MutationTaster & RegulationSpotter

Pathogenicity Prediction of Sequence Variants: Past, Present and Future

Dr. rer. nat. Jana Marie Schwarz

Klinik für Pädiatrie m. S. Neurologie
Exzellenzcluster NeuroCure
Charité Universitätsmedizin Berlin

AG Translational Genomics

07.11.2017

AG Translational Genomics

Focus on rare (mendelian) diseases in children

- muscular, neurological and neuro-muscular system

Next Generation Sequencing (NGS)

- software development for easy and comprehensive data analysis

Interdisciplinary group

- close collaboration between clinic and research
- clinicians, biochemists, biologists, bioinformaticians

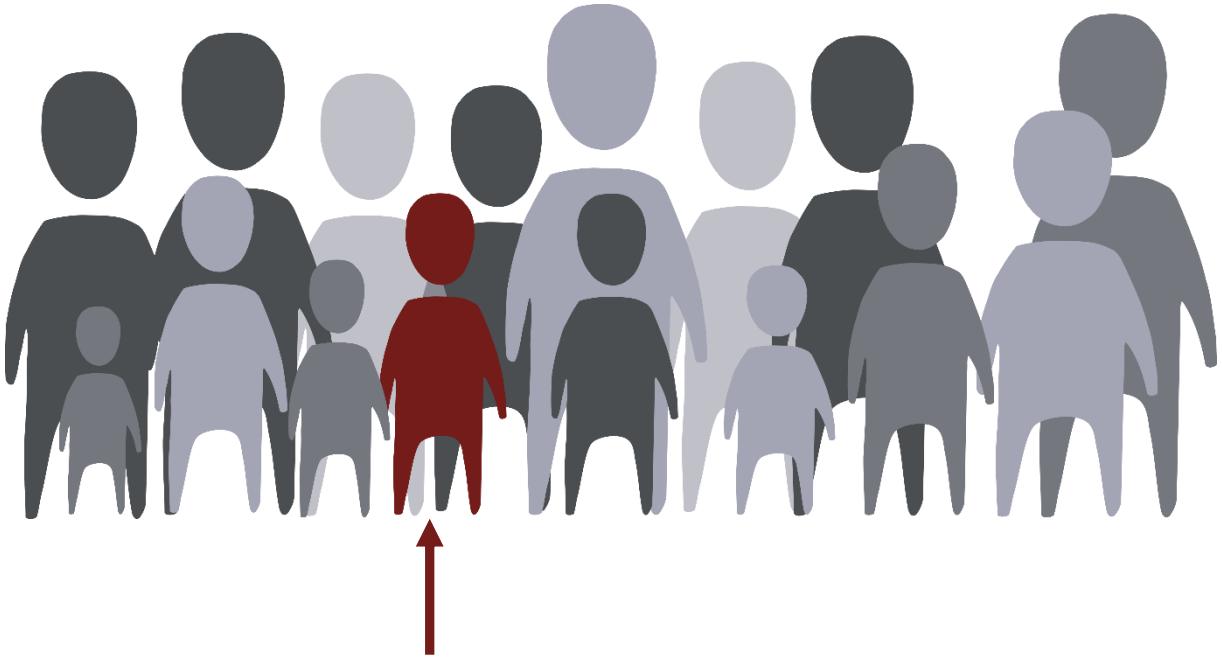
Follow-up functional studies in the wet-lab

CHARITÉ
UNIVERSITÄTSMEDIZIN BERLIN

BERLIN INSTITUTE OF HEALTH
Charité & Max Delbrück Center

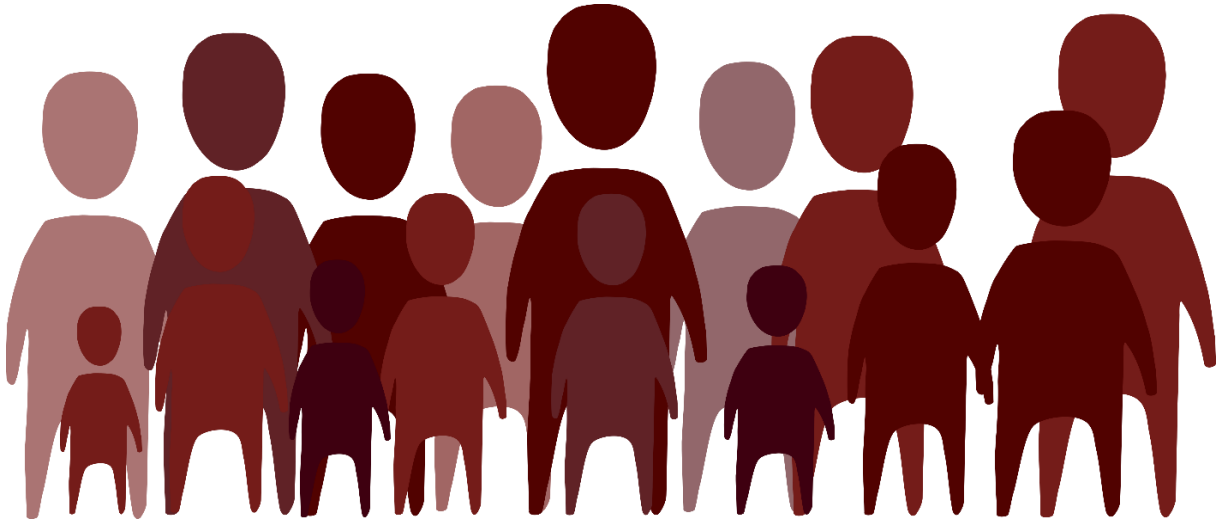
NEUROCURE
Exzellenzcluster

Rare diseases



Less than **1** affected in 2.000 individuals

Rare diseases

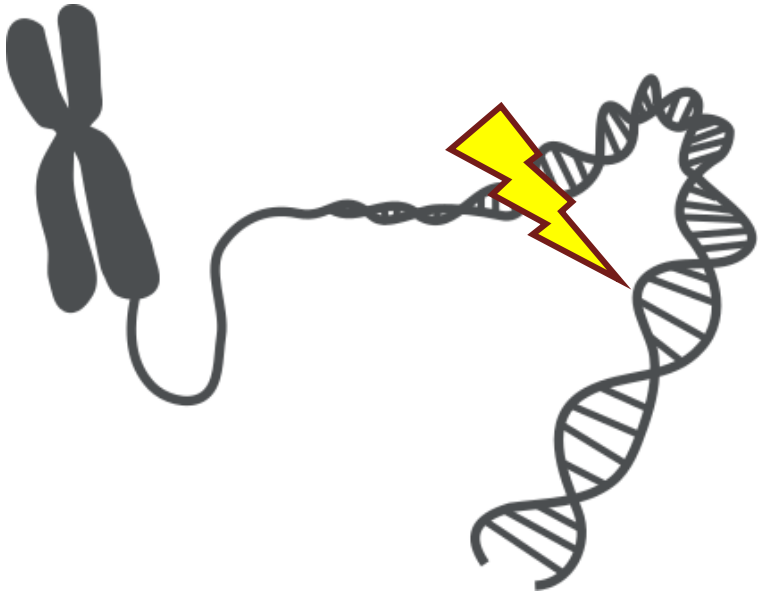


But: **7000 – 8000** rare diseases are known.

>> World-wide more than **300 Millions** patients

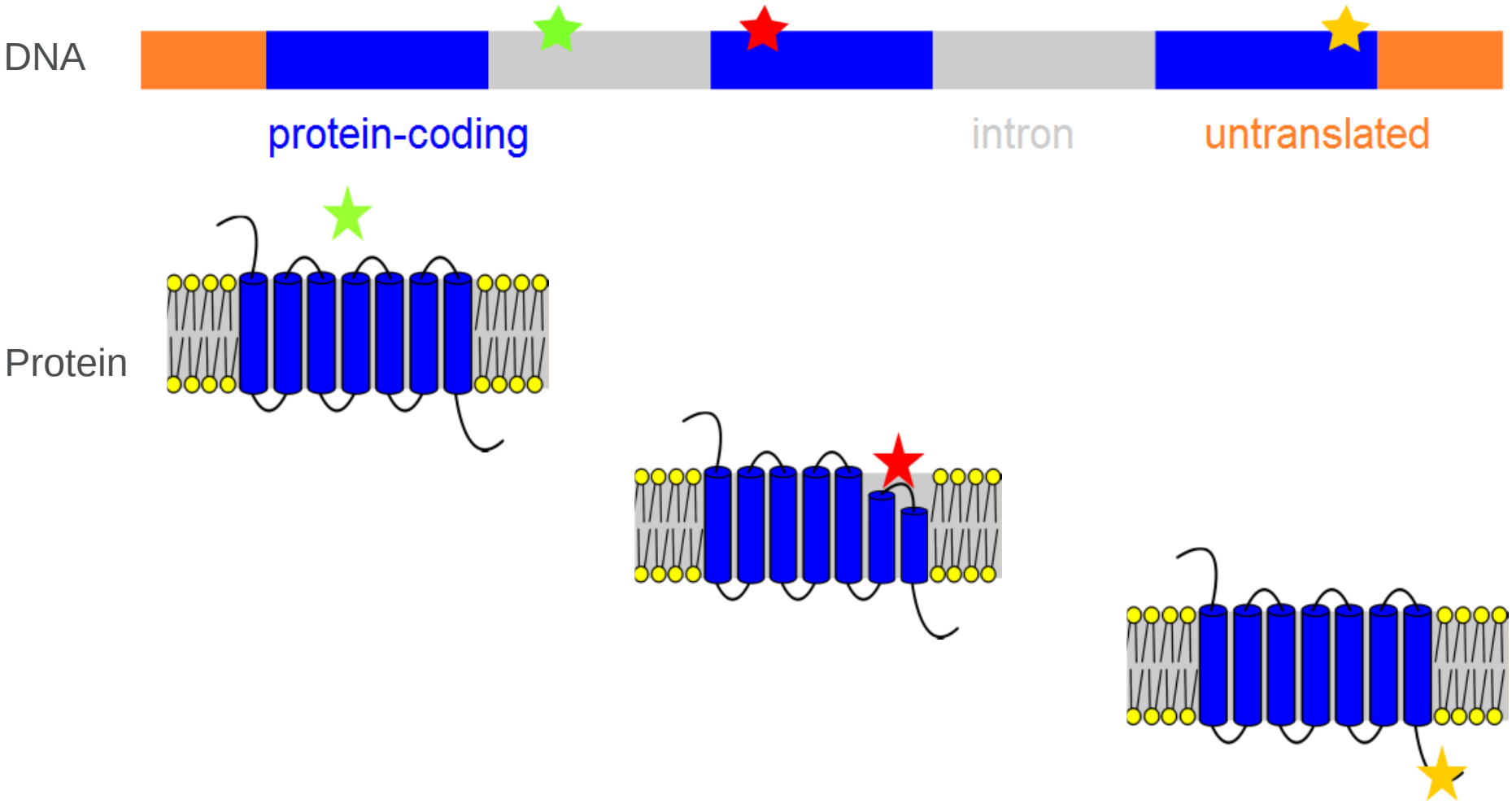
Rare diseases

80% are genetically caused.



Mutations in the genome lead to a disease.

Rare diseases – genetically caused.

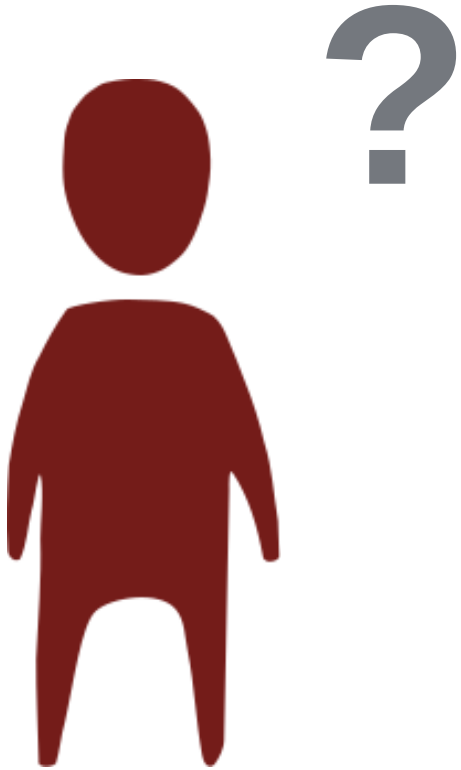


Rare diseases

Severe symptoms occur congenital or in early childhood.

Rare diseases

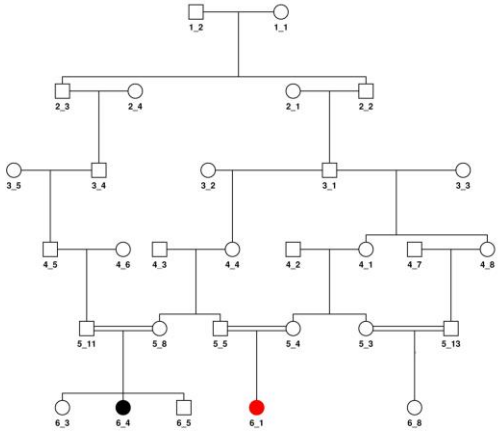
Diagnostics of rare diseases are often difficult and time-intensive



Many patients don‘t get a clear molecular diagnosis or no diagnosis at all.

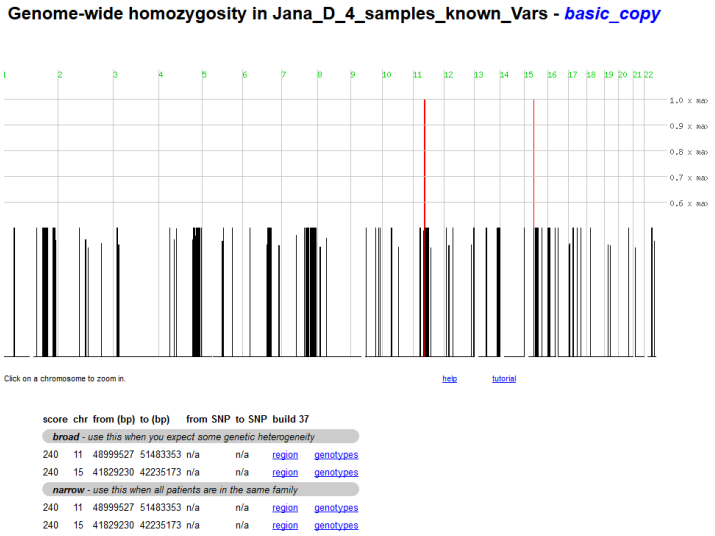
The case of Family D.

- huge, consanguinous family from libanon
- two affected cousins
- development initially normal
- gait ataxia, distal muscle atrophy and developmental speech delays at the age of 3-5 yrs
- neurography: combined demyelinating / axonal neuropathy
- progredient course
- both patients meanwhile died



Strategies to find the disease mutation

Homozygosity mapping

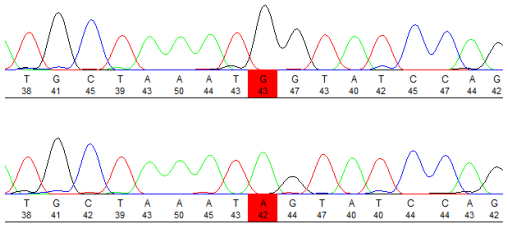


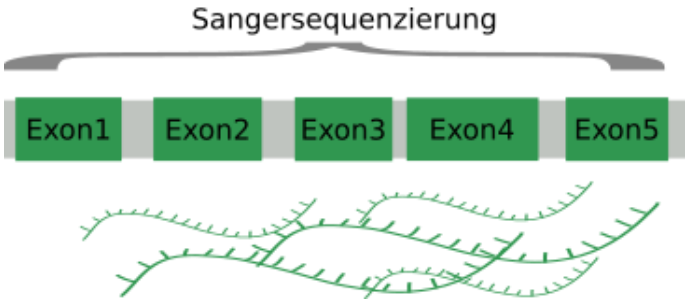
score	chr	from (bp)	to (bp)	from SNP	to SNP	build 37	
broad *- use this when you expect some genetic heterogeneity*							
240	11	48999527	51483353	n/a	n/a	region	genotypes
240	15	41829230	42235173	n/a	n/a	region	genotypes
narrow *- use this when all patients are in the same family*							
240	11	48999527	51483353	n/a	n/a	region	genotypes
240	15	41829230	42235173	n/a	n/a	region	genotypes

Strategies to find the disease mutation

Homozygosity mapping

Sequencing of single candidate genes





Questions in 2009

- How to sequence many genes at a time?

Next Generation Sequencing (NGS)

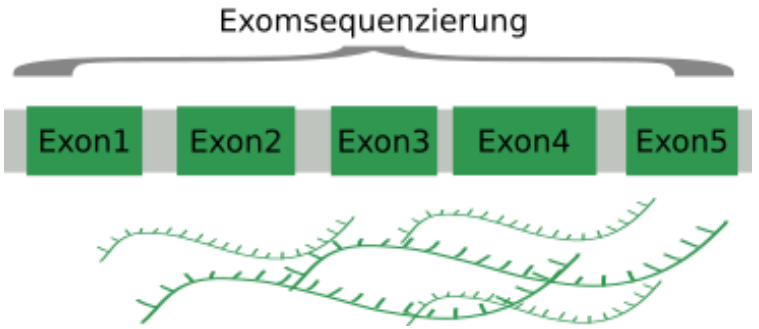
- massive parallel sequencing
- detection of tens of thousands of variants

Strategies to find the disease mutation

Homozygosity mapping

Sequencing of single candidate genes

Whole Exome Sequencing



1 variant / 1.000bp > 30.000 variants / exome

Questions in 2009

- How to sequence many genes at a time?

- How to automatically evaluate the disease potential of found DNA variants?

http://www.mutationtaster.org/

Schwarz JM, Roedelsperger C, Schuelke M, Seelow D:
MutationTaster evaluates disease-causing potential of sequence alterations.
Nat Methods. 2010 Aug;7(8):575-6. doi: 10.1038/nmeth0810-575.

mutation t@sting

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

Gene SOD1 HGNC gene symbol, NCBI Gene ID, Ensembl gene ID show available transcripts clear input

Transcript ENST00000270142 Ensembl transcript ID

Choose the transcript:

- ENST00000470944 (*processed_transcript*, 1746 bases)
- ENST00000270142 (*protein_coding*, 966 bases) NM_000454
- ENST00000389995 (*protein_coding*, 865 bases)
- ENST00000476106 (*processed_transcript*, 586 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

options

☐ show nucleotide alignment

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)

Name of alteration SOD1_ALS if you would like to have a name for this alteration in the output later on, please type in here

continue

If you use MutationTaster, please cite our publication: Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. **MutationTaster evaluates disease-causing potential of sequence alterations.** *Nat Methods. 2010 Aug;7(8):575-6.*

Current build: NCBI 37 / Ensembl 69

mutation t@sting

documentation

Alteration SOD1_ALS

Prediction **disease causing** Model: *simple_aae*, prob: 0.999999999908447 (classification due to ClinVar, real probability is shown anyway) (explain)

Summary

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

analysed issue	analysis result
name of alteration	SOD1_ALS
alteration (phys. location)	chr21:33036170A>G
HGNC symbol	SOD1
Ensembl transcript ID	ENST00000270142
UniProt peptide	P00441
alteration type	single base exchange
alteration region	CDS
DNA changes	c.140A>G cDNA.288A>G g.4236A>G
AA changes	H47R Score: 0.79 explain score(s)
position(s) of altered AA if AA alteration in CDS	47
frameshift	no
dbSNP / TGP / HGMD(public) / ClinVar	**known disease mutation: rs121912443 (pathogenic for *Amyotrophic lateral sclerosis type 1*)** dbSNP NCBI variation viewer known disease mutation at this position, please check HGMD for details (HGMD ID CM930680)
regulatory features	H3K9me1, Histone, Histone 3 Lysine 9 mono-methylation H3K36me3, Histone, Histone 3 Lysine 36 Tri-Methylation H3K79me2, Histone, Histone 3 Lysine 79 di-methylation H4K20me1, Histone, Histone 4 Lysine 20 mono-methylation
phyloP / phastCons	5.162, 1 (flanking) 4.283, 0.996

phyloP / phastCons	5.162, 1 (flanking) 4.283, 0.996 -2.023, 0.059 (flanking) explain score(s) and/or inspect your position(s) in in UCSC Genome Browser

splice sites

effect	gDNA position	score	wt detection sequence	exon-intron border
Acc increased	4243	wt: 0.31 / mu: 0.44	wt: CTGCATGGATTCCATGTTCATGAGTTTGGAGATAATACAGC mu: CTGCATGGATTCCGTGTTCATGAGTTTGGAGATAATACAGC	tcat|GAGT
Acc marginally increased	4245	wt: 0.8500 / mu: 0.9286 (marginal change - not scored)	wt: GCATGGATTCCATGTTCATGAGTTTGGAGATAATACAGCAG mu: GCATGGATTCCGTGTTCATGAGTTTGGAGATAATACAGCAG	atga|GTTT

Kozak consensus sequence altered? N/A

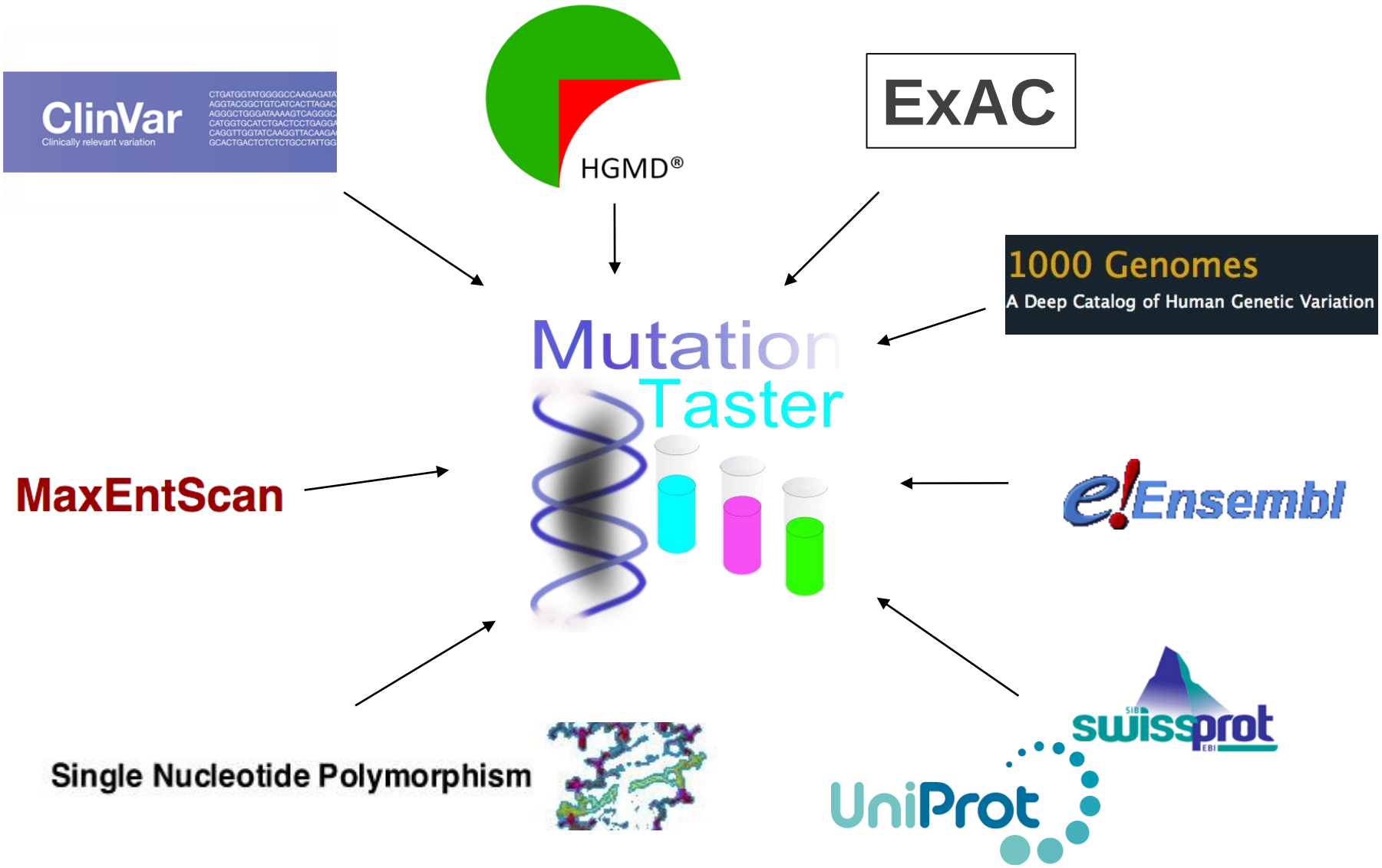
conservation
protein level for non-synonymous changes

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
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	48 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
Fcatus	no homologue		
Dmelanogaster	all identical	FBgn0003462	47 V C G L A K G L H G F H V H E F G D N T N G C
Celegans	no homologue		
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
49	49	METAL	Copper; catalytic.	might get lost (downstream of altered splice site)
51	52	MUTAGEN	FG->EE: Abolishes dimerization; when associated with Q-134.	might get lost (downstream of altered splice site)
54	56	HELIX		might get lost (downstream of altered splice site)
57	61	HELIX		might get lost (downstream of altered splice site)
58	58	MUTAGEN	C->S: Enhances formation of fibrillar aggregates in the absence of bound zinc; when associated with S-7; S-112 and S- 147.	might get lost (downstream of altered splice site)
58	58	DISULFID		might get lost (downstream of altered splice site)
63	65	STRAND		might get lost (downstream of altered splice site)

Integrated data (examples)



General tests – entire transcript

test	**data source / software**
conservation (DNA)	BLASTN / phyloP / phastCons
splicing	MaxEntScan
occurence in data bases - known as disease mutation?	NCBI ClinVar / HGMD
occurence in data bases - known as polymorphism?	dbSNP / 1000G / ExAC
regulatory elements	Ensembl Regulation

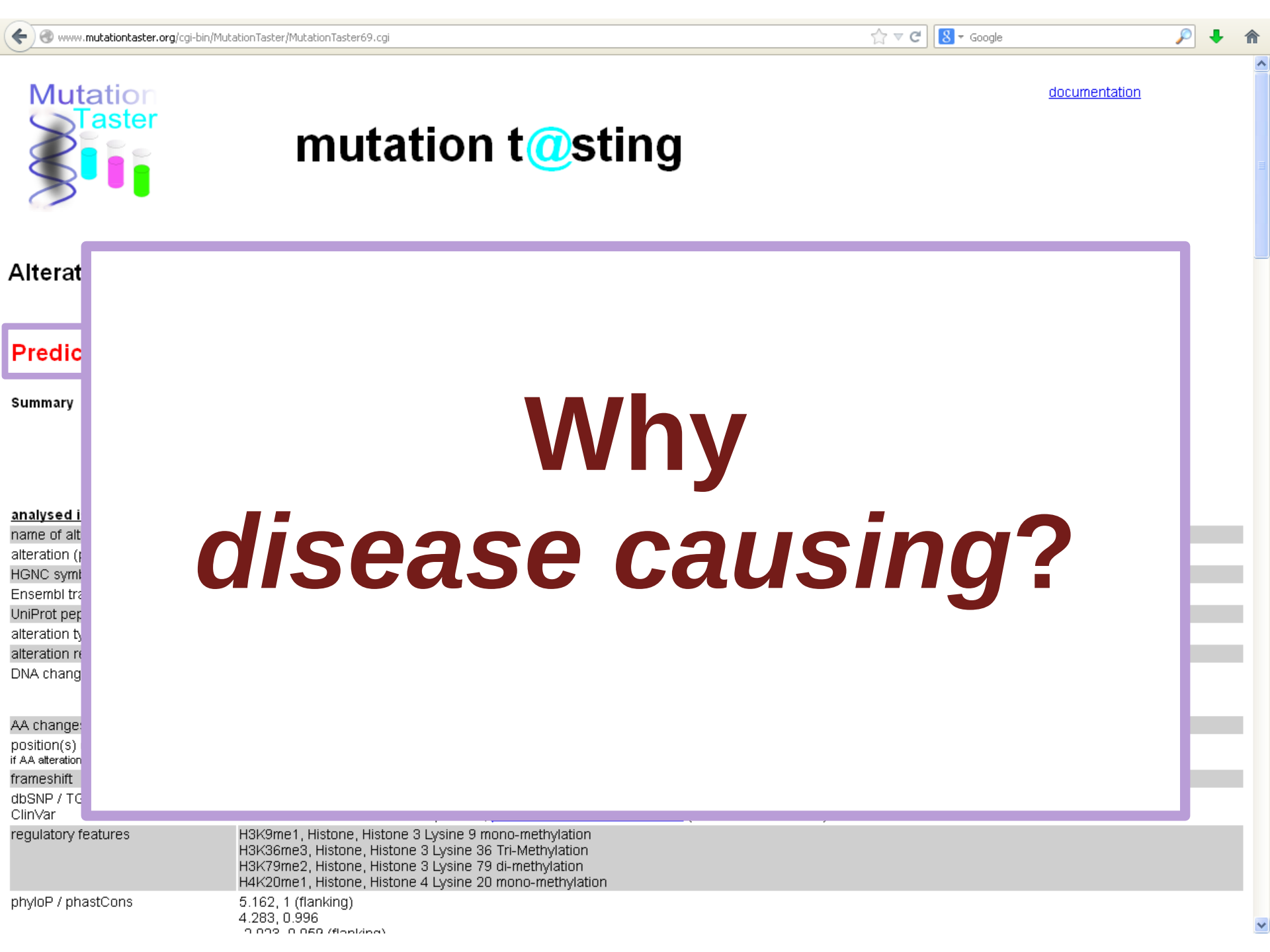
HGMD = Human Gene Mutation Database; 1000G = 1000 Genomes Project

Tests for coding sequence (CDS)

test	data / software
amino acid exchange	MutationTaster
conservation (protein)	BLASTP
protein features	UniProtKB / SwissProt
frameshift	MutationTaster
length of protein	MutationTaster

Tests for untranslated regions (UTR)

region	test	data source / software
5‘UTR	Kozak consensus sequence	MutationTaster
3‘UTR	Poly(A)-Signal	polyadq


www.mutationtaster.org/cgi-bin/MutationTaster/MutationTaster69.cgi
Google
Mutation Taster
mutation t@sting
documentation
Alterat
Predic
Summary
analysed i
name of alt
alteration (
HGNC symb
Ensembl tra
UniProt pep
alteration ty
alteration re
DNA chang
AA change
position(s)
if AA alteration
frameshift
dbSNP / TG
ClinVar
regulatory features
H3K9me1, Histone, Histone 3 Lysine 9 mono-methylation
H3K36me3, Histone, Histone 3 Lysine 36 Tri-Methylation
H3K79me2, Histone, Histone 3 Lysine 79 di-methylation
H4K20me1, Histone, Histone 4 Lysine 20 mono-methylation
phyloP / phastCons
5.162, 1 (flanking)
4.283, 0.996
Why
disease causing?

Computer based classification

- „either-or“ decision

- automated classification by Bayes classifier

- **simple handling**
- **CPU/RAM friendly**
- **fast, robust results**

Bayes classifier in MutationTaster

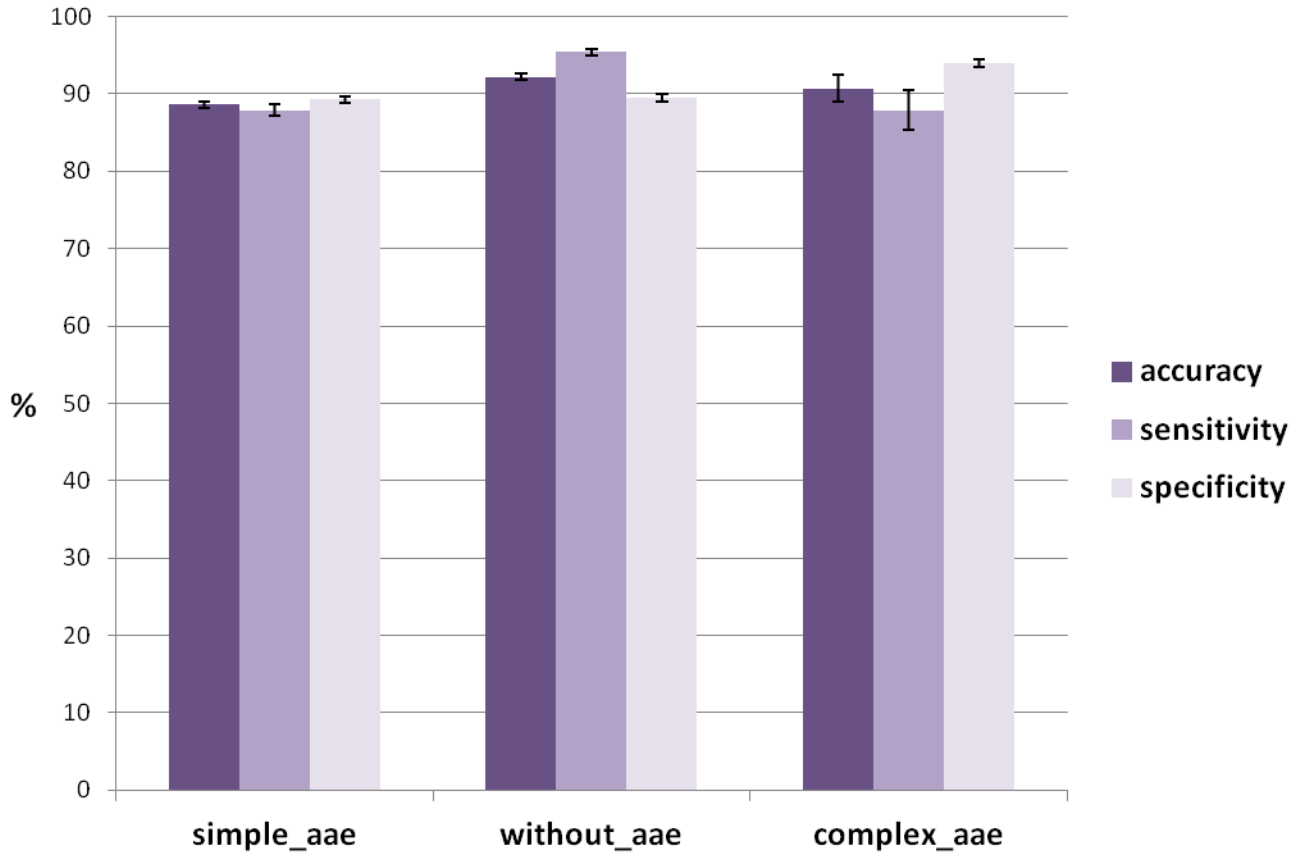
- training with

  a) > 100,000 known disease mutations
  Human Gene Mutation Database (HGMD)

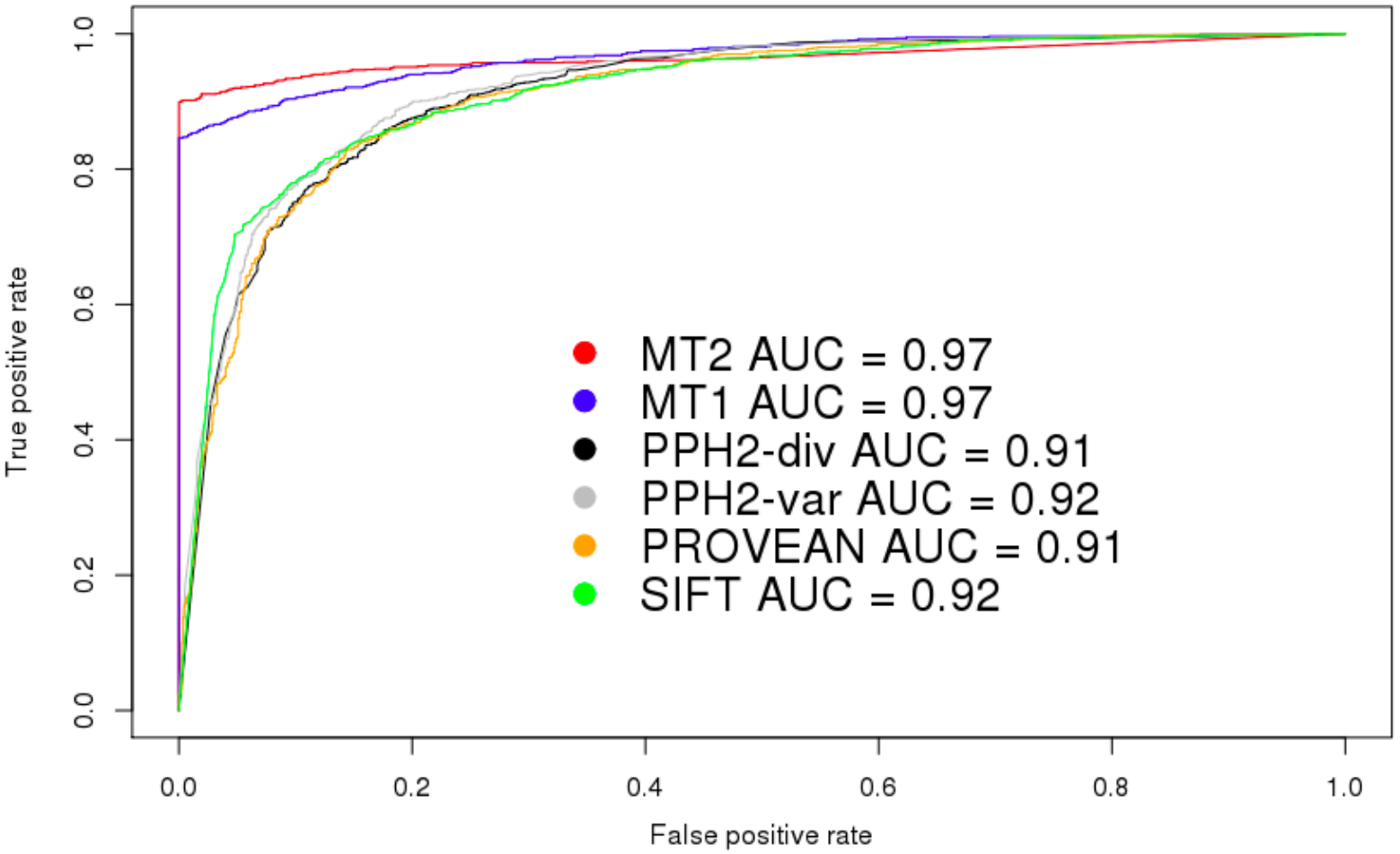
  b) > 6,000,000 harmless polymorphisms
  1000 Genomes Project (1000G)

- classification model:
  stores frequencies of different values / test results

5-fold cross validation: results



Comparison with similar programs



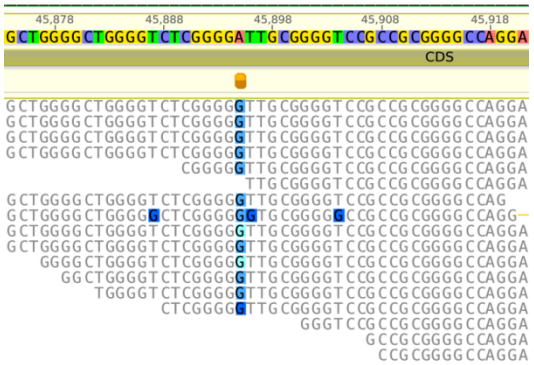
test set: 1100 polymorphisms from *1000 Genomes Project*
1100 disease mutations from *Human Gene Mutation Database* (HGMD)

abbreviations: *PPH2-var* = PolyPhen2 HumVar model, *PPH2-div* = PolyPhen2 HumDiv model

MutationTaster's, automatic classifications based on public variant data from 1000G and ClinVar were ignored for this comparison.

NGS data and MutationTaster

1) sequencing → *2) alignment* → *3) variant calling*



?

MutationTaster QueryEngine

mutation t@sting

QueryEngine

- documentation | FAQs
- single query
- query chromosomal positions
- QueryEngine
- MutationDistiller (public beta)
- RegulationSpotter (public beta)
- other applications | team
- slides ESHG2017 Copenhagen

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.

VCF file Durchsuchen... MTQE_example_vcf.vcf sample file clear input

Please zip or gzip large files!

Format:

```
#CHROM   POS    ID  REF  ALT  QUAL  FILTER  INFO                                                                  FORMAT       SAMPLE
chr1     10199  .   A    C    4.77  .       DP=2;AF1=0.5003;CI95=0.25,0.75;DP4=1,0,0,1;MQ=60;FQ=-3.1;PV4=1,1,1,1  GT:PL:DP:GQ  0/1:33,0,28:2:30
```

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

Project name Example

Unfortunately, E-Mail notification does currently not work. We are sorry for the inconvenience and try to solve the problem as soon as possible

E-mail address

generate HTML files (slower)

Analysis settings

search for homozygous variants yes

combine neighbouring variants yes

filter against TGP homozygous in 4 or more TGP samples

heterozygous in 20 or more TGP samples

minimum coverage 4

Submit

queue status 2017-09-21 17:42 CEST (refresh)

low load - our bored server is looking forward to analysing your data

0 jobs running, 0 queued, 1511.2 millions alterations analysed.

DB queries	free slots (0 running, 16 of 16 available - at 0% capacity)
large jobs	free slots (0 running, 20 of 20 available - at 0% capacity)
small jobs	free slots (0 running, 30 of 30 available - at 0% capacity)

- 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)

MTQE documentation

[MutationTaster] for Example - explore your results

highly-parallel fashion (500,000 variants / hour)

Statistics

submitted variants		4
pre-discarded variants		0
analysable alterations		4
discarded (TGP)		1
discarded (out of gene/exon/region)		0
alterations analysed		3
MT cases		9
type	without_aae	6
type	simple_aae	3
prediction	disease_causing_automatic	4
prediction	polymorphism	5
genes hit (except HapMap/TGP)		
...non-synonymous		2
...disease causing		2
...disease causing, non-synonymous		2

Analysis options

coverage threshold	4
filter out variants homozygous in TGP	4
>	

Display / filter / export results

sort & group

- (•) sort & group by prediction | model | gene symbol
- ( ) sort & group by prediction | model | gene symbol | variation
- ( ) *sort by these attributes*
  - 1: [ ] descending
  - 2: [ ] descending
  - 3: [ ] descending

hide

- [ ] silent alterations
- [ ] all predicted polymorphisms
- [ ] known polymorphisms
- [ ] prediction problems

show

- [ ] only homozygous alterations
- [ ] all recessive (homozygous, comp-het, genes with >1 disease-causing variant)
- [ ] only genes relevant for mitochondria

get the data

display (5000 results per page) export as TSV

Query GeneDistiller

call GeneDistiller for all...

- disease-causing mutations
- non-synonymous disease-causing mutations
- non-synonymous alterations (excl. known polymorphisms)

- see QueryEngine log for further details on the analysis run
- were discarded before mapping
- were filtered out (TGP)
- were not mapped to a transcript
- got no MutationTaster prediction

http://www.mutationtaster.org/

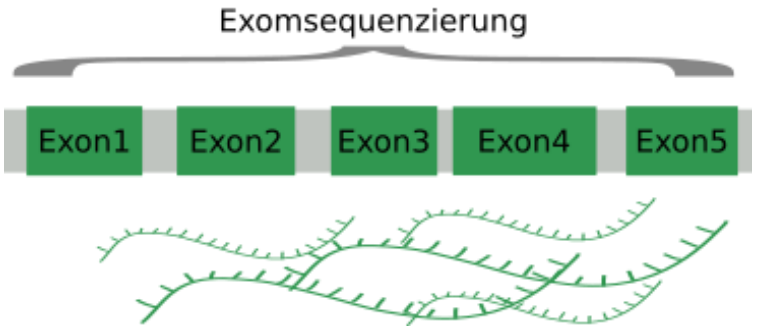
Schwarz JM, Cooper DN, Schuelke M, Seelow D:
MutationTaster2: mutation prediction for the deep-sequencing age.
Nature Methods (April 2014) 11(4):361-2.

Strategies to find the disease mutation

Homozygosity mapping

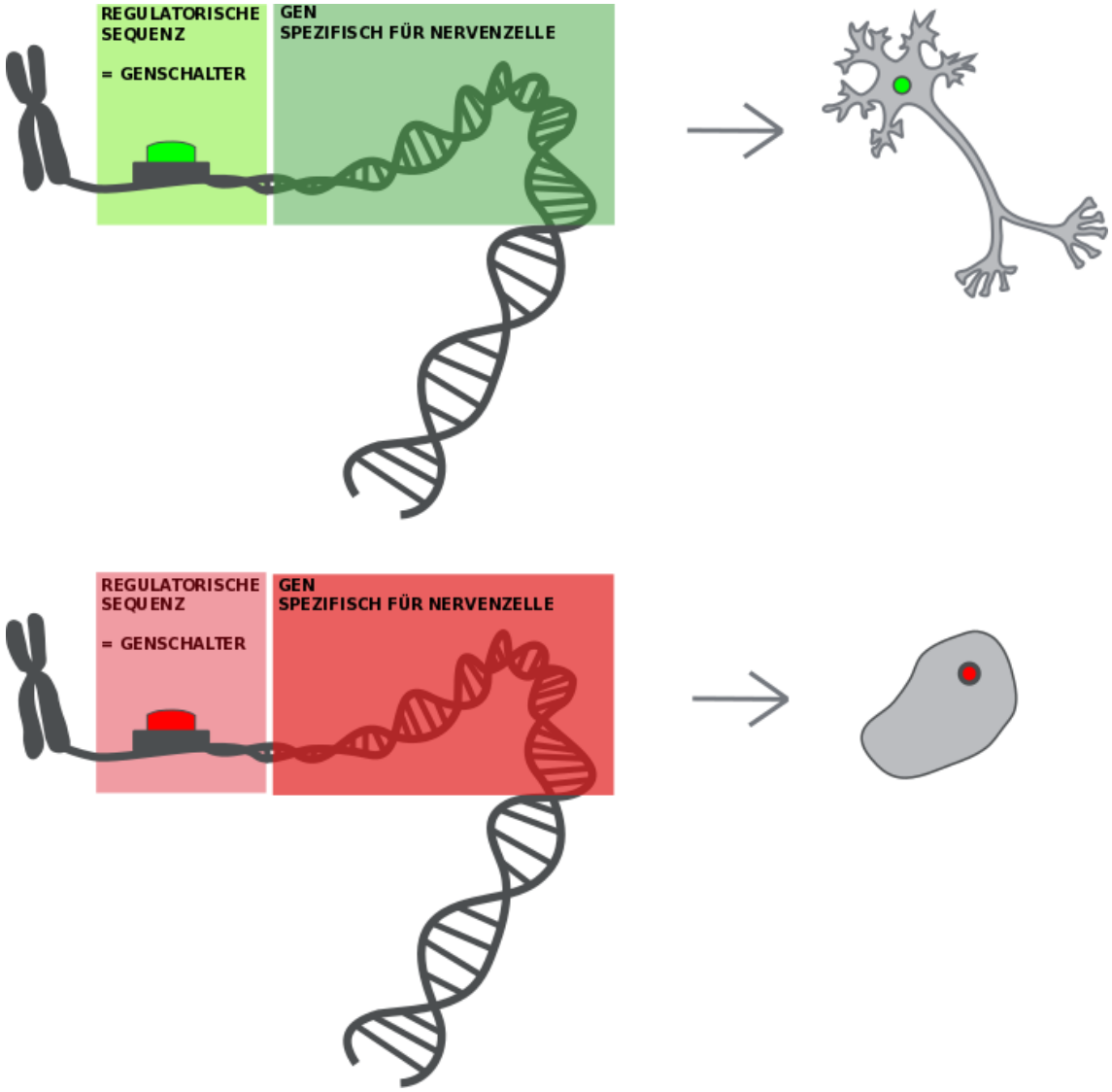
Sequencing of single candidate genes

Whole Exome Sequencing

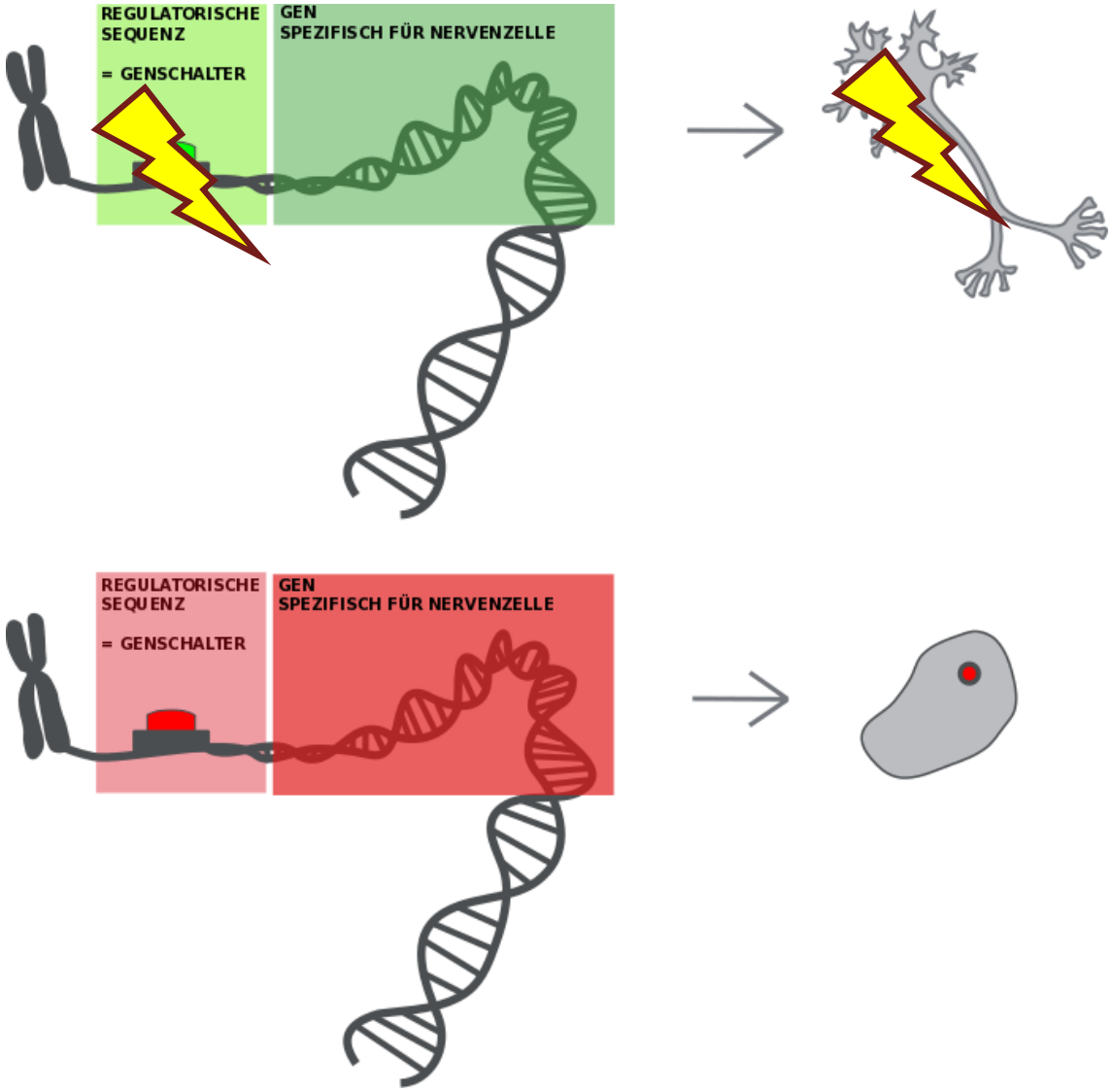


....did not detect the disease mutation!

Extragenic / regulatory DNA variants

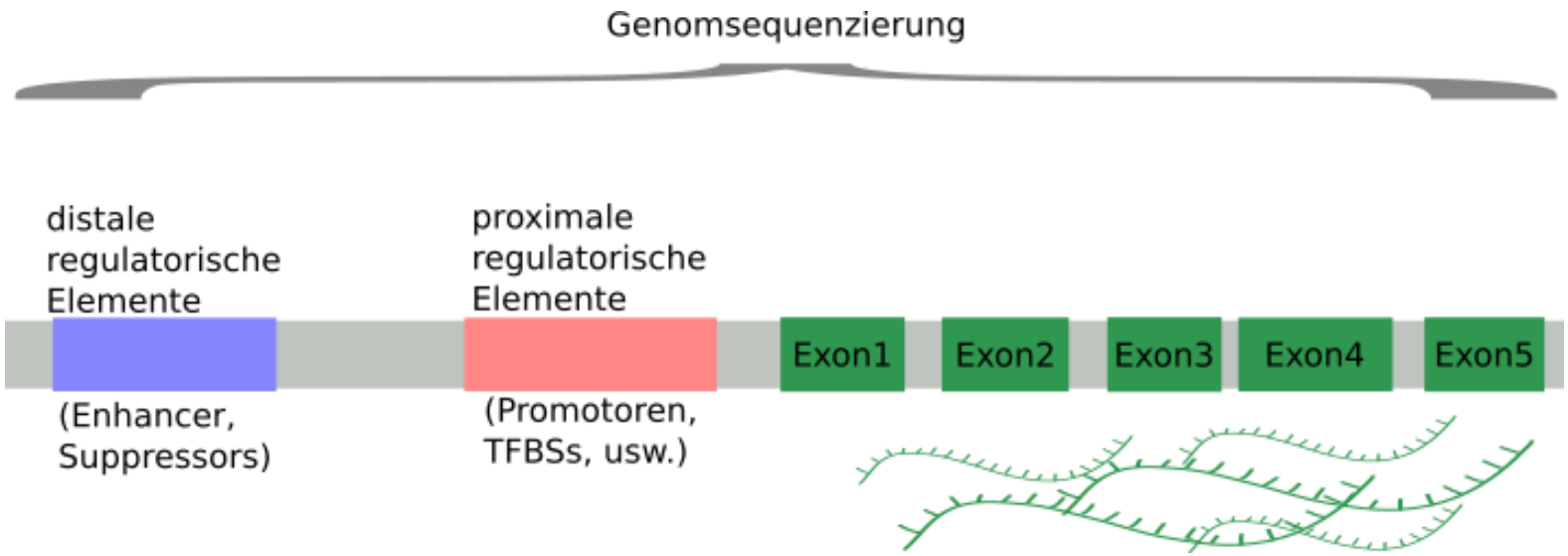


Extragenic / regulatory DNA variants



Mutations in regulatory elements can lead to onset of a genetic diseases.

Whole Genome Sequencing



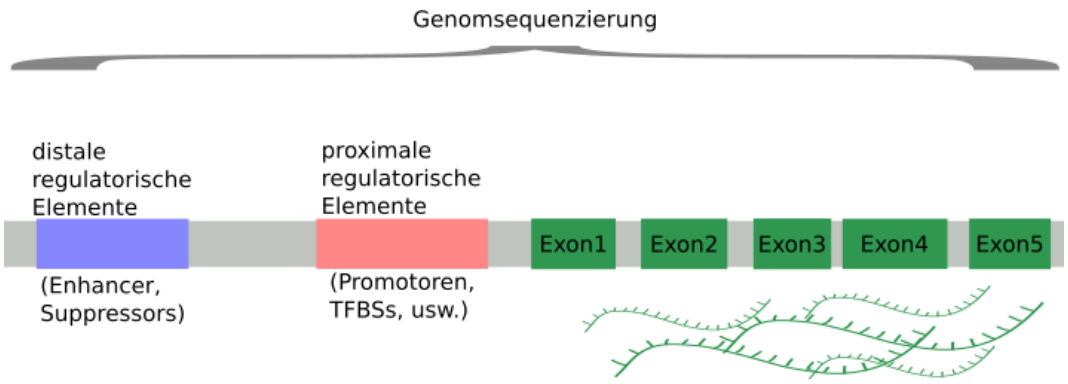
Strategies to find disease mutations

Homozygosity mapping

Sequencing of single candidate genes

Whole Exome Sequencing

Whole Genome Sequencing



~ 3 million variants

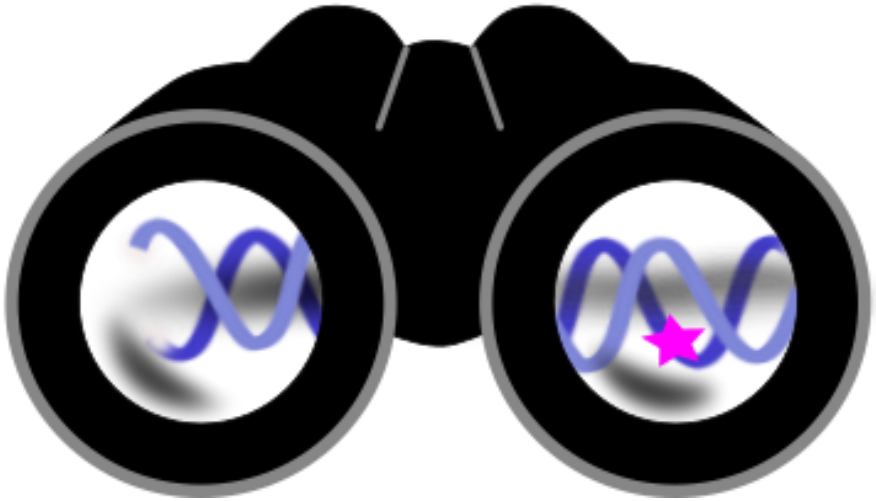
Mutation Taster

Only for genes.

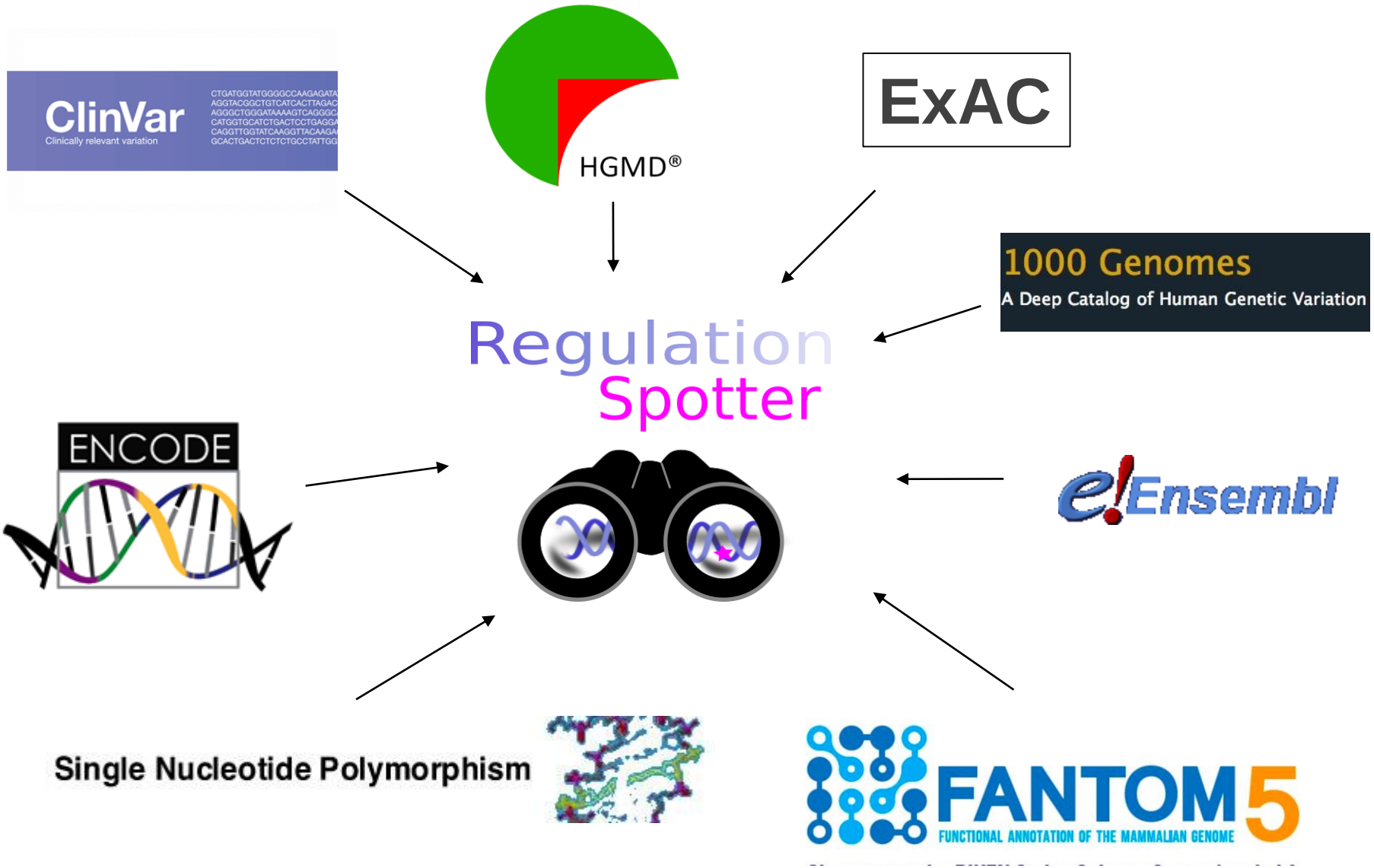
Schwarz JM, Cooper DN, Schuelke M, Seelow D: MutationTaster2: mutation prediction for the deep-sequencing age. *Nature Methods* (April 2014) 11(4):361-2.

RegulationSpotter

Regulation
Spotter



Integrated data (examples)



Integrated data (examples)

Integrated regulatory features

- 125 regulatory features including
  - promoter and enhancer annotations
  - histone modifications
  - DNAseI hypersensitive sites
  - Transcription factor binding sites

X-score

- reflects amount of evidence that a variant is located in a regulatory relevant region
- calculated based on 54 most meaningful regulatory annotations / features

feature weight:

- reflects assumed impact of each feature
- determined based on relative risk to occur in a testset of functional vs. non-functional extragenic variants

- X-score represents the sum of the weights for all features annotated for a variant

X-score

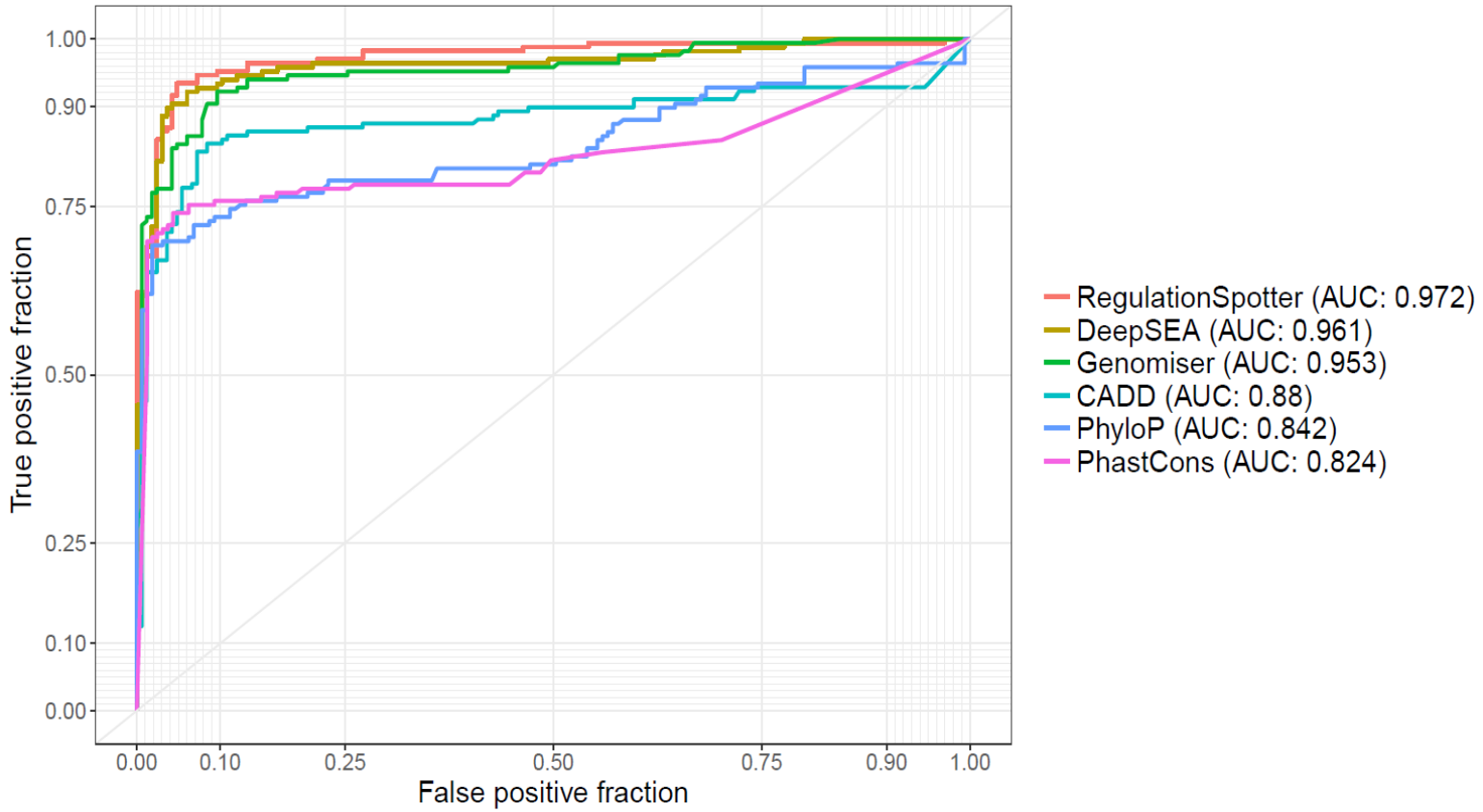
- low number of extragenic variants validated as regulatory disease mutations

  ➡ not enough for classifier training

positive: 295 disease variants from HGMDpro

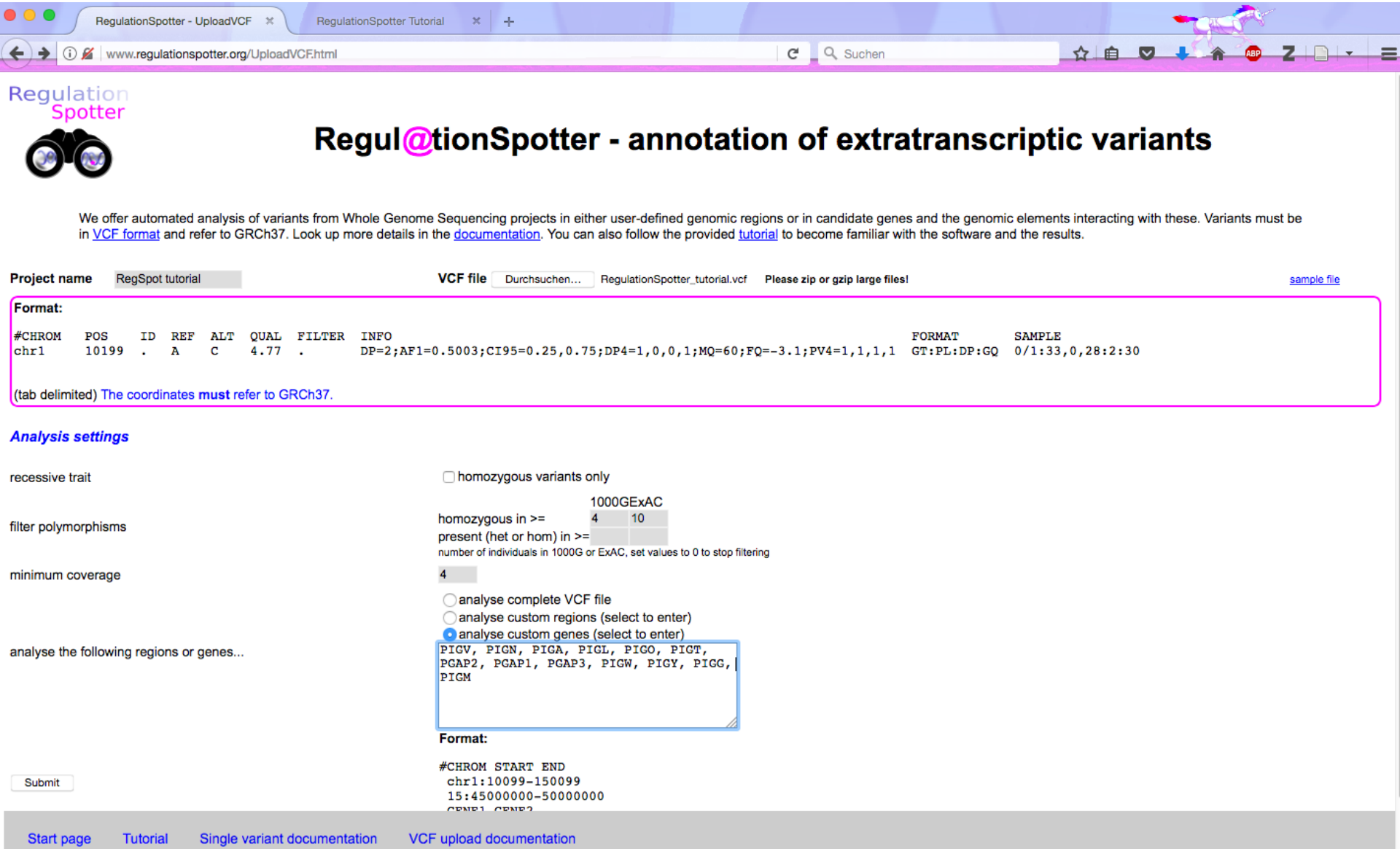
negative: millions of extragenic non-functional variants

Comparison with similar programs

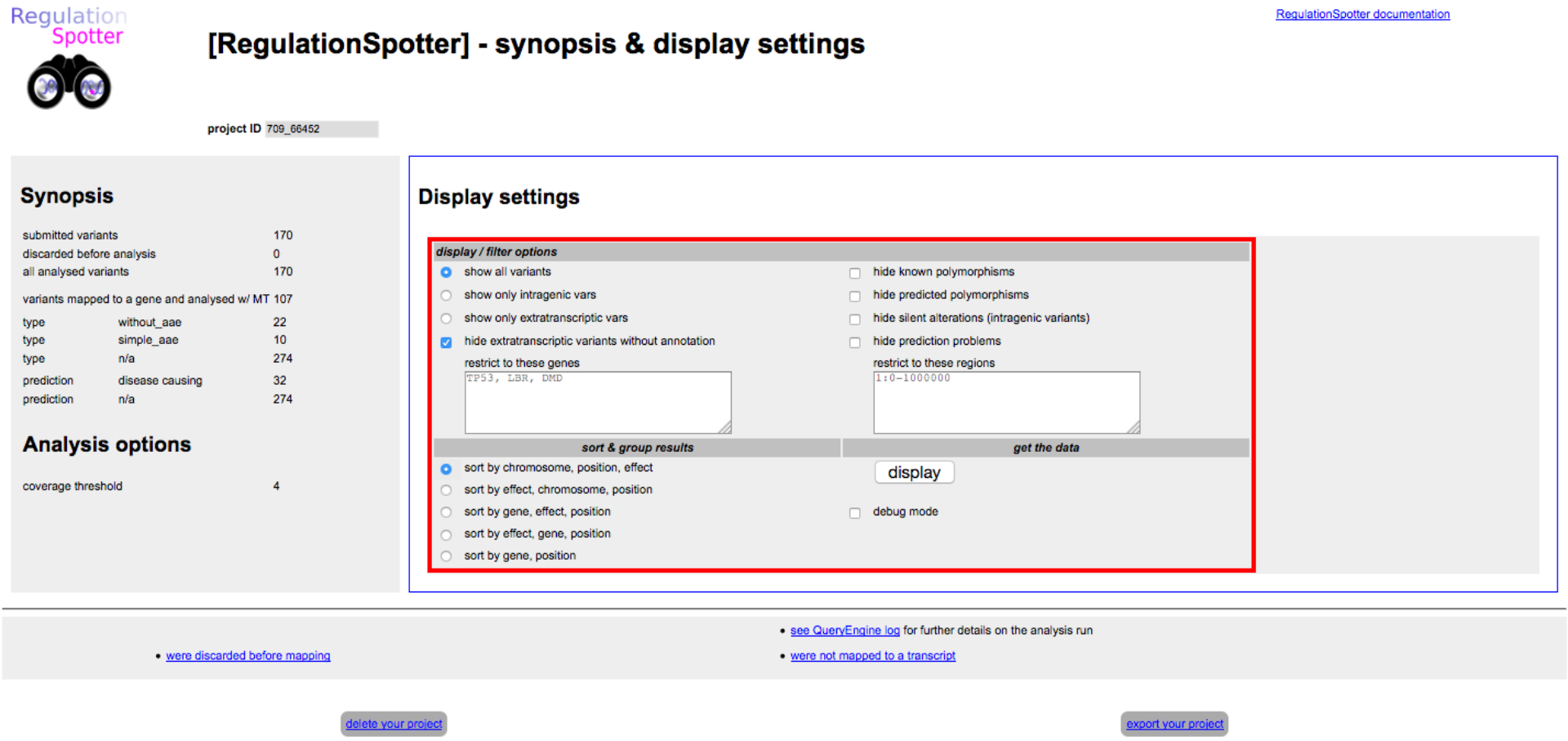


test set: 166 extragenic harmless polymorphisms from *1000 Genomes Project*
166 extragenic disease mutations from NCBI *ClinVar*

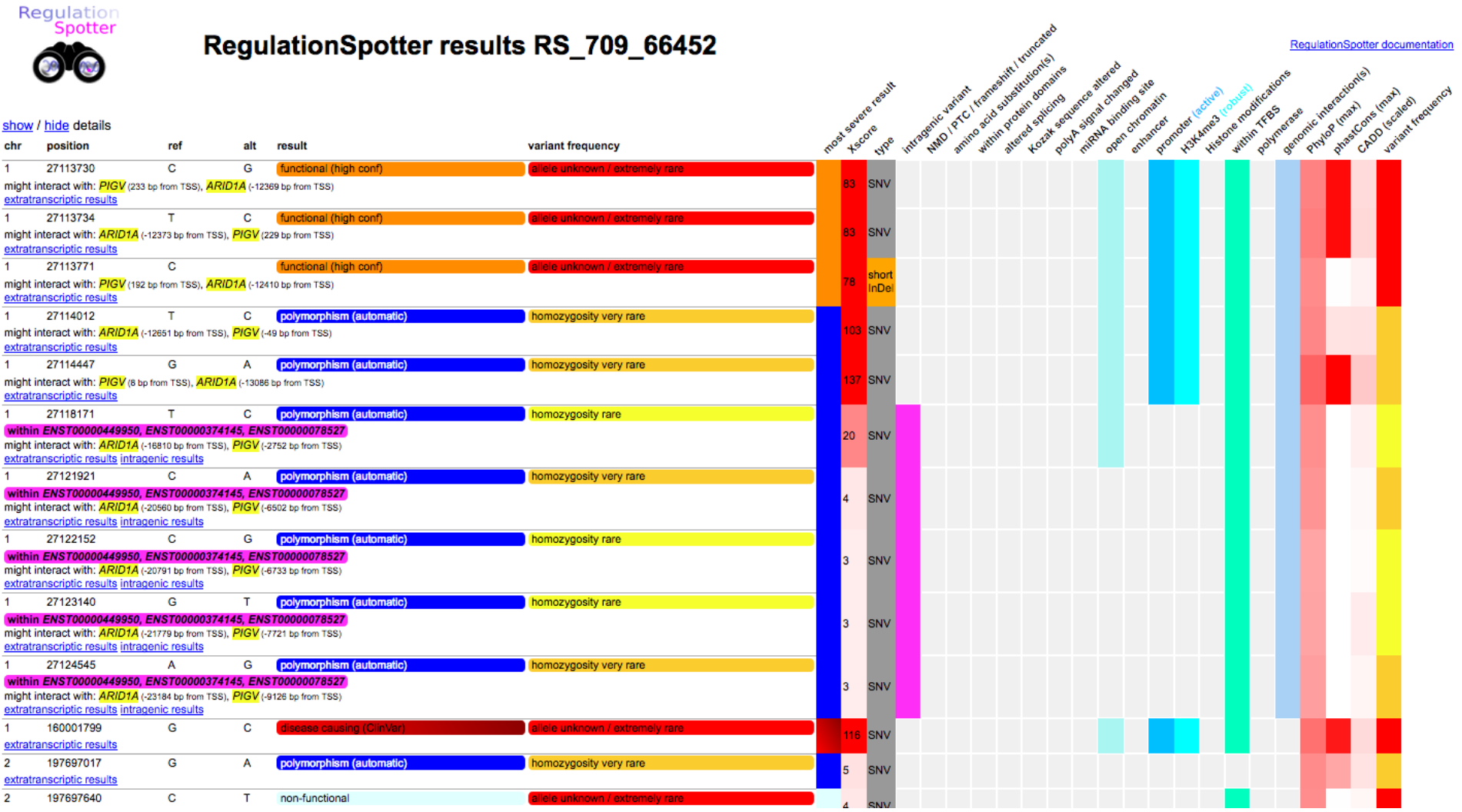
RegulationSpotter – submit VCF file



RegulationSpotter – results synopsis



RegulationSpotter – results matrix



RegulationSpotter – detailed results

regul@tion spotting

Likely effect **functional (high confidence)** **Model: *extratranscriptic*, score: 117.35756097561**

Summary hyperlink

- **within TFBS**
- **within active promoter**

analysed issue	**analysis result**
name of alteration	no title
alteration (phys. location)	chr17:37844280G>C IGV
alteration type	single base exchange
alteration region	extratranscriptic
known variant	Reference ID: rs564810973

database	homozygous (C/C)	heterozygous	allele carriers
1000G	0	1	1
ExAC	-	-	-

promoters

Ensembl Promoter ENSR00000126595

active A549, Aorta, B cells (PB) Roadmap, CD14+CD16- monocyte (CB), CD14+CD16- monocyte (VB), CD4+ ab T cell (VB), CD8+ ab T cell (CB), CM CD4+ ab T cell (VB), DND-41, eosinophil (VB), EPC (VB), erythroblast (CB), Fetal Adrenal Gland, Fetal Intestine Large, Fetal Intestine Small, Fetal Muscle Leg, Fetal Muscle Trunk, Fetal Stomach, Gastric, GM12878, H1ESC, H1-mesenchymal, H1-neuronal progenitor, H9, HeLa-S3, HepG2, HMEC, HSMM, HSMMtube, HUVEC, HUVEC prol (CB), IMR90, iPS-20b, iPS DF 19.11, iPS DF 6.9, K562, Left Ventricle, Lung, M0 macrophage (CB), M0 macrophage (VB), M1 macrophage (CB), M1 macrophage (VB), M2 macrophage (CB), M2 macrophage (VB), Monocytes-CD14+, Monocytes-CD14+ (PB) Roadmap, MSC (VB), naive B cell (VB), Natural Killer cells (PB), neutrophil myelocyte (BM), NH-A, NHDF-AD, NHEK, NHLF, Osteobl, Ovary, Pancreas, Placenta, Psoas Muscle, Right Atrium, Small Intestine, Spleen, T cells (PB) Roadmap

poised Fetal Thymus, H1-trophoblast, neutrophil (CB), neutrophil (VB), Thymus

FANTOM predictions

FANTOM TSS relaxed 17:37844268-37844316(-)

RegulationSpotter

17: 37843854-37844404	DNase1	H3K4me3	ERBB2	ENST00000584014
17: 37843893-37844443	DNase1	H3K4me3	ERBB2	ENST00000406381
17: 37843872-37844422	DNase1	H3K4me3	ERBB2	ENST00000578199
17: 37844252-37844802	DNase1	H3K4me3	PGAP3	ENST00000582276
17: 37844204-37844754	DNase1	H3K4me3	PGAP3	ENST00000584620
17: 37844260-37844810	DNase1	H3K4me3	PGAP3	ENST00000579146, ENST00000429199, ENST00000378011, ENST00000309862

enhancers	none found	
epigenetic marks (RegulationSpotter)	DNase1 DNase1 hypersensitive site in promoter	17:37843652-37844692
	H3K4me3 H3K4me3 in 3 cell lines, overlap with promoter	17:37843348-37845639
histone modifications	H2AZ 5 cell types: GM12878, HSMM, HSMMtube, K562, Osteobl	explain

Summary

MutationTaster

- evaluates disease potential of **intragenic** DNA variants
- intronic & exonic
- non-synonymous / synonymous and silent single base exchanges and indels
- classification by Bayes Classifier
- for single variants or complete vcf files
- for variants involved in Mendelian disorders!

http://www.mutationtaster.org/

RegulationSpotter

- evaluates functional / regulatory impact of **extragenic** DNA variants
- single base exchanges and indels
- classification with help of X-score
- for single variants or complete vcf files

http://www.regulationspotter.org/

Thank you!

AG Seelow (Charité Berlin / BIH)
Daniela Hombach
Sebastian Köhler
Dominik Seelow

https://translationalgenomics.charite.de/

AG Schülke (Charité Berlin / Neuropeadiatrics)
Ellen Knierim
Gudrun Schottmann
Markus Schülke

HGMD
David Cooper (University of Cardiff)

Practicals:
Wednesday
11.45-13.15
14.15-15.45
Zürich 3

Schwarz JM, Rödelsperger C, Schuelke M, Seelow D.
MutationTaster evaluates disease-causing potential of sequence alterations.
Nature Methods. 2010 Aug;7(8):575-6

Schwarz JM, Cooper DN, Schuelke M, Seelow D.
MutationTaster2: mutation prediction for the deep-sequencing age.
Nature Methods. 2014, Apr;11(4):361-2.

Schwarz JM, Hombach D, Koehler S, Cooper DN, Schuelke M, Seelow D.
RegulationSpotter: Easy and comprehensive analysis of extratranscriptic sequence variants.
Under revision.

http://www.mutationtaster.org/ http://www.regulationspotter.org/

The case of Family D. – what comes up next

Homozygosity mapping

Sequencing of single candidate genes

Whole Exome Sequencing

Whole Genome Sequencing

iPSCs

Whole Transcriptome Sequencing