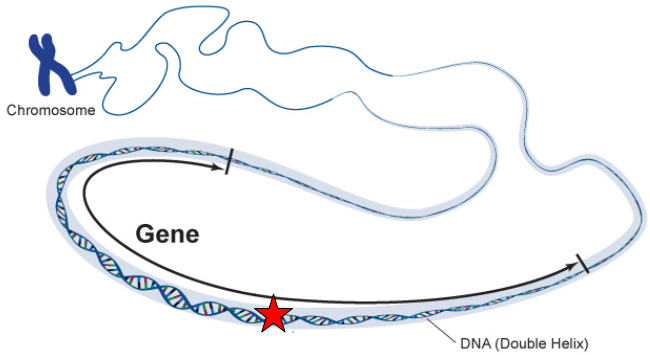
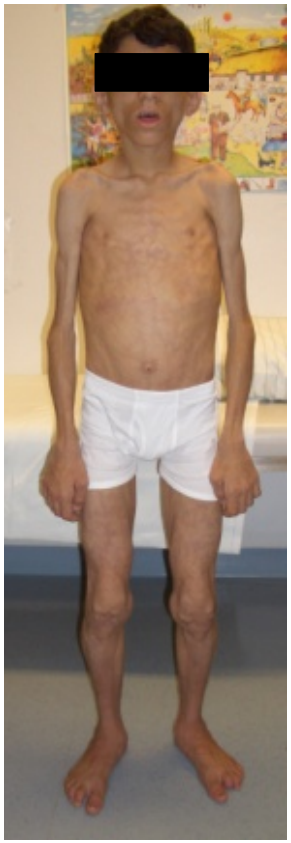
Jana Marie Schwarz
Markus Schülke
Dominik Seelow
Charité –
Universitätsmedizin Berlin

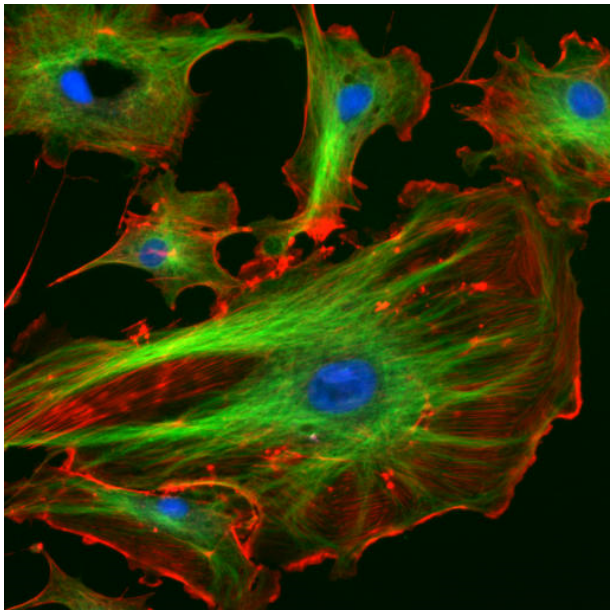
Automatische Prozessierung von NGS-Daten mit MutationTaster2

Mutation → Phänotyp





Experimentelle Validierung...



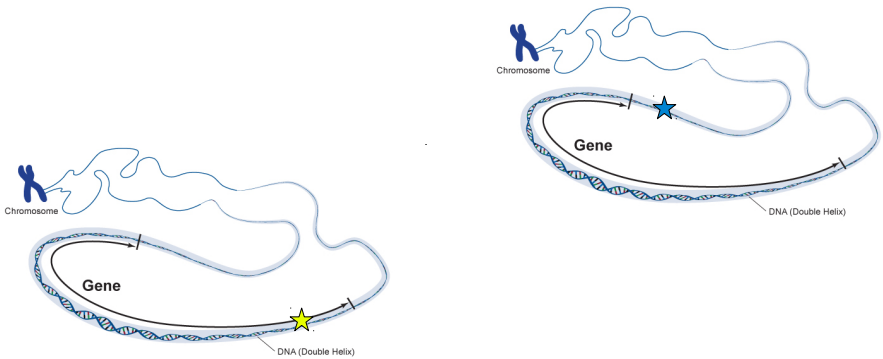
© NIH, USA

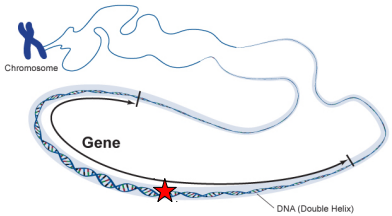
… kostet Zeit und Geld

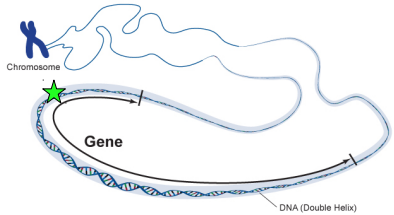
© Der Tagesspiegel

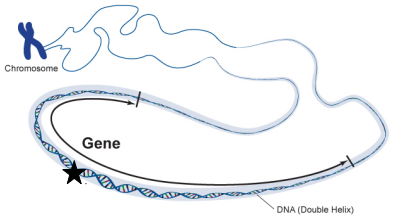
© www.wirklichgutes.de

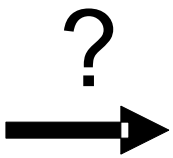
Und bei mehreren Veränderungen?

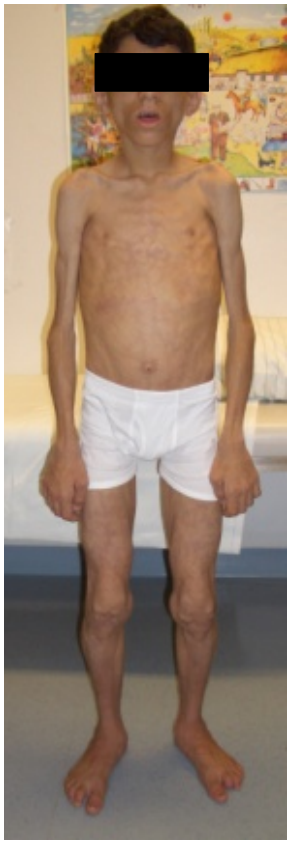




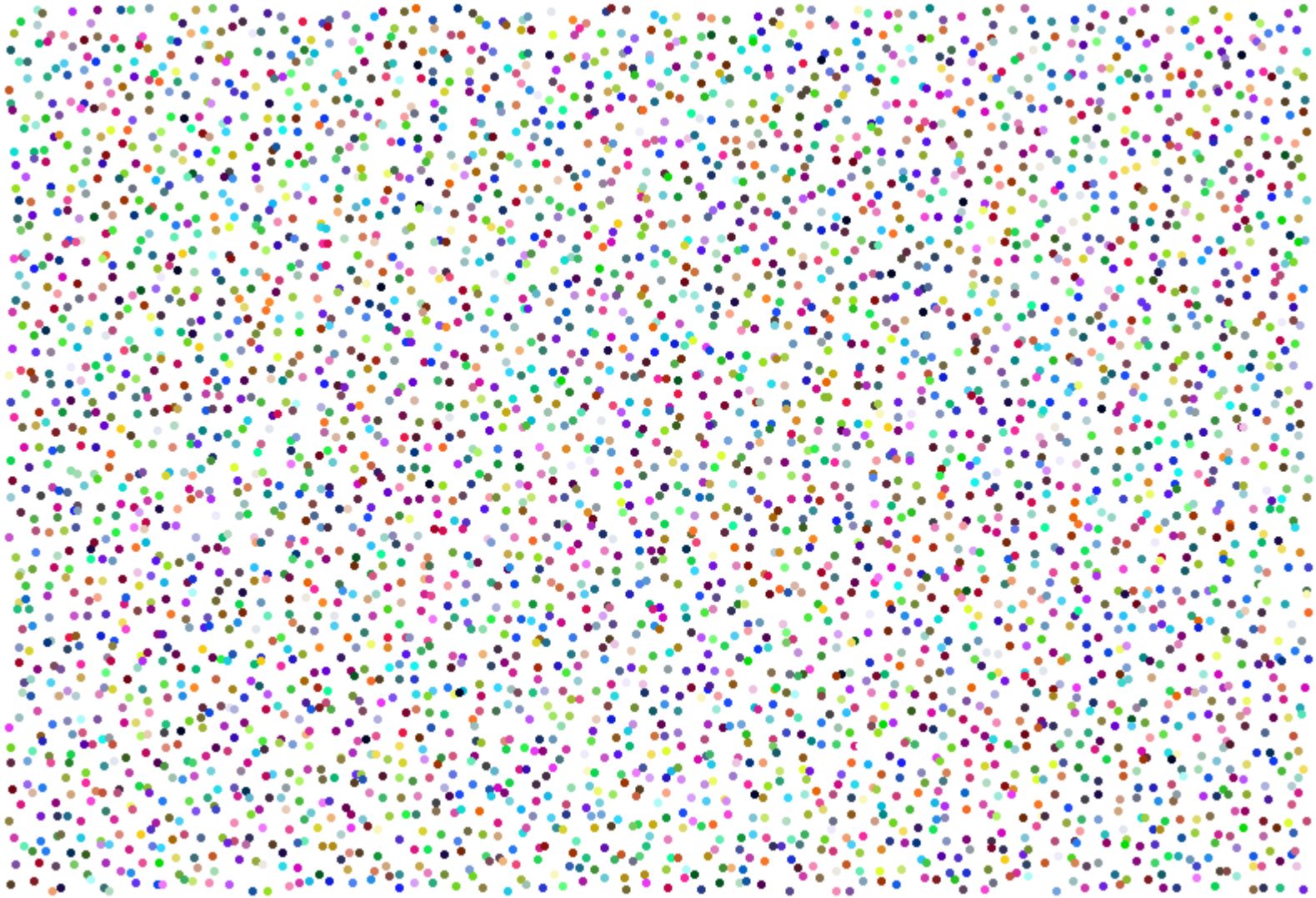




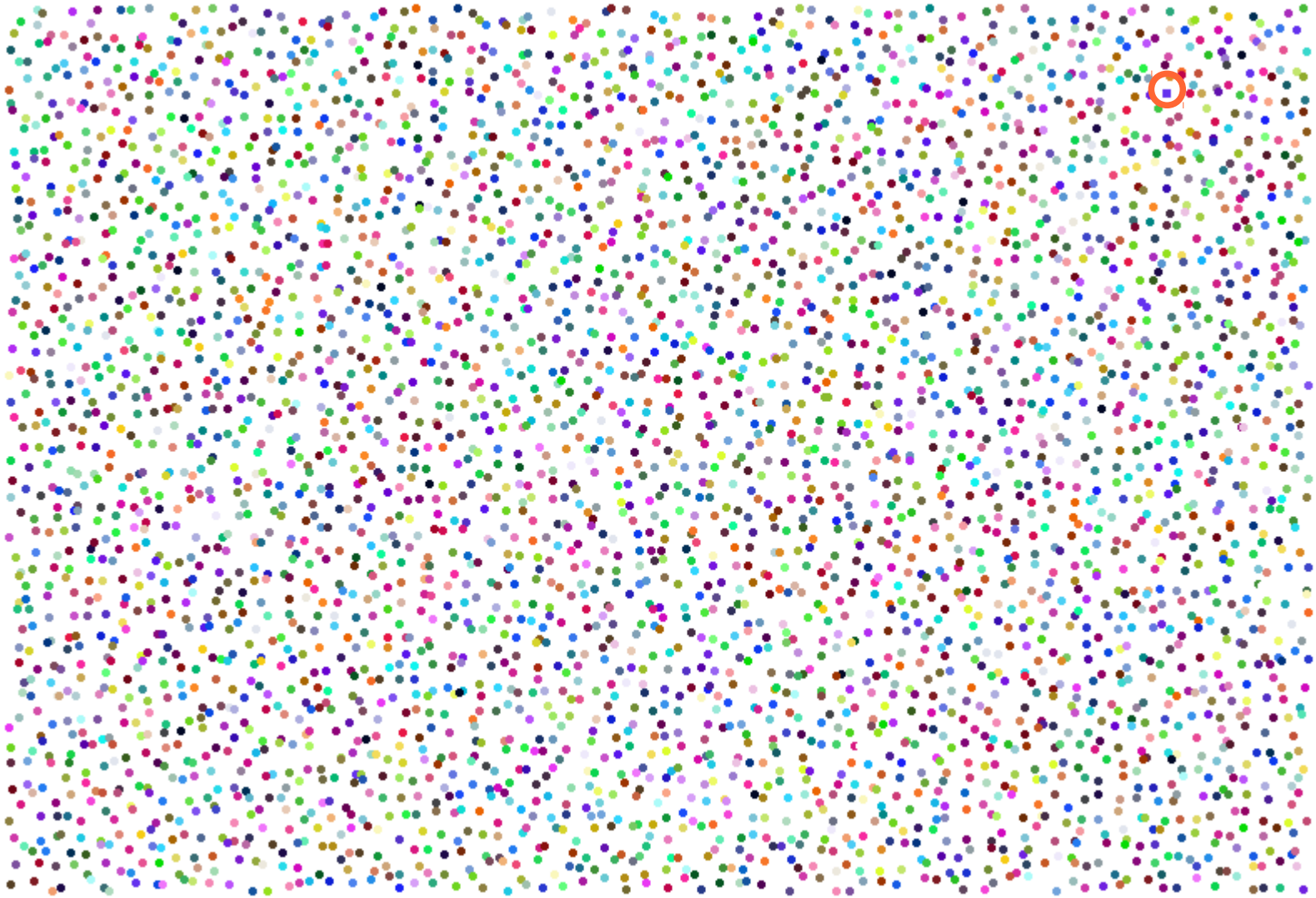




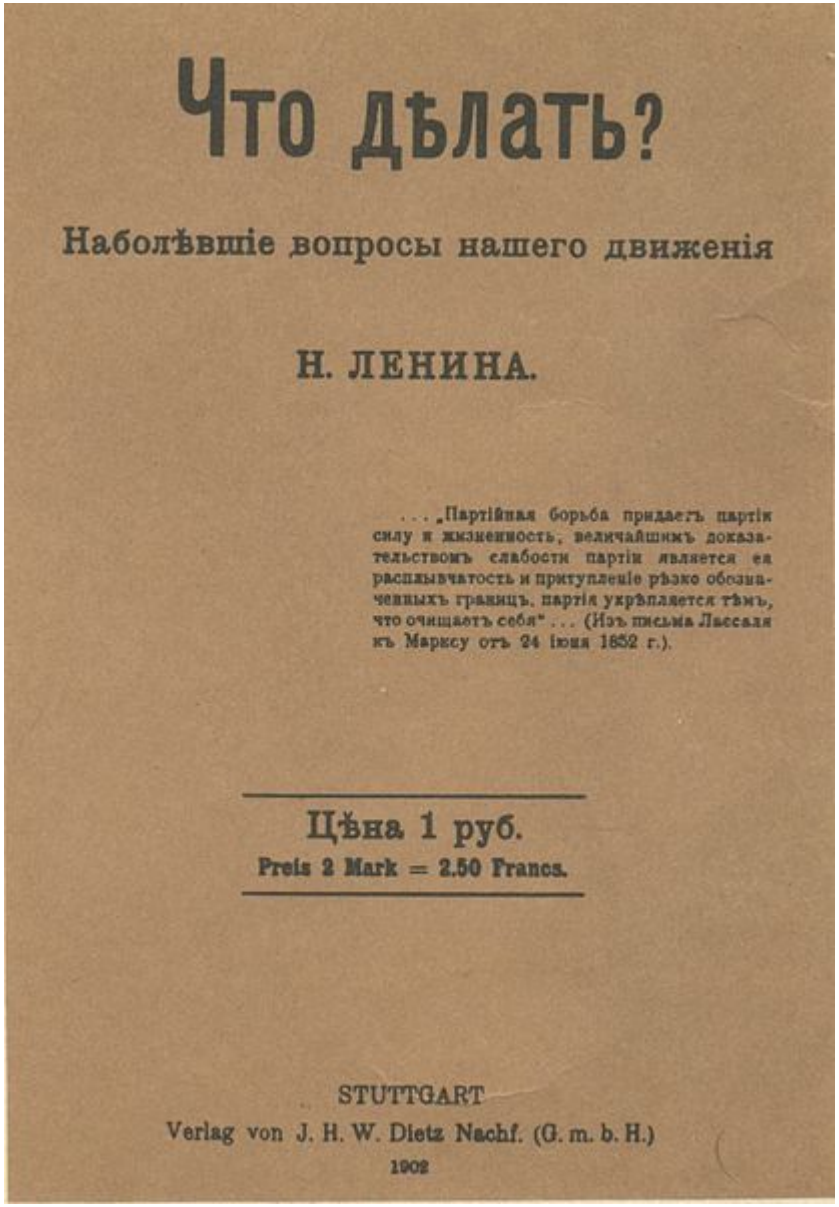
Die Suche nach der Einen...



ist – naja - aufwendig...



Was tun?



Что дѣлать?

Наболѣвшіе вопросы нашего движенія

Н. ЛЕНИНА.

. . . „Партійная борьба придаетъ партіи силу и жизненность, величайшимъ доказательствомъ слабости партіи является ея расплывчатость и притупленіе рѣзко обозначенныхъ границъ, партія укрѣпляется тѣмъ, что очищаетъ себя" . . . (Изъ письма Лассаля къ Марксу отъ 24 іюня 1852 г.).

Цѣна 1 руб.
Preis 2 Mark = 2.50 Francs.

STUTTGART
Verlag von J. H. W. Dietz Nachf. (G. m. b. H.)
1902

Filterstrategien: negativ

Filterstrategien: negativ

Probleme

- dbSNP enthält Krankheitsmutationen
- viele Polymorphismen (noch) nicht im TGP

- 'persönliche' Polymorphismen fehlen natürlich

Filterstrategien: positiv

HGMD®

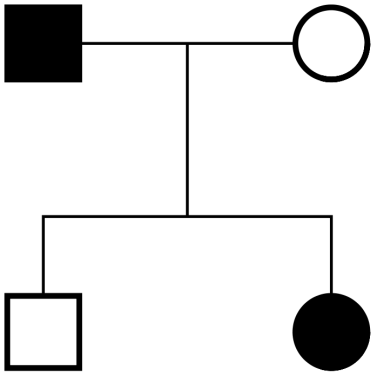
Filterstrategien: positiv

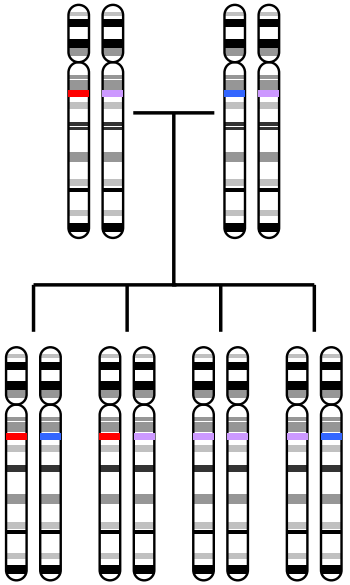
Probleme

- ClinVar ist ziemlich leer
- HGMD ist lizenzpflichtig (und auch nicht komplett)

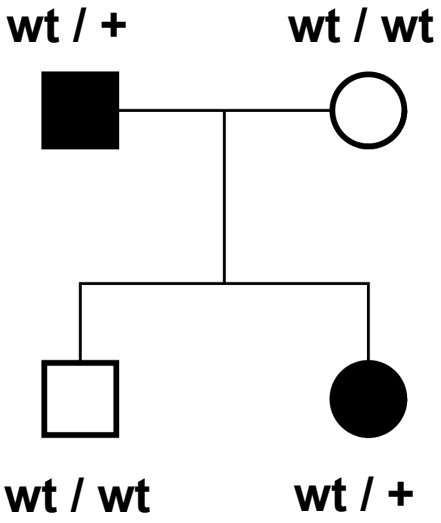
- neue Mutationen fehlen natürlich

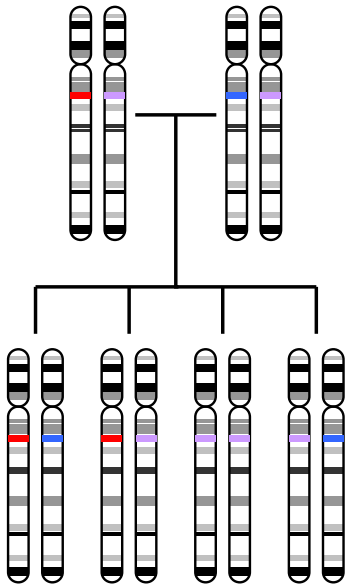
Filterstrategien: Vererbung / Kopplungsregionen





Filterstrategien: Vererbung / Kopplungsregionen

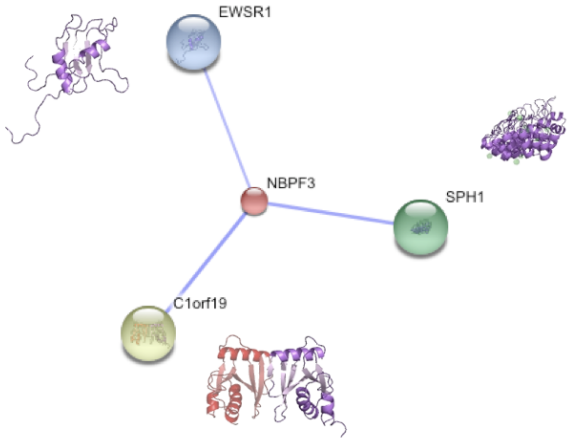




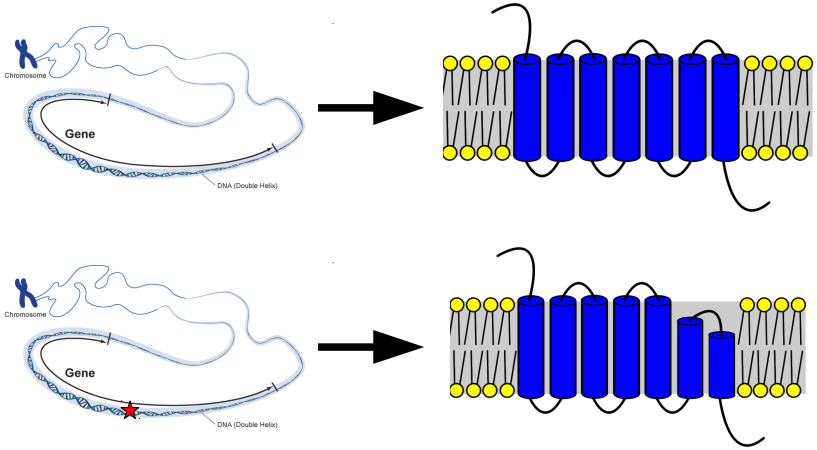
Probleme

- für Filterung wt/hom/het werden auch die RefSeq-Allele benötigt
- in Kopplungsregionen / homozygoten Regionen paßt natürlich vieles

Filterstrategien: nach Zielgenen



Filterstrategien: nach Mutationseffekt



MutationTaster evaluates disease-causing potential of sequence alterations.
Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. *Nat Methods*. 2010
http://www.mutationtaster.org/

MutationTaster

- diverse *in silico* Tests zur Vorhersage des Krankheitspotentials
- richtige Vorhersagen: ~90%
- besser als PolyPhen, PolyPhen-2 etc.
- nicht nur Aminosäureaustausche
- ca. 0,3 s / Variante
- schlecht bei 'stillen' Mutationen

MutationTaster - Mozilla Firefox

mutation t@sting

- MutationTaster - stable version
- **documentation**
- **FAQs**
- batch queries
- Next Generation Sequencing pipeline
- supplementary data
- 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 (1746 bases)
- ENST00000270142 (966 bases) NM_000454
- ENST00000389995 (865 bases)
- ENST00000476106 (586 bases)

Position / snippet refers to coding sequence (ORF) transcript (cDNA sequence) gene (genomic sequence)

Alteration

all types by sequence

TGCATGGATTCC[A/G]TGTTCATGAGTTT

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)

Name of alteration SOD1 H47R 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.*

MutationTaster - Mozilla Firefox

mutation t@sting

- MutationTaster - stable version
- **documentation**
- **FAQs**
- batch queries
- Next Generation Sequencing pipeline
- supplementary data
- other applications
- team

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

SOD1

Transcript ENST00000270142 Ensembl

Choose the transcript:

- ENST00000470944 (1746 bases)
- ENST00000270142 (966 bases) NM_000454
- ENST00000389995 (865 bases)
- ENST00000476106 (586 bases)

Position / snippet refers to coding sequence (ORF) transcript (cDNA sequence) gene (genomic sequence)

Alteration

all types by sequence

TGCATGGATTCC[A/G]TGTTCATGAGTTT

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)

Name of alteration SOD1 H47R 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.

MutationTaster - Mozilla Firefox

mutation t@sting

- MutationTaster - stable version
- **documentation**
- **FAQs**
- batch queries
- Next Generation Sequencing pipeline
- supplementary data
- 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 (1746 bases)
- ENST00000270142 (966 bases) NM_000454
- ENST00000389995 (865 bases)
- ENST00000476106 (586 bases)

Position / snippet refers to coding sequence (ORF) transcript (cDNA sequence) gene (genomic sequence)

Alteration

all types by sequence

TGCATGGATTCC[A/G]TGTTCATGAGTTT

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

ATGGATTCC**[A/G]**TGTTCATGAG

single base exchange by position

enter **position**

and **new base**

(causes ALS)

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 H47R 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.

Mutation Taster

documentation

mutation t@sting

Alteration SOD1H47R

Prediction **disease causing** Model: *simple_aae*, p (probability): 0.999999345096013 (explain)

Summary

- amino acid sequence changed
- protein features (might be) affected
- splice site changes

analysed issue	analysis result
name of alteration	SOD1H47R
HGNC symbol	SOD1
Ensembl transcript ID	ENST00000270142
UniProt peptide	P00441
alteration type	single base exchange
alteration region	CDS
alteration on nucleotide level	A>G
coding sequence (CDS) position	140
AA changes	H47R Score: 0.79 explain score(s)
frameshift	no
SNPs	no SNPs in altered region found

splice sites

effect	gDNA position	score	sequence
Acc increased	4243	wt: 0.31 / mu: 0.44	CTGCATGGATTCCATGTTCATGAGTTTGGAGATAATACAGC / tcatGAGT
Acc marginally increased	4245	wt: 0.8500 / mu: 0.9286 (marginal change - not scored)	GCATGGATTCCATGTTCATGAGTTTGGAGATAATACAGCAG / atgaGTTT

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
Chimp	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
Rhesus	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
Cat	no homologue		
Mouse	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

Mutation Taster

documentation

mutation t@sting

Alteration SOD1H47R

Prediction **disease causing** Model: *simple_aae*, p (probability): 0.999999345096013 (explain)

Summary

- amino acid sequence changed
- protein features (might be) affected
- splice site changes

analysed issue	analysis result
name of alteration	SOD1H47R
HGNC symbol	SOD1
Ensembl transcript ID	ENST00000270142
UniProt peptide	P00441
alteration type	single base exchange
alteration region	CDS
alteration on nucleotide level	A>G
coding sequence (CDS) position	140
AA changes	H47R Score: 0.79 explain score(s)
frameshift	no
SNPs	no SNPs in altered region found

splice sites

effect	gDNA position	score	sequence
Acc increased	4243	wt: 0.31 / mu: 0.44	CTGCATGGATTCCATGTTCATGAGTTTGGAGATAATACAGC / tcatGAGT
Acc marginally increased	4245	wt: 0.8500 / mu: 0.9286 (marginal change - not scored)	GCATGGATTCCATGTTCATGAGTTTGGAGATAATACAGCAG / atgaGTTT

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
Chimp	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
Rhesus	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
Cat	no homologue		
Mouse	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

documentation

mutation t@sting

Alteration SOD1H47R

Prediction **disease causing** Model: *simple_aae*, p (probability): 0.999999345096013 (explain)

Summary

- amino acid sequence changed
- protein features (might be) affected
- splice site changes

analysed issue	analysis result
name of alteration	SOD1H47R
HGNC symbol	SOD1
Ensembl transcript ID	ENST00000270142
UniProt peptide	P00441
alteration type	single base exchange
alteration region	CDS
alteration on nucleotide level	A>G
coding sequence (CDS) position	140
AA changes	H47R Score: 0.79 explain score(s)
frameshift	no
SNPs	no SNPs in altered region found

splice sites	effect	gDNA position	score	sequence
	Acc increased	4243	wt: 0.31 / mu: 0.44	CTGCATGGATTCCATGTTCATGAGTTTGGAGATAATACAGC / tcatGAGT
	Acc marginally increased	4245	wt: 0.8500 / mu: 0.9286 (marginal change - not scored)	GCATGGATTCCATGTTCATGAGTTTGGAGATAATACAGCAG / atgaGTTT
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
Chimp	all identical	ENSPTRG00000013847	47 I K G L T E G L H G F **I** V H E F G D N T A G C
Rhesus	all identical	ENSMMUG00000001711	47 I T G L T E G L H G F **I** V H Q F G D N T Q G C
Cat	no homologue		
Mouse	all identical	ENSMUSG00000022982	47 I T G L T E G Q H G F **I** V H Q Y G D N T Q G C

	increased - not scored) atgaGTTT
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
Chimp	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
Rhesus	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
Cat	no homologue		
Mouse	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
Chicken	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
Xenopus	all identical	ENSXETG00000007350	46 I Y G L T D G K H G F **H** I H E F G D N T N G C
Zebrafish	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
Fugu	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
Elegans	no homologue		
Drosophila	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

variants	N/A

protein features

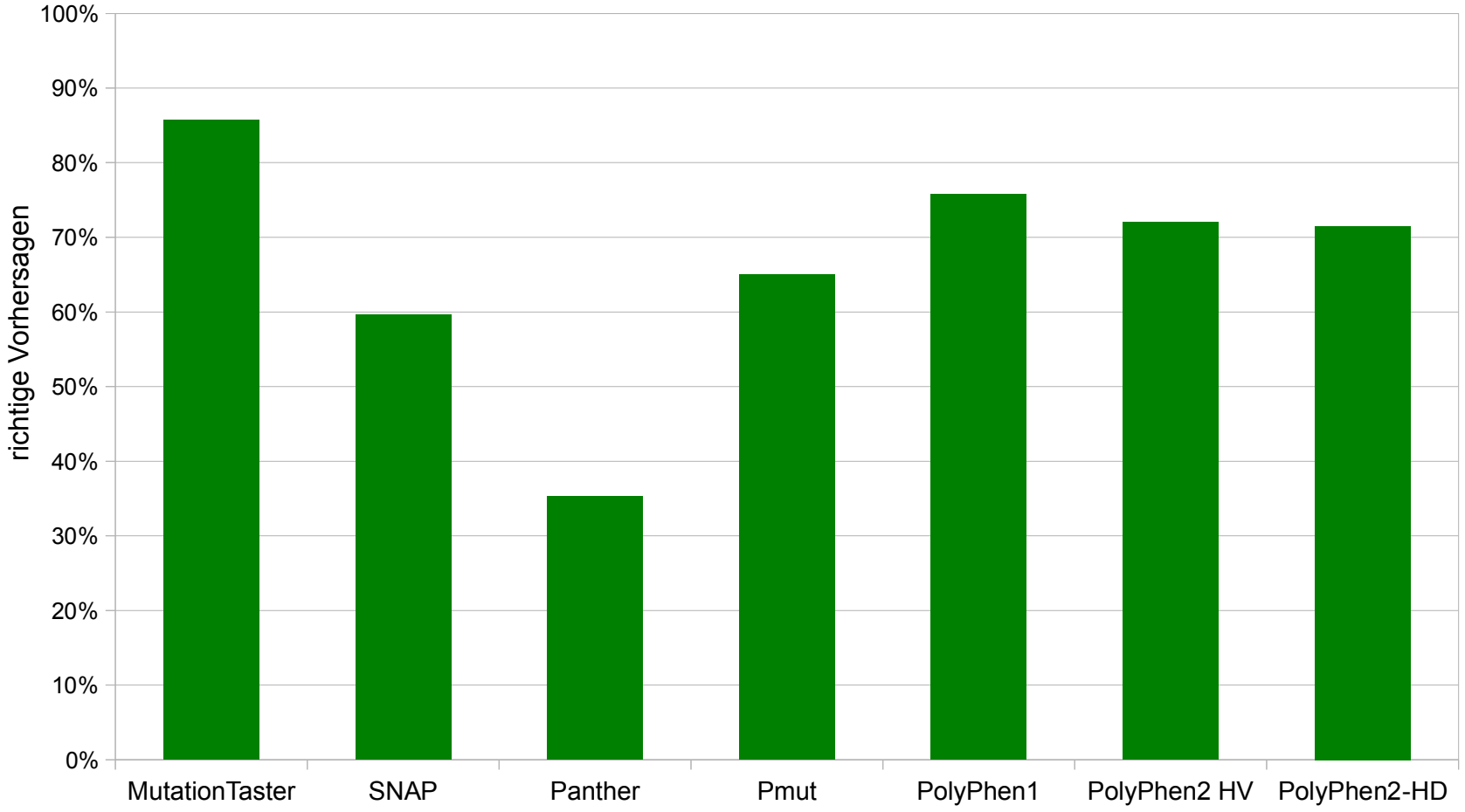
start (aa)	end (aa)	feature	details	
41	50	STRAND		lost
47	47	METAL	Copper; catalytic.	lost
116	123	STRAND		might get lost (downstream of altered splice site)
123	123	MOD_RES	N6-acetyllysine.	might get lost (downstream of altered splice site)
130	132	STRAND		might get lost (downstream of altered splice site)
134	134	MUTAGEN	E->Q: Abolishes dimerization; when associated with E-50 and E-51.	might get lost (downstream of altered splice site)
135	138	HELIX		might get lost (downstream of altered splice site)
143	149	STRAND		might get lost (downstream of altered splice site)
147	147	MUTAGEN	C->S: Enhances formation of fibrillar aggregates in the absence of bound zinc; when associated with S-7; S-58 and S-112.	might get lost (downstream of altered splice site)
147	147	DISULFID		might get lost (downstream of altered splice site)

length of protein	Protein of normal length (unless alteration is in UTR...)
AA sequence altered	yes
position(s) of altered AA if AA alteration in CDS	N/A
position of stopcodon in wt / mu CDS	465 / 465
position (AA) of stopcodon in wt / mu AA sequence	155 / 155
poly(A) signal	N/A

conservation
nucleotide level for all changes - no scoring up to now

species	match	gene	aa alignment
Human			4236 G C A T G G A T T C C A T G T T C A T G A G T T

Vergleich mit ähnlichen Programmen



1200 Einzelaminosäurenaustausche
600 bekannte Krankheitsmutationen vs. 600 validierte Polymorphismen

MutationTaster2

- mehr Tests für 'stille' Mutationen
- Integration von TGP, HapMap & ClinVar
- richtige Vorhersagen: ~90% über alle Modelle
- ca. 0,3 s / Variante
- Training mit TGP & HGMD (ca. 6M Varianten)

MutationTaster2: QueryEngine

- Hochladen von VCF-Dateien
- Einstellen von Filteroptionen

  (TGP, hom/het, Regionen)
- parallele Prozessierung der Daten
- Portal zur Auswertung der Ergebnisse

mutation t@sting

QueryEngine

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 Browse... 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

E-mail address

no HTML files (faster) ☑

Analysis settings

search for homozygous variants ☐ yes

search for compound heterozygous 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

- ◉ analyse complete VCF
- ○ analyse only exonic variants (20 bp intronic flanking)
- ○ analyse only the following regions
- ○ exclude the following regions

Format:

DM_SNP - explore the results

documentation

Statistics

analysable alterations		120769
...in TGP		0
...outside genes		366
alterations analysed		120403
MT cases		444618
type	without_aae	134110
type	simple_aae	165103
type	complex_aae	142365
type	n/a	3040
prediction	disease_causing	245061
prediction	polymorphism	196279
prediction	polymorphism_automatic	238
prediction	n/a	3040
genes hit (except HapMap/TGP)		
...non-synonymous		2532
...disease causing		2466
...disease causing, non-synonymous		2305

Display / filter / export data

sort & group

- (•) sort & group by prediction | model | gene symbol
- ( ) sort & group by prediction | model | gene symbol | variation
- ( ) *sort by these attributes*

1: [ ]
2: [ ]
3: [ ]

hide

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

get the data

display

(5000 results per page)

export as TSV

Show alterations which

- had input problems
- could not mapped to a transcript
- were filtered out (TGP)
- could not be predicted

Query GeneDistiller

call GeneDistiller for all...

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

documentation

MT results

back to overview

show next

chromosome	position	genesymbol	pred_index	model	probability	alt_type	AAE	snp_id	allele_ref	allele_alt	f_ClinVar	features	coverage	gt/freq(var)	filename
12	9232351	A2M	disease_causing	simple_aae	1	single base exchange	C972Y	1800433	C	T		splicing affected; protein features affected	154	het	1
22	43089410	A4GALT	disease_causing	simple_aae	0.684585	single base exchange	M183K	74315453	A	T		splicing affected; protein features affected	154	het	2
22	43089410	A4GALT	disease_causing	simple_aae	0.684585	single base exchange	M183K	74315453	A	T		splicing affected; protein features affected	154	het	3
22	43089410	A4GALT	disease_causing	simple_aae	0.684585	single base exchange	M183K	74315453	A	T		splicing affected; protein features affected	154	het	4
22	43089207	A4GALT	polymorphism	simple_aae	0.604124	single base exchange	P251S		G	A		protein features affected	154	het	5
22	43089207	A4GALT	polymorphism	simple_aae	0.604124	single base exchange	P251S		G	A		protein features affected	154	het	6
22	43089207	A4GALT	polymorphism	simple_aae	0.604124	single base exchange	P251S		G	A		protein features affected	154	het	7
22	43089206	A4GALT	polymorphism	complex_aae	0.993014	single base exchange		28940571	G	A		protein features affected	154	het	8
22	43089206	A4GALT	polymorphism	complex_aae	0.993014	single base exchange		28940571	G	A		protein features affected	154	het	9
22	43089206	A4GALT	polymorphism	complex_aae	0.993014	single base exchange		28940571	G	A		protein features affected	154	het	10
22	43089302	A4GALT	polymorphism	simple_aae	0.600435	single base	A219V		G	A		splicing affected; protein features	154	het	11

Ausblick & Visionen

- Filterung nach Vererbung / Status
- Krankheits-Datenbank
- weitere Polymorphismen-Datenbanken

- eigene Datenbank via MT?

Danke fürs Zuhören!