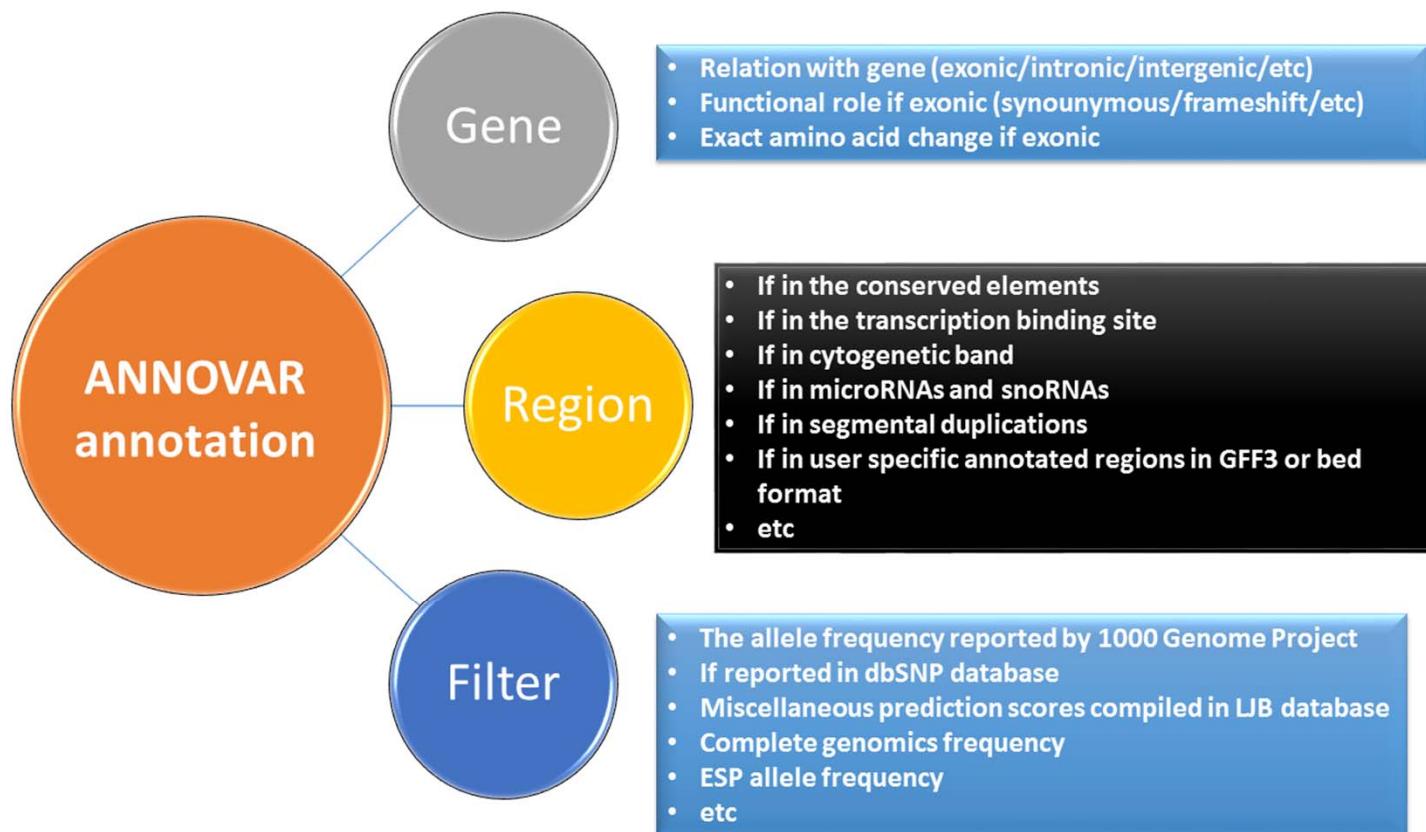


# Annotation and phenotype- driven interpretation of genetic variants

2019 Dragon Star Bioinformatics Course (Day 3)

# ANNOVAR for variant annotation



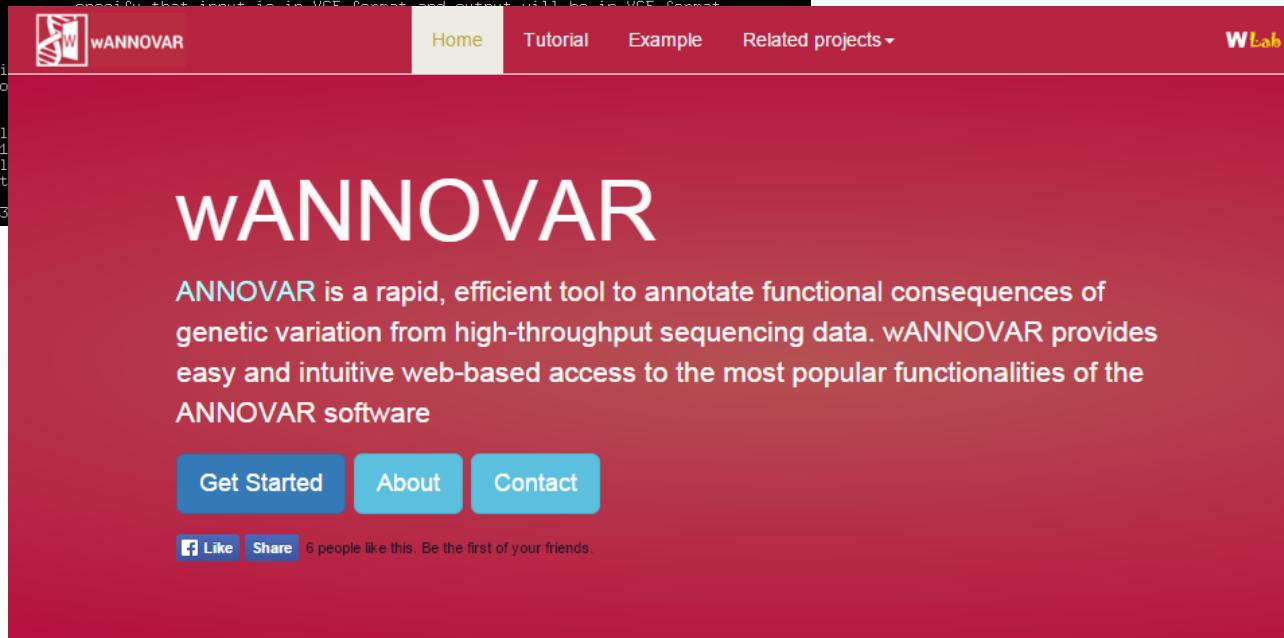
# ANNOVAR/wANNOVAR

```
[kaiwang@biocluster ~]$ table_annovar.pl
Usage:
  table_annovar.pl [arguments] <query-file> <database-location>

Optional arguments:
  -h, --help          print help message
  -m, --man           print complete documentation
  -v, --verbose       use verbose output
  --protocol <string> comma-delimited string specifying database protocol
  --operation <string> comma-delimited string specifying type of operation
  --outfile <string>  output file name prefix
  --buildver <string> genome build version (default: hg18)
  --remove           remove all temporary files
  --(no)checkfile    check if database file exists (default: ON)
  --genericdbfile <files> specify comma-delimited generic db files
  --gff3dbfile <files> specify comma-delimited GFF3 files
  --bedfile <files>   specify comma-delimited BED files
  --vcfdbfile <files> specify comma-delimited VCF files
  --otherinfo        print out otherinfo (information after fifth column in queryfile)
  --onetranscript    print out only one transcript for exonic variants (default: all transcripts)
  --nastring <string> string to display when a score is not available (default: null)
  --csvout          generate comma-delimited CSV file (default: tab-delimited txt file)
  --argument <string> comma-delimited strings as optional argument for each operation
  --tempdir <dir>    directory to store temporary files (default: --outfile)
  --vcfinput         specify that input is in VCF format and output will be in VCF format
  --dot2underline    Function: automatically run a pipeline to predict protein functional effects in a
                     their functional effects in a
                     manual filtering
  --manualfiltering

Example: table_annovar.pl example
2014oct_all,1000g2014oct_afr,1000g201
table_annovar.pl example
oct_all,1000g2014oct_afr,1000g2014oct

Version: $Date: 2015-06-17 21:43
```

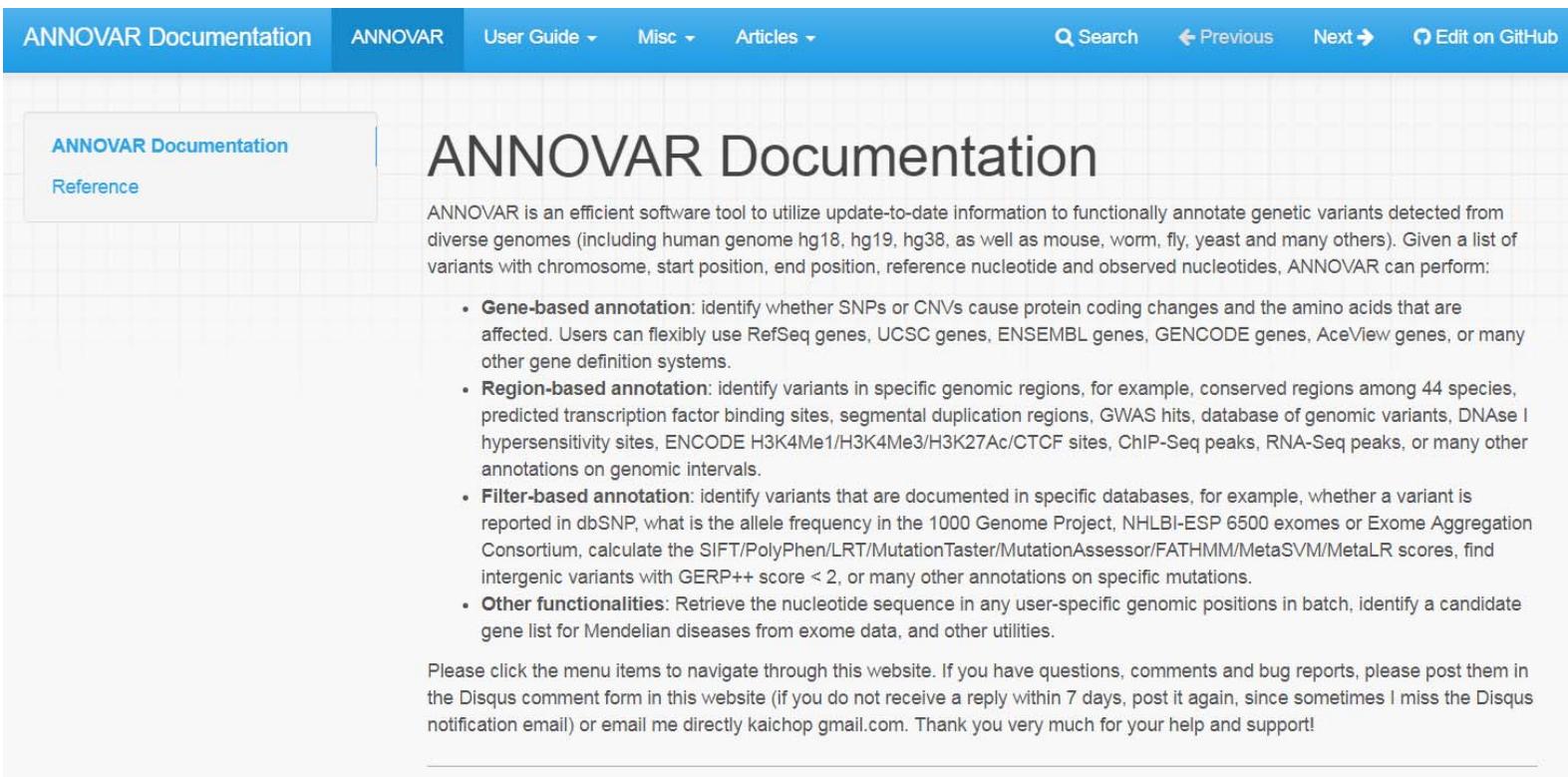


The screenshot shows the wANNOVAR website. At the top, there is a navigation bar with links for Home, Tutorial, Example, Related projects, and WLab. The main title "wANNOVAR" is prominently displayed. Below the title, a brief description of the tool is provided: "ANNOVAR is a rapid, efficient tool to annotate functional consequences of genetic variation from high-throughput sequencing data. wANNOVAR provides easy and intuitive web-based access to the most popular functionalities of the ANNOVAR software". There are three blue buttons labeled "Get Started", "About", and "Contact". At the bottom, there are social media sharing options: "Like" and "Share", with a note that 6 people like this.

# Similar tools

- Besides ANNOVAR, several other similar annotation tools have also been developed
- Command line programs include:
  - VEP
  - snpEff
  - VAAST
  - AnnTools
  - Jannovar
- Web servers include:
  - VAT
  - SeattleSeq
  - AVIA
  - VARIANT

# ANNOVAR website today



The screenshot shows the ANNOVAR Documentation website. The top navigation bar includes links for ANNOVAR Documentation, ANNOVAR, User Guide, Misc, Articles, Search, Previous, Next, and Edit on GitHub. A sidebar on the left has links for ANNOVAR Documentation and Reference. The main content area features a large title "ANNOVAR Documentation". Below the title, a paragraph describes ANNOVAR as an efficient software tool for annotating genetic variants across diverse genomes. A bulleted list details its functionalities: Gene-based annotation, Region-based annotation, Filter-based annotation, and Other functionalities. At the bottom, a note encourages users to click menu items for navigation and provides contact information for support.

ANNOVAR Documentation

Reference

## ANNOVAR Documentation

ANNOVAR is an efficient software tool to utilize update-to-date information to functionally annotate genetic variants detected from diverse genomes (including human genome hg18, hg19, hg38, as well as mouse, worm, fly, yeast and many others). Given a list of variants with chromosome, start position, end position, reference nucleotide and observed nucleotides, ANNOVAR can perform:

- **Gene-based annotation:** identify whether SNPs or CNVs cause protein coding changes and the amino acids that are affected. Users can flexibly use RefSeq genes, UCSC genes, ENSEMBL genes, GENCODE genes, AceView genes, or many other gene definition systems.
- **Region-based annotation:** identify variants in specific genomic regions, for example, conserved regions among 44 species, predicted transcription factor binding sites, segmental duplication regions, GWAS hits, database of genomic variants, DNase I hypersensitivity sites, ENCODE H3K4Me1/H3K4Me3/H3K27Ac/CTCF sites, ChIP-Seq peaks, RNA-Seq peaks, or many other annotations on genomic intervals.
- **Filter-based annotation:** identify variants that are documented in specific databases, for example, whether a variant is reported in dbSNP, what is the allele frequency in the 1000 Genome Project, NHLBI-ESP 6500 exomes or Exome Aggregation Consortium, calculate the SIFT/PolyPhen/LRT/MutationTaster/MutationAssessor/FATHMM/MetaSVM/MetaLR scores, find intergenic variants with GERP++ score < 2, or many other annotations on specific mutations.
- **Other functionalities:** Retrieve the nucleotide sequence in any user-specific genomic positions in batch, identify a candidate gene list for Mendelian diseases from exome data, and other utilities.

Please click the menu items to navigate through this website. If you have questions, comments and bug reports, please post them in the Disqus comment form in this website (if you do not receive a reply within 7 days, post it again, since sometimes I miss the Disqus notification email) or email me directly kaichop@gmail.com. Thank you very much for your help and support!

<http://annovar.openbioinformatics.org/en/latest/>

# Example command and output

- [kaiwang@biocluster ~]\$ table\_annovar.pl example/ex2.vcf humandb/ -buildver hg19 -out myanno -remove **-protocol refGene,cytoBand,genomicSuperDups,esp6500siv2\_all,1000g2015aug\_all,1000g2015aug\_eur,exac03,avsnpl47,dbnsfp30a** -operation g,r,r,f,f,f,f,f,f -nastring . -vcfinput

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	
1	Chr	Start	End	Ref	Alt	Func.ref	Gene.ref	GeneDet	ExonicFv	AAChange	cytoBand	genomic	esp6500	1000g20	1000g20	1000g20	1000g20	snp138	SIFT_soc	SIFT_pre	
2	1	948921	948921	T	C	UTR5	ISG15	NM_0051	.	.	1p36.33	.	0.8858	0.9032	0.7201	0.999	0.9573	rs15842	.	.	
3	1	1E+06	1E+06	G	T	UTR3	ATAD3C	NM_0010	.	.	1p36.33	Score=0.	0.0279	0.0677	0.031	0.126	0.0388	rs149123	.	.	
4	1	6E+06	6E+06	A	T	splicing	NPHP4	NM_0012	.	.	1p36.31	.	0.8431	0.8433	0.91	0.7907	0.8111	rs128763	.	.	
5	1	2E+08	2E+08	C	T	intronic	DDR2	.	.	.	1q23.3	.	0.6206	0.385	0.5923	0.8529	0.8529	rs100005	.	.	
6	1	8E+07	8E+07	C	T	intronic	DNASE2I	.	.	.	1p31.1	.	0.5489	0.5303	0.6062	0.5239	0.5239	rs657670	.	.	
7	1	1E+07	1E+07	TC	-	intergenic	LOC6451	dist=1156	.	.	1p36.21	Score=0.	.	.	.	.	.	.	.	.	
8	1	1E+07	1E+07	-	AT	intergenic	UBIAD1_F	dist=5510	.	.	1p36.22	.	.	.	.	.	.	rs355611	.	.	
9	1	1E+08	1E+08	A	ATAAA	intergenic	LOC1001	dist=8721	.	.	1p21.1	.	.	.	.	.	.	.	.	.	
10	1	7E+07	7E+07	G	A	exonic	IL23R	.	nonsynon	IL23R:NM_1p31.3	.	0.0469	0.0228	0.003	.	0.0616	rs112090	0.1	T		
11	2	2E+08	2E+08	A	G	exonic	ATG16L1	.	nonsynon	ATG16L1_2q37.1	.	0.4563	0.396	0.3101	0.3224	0.5368	rs224188	0.57	T		
12	16	5E+07	5E+07	C	T	exonic	NOD2	.	nonsynon	NOD2:NM_16q12.1	.	0.0316	0.0144	0.0023	.	0.0507	rs206684	0.01	D		
13	16	5E+07	5E+07	G	C	exonic	NOD2	.	nonsynon	NOD2:NM_16q12.1	.	0.0102	0.0046	.	.	0.0099	rs206684	0.02	D		
14	16	5E+07	5E+07	-	C	exonic	NOD2	.	frameshift	NOD2:NM_16q12.1	.	0.0162	0.006	0.0038	.	0.0139	rs206684	.	.		
U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	
Polyph	Polyph	Polyph	Polyph	LRT_soc	LRT_pre	Mutation	Mutation	Mutation	FATHMM	FATHMM	RadialSv	RadialSv	LR_score	LR_pred	VEST3_s	CADD_ra	CADD_pl	GERP++_phyloP46			
.	.	.	.	.	.	0	P	.	.	.	.	.	.	.	.	0.905	8.681	4.93	0.745		
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.		
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.		
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.		
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.		
1 D	0.999	D	0 D	1 D	1.935	M	0.71	T	-0.599	T	0.275	T	0.655	4.183	21.7	5.19	2.865				
0.001	B	0.005	B	0.07	N	0.998	P	-0.255	N	1.71	T	-1.007	T	0.192	1.505	10.98	-11.4	-1.999			
0.999	D	0.901	P	0.993	N	1 N	2.32	M	-0.62	T	-0.855	T	0.138	T	0.219	3.578	18.23	3.66	1.421		
1 D	0.986	D	0 D	1 D	1.79	L	0.57	T	-0.696	T	0.138	T	0.94	4.373	23.1	5.91	2.813				

# New user: what are r,g,f?

- ANNOVAR supports 3 types of annotations:
  - Gene-based annotation (g)
    - What is the consequence of a variant on gene: intronic, intergenic, non-synonymous, frameshift, etc.
  - Region-based annotation (r)
    - Whether the region overlaps with specific genomic regions, such as conserved sites, ENCODE peaks, microRNA target sites, cytogenetic bands or common structural variations
  - Filter-based annotation (f)
    - Whether the exact variant has been reported in databases, such as dbSNP, 1000 Genomes Project, NHLBI-ESP6500 project, COSMIC database, NCI60 database
    - What are the SIFT, PolyPhen, MutationTaster, MutationAssessor, LRT scores for a non-synonymous variant

# Gene-based annotation

- Several commonly used gene definitions:
  - RefSeq gene
  - UCSC known gene
  - ENSEMBL gene
  - GENCODE gene
- The annotation is based on precedence rules
  - Variant\_function (exonic, intronic, intergenic, etc)
  - Exonic\_variant\_function (nonsense, synonymous, etc)

# What is RefSeqGene?

Display Settings:  Graphics      Send:

## Homo sapiens hemochromatosis (HFE), RefSeqGene on chromosome 6

NCBI Reference Sequence: NG\_008720.1  
[GenBank](#) [FASTA](#)

[Link To This Page](#) | [Help](#) | [Feedback](#) | [Printer-Friendly Page](#)

**NG\_008720.1 (14,961 bases)**

Sequence NG\_008720.1: Homo sapiens hemochromatosis (HFE), RefSeqGene on chromosome 6

**1 : 14,961 (14,961 bases shown, positive strand)**

**SNPs** **Clinically Associated Variants** **Cited Variants** **Gene** **Alignments**

**Marker** **Search...**

**Analyze this sequence**

- Run BLAST
- Pick Primers
- Highlight Sequence Features

**Articles about the HFE gene**

- Role of HFE gene mutations on developing iron overload in beta-thal [East Mediterr Health J. 2011]
- Iron overload and HFE gene mutations in Polish patients with [Hepatobiliary Pancreat Dis Int. 2011]
- Genome-wide association study identifies two loci strongly affecting transferrin [Hum Mol Genet. 2011]

[See all...](#)

**Variation viewer**

See a summary of HFE variations, including those of clinical significance.

**Reference sequence information**

RefSeq alternative splicing  
See 9 reference mRNA sequence splice variants for the HFE gene.

**More about the HFE gene**

The protein encoded by this gene is a membrane protein that is similar to MHC class I-type proteins and associates with beta2-microglobulin ...  
Also Known As: HFE1, HH, HLA-H, IMAGE:...

# Gene model differences

- **RefSeq gene**: a collection of non-redundant, curated mRNA models
- UCSC gene: constructed by a fully automated process, based on protein data from Swiss-Prot/TrEMBL (UniProt) and the associated mRNA data from Genbank.
- Ensembl: contains more gene models from multiple sources (including RefSeq) mapped to the reference genome.
- **GENCODE gene**: combination of computational analysis, manual annotation, and experimental validation.

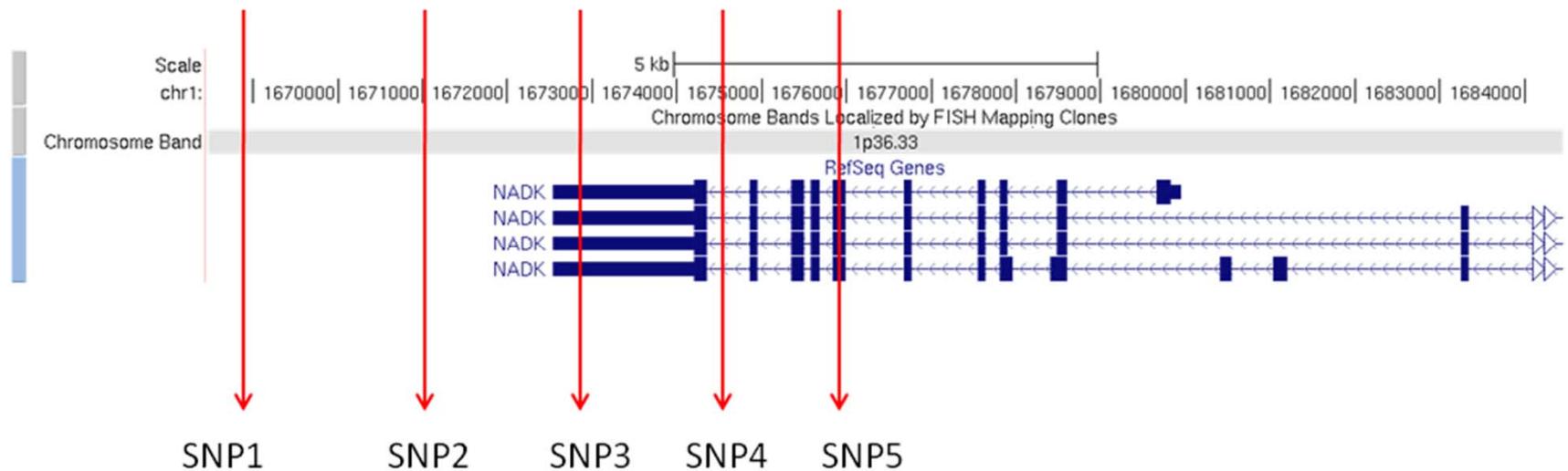
**There are ~50K, ~80K, ~200K, ~190K transcripts in the four gene models, respectively**

# Variant\_function precedence

Value	Default precedence	Explanation	Sequence Ontology
exonic	1	variant overlaps a coding	exon_variant (SO:0001791)
splicing	1	variant is within 2-bp of a splicing junction (use -splicing_threshold to change this)	splicing_variant (SO:0001568)
ncRNA	2	variant overlaps a transcript without coding annotation in the gene definition (see Notes below for more explanation)	non_coding_transcript_variant (SO:0001619)
UTR5	3	variant overlaps a 5' untranslated region	5_prime_UTR_variant (SO:0001623)
UTR3	3	variant overlaps a 3' untranslated region	3_prime_UTR_variant (SO:0001624)
intronic	4	variant overlaps an intron	intron_variant (SO:0001627)
upstream	5	variant overlaps 1-kb region upstream of transcription start site	upstream_gene_variant (SO:0001631)
downstream	5	variant overlaps 1-kb region downstream of transcription end site (use -neargene to change this)	downstream_gene_variant (SO:0001632)
intergenic	6	variant is in intergenic region	intergenic_variant (SO:0001628)

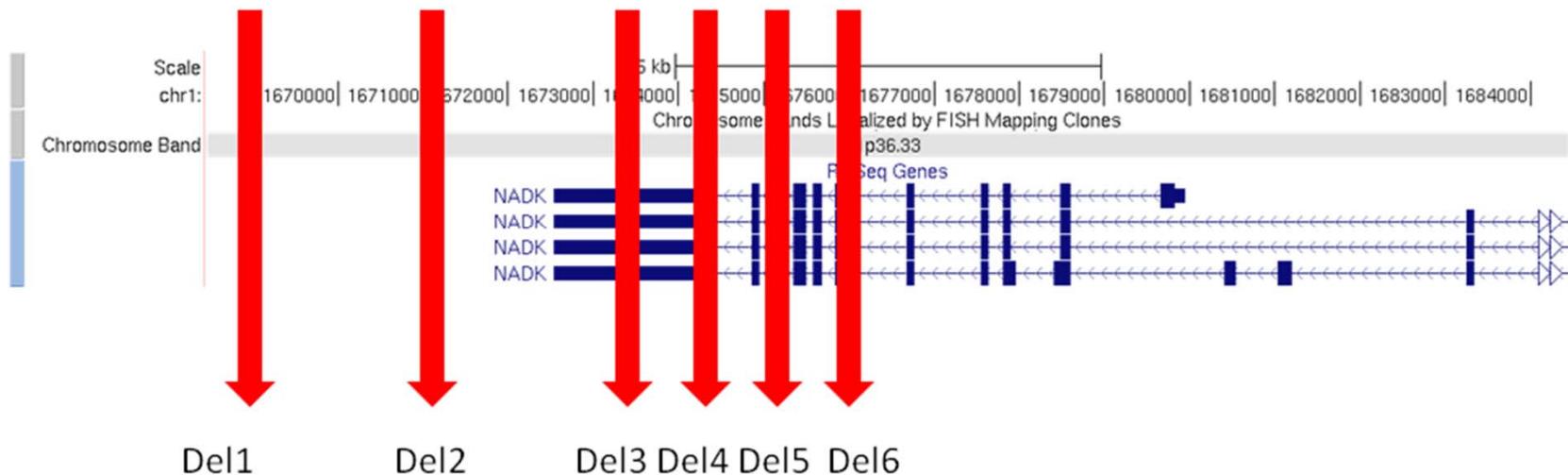
## Example: SNVs

- SNP1 is an intergenic variant, as it is >1kb away from any gene
- SNP2 is a downstream variant, as it is 1kb from the 3'end of the NADK gene
- SNP3 is a UTR3 variant
- SNP4 is an intronic variant
- SNP5 is an exonic variant



## Example: indels

- Deletion 1 is an intergenic variant;
- Deletion 2 is a downstream variant;
- Deletion 3 is a UTR3 variant;
- Deletion 4 overlaps both with UTR3 and intron, and based on the precedence rule, it is a UTR3 variant;
- Deletion 5 is an intronic variant;
- Deletion 6 overlaps with both an exon and an intron, and based on the precedence rule, it is an exonic variant.



# Exonic\_variant\_function precedence

Annotation	Precedence	Explanation	Sequence Ontology
frameshift insertion	1	an insertion of one or more nucleotides that cause frameshift changes in protein coding sequence	frameshift_elongation (SO:0001909)
frameshift deletion	2	a deletion of one or more nucleotides that cause frameshift changes in protein coding sequence	frameshift_truncation (SO:0001910)
frameshift block substitution	3	a block substitution of one or more nucleotides that cause frameshift changes in protein coding sequence	frameshift_variant (SO:0001589)
stopgain	4	a nonsynonymous SNV, frameshift insertion/deletion, nonframeshift insertion/deletion or block substitution that lead to the immediate creation of stop codon at the variant site. For frameshift mutations, the creation of stop codon downstream of the variant will not be counted as "stopgain"!	stop_gained (SO:0001587)
stoploss	5	a nonsynonymous SNV, frameshift insertion/deletion, nonframeshift insertion/deletion or block substitution that lead to the immediate elimination of stop codon at the variant site	stop_lost (SO:0001578)
nonframeshift insertion	6	an insertion of 3 or multiples of 3 nucleotides that do not cause frameshift changes in protein coding sequence	inframe_insertion (SO:0001821)
nonframeshift deletion	7	a deletion of 3 or multiples of 3 nucleotides that do not cause frameshift changes in protein coding sequence	inframe_deletion (SO:0001822)
nonframeshift block substitution	8	a block substitution of one or more nucleotides that do not cause frameshift changes in protein coding sequence	inframe_variant (SO:0001650)
nonsynonymous SNV	9	a single nucleotide change that cause an amino acid change	missense_variant (SO:0001583)
synonymous SNV	10	a single nucleotide change that does not cause an amino acid change	synonymous_variant (SO:0001819)
unknown	11	unknown function (due to various errors in the gene structure definition in the database file)	sequence_variant (SO:0001060)

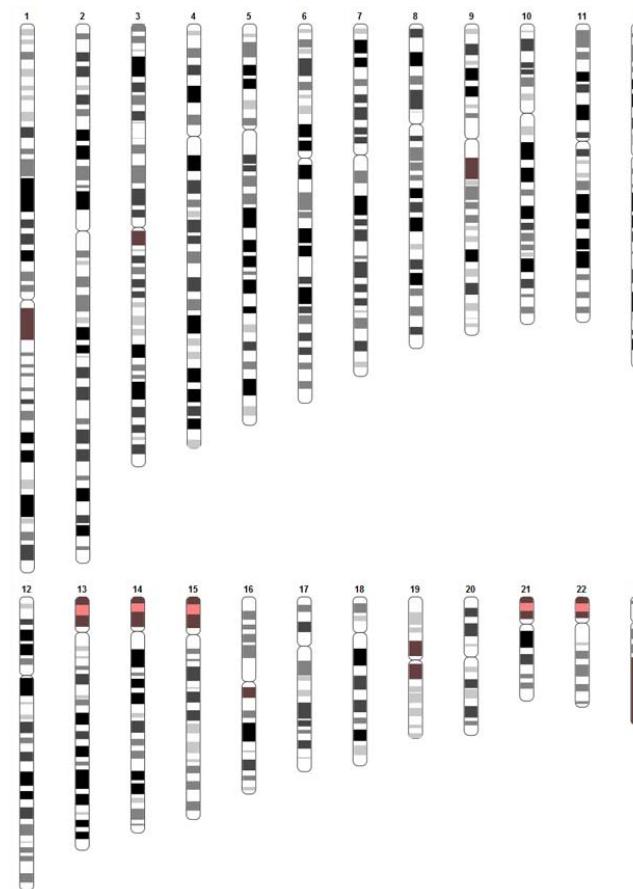
# Region-based annotation

- Several commonly used filter annotation:
  - Cytogenetic band
  - Located in a ChIP-Seq peaks from ENCODE
  - Located in a predicted repressor, promoter, enhancer, etc
  - Overlap with conserved genomic regions

# Cytogenetic band

- Keyword is cytoBand.

1	Chr	Start	End	Ref	Alt	cytoBand
2	1	948921	948921	T	C	1p36.33
3	1	1404001	1404001	G	T	1p36.33
4	1	5935162	5935162	A	T	1p36.31
5	1	162736463	162736463	C	T	1q23.3
6	1	84875173	84875173	C	T	1p31.1
7	1	13211293	13211294	TC	-	1p36.21
8	1	11403596	11403596	-	AT	1p36.22
9	1	105492231	105492231	A	ATAAA	1p21.1
10	1	67705958	67705958	G	A	1p31.3
11	2	234183368	234183368	A	G	2q37.1
12	16	50745926	50745926	C	T	16q12.1
13	16	50756540	50756540	G	C	16q12.1
14	16	50763778	50763778	-	C	16q12.1
15	13	20763686	20763686	G	-	13q12.11
16	13	20797176	21105944	O	-	13q12.11



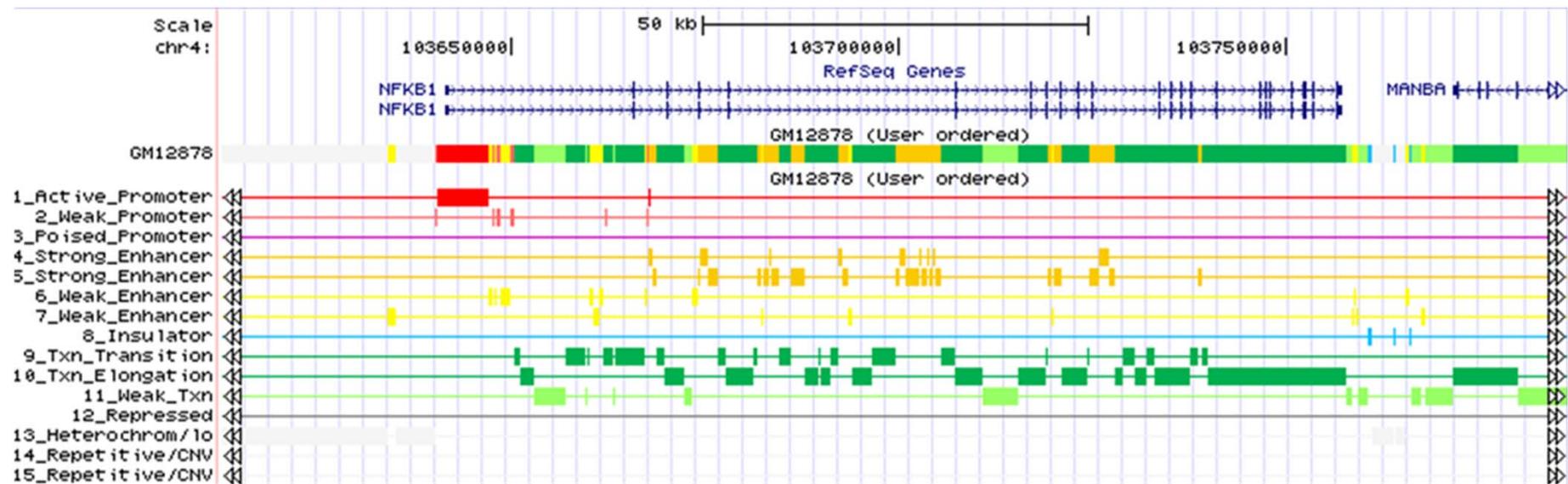
# ENCODE ChIP-Seq peaks

- General guideline:
  - Active promoter: H3K4me3, H3K9Ac
  - Active enhancer: H3K4me1, H3K27Ac
  - Active elongation: H3K36me3, H3K79me2
  - Repressed promoters and broad regions: H3K27me3, H3K9me3

Cell	Description	Lineage	Tissue	Karyotype
GM12878	B-lymphocyte, lymphoblastoid, International HapMap Project - CEPH/Utah - European Caucasian, Epstein-Barr Virus	mesoderm	blood	normal
H1-hESC	embryonic stem cells	inner cell mass	embryonic stem cell	normal
K562	leukemia, "The continuous cell line K-562 was established by Lozzio and Lozzio from the pleural effusion of a 53-year-old female with chronic myelogenous leukemia in terminal blast crises." - ATCC	mesoderm	blood	cancer
HepG2	hepatocellular carcinoma	endoderm	liver	cancer
HUVEC	umbilical vein endothelial cells	mesoderm	blood vessel	normal
HMEC	mammary epithelial cells	ectoderm	breast	normal
HSMM	skeletal muscle myoblasts	mesoderm	muscle	normal
NHEK	epidermal keratinocytes	ectoderm skin	normal	
NHLF	lung fibroblasts	endoderm	lung	normal

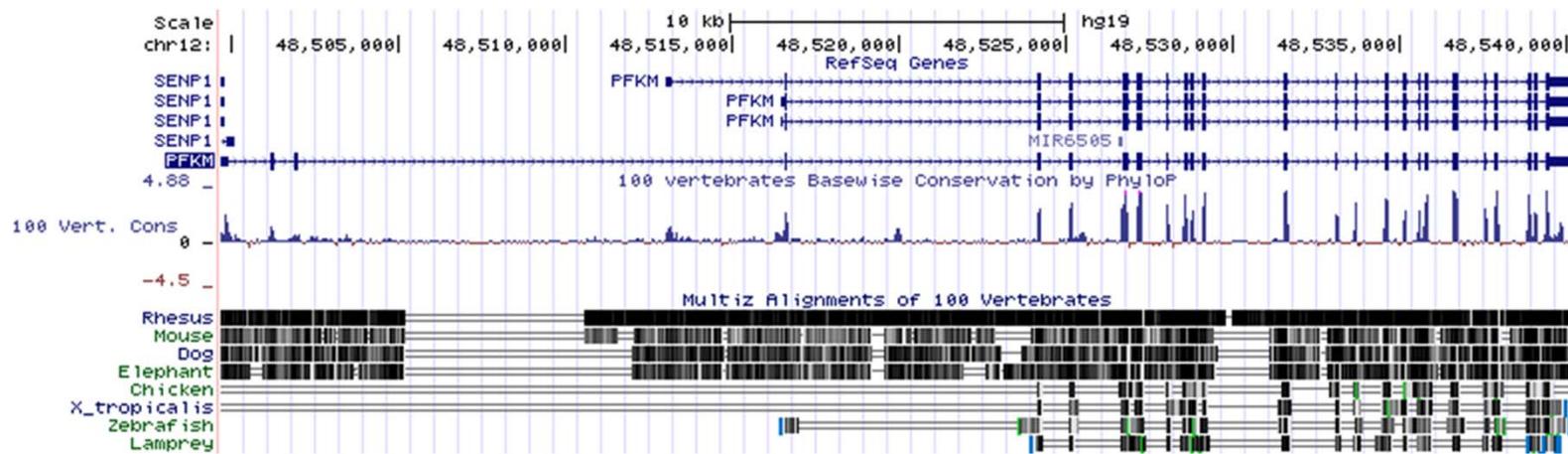
# chromHMM predictions

- ChromHMM integrate multiple ChIP-Seq datasets of various histone modifications to discover de novo the major re-occurring combinatorial and spatial patterns of marks.
- 15 different “states” are provided



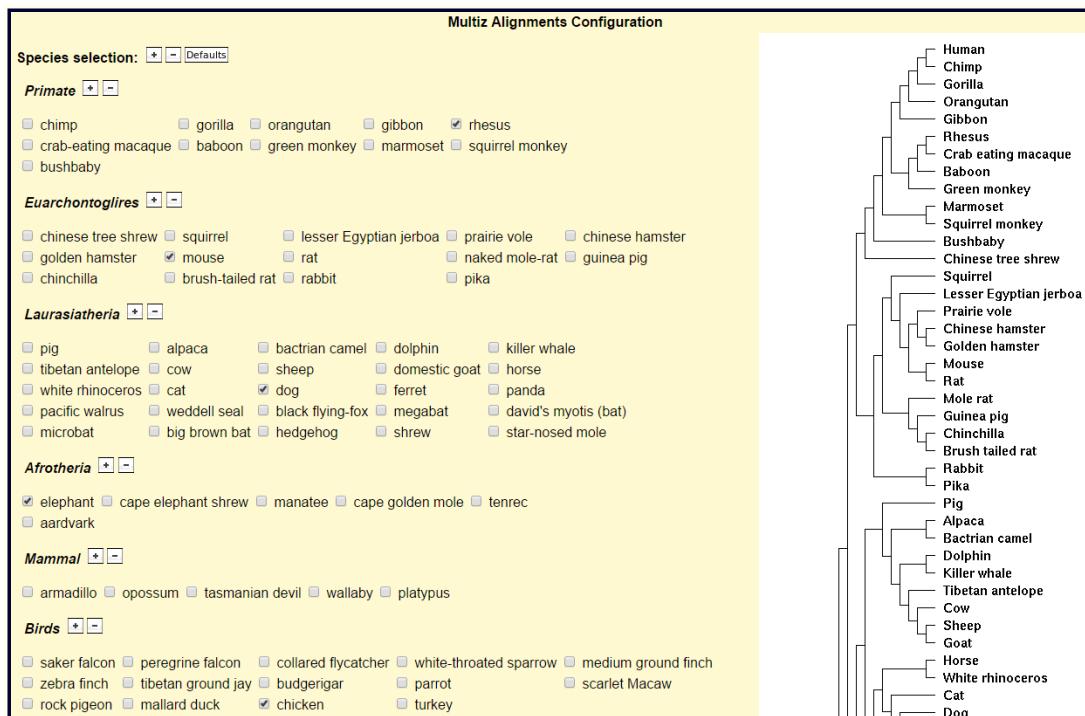
# Conserved genomic regions

- UCSC Genome Browser provides multi-species alignment for human genome sequence
- Conserved region may indicate functionally important regions



# Selection of alignment

- Several tracks to choose from (on hg19 coordinate):
  - 100-way alignment
  - 46-way alignment
  - GERP++ conserved elements



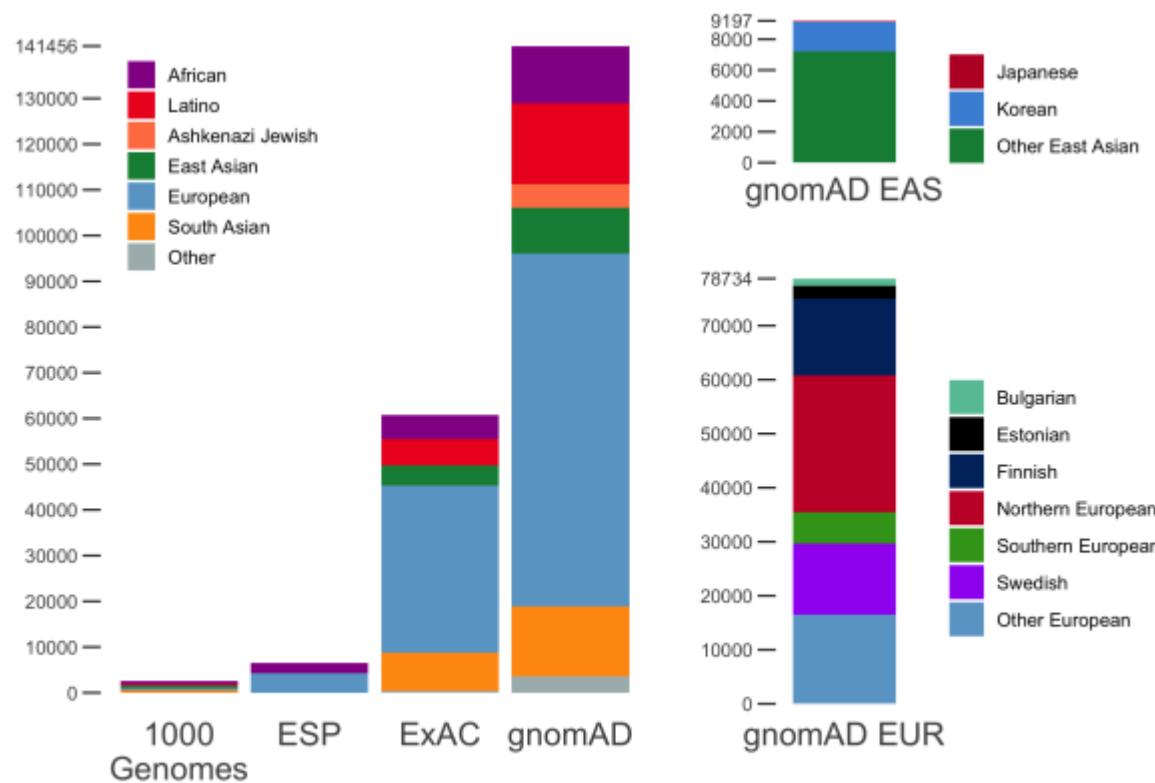
# Filter-based annotation

- Several commonly used filter annotation:
  - Allele frequency in various databases
  - Presence in various association databases
  - Functional prediction scores
  - dbSNP identifier

# Allele frequency databases

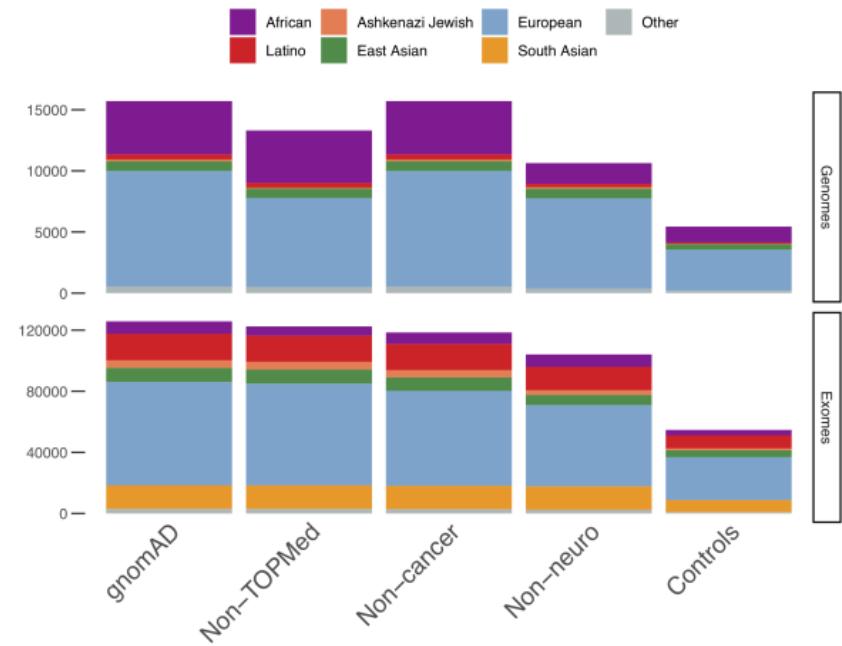
- 1000 Genomes Project: ~2500 genomes
  - Keyword is 1000g2015aug\_all
- ExAC (exome aggregation consortium): ~65000 exomes
  - Keyword is exac03
- NHLBI-ESP6500 (European Americans and African Americans): ~6500 exomes
  - Keyword is esp6500siv2\_aa, esp6500siv2\_ea
- gnomAD (genome aggregation consortium): 123,136 exome sequences and 15,496 whole-genome
  - Keyword is gnomad\_exome and genomad\_genome
- gnomAD2.1.1 (genome aggregation consortium): 125,748 exome sequences and 15,708 whole-genome
  - Keyword is gnomad221\_exome and genomad221\_genome

# Allele frequencies in different ethnicity groups



# Allele frequencies from different subset of data

- Non-TOPMed: only samples that are not present in the Trans-Omics for Precision Medicine (TOPMed)-[BRAVO](#) release. The allele counts in this subset can thus be added to those of [BRAVO](#) to federate both datasets.
- Non-cancer: Only samples from individuals who were not ascertained for having cancer in a cancer study
- Non-neuro: Only samples from individuals who were not ascertained for having a neurological condition in a neurological case/control study
- Controls-only: Only samples from individuals who were not selected as a case in a case/control study of common disease



# Disease association databases

- ClinVar: ClinVar archives and aggregates information about relationships among variation and human health
  - Keyword: clinvar\_20170130
- COSMIC: somatic mutations in various cancer types
  - keyword: cosmic72, cosmic76, cosmic80, etc
- ICGC: somatic mutations in the International Cancer Genome Consortium
  - keyword: icgc21
- HGMD and others
  - If you have them as VCF files, you can annotate variants using ‘vcf’ as the keyword
  - If you have them as “generic” files, you can annotate variants using ‘generic’ as the keyword

# Functional prediction scores

- Exome scores: all scores for non-synonymous variants are taken from dbNSFP database now (keyword is dbnsfp33a)
  - SIFT
  - PolyPhen
  - LRT
  - MutationTaster
  - MutationAssessor
  - MetaSVM, etc.
- Genome scores: each database is over 200Gb
  - GERP++,
  - CADD,
  - DANN,
  - FATHMM,
  - GWAVA,
  - FunSeq2, etc.

# Review: Example command and output

- [kaiwang@biocluster ~]\$ table\_annovar.pl example/ex2.vcf humandb/ - buildver hg19 -out myanno -remove **-protocol** refGene,cytoBand,genomicSuperDups,esp6500siv2\_all,1000g2015aug\_all,1000g2015aug\_eur,exac03,avsnpl47,dbnsfp30a **-operation** g,r,r,f,f,f,f,f,f -nastring . -vcfinput
- We requested to generate 9 annotations (1 gene-based, 2 region-based, 6 filter-based)
- The input file is in VCF format
- The output file name prefix is “myanno”
- The genome build is hg19
- The ANNOVAR database is stored at humandb/ directory
- The Nastring is “.” when an annotation is not available

# Examine the output

- myanno.hg19\_multianno.txt
  - Each line in the file represents one variant from the input file.
  - It is a tab-delimited file with added annotations represented as extra columns, by the same order as the annotation types following the '--protocol' argument.
- myanno.hg19\_multianno.vcf
  - This will be a VCF file in which the INFO column has extra fields in the form 'key=value' separated by ';'. For example, 'Func.refGene=intronic;Gene.refGene=SAMD11'.
  - Each key-value pair represents one piece of ANNOVAR annotation. The output file can be further processed by genetic analysis software tools that are designed for the VCF file format.

# Input and output file

	CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA06986	NA06994	
	A	B	C	D	E	F	G	H	I	J	K	
1	Chr	Start	End	Ref	Alt	Func.refGene	Gene.refGene	GeneDetail.refGene	ExonicFunc.refGene	AAChange.refGene	cytoBand	genomicSt
2	1	948921	948921	T	C	UTR5	ISG15	NM_005101:c.-33T>C	.	.	1p36.33	.
3	1	1404001	1404001	G	T	UTR3	ATAD3C	NM_001039211:c.*91G>T	.	.	1p36.33	Score=0.90
4	1	5935162	5935162	A	T	splicing	NPHP4	NM_015102:exon22:c.2818-2T>A	.	.	1p36.31	.
5	1	162736463	162736463	C	T	intronic	DDR2	.	.	.	1q23.3	.
6	1	84875173	84875173	C	T	intronic	DNASE2B	.	.	.	1p31.1	.
7	1	13211293	13211294	TC	-	intergenic	HNRNPCP5,PRAMEF3	dist=26967;dist=116902	.	.	1p36.21	Score=0.99
8	1	11403596	11403596	-	AT	intergenic	UBIAD1,PTCHD2	dist=55105;dist=135699	.	.	1p36.22	.
9	1	105492231	105492231	A	ATAAA	intergenic	LOC100129138,NONE	dist=872538;dist=NONE	.	.	1p21.1	.
10	1	67705958	67705958	G	A	exonic	IL23R	.	nonsynonymous SNV	IL23R:NM_1447	1p31.3	.
11	2	234183368	234183368	A	G	exonic	ATG16L1	.	nonsynonymous SNV	ATG16L1:NM_1	2q37.1	.
12	16	50745926	50745926	C	T	exonic	NOD2	.	nonsynonymous SNV	NOD2:NM_022116q12.1	.	.
13	16	50756540	50756540	G	C	exonic	NOD2	.	nonsynonymous SNV	NOD2:NM_022116q12.1	.	.
14	16	50763778	50763778	-	C	exonic	NOD2	.	frameshift insertion	NOD2:NM_022116q12.1	.	.
15	13	20763686	20763686	G	-	exonic	GJB2	.	frameshift deletion	GJB2:NM_00401	13q12.11	.
16	13	20797176	21105944	0	-	exonic	CRY1,GIR6	.	frameshift deletion	GIR6:NM_001113q12.11	.	.

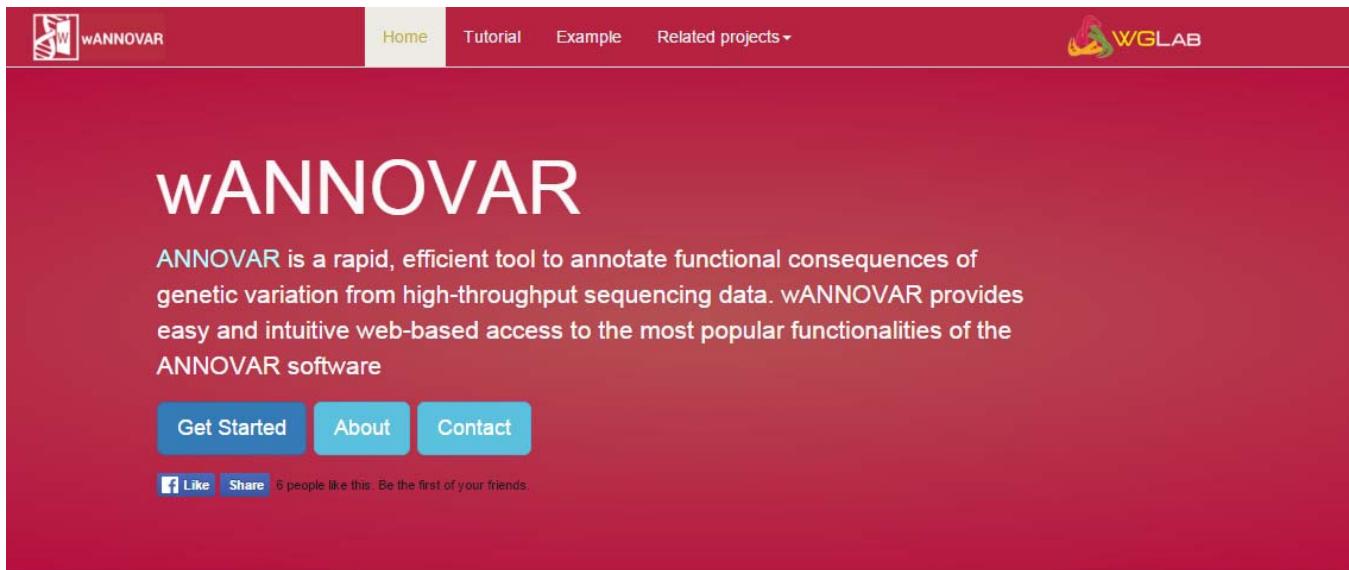
# Using web version of ANNOVAR

- Main advantage of wANNOVAR
  - No need to run command line
- Main limitations
  - Only a limited number of annotations are available
  - Many annotation databases are outdated, compared the latest ones that are available in ANNOVAR

# wANNOVAR

- URL:
  - Stable version: <http://wannovar.wglab.org>
- Function:
  - Users provide VCF file, returns annotated output in CSV format or tab-delimited format
  - Perform variants reduction to find disease genes
  - Display annotated output in web interface

# Main interface



The screenshot shows the wANNOVAR web application. At the top, there is a navigation bar with links for Home, Tutorial, Example, and Related projects. On the right side of the top bar is the WGLAB logo. The main title "wANNOVAR" is prominently displayed in large white letters on a red background. Below the title, a descriptive text explains what wANNOVAR is: "ANNOVAR is a rapid, efficient tool to annotate functional consequences of genetic variation from high-throughput sequencing data. wANNOVAR provides easy and intuitive web-based access to the most popular functionalities of the ANNOVAR software". There are three blue buttons labeled "Get Started", "About", and "Contact". Below these buttons, there are social sharing links for Facebook and Twitter, with a note that 6 people like this.

### Basic Information

Email

Sample Identifier

Input File

or Paste Variant Calls

I agree to the [Terms of Use](#)

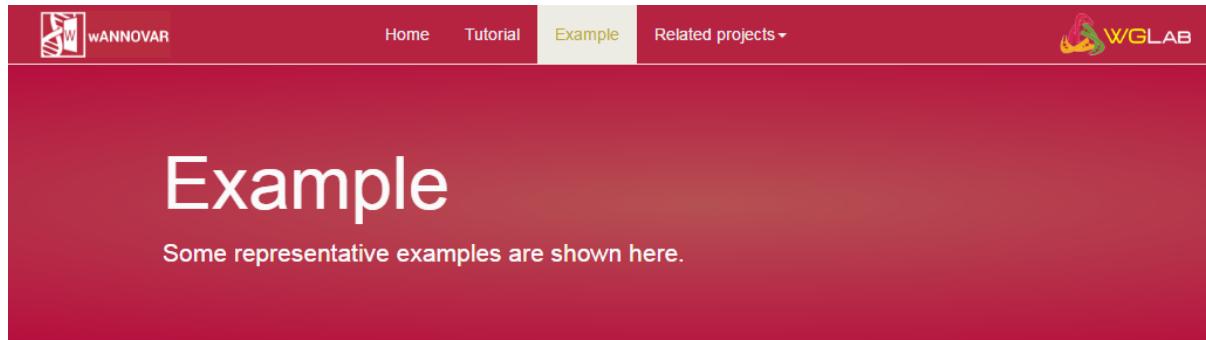
### Recent Updates

[10/22/2015] Now the filter is working for hg38! However, the custom filter is still not supported

[07/16/2015] Now we added another select called 'Individual Analysis', which is designed for VCF files. If you want to include all the individuals in your VCF file, please choose 'All annotations'. If you want to conduct individual based analysis (the first one if multiple samples are present), please choose 'Individual analysis'.

# Demo

- Go to <http://wannovar.wglab.org/example.html>



## Example

### Example 1

#### Exome sequencing data

we previously reported an exome sequencing study identifying a mutation in PKLR as 'unrelated finding' in a patient with hemolytic anemia, through a study originally designed to uncover the genetic basis of attention deficit/hyperactivity disorder (ADHD) 5. The VCF file is used as the input into wANNOVAR, with 'rare dominant Mendelian disease' selected as disease model. In total, 87 variants were left after the filtration, whose corresponding genes are then submitted automatically as input into Phenolyzer together with the term 'anemia' or 'hemolytic anemia', by wANNOVAR. From the result network, the PKLR gene is ranked top with the term 'hemolytic anemia'

#### Input:

[anemia.vcf](#)

#### Output:

[link to the result](#)

# Job submission

## Basic Information

Email

Sample Identifier

Input File  1) Upload variant file

or Paste Variant Calls  8) Submit

2) Submit

I agree to the Terms of Use

## Disease/Phenotype

Enter Disease or Phenotype Terms  2) Enter Phenotype or Disease terms

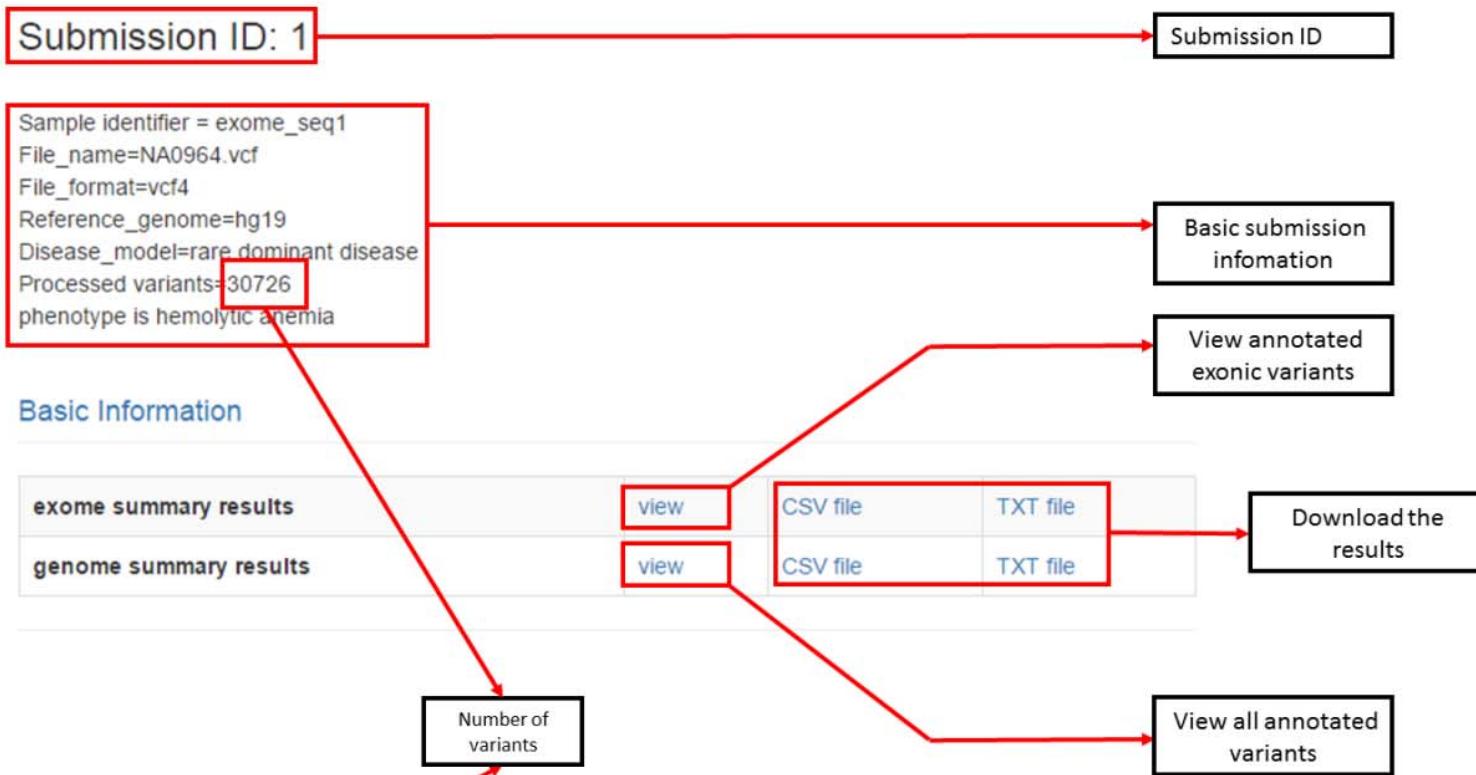
Please use semicolon or enter as separators. Like "alzheimer;brain".  
Try to use multiple terms instead of a super long term  
OMIM IDs are also accepted, like 114480 for 'Breast cancer'  
Better Combined with wANNOVAR's disease model.

## Parameter Settings

Result duration	<input type="text" value="2 months"/> 3) Choose how long you want your result reserved in the server
Reference Genome	<input type="text" value="hg19"/> 4) Choose Reference Genome version
Input Format	<input type="text" value="VCF"/> 5) Choose Input Format
Gene Definition	<input type="text" value="RefSeq Gene"/> 6) Choose gene definition
Individual analysis	<input type="text" value="Individual analysis"/> 7) Choose whether you want all the variants or only for one individual
Disease Model	<input type="text" value="none"/> 8) Choose disease model for variant filtration

# Results page

a



# Results page

## ANNOVAR filtering results:

(click to view details about this pipeline)

Initially 30726 variants were fed into the annotation pipeline and 0 variants were detected as invalid input.

[download all filtering results](#)

<b>Step1:7963 variants</b>	Identify missense, nonsense and splicing variants	<a href="#">download</a>
<b>Step2:584 variants</b>	Remove variants in the 1000 Genomes Project(ALL) with MAF>0.01	<a href="#">download</a>
<b>Step3:421 variants</b>	Remove variants in NHLBI-ESP 6500 exomes with MAF>0.01	<a href="#">download</a>
<b>Step4:80 variants</b>	Remove variants in dbSNP138 (excluding clinically associated SNPs)	<a href="#">download</a>
<b>Step5:76 genes</b>	Compile a list of candidate genes based on disease model	<a href="#">download</a>

download

download

download

download

Download the filtered variants

## Phenotype/disease Prioritization Result:

Exonic variant list from the wANNOVAR output (Total: 76)

[Variant List](#)

Download the variant list from ANNOVAR annotation or filter pipeline

Gene list from the wANNOVAR output, input into Phenolyzer (Total: 76)

[Input Gene List](#)

Download the input gene list for Phenolyzer

The prioritized genes from Phenolyzer (Total: 72)

[Result Gene List](#)

Download the prioritized gene list by Phenolyzer

The network visualization

[Show](#)

Link to the Phenolyzer Network

# From functional annotation to clinical interpretation

- In 2015, ACMG and AMP jointly developed standards and guidelines for the clinical interpretation of sequence variants with respect to human diseases.



## Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology

Sue Richards, PhD<sup>1</sup>, Nazneen Aziz, PhD<sup>2,16</sup>, Sherri Bale, PhD<sup>3</sup>, David Bick, MD<sup>4</sup>, Soma Das, PhD<sup>5</sup>, Julie Gastier-Foster, PhD<sup>6,7,8</sup>, Wayne W. Grody, MD, PhD<sup>9,10,11</sup>, Madhuri Hegde, PhD<sup>12</sup>, Elaine Lyon, PhD<sup>13</sup>, Elaine Spector, PhD<sup>14</sup>, Karl Voelkerding, MD<sup>13</sup> and Heidi L. Rehm, PhD<sup>15</sup>; on behalf of the ACMG Laboratory Quality Assurance Committee

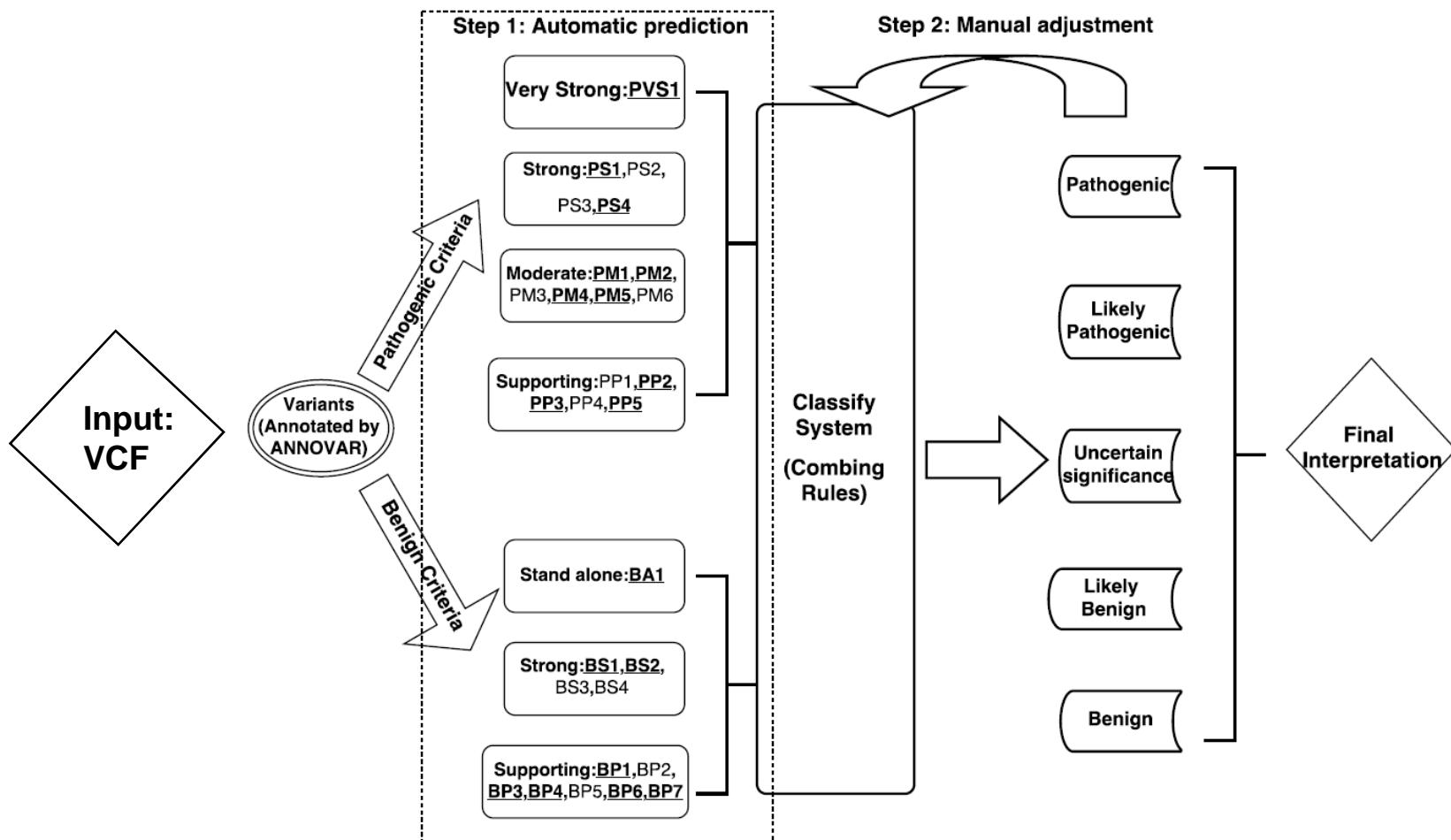
**Disclaimer:** These ACMG Standards and Guidelines were developed primarily as an educational resource for clinical laboratory geneticists to help them provide quality clinical laboratory services. Adherence to these standards and guidelines is voluntary and does not necessarily assure a successful medical outcome. These Standards and Guidelines should not be considered inclusive of all proper procedures and tests or exclusive of other procedures and tests that are reasonably directed to obtaining the same results. In determining the propriety of any specific procedure or test, the clinical laboratory geneticist should apply his or her own professional judgment to the specific circumstances presented by the individual patient or specimen. Clinical laboratory geneticists are encouraged to document in the patient's record the rationale for the use of a particular procedure or test, whether or not it is in conformance with these Standards and Guidelines. They also are advised to take notice of the date any particular guideline was adopted and to consider other relevant medical and scientific information that becomes available after that date. It also would be prudent to consider whether intellectual property interests may restrict the performance of certain tests and other procedures.

# 5-tier system for clinical interpretation

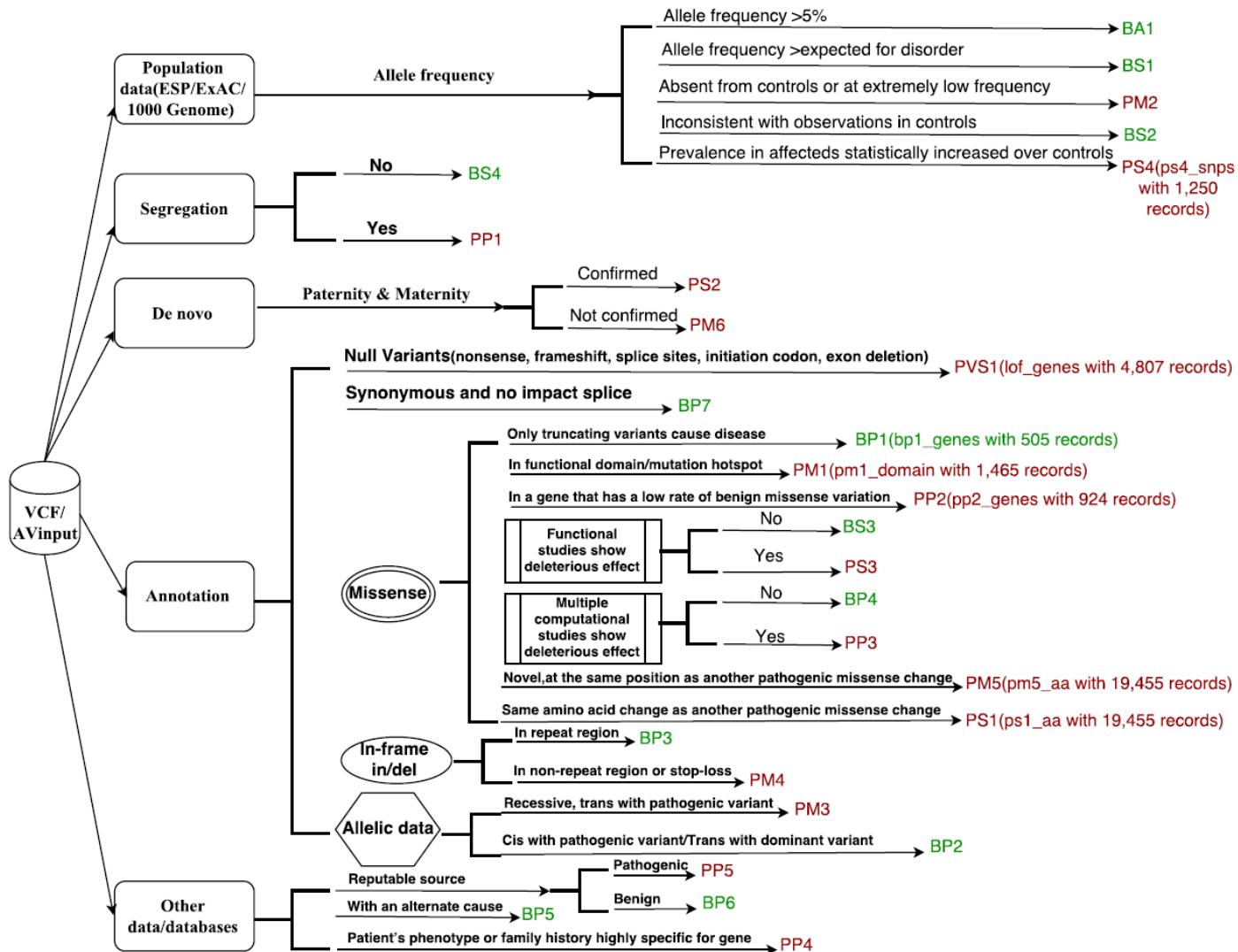
Pathogenic	(i) 1 Very strong (PVS1) AND (a) $\geq 1$ Strong (PS1–PS4) OR (b) $\geq 2$ Moderate (PM1–PM6) OR (c) 1 Moderate (PM1–PM6) and 1 supporting (PP1–PP5) OR (d) $\geq 2$ Supporting (PP1–PP5) (ii) $\geq 2$ Strong (PS1–PS4) OR (iii) 1 Strong (PS1–PS4) AND (a) $\geq 3$ Moderate (PM1–PM6) OR (b) 2 Moderate (PM1–PM6) AND $\geq 2$ Supporting (PP1–PP5) OR (c) 1 Moderate (PM1–PM6) AND $\geq 4$ supporting (PP1–PP5)	Benign	(i) 1 Stand-alone (BA1) OR (ii) $\geq 2$ Strong (BS1–BS4) Likely benign (i) 1 Strong (BS1–BS4) and 1 supporting (BP1–BP7) OR (ii) $\geq 2$ Supporting (BP1–BP7) Uncertain significance (i) Other criteria shown above are not met OR (ii) the criteria for benign and pathogenic are contradictory
Likely pathogenic	(i) 1 Very strong (PVS1) AND 1 moderate (PM1–PM6) OR (ii) 1 Strong (PS1–PS4) AND 1–2 moderate (PM1–PM6) OR (iii) 1 Strong (PS1–PS4) AND $\geq 2$ supporting (PP1–PP5) OR (iv) $\geq 3$ Moderate (PM1–PM6) OR (v) 2 Moderate (PM1–PM6) AND $\geq 2$ supporting (PP1–PP5) OR (vi) 1 Moderate (PM1–PM6) AND $\geq 4$ supporting (PP1–PP5)		

**The system uses a total of 28 criteria and a five-tiered categorization for classifying variants.**

# InterVar: Automated implementation of the ACMG-AMP clinical interpretation guidelines



# InterVar scoring logic



# Web implementation of InterVar

The American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) published in 2015 the updated standards and guidelines for the clinical interpretation of sequence variants, based on 28 criteria. However, variability between individual interpreters may be extensive due to lack of standard algorithms that implement these guidelines. This ACMG/AMP2015 guideline is [at here](#)

## Clinical Interpretation of genetic variants by ACMG/AMP 2015 guideline

InterVar is a bioinformatics software tool for clinical interpretation of genetic variants by the ACMG/AMP 2015 guideline. The input to InterVar is an annotated file generated from ANNOVAR, while the output of InterVar is the classification of variants into 'Benign', 'Likely benign', 'Uncertain significance', 'Likely pathogenic' and 'Pathogenic', together with detailed evidence code.

Search your **exonic** variants from pre-built wIntervar databases(built on 2017-April-30 22:58:49 80,077,300 records):

If you already know the criteria of your variant, you can [click here](#) to interpret your variant directly.

This server is for exon variants interpretation only, if you have indels, you need to download the intervar tool from [github](#), then interpret your variant on local.

Query by genomic coordinate

hg19 ▾ Chr 1 ▾ 115828756 Ref: G Alt: A

Query by dbSNP ID

rs.: rs373849532

Query by HGNC gene symbol

Gene: LEP cDNA change: c. G298A

<http://wintervar.wglab.org>

# Web implementation of InterVar

The American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) published in 2015 the updated standards and guidelines for the clinical interpretation of sequence variants, based on 28 criteria. However, variability between individual interpreters may be extensive due to lack of standard algorithms that implement these guidelines. This ACMG/AMP2015 guideline is [at here](#)

## Clinical Interpretation of genetic variants by ACMG/AMP 2015 guideline

InterVar is a bioinformatics software tool for clinical interpretation of genetic variants by the ACMG/AMP 2015 guideline. The input to InterVar is an annotated file generated from ANNOVAR, while the output of InterVar is the classification of variants into 'Benign', 'Likely benign', 'Likely significant', 'Likely pathogenic' and 'Pathogenic', together with detailed evidence code.

**Warning: All listed results were from default parameters!**  
Users are advised to examine detailed evidence and disease to perform manual adjustments.

You searched by chromosomal coordinates and Alleles  
build:hg19 Chr:1 Pos:115828756 Ref:G Alt:A

Show/hide columns    Restore columns    Copy to clipboard    Download result

Chr	Position	Ref	Alt	Gene (refGene)	InterVar
1	115828756	G	A	NGF	Likely pathogenic (Details&Adjust)

Showing 1 to 1 of 1 entries  
(Move mouse to popover or click the button of "Show/hide columns" for more info)

List of evidence in 28 criteria  
. means absent

**PM1:** Located in a mutational hot spot and/or critical and well-established functional domain (e.g., active site of an enzyme) without benign variation.  
**Cystine-knot cytokine:Nerve growth factor-related**

**PM2:** Absent from controls (or at extremely low frequency if recessive) in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium. **Allele Frequencies in ExAC:3.308E-5;in 1000 Genome:;in ESP:7.7E-5**

**PP2:** Missense variant in a gene that has a low rate of benign missense variation and in which missense variants are a common mechanism of disease.

**PP3:** Multiple lines of computational evidence support a deleterious effect on the gene or gene product (conservation, evolutionary, splicing)

Search:

Transcripts (Ref)	MAF in ExAC_ALL	Disease in OrphaNet	OMIM
NM_002506 p.R221W	3.308E-5 (show in 7 POPs)	64752	162030

Previous    1    Next

# Web implementation of InterVar

The Classify System is combining the rules from the Evidence System. The execution of our InterVar mainly consists of two major steps: 1) automatically interpretation by 28 criteria; and 2) manual adjustment by users to re-interpret the clinical significance.

Start wInterVar   About   Services ▾   Contact   Related projects ▾

**WGLAB**

**Clinical Interpretation of genetic variants by ACMG/AMP 2015 guideline**

InterVar is a bioinformatics software tool for clinical interpretation of genetic variants by the ACMG/AMP 2015 guideline. The input to InterVar is an annotated file generated from ANNOVAR, while the output of InterVar is the classification of variants into 'Benign', 'Likely benign', 'Uncertain significance', 'Likely pathogenic' and 'Pathogenic', together with detailed evidence code.

**Warning:** All listed results were from the automated interpretation on default parameters!  
Users are advised to examine detailed evidence and use prior knowledge on ethnicity/disease to perform manual adjustments.

You searched by chromosomal coordinates and Alleles  
build:**hg19** Chr:1 Pos:**115828756** Ref:**G** Alt:**A**

Show/hide columns   Restore columns   Copy to clipboard   Download result as CSV   Search:

Chr	Position	Ref	Alt	Gene (refGene)	Intervar	ExonicFunc (refGene)	SNP	Transcripts (Ref)	MAF in ExAC_ALL	Disease in OrphaNet	OMIM
1	115828756	G	A	NGF	Likely pathogenic <a href="#">(Details&amp;Adjust)</a>	nonsynonymous SNV	rs11466112(details of MAF)	NM_002506 p.R221W	3.308E-5 <a href="#">(show in 7 variants)</a>	64752	162030

Showing 1 to 1 of 1 entries  
(Move mouse to popover or click the button of "Show/hide columns" for more information)

Go back!

Disease information  
(- means absent, click to OrphaNet)

Orpha No: ORPHA64752

Syndrome(s): Congenital insensitivity  
to pain and thermal analgesia  
HSAN5

Prevalence: <1 / 1 000 000

Inheritance: Autosomal recessive

Age of onset: Infancy

Neonatal

OMIMs: 608654

# Web implementation of InterVar

Start wInterVar   About   Services ▾   Contact   Related projects ▾

 WGLAB

InterVar is a bioinformatics software tool for clinical interpretation of genetic variants by the ACMG/AMP 2015 guideline. The input to InterVar is an annotated file generated from ANNOVAR, while the output of InterVar is the classification of variants into 'Benign', 'Likely benign', 'Uncertain significance', 'Likely pathogenic' and 'Pathogenic', together with detailed evidence code.

**Re-Interpret your variant with position: 1:115828756 Ref:G Alt:A Gene: NGF**  
The automated clinical interpretation is : **Likely pathogenic**, but you can manually adjust it by checking/unchecking the criteria below

The blue color represents the criteria that need manual adjustment

PVS1: null variant (nonsense, frameshift, canonical +/- 2 splice sites, initiation codon, single or multiexon deletion) in a gene where LOF is a known mechanism of disease

Strong ▾ PS1: Same amino acid change as a previously established pathogenic variant regardless of nucleotide change

Strong ▾ PS2: De novo (both maternity and paternity confirmed) in a patient with the disease and no family history

Strong ▾ PS3: Well-established in vitro or in vivo functional studies supportive of a damaging effect on the gene or gene product

Strong ▾ PS4: The prevalence of the variant in affected individuals is significantly increased compared with the prevalence in controls

Strong ▾ PS5: The user has additional 1 ▾ strong pathogenic evidence

Moderate ▾ PM1: Located in a mutational hot spot and/or critical and well-established functional domain (e.g., active site of an enzyme) without benign variation

Moderate ▾ PM2: Absent from controls (or at extremely low frequency if recessive) in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium

Moderate ▾ PM3: For recessive disorders, detected in trans with a pathogenic variant

Moderate ▾ PM4: Protein length changes as a result of in-frame deletions/insertions in a nonrepeat region or stop-loss variants

Moderate ▾ PM5: Novel missense change at an amino acid residue where a different missense change determined to be pathogenic has been seen before

Moderate ▾ PM6: Assumed de novo, but without confirmation of paternity and maternity

Moderate ▾ PM7: The user has additional 1 ▾ moderate pathogenic evidence

Supporting ▾ PP1: Cosegregation with disease in multiple affected family members in a gene definitively known to cause the disease

Supporting ▾ PP2: Missense variant in a gene that has a low rate of benign missense variation and in which missense variants are a common mechanism of disease

Supporting ▾ PP3: Multiple lines of computational evidence support a deleterious effect on the gene or gene product (conservation, evolutionary, splicing impact, etc.)

Supporting ▾ PP4: Patient's phenotype or family history is highly specific for a disease with a single genetic etiology

Supporting ▾ PP5: Reputable source recently reports variant as pathogenic, but the evidence is not available to the laboratory to perform an independent evaluation

Supporting ▾ PP6: The user has additional 1 ▾ supporting pathogenic evidence

BA1: Allele frequency is >5% in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium

Strong ▾ BS1: Allele frequency is greater than expected for disorder

Strong ▾ BS2: Observed in a healthy adult individual for a recessive (homozygous), dominant (heterozygous), or X-linked (hemizygous) disorder, with full penetrance expected at an early age

Strong ▾ BS3: Well-established in vitro or in vivo functional studies show no damaging effect on protein function or splicing

Strong ▾ BS4: Lack of segregation in affected members of a family

Strong ▾ BS5: The user has additional 1 ▾ strong benign evidence

Supporting ▾ BP1: Missense variant in a gene for which primarily truncating variants are known to cause disease

Supporting ▾ BP2: Observed in trans with a pathogenic variant for a fully penetrant dominant gene/disorder or observed in cis with a pathogenic variant in any inheritance pattern

Supporting ▾ BP3: In-frame deletions/insertions in a repetitive region without a known function

Supporting ▾ BP4: Multiple lines of computational evidence suggest no impact on gene or gene product (conservation, evolutionary, splicing impact, etc.)

# Introduction to HPO

- The Human Phenotype Ontology (HPO)
  - Aims to provide a standardized vocabulary of phenotypic abnormalities encountered in human disease.
  - Each term in the HPO describes a phenotypic abnormality, such as atrial septal defect.
  - Currently contains approximately 13,000 terms (still growing) and over 156,000 annotations to hereditary diseases.
  - Also provides a large set of HPO annotations to approximately 4000 common diseases.

Human Phenotype Ontology

Home About Downloads Tools Documentation Users History  
FAQ License Citation Contact

March 2018 release  
March 9, 2018 March 2018 release

Join our mailing list  
February 14, 2018 HPO mailing list for news announcements

New format for the HPO annotation data  
February 7, 2018 New format for the HPO annotation data

January 2018 release  
January 26, 2018 January 2018 release

Page: 1 of 9 [Next»](#)



hpo

Add content to HPO  
Suggest change to HPO  
HPO mailinglist  
HPO browser  
Twitter  
Contact

# HPO is widely used

- Original paper published in 2008, cited > 500 times
- Progress paper published in 2014, cited > 500 times
- Many databases and tools are supporting or adopting it (see Table in the 2017 HPO paper)

AJHG

Volume 83, Issue 5, 17 November 2008, Pages 610–615



Report

## The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease

Peter N. Robinson<sup>1, 2</sup>, Sebastian Köhler<sup>1, 2</sup>, Sebastian Bauer<sup>1</sup>, Dominik Seelow<sup>1, 3</sup>, Denise Horn<sup>1</sup>, Stefan Mundlos<sup>1, 2, 4</sup>

D966–D974 *Nucleic Acids Research*, 2014, Vol. 42, Database issue  
doi:10.1093/nar/gkt1026

Published online 11 November 2013

## The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data

Sebastian Köhler<sup>1,2,\*</sup>, Sandra C. Doelken<sup>1</sup>, Christopher J. Mungall<sup>3</sup>, Sebastian Bauer<sup>1</sup>, Helen V. Firth<sup>4,5</sup>, Isabelle Bailleul-Forestier<sup>6</sup>, Graeme C. M. Black<sup>7,8</sup>, Danielle L. Brown<sup>9</sup>, Michael Brudno<sup>10,11</sup>, Jennifer Campbell<sup>9,12</sup>, David R. FitzPatrick<sup>13</sup>, Janan T. Eppig<sup>14</sup>, Andrew P. Jackson<sup>13</sup>, Kathleen Freson<sup>15</sup>, Marta Girdea<sup>10,11</sup>, Ingo Helbig<sup>16</sup>, Jane A. Hurst<sup>17</sup>, Johanna Jähn<sup>16</sup>, Laird G. Jackson<sup>18</sup>, Anne M. Kelly<sup>19</sup>, David H. Ledbetter<sup>20</sup>, Sahar Mansour<sup>21</sup>, Christa L. Martin<sup>20</sup>, Celia Moss<sup>22</sup>, Andrew Mumford<sup>23</sup>, Willem H. Ouwehand<sup>4,19</sup>, Soo-Mi Park<sup>6</sup>, Erin Rooney Riggs<sup>20</sup>, Richard H. Scott<sup>24</sup>, Sanjay Sisodiya<sup>25</sup>, Steven Van Vooren<sup>26</sup>, Ronald J. Wapner<sup>27</sup>, Andrew O. M. Wilkie<sup>28</sup>, Caroline F. Wright<sup>4</sup>, Anneke T. Vulto-van Silfhout<sup>29</sup>, Nicole de Leeuw<sup>29</sup>, Bert B. A. de Vries<sup>29</sup>, Nicole L. Washington<sup>3</sup>, Cynthia L. Smith<sup>14</sup>, Monte Westerfield<sup>30</sup>, Paul Schofield<sup>14,31</sup>, Barbara J. Ruef<sup>30</sup>, Georgios V. Gkoutos<sup>32</sup>, Melissa Haendel<sup>33</sup>, Damian Smedley<sup>4</sup>, Suzanna E. Lewis<sup>3</sup> and Peter N. Robinson<sup>1,2,34,\*</sup>

Published online 24 November 2016

*Nucleic Acids Research*, 2017, Vol. 45, Database issue D865–D876  
doi: 10.1093/nar/gkw1039

## The Human Phenotype Ontology in 2017

Sebastian Köhler<sup>1,\*</sup>, Nicole A. Vasilevsky<sup>2</sup>, Mark Engelstad<sup>2</sup>, Erin Foster<sup>2</sup>, Julie McMurry<sup>2</sup>, Ségoïène Aymé<sup>3</sup>, Gareth Baynam<sup>4,5</sup>, Susan M. Bell<sup>6</sup>, Cornelius F. Boerkel<sup>7</sup>, Kym M. Boycott<sup>8</sup>, Michael Brudno<sup>9</sup>, Orion J. Buske<sup>9</sup>, Patrick F. Chinnery<sup>10,11</sup>, Valentina Cipriani<sup>12,13</sup>, Laureen E. Connell<sup>14</sup>, Hugh J.S. Dawkins<sup>15</sup>, Laura E. DeMare<sup>14</sup>, Andrew D. Devereau<sup>16</sup>, Bert B.A. de Vries<sup>17</sup>, Helen V. Firth<sup>18</sup>, Kathleen Freson<sup>19</sup>, Daniel Greene<sup>20,21</sup>, Ada Hamosh<sup>22</sup>, Ingo Helbig<sup>23,24</sup>, Courtney Hum<sup>25</sup>, Johanna A. Jähn<sup>24</sup>, Roger James<sup>11,21</sup>, Roland Krause<sup>26</sup>, Stanley J. F. Laulederkind<sup>27</sup>, Hanns Lochmüller<sup>28</sup>, Gholson J. Lyon<sup>29</sup>, Soichi Ogishima<sup>30</sup>, Annie Olry<sup>31</sup>, Willem H. Ouwehand<sup>20</sup>, Nikolas Pontikos<sup>12,13</sup>, Ana Rath<sup>31</sup>, Franz Schaefer<sup>32</sup>, Richard H. Scott<sup>16</sup>, Michael Segal<sup>33</sup>, Panagiotis I. Sergouniotis<sup>34</sup>, Richard Sever<sup>14</sup>, Cynthia L. Smith<sup>6</sup>, Volker Straub<sup>28</sup>, Rachel Thompson<sup>28</sup>, Catherine Turner<sup>28</sup>, Ernest Turro<sup>20,21</sup>, Marijcke W.M. Veltman<sup>11</sup>, Tom Vulliamy<sup>35</sup>, Jing Yu<sup>36</sup>, Julie von Ziegenweidt<sup>20</sup>, Andreas Zankl<sup>37,38</sup>, Stephan Züchner<sup>39</sup>, Tomasz Zemojtel<sup>1</sup>, Julius O.B. Jacobsen<sup>16</sup>, Tudor Groza<sup>40,41</sup>, Damian Smedley<sup>16</sup>, Christopher J. Mungall<sup>42</sup>, Melissa Haendel<sup>2</sup> and Peter N. Robinson<sup>43,44,\*</sup>

# HPO in 2019: new website and expanded knowledge base

D1018–D1027 *Nucleic Acids Research*, 2019, Vol. 47, Database issue  
doi: 10.1093/nar/gky1105

Published online 22 November 2018

## Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources

Sebastian Köhler<sup>①,2,3</sup>, Leigh Carmody<sup>3,4</sup>, Nicole Vasilevsky<sup>③,5</sup>, Julius O.B. Jacobsen<sup>3,6</sup>, Daniel Danis<sup>3,4</sup>, Jean-Philippe Gourdine<sup>③,5</sup>, Michael Gargano<sup>3,4</sup>, Nomi L. Harris<sup>3,7</sup>, Nicolas Matentzoglu<sup>3,8</sup>, Julie A. McMurry<sup>③,9</sup>, David Osumi-Sutherland<sup>3,8</sup>, Valentina Cipriani<sup>③,10,11,12</sup>, James P. Balhoff<sup>③,13</sup>, Tom Conlin<sup>③,9</sup>, Hannah Blau<sup>③,4</sup>, Gareth Baynam<sup>14,15,16,17,18</sup>, Richard Palmer<sup>17</sup>, Dylan Gratian<sup>14</sup>, Hugh Dawkins<sup>18</sup>, Michael Segal<sup>19</sup>, Anna C. Jansen<sup>20,21</sup>, Ahmed Muaz<sup>3,22</sup>, Willie H. Chang<sup>23</sup>, Jenna Bergerson<sup>24</sup>, Stanley J.F. Laulederkind<sup>②5</sup>, Zafer Yüksel<sup>②6</sup>, Sergi Beltran<sup>②7,28</sup>, Alexandra F. Freeman<sup>24</sup>, Panagiotis I. Sergouniotis<sup>29</sup>, Daniel Durkin<sup>4</sup>, Andrea L. Storm<sup>30,31</sup>, Marc Hanauer<sup>32</sup>, Michael Brudno<sup>23</sup>, Susan M. Bello<sup>③33</sup>, Murat Sincan<sup>34</sup>, Kayli Rageth<sup>34</sup>, Matthew T. Wheeler<sup>③35</sup>, Renske Oegema<sup>36</sup>, Halima Lourghi<sup>32</sup>, Maria G. Della Rocca<sup>30,31</sup>, Rachel Thompson<sup>③37</sup>, Francisco Castellanos<sup>4</sup>, James Priest<sup>38</sup>, Charlotte Cunningham-Rundles<sup>39</sup>, Ayushi Hegde<sup>4</sup>, Ruth C. Lovering<sup>④0</sup>, Catherine Hajek<sup>34</sup>, Annie Olry<sup>32</sup>, Luigi Notarangelo<sup>24</sup>, Morgan Similuk<sup>24</sup>, Xingmin A. Zhang<sup>③,4</sup>, David Gómez-Andrés<sup>41</sup>, Hanns Lochmüller<sup>②7,42,43,44</sup>, Hélène Dollfus<sup>45</sup>, Sergio Rosenzweig<sup>46</sup>, Shruti Marwaha<sup>35</sup>, Ana Rath<sup>③2</sup>, Kathleen Sullivan<sup>47</sup>, Cynthia Smith<sup>③33</sup>, Joshua D. Milner<sup>24</sup>, Dorothée Leroux<sup>45</sup>, Cornelius F. Boerkoel<sup>34</sup>, Amy Klion<sup>24</sup>, Melody C. Carter<sup>24</sup>, Tudor Groza<sup>3,22</sup>, Damian Smedley<sup>3,6</sup>, Melissa A. Haendel<sup>③,5,9</sup>, Chris Mungall<sup>3,7</sup> and Peter N. Robinson<sup>③,4,48,\*</sup>

# Various tools and data sets can be downloaded from the new HPO site

The screenshot shows the homepage of the Human Phenotype Ontology (HPO) website. At the top, there is a dark blue header bar with the HPO logo (a stylized antenna icon), the text "human phenotype ontology", and three dropdown menus: "Tools", "Downloads", and "Help". Below the header is a large search bar with a dropdown menu set to "All" and a placeholder text "Search for phenotypes, diseases, genes...". Below the search bar is an example search query: "e.g. Arachnodactyly | Marfan syndrome | FBN1". The main content area has a white background. On the left, there is a section titled "The Human Phenotype Ontology" which contains a detailed description of what the HPO is and its purpose. On the right, there is a section titled "News & Updates" which lists recent releases and updates. The first item in the news list is "June 2019 release" dated June 16, 2019. The second item is "April 2019 HPO Release" dated April 16, 2019. The third item is "hpo-web 1.5.0" dated April 16, 2019. At the bottom left is a button labeled "Learn More About HPO" and at the bottom right is a button labeled "View All News".

The Human Phenotype Ontology (HPO) provides a standardized vocabulary of phenotypic abnormalities encountered in human disease. Each term in the HPO describes a phenotypic abnormality, such as [Atrial septal defect](#). The HPO is currently being developed using the medical literature, Orphanet, DECIPHER, and OMIM. HPO currently contains over 13,000 terms and over 156,000 annotations to hereditary diseases. The HPO project and others have developed software for phenotype-driven differential diagnostics, genomic diagnostics, and translational research. The HPO is a flagship product of the [Monarch Initiative](#), an NIH-supported international consortium dedicated to semantic integration of biomedical and model organism data with the ultimate goal of improving biomedical research. The HPO, as a part of the Monarch Initiative, is a central component of one of the [13 driver projects](#) in the [Global Alliance for Genomics and Health \(GA4GH\)](#) strategic roadmap.

[Learn More About HPO](#)

[View All News](#)

**News & Updates**

[June 2019 release](#) June 16, 2019

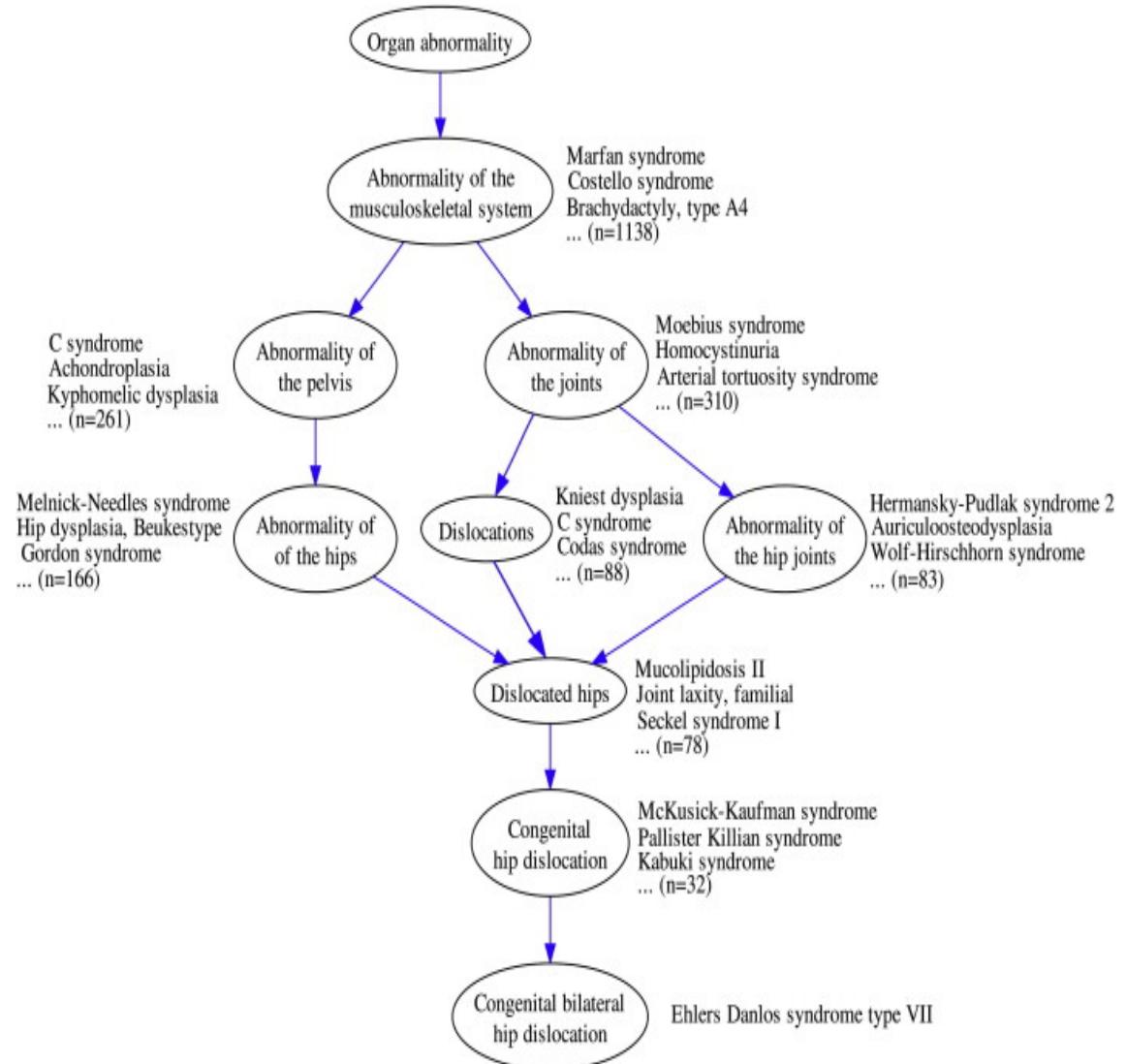
[April 2019 HPO Release](#) April 16, 2019

[hpo-web 1.5.0](#) April 16, 2019

<https://hpo.jax.org/> (new website)

# What does HPO look like?

- Each term in HPO can have multiple parents and children
- Tree structure
- Each HPO term can be mapped to multiple diseases with a frequency measure



# HPO browser

- Currently, the “Phenotypic abnormality” term has 26 subclasses
- The terms are still under active development
- Most ontologies are structured as directed acyclic graphs (DAG)
  - Similar to hierarchies but
  - Differ in that a more specialized term (child) can be related to more than one less specialized term (parent).

## Subclasses

[Abnormal test result](#)  
[Abnormality of the voice](#)  
[Abnormality of the endocrine system](#)  
[Abnormality of the skeletal system](#)  
[Abnormality of the breast](#)  
[Abnormality of limbs](#)  
[Abnormality of blood and blood-forming tissues](#)  
[Abnormality of the integument](#)  
[Neoplasm](#)  
[Constitutional symptom](#)  
[Abnormality of the respiratory system](#)  
[Abnormality of prenatal development or birth](#)  
[Abnormality of the musculature](#)  
[Abnormality of the digestive system](#)  
[Abnormality of metabolism/homeostasis](#)  
[Abnormality of the nervous system](#)  
[Abnormality of the cardiovascular system](#)  
[Abnormality of the genitourinary system](#)  
[Growth abnormality](#)  
[Abnormality of the immune system](#)  
[Abnormality of the eye](#)  
[Abnormality of connective tissue](#)  
[Abnormality of the ear](#)  
[Abnormal cellular phenotype](#)  
[Abnormality of the thoracic cavity](#)  
[Abnormality of head or neck](#)

# Examples of HPO terms

- Each has a unique and stable identifier (e.g. HP:0001251), a label and a list of synonyms.

Infopage for HPO class	Ataxia	S
<p>Primary ID HP:0001251 Alternative IDs HP:0007050, HP:0002513, HP:0001253, HP:0007157 PURL <a href="http://purl.obolibrary.org/obo/HP_0001251">http://purl.obolibrary.org/obo/HP_0001251</a></p>	<p>Synonyms Cerebellar ataxia</p>	<p>Textual definition Cerebellar ataxia refers to ataxia due to dysfunction of the cerebellum. This causes a variety of elementary neurological deficits including asynergy (lack of coordination between muscles, limbs and joints), dysmetria (lack of ability to judge distances that can lead to under- oder overshoot in grasping movements), and dysdiadochokinesia (inability to perform rapid movements requiring antagonizing muscle groups to be switched on and off repeatedly). Logical definition 'has part' some Intersection of - <a href="#">increased amount</a> - 'inheres in' some <a href="#">ataxia</a> - 'has modifier' some <a href="#">abnormal</a></p>

# Ataxia (HP:0001251)

- HPO terms have superclasses (possibly more than one) and subclasses

Superclasses

[Abnormality of the cerebellum](#)  
[Abnormality of coordination](#)

Subclasses

[Progressive cerebellar ataxia](#)  
[Nonprogressive cerebellar ataxia](#)  
[Dyssynergia](#)  
[Cerebellar ataxia associated with quadrupedal gait](#)  
[Limb ataxia](#)  
[Dysmetria](#)  
[Spastic ataxia](#)  
[Gait ataxia](#)  
[Truncal ataxia](#)  
[Dysdiadochokinesis](#)  
[Episodic ataxia](#)

777 associated diseases

Disease id	Disease name
ORPHA:3350	Tremor-nystagmus-duodenal ulcer syndrome
OMIM:312170	PYRUVATE DEHYDROGENASE E1-ALPHA DEFICIENCY
OMIM:305000	DYSKERATOSIS CONGENITA, X-LINKED
OMIM:604273	MITOCHONDRIAL COMPLEX V (ATP SYNTHASE) DEFICIENCY, NUCLEAR TYPE 1
OMIM:213600	BASAL GANGLIA CALCIFICATION, IDIOPATHIC, 1
OMIM:606554	EPISODIC ATAXIA, TYPE 3

# Phenomizer and HPO

The Phenomizer

Patient's Features. Diagnosis.

HPO.	Feature. ▲	Modifier.	Num diseas...
category.: Abnormality of metabolism/homeostasis (1 Item)			
HP:0003236	Elevated serum creatine phosphokin...	observed.	190 of 7994
category.: Abnormality of the musculature (1 Item)			
HP:0030224	Abnormal muscle fiber desmin	observed.	0 of 7994
category.: Abnormality of metabolism/homeostasis (1 Item)			
HP:0012113	Abnormality of creatine metabolism	observed.	2 of 7994

Clear. Mode of inheritance. ▾ Get diagnosis.

**Input: HPO terms**

The Phenomizer

Patient's Features. Diagnosis.

Algorithm: resnik (Unsymmetric). 3 Features.

	p-value. ▲	Disease Id.	Disease name.	Genes.
<input checked="" type="checkbox"/>	0.0160	OMIM:123...	123270 CREATINE KINASE, BRAIN TYPE, ECT...	
<input checked="" type="checkbox"/>	0.0484	OMIM:616...	#616052 MUSCULAR DYSTROPHY-DYSTROG...	ISPD
<input checked="" type="checkbox"/>	0.0484	OMIM:612...	#612718 CEREBRAL CREATINE DEFICIENCY ...	GATM
<input checked="" type="checkbox"/>	0.0484	OMIM:309...	MUSCULAR DYSTROPHY, CARDIAC TYPE	
<input checked="" type="checkbox"/>	0.0484	OMIM:614...	#614408 MYOPATHY, CENTRONUCLEAR, 3; C...	BIN1, MTMR1...
<input checked="" type="checkbox"/>	0.0484	OMIM:605...	NONAKA MYOPATHY	GNE
<input checked="" type="checkbox"/>	0.0484	OMIM:253...	#253601 MUSCULAR DYSTROPHY, LIMB-GIR...	DYSF
<input checked="" type="checkbox"/>	0.0484	OMIM:612...	GLYCOGEN STORAGE DISEASE XIII; GSD13	ENO3
<input checked="" type="checkbox"/>	0.0484	OMIM:609...	MYOPATHY, AUTOPHAGIC VACUOLAR, INF...	
<input checked="" type="checkbox"/>	0.0484	OMIM:600...	INCLUSION BODY MYOPATHY 2, AUTOSOMA...	GNE
<input checked="" type="checkbox"/>	0.0484	OMIM:604...	#604454 WELANDER DISTAL MYOPATHY; WD...	TIA1
<input checked="" type="checkbox"/>	0.0484	OMIM:615...	#615422 INCLUSION BODY MYOPATHY WITH...	VCP, HNRNPA...
<input checked="" type="checkbox"/>	0.0484	OMIM:613...	#613204 MUSCULAR DYSTROPHY, CONGEN...	ITGA7
<input checked="" type="checkbox"/>	0.0484	OMIM:300...	#300717 MYOPATHY, REDUCING BODY, X-LIN...	FHL1
<input checked="" type="checkbox"/>	0.0484	OMIM:609...	FILAMINOPATHY, AUTOSOMAL DOMINANT	FLNC

Page 1 of 268 Improve Differential Diagnosis. Download Results.

**Output: Disease diagnosis and p-values**

# Quick introduction to Gene Ontology

- The goal of the Gene Ontology Consortium is to produce a dynamic, controlled vocabulary that can be applied to all eukaryotes even as knowledge of gene and protein roles in cells is accumulating and changing.

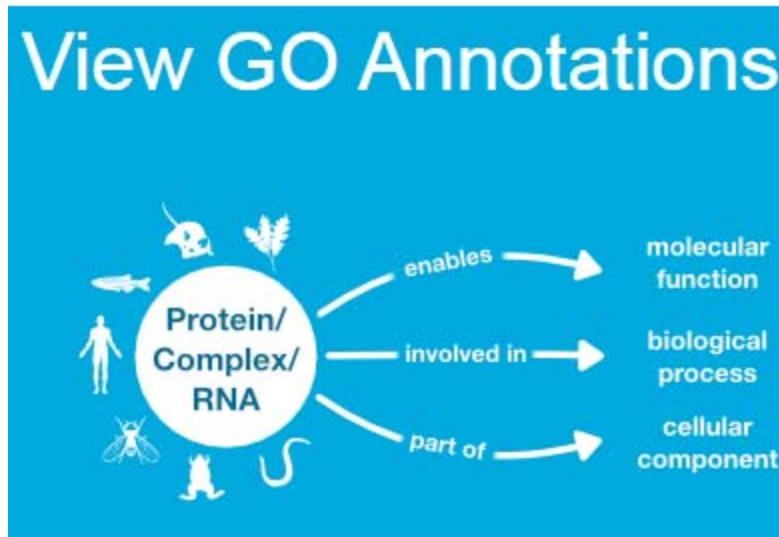


Commentary

## Gene Ontology: tool for the unification of biology

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein<sup>✉</sup>, Heather Butler, J. Michael Cherry<sup>✉</sup>, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin & Gavin Sherlock

# Gene Ontology represents 3 different ontologies

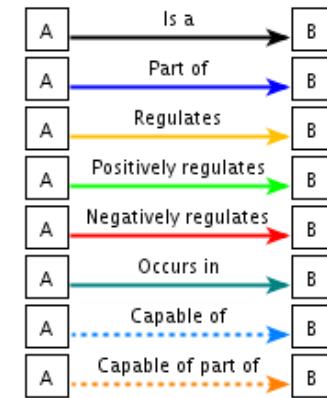
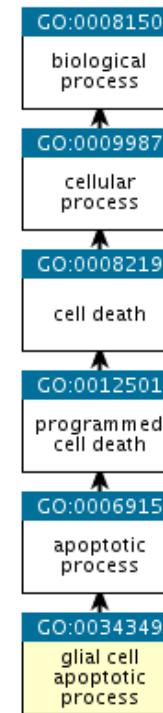


- Cellular components:
  - The parts of a cell or its extracellular environment;
- Molecular functions:
  - The elemental activities of a gene product at the molecular level, such as binding or catalysis;
- Biological processes:
  - Operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms.

<https://www.ebi.ac.uk/QuickGO/>

# Example: glial cell apoptotic process

- Biological Process
- GO:0034349
- Definition: Any apoptotic process in a glial cell, a non-neuronal cell of the nervous system.



QuickGO - <https://www.ebi.ac.uk/QuickGO>

## Child Terms

This table lists all terms that are direct descendants (child terms) of GO:0034349

Child Term	Relationship to GO:0034349
GO:0097252 (P) 🏫 ✅ oligodendrocyte apoptotic process	is_a
GO:0034352 (P) 🏫 ✅ positive regulation of glial cell apoptotic process	positively_regulates
GO:0034350 (P) 🏫 ✅ regulation of glial cell apoptotic process	regulates
GO:0034351 (P) 🏫 ✅ negative regulation of glial cell apoptotic process	negatively_regulates

# An example on the TCF4 gene

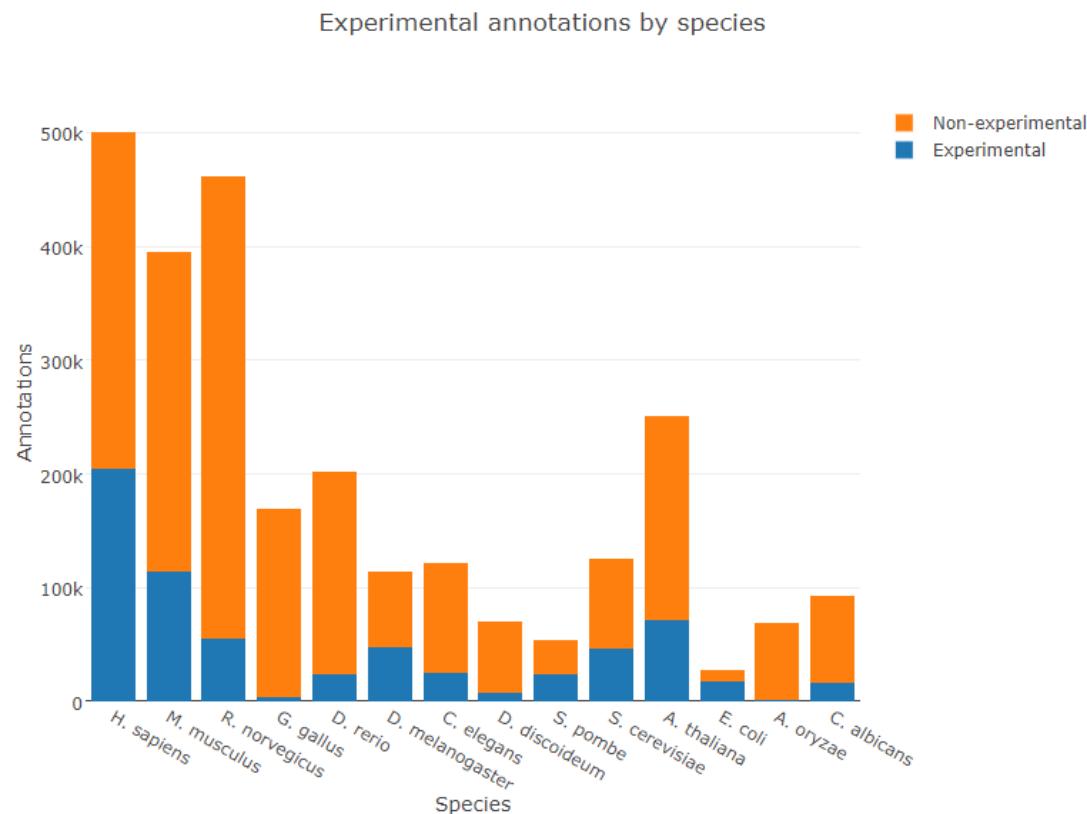
Total annotations: 198; showing: 1-10 Results count <input type="button" value="10"/>								<a href="#">«First</a> <a href="#">&lt;Prev</a> <a href="#">Next&gt;</a> <a href="#">Last»</a> <a href="#">Download (up to 100000)</a>				
	Gene/product name	Annotation qualifier	GO class (direct)	Annotation extension	Contributor	Organism	Evidence	Evidence with	PANTHER family	Isoform	Reference	Date
<input type="checkbox"/>	TCF4	Transcription factor 4	nuclear chromatin	coincident with ENSEMBL:ENSG00000168646 coincident with SO:0000167	ParkinsonsUK-UCL	Homo sapiens	IDA		basic helix-loop-helix transcription factor pthr11793		PMID:21880741	20160517
<input type="checkbox"/>	TCF4	Transcription factor 4	RNA polymerase II proximal promoter sequence-specific DNA binding		UniProt	Homo sapiens	IDA		basic helix-loop-helix transcription factor pthr11793		PMID:12651860	20170310
<input type="checkbox"/>	TCF4	Transcription factor 4	RNA polymerase II proximal promoter sequence-specific DNA binding		BHF-UCL	Homo sapiens	ISS	UniProtKB:Q60722	basic helix-loop-helix transcription factor pthr11793		PMID:8978694	20101210

Inferred from Direct Assay (IDA)  
Inferred from Sequence or  
structural Similarity (ISS)

<http://www.geneontology.org/page/guide-go-evidence-codes>

# GO associates annotations to genes

- Human, mouse and rat and Arabidopsis represent the species with the most annotations



# OMIM database

- A continuously updated catalog of human genes and genetic disorders and traits

**OMIM®**

**Online Mendelian Inheritance in Man®**

**An Online Catalog of Human Genes and Genetic Disorders**

Updated April 4, 2018

Search OMIM for clinical features, phenotypes, genes, and more...



[Advanced Search : OMIM, Clinical Synopses, Gene Map](#)

[Need help? : Example Searches, OMIM Search Help, OMIM Tutorial](#)

[Mirror site : mirror.omim.org](#)

OMIM is supported by a grant from NHGRI, licensing fees, and generous contributions from people like you.

**Make a donation!**



# OMIM's statistics of known gene/phenotype relationships

## OMIM Gene Map Statistics

OMIM Morbid Map Scorecard (Updated July 2nd, 2019) :

Total number of phenotypes* for which the molecular basis is known	6,452
Total number of genes with phenotype-causing mutation	4,114

\* Phenotypes include (1) single-gene mendelian disorders and traits; (2) susceptibilities to cancer and complex disease (e.g., BRCA1 and familial breast-ovarian cancer susceptibility, [113705.0001](#), and CFH and macular degeneration, [134370.0008](#)); (3) variations that lead to abnormal but benign laboratory test values ("nondiseases") and blood groups (e.g., lactate dehydrogenase B deficiency, [150100.0001](#) and ABO blood group system, [110300.0001](#)); and (4) select somatic cell genetic disease (e.g., GNAS and McCune-Albright syndrome, [139320.0008](#) and IDH1 and glioblastoma multiforme, [147700.0001](#).)

Distribution of Phenotypes across Genes (Updated July 2nd, 2019) :

Number of genes with 1 phenotype	2,837
Number of genes with 2 phenotypes	773
Number of genes with 3 phenotypes	269
Number of genes with 4+ phenotypes	235

# Example of an OMIM entry

\*602272  
Table of Contents

Title

Gene-Phenotype Relationships

Text

Description

Cloning and Expression

Gene Structure

Mapping

Gene Function

Molecular Genetics

Animal Model

Allelic Variants

Table View

References

Contributors

Creation Date

Edit History

\* 602272

## TRANSCRIPTION FACTOR 4; TCF4

*Alternative titles; symbols*

IMMUNOGLOBULIN TRANSCRIPTION FACTOR 2; ITF2  
SEF2-1B  
SEF2  
E2-2

*HGNC Approved Gene Symbol:* **TCF4**

*Cytogenetic location:* [18q21.2](#)   *Genomic coordinates (GRCh38):* [18:55,222,330-55,635,992](#) (from NCBI)

### Gene-Phenotype Relationships

Location	Phenotype	Phenotype MIM number	Inheritance	Phenotype mapping key
18q21.2	Corneal dystrophy, Fuchs endothelial, 3	613267	AD	3
	Pitt-Hopkins syndrome	610954	AD	3

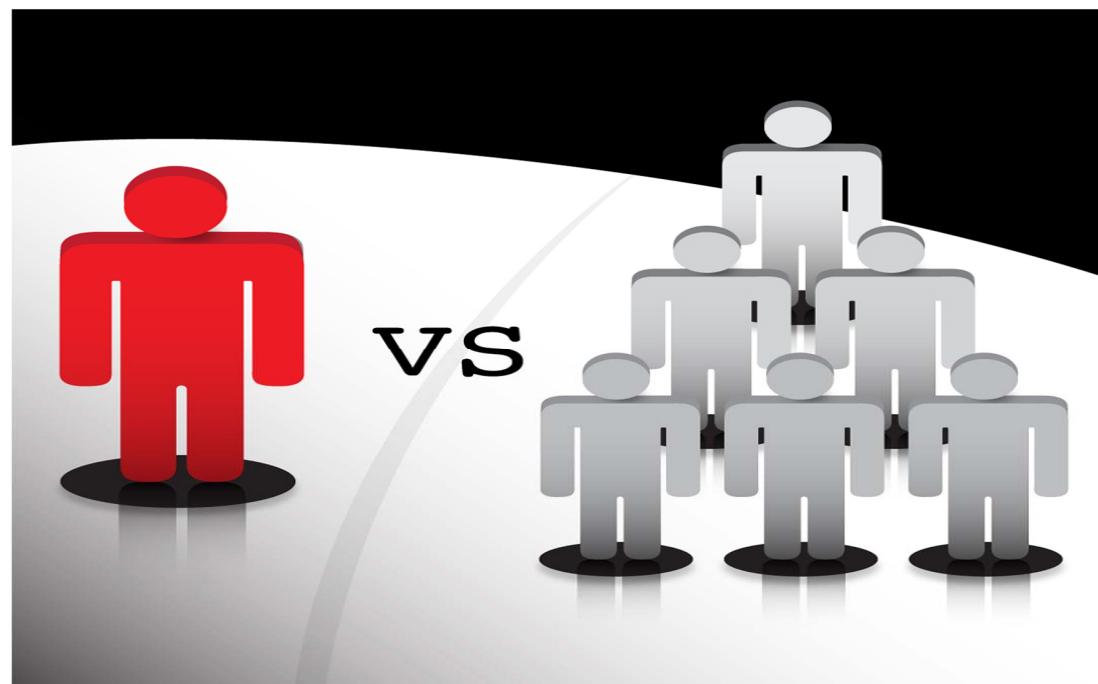
### TEXT

#### ▼ Description

TCF4 is a broadly expressed basic helix-loop-helix (bHLH) protein that functions as a homodimer or as a heterodimer with other bHLH proteins. These dimers bind DNA at E-box sequences. Alternative splicing produces numerous N-terminally distinct TCF4 isoforms that differ in their subcellular localization and transactivational capacity (summary by Sepp et al., 2012). 

# Challenges in genomic medicine

- Identify disease causal genes from **population cohort**
- Identify disease causal genes for **individual patients** with suspected Mendelian diseases



# Clinical phenotype facilitates differential diagnosis of rare diseases

- Several phenotype analysis tools exist, such as
  - Linking Open data for Rare Diseases (LORD)
  - Online Mendelian Inheritance in Man (OMIM)
  - Orphanet (<http://www.orpha.net/consor/cgi-bin/index.php>)
  - Phenomizer (<http://compbio.charite.de/phenomizer/>)
  - Phevor (<http://weatherby.genetics.utah.edu/phevor2/index.html>)
  - Rare Disease Discovery (<http://disease-discovery.udl.cat/>)
  - .....
- Most systems do not provide explicit rankings or probability scores
- Most systems do not incorporate disease-gene relationships to facilitate gene-finding

# Our goal: from clinical phenotypes to genes

## Phenotypic features

short stature  
obesity  
short fingers  
overlapping toe  
short toes  
macrocephaly  
strabismus  
upturned earlobe

## Candidate disease gene lists

GNAS
SMAD1
TGFBR1
SMAD9
CREBBP
BMP2
NRAS
CHN1
IKBKB

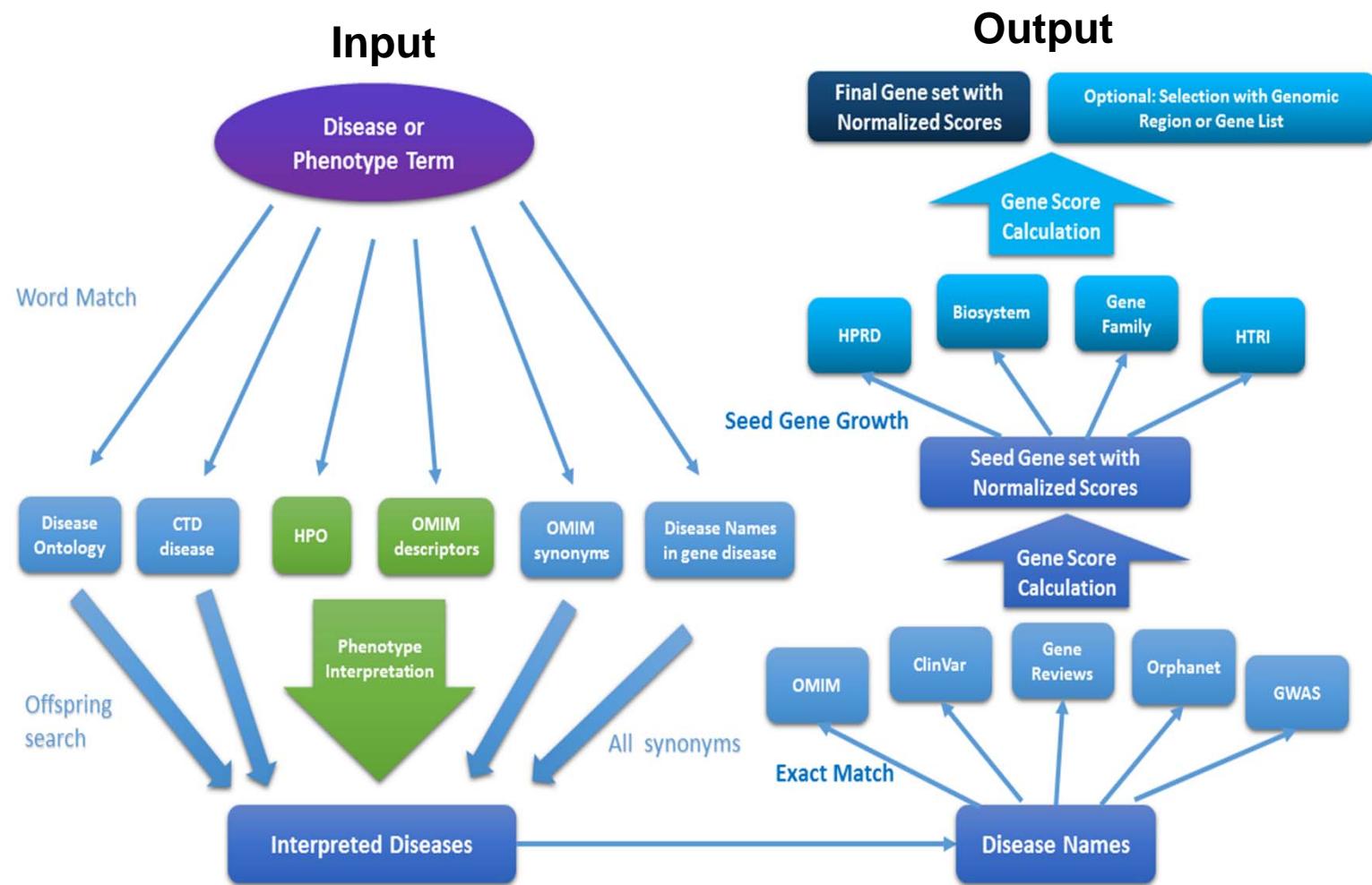
**Expedite the discovery of disease causal variants from exome/genome sequencing data**



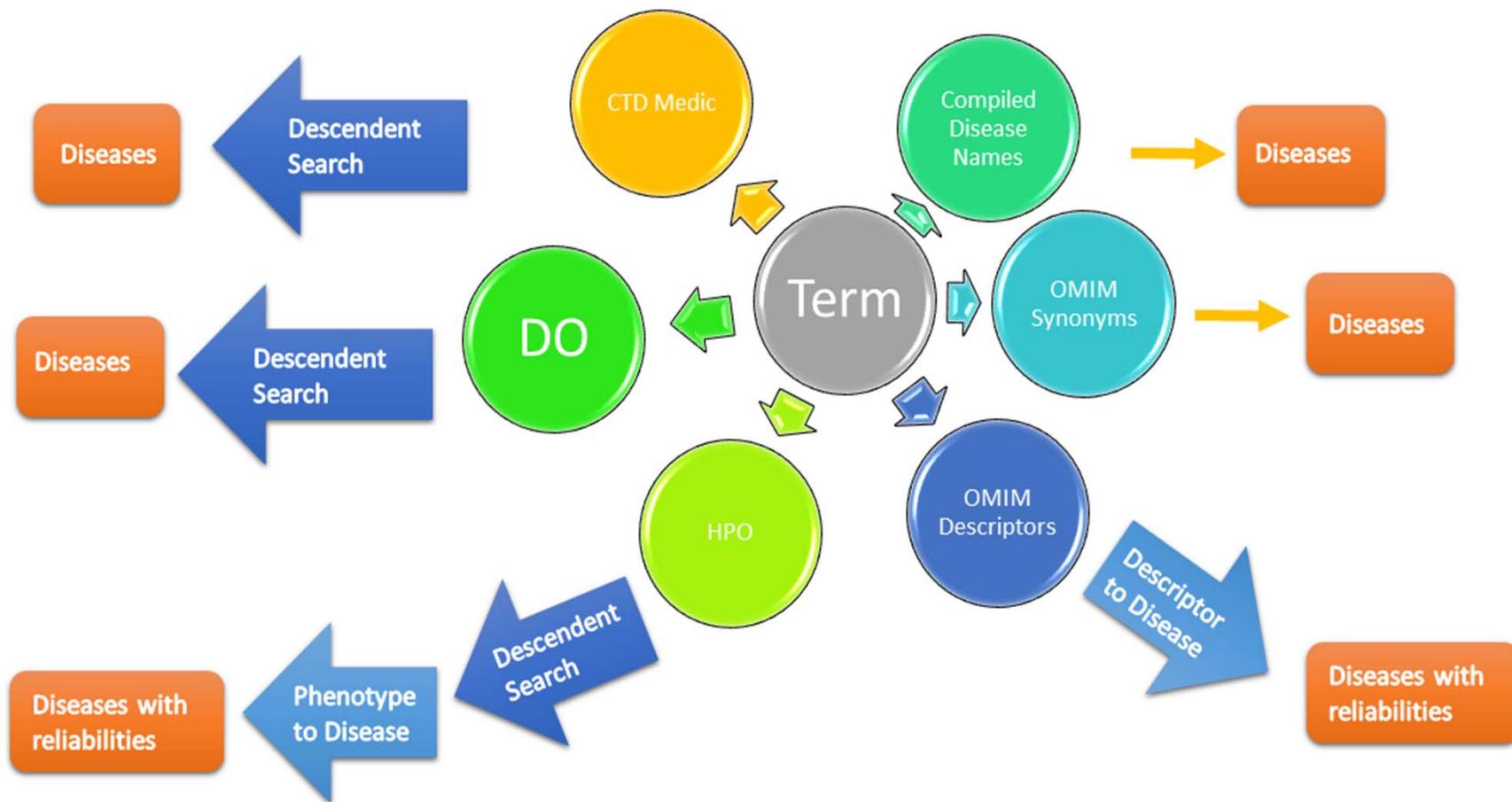
# What is Phenolyzer

- A tool for phenotype analysis:
  - 1) To map user-supplied phenotypes to diseases and candidate genes
  - 2) A resource that integrates existing biological knowledge to identify known disease genes
  - 3) A prediction algorithm to predict novel disease genes
  - 4) A model to integrate multiple features to score and prioritize genes
  - 5) A network visualization tool to explore the disease-term, disease-gene and gene-gene relations

# Detailed work flow of Phenolyzer



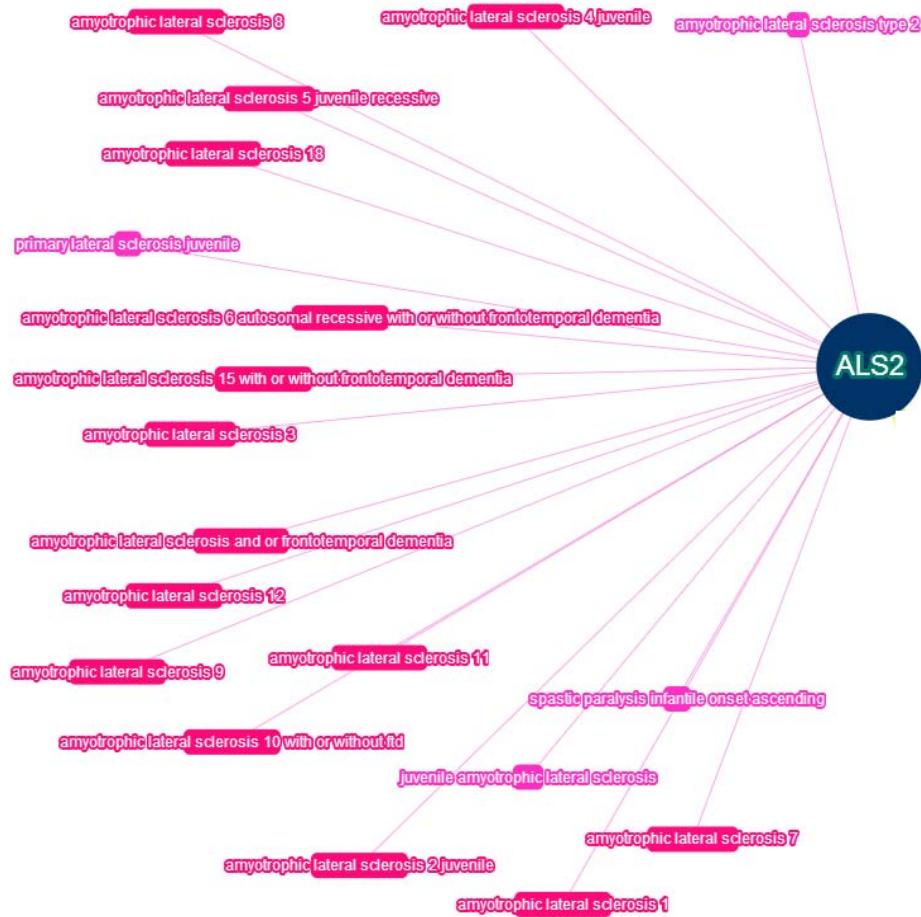
# Step 1: Term interpretation



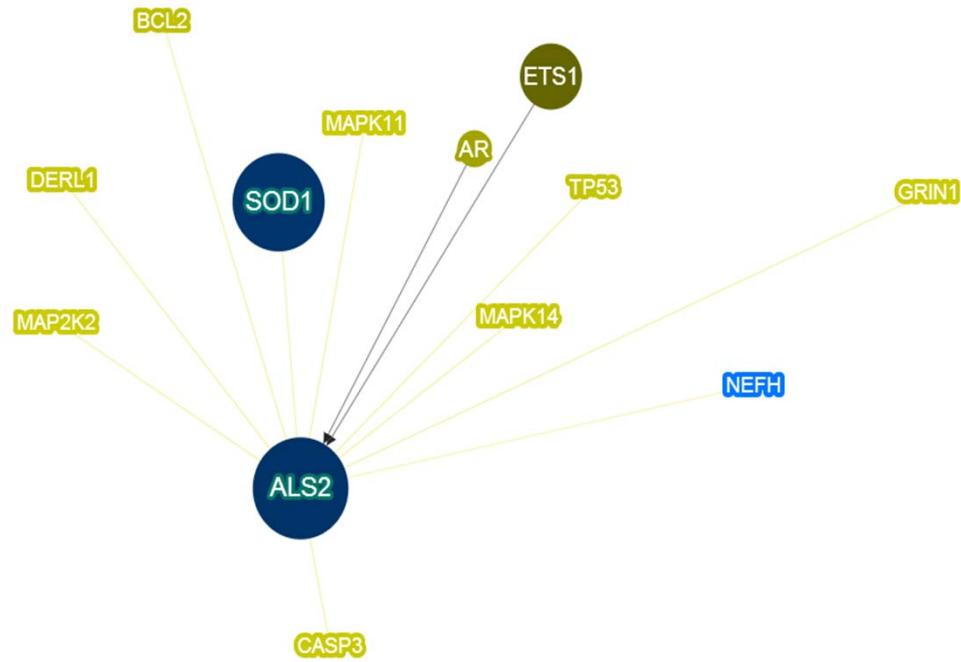
“Term” can be any phenotype term, such as “developmental delay”, “hearing loss”, etc

# **‘Amyotrophic lateral sclerosis’ interpreted**

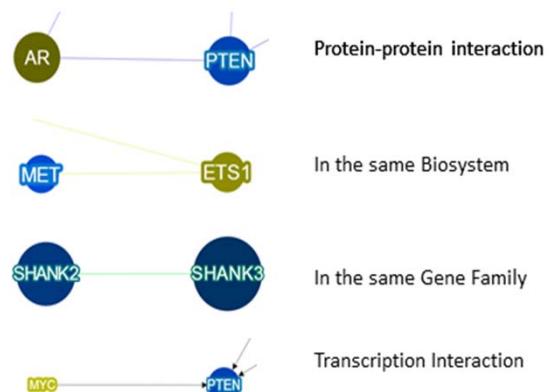
## Step 2: Seed gene set generation



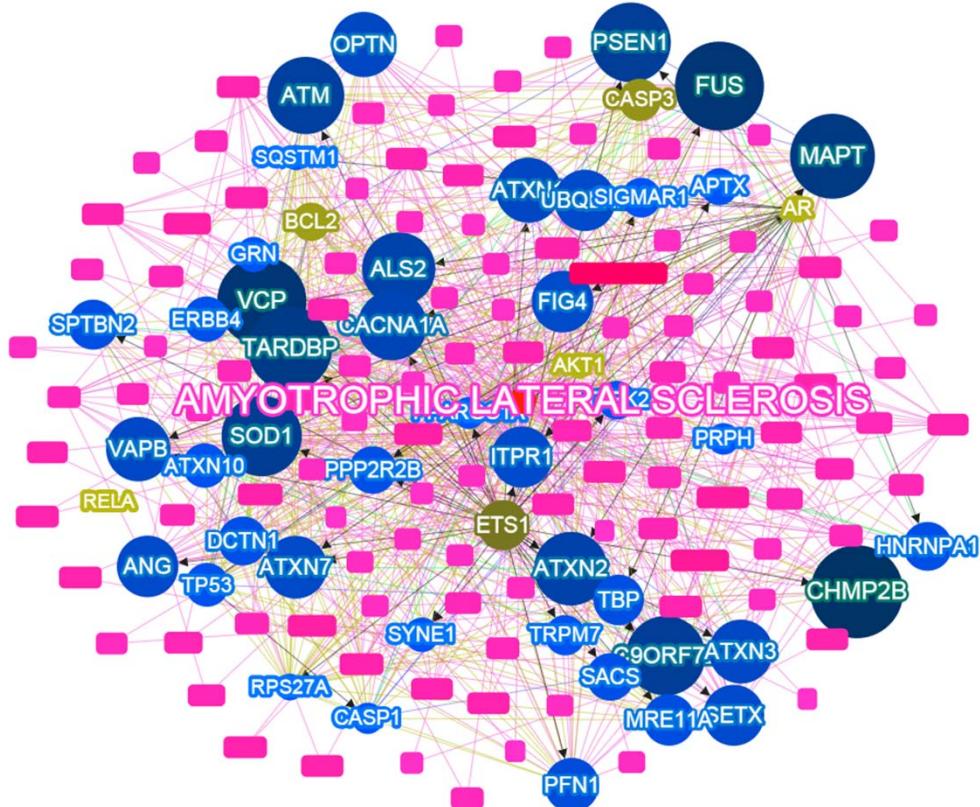
## Step 3: Seed gene set growth



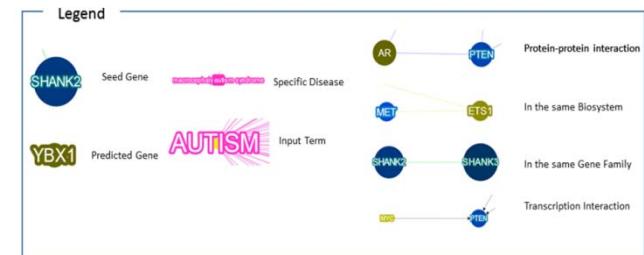
**ALS2 is grown based on four type of gene-gene relations,:  
1. protein interaction  
2. pathway  
3. gene family  
4. transcription interaction**



# Step 4: Data integration and scoring



**Gene-gene and gene-disease network for 'ALS'**

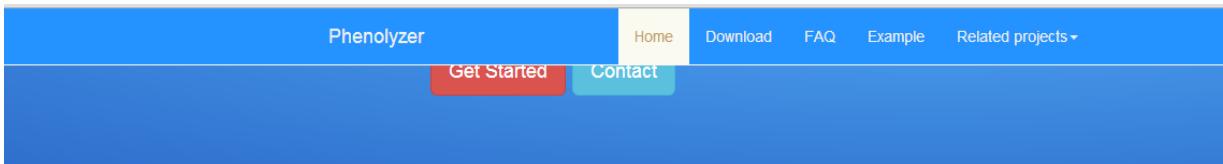


Phenolyzer generates a gene-phenotype-disease network for ALS

# Web implementation

<http://phenolyzer.wglab.org>

1) Enter the website



## Basic Information

Email

please enter your focused disease/phenotype terms

Please use semicolon or enter as separators. Like "disease,brain".  
Try to use multiple terms instead of a super long term  
OMIM IDs are also accepted, like 114480 for 'Breast cancer'

Submit

Reset

2) Enter the disease/phenotype terms, like 'Autism'

3) Submit, done!

## Options

Gene Selection

No



Region Selection

No



Advanced Options

Phenotype Interpretation



Weight Adjust

No



Word Cloud

No



Optional 4) Turn on Word Cloud

**Click to see all the interpreted diseases**

**Click to see the detailed report of all genes**

**Click to see the whole gene list**

**Click to see the detailed report of seed genes**

**Click to see the seed gene list**

Submission ID: 2007

**Click to see the Wordcloud of the term before.**

Summary    Network    Barplot    Details

### Submission information

Phenotypes are interpreted.

At most 2000 genes will be found in details, for the complete list, please download the report here.

1 disease terms have been entered, among which, 1 terms have corresponding records in our database.

They are [huntington](#) [WordCloud](#)

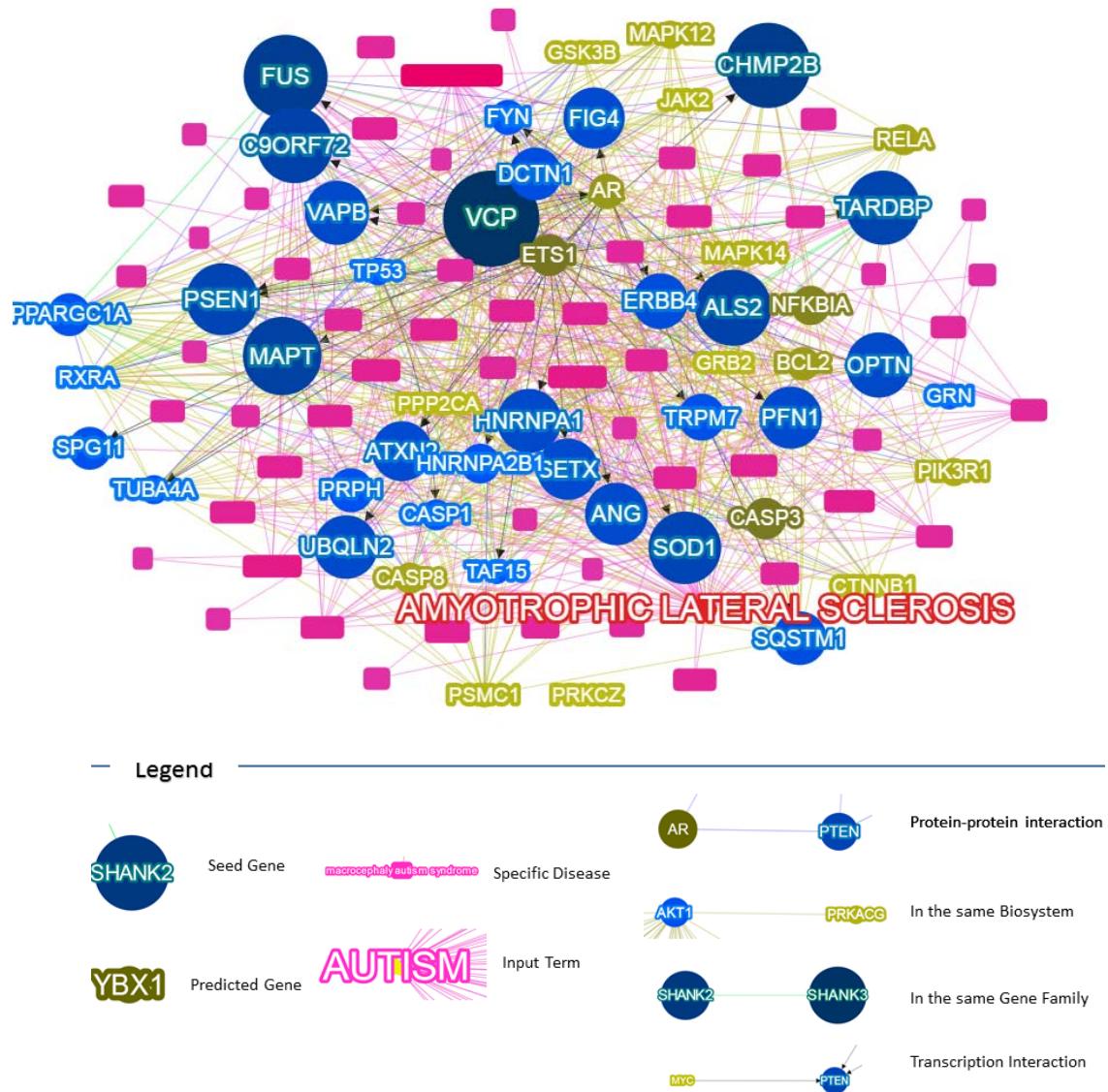
The whole report could be found [Here](#).

The normalized gene scores could be found [Here](#).

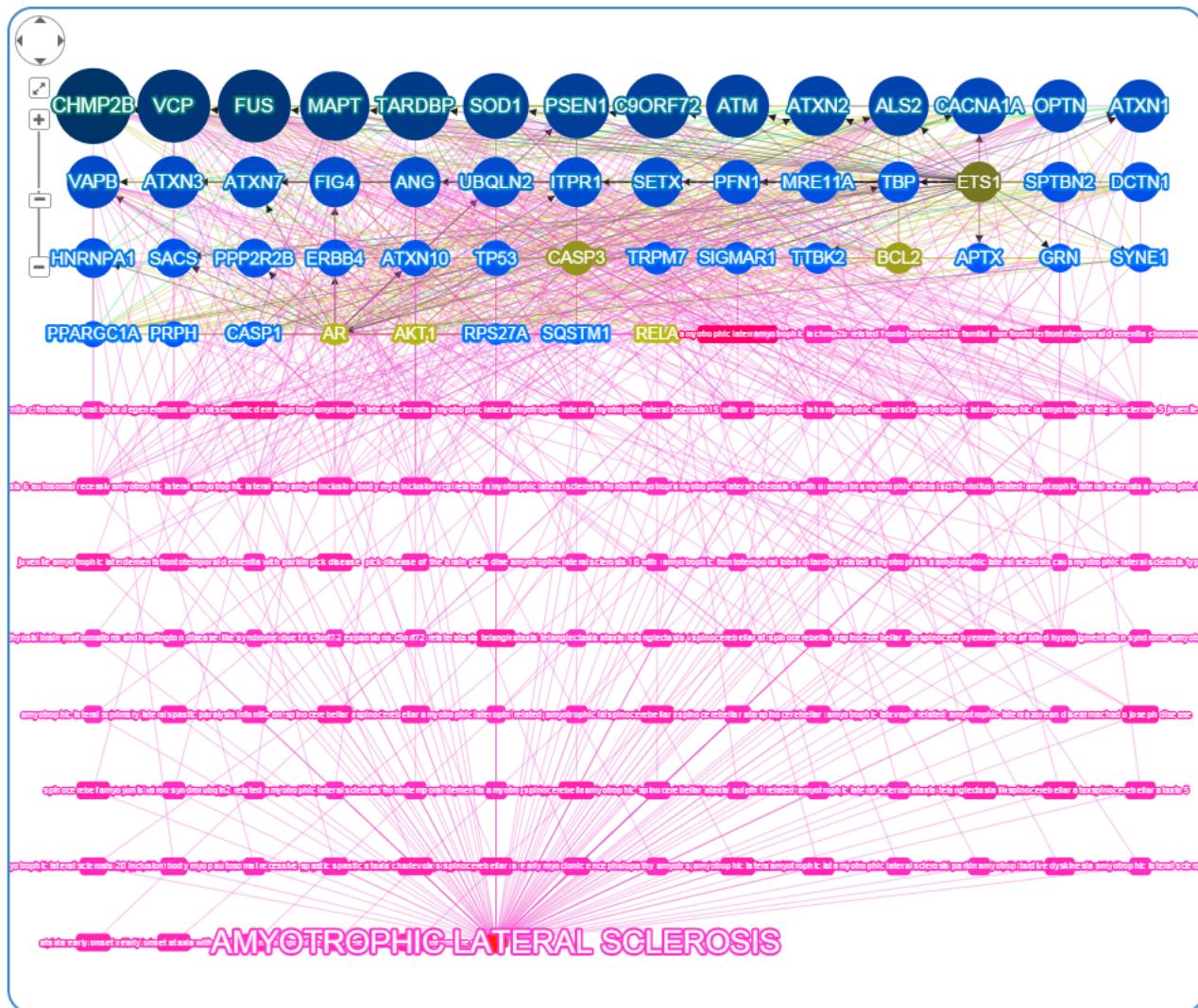
The report without prediction could be found [Here](#).

The normalized gene scores without prediction could be found [Here](#).

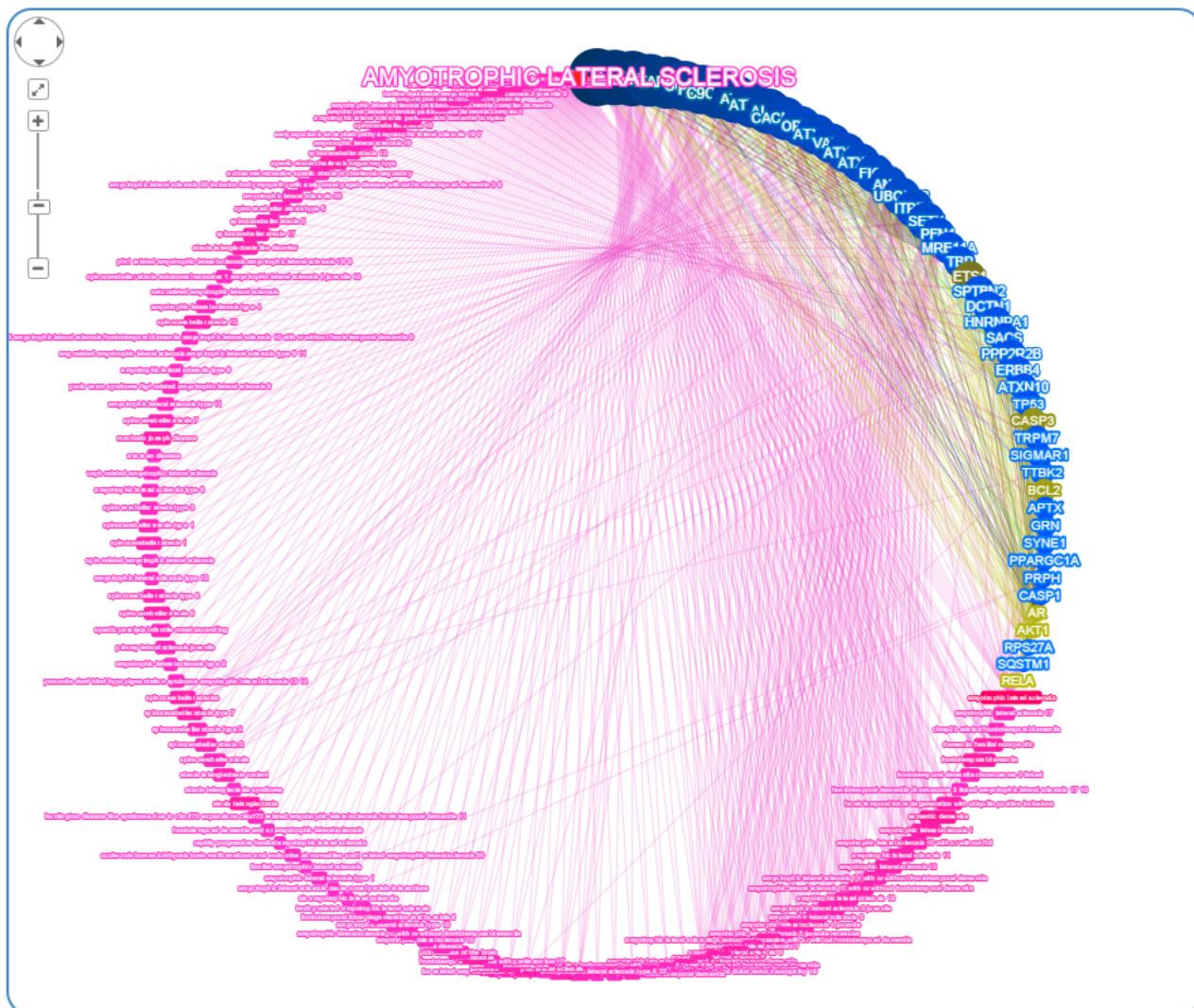
# Example output for ALS



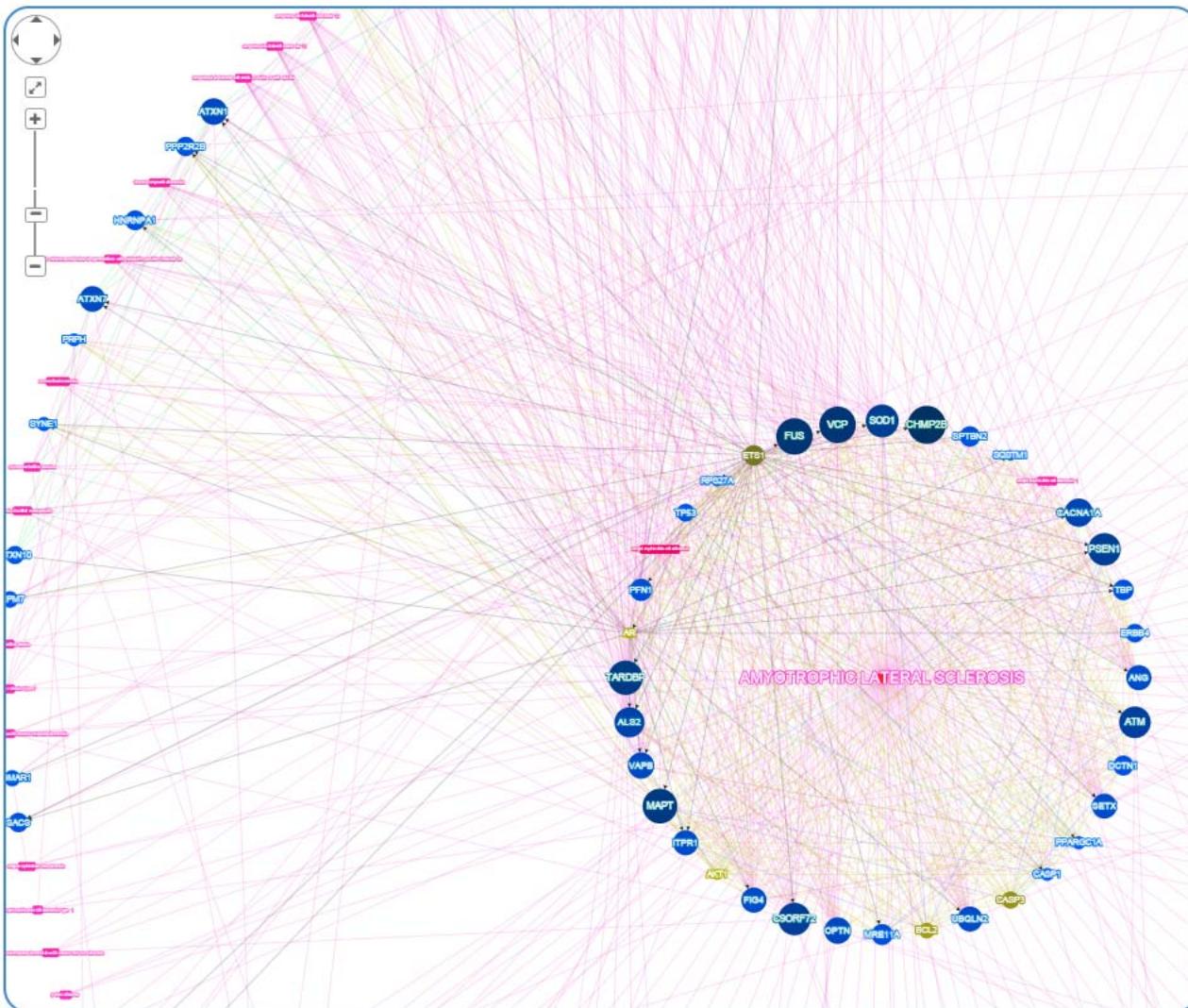
# Different layouts for ‘top gene list’



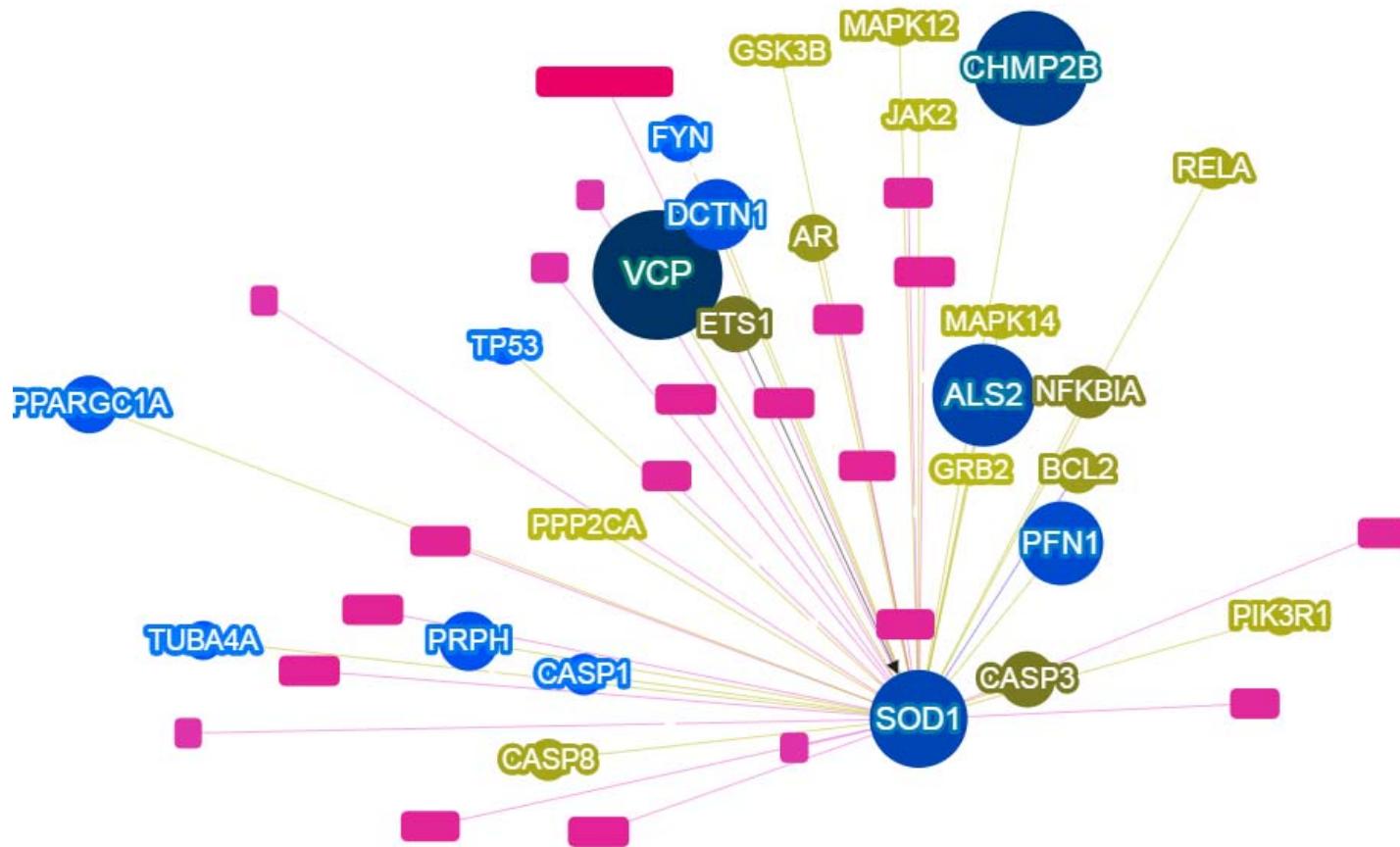
# Different layouts for ‘top gene list’



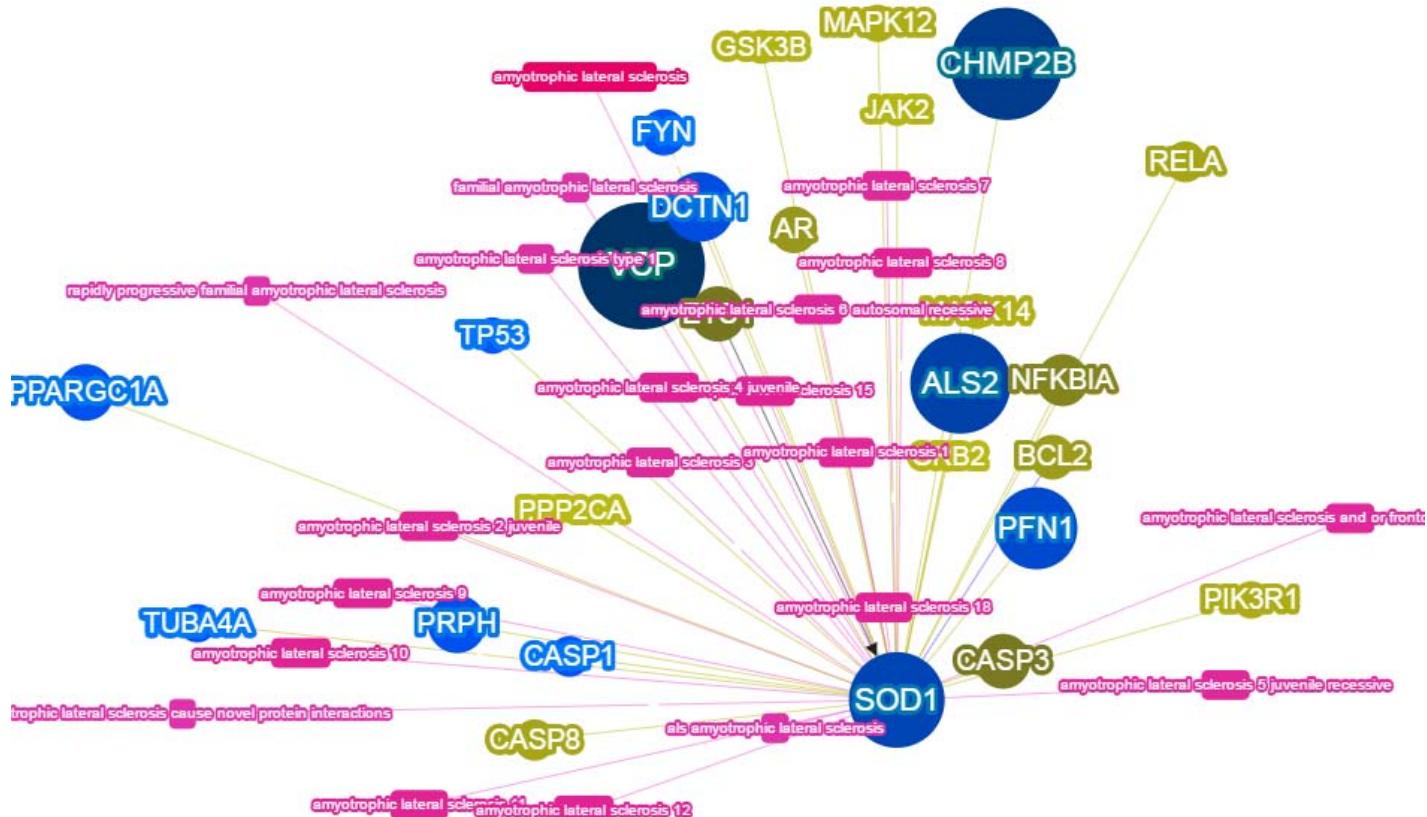
# Different layouts for ‘top gene list’



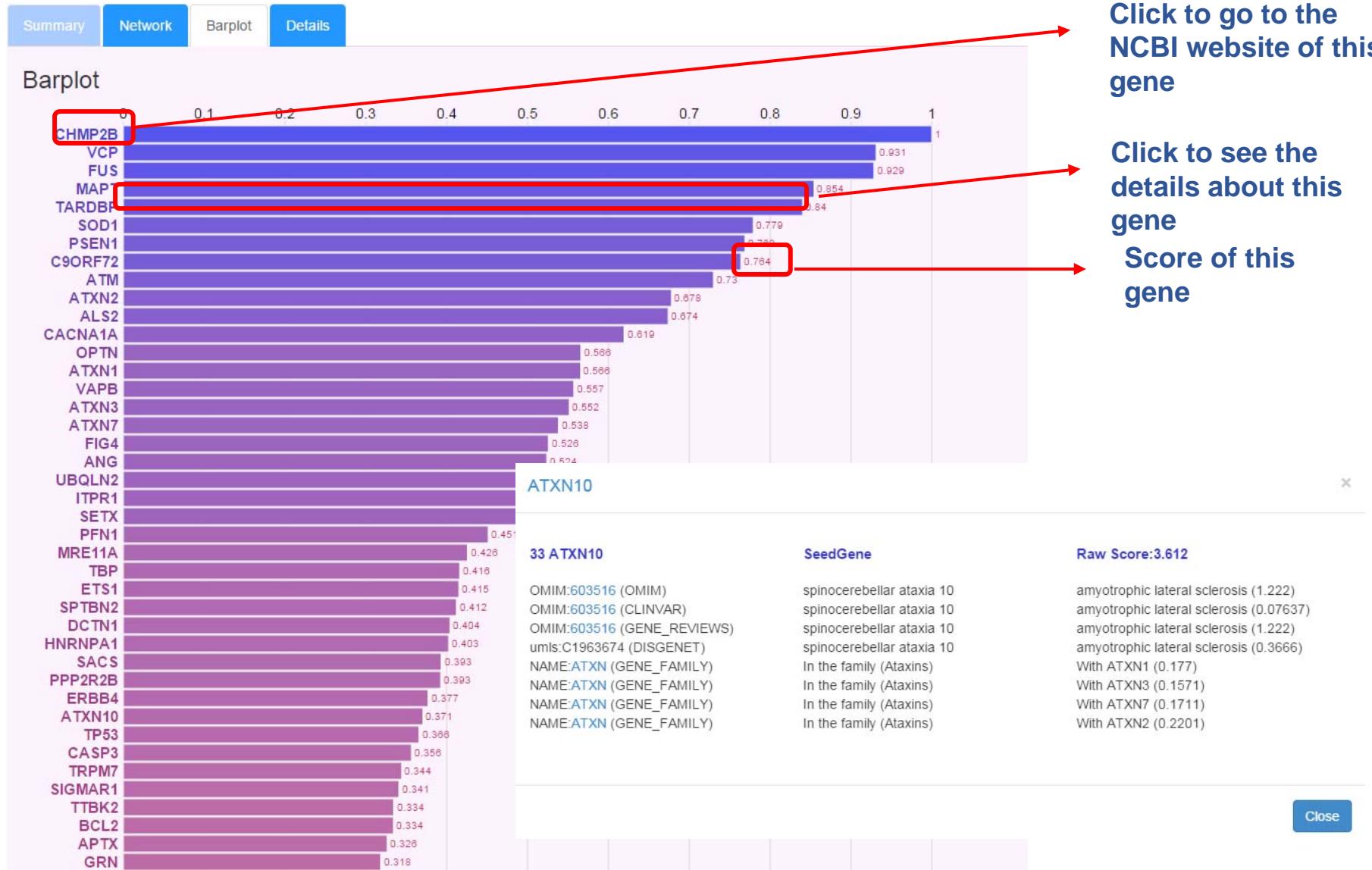
# Only connections for a specific gene



# Turn on all labels



Disease ON  Gene ON  Gene Name ON  Disease Name ON



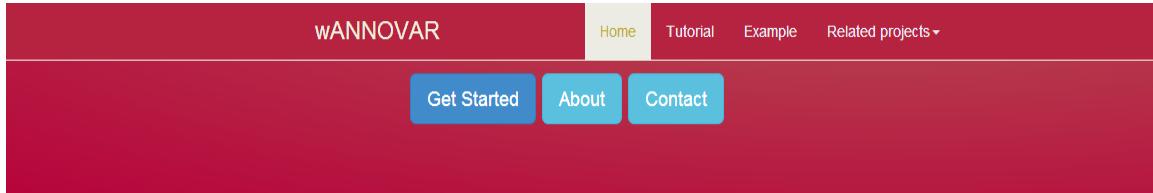
# Detailed examination of the evidence

Summary	Network	Barplot	Details
<a href="#">Start</a>	<a href="#">Previous</a>	1	<a href="#">Go</a> <a href="#">Next</a> <a href="#">End</a>
▶ 1 CHMP2B	SeedGene	Raw Score:9.746	
▶ 2 VCP	SeedGene	Raw Score:9.075	
▶ 3 FUS	SeedGene	Raw Score:9.049	
▶ 4 MAPT	SeedGene	Raw Score:8.326	
▼ 5 TARDBP	SeedGene	Raw Score:8.188	
<b>TARDBP</b>			
ORPHANET: <a href="#">803</a> (ORPHANET)	amyotrophic lateral sclerosis	amyotrophic lateral sclerosis (0.8728)	
umls:C0002736 (DISGENET)	amyotrophic lateral sclerosis	amyotrophic lateral sclerosis (0.5071)	
unknown (GENE_CARDS)	amyotrophic lateral sclerosis	amyotrophic lateral sclerosis (0.3055)	
OMIM: <a href="#">105400</a> (GENE_REVIEWS)	amyotrophic lateral sclerosis 1	amyotrophic lateral sclerosis (0.08146)	
umls:C3502417 (DISGENET)	amyotrophic lateral sclerosis 10	amyotrophic lateral sclerosis (0.3666)	
umls:C2677565 (DISGENET)	amyotrophic lateral sclerosis 10 with or without frontotemporal dementia	amyotrophic lateral sclerosis (0.3666)	
OMIM: <a href="#">612069</a> (OMIM)	amyotrophic lateral sclerosis 10 with or without ftd	amyotrophic lateral sclerosis (1.222)	
OMIM: <a href="#">612069</a> (GENE_REVIEWS)	amyotrophic lateral sclerosis 10 with or without ftd	amyotrophic lateral sclerosis (0.08146)	
unknown (GENE_CARDS)	amyotrophic lateral sclerosis 10 with or without ftd	amyotrophic lateral sclerosis (0.3055)	
OMIM: <a href="#">612577</a> (GENE_REVIEWS)	amyotrophic lateral sclerosis 11	amyotrophic lateral sclerosis (0.08146)	
OMIM: <a href="#">613435</a> (GENE_REVIEWS)	amyotrophic lateral sclerosis 12	amyotrophic lateral sclerosis (0.08146)	
OMIM: <a href="#">300857</a> (GENE_REVIEWS)	amyotrophic lateral sclerosis 15 with or without	amyotrophic lateral sclerosis (0.08146)	

# Integration with wANNOVAR

<http://wannovar.wglab.org>

1) Enter wANNOVAR website address



4) Submit, done!

2) Enter variant file, email and all the information wANNOVAR requires.

3) Enter disease/phenotype terms here

The submission form has several fields:

- Email
- Sample Identifier
- Input File (with a '+ Input File' button)
- or Paste Variant Calls (with a text area labeled 'paste your variant call here')

At the bottom are three buttons: Submit (highlighted with a red box), Reset, and Monitor Progress.

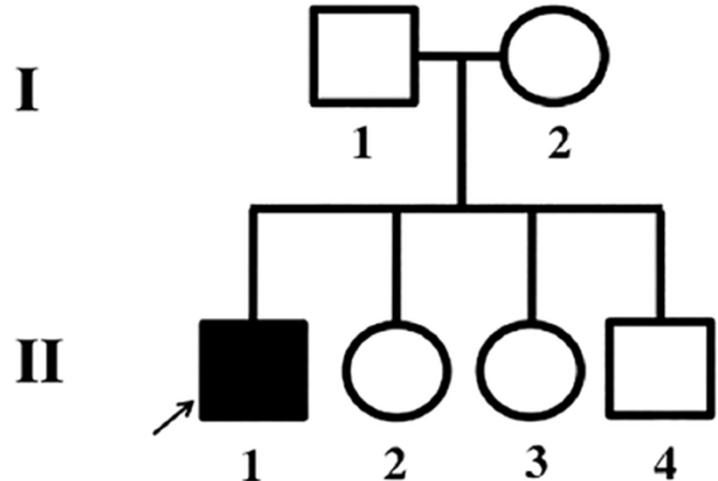
## Disease/Phenotype (Optional)

Enter Disease or Phenotype Terms

A large text input field with placeholder text: 'please enter your focused disease/phenotype terms'. Below the field is a note: 'Please use semicolon or enter as separators. Like "alzheimer;brain". Try to use multiple terms instead of a super long term OMIM IDs are also accepted, like 114480 for "Breast cancer"'.

# Case study: how to use Phenolyzer in practice?

- The proband had his first epileptic episode at 3 years of age. After this episode, he lost all speech, began exhibiting autistic behavior, and also started to have frequent generalized tonic–clonic seizures.
- Other developmental skills, including throwing a ball, responding to his name, feeding himself with utensils, and self-care skills were lost by 4 years of age.
- He attended a week in an autism evaluation classroom where he was diagnosed with ASD and considered severe and qualified for every service offered.



# Phenotype features of an undiagnosed case



**bilateral clinodactyly of the fifth finger, brachydactyly, and bilateral single transverse palmar creases**

**rounded face, bushy eyebrows, broad nasal tip, short philtrum, thick lips, and prognathism**

# Phenotype translation into HPO terms

Features (Human Phenotype Ontology)	Proband
Facial dysmorphism	
Large fontanelle (HP:0000239)	+
Rounded face (HP:0000311)	+
Bushy eyebrows (HP:0000574)	+
Broad nasal Tip (HP:0000455)	+
Short philtrum (HP:0000322)	+
Full/thick lips (HP:0012471)	+
Cupid bow upper lip (HP:0002263)	+
Macrodontia of upper central incisors (HP:0000675)	+
Prognathism (HP:0000303)	+
Developmental/intellectual disability	
Intellectual disability (HP:0001249)	+
Absent speech (HP:0001344)	+
Skeletal	
Clinodactyly of the fifth finger (HP:0004209)	+
Brachydactyly (HP:0009803)	+
Bilateral single transverse palmar creases (HP:0007598)	+
Short toes (HP:0001831)	+
Pes planus (HP:0001763)	+
Neurological	
Seizures (T/C, atonic, complex, partial, tonic, gelastic) (HP:0001250)	+
Growth	
Currently short stature (10th percentile) (HP:0004322)	+
Behavioral	
Autistic behavior (HP:0000729)	+

# Analysis of genetic variants

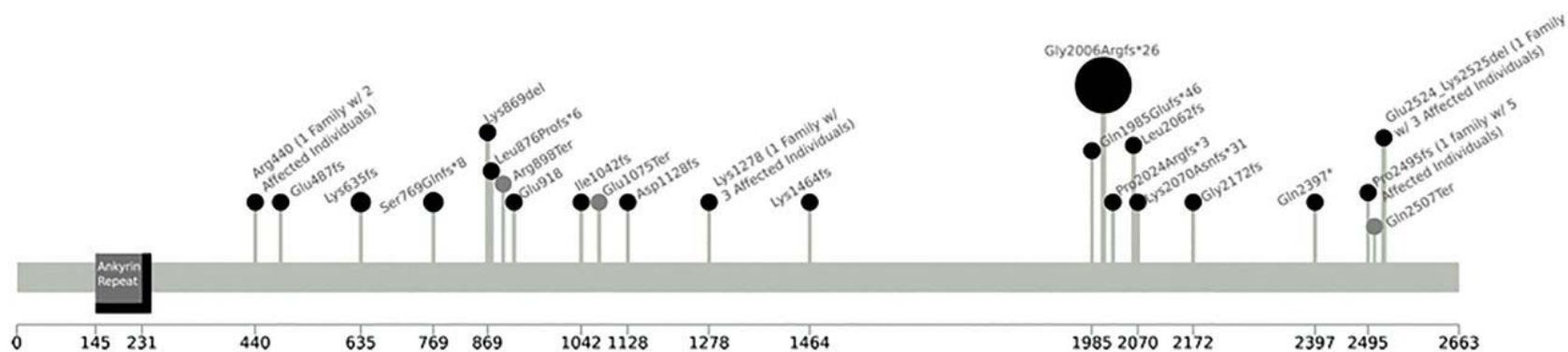
- Sequencing performed on Ion Proton with AmpliSeq Exome panel
- With standard analytical protocol, more than 1000 variants were recognized as *de novo*, well above the expected number of *de novo* mutations, suggesting low quality of sequencing and variant calling

**Table 2.** Count of single-nucleotide polymorphisms (SNPs), insertions and deletions (indels), and the total number of variants for each sequenced family member

Individual	Number of single-nucleotide polymorphisms	Number of insertions/deletions	Total number of variants
Proband	21,014	769	21,783
Mother	21,224	1011	22,235
Father	20,203	953	21,156
Sister 1	21,030	959	21,989
Sister 2	21,458	1046	22,504
Brother	20,163	1253	21,416

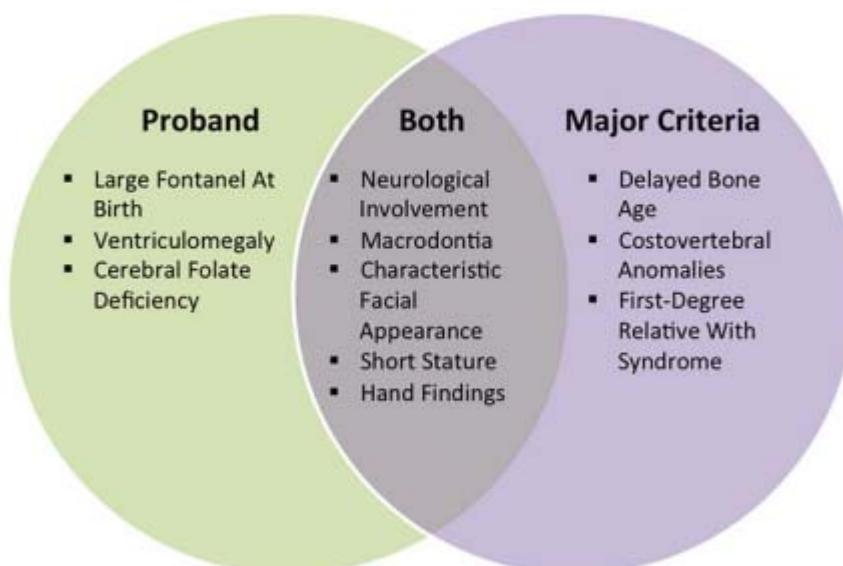
# Phenolyzer+wANNOVAR joint analysis

- Despite all the noises, Phenolyzer and wANNOVAR indicated that a heterozygous frameshift mutation in *ANKRD11* is the top candidate mutation.



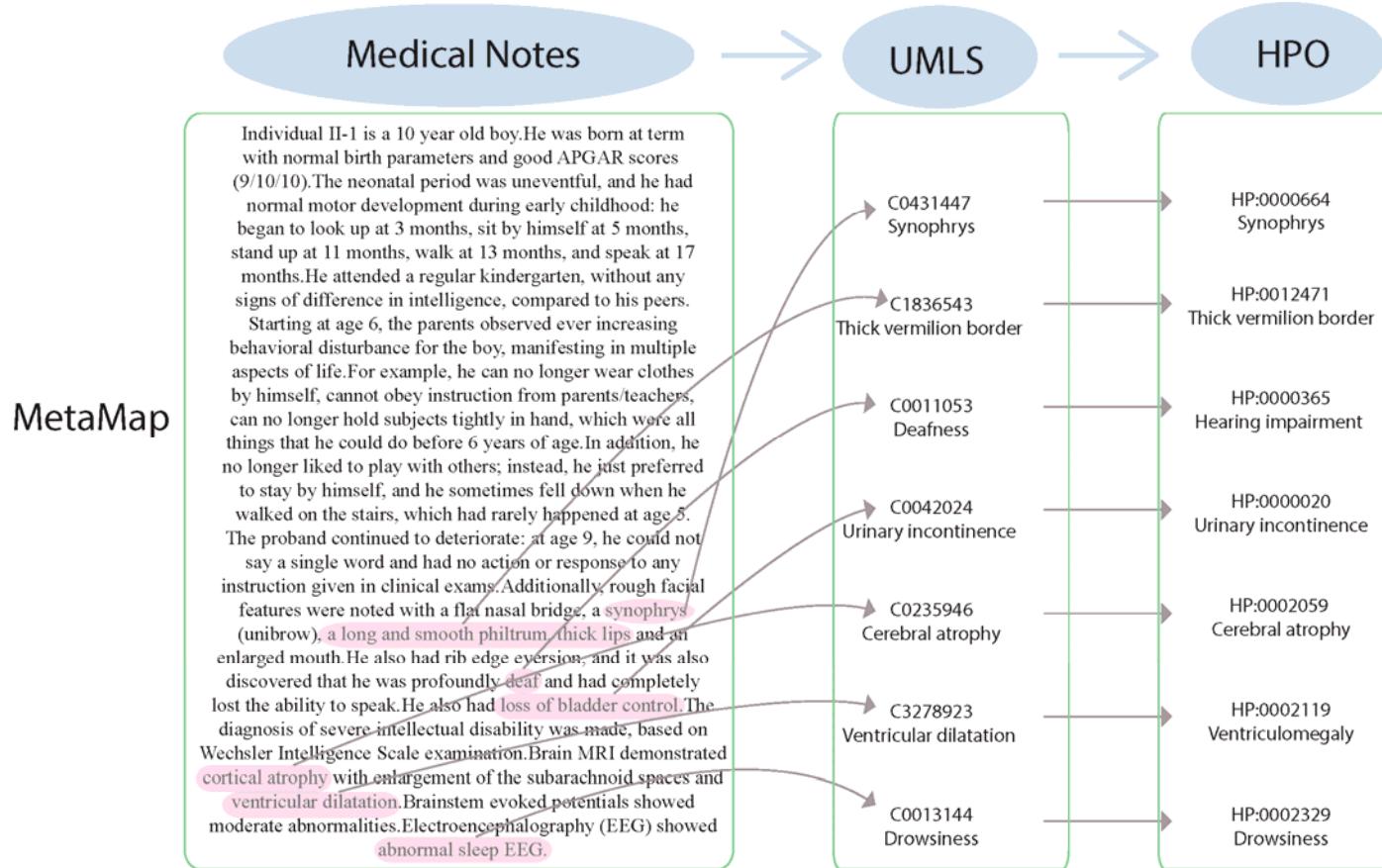
# Molecular diagnosis of KBG syndrome

- KBG syndrome is a rare autosomal dominant genetic condition characterized by neurological involvement and distinct facial, hand, and skeletal features.
- 70 cases have been reported
- Highly heterogeneous phenotypic features



**Proband met 5 of the 8 phenotypic criteria previous suggested to diagnose KBG syndrome**

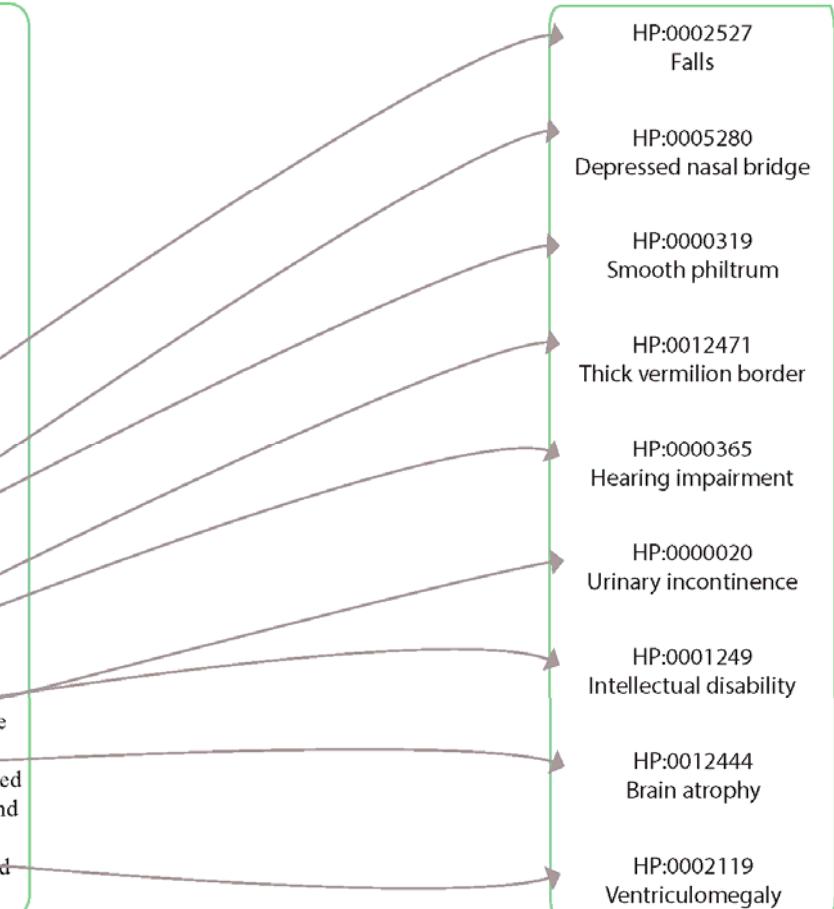
# Computational generation of HPO terms?



# Computational generation of HPO terms?

MedLEE

Individual II-1 is a 10 year old boy. He was born at term with normal birth parameters and good APGAR scores (9/10/10). The neonatal period was uneventful, and he had normal motor development during early childhood: he began to look up at 3 months, sit by himself at 5 months, stand up at 11 months, walk at 13 months, and speak at 17 months. He attended a regular kindergarten, without any signs of difference in intelligence, compared to his peers. Starting at age 6, the parents observed ever increasing behavioral disturbance for the boy, manifesting in multiple aspects of life. For example, he can no longer wear clothes by himself, cannot obey instruction from parents/teachers, can no longer hold subjects tightly in hand, which were all things that he could do before 6 years of age. In addition, he no longer liked to play with others; instead, he just preferred to stay by himself, and he sometimes fell down when he walked on the stairs, which had rarely happened at age 5. The proband continued to deteriorate: at age 9, he could not say a single word and had no action or response to any instruction given in clinical exams. Additionally, rough facial features were noted with a flat nasal bridge, a synophrys (unibrow), a long and smooth philtrum, thick lips and an enlarged mouth. He also had rib edge eversion, and it was also discovered that he was profoundly deaf and had completely lost the ability to speak. He also had loss of bladder control. The diagnosis of severe intellectual disability was made, based on Wechsler Intelligence Scale examination. Brain MRI demonstrated cortical atrophy with enlargement of the subarachnoid spaces and ventricular dilatation. Brainstem evoked potentials showed moderate abnormalities. Electroencephalography (EEG) showed abnormal sleep EEG.



# Go back to case study #1

- We could have used natural language processing to find the disease causal gene automatically

