ~18k single nucleotide variants

→ CAGI challenge for 2017

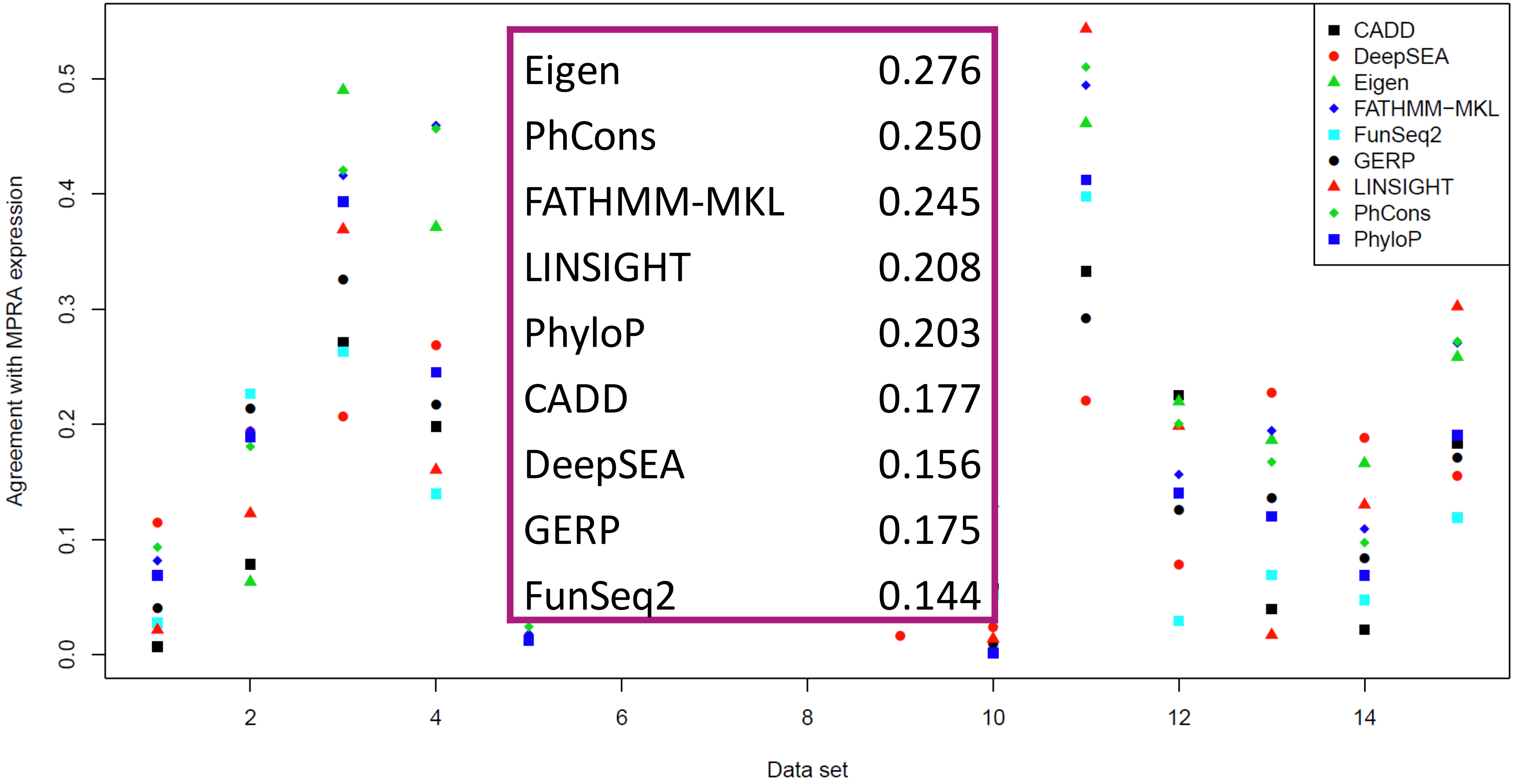
What about the variant scores?

BERLIN
INSTITUTE
OF HEALTH
Charité & Max Delbrück Center



What about the variant scores?

BERLIN INSTITUTE OF HEALTH
Charité & Max Delbrück Center



Eigen	0.276
PhCons	0.250
FATHMM-MKL	0.245
LINSIGHT	0.208
PhyloP	0.203
CADD	0.177
DeepSEA	0.156
GERP	0.175
FunSeq2	0.144

How should I consider regulatory mutations in my projects for now?

BERLIN INSTITUTE OF HEALTH
Charité & Max Delbrück Center

1. **Use available element annotations**
   - Enhancer, Promoter annotations, e.g.
     - Ensembl Regulatory Build:

       ftp://ftp.ensembl.org/pub/current_regulation/homo_sapiens/RegulatoryFeatureActivity/
     - Epigenomics RoadMap:

       http://egg2.wustl.edu/roadmap/web_portal/predict_reg_motif.html#predicting_reg
     - Fantom5: http://enhancer.binf.ku.dk/presets/
   - DHS sites, e.g. http://egg2.wustl.edu/roadmap/web_portal/DNase_reg.html#delieation
   - Segmentation (e.g. Epigenomics RoadMap)
2. **Use available combined scores within these elements**

QUESTIONS AND PARTICIPANT FEEDBACK

THANK YOU!

BERLIN INSTITUTE OF HEALTH
Charité & Max Delbrück Center

CONTACT

Malte Spielmann	spielman@uw.edu
Martin Kircher	martin.kircher@bihealth.de
Dominik Seelow	dominik.seelow@charite.de