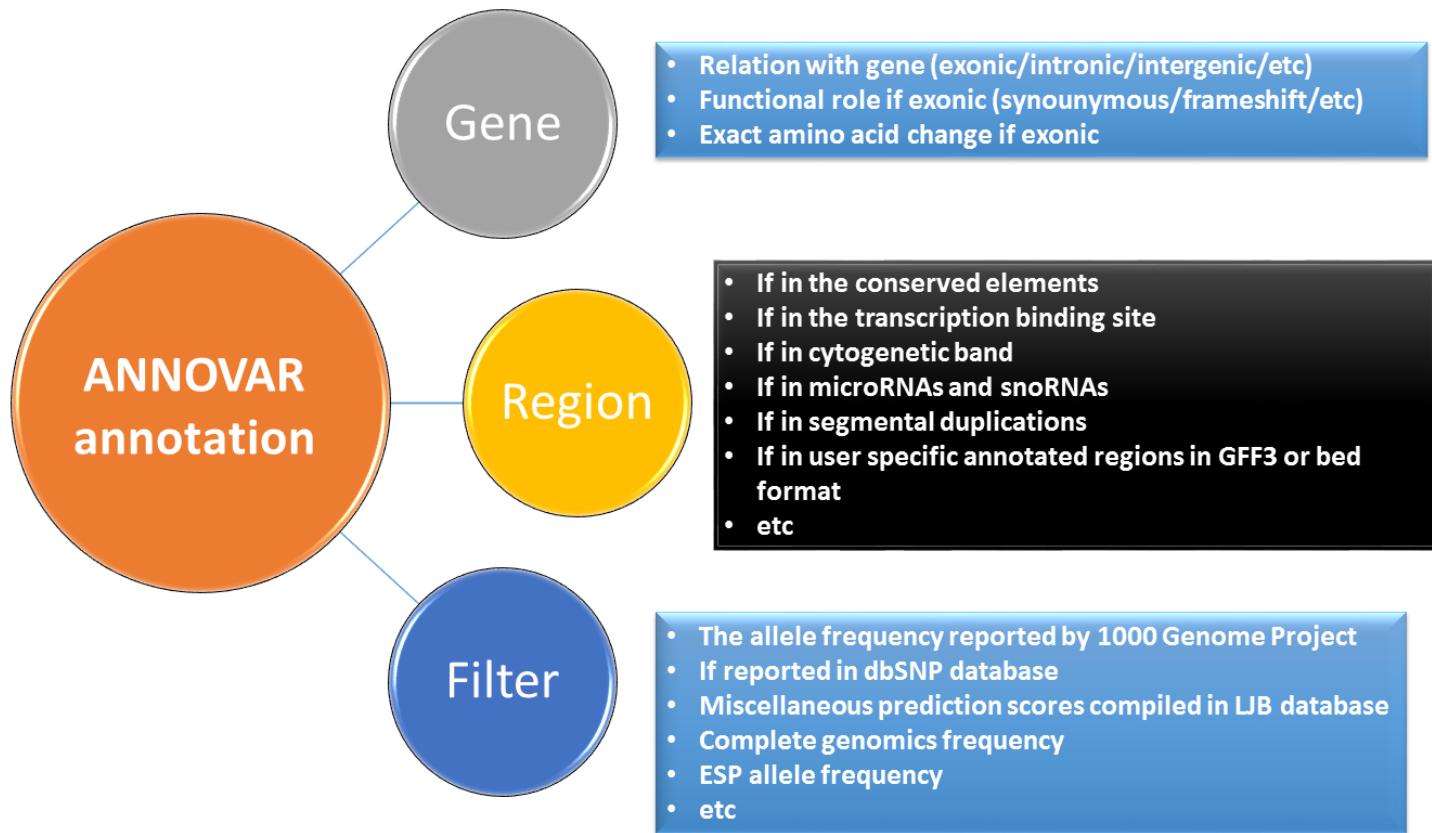


Annotation and phenotype- driven interpretation of genetic variants

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
-m, --man
-v, --verbose
--protocol <string>
--operation <string>
--outfile <string>
--buildver <string>
--remove
--(no)checkfile
--genericdbfile <files>
--gff3dbfile <files>
--bedfile <files>
--vcfdbfile <files>
--otherinfo
--onetranscript
--nastranscript
--csvout
--argument <string>
--tempdir <dir>
--vcfinput
--dot2underline
```



wANNOVAR

Home

Tutorial

Example

Related projects ▾

WLab

Function: automatically run a pi
their functional effects in a co
manual filtering

Example: table_annovar.pl example1
2014oct_all,1000g2014oct_afn,1000g201
table_annovar.pl example1
oct_all,1000g2014oct_afn,1000g2014oct

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

wANNOVAR

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

Get Started

About

Contact

Like Share 6 people like this. Be the first of your friends.

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. At the top, there is a blue header bar with the following navigation items: 'ANNOVAR Documentation' (highlighted in blue), 'ANNOVAR' (highlighted in blue), 'User Guide ▾', 'Misc ▾', 'Articles ▾', a search bar with a magnifying glass icon labeled 'Search', a 'Previous' button with a left arrow, a 'Next' button with a right arrow, and a 'Edit on GitHub' button.

In the main content area, there is a sidebar on the left with the title 'ANNOVAR Documentation' and a link 'Reference'. The main title 'ANNOVAR Documentation' is displayed prominently in large, bold, dark gray font. Below the title, a paragraph of text describes the tool's purpose and capabilities. A bulleted list follows, detailing specific annotation types: Gene-based annotation, Region-based annotation, Filter-based annotation, and Other functionalities. At the bottom of the page, a note encourages users to navigate the site via the menu and provides contact information for support.

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!

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 View: Sequence | Set Origin | Views & Tools ▾

Marker View: Markers | Search... ▾

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

Sequence View: Sequence | Flip Strands | Tools ▾

Marker View: Markers | Configure ▾

LSDB Clinically Associated Variants

Cited Variants

Genes

Alignments

1 K 2 K 3 K 4 K 5 K 6 K 7 K 8 K 9 K 10 K 11 K 12 K 13 K 14 K 14,961

Sequence NG_008720.1: Homo sapiens hemochromatosis (HFE), RefSeqGene on chromosome 6

2 + 1 2 1 1 1 1 1

HFE NM_000410.3 NP_000401.1 exon 1 exon 2a exon 4a exon 6

NM_000410.3 NM_139003.2 NM_139004.2

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.

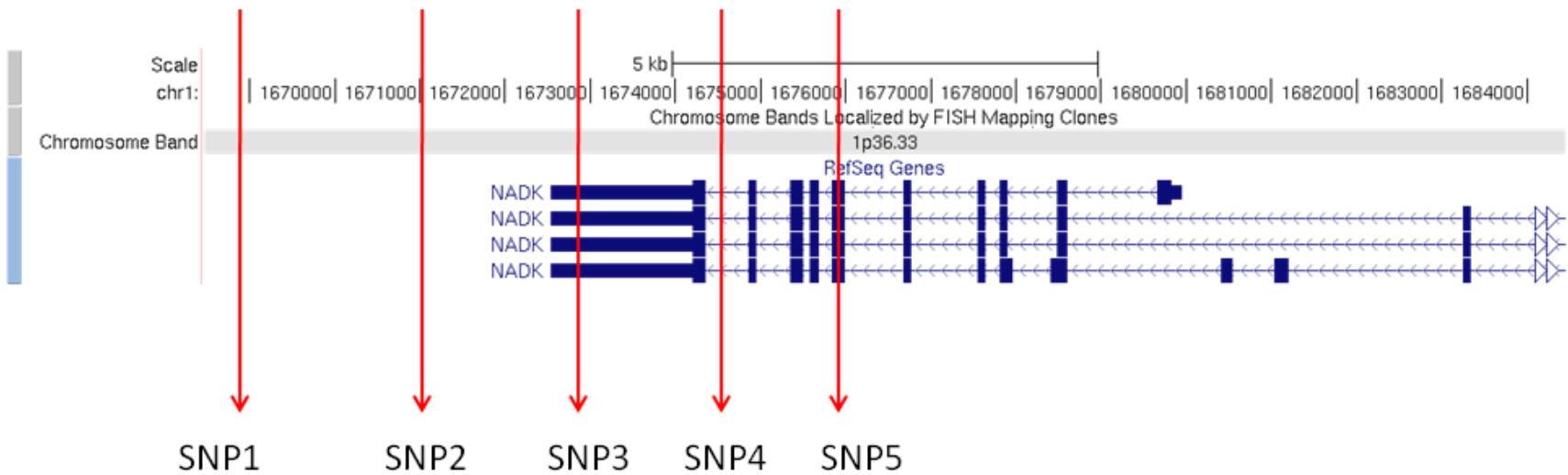
There are ~50K, ~80K, ~200K, ~190K transcripts in the four gene models, respectively

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)

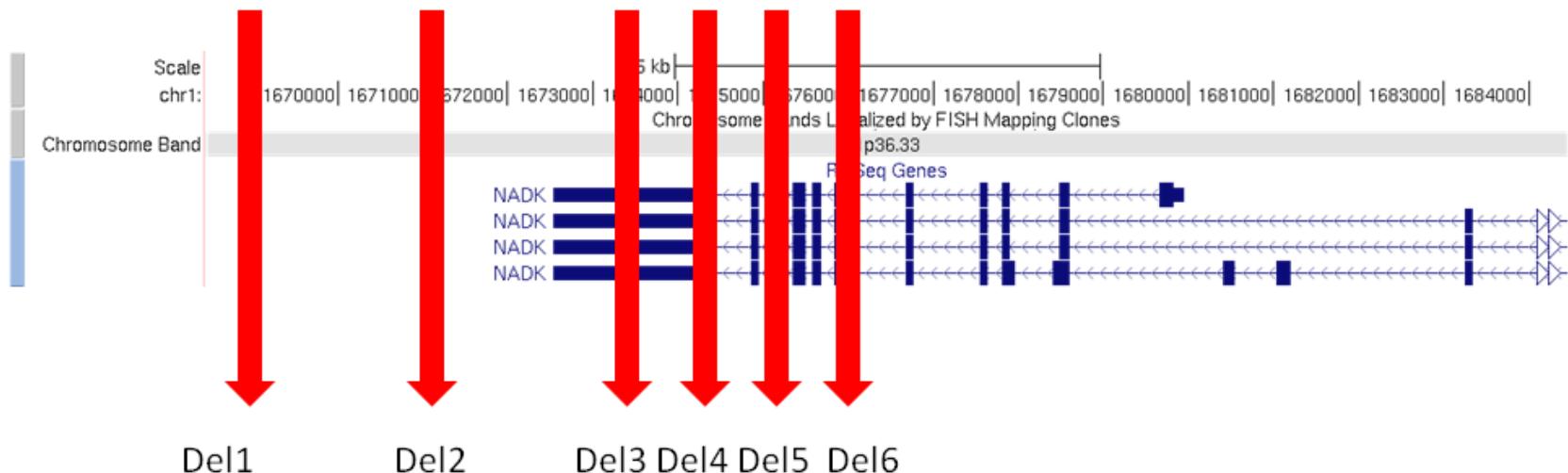
Example: SNVs

- 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 mutliples 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)

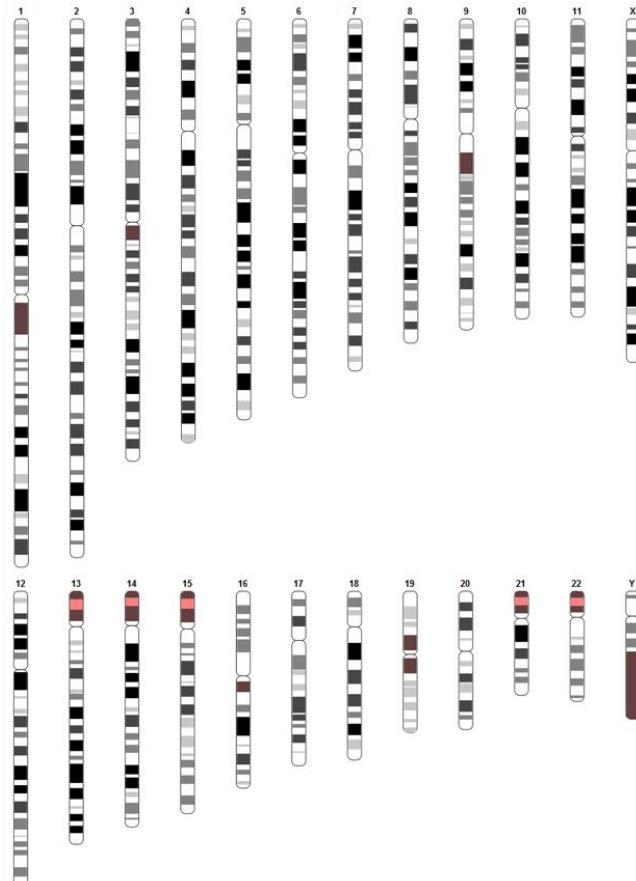
Region-based annotation

- 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



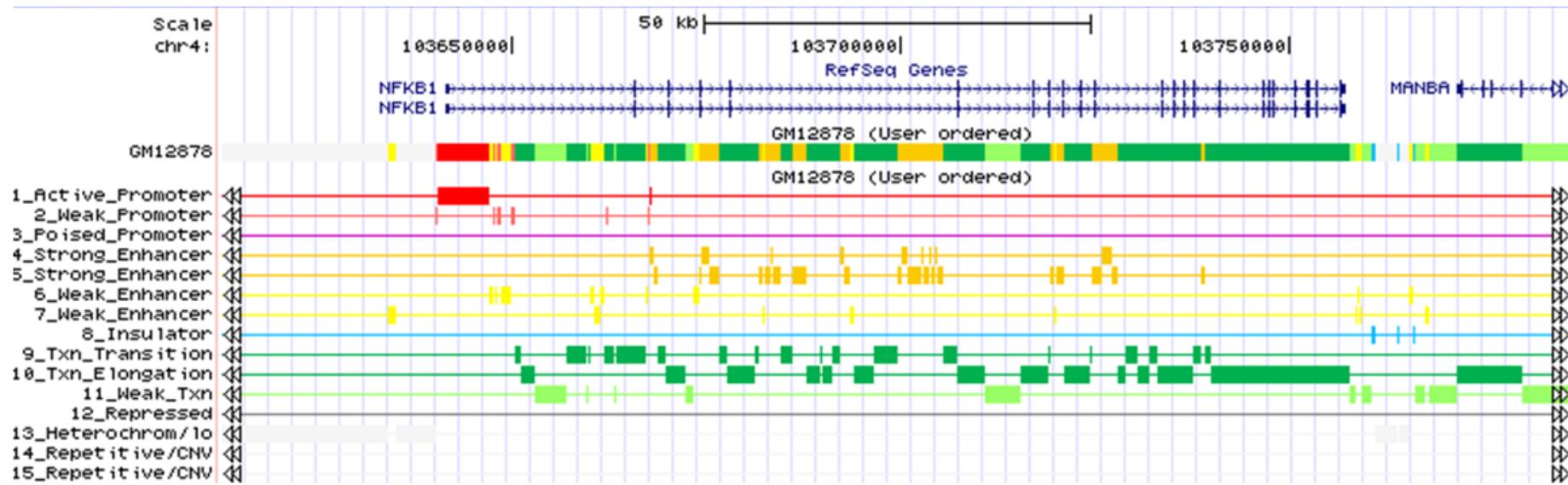
ENCODE ChIP-Seq peaks

- 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

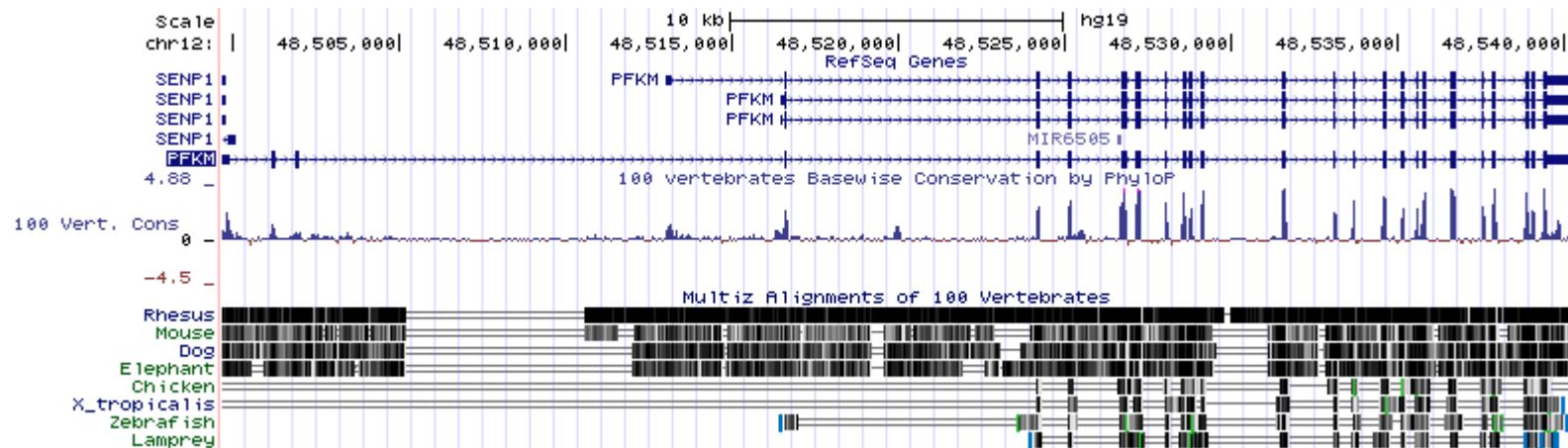
chromHMM predictions

- 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



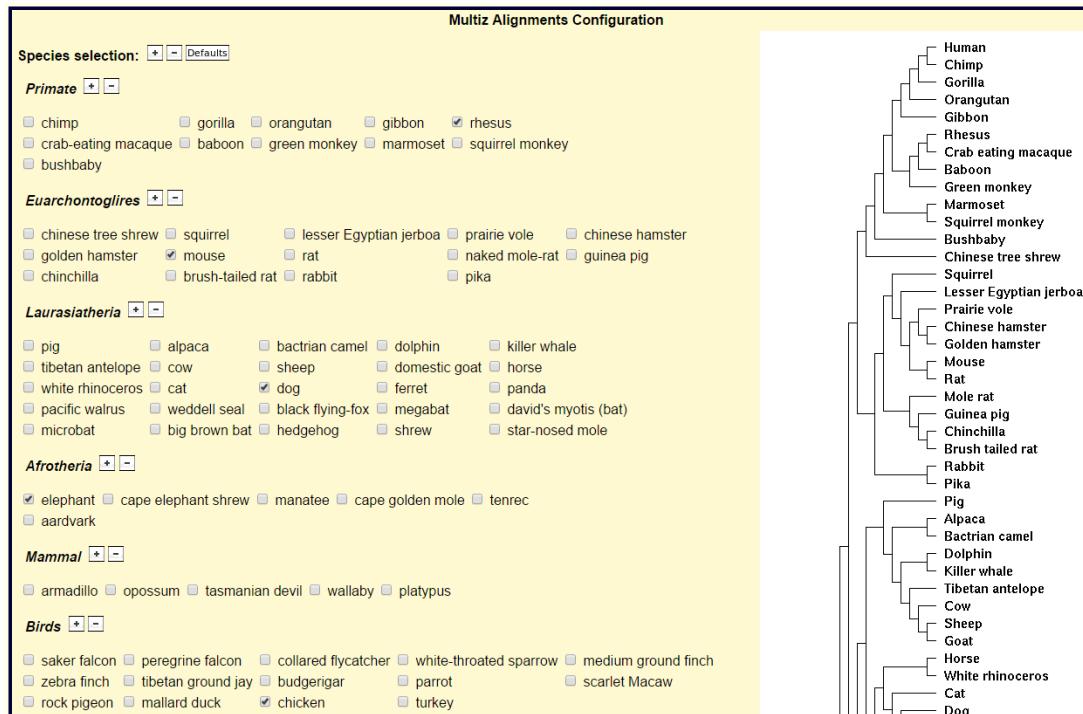
Conserved genomic regions

- 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



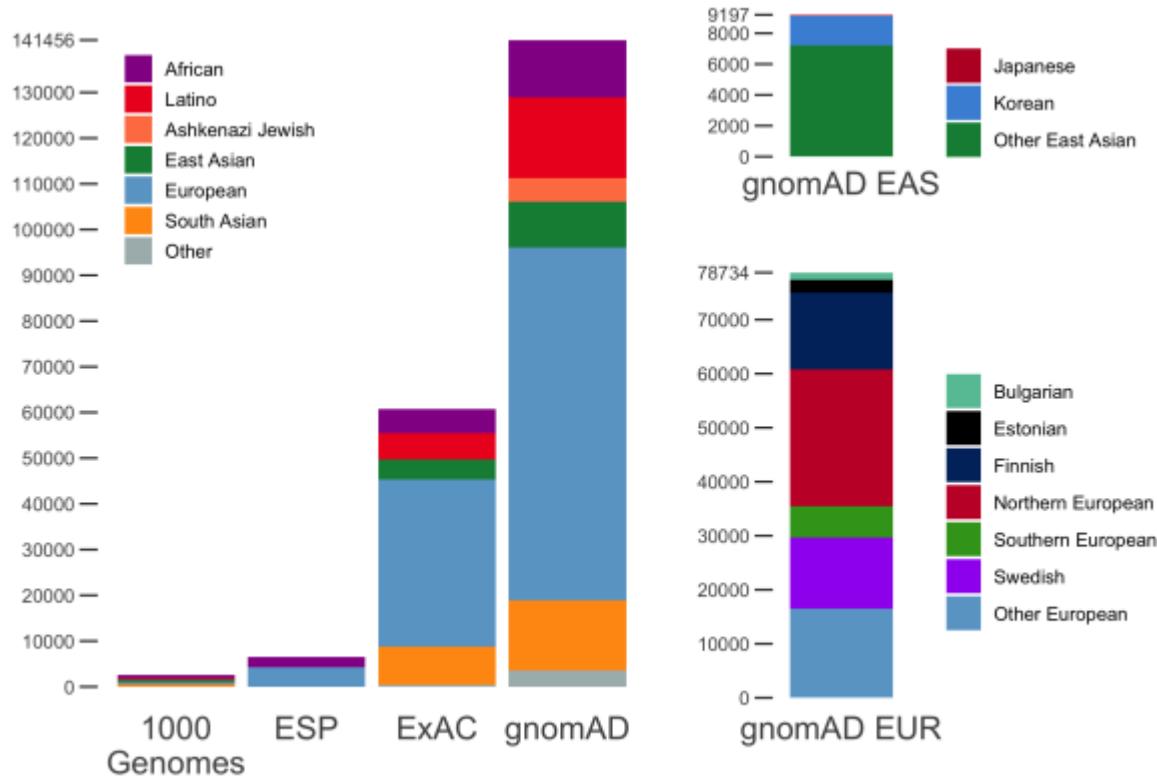
Filter-based annotation

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

Allele frequency databases

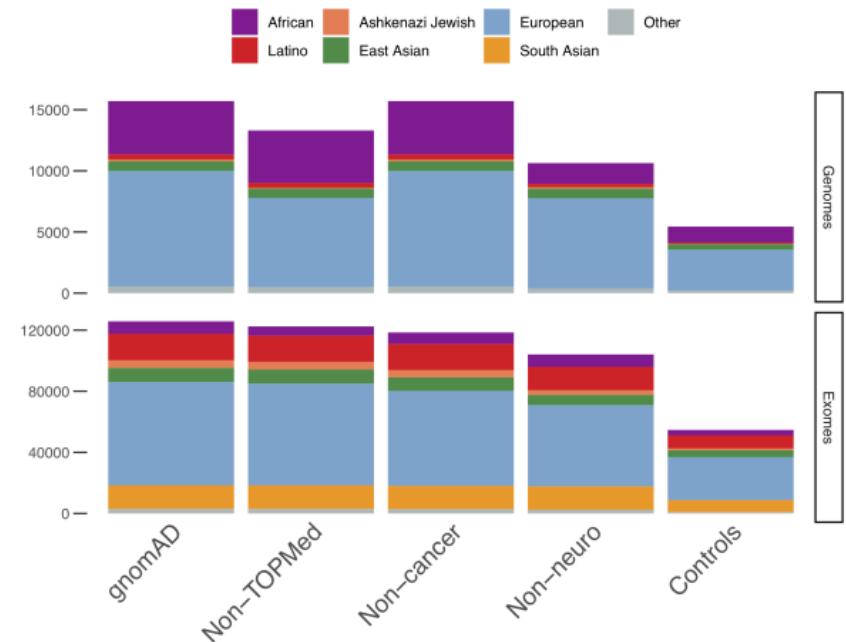
- 1000 Genomes Project: ~2500 genomes
 - Keyword is 1000g2015aug_all
- ExAC (exome aggregation consortium): ~65000 exomes
 - Keyword is exac03
- NHLBI-ESP6500 (European Americans and African Americans): ~6500 exomes
 - Keyword is esp6500siv2_aa, esp6500siv2_ea
- gnomAD (genome aggregation consortium): 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,avsnp147,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	
1	Y	2715180	rs11575897	G	A	39	.	AC=5;AN=58;DB;DP=168;NS=61;NR	GT:GQ:DP	0:60:5	0:19:1	
2	Y	2728456	rs2058276	T	C	32	.	AC=42;AN=61;DB;DP=182;H2;NS=65;NR	GT:GQ:DP	0:63:6	0:0:0	
3	Y	2731229	.	C	T	23	.	AC=6;AN=55;DP=191;NS=59;NR	GT:GQ:DP	0:69:8	0:10:3	
4	Y	2734240	.	G	A	31	.	AC=5;AN=59;DP=196;NS=63;NR	GT:GQ:DP	0:46:3	1:22:2	
5	Y	2740720	rs9785893	C	T	18	.	AC=10;AN=59;DB;DP=163;NS=64;NR	GT:GQ:DP	0:54:5	0:57:4	
6	Y	2743014	.	G	T	32	.	AC=4;AN=65;DP=224;NS=68;NR	GT:GQ:DP	0:72:10	0:69:8	
7	Y	2743242	.	C	T	25	.	AC=4;AN=62;DP=275;NS=66;NR	GT:GQ:DP	0:43:4	1:14:4	
8	Y	2746727	.	A	G	34	.	AC=42;AN=58;DP=179;NS=64;NR	GT:GQ:DP	0:66:7	1:2:1	
9	Y	2765306	.	A	G	34	.	AC=4;AN=57;DP=188;NS=61;NR	GT:GQ:DP	0:66:7	0:51:2	
10	Y	2782506	rs2075640	A	G	38	.	AC=29;AN=62;DB;DP=254;H2;NS=66;NR	GT:GQ:DP	0:60:5	0:51:7	
11	Y	2783755	.	G	A	51	.	AC=5;AN=63;DP=217;NS=67;NR	GT:GQ:DP	0:60:5	1:60:6	
12	Y	2786402	.	T	G	41	.	AC=7;AN=61;DP=209;NS	Allele count = 29 Total number of alleles = 62 dbSNP membership	:7	0:63:6	
13	Y	2788927	rs56004558	A	G	38	.	AC=24;AN=57;DB;DP=17		:2	0:54:3	
14	Y	2794854	rs35284970	C	T	38	.	AC=2;AN=62;DB;DP=181	Combined depth across samples = 254 hapmap2 membership	:4	0:57:4	
15	Y	2798273	.	G	A	35	.	AC=13;AN=62;DP=189;N		:3	0:28:3	
16	Y	2799468	rs9786209	C	T	36	.	AC=18;AN=62;DB;DP=24	Number of samples = 66	:8	0:54:3	
17	Y	2812887	.	G	T	35	.	AC=4;AN=52;DP=189;NS	NR	:7	0:51:2	
18	Y	2813096	.	G	A	54	.	AC=1;AN=58;DP=167;NS=62;NR	GT:GQ:DP	0:51:2	0:54:5	
1	A	B	C	D	E	F	G	H	I	J	K	
2	Chr	Start	End	Ref	Alt	Func.refGene	Gene.refGene	GeneDetail.refGene	ExonicFunc.refGene	AAChange.refGene	cytoBand	genomicS
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_192q37.1	.	.
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_0040013q12.11	.	.
16	13	20797176	21105944	0	-	exonic	CRY1 GIB6	.	frameshift deletion	GIB6:NM_001113q12.11	.	.

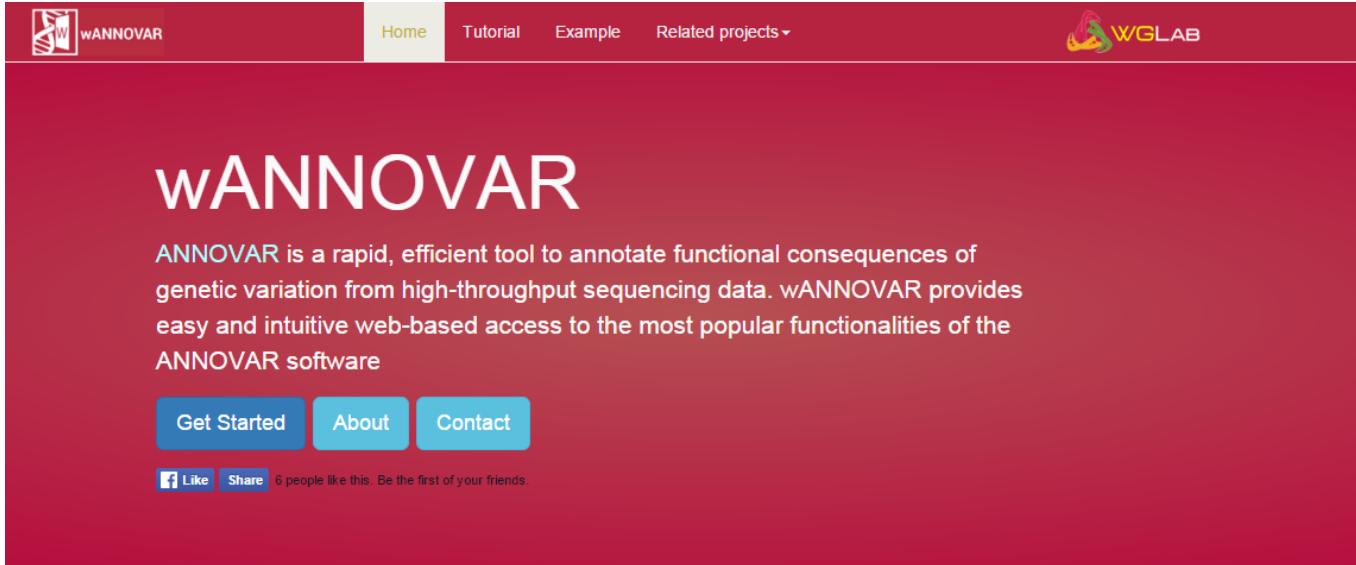
Using web version of ANNOVAR

- 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 navigation bar is the WGLAB logo. The main content area has a red background. The title "wANNOVAR" is displayed prominently in large white letters. 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". Below this text are three blue buttons labeled "Get Started", "About", and "Contact". At the bottom of the main content area, there are social sharing links for Facebook and Twitter, followed by the text "6 people like this. Be the first of your friends."

Basic Information

Email	<input type="text" value="Email"/>
Sample Identifier	<input type="text" value="Sample Identifier"/>
Input File	<input type="button" value="+ Input File"/>
or Paste Variant Calls	<input type="text" value="paste your variant call here"/>

I agree to the Terms of Use

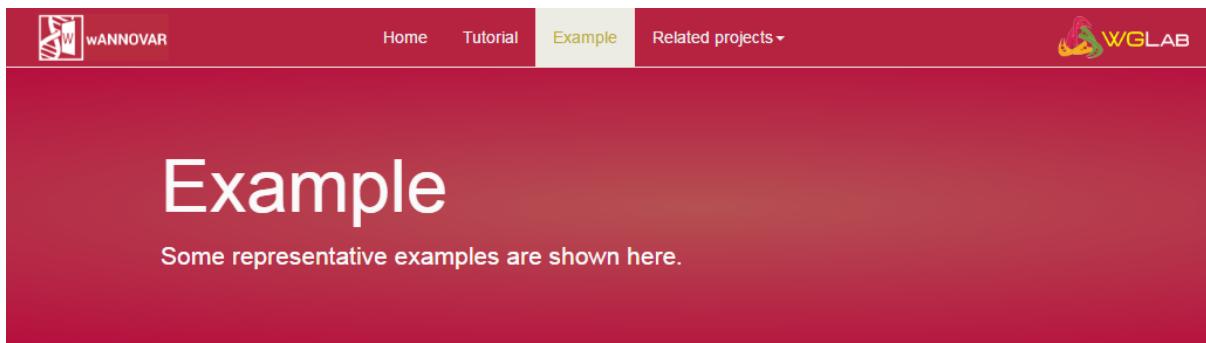
Recent Updates

[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

alzheimer

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.

3) Choose how long you want your result reserved in the server

Parameter Settings

Result duration

2 months



4) Choose Reference Genome version

Reference Genome

hg19



5) Choose Input Format

Input Format

VCF



6) Choose gene definition

Gene Definition

RefSeq Gene



7) Choose whether you want all the variants or only for one individual

Individual analysis

Individual analysis



8) Choose disease model for variant filtration

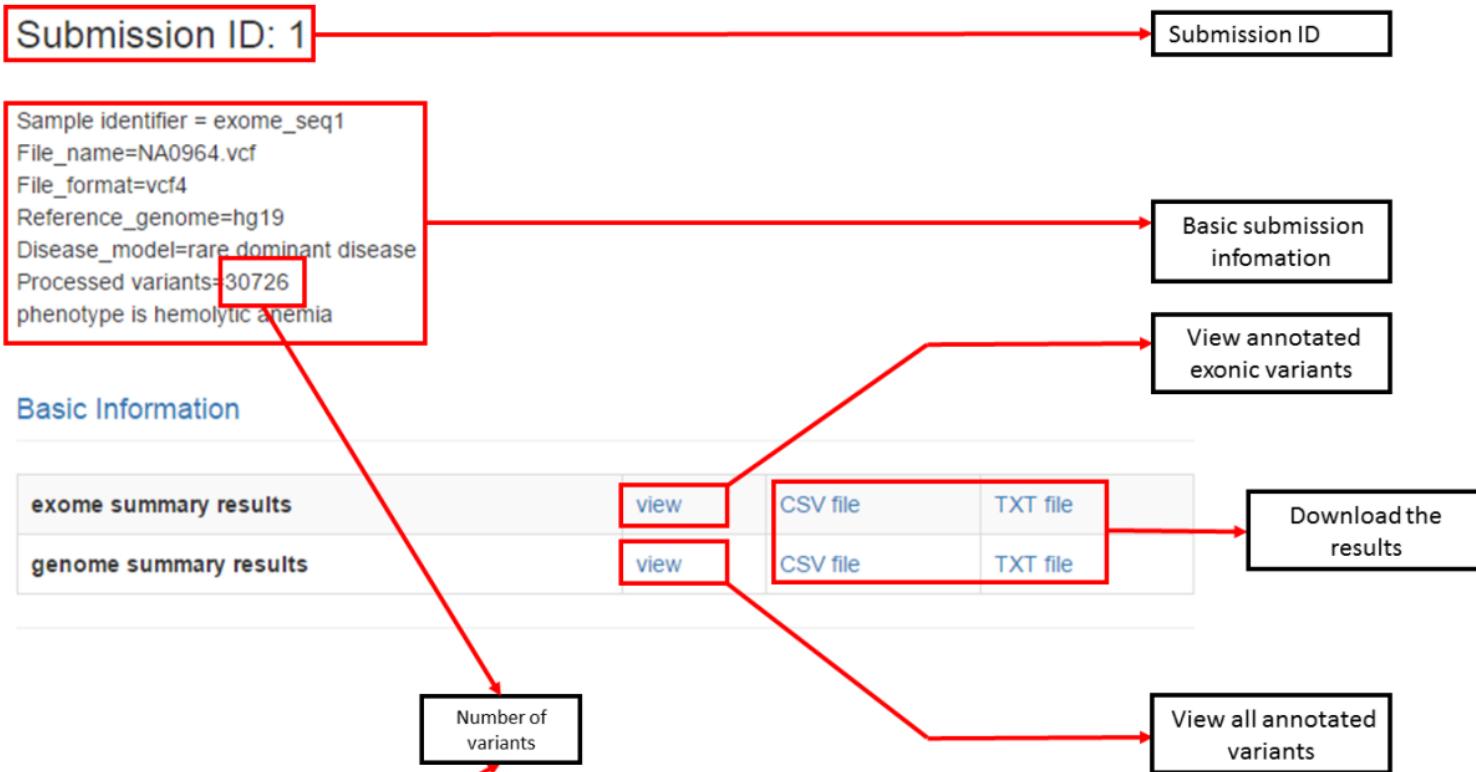
Disease Model

none



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](#)

Step1:7963 variants	Identify missense, nonsense and splicing variants	download
Step2:584 variants	Remove variants in the 1000 Genomes Project(ALL) with MAF>0.01	download
Step3:421 variants	Remove variants in NHLBI-ESP 6500 exomes with MAF>0.01	download
Step4:80 variants	Remove variants in dbSNP138 (excluding clinically associated SNPs)	download
Step5:76 genes	Compile a list of candidate genes based on disease model	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, PhD¹, Nazneen Aziz, PhD^{2,16}, Sherri Bale, PhD³, David Bick, MD⁴, Soma Das, PhD⁵, Julie Gastier-Foster, PhD^{6,7,8}, Wayne W. Grody, MD, PhD^{9,10,11}, Madhuri Hegde, PhD¹², Elaine Lyon, PhD¹³, Elaine Spector, PhD¹⁴, Karl Voelkerding, MD¹³ and Heidi L. Rehm, PhD¹⁵, on behalf of the ACMG Laboratory Quality Assurance Committee

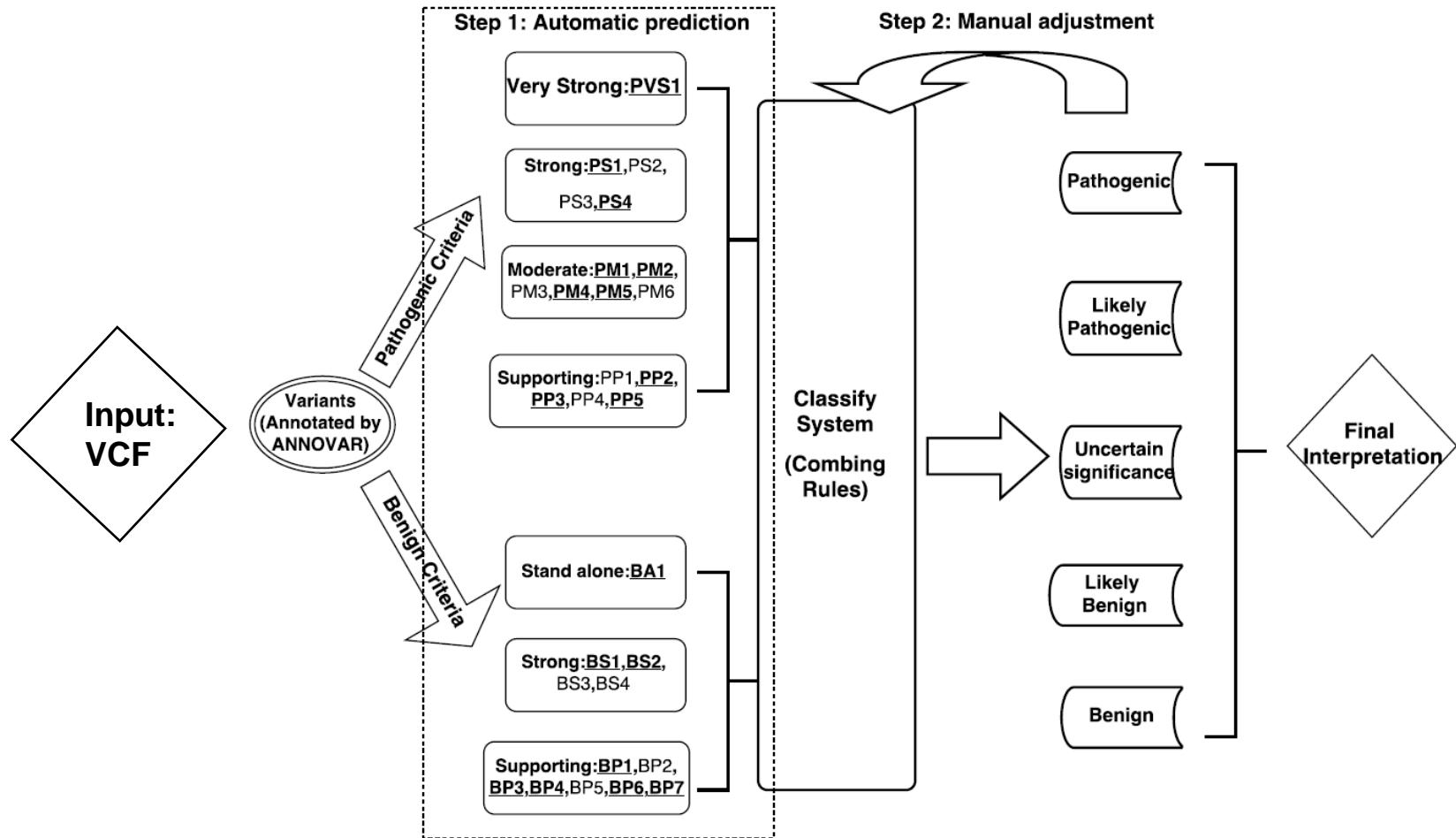
Disclaimer: These ACMG Standards and Guidelines were developed primarily as an educational resource for clinical laboratory geneticists to help them provide quality clinical laboratory services. Adherence to these standards and guidelines is voluntary and does not necessarily assure a successful medical 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

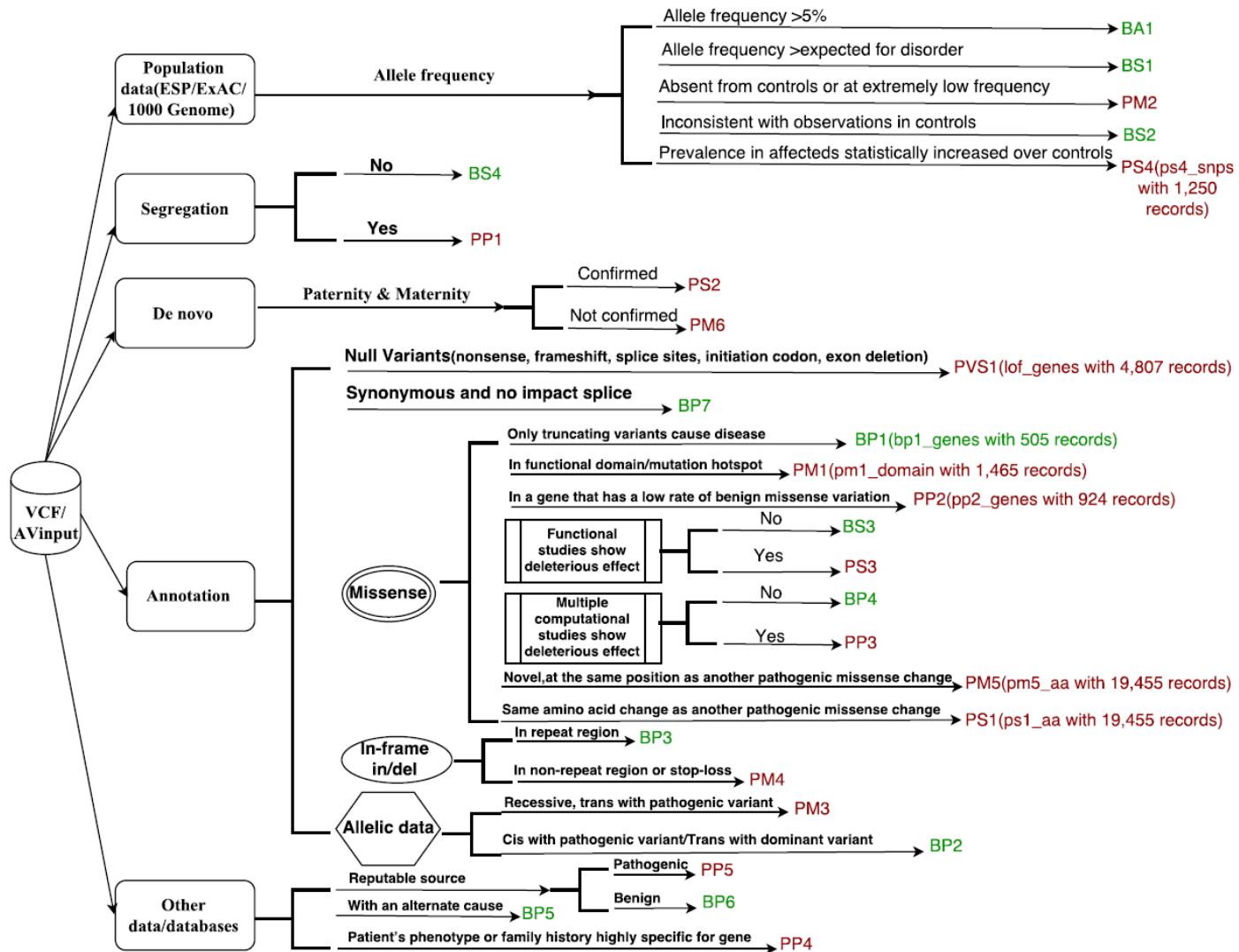
Pathogenic	(i) 1 Very strong (PVS1) AND (a) ≥ 1 Strong (PS1–PS4) OR (b) ≥ 2 Moderate (PM1–PM6) OR (c) 1 Moderate (PM1–PM6) and 1 supporting (PP1–PP5) OR (d) ≥ 2 Supporting (PP1–PP5) (ii) ≥ 2 Strong (PS1–PS4) OR (iii) 1 Strong (PS1–PS4) AND (a) ≥ 3 Moderate (PM1–PM6) OR (b) 2 Moderate (PM1–PM6) AND ≥ 2 Supporting (PP1–PP5) OR (c) 1 Moderate (PM1–PM6) AND ≥ 4 supporting (PP1–PP5)	Benign	(i) 1 Stand-alone (BA1) OR (ii) ≥ 2 Strong (BS1–BS4) Likely benign	(i) 1 Strong (BS1–BS4) and 1 supporting (BP1–BP7) OR (ii) ≥ 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 ≥ 2 supporting (PP1–PP5) OR (iv) ≥ 3 Moderate (PM1–PM6) OR (v) 2 Moderate (PM1–PM6) AND ≥ 2 supporting (PP1–PP5) OR (vi) 1 Moderate (PM1–PM6) AND ≥ 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 115828756 Ref: Alt:

Query by dbSNP ID

rs.:

Query by HGNC gene symbol

Gene: cDNA change:

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.

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 information)

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)

Transcripts (Ref) MAF in ExAC_ALL Disease in OrphaNet OMIM

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

Search:

Previous 1 Next

Web implementation of InterVar

Start wInterVar About Services ▾ Contact Related projects ▾

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.

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 (Details&Adjust)	nonsynonymous SNV	rs11466112(details of MAF)	NM_002506 p.R221W	3.308E-5 (show in 7 records)	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 multixon 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»](#)

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)



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^{1, 2}, Sebastian Köhler^{1, 2}, Sebastian Bauer¹, Dominik Seelow^{1, 3}, Denise Horn¹, Stefan Mundlos^{1, 2, 4}

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^{1,2,*}, Sandra C. Doelken¹, Christopher J. Mungall³, Sebastian Bauer¹, Helen V. Firth^{4,5}, Isabelle Bailleul-Forestier⁶, Graeme C. M. Black^{7,8}, Danielle L. Brown⁹, Michael Brudno^{10,11}, Jennifer Campbell^{9,12}, David R. FitzPatrick¹³, Janan T. Eppig¹⁴, Andrew P. Jackson¹³, Kathleen Freson¹⁵, Marta Girdea^{10,11}, Ingo Helbig¹⁶, Jane A. Hurst¹⁷, Johanna Jähn¹⁶, Laird G. Jackson¹⁸, Anne M. Kelly¹⁹, David H. Ledbetter²⁰, Sahar Mansour²¹, Christa L. Martin²⁰, Celia Moss²², Andrew Mumford²³, Willem H. Ouwehand^{4,19}, Soo-Mi Park⁵, Erin Rooney Riggs²⁰, Richard H. Scott²⁴, Sanjay Sisodiya²⁵, Steven Van Vooren²⁶, Ronald J. Wapner²⁷, Andrew O. M. Wilkie²⁸, Caroline F. Wright⁴, Anneke T. Vulto-van Silfhout²⁹, Nicole de Leeuw²⁹, Bert B. A. de Vries²⁹, Nicole L. Washington³, Cynthia L. Smith¹⁴, Monte Westerfield³⁰, Paul Schofield^{14,31}, Barbara J. Ruet³⁰, Georgios V. Gkoutos³², Melissa Haendel³³, Damian Smedley⁴, Suzanna E. Lewis³ and Peter N. Robinson^{1,2,34,*}

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^{1,*}, Nicole A. Vasilevsky², Mark Engelstad², Erin Foster², Julie McMurry², Ségolène Aymé³, Gareth Baynam^{4,5}, Susan M. Bello⁶, Cornelius F. Boerkoel⁷, Kym M. Boycott⁸, Michael Brudno⁹, Orion J. Buske⁹, Patrick F. Chinnery^{10,11}, Valentina Cipriani^{12,13}, Laureen E. Connell¹⁴, Hugh J.S. Dawkins¹⁵, Laura E. DeMare¹⁴, Andrew D. Devereau¹⁶, Bert B.A. de Vries¹⁷, Helen V. Firth¹⁸, Kathleen Freson¹⁹, Daniel Greene^{20,21}, Ada Hamosh²², Ingo Helbig^{23,24}, Courtney Hum²⁵, Johanna A. Jähn²⁴, Roger James^{11,21}, Roland Krause²⁶, Stanley J. F. Laulederkind²⁷, Hanns Lochmüller²⁸, Gholson J. Lyon²⁹, Soichi Ogishima³⁰, Annie Olry³¹, Willem H. Ouwehand²⁰, Nikolas Pontikos^{12,13}, Ana Rath³¹, Franz Schaefer³², Richard H. Scott¹⁶, Michael Segal³³, Panagiota I. Sergouniotis³⁴, Richard Sever¹⁴, Cynthia L. Smith⁶, Volker Straub²⁸, Rachel Thompson²⁸, Catherine Turner²⁸, Ernest Turro^{20,21}, Marijcke W.M. Veltman¹¹, Tom Vulliamy³⁵, Jing Yu³⁶, Julie von Ziegenweidt²⁰, Andreas Zankl^{37,38}, Stephan Züchner³⁹, Tomasz Zemojtel¹, Julius O.B. Jacobsen¹⁶, Tudor Groza^{40,41}, Damian Smedley¹⁶, Christopher J. Mungall⁴², Melissa Haendel² and Peter N. Robinson^{43,44,*}

HPO in 2019: new website and expanded knowledge base

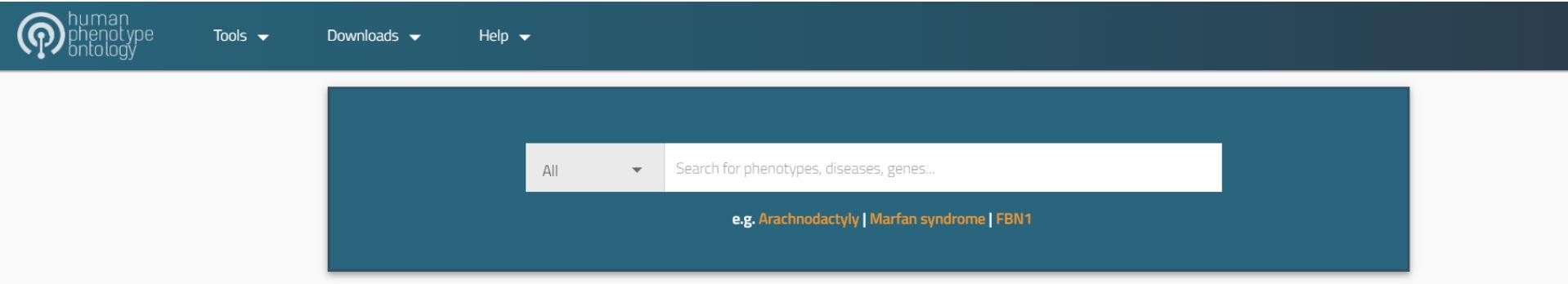
D1018–D1027 *Nucleic Acids Research*, 2019, Vol. 47, Database issue
doi: 10.1093/nar/gky1105

Published online 22 November 2018

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

Sebastian Köhler^{①,2,3}, Leigh Carmody^{3,4}, Nicole Vasilevsky^{③,5}, Julius O.B. Jacobsen^{3,6}, Daniel Danis^{3,4}, Jean-Philippe Gourdine^{③,5}, Michael Gargano^{3,4}, Nomi L. Harris^{3,7}, Nicolas Matentzoglu^{3,8}, Julie A. McMurry^{③,9}, David Osumi-Sutherland^{3,8}, Valentina Cipriani^{③,10,11,12}, James P. Balhoff^{③,13}, Tom Conlin^{③,9}, Hannah Blau^{③,4}, Gareth Baynam^{14,15,16,17,18}, Richard Palmer¹⁷, Dylan Gratian¹⁴, Hugh Dawkins¹⁸, Michael Segal¹⁹, Anna C. Jansen^{20,21}, Ahmed Muaz^{3,22}, Willie H. Chang²³, Jenna Bergerson²⁴, Stanley J.F. Laulederkind^{②5}, Zafer Yüksel^{②6}, Sergi Beltran^{②7,28}, Alexandra F. Freeman²⁴, Panagiotis I. Sergouniotis²⁹, Daniel Durkin⁴, Andrea L. Storm^{30,31}, Marc Hanauer³², Michael Brudno²³, Susan M. Bello^{③33}, Murat Sincan³⁴, Kayli Rageth³⁴, Matthew T. Wheeler^{③5}, Renske Oegema³⁶, Halima Lourghi³², Maria G. Della Rocca^{30,31}, Rachel Thompson^{③7}, Francisco Castellanos⁴, James Priest³⁸, Charlotte Cunningham-Rundles³⁹, Ayushi Hegde⁴, Ruth C. Lovering^{④0}, Catherine Hajek³⁴, Annie Olry³², Luigi Notarangelo²⁴, Morgan Similuk²⁴, Xingmin A. Zhang^{③,4}, David Gómez-Andrés⁴¹, Hanns Lochmüller^{②7,42,43,44}, Hélène Dollfus⁴⁵, Sergio Rosenzweig⁴⁶, Shruti Marwaha³⁵, Ana Rath^{③2}, Kathleen Sullivan⁴⁷, Cynthia Smith^{③33}, Joshua D. Milner²⁴, Dorothée Leroux⁴⁵, Cornelius F. Boerkoel³⁴, Amy Klion²⁴, Melody C. Carter²⁴, Tudor Groza^{3,22}, Damian Smedley^{3,6}, Melissa A. Haendel^{③,5,9}, Chris Mungall^{3,7} and Peter N. Robinson^{③,4,48,*}

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



The Human Phenotype Ontology

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](#)

News & Updates

[June 2019 release](#)

June 16, 2019

[April 2019 HPO Release](#)

April 16, 2019

[hpo-web 1.5.0](#)

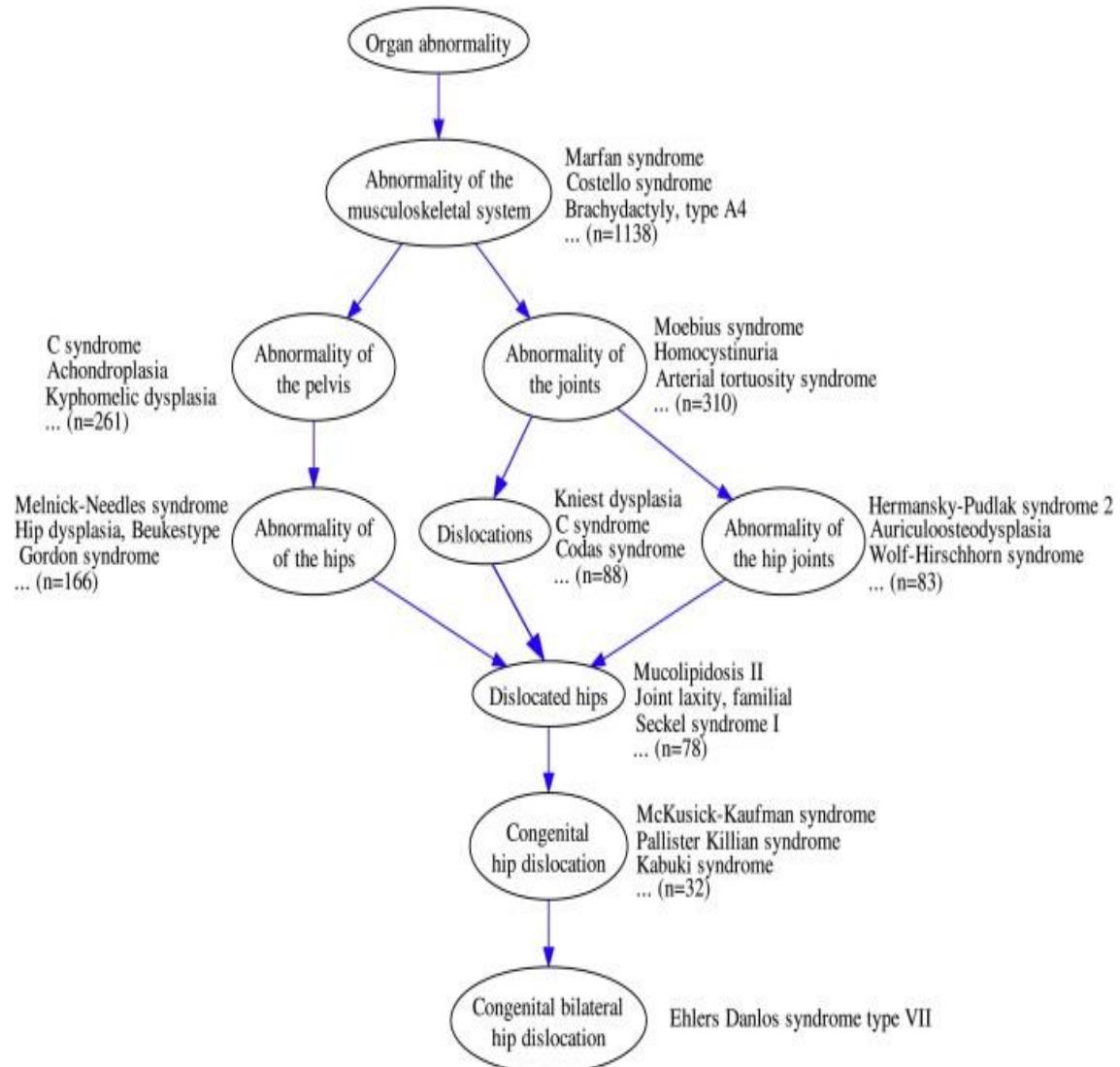
April 16, 2019

[View All News](#)

<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</p> <p>Alternative IDs HP:0007050, HP:0002513, HP:0001253, HP:0007157</p> <p>PURL http://purl.obolibrary.org/obo/HP_0001251</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).</p> <p>Logical definition 'has part' some</p> <p>Intersection of</p> <ul style="list-style-type: none">- increased amount- 'inheres in' some ataxia- 'has modifier' some abnormal

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.

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	EN03
<input checked="" type="checkbox"/>	0.0484	OMIM:609...	MYOPATHY, AUTOPHAGIC VACUOLAR, INFAR...	
<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, CONGENI...	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
<input checked="" type="checkbox"/>	0.0484	OMIM:600...	#600710 MYOPATHY, REDUCING BODY, X-LIN...	FHL1

◀ | Page 1 of 268 | ▶ | ⌂ | Improve Differential Diagnosis. Download Results.

Input: HPO terms

Output: Disease diagnosis and p-values

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

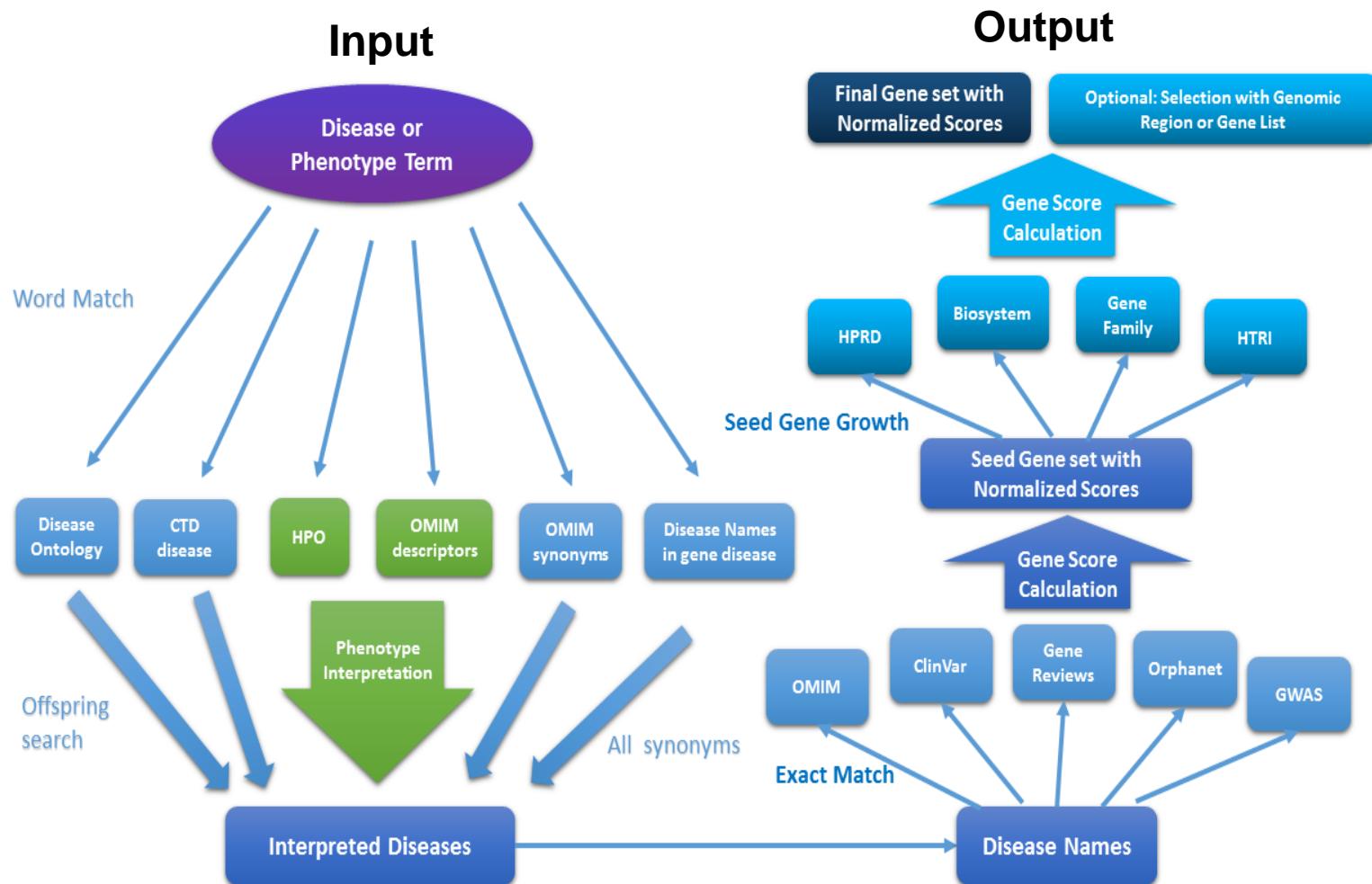


Expedite the discovery of disease causal variants from exome/genome sequencing data

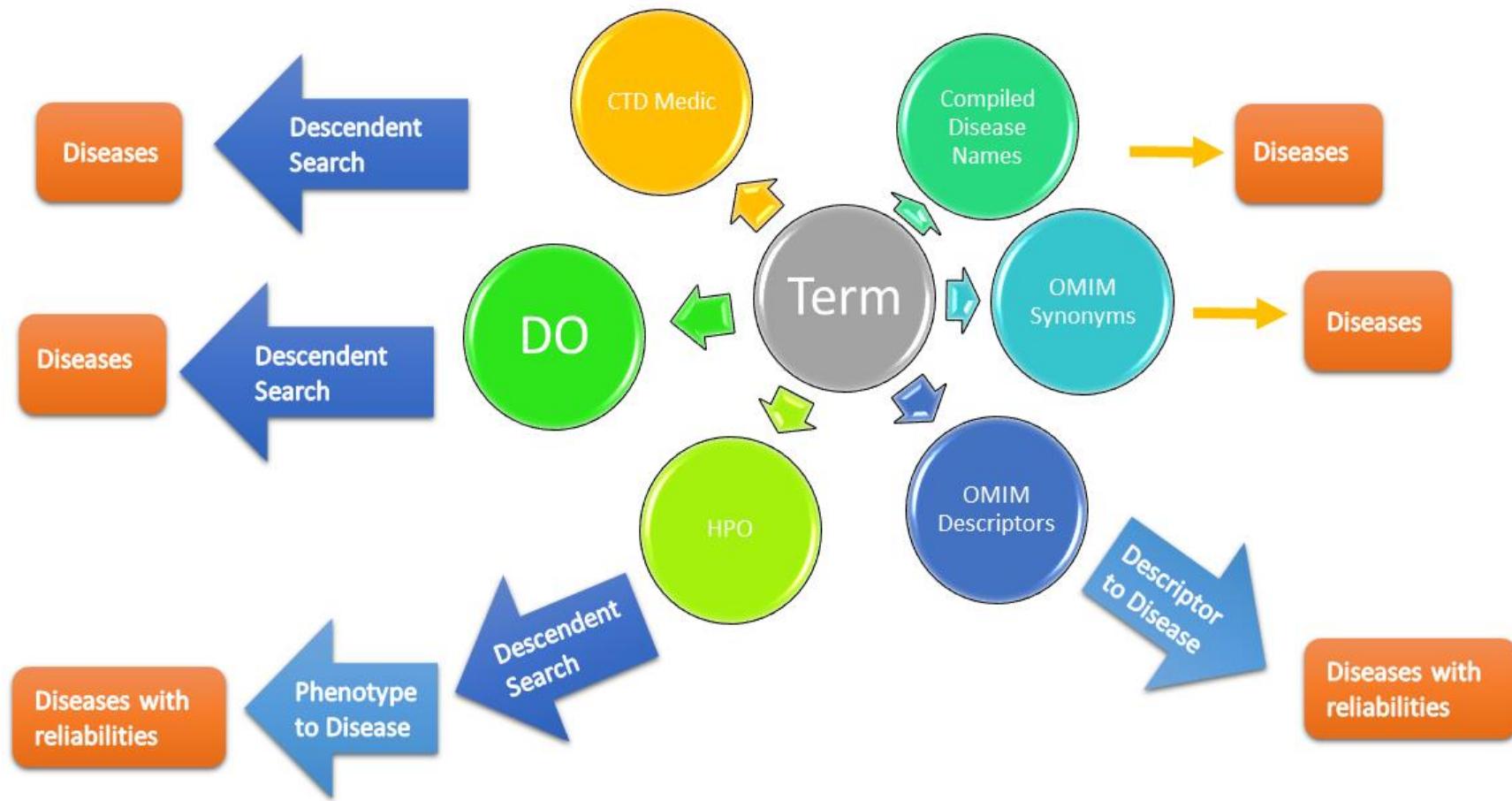
What is Phenolyzer

- 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



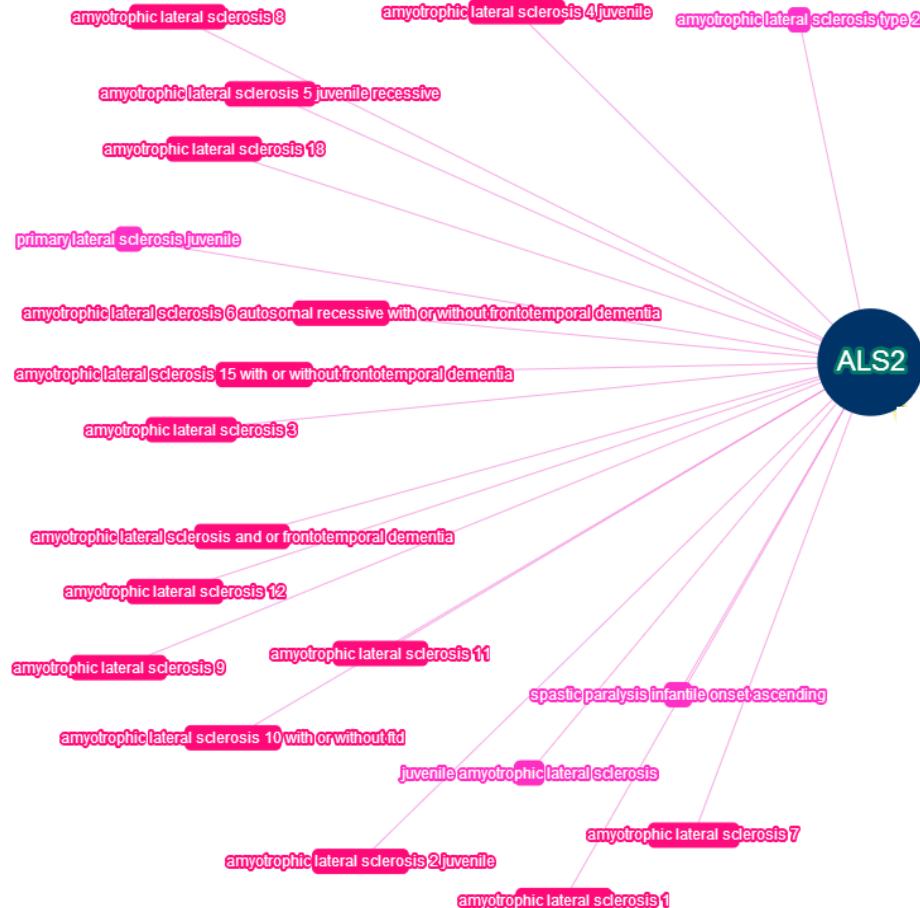
Step 1: Term interpretation



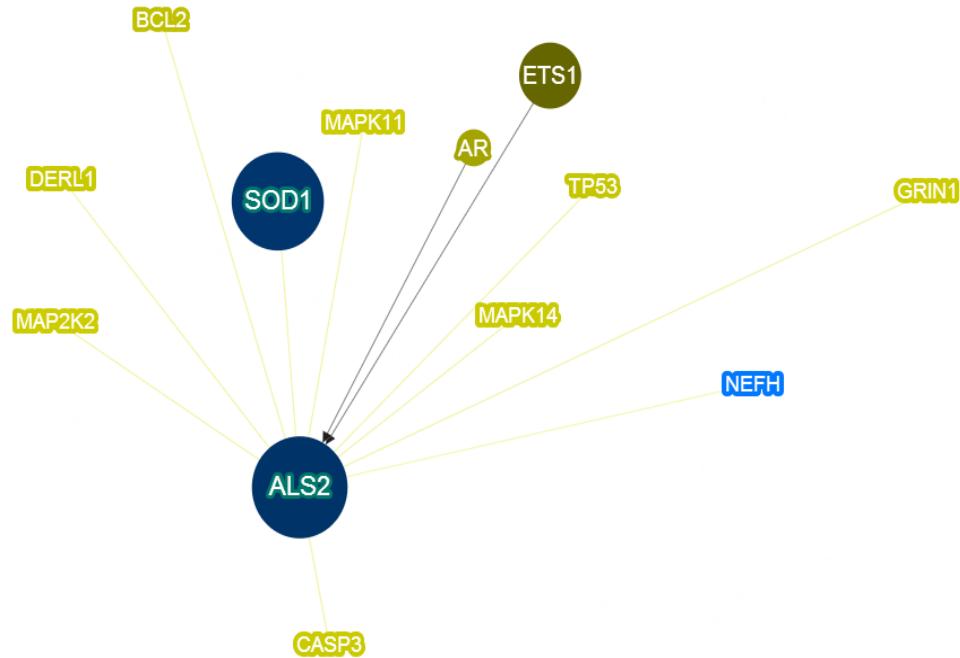
"Term" can be any phenotype term, such as "developmental delay", "hearing loss", etc

'Amyotrophic lateral sclerosis' interpreted

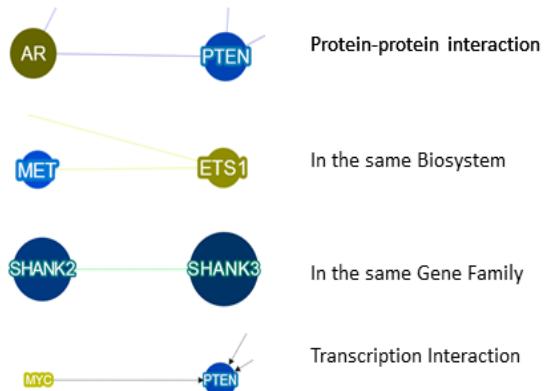
Step 2: Seed gene set generation



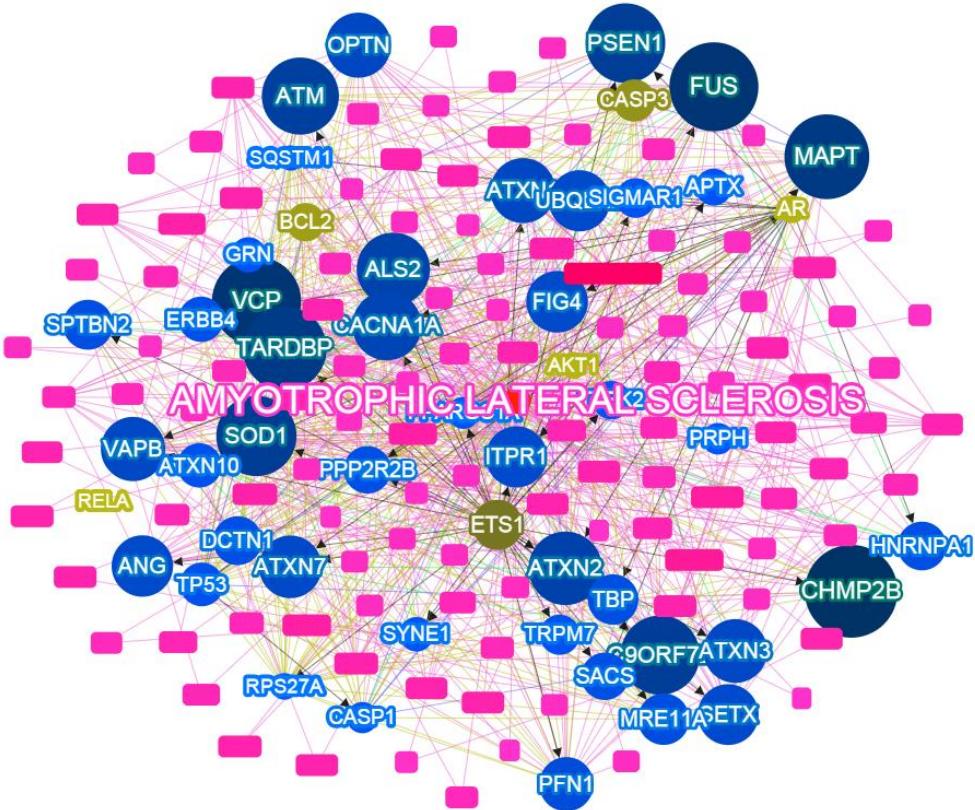
Step 3: Seed gene set growth



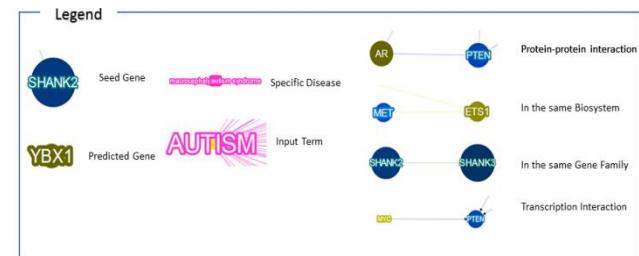
**ALS2 is grown based on four type of gene-gene relations,:
1. protein interaction
2. pathway
3. gene family
4. transcription interaction**



Step 4: Data integration and scoring



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

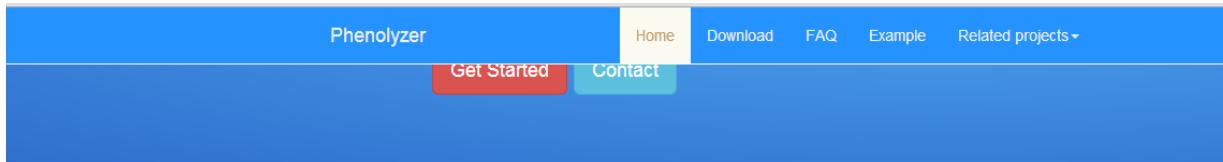


Phenolyzer generates a gene-phenotype-disease network for ALS

Web implementation

<http://phenolyzer.wglab.org>

1) Enter the website



Basic Information

Email

Diseases/Phenotypes please enter your focused disease/phenotype terms

Please use semicolon or enter as separators. Like "disease1;brain"
Try to use multiple terms instead of a super long term
OMIM IDs are also accepted, like 114480 for 'Breast cancer'

A red box highlights the input field for "Diseases/Phenotypes", and a red arrow points from it to the instruction "2)Enter the disease/phenotype terms, like 'Autism'". Another red arrow points from the "Submit" button to the instruction "3)Submit, done!".

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

3)Submit, done!

Options

Gene Selection

Region Selection

Advanced Options

Weight Adjust

Word Cloud

A red box highlights the "Word Cloud" dropdown menu, and a red arrow points from it to the instruction "Optional 4)Turn on Word Cloud".

Optional 4)Turn on Word Cloud

Click to see all the interpreted diseases

Click to see the Wordcloud of the term before.

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

Dear Phenolyzer user, your submission (identifier: 1935) was received at Wed Aug 20 22:23:48 2014 and processed at Wed Aug 20 22:23:48 2014.

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](#)

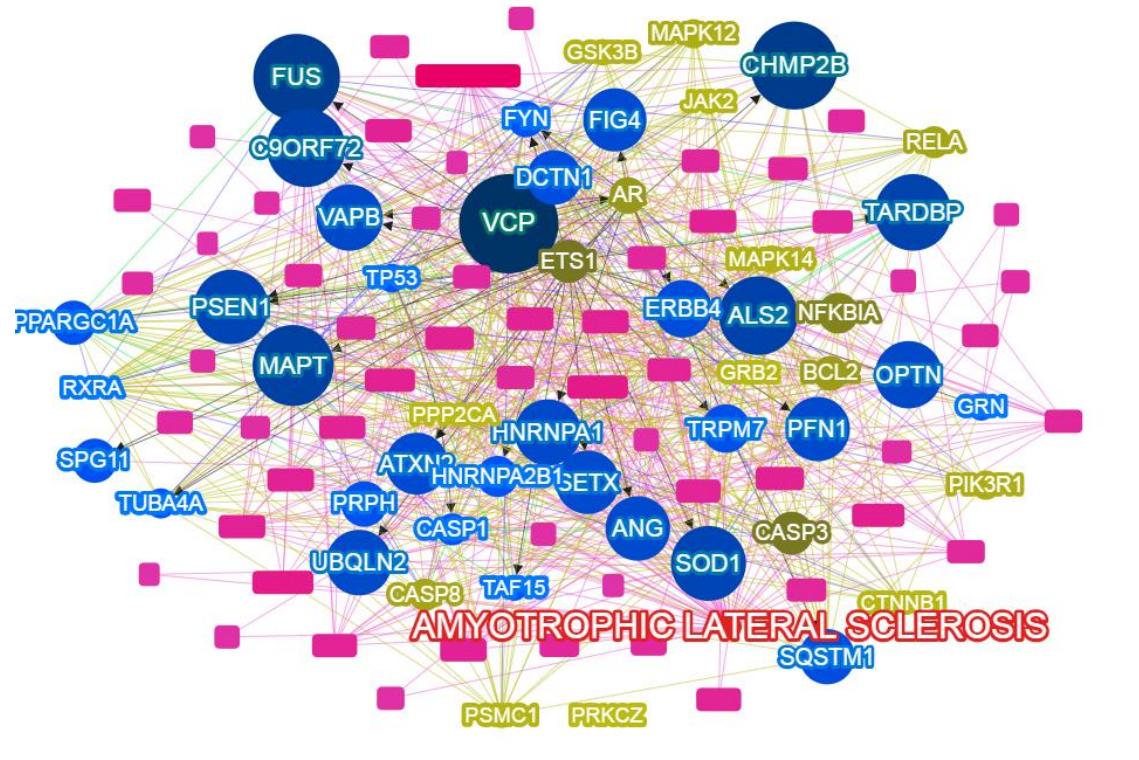
The whole report could be found [Here](#)

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](#)

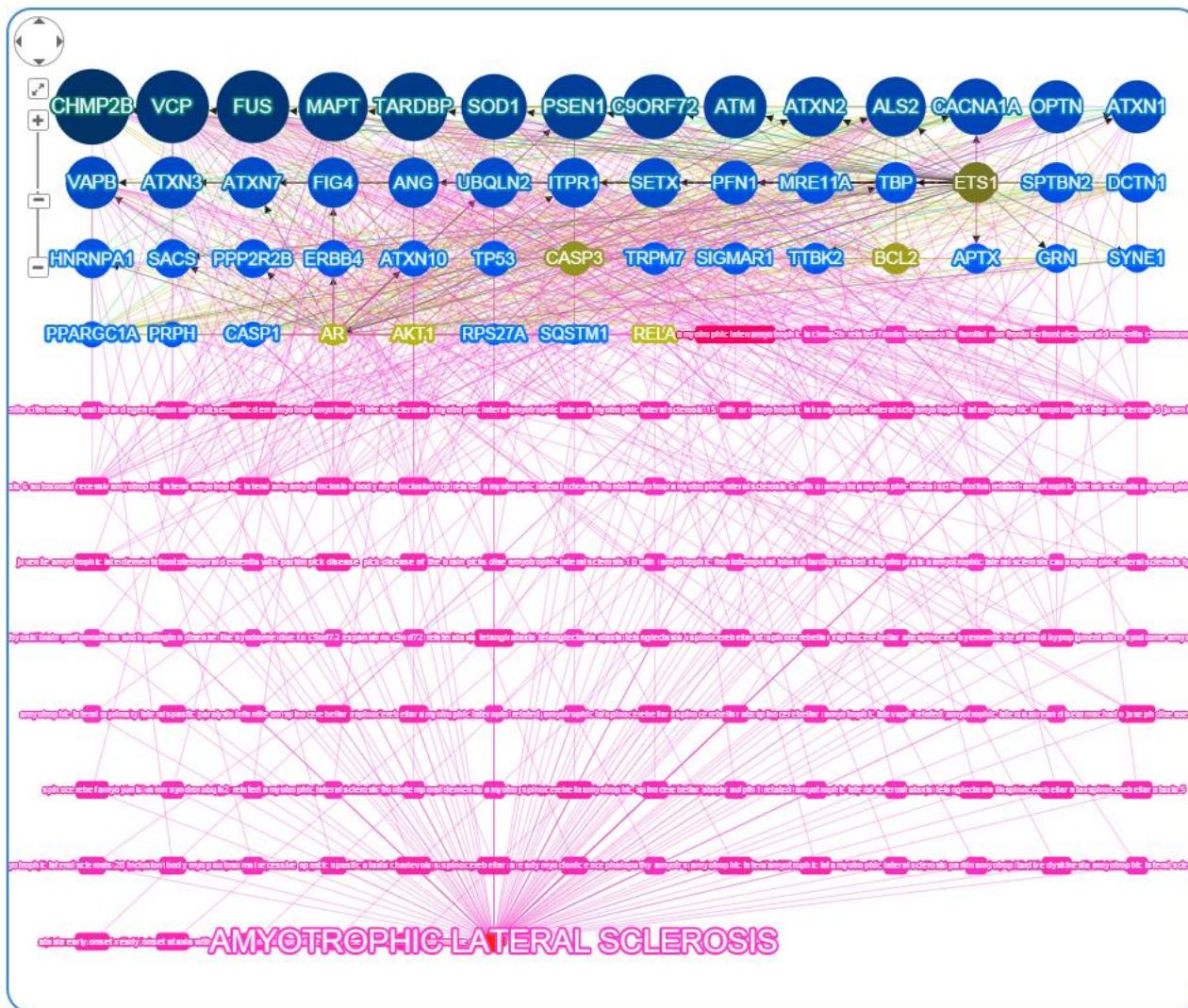
Example output for ALS



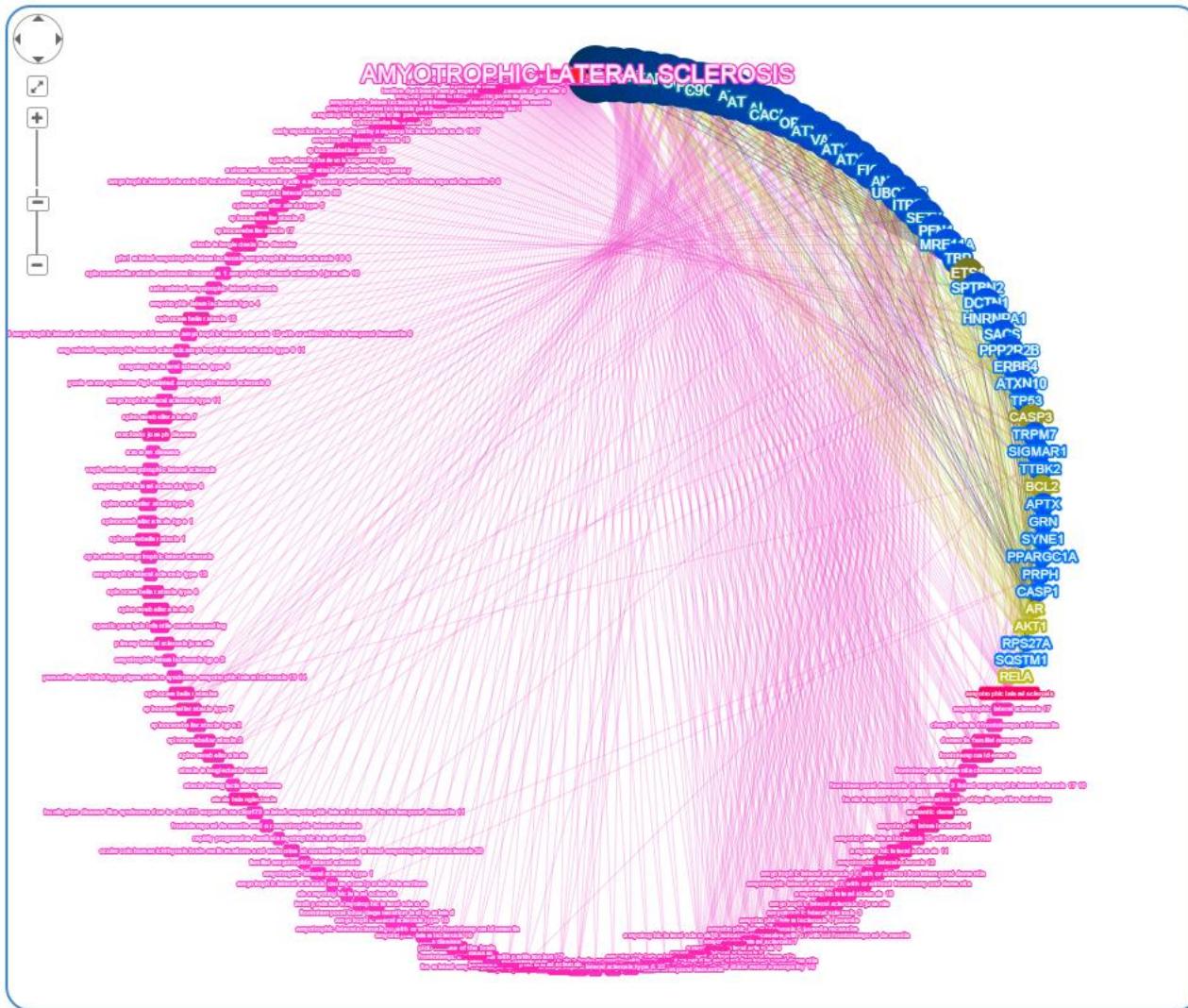
Legend

- Seed Gene:** SHANK2
- Predicted Gene:** YBX1
- Input Term:** AUTISM
- Specific Disease:** macrocephaly autism syndrome
- Protein-protein interaction:** AR — PTEN
- In the same Biosystem:** AKT1 — PRKACG
- In the same Gene Family:** SHANK2 — SHANK3
- Transcription Interaction:** MYC — PTEN

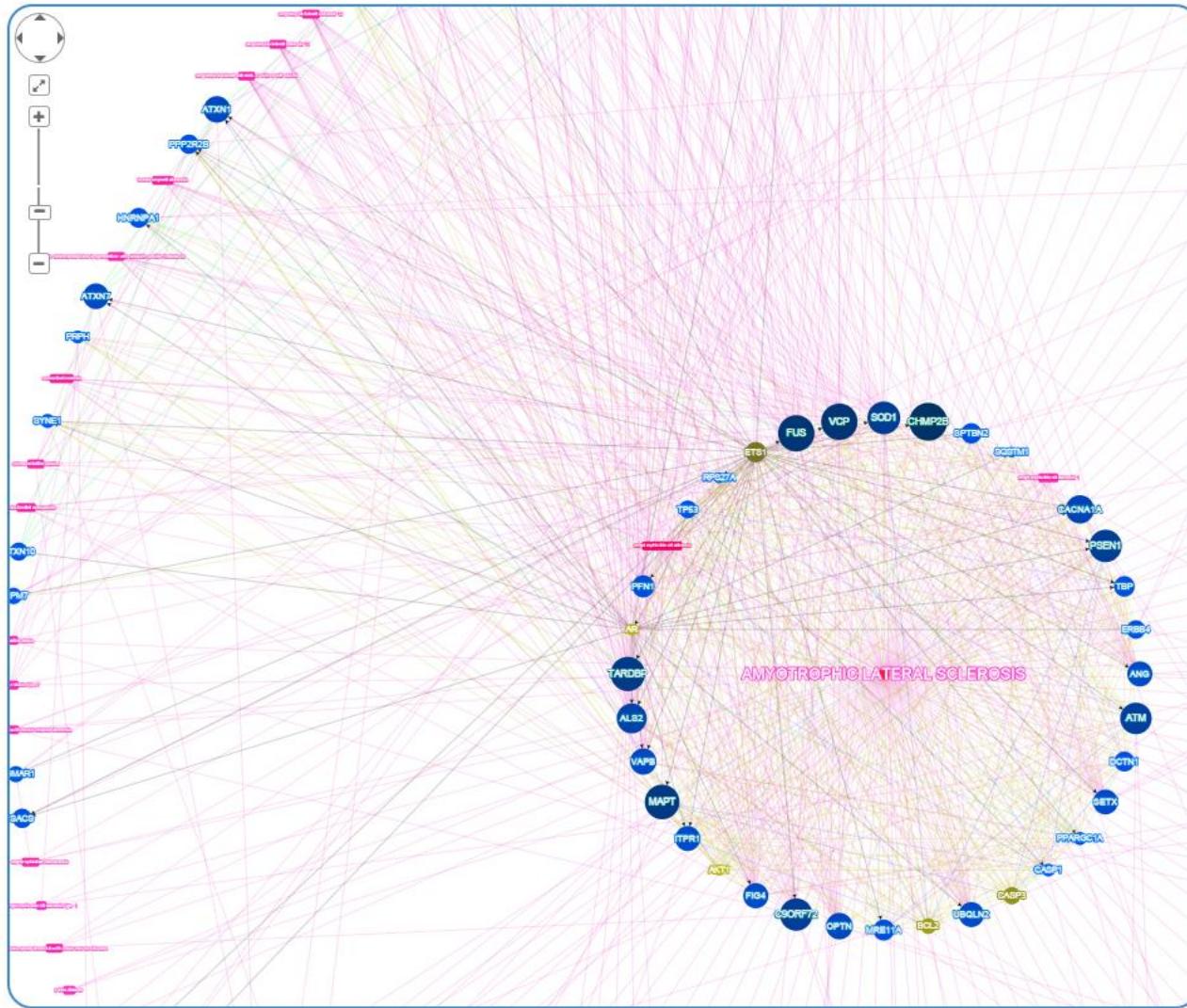
Different layouts for ‘top gene list’



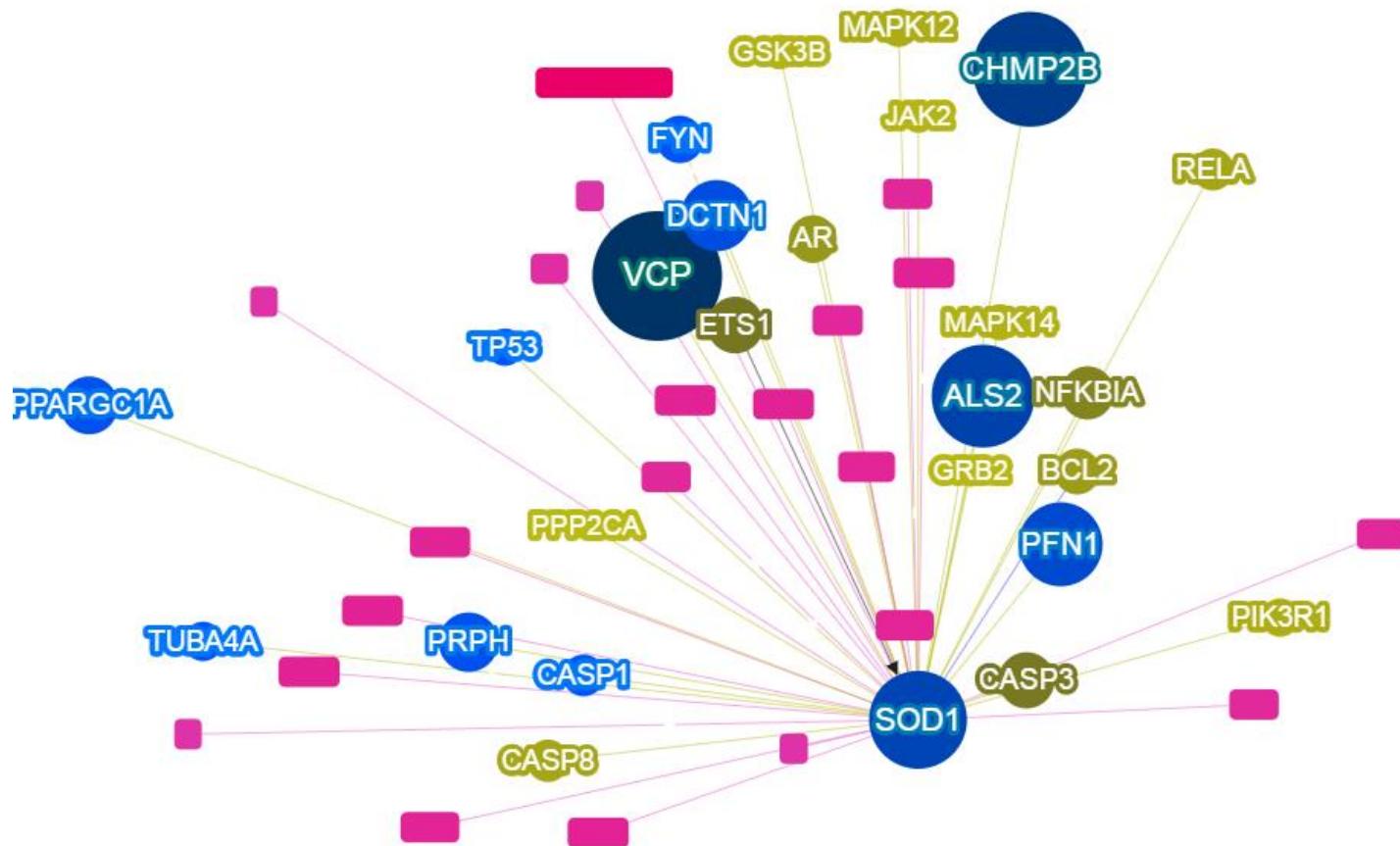
Different layouts for ‘top gene list’



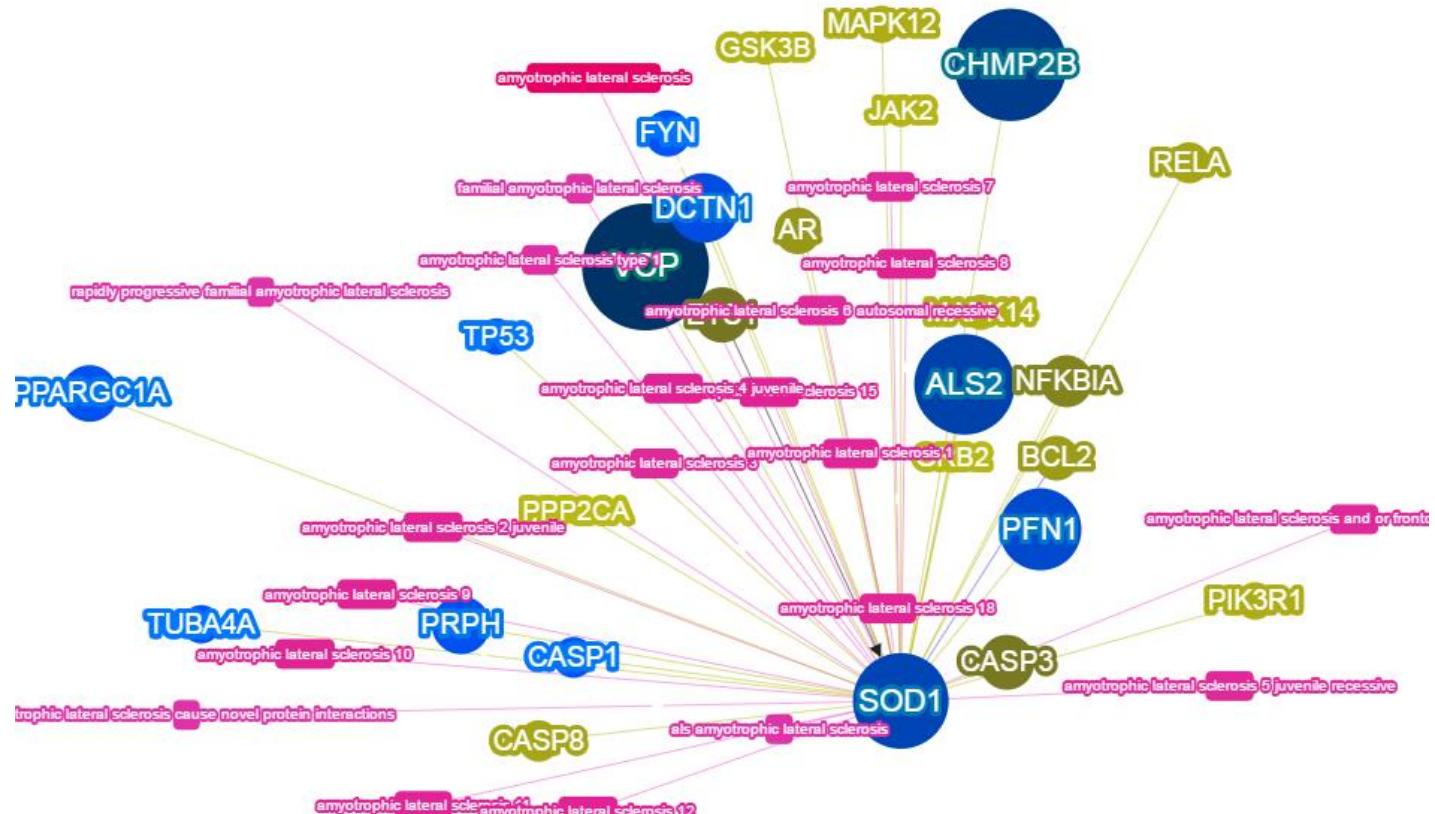
Different layouts for ‘top gene list’



Only connections for a specific gene



Turn on all labels



Disease

ON

Gene

ON

Gene Name

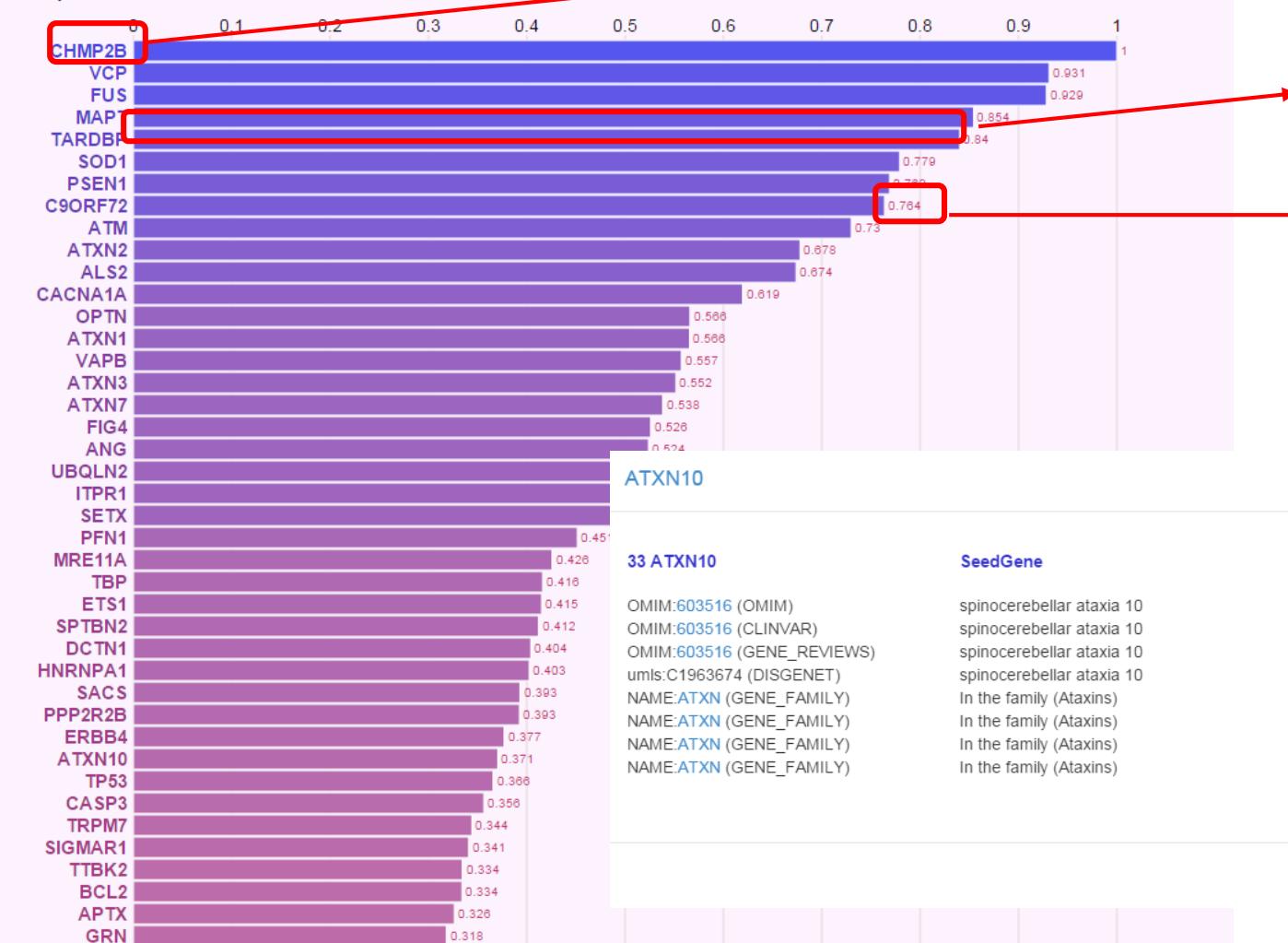
ON

Disease Name

ON

Summary Network Barplot Details

Barplot



Click to go to the NCBI website of this gene

Click to see the details about this gene

Score of this gene

ATXN10

33 ATXN10

SeedGene

Raw Score: 3.612

OMIM:603516 (OMIM)
OMIM:603516 (CLINVAR)
OMIM:603516 (GENE_REVIEWS)
umls:C1963674 (DISGENET)
NAME:ATXN (GENE_FAMILY)
NAME:ATXN (GENE_FAMILY)
NAME:ATXN (GENE_FAMILY)
NAME:ATXN (GENE_FAMILY)

spinocerebellar ataxia 10
spinocerebellar ataxia 10
spinocerebellar ataxia 10
spinocerebellar ataxia 10
In the family (Ataxins)
In the family (Ataxins)
In the family (Ataxins)
In the family (Ataxins)

amyotrophic lateral sclerosis (1.222)
amyotrophic lateral sclerosis (0.07637)
amyotrophic lateral sclerosis (1.222)
amyotrophic lateral sclerosis (0.3666)
With ATXN1 (0.177)
With ATXN3 (0.1571)
With ATXN7 (0.1711)
With ATXN2 (0.2201)

Close

Detailed examination of the evidence

Summary Network Barplot Details

Start Previous 1 Go Next End

▶ 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

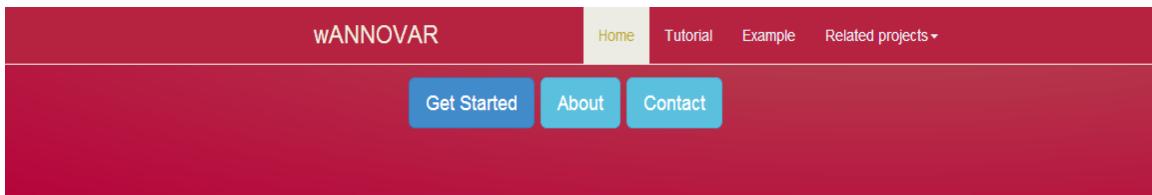
TARDBP

ORPHANET:803 (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:105400 (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:612069 (OMIM)	amyotrophic lateral sclerosis 10 with or without ftd	amyotrophic lateral sclerosis (1.222)
OMIM:612069 (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:612577 (GENE_REVIEWS)	amyotrophic lateral sclerosis 11	amyotrophic lateral sclerosis (0.08146)
OMIM:613435 (GENE_REVIEWS)	amyotrophic lateral sclerosis 12	amyotrophic lateral sclerosis (0.08146)
OMIM:300857 (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!

Basic Information

Email

Sample Identifier

Input File

or Paste Variant Calls

+ Input File

paste your variant call here

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

3) Enter disease/phenotype terms here

Disease/Phenotype (Optional)

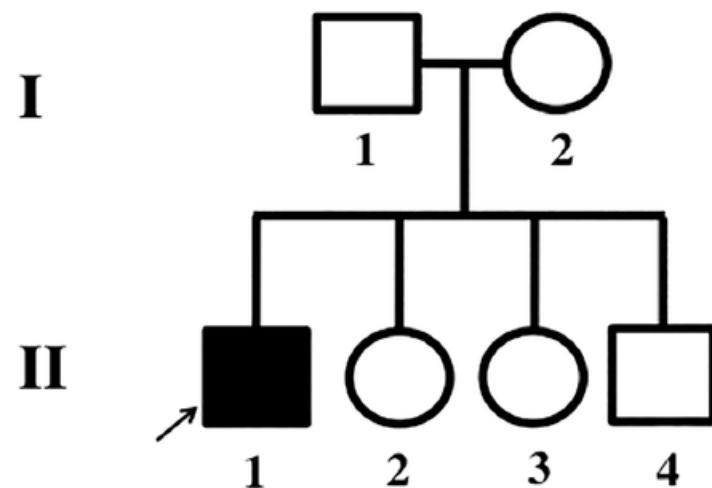
Enter Disease or Phenotype Terms

please enter your focused disease/phenotype 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'

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

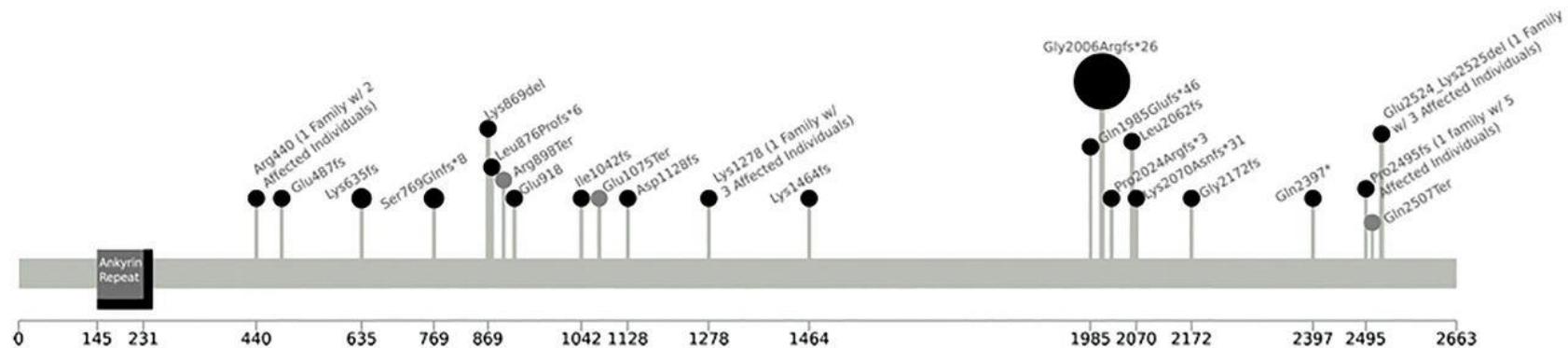
- Sequencing performed on Ion Proton with AmpliSeq Exome panel
- With standard analytical protocol, more than 1000 variants were recognized as *de novo*, well above the expected number of *de novo* mutations, suggesting low quality of sequencing and variant calling

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

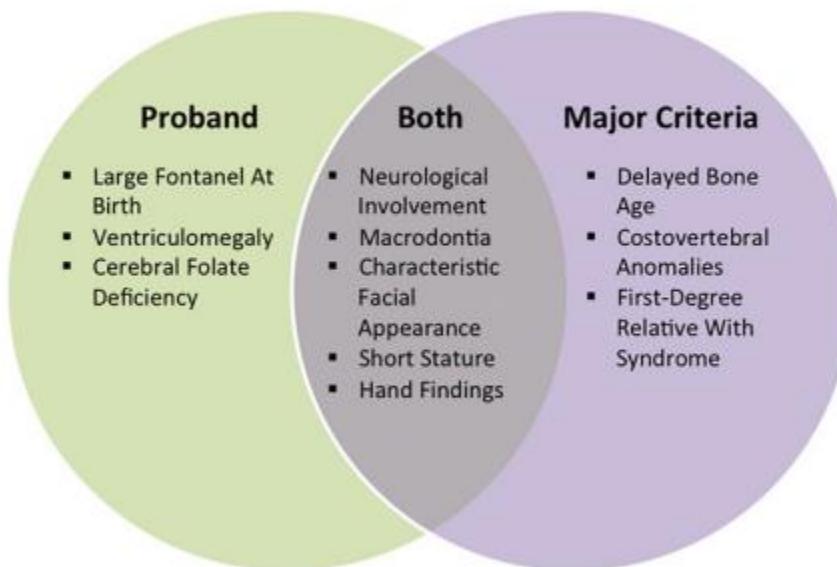
Phenolyzer+wANNOVAR joint analysis

- 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

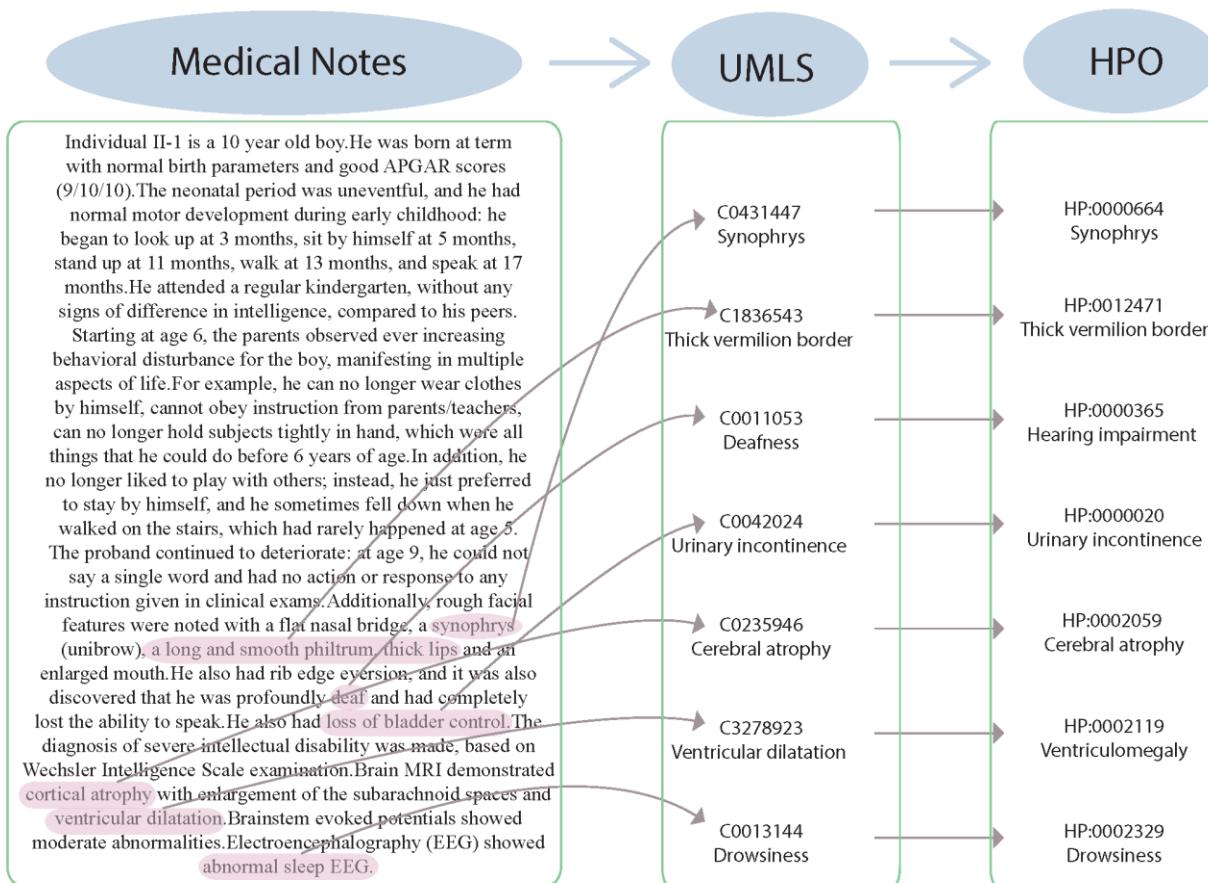
- KBG syndrome is a rare autosomal dominant genetic condition characterized by neurological involvement and distinct facial, hand, and skeletal features.
- 70 cases have been reported
- Highly heterogeneous phenotypic features



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

Computational generation of HPO terms?

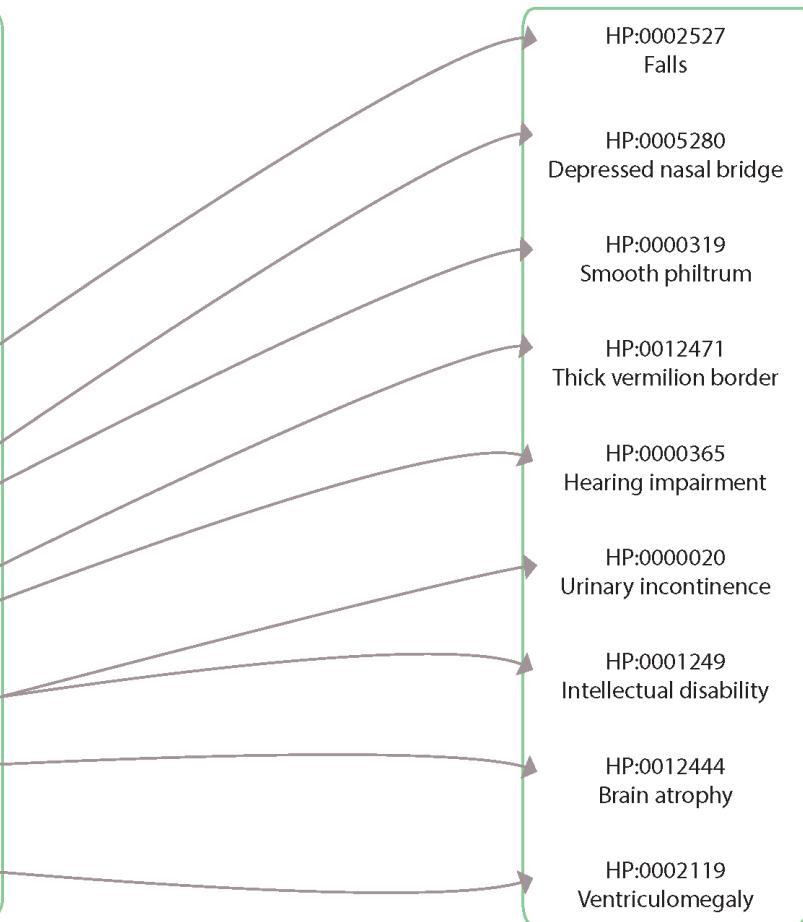
MetaMap



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

