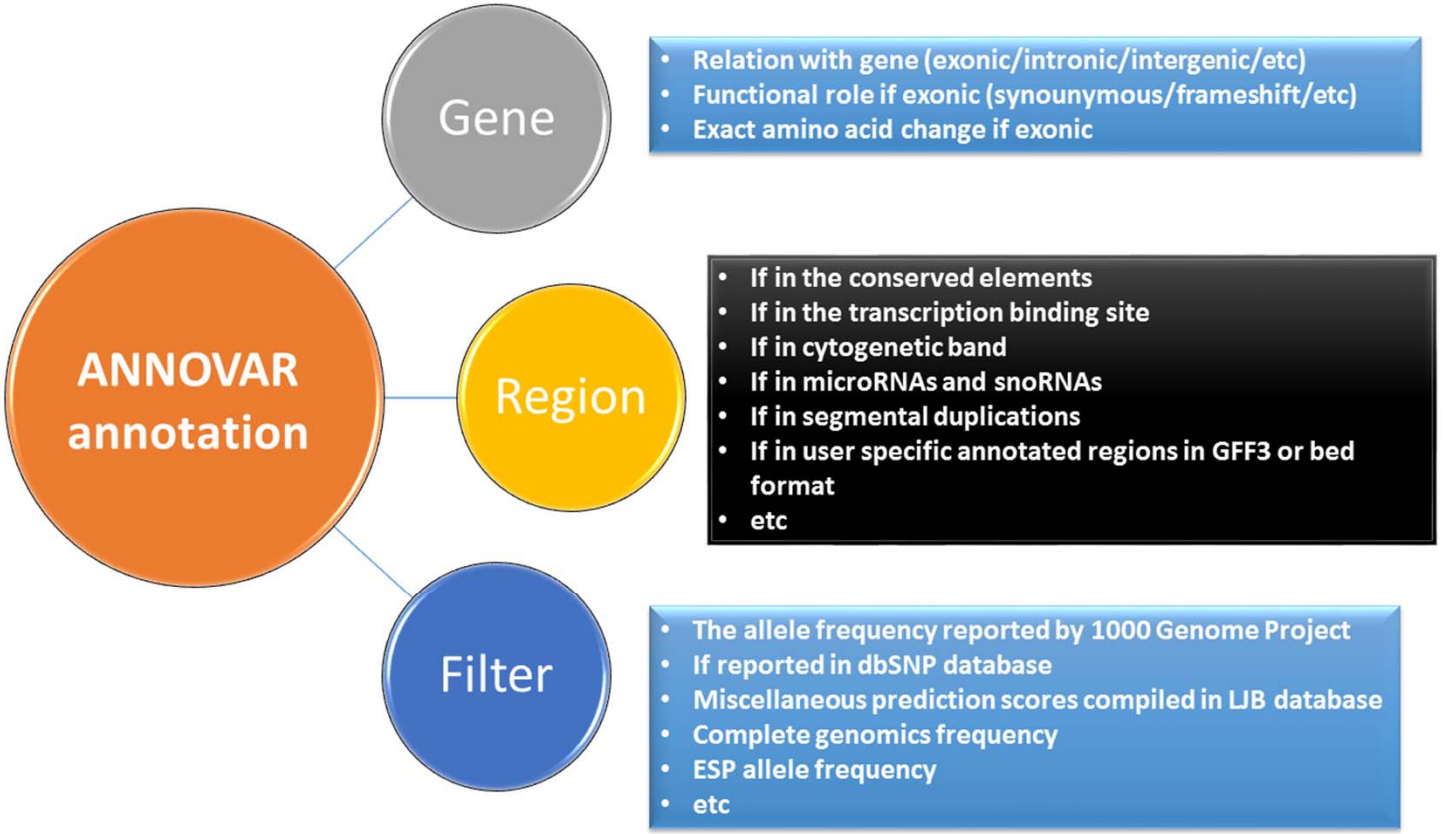
Annotation and phenotype-driven interpretation of genetic variants

2019 Dragon Star Bioinformatics Course (Day 3)

ANNOVAR for variant annotation



ANNOVAR/wANNOVAR

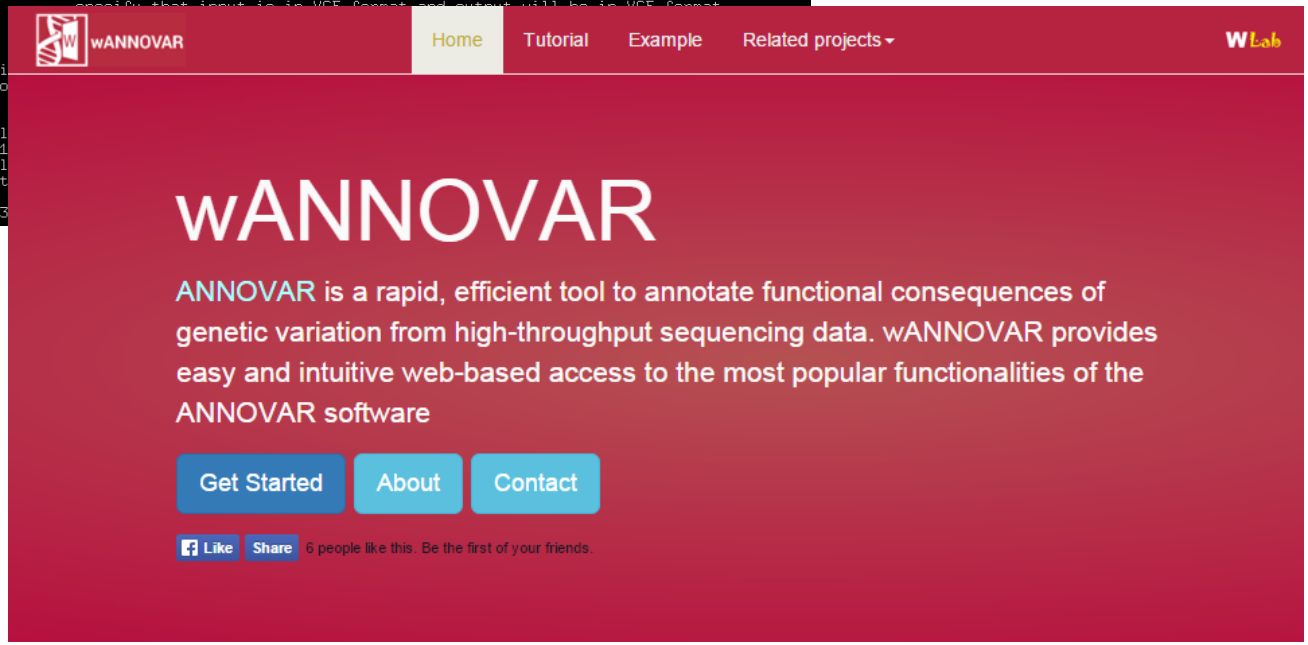
```
[kaiwang@biocluster ~]$ table_annovar.pl
Usage:
     table_annovar.pl [arguments] <query-file> <database-location>

     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 (infomration 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
                --dot2underline


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

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

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

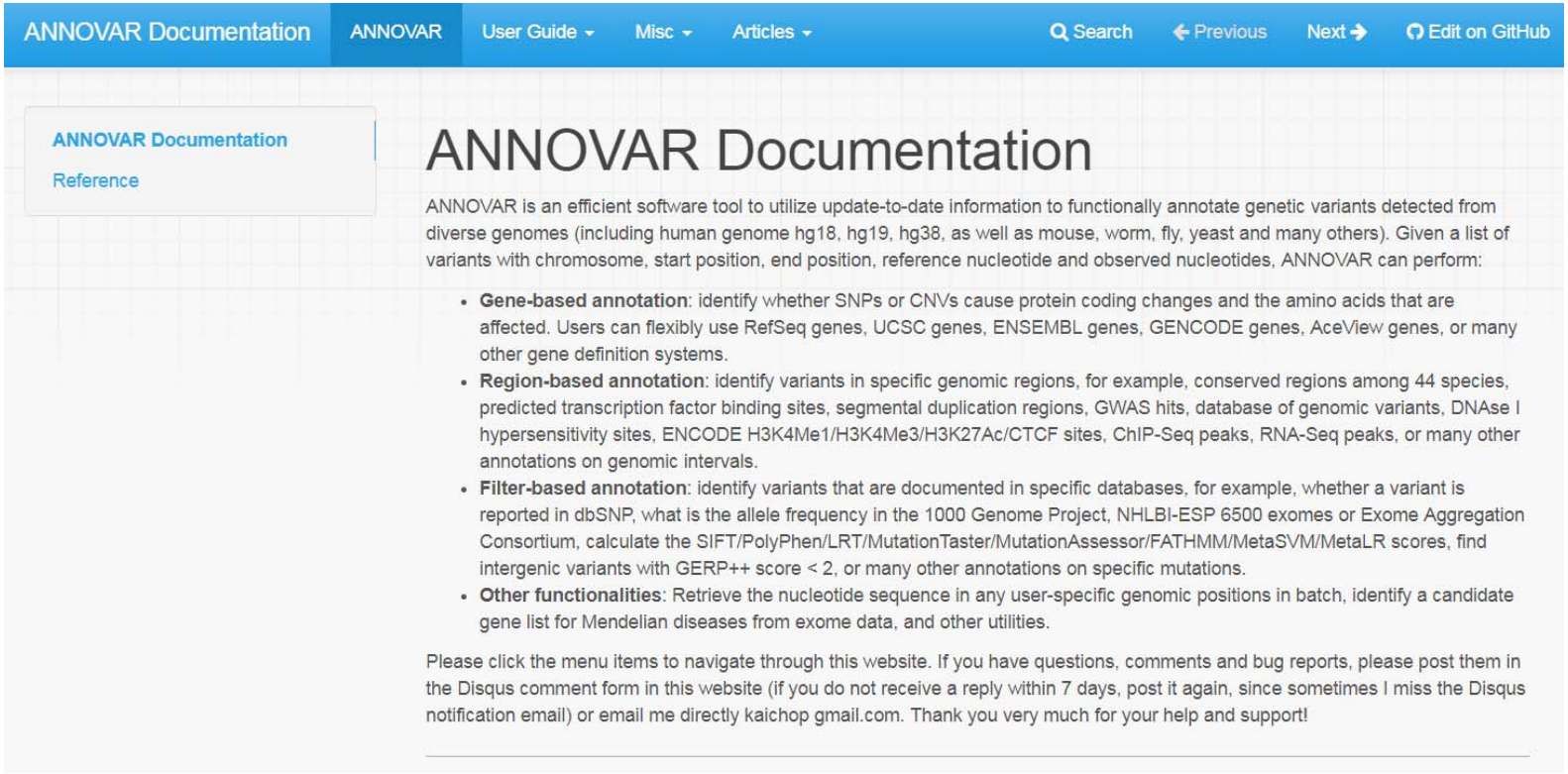


Yang H, Wang K. *Nature Protocols*, 10:1556-1566, 2015

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



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,avsnp147,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	ExonicFu	AAChang	cytoBand	genomicS	esp6500	1000g20	1000g20	1000g20	1000g20	snp138	SIFT_sco	SIFT_pre
2	1	948921	948921	T	C	UTR5	ISG15	NM_005	.	.	1p36.33	.	0.8858	0.9032	0.7201	0.999	0.9573	rs15842	.	.
3	1	1E+06	1E+06	G	T	UTR3	ATAD3C	NM_001(	.	.	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_001:	.	.	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	rs100005	.	.
6	1	8E+07	8E+07	C	T	intronic	DNASE2I	.	.	.	1p31.1	.	.	0.5489	0.5303	0.6062	0.5239	rs657670	.	.
7	1	1E+07	1E+07	TC	-	intergeni	LOC645:	dist=1156	.	.	1p36.21	Score=0.	.	.	.	.	.	.	.	.
8	1	1E+07	1E+07	-	AT	intergeni	UBIAD1,F	dist=551(	.	.	1p36.22	.	.	.	.	.	.	rs355611	.	.
9	1	1E+08	1E+08	A	ATAAA	intergeni	LOC1001	dist=872	.	.	1p21.1	.	.	.	.	.	.	.	.	.
10	1	7E+07	7E+07	G	A	exonic	IL23R	.	nonsynor	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	.	nonsynor	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	.	nonsynor	NOD2:NM	16q12.1	.	0.0316	0.0144	0.0023	.	0.0507	rs20668	0.01	D
13	16	5E+07	5E+07	G	C	exonic	NOD2	.	nonsynor	NOD2:NM	16q12.1	.	0.0102	0.0046	.	.	0.0099	rs20668	0.02	D
14	16	5E+07	5E+07	-	C	exonic	NOD2	.	frameshif	NOD2:NM	16q12.1	.	0.0162	0.006	0.0038	.	0.0139	rs20668	.	.

U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO
Polypher	Polypher	Polypher	Polypher	LRT_sco	LRT_pre	Mutation	Mutation	Mutation	Mutation	FATHMM	FATHMM	RadialSV	RadialSV	LR_score	LR_pred	VEST3_s	CADD_ra	CADD_ph	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	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
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

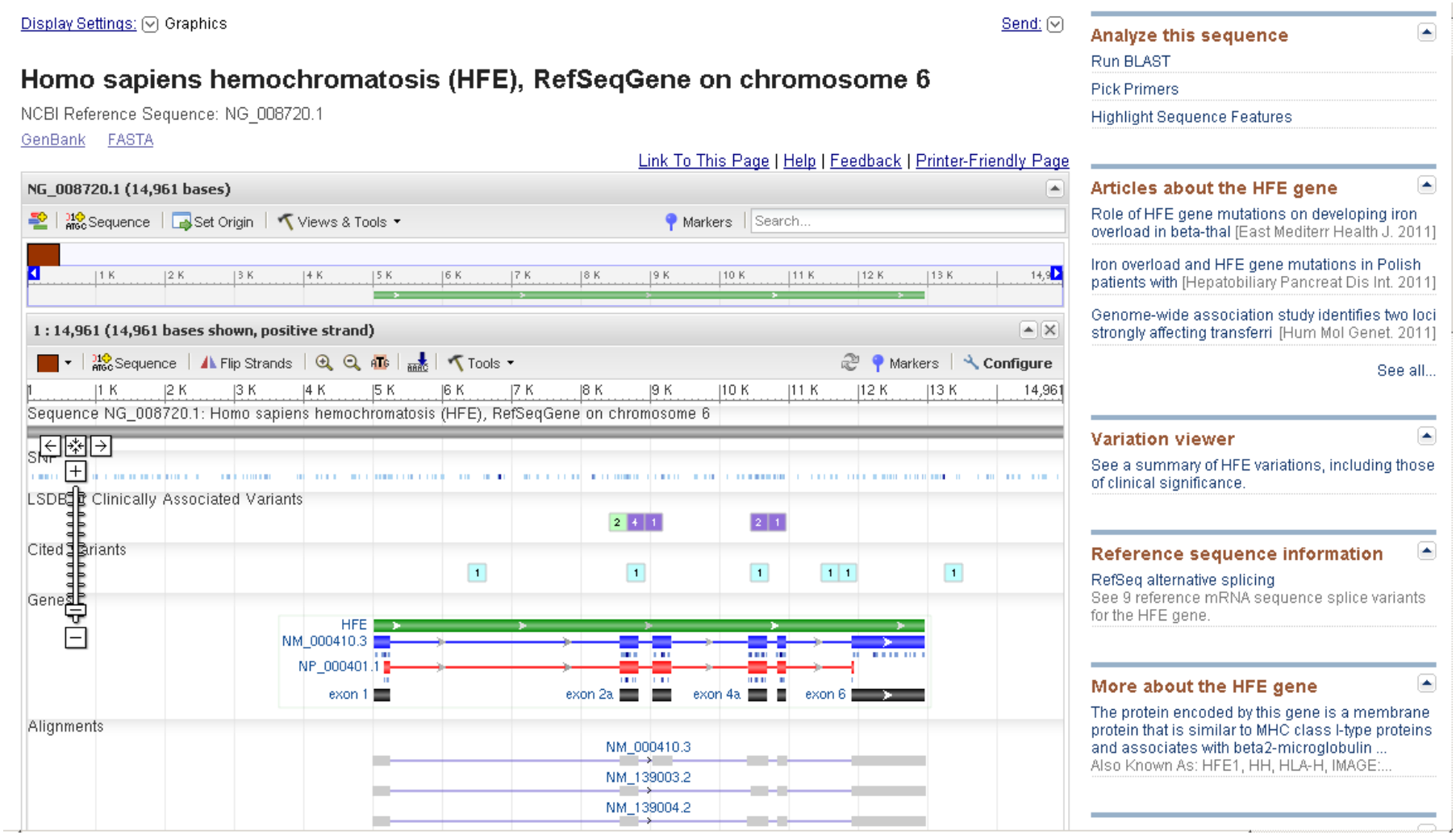
New user: what are r,g,f?

- 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?



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.

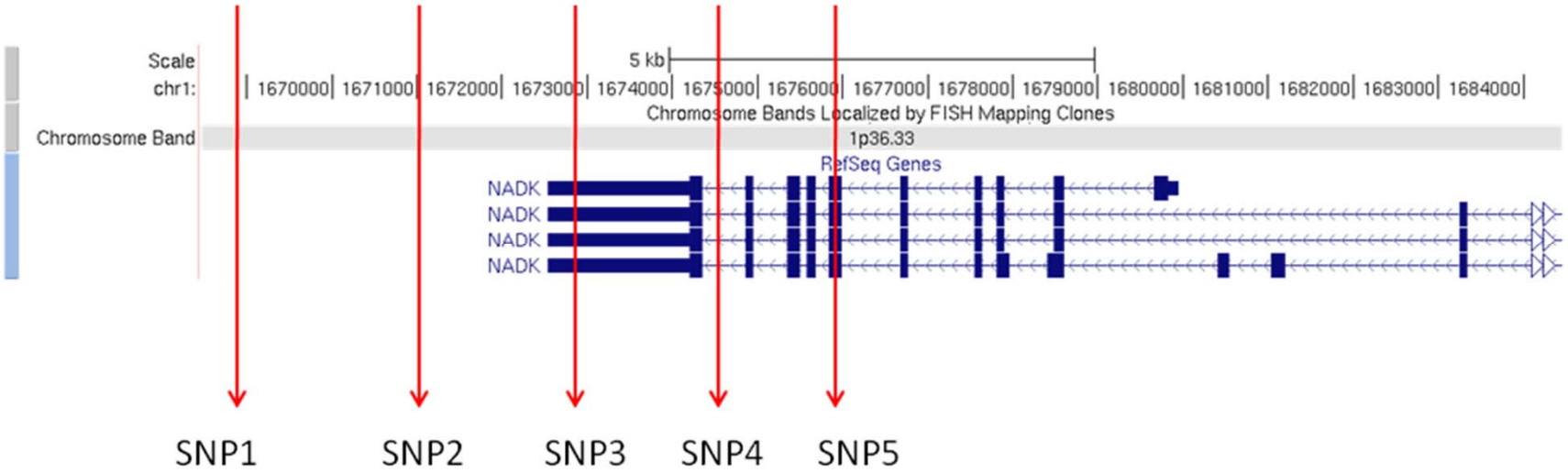
There are ~50K, ~80K, ~200K, ~190K transcripts in the four gene models, respectively

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 downtream 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)

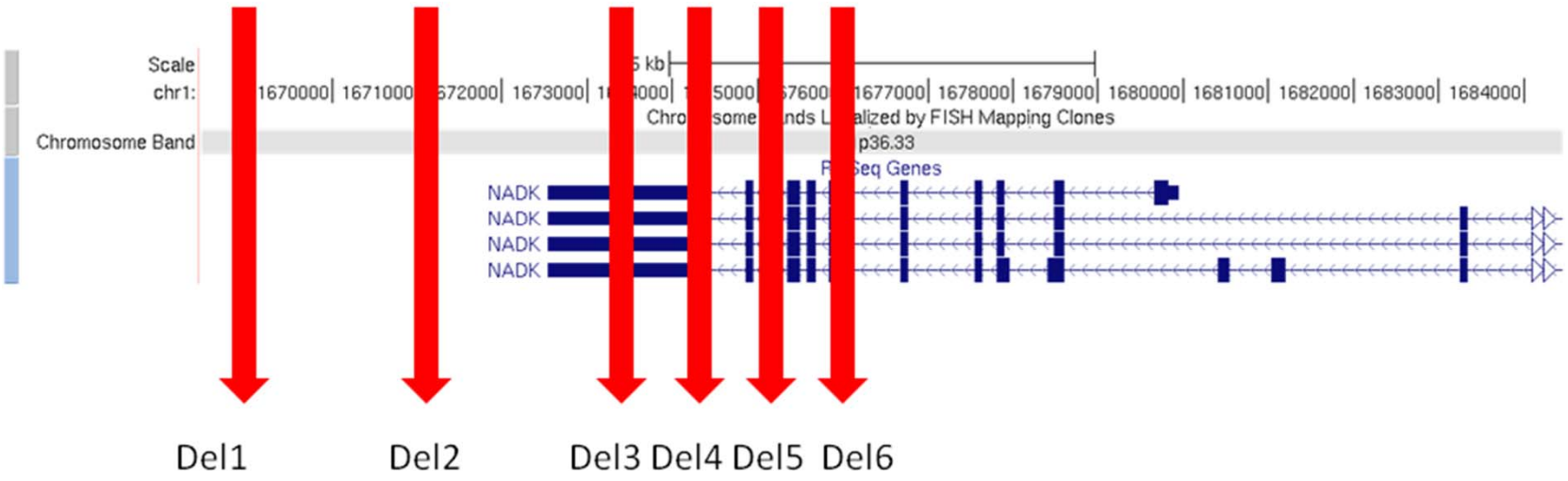
Example: SNVs

- 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)

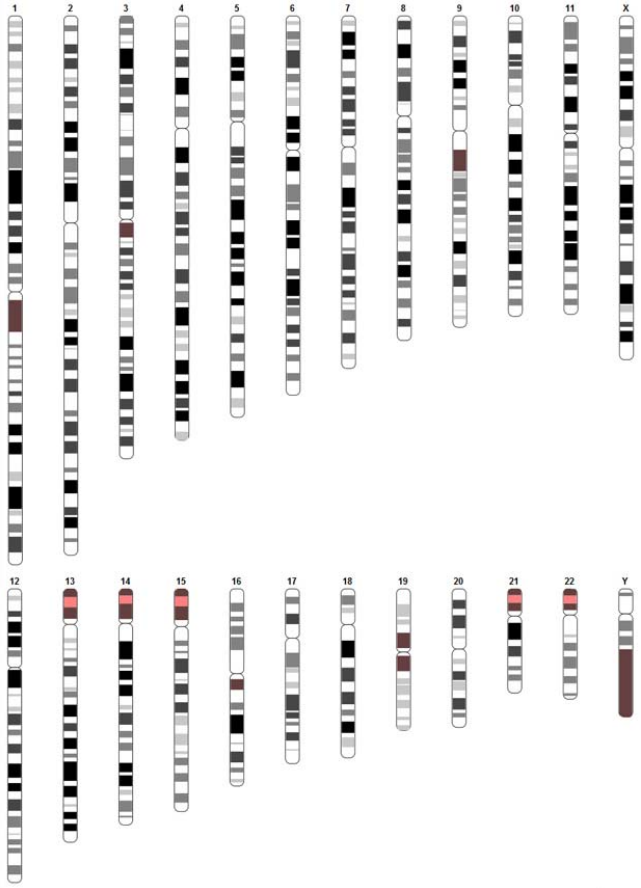
Region-based annotation

- 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.

	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	0	-	13q12.11



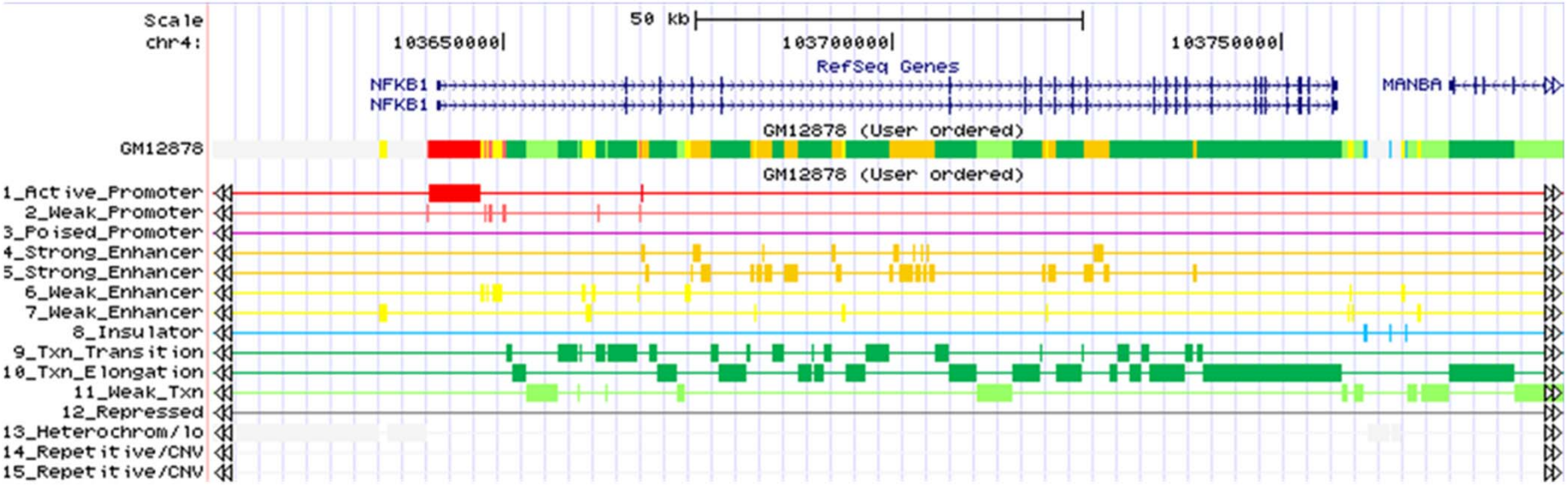
ENCODE ChIP-Seq peaks

- 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 Caucasion, 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

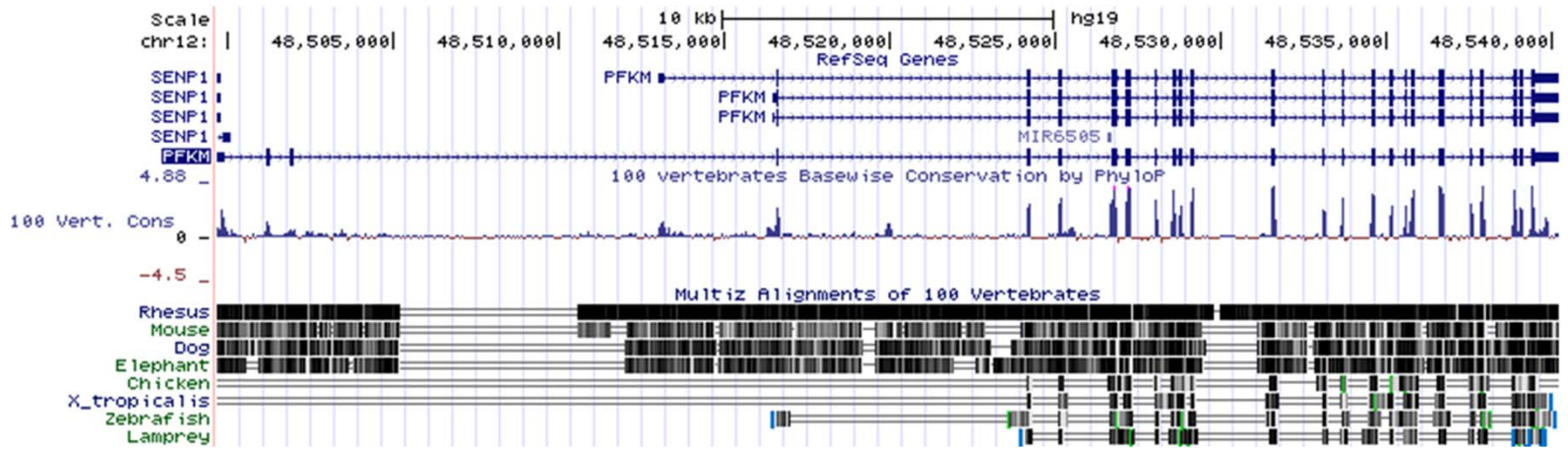
chromHMM predictions

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



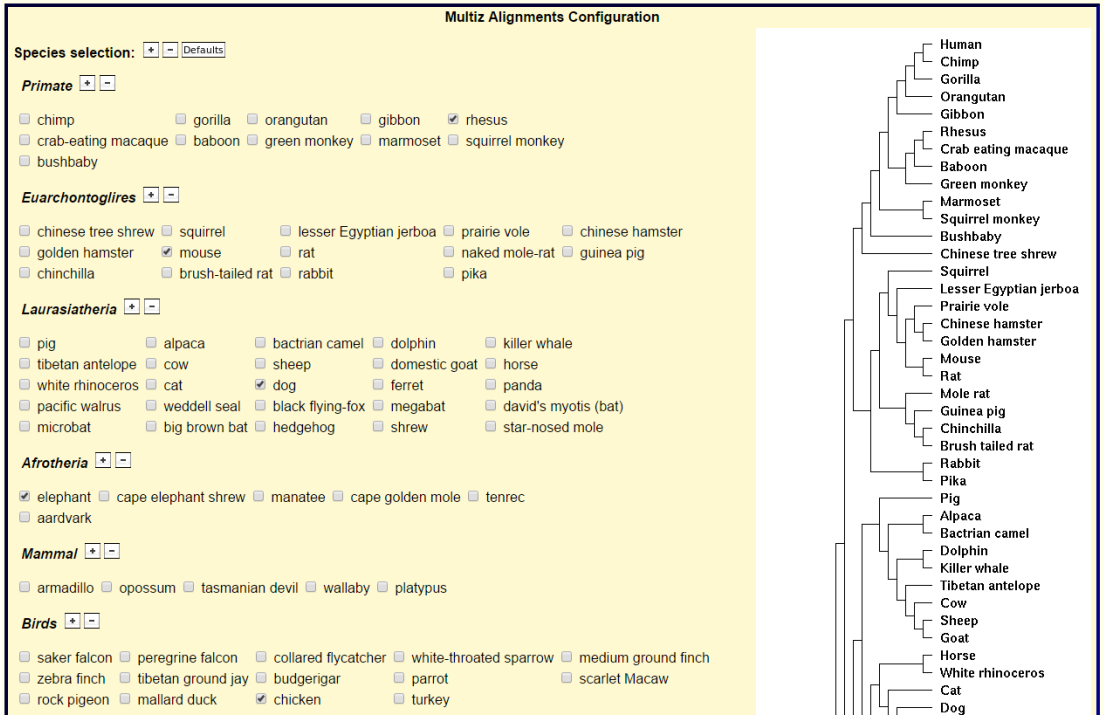
Conserved genomic regions

- 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



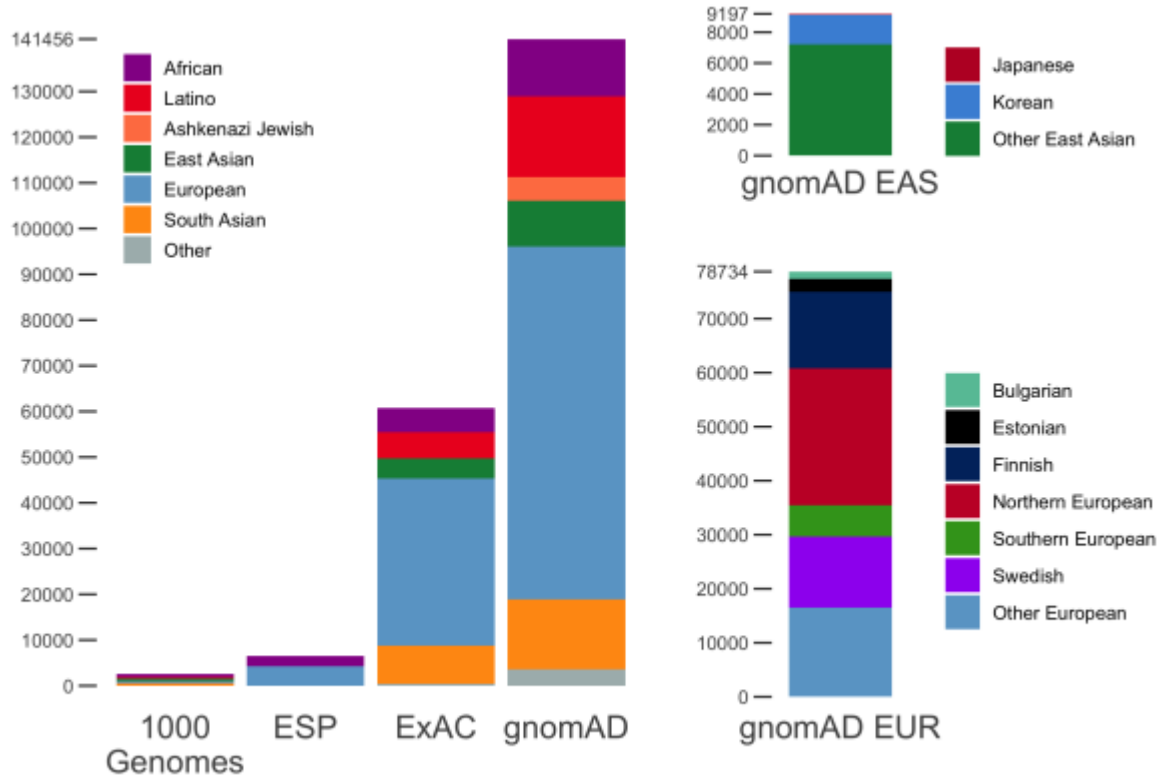
Filter-based annotation

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

Allele frequency databases

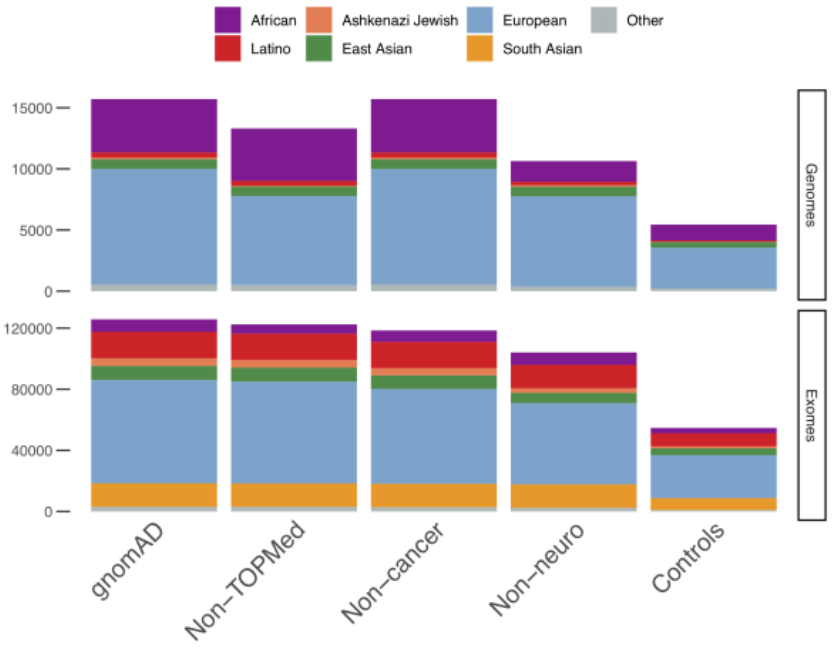
- 1000 Genomes Project: ~2500 genomes
  - Keyword is 1000g2015aug_all
- ExAC (exome aggregation consortium): ~65000 exomes
  - Keyword is exac03
- NHLBI-ESP6500 (European Americans and African Americans): ~6500 exomes
  - Keyword is esp6500siv2_aa, esp6500siv2_ea
- gnomAD (genome aggregation consortium): 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
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			0:63:6
13	Y	2788927	rs56004558	A	G	38	.	AC=24;AN=57;DB;DP=17			0:54:3
14	Y	2794854	rs35284970	C	T	38	.	AC=2;AN=62;DB;DP=181			0:57:4
15	Y	2798273	.	G	A	35	.	AC=13;AN=62;DP=189;N			0:28:3
16	Y	2799468	rs9786209	C	T	36	.	AC=18;AN=62;DB;DP=24			0:54:3
17	Y	2812887	.	G	T	35	.	AC=4;AN=52;DP=189;NS			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

Allele count = 29
Total number of alleles = 62
dbSNP membership
Combined depth across samples = 254
hapmap2 membership
Number of samples = 66
NR

	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.refGe	cytoBand	genomicSu
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=0.9(
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=0.9
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_022	16q12.1	.
13	16	50756540	50756540	G	C	exonic	NOD2	.	nonsynonymous SNV	NOD2:NM_022	16q12.1	.
14	16	50763778	50763778	-	C	exonic	NOD2	.	frameshift insertion	NOD2:NM_022	16q12.1	.
15	13	20763686	20763686	G	-	exonic	GJB2	.	frameshift deletion	GJB2:NM_0040(	13q12.11	.
16	13	20797176	21105944	0	-	exonic	CRYL1,GJB6		frameshift deletion	GJB6:NM_0011	13q12.11	

Using web version of ANNOVAR

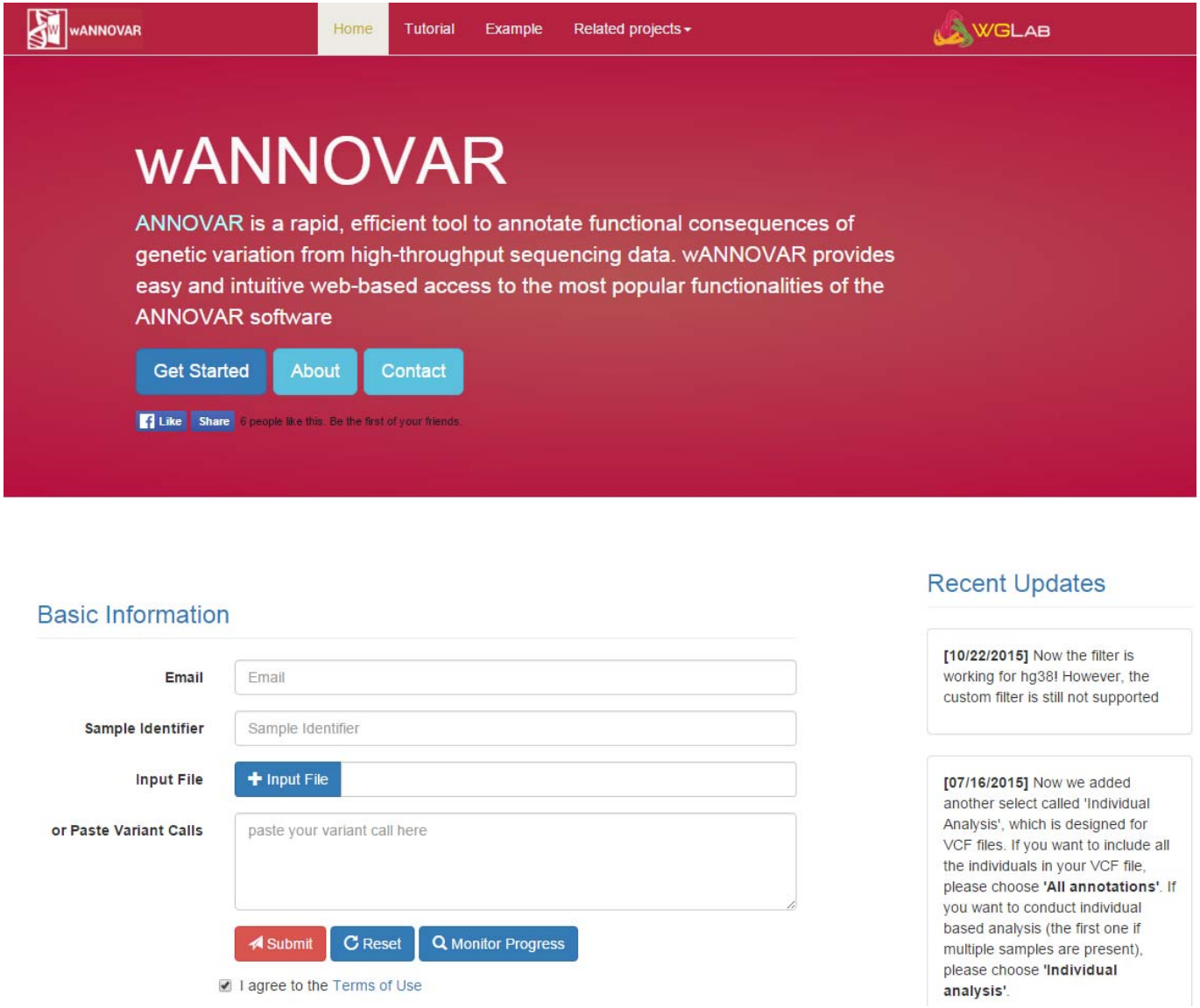
- Main advantage of wANNOVAR
  - No need to run command line
- Main limitations
  - Only a limited number of annotations are available
  - Many annotation databases are outdated, compared the latest ones that are available in ANNOVAR

wANNOVAR

- URL:
  - Stable version: http://wannovar.wglab.org

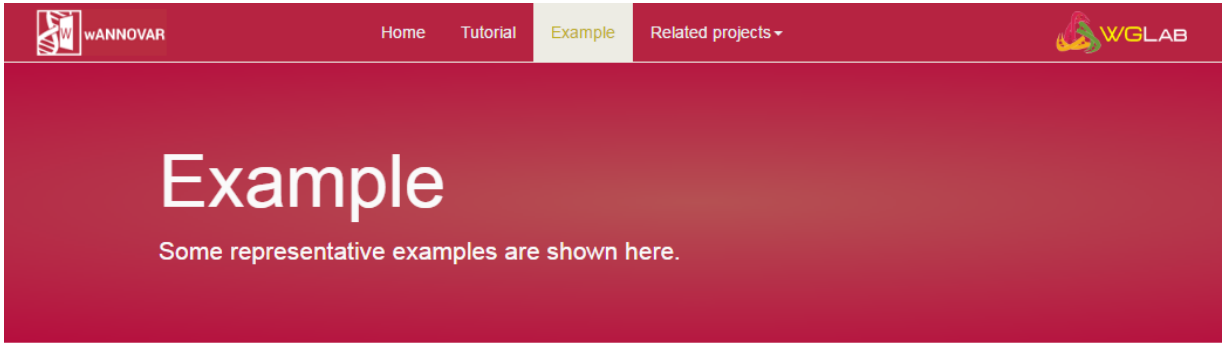
- Function:
  - Users provide VCF file, returns annotated output in CSV format or tab-delimited format
  - Perform variants reduction to find disease genes
  - Display annotated output in web interface

Main interface



Demo

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



Example

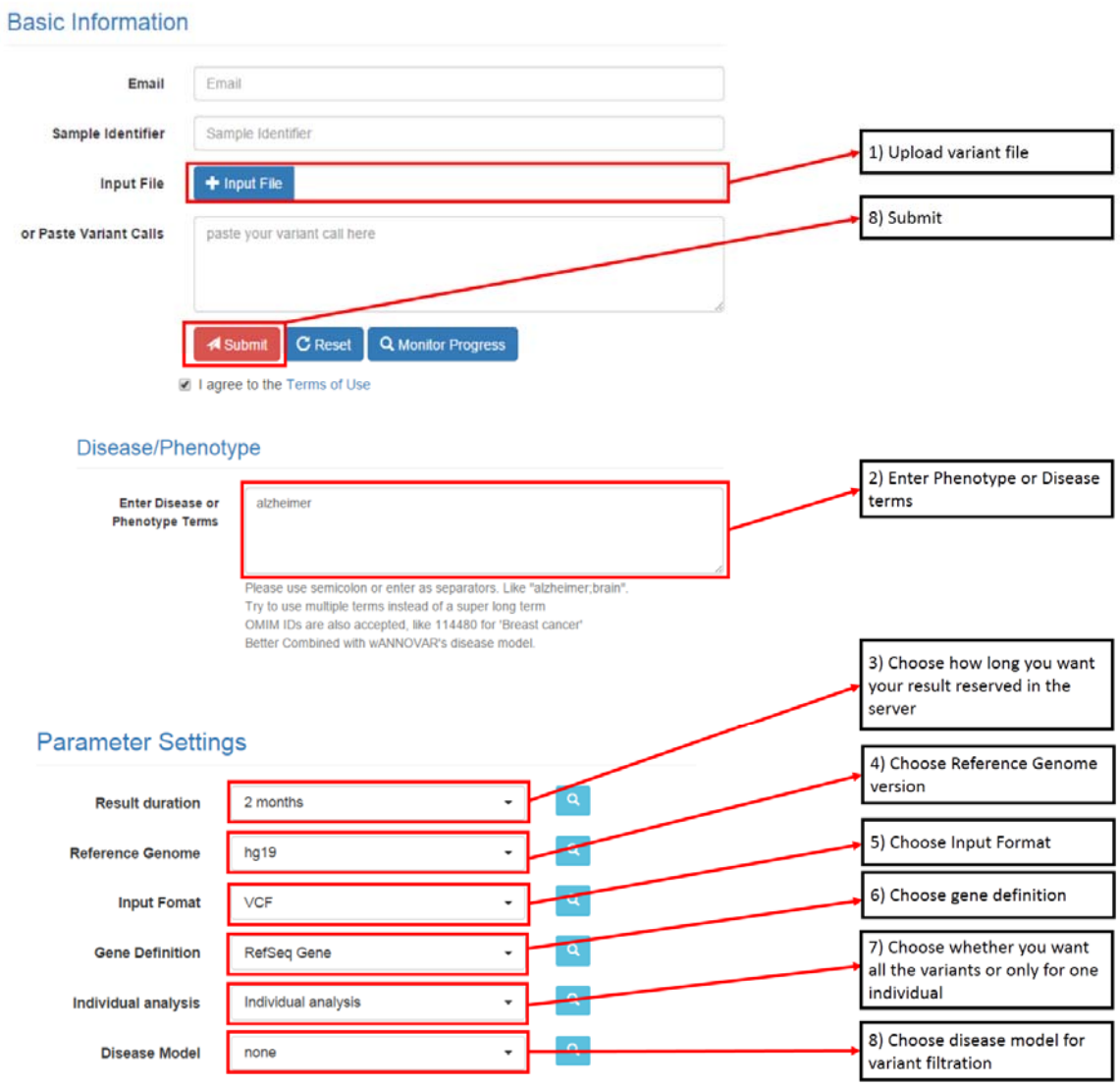
Example 1

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'

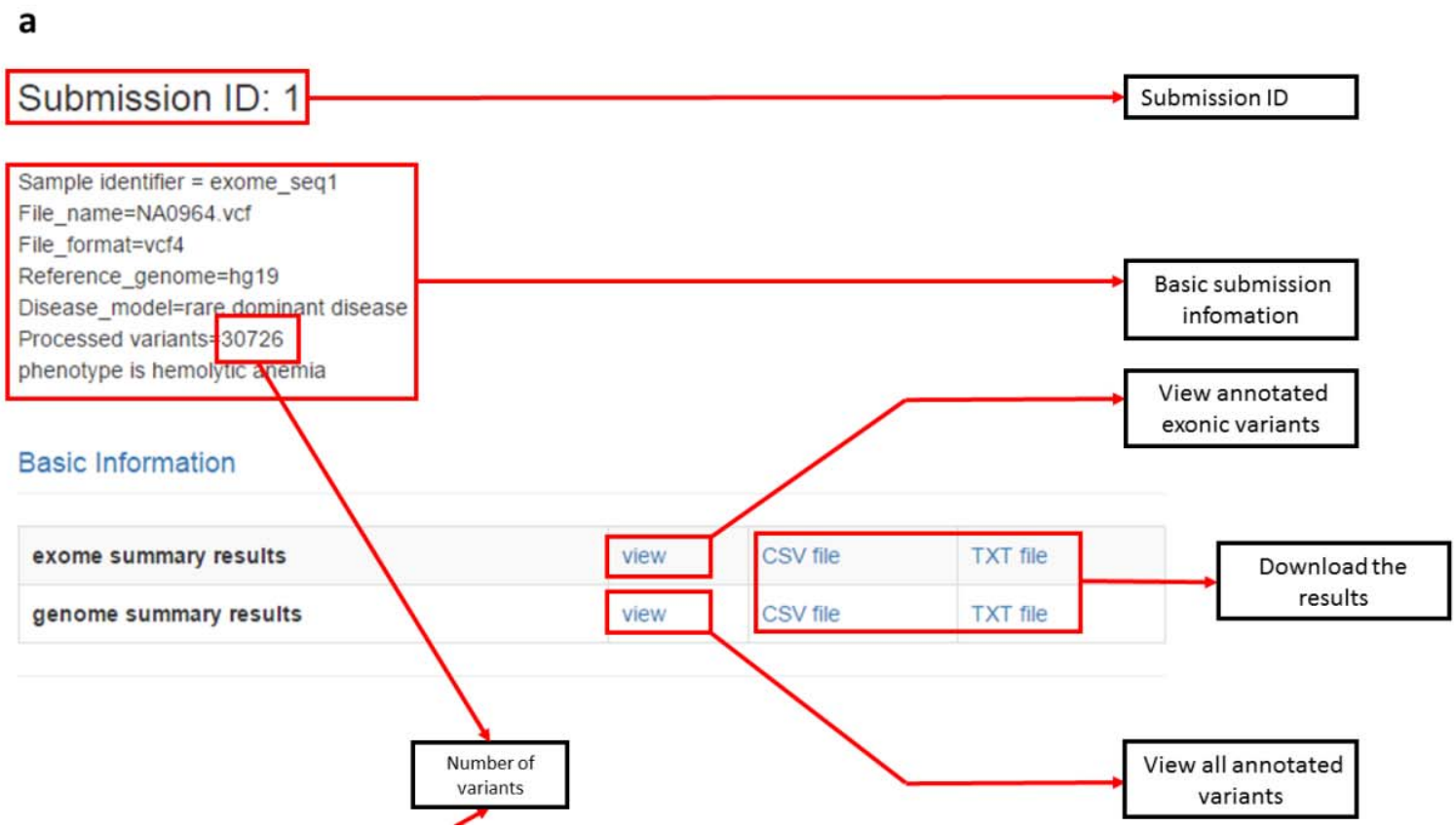
Input:
anemia.vcf

Output:
link to the result

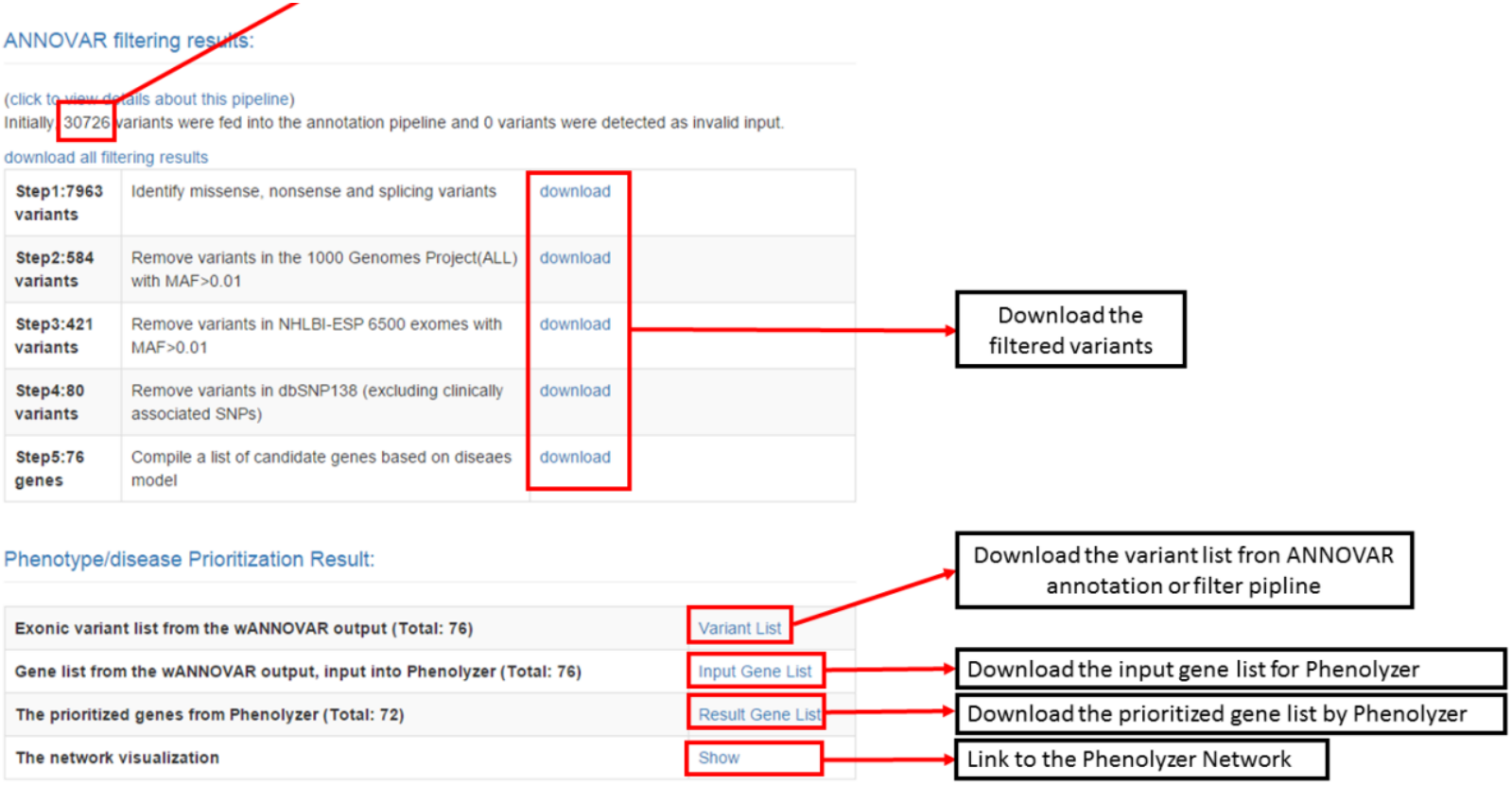
Job submission



Results page

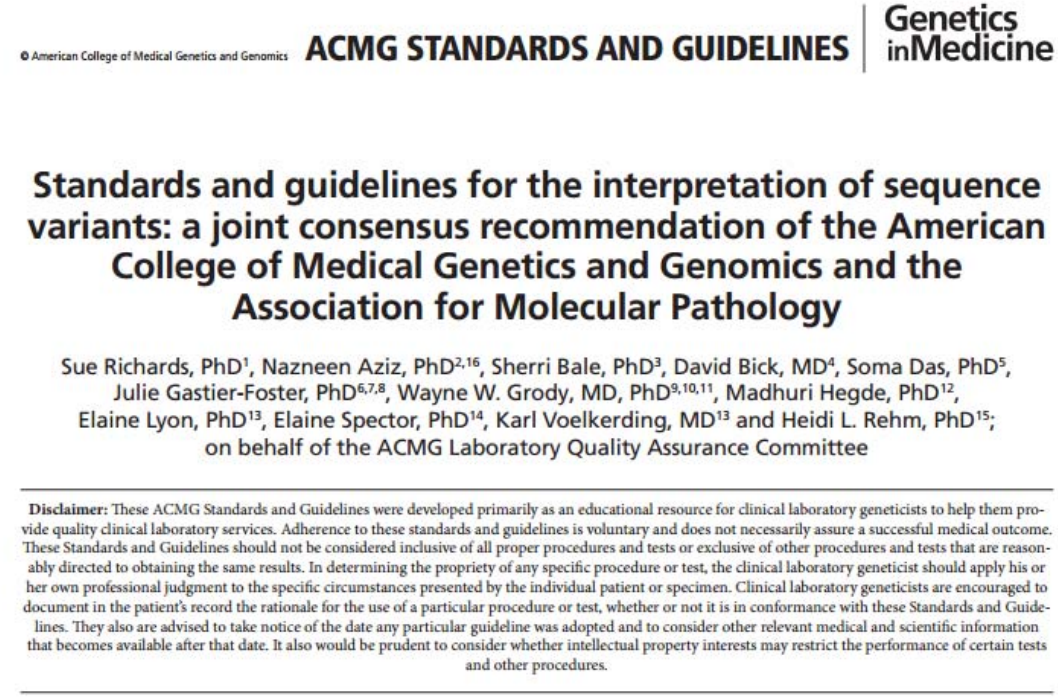


Results page



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.



ACMG STANDARDS AND GUIDELINES | Genetics inMedicine

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

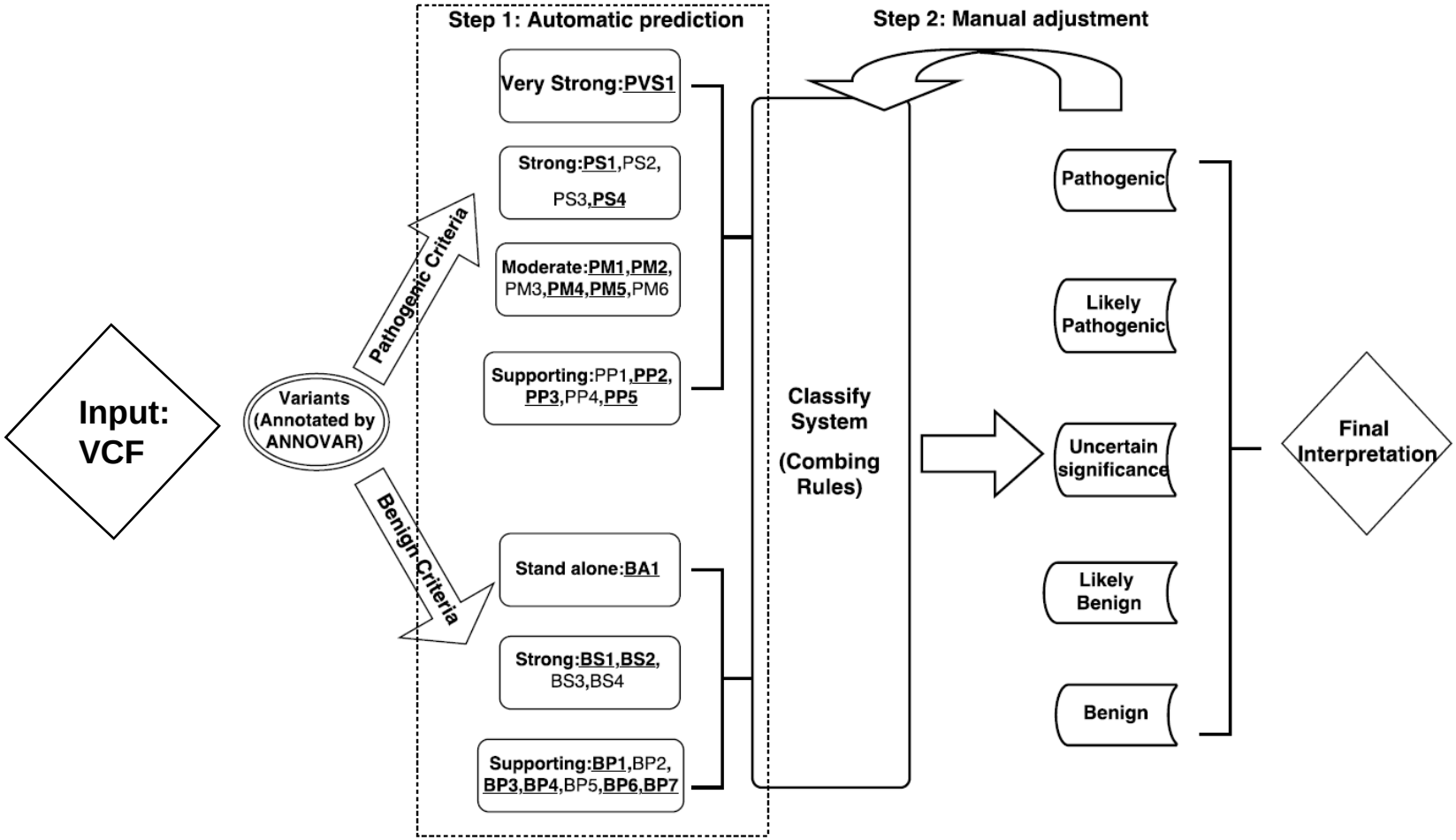
Disclaimer: These ACMG Standards and Guidelines were developed primarily as an educational resource for clinical laboratory geneticists to help them provide quality clinical laboratory services. Adherence to these standards and guidelines is voluntary and does not necessarily assure a successful medical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

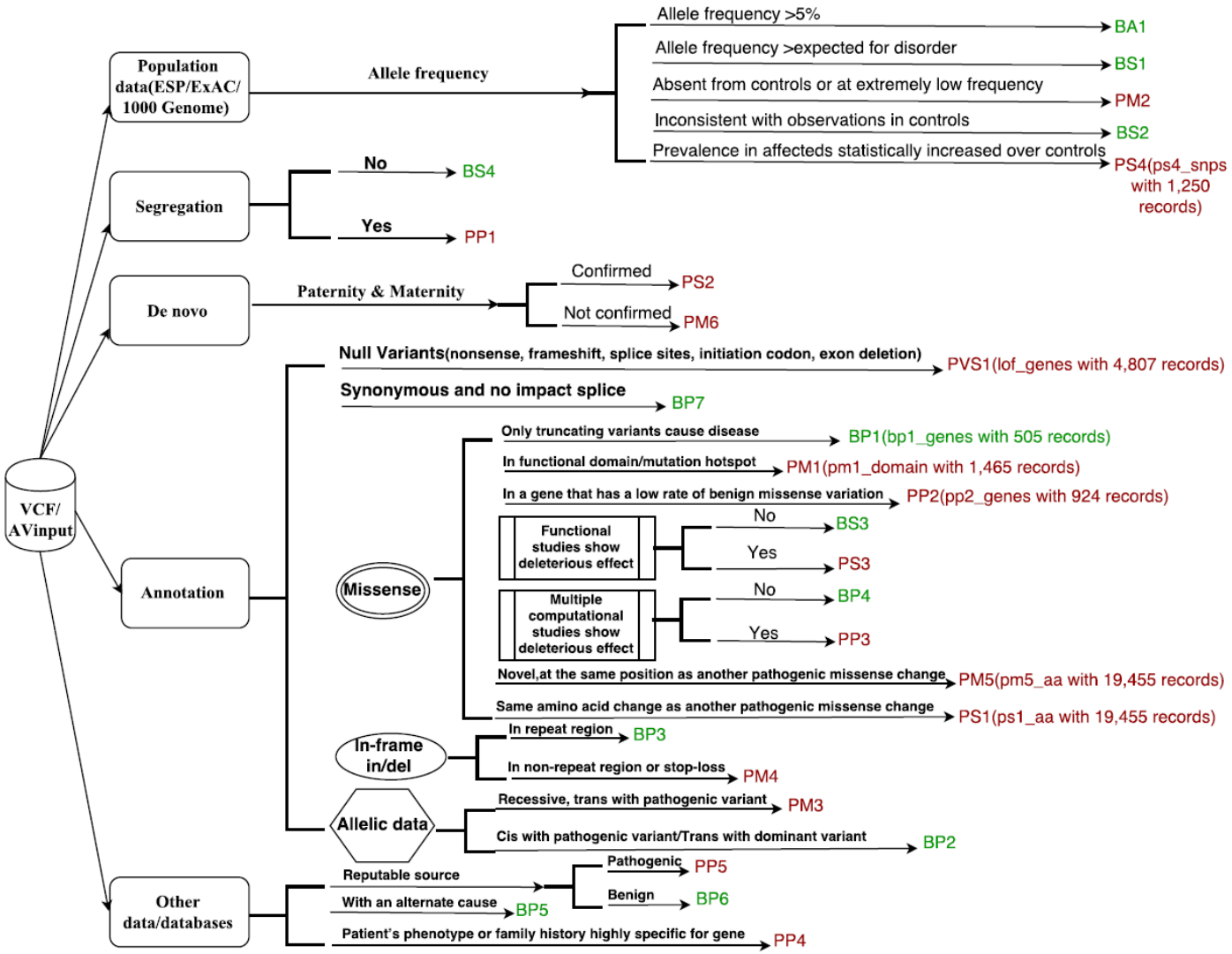
The system uses a total of 28 criteria and a five-tiered categorization for classifying variants.

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

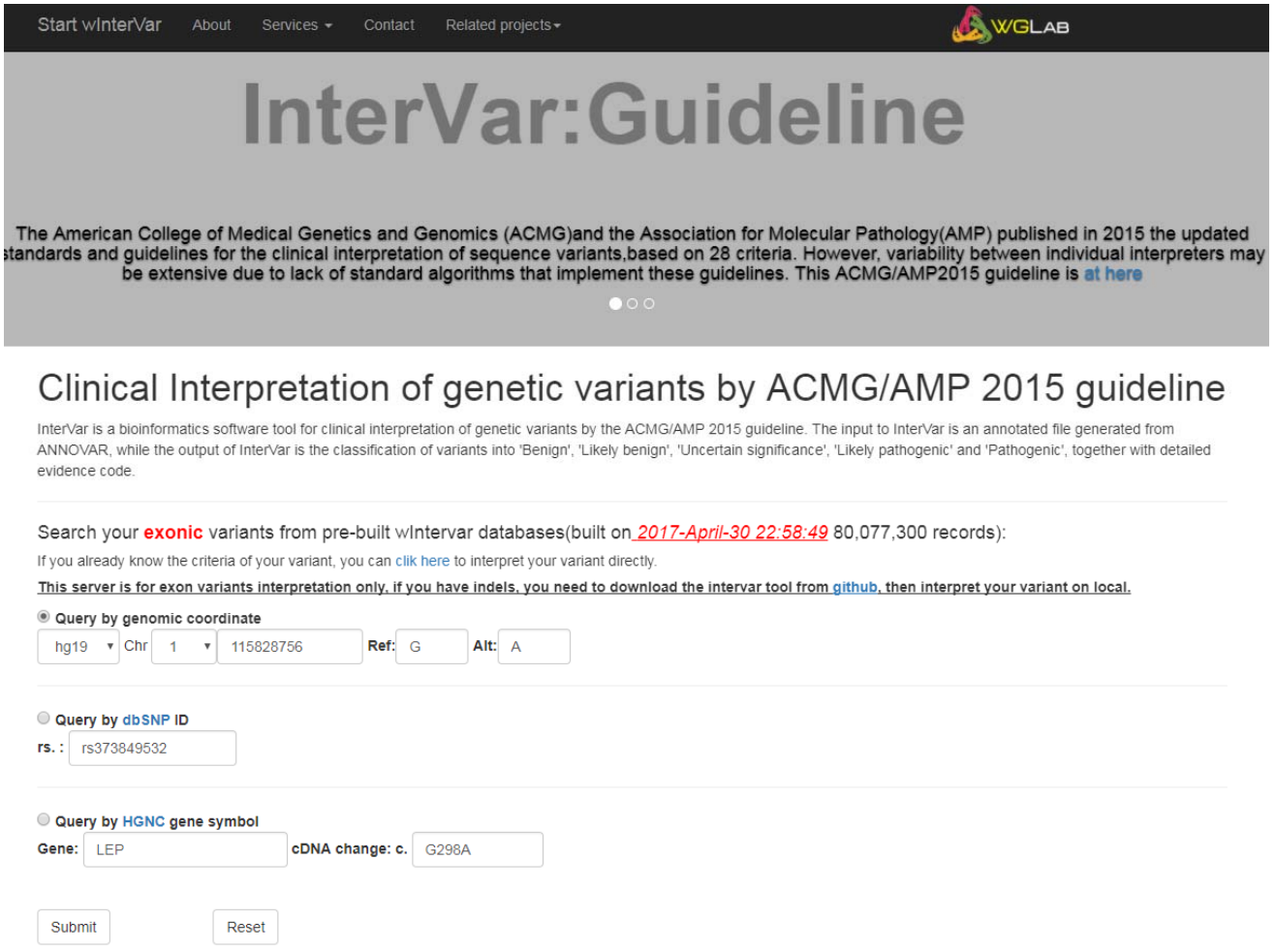


Li et al, *Am J Hum* Genet, 2:267-280, 2017

InterVar scoring logic

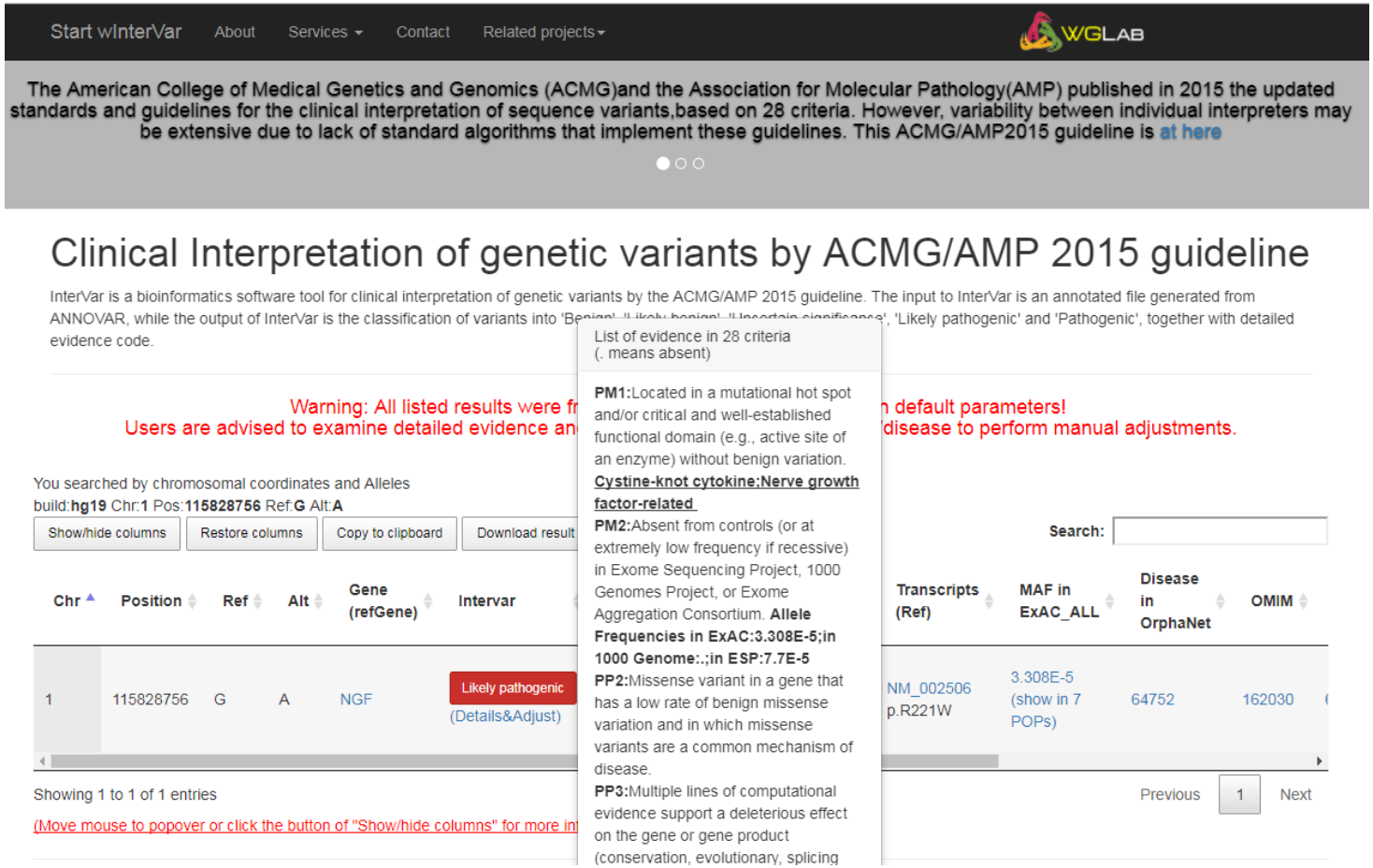


Web implementation of InterVar

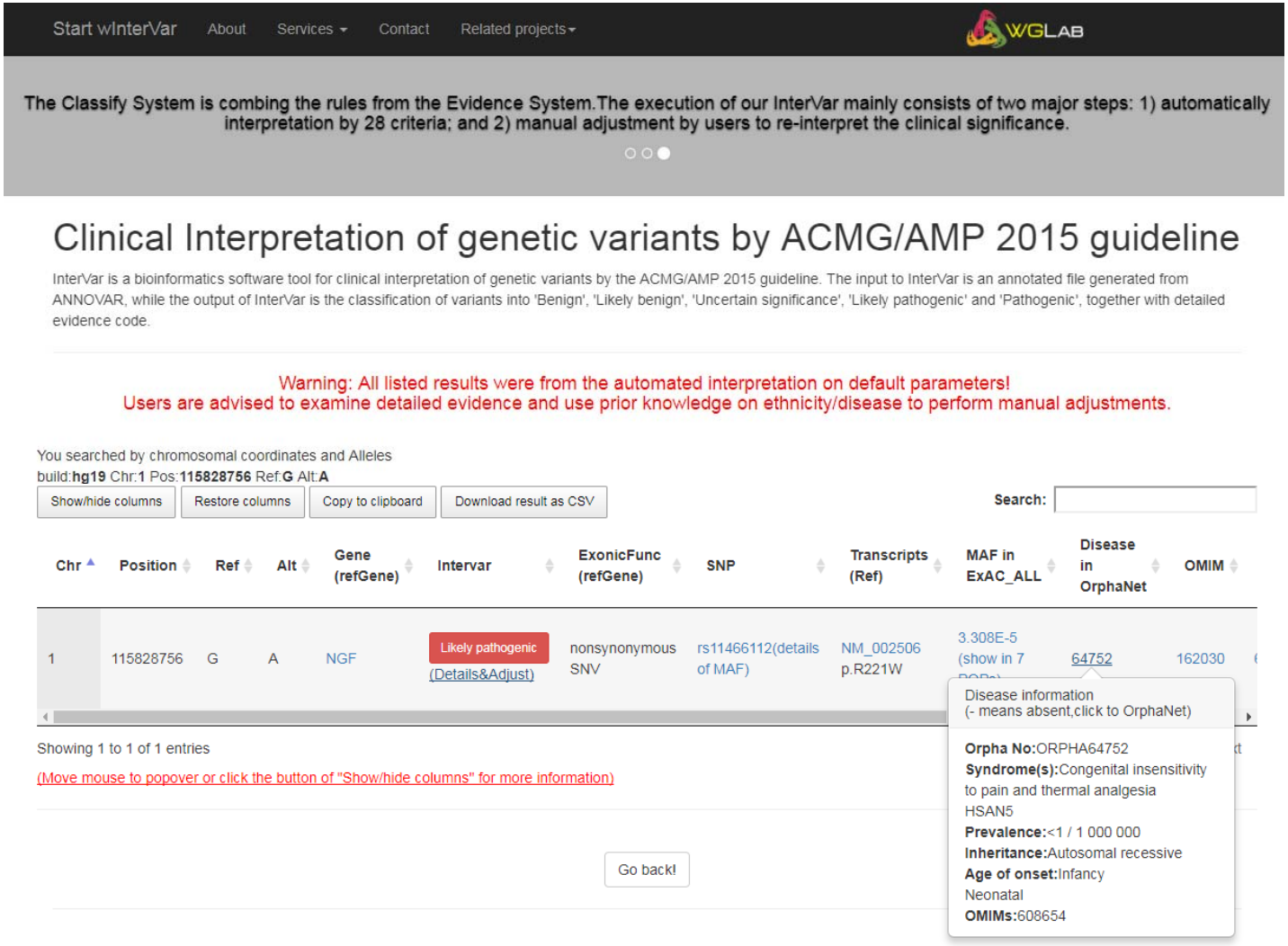


http://wintervar.wglab.org

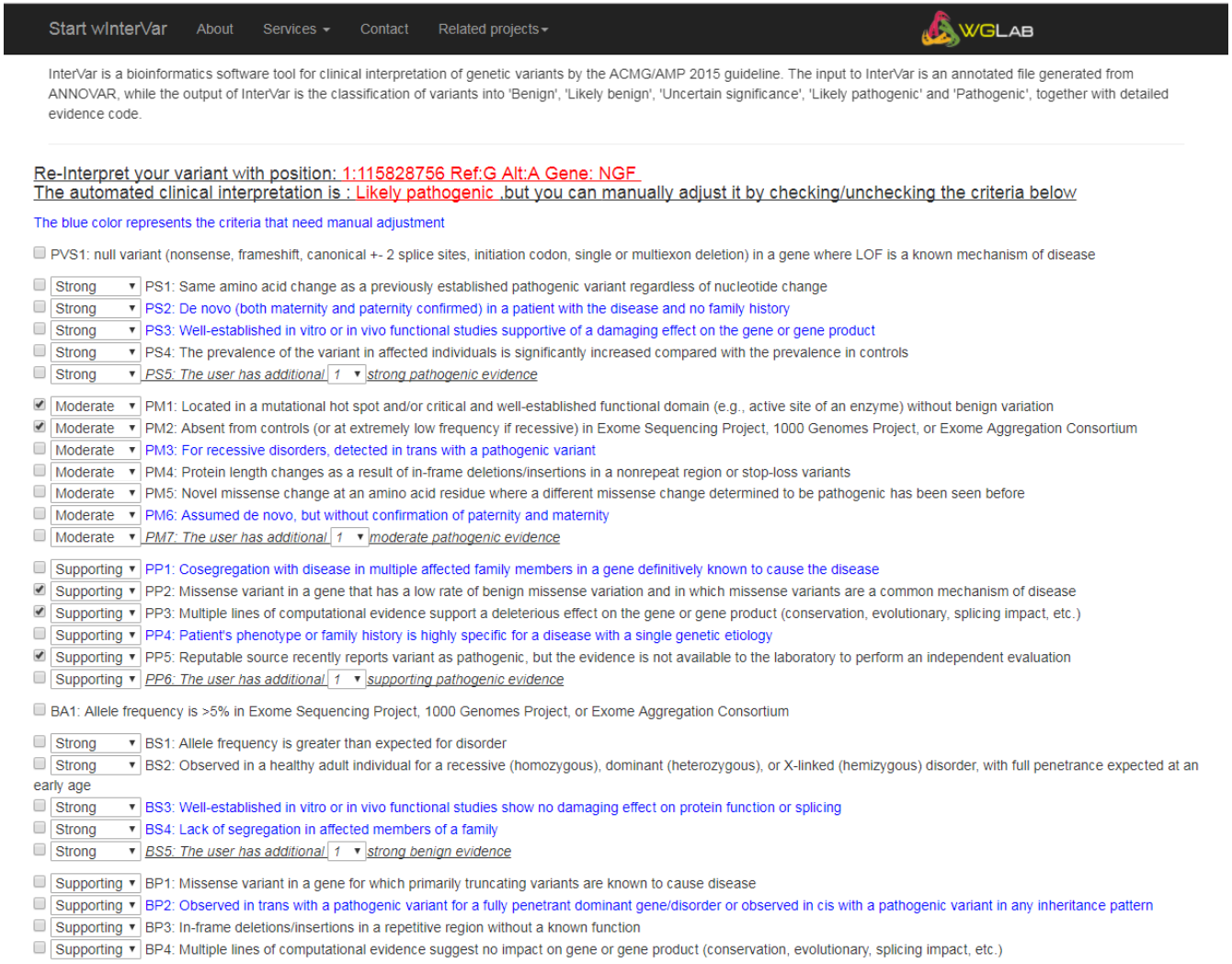
Web implementation of InterVar



Web implementation of InterVar



Web implementation of InterVar



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.

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[1, 2], Sebastian Köhler[1, 2], Sebastian Bauer[1], Dominik Seelow[1, 3], Denise Horn[1], Stefan Mundlos[1, 2, 4]

D966–D974 Nucleic Acids Research, 2014, Vol. 42, Database issue *Published online 11 November 2013*
doi:10.1093/nar/gkt1026

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

Sebastian Köhler[1,2,*], Sandra C. Doelken[1], Christopher J. Mungall[3], Sebastian Bauer[1], Helen V. Firth[4,5], Isabelle Bailleul-Forestier[6], Graeme C. M. Black[7,8], Danielle L. Brown[9], Michael Brudno[10,11], Jennifer Campbell[9,12], David R. FitzPatrick[13], Janan T. Eppig[14], Andrew P. Jackson[13], Kathleen Freson[15], Marta Girdea[10,11], Ingo Helbig[16], Jane A. Hurst[17], Johanna Jähn[16], Laird G. Jackson[18], Anne M. Kelly[19], David H. Ledbetter[20], Sahar Mansour[21], Christa L. Martin[20], Celia Moss[22], Andrew Mumford[23], Willem H. Ouwehand[4,19], Soo-Mi Park[5], Erin Rooney Riggs[20], Richard H. Scott[24], Sanjay Sisodiya[25], Steven Van Vooren[26], Ronald J. Wapner[27], Andrew O. M. Wilkie[28], Caroline F. Wright[4], Anneke T. Vulto-van Silfhout[29], Nicole de Leeuw[29], Bert B. A. de Vries[29], Nicole L. Washingthon[3], Cynthia L. Smith[14], Monte Westerfield[30], Paul Schofield[14,31], Barbara J. Ruef[30], Georgios V. Gkoutos[32], Melissa Haendel[33], Damian Smedley[4], Suzanna E. Lewis[3] 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. Vasilevsky[2], Mark Engelstad[2], Erin Foster[2], Julie McMurry[2], Ségolène Aymé[3], Gareth Baynam[4,5], Susan M. Bello[6], Cornelius F. Boerkoel[7], Kym M. Boycott[8], Michael Brudno[9], Orion J. Buske[9], Patrick F. Chinnery[10,11], Valentina Cipriani[12,13], Laureen E. Connell[14], Hugh J.S. Dawkins[15], Laura E. DeMare[14], Andrew D. Devereau[16], Bert B.A. de Vries[17], Helen V. Firth[18], Kathleen Freson[19], Daniel Greene[20,21], Ada Hamosh[22], Ingo Helbig[23,24], Courtney Hum[25], Johanna A. Jähn[24], Roger James[11,21], Roland Krause[26], Stanley J. F. Laulederkind[27], Hanns Lochmüller[28], Gholson J. Lyon[29], Soichi Ogishima[30], Annie Olry[31], Willem H. Ouwehand[20], Nikolas Pontikos[12,13], Ana Rath[31], Franz Schaefer[32], Richard H. Scott[16], Michael Segal[33], Panagiotis I. Sergouniotis[34], Richard Sever[14], Cynthia L. Smith[6], Volker Straub[28], Rachel Thompson[28], Catherine Turner[28], Ernest Turro[20,21], Marijcke W.M. Veltman[11], Tom Vulliamy[35], Jing Yu[36], Julie von Ziegenweidt[20], Andreas Zankl[37,38], Stephan Züchner[39], Tomasz Zemojtel[1], Julius O.B. Jacobsen[16], Tudor Groza[40,41], Damian Smedley[16], Christopher J. Mungall[42], Melissa Haendel[2] and Peter N. Robinson[43,44,*]

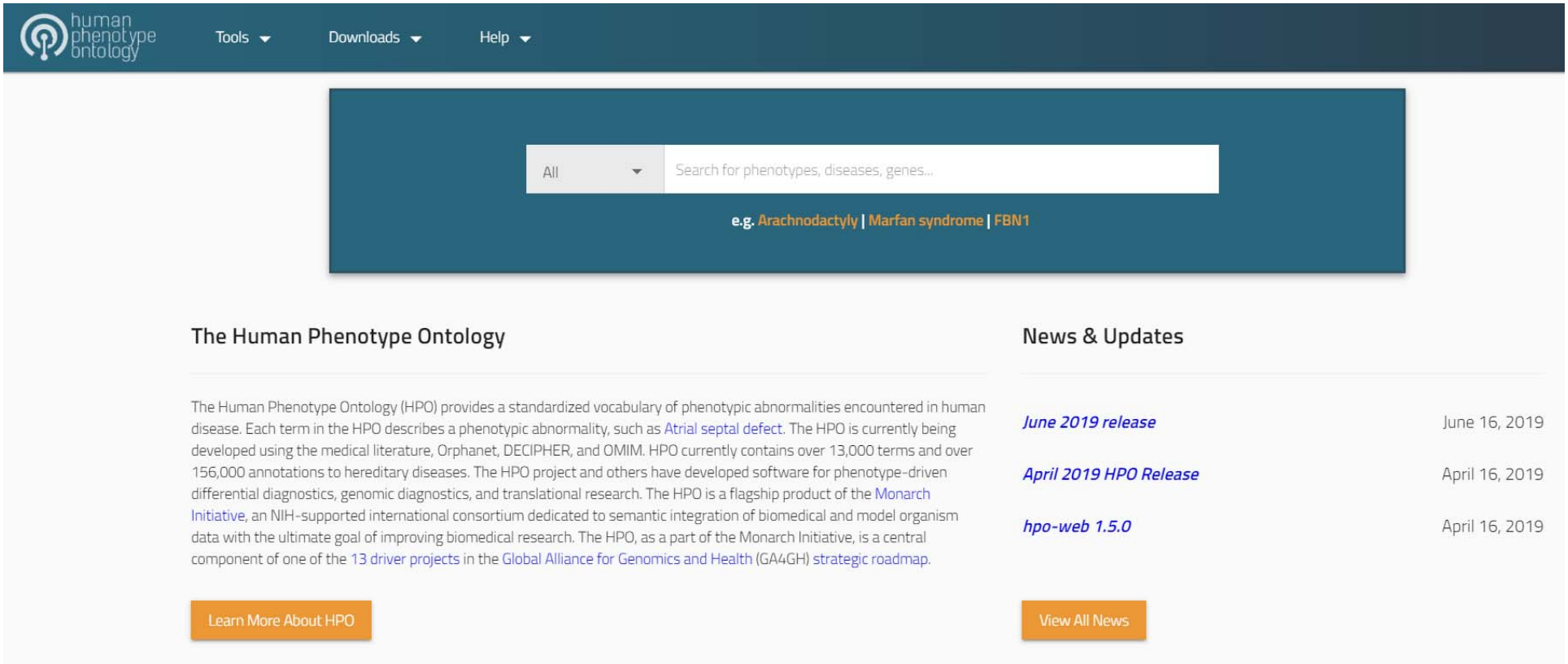
HPO in 2019: new website and expanded knowledge base

D1018–D1027 *Nucleic Acids Research, 2019, Vol. 47, Database issue* *Published online 22 November 2018*
doi: 10.1093/nar/gky1105

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 Palmer[17], Dylan Gratian[14], Hugh Dawkins[18], Michael Segal[19], Anna C. Jansen[20,21], Ahmed Muaz[3,22], Willie H. Chang[23], Jenna Bergerson[24], Stanley J.F. Laulederkind[25], Zafer Yüksel[26], Sergi Beltran[27,28], Alexandra F. Freeman[24], Panagiotis I. Sergouniotis[29], Daniel Durkin[4], Andrea L. Storm[30,31], Marc Hanauer[32], Michael Brudno[23], Susan M. Bello[33], Murat Sincan[34], Kayli Rageth[34], Matthew T. Wheeler[35], Renske Oegema[36], Halima Lourghi[32], Maria G. Della Rocca[30,31], Rachel Thompson[37], Francisco Castellanos[4], James Priest[38], Charlotte Cunningham-Rundles[39], Ayushi Hegde[4], Ruth C. Lovering[40], Catherine Hajek[34], Annie Olry[32], Luigi Notarangelo[24], Morgan Similuk[24], Xingmin A. Zhang[3,4], David Gómez-Andrés[41], Hanns Lochmüller[27,42,43,44], Hélène Dollfus[45], Sergio Rosenzweig[46], Shruti Marwaha[35], Ana Rath[32], Kathleen Sullivan[47], Cynthia Smith[33], Joshua D. Milner[24], Dorothée Leroux[45], Cornelius F. Boerkoel[34], Amy Klion[24], Melody C. Carter[24], 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,*]

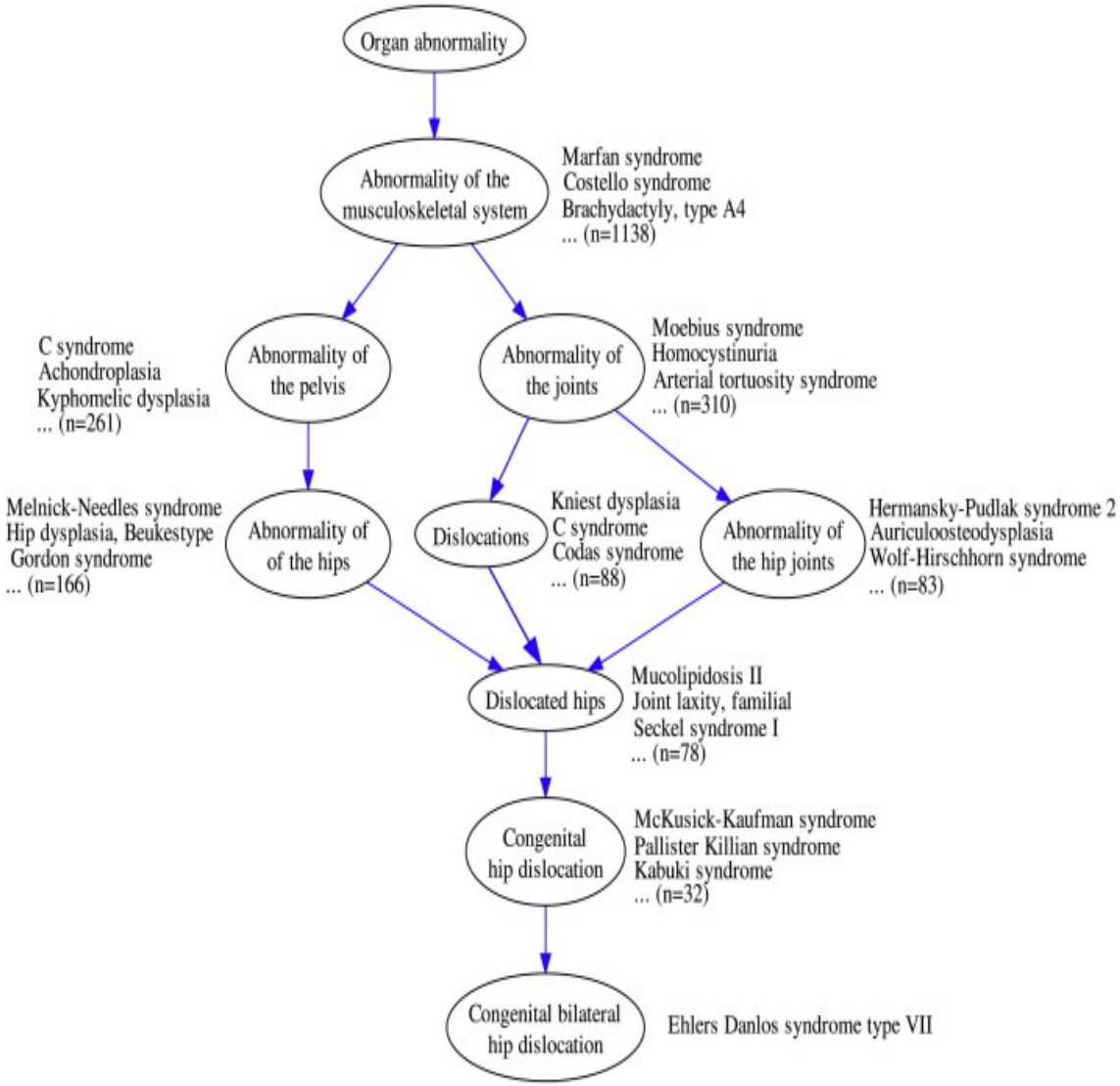
Various tools and data sets can be downloaded from the new HPO site



https://hpo.jax.org/ (new website

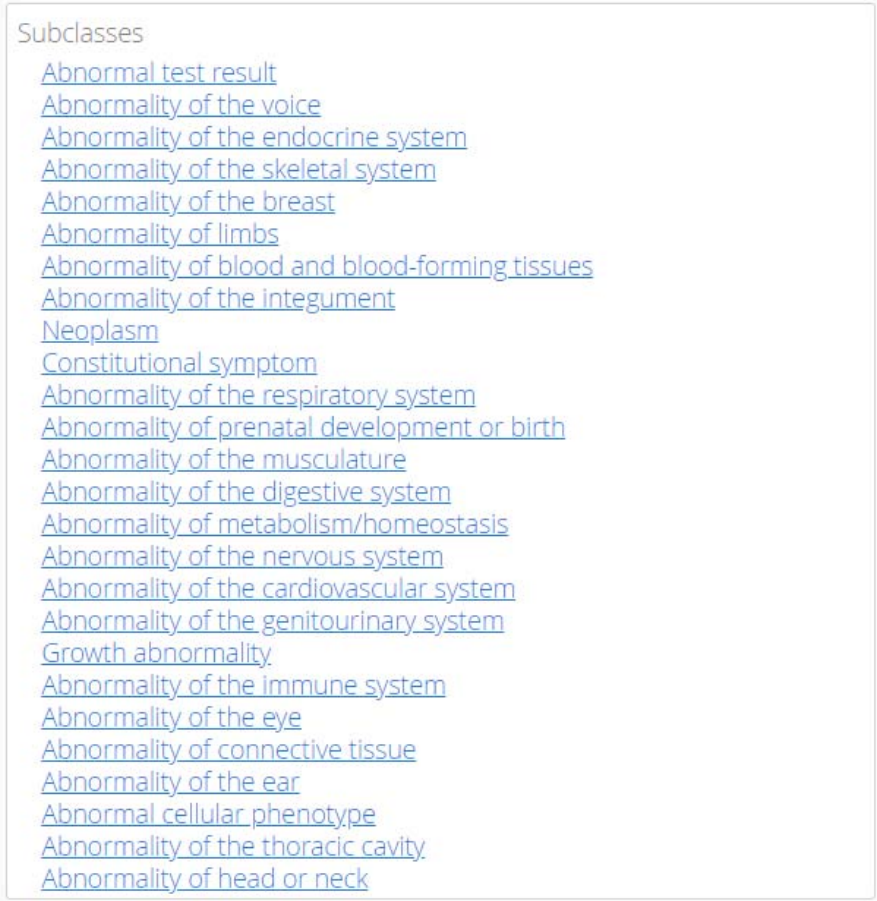
What does HPO look like?

- 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



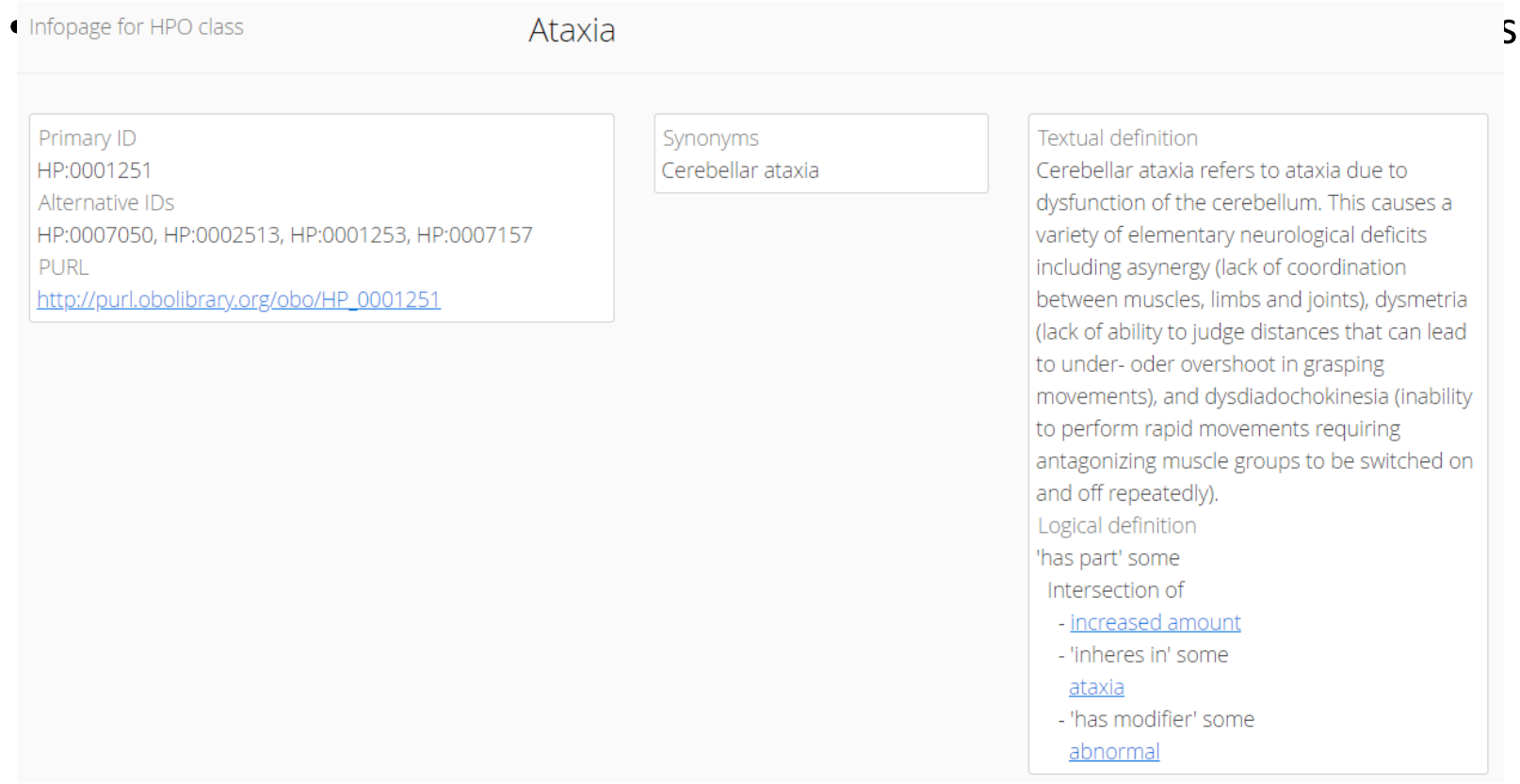
HPO browser

- 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).



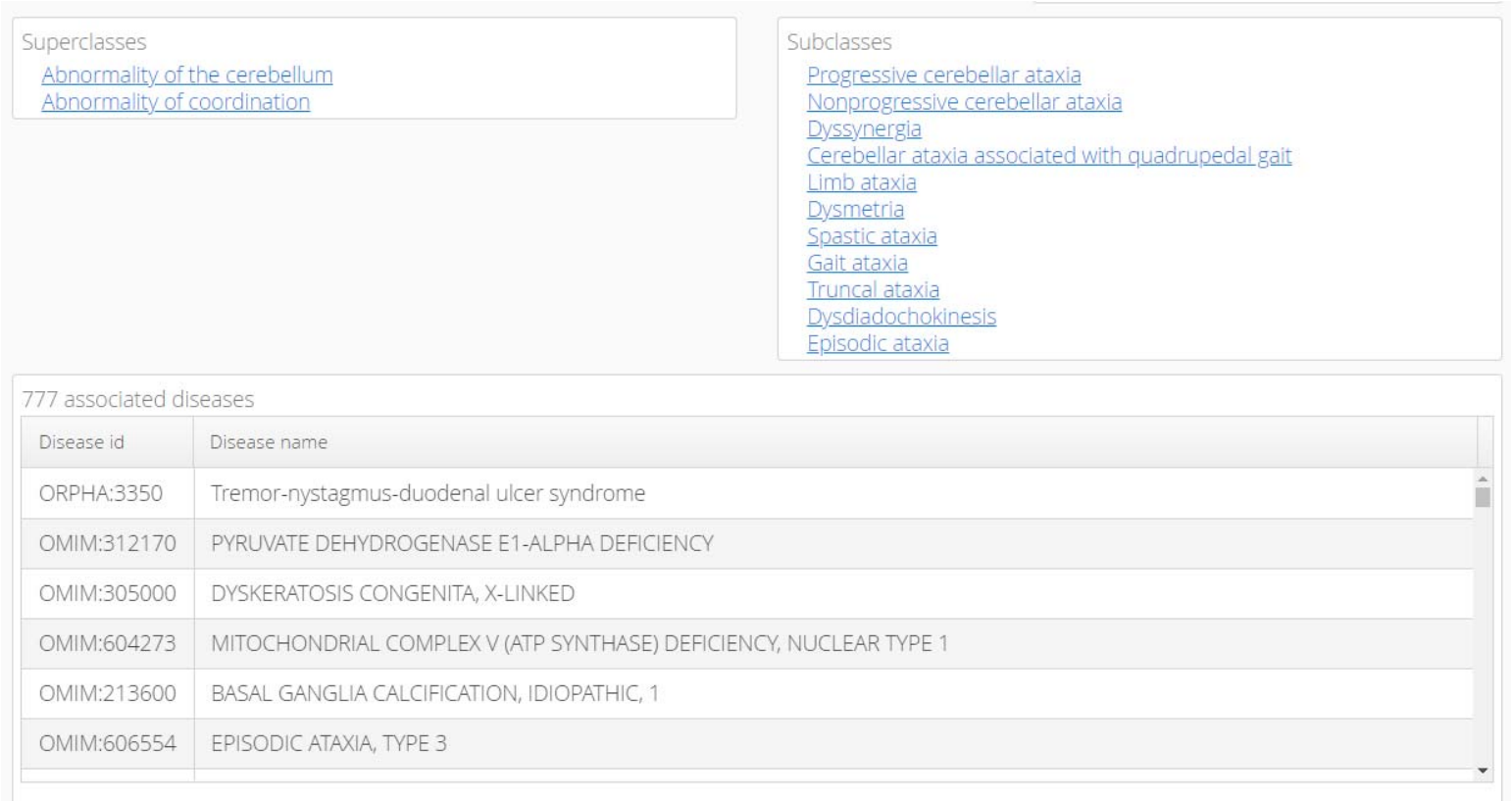
Examples of HPO terms

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



Ataxia (HP:0001251)

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



Superclasses

Abnormality of the cerebellum
Abnormality of coordination

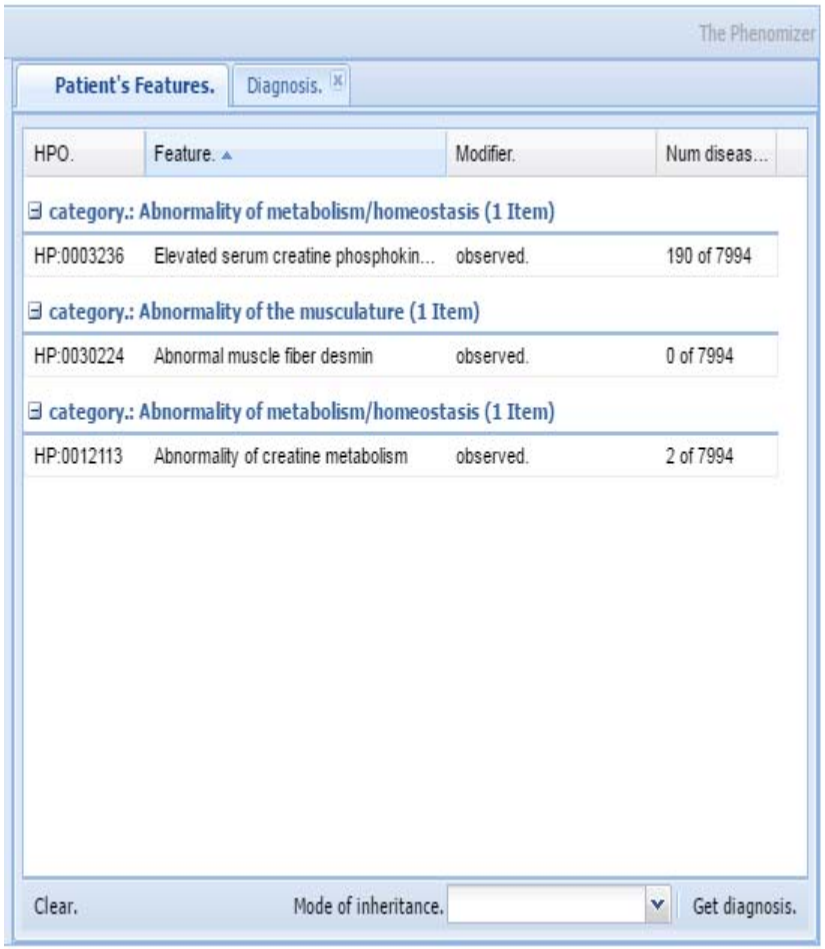
Subclasses

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

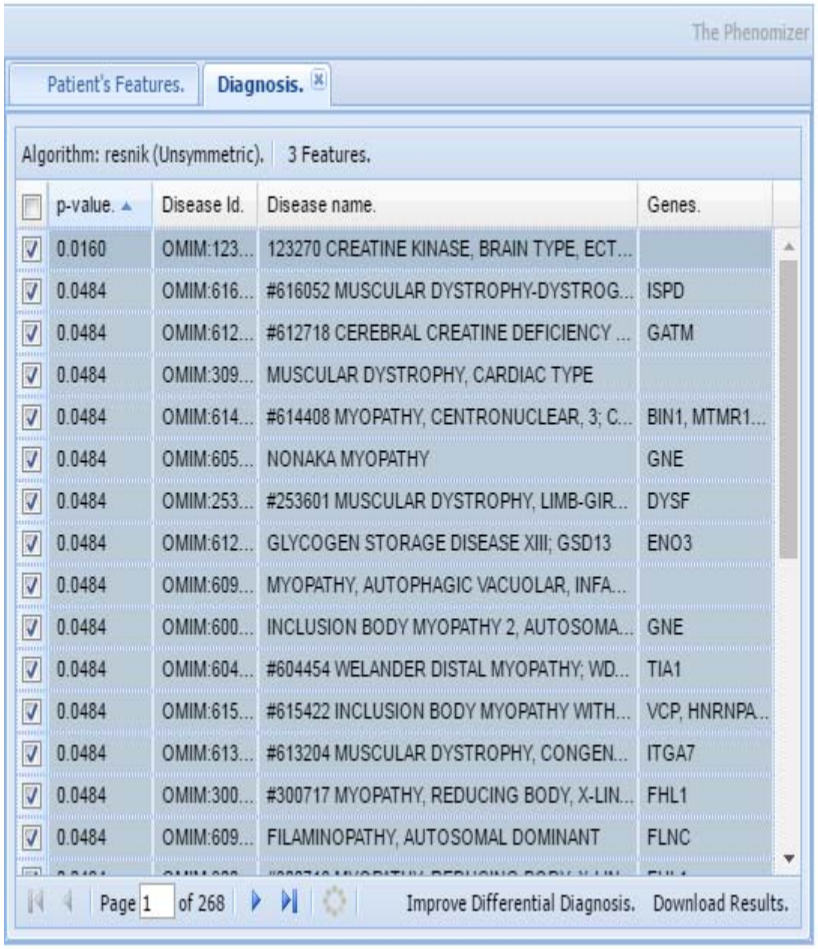
777 associated diseases

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



Input: HPO terms

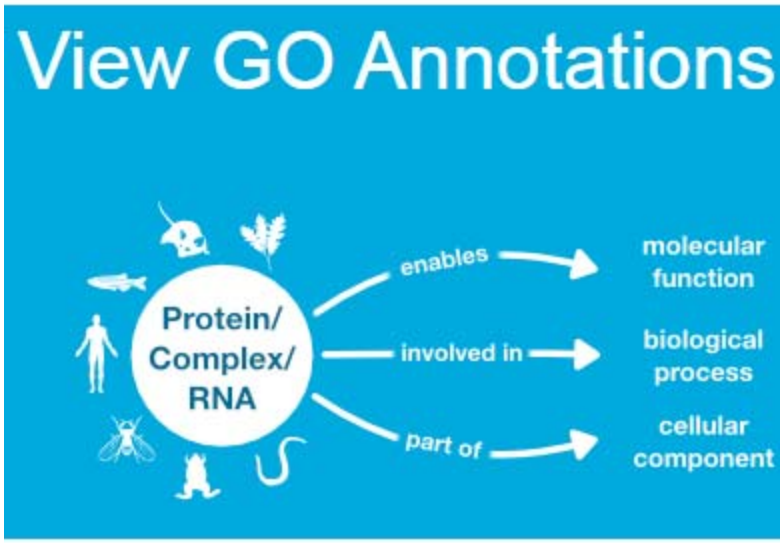


Output: Disease diagnosis and p-values

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.**

Gene Ontology represents 3 different ontologies

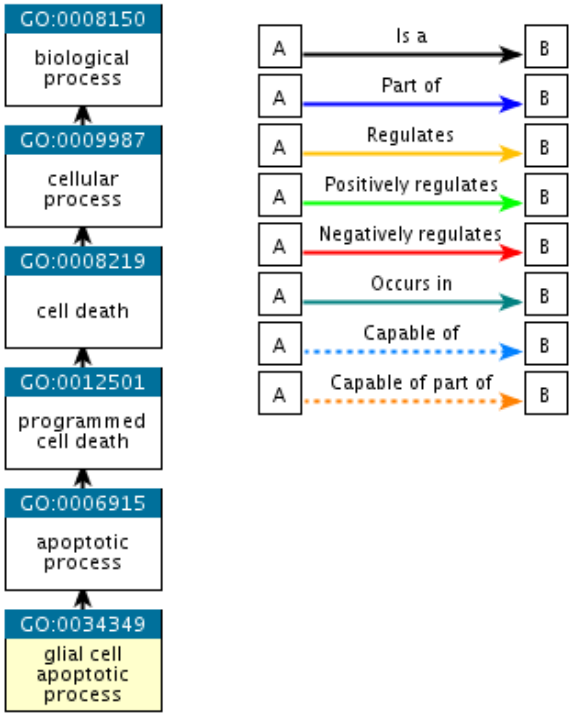


- Cellular components:
  - The parts of a cell or its extracellular environment;
- Molecular functions:
  - The elemental activities of a gene product at the molecular level, such as binding or catalysis;
- Biological processes:
  - Operations or sets of molecular events with a defined beginning and 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 10

«First | <Prev | Next> | Last» | Download (up to 100000)

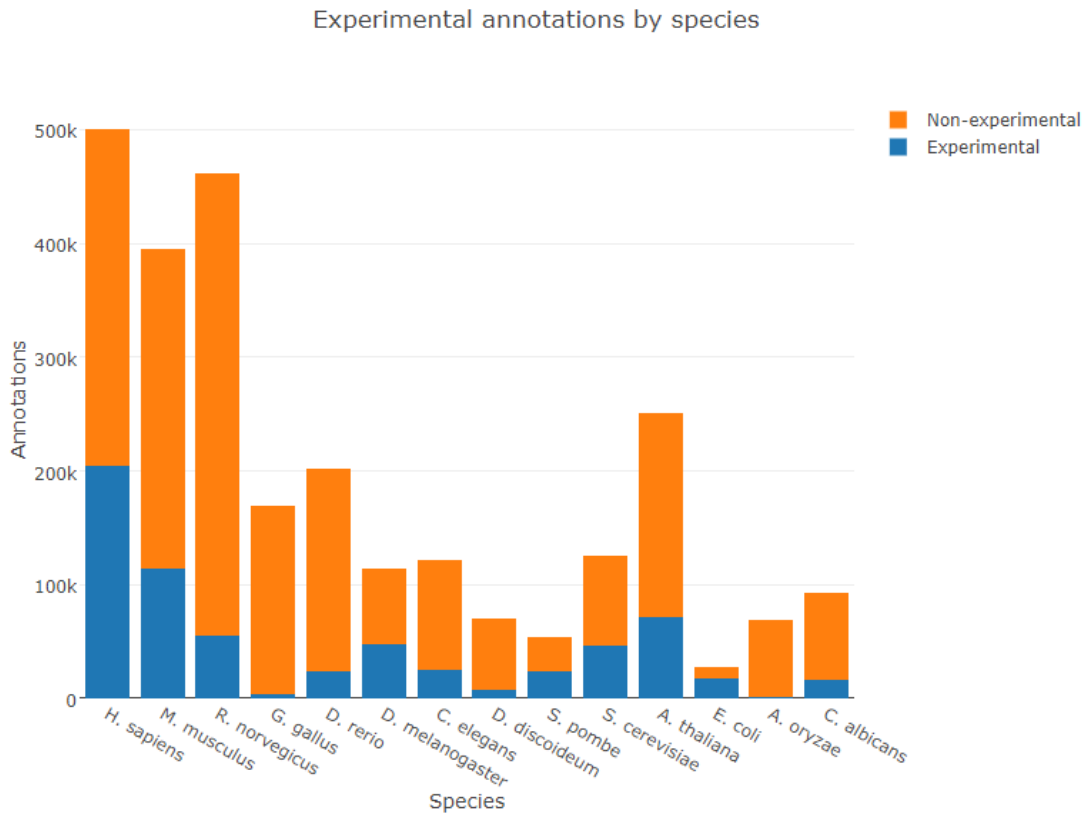
Gene/product	Gene/product name	Annotation qualifier	GO class (direct)	Annotation extension	Contributor	Organism	Evidence	Evidence with	PANTHER family	Isoform	Reference	Date
TCF4	Transcription factor 4		nuclear chromatin	coincident with ENSEMBL:ENSG00000168646 coincident with SO:0000167	ParkinsonsUK-UCL	Homo sapiens	IDA		basic helix-loop-helix transcription factor pthr11793		PMID:21880741	20160517
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
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

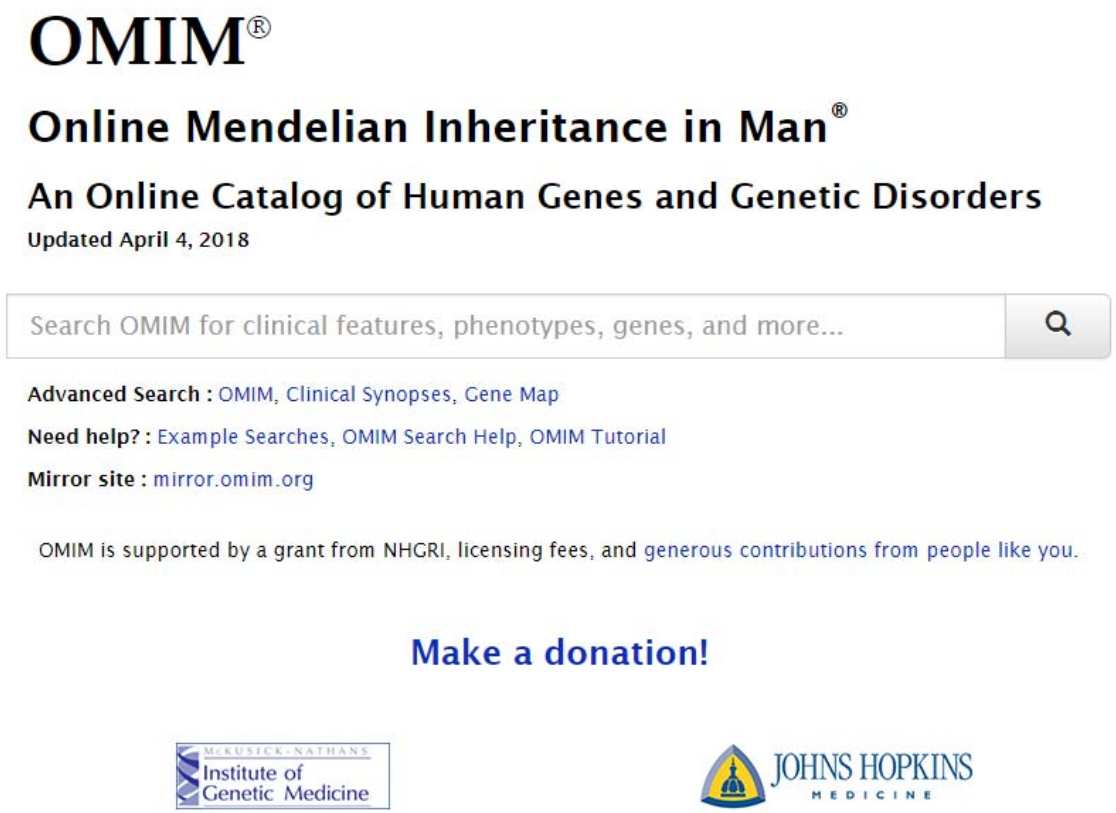
GO associates annotations to genes

- 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'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

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

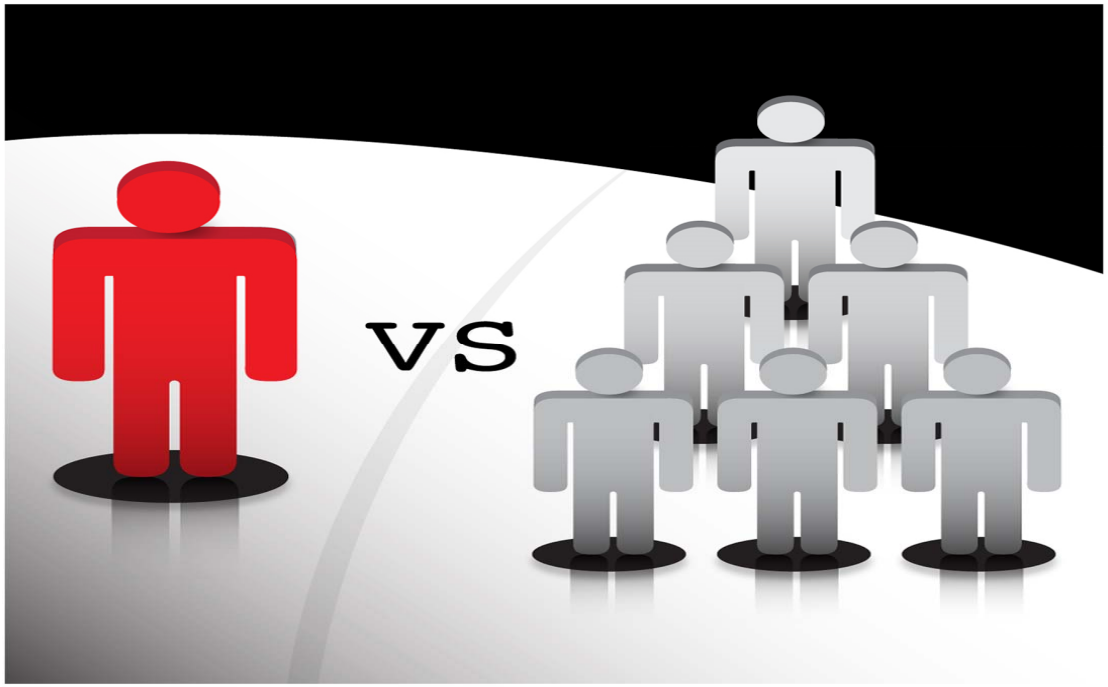
TEXT

▼ 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 Ephrussi (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

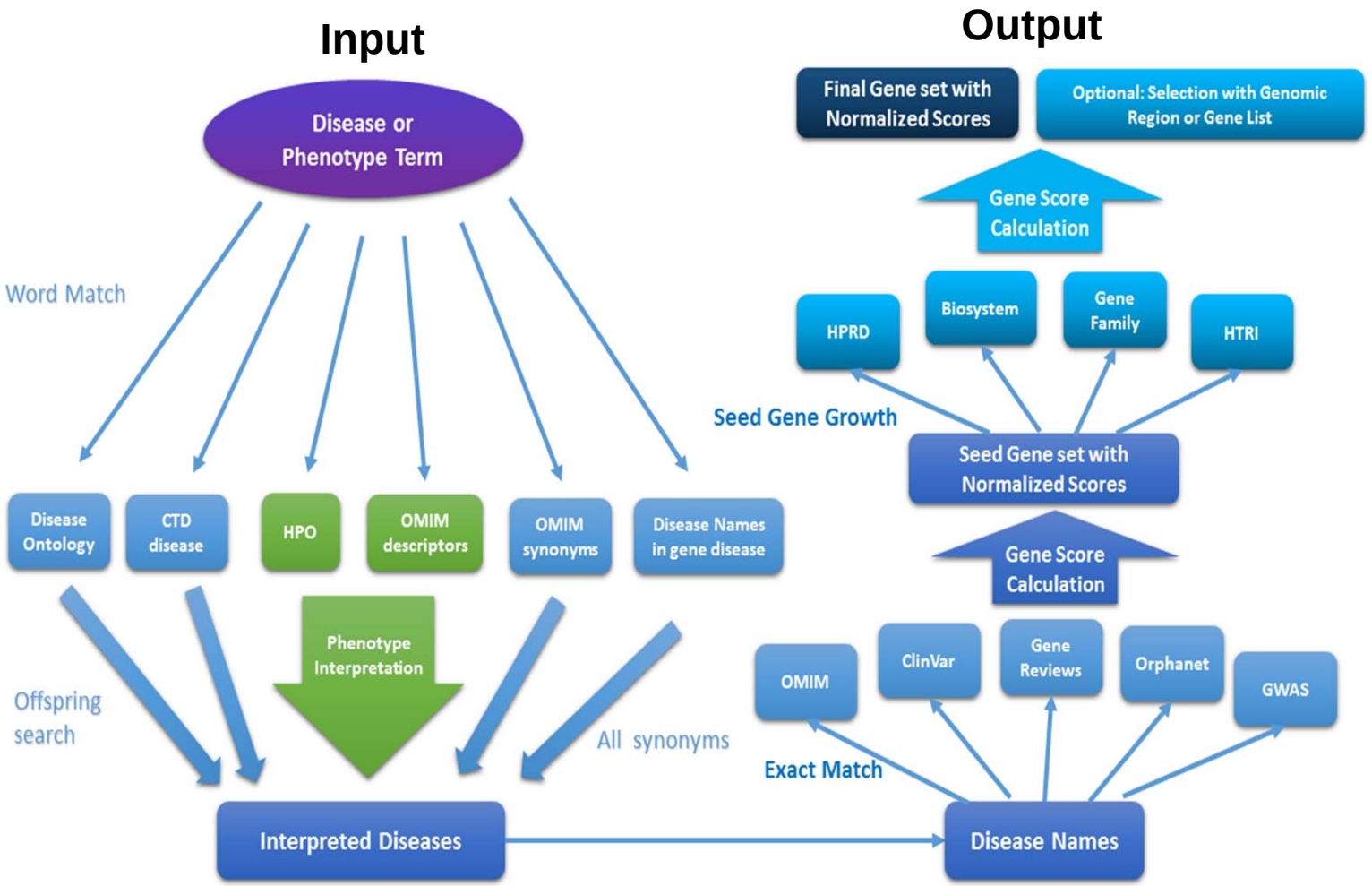
GNAS
SMAD1
TGFBR1
SMAD9
CREBBP
BMP2
NRAS
CHN1
IKBKB

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

What is Phenolyzer

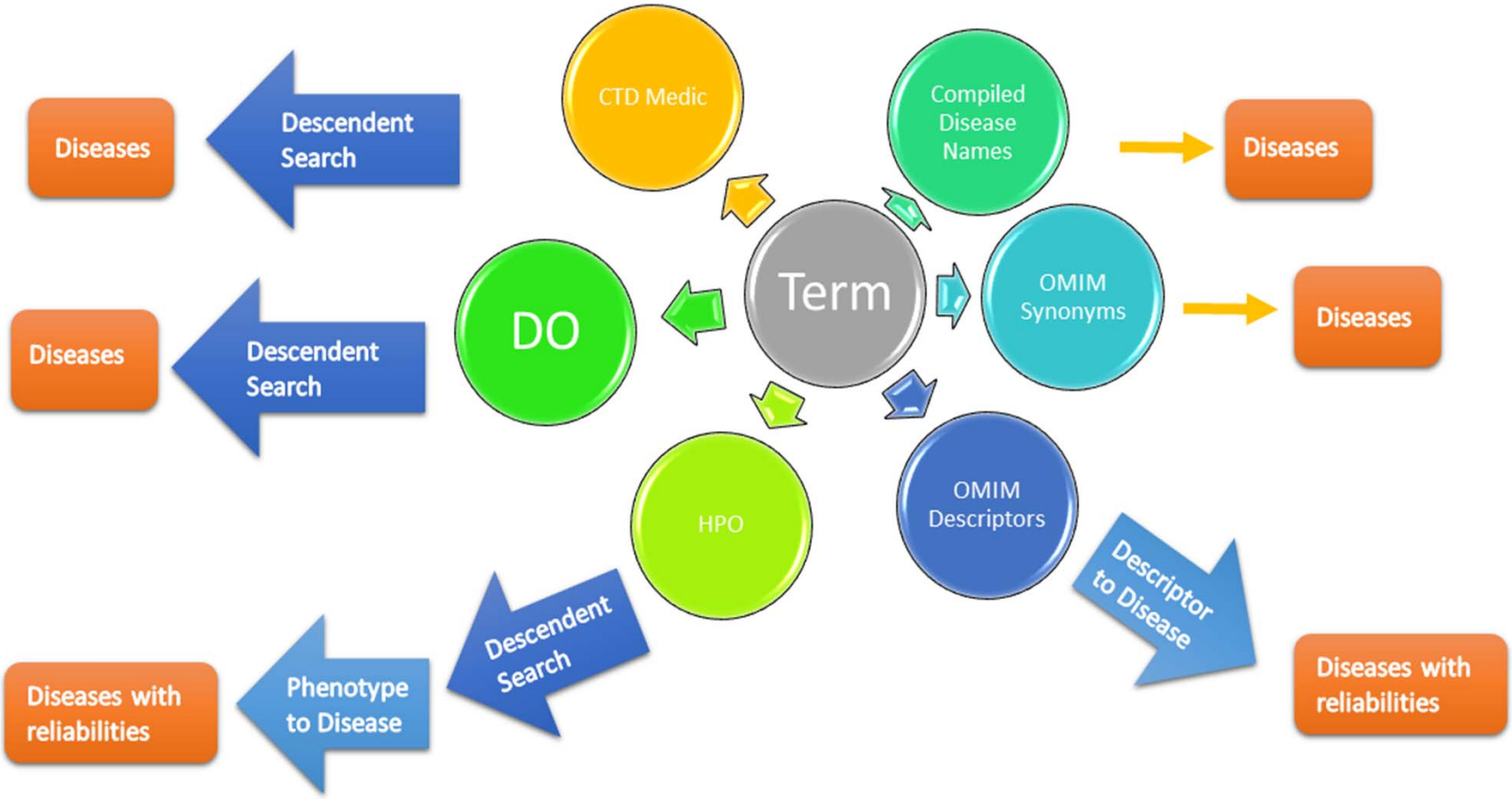
- A tool for phenotype analysis:
  - 1) To map user-supplied phenotypes to diseases and candidate genes
  - 2) A resource that integrates existing biological knowledge to identify known disease genes
  - 3) A prediction algorithm to predict novel disease genes
  - 4) A model to integrate multiple features 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



Yang et al, *Nature Methods* (2015) 12:841-843

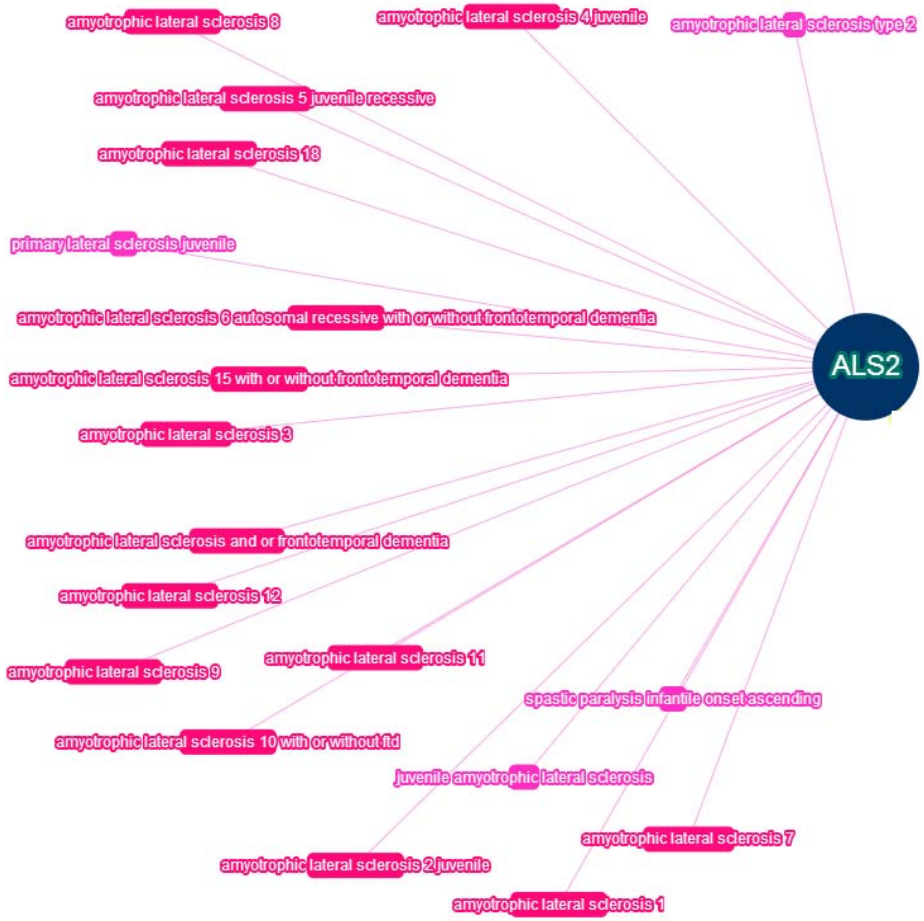
Step 1: Term interpretation



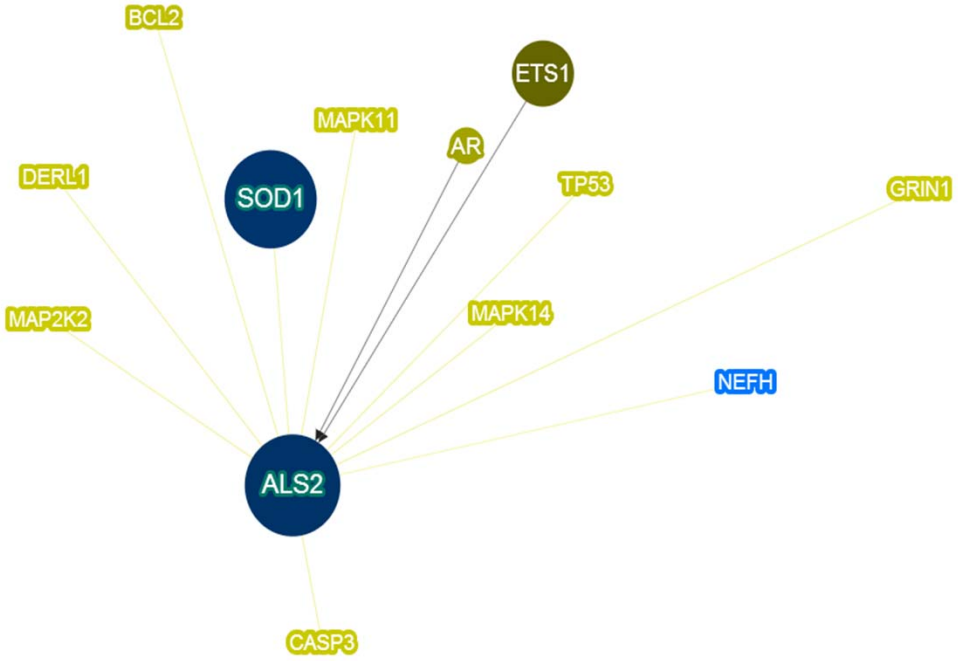
“Term” can be any phenotype term, such as “developmental delay”, “hearing loss”, etc

‘Amyotrophic lateral sclerosis’ interpreted

Step 2: Seed gene set generation

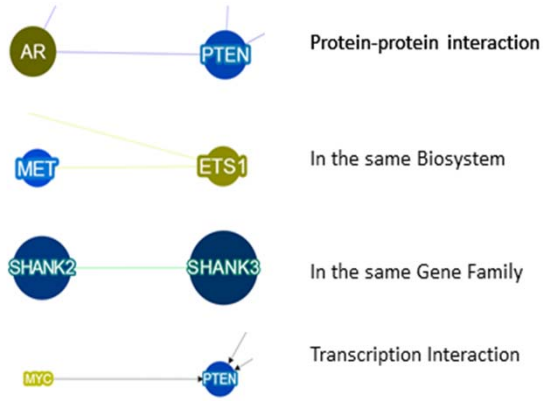


Step 3: Seed gene set growth

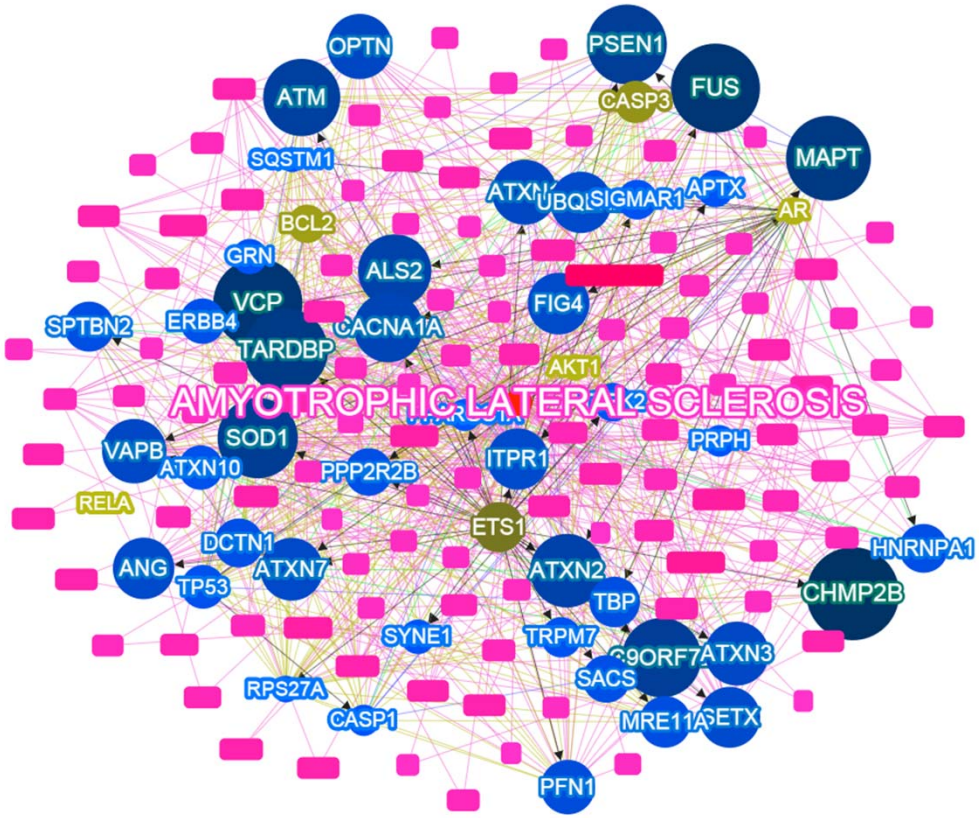


ALS2 is grown based on four type of gene-gene relations,:

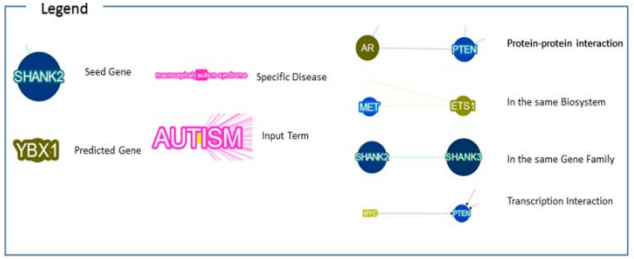
1. **protein interaction**
2. **pathway**
3. **gene family**
4. **transcription interaction**



Step 4: Data integration and scoring



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

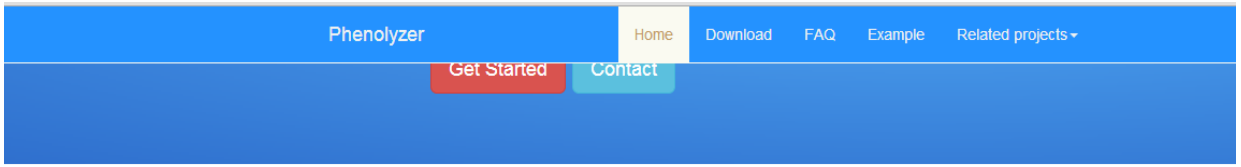


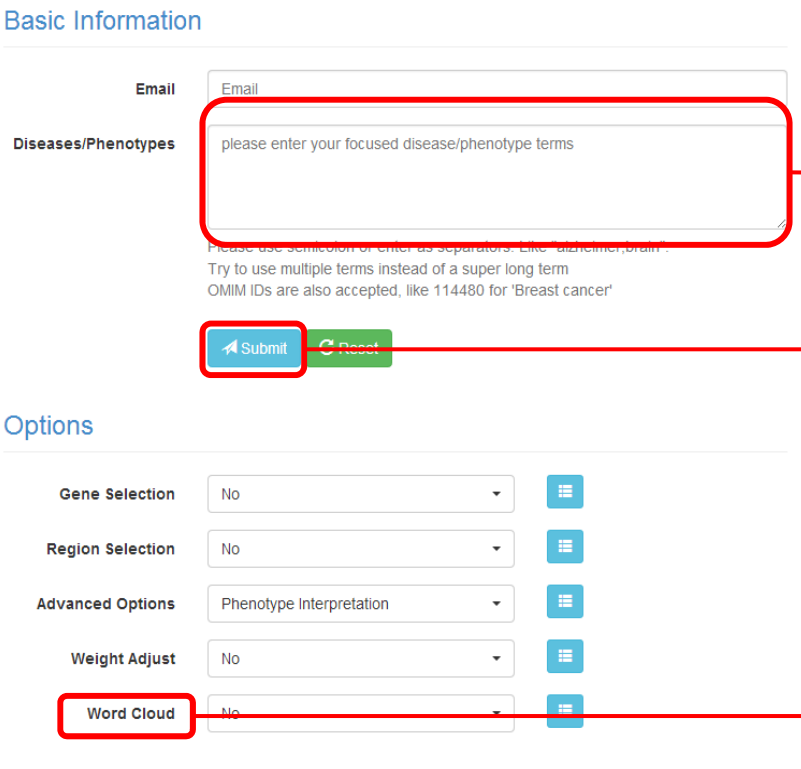
Phenolyzer generates a gene-phenotype-disease network for ALS

Web implementation

http://phenolyzer.wglab.org

1) Enter the website

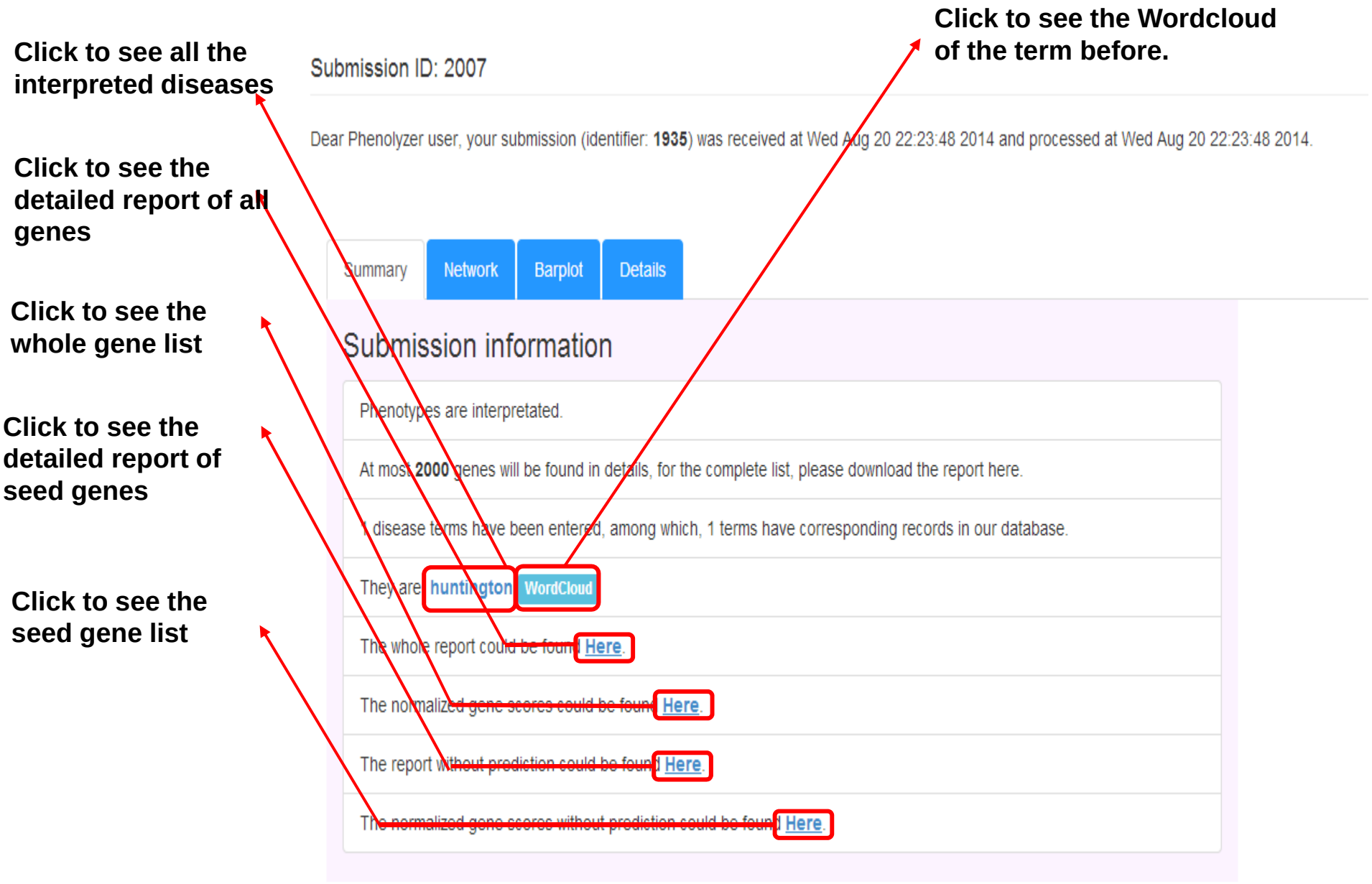




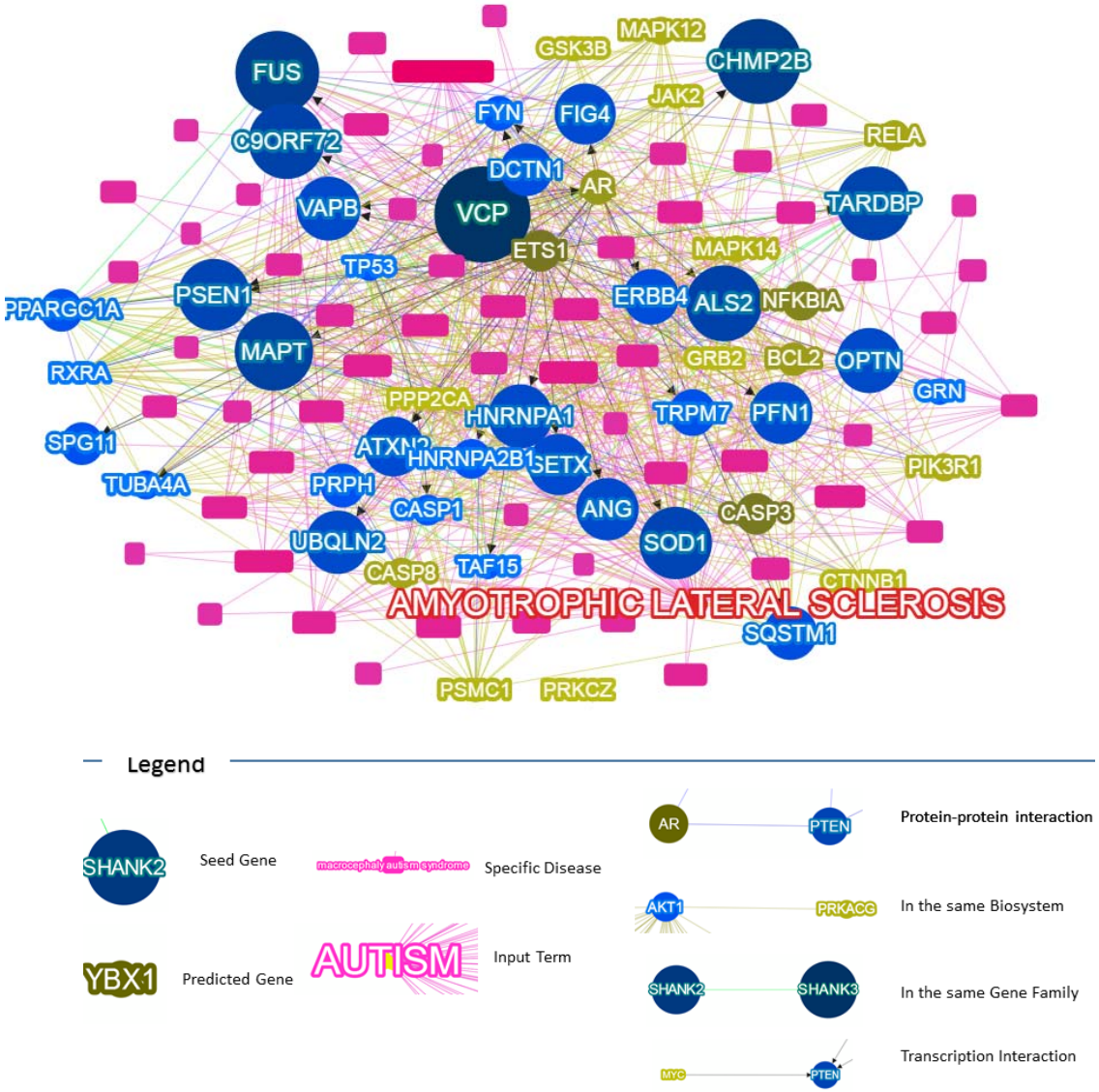
2)Enter the disease/phenotype terms, like ‘Autism’

3)Submit, done!

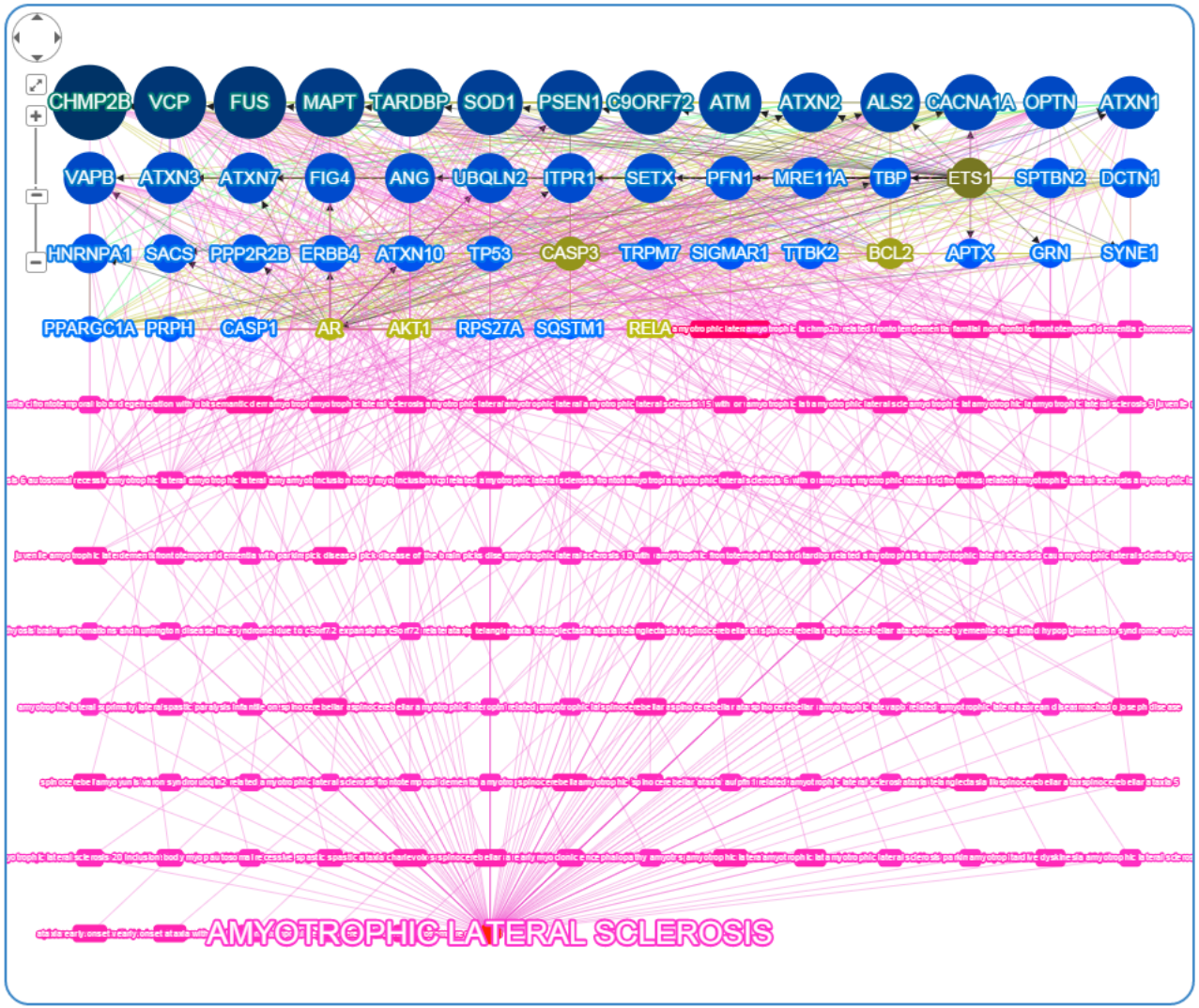
Optional 4)Turn on Word Cloud


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

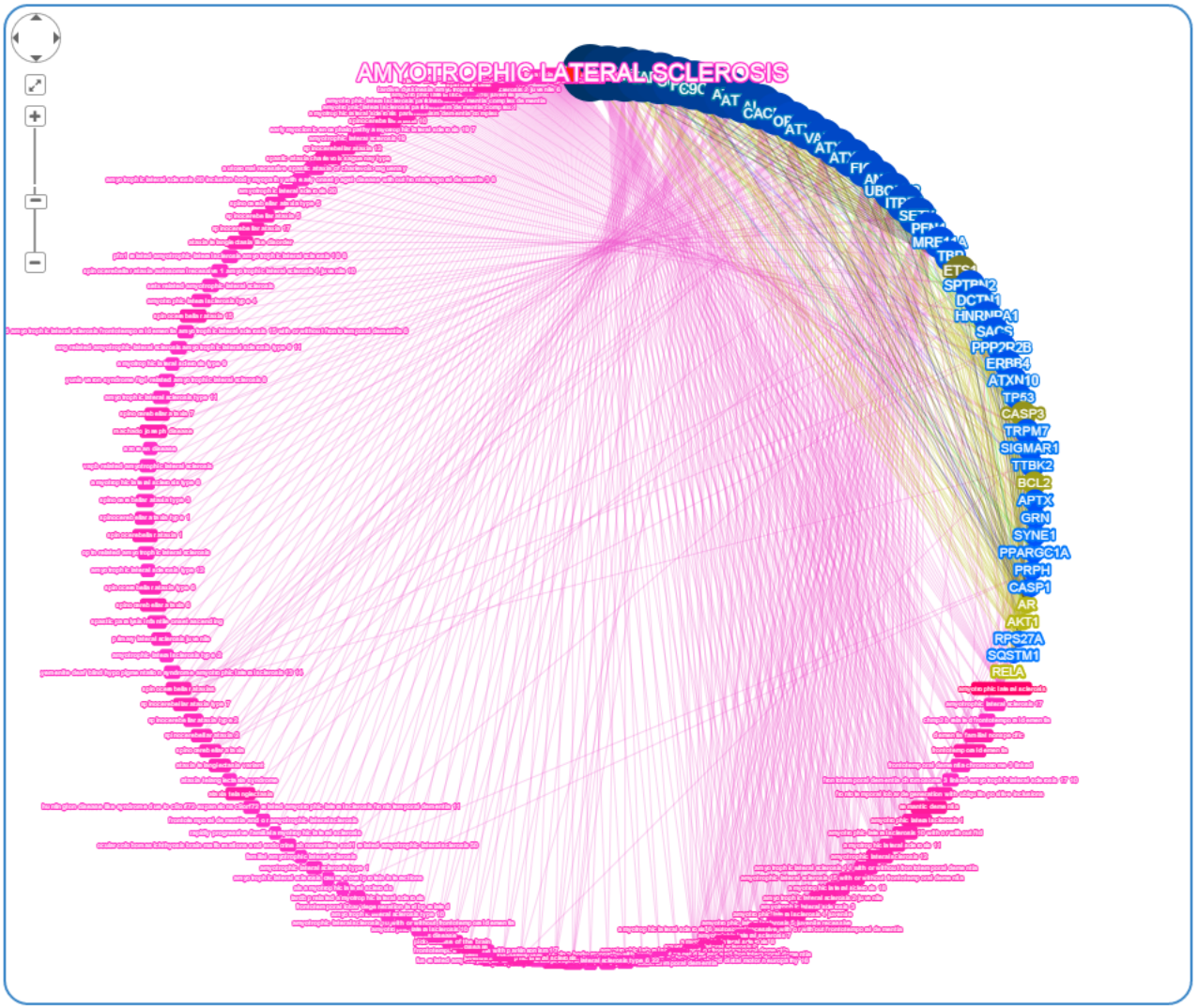
Example output for ALS



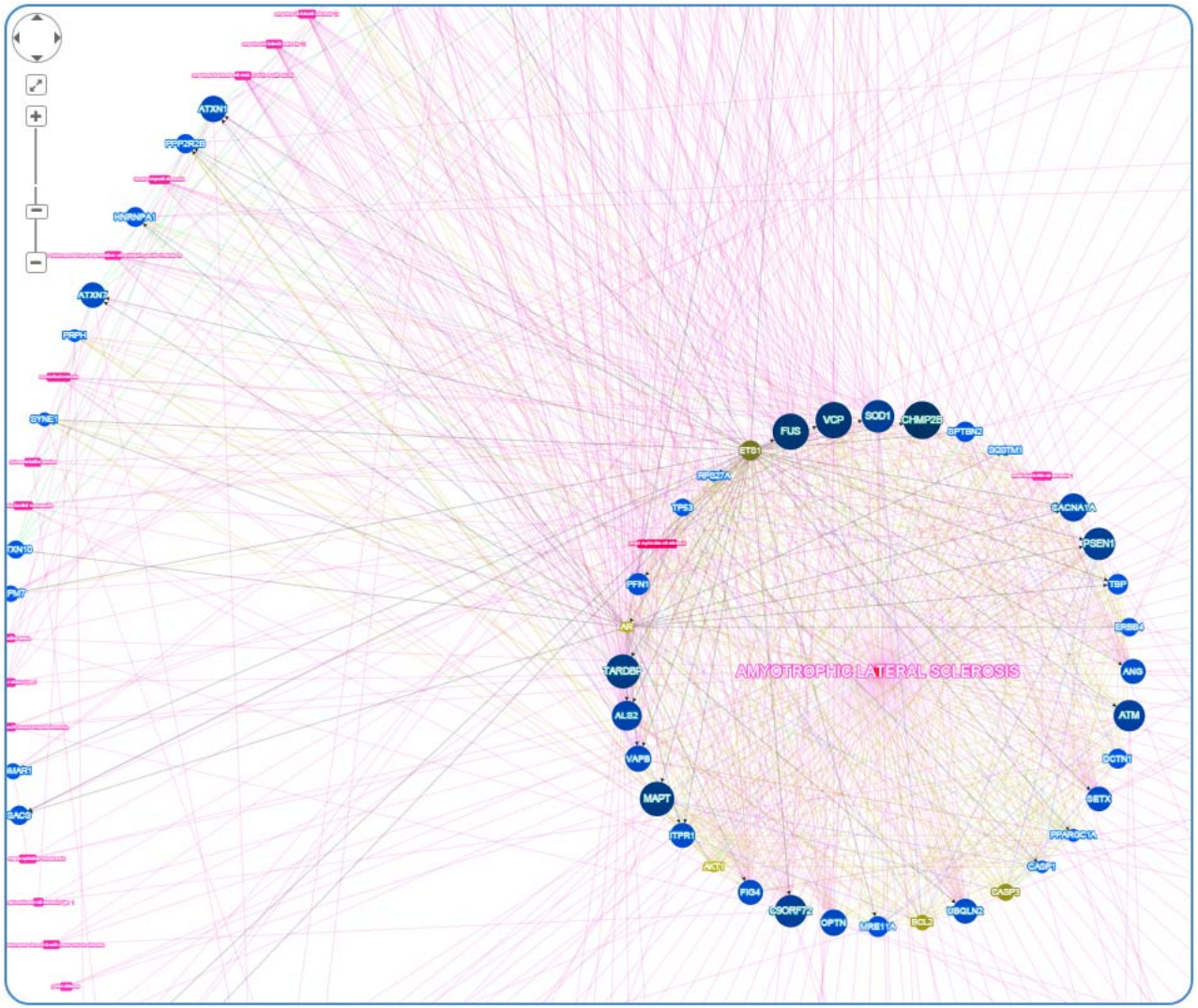
Different layouts for ‘top gene list’



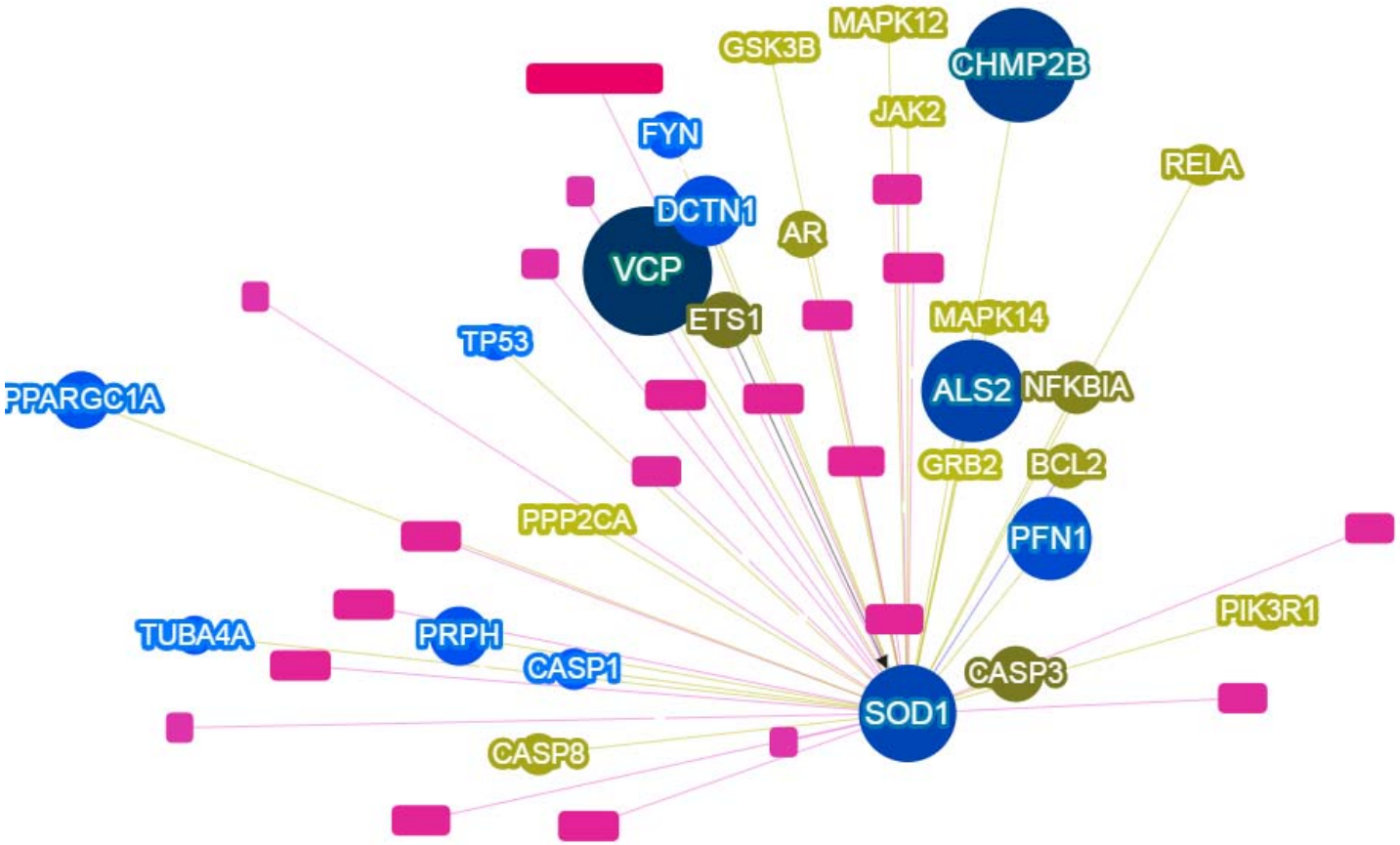
Different layouts for ‘top gene list’



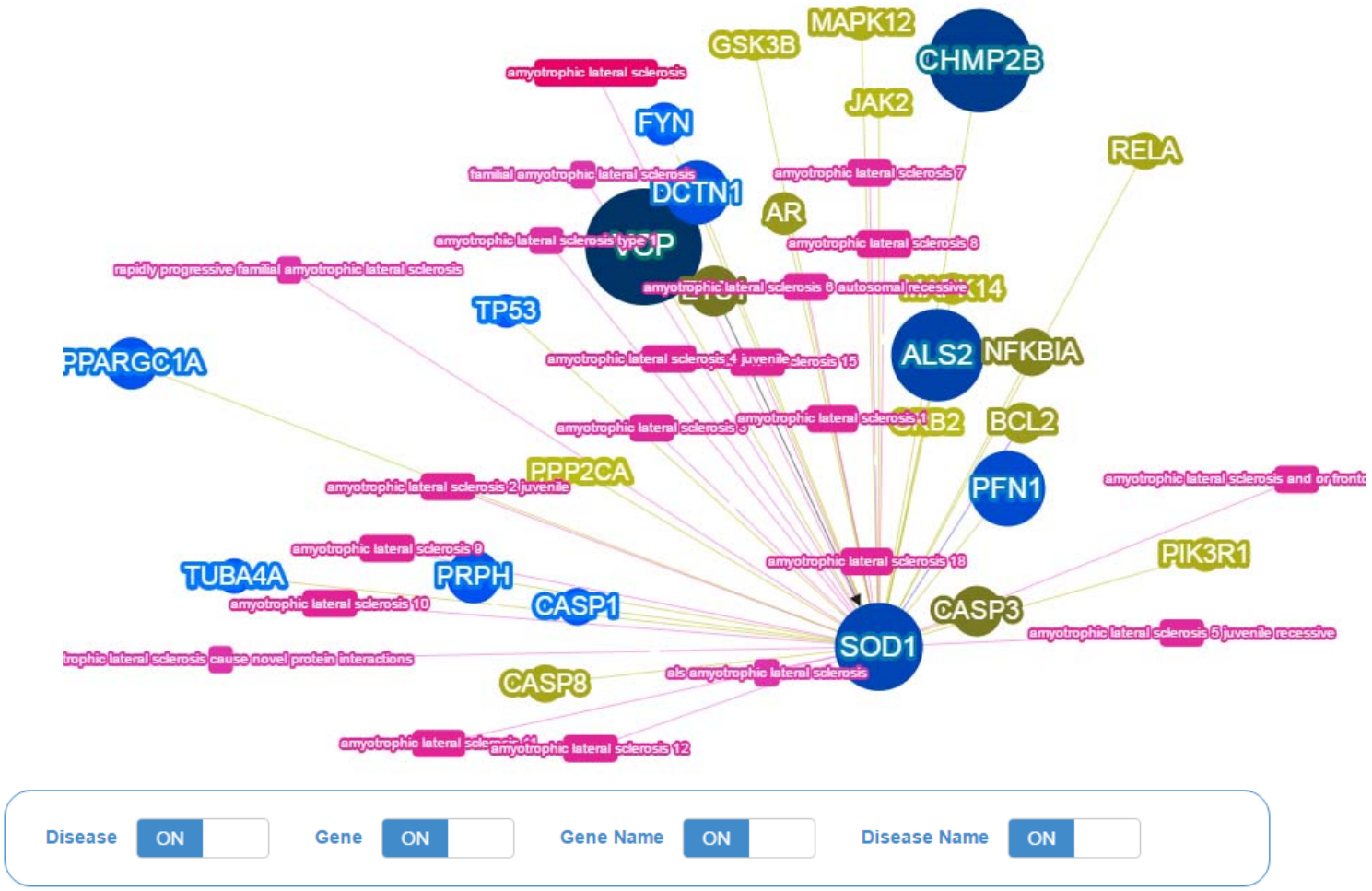
Different layouts for ‘top gene list’

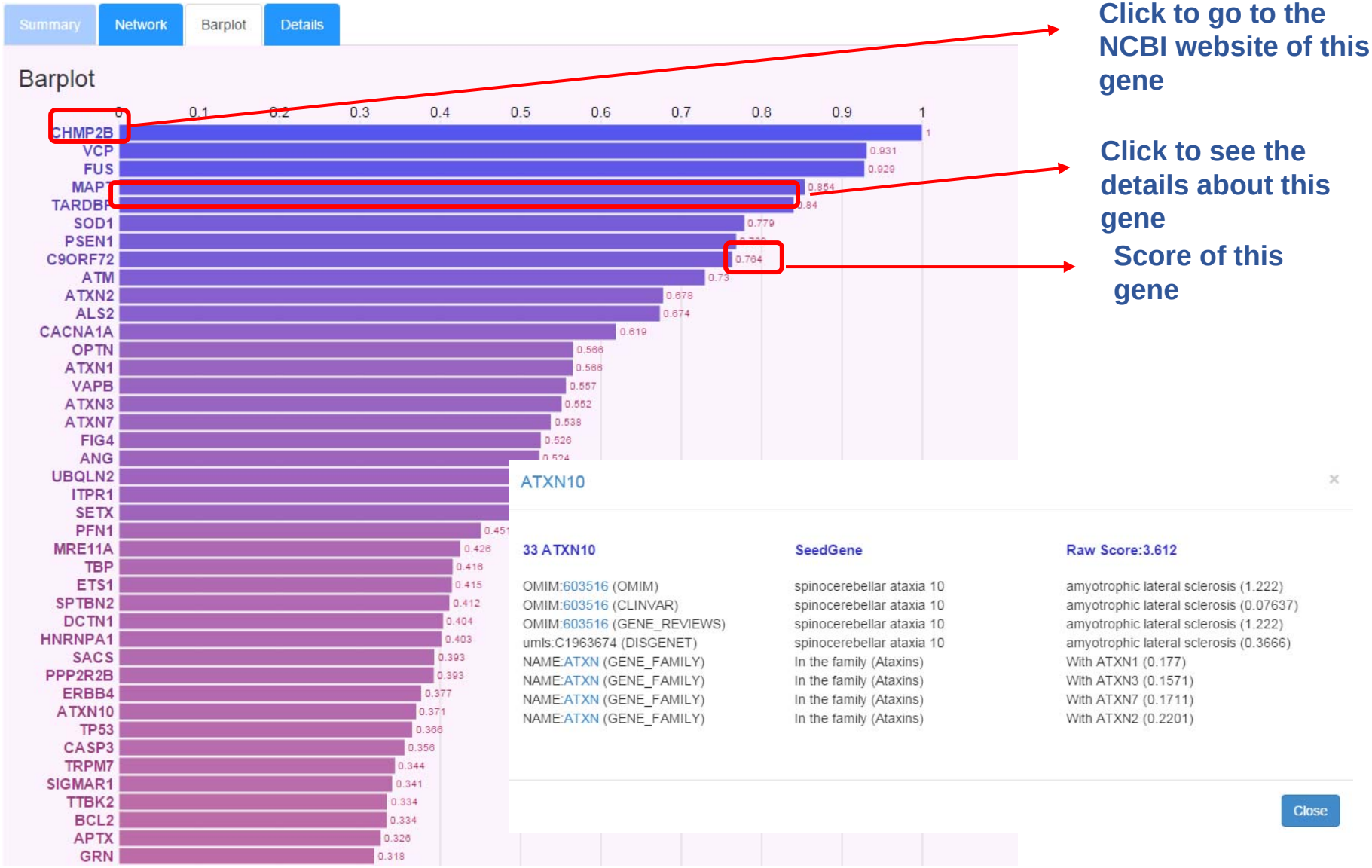


Only connections for a specific gene



Turn on all labels



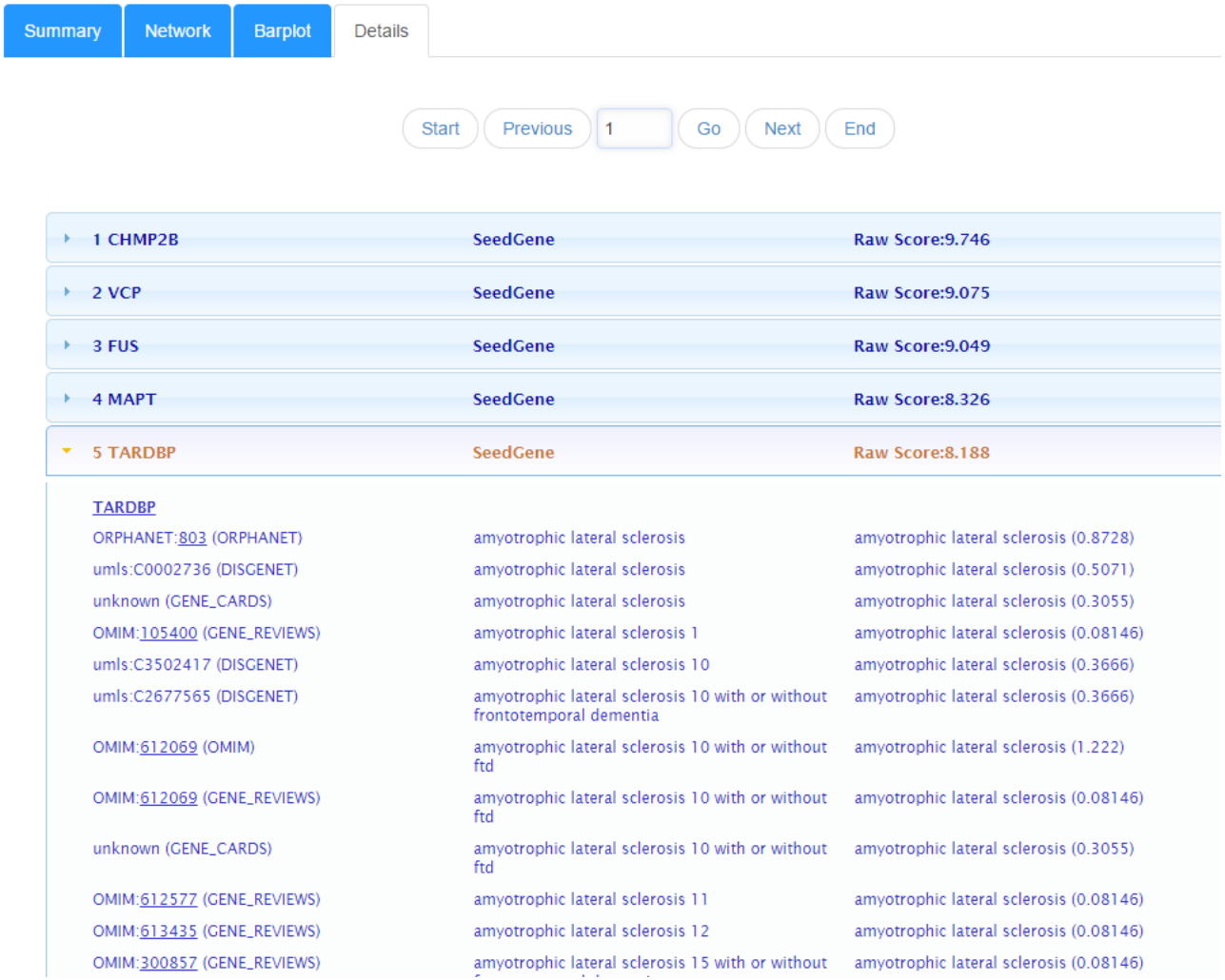


Click to go to the NCBI website of this gene

Click to see the details about this gene

Score of this gene

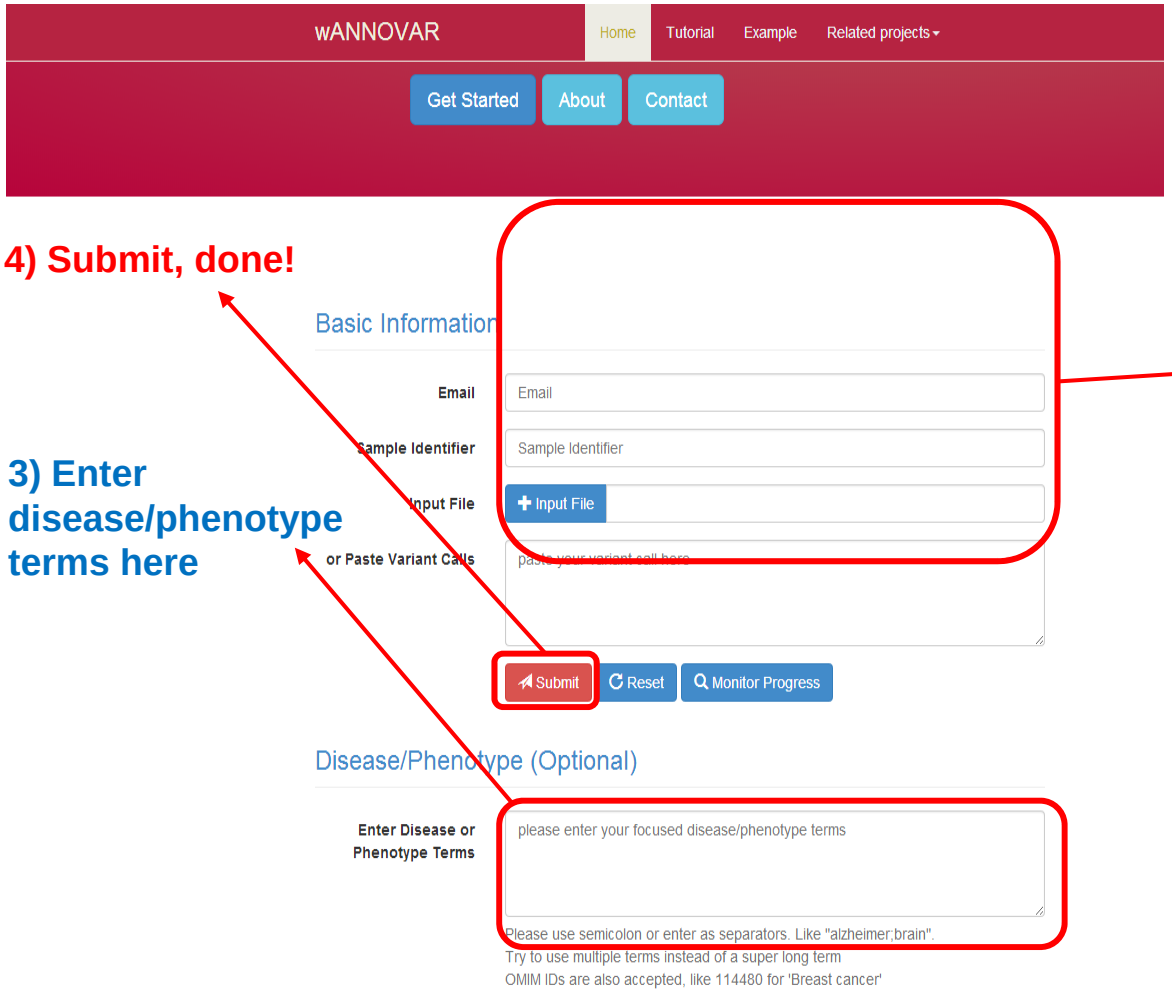
Detailed examination of the evidence



Integration with wANNOVAR

http://wannovar.wglab.org

1) Enter wANNOVAR website address



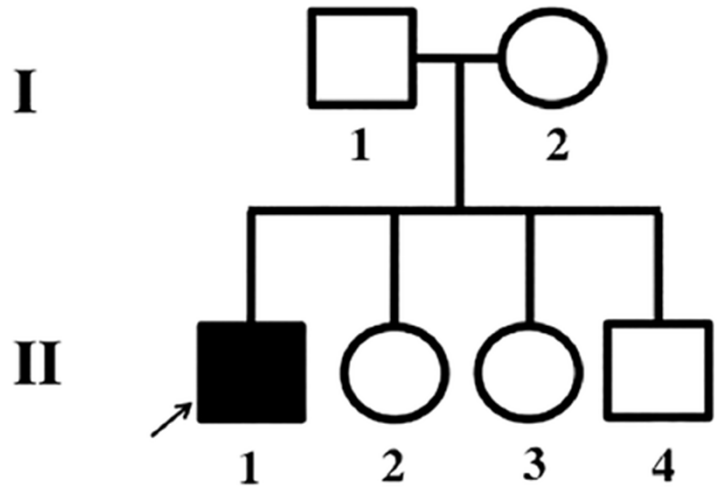
2) Enter variant file, email and all the information wANNOVAR requires.

3) Enter disease/phenotype terms here

4) Submit, done!

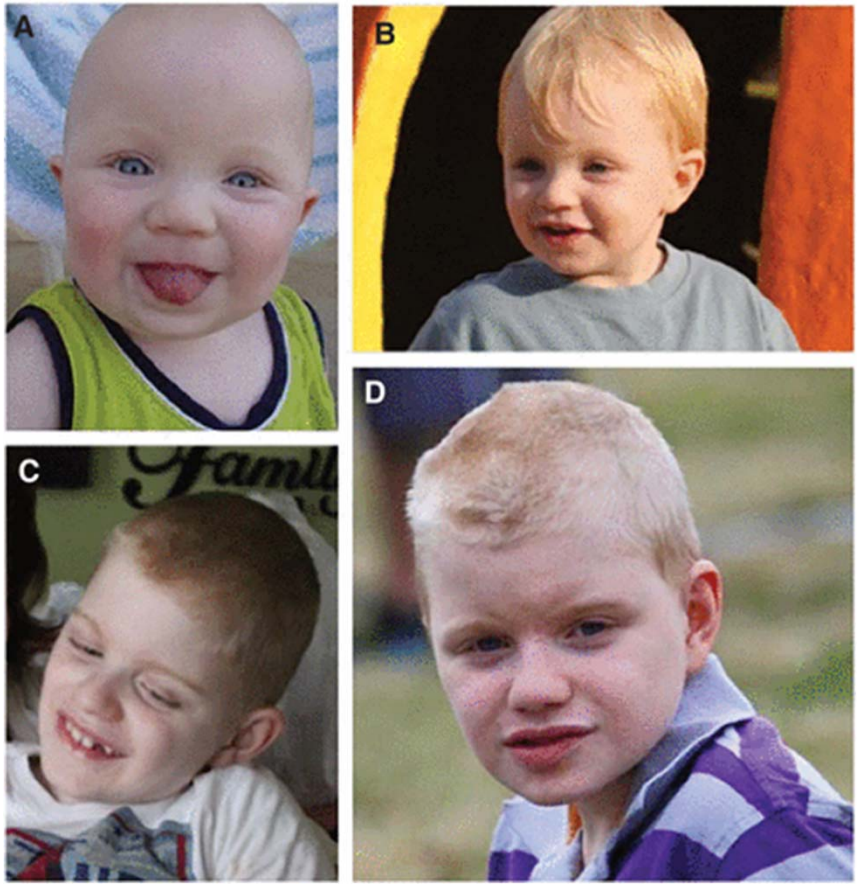
Case study: how to use Phenolyzer in practice?

- 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.

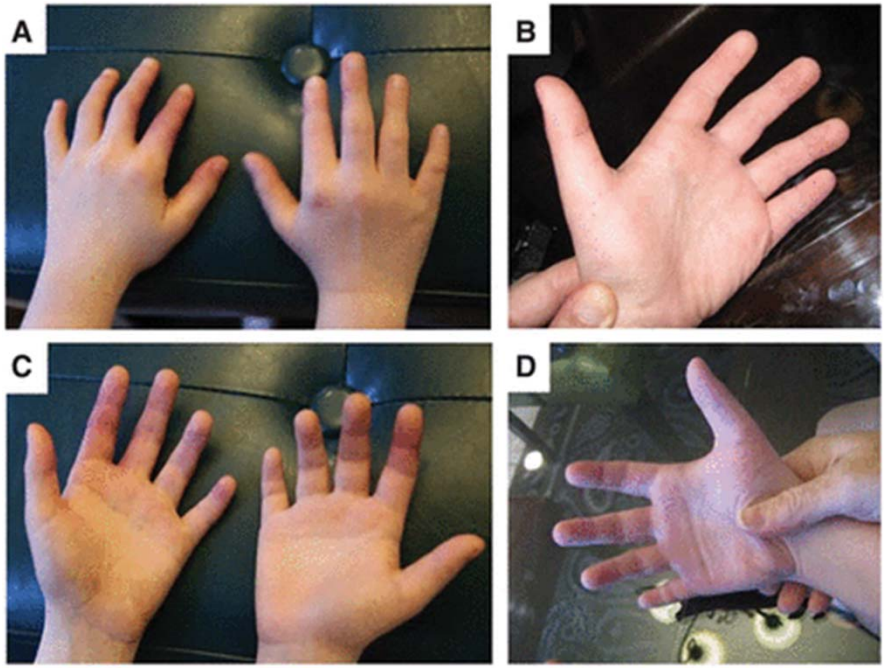


Kleyner et al, Mol Case Studies, 2:a001131, 2016

Phenotype features of an undiagnosed case



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



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

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

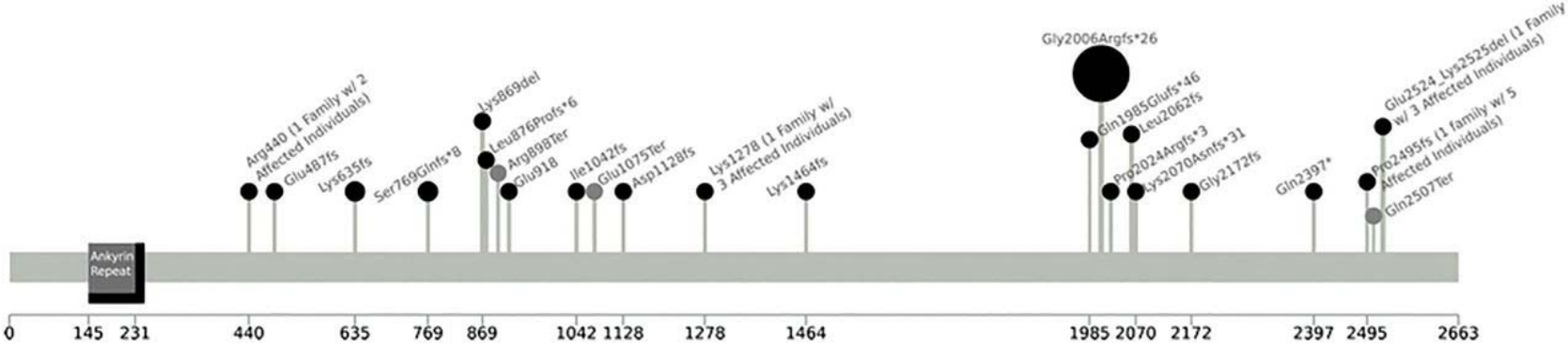
- Sequencing performed on Ion Proton with AmpliSeq Exome panel
- With standard analytical protocol, more than 1000 variants were recognized as *de novo*, well above the expected number of *de novo* mutations, suggesting low quality of sequencing and variant calling

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

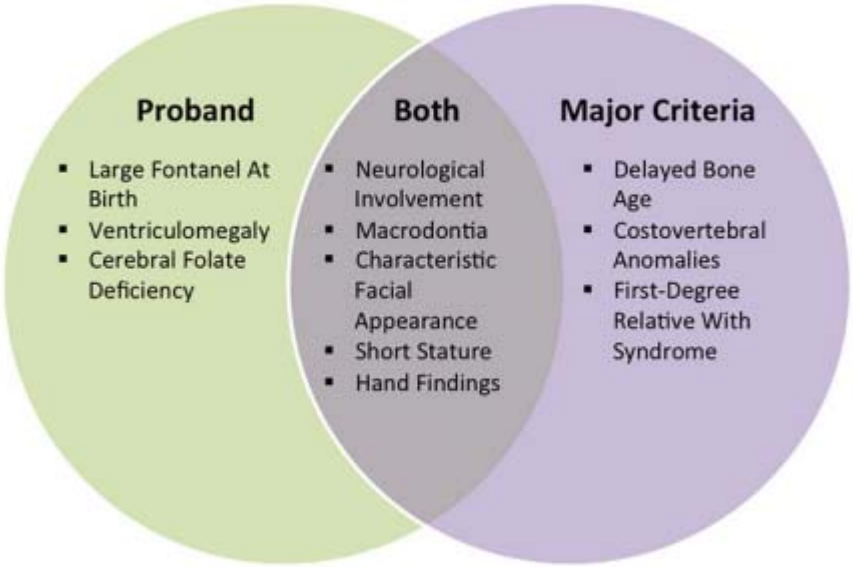
Phenolyzer+wANNOVAR joint analysis

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



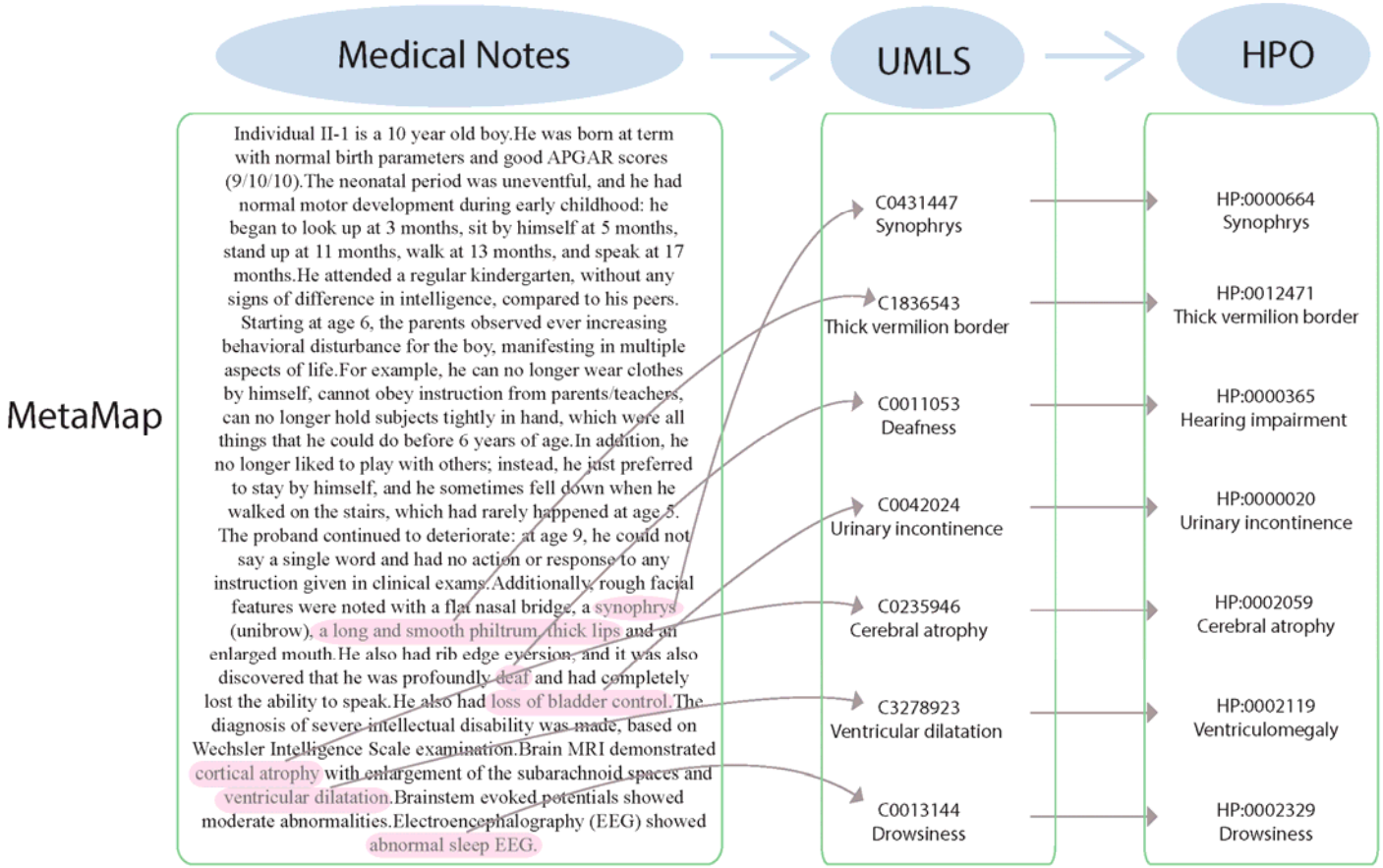
Molecular diagnosis of KBG syndrome

- 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

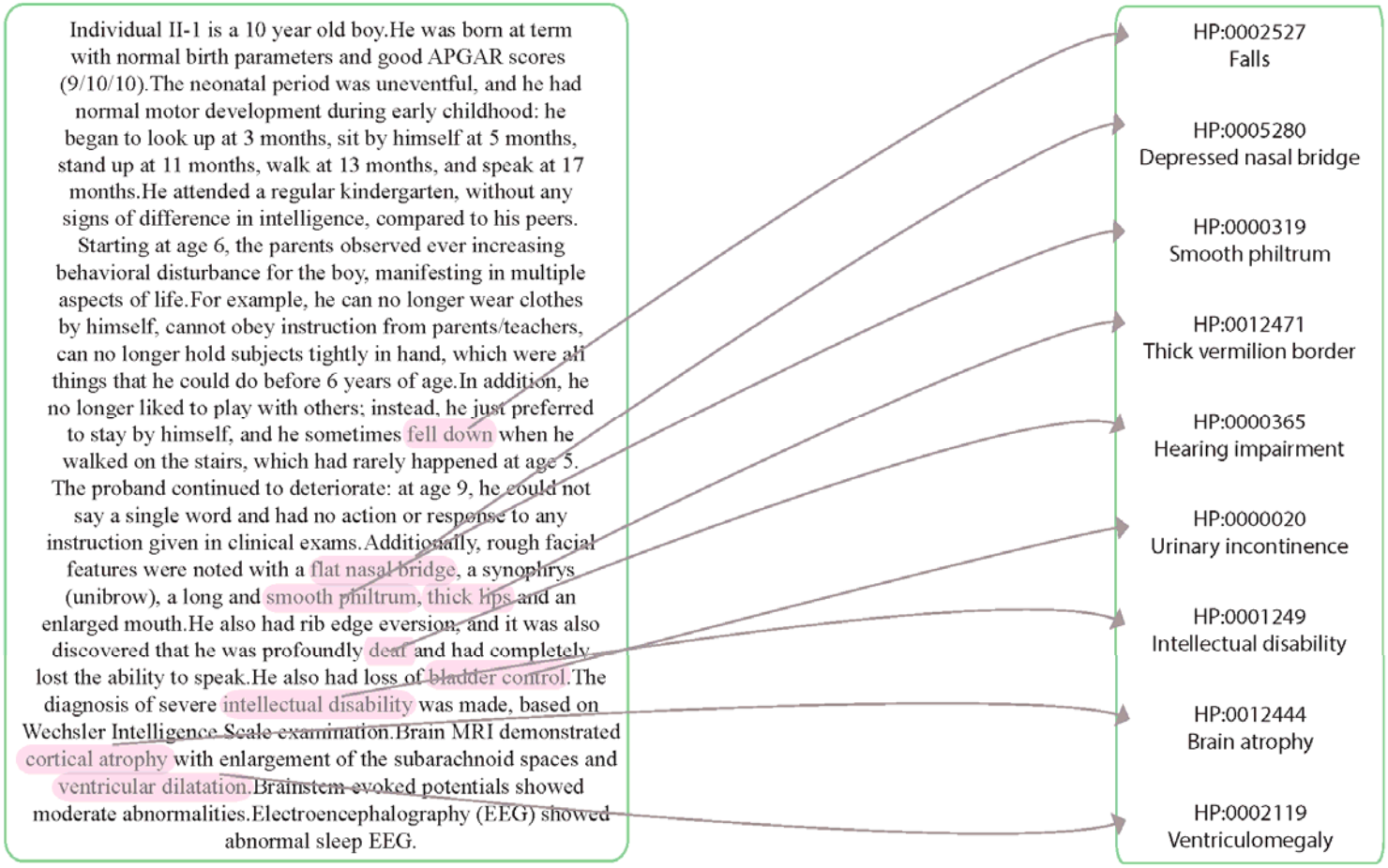
Computational generation of HPO terms?



Son et al, Am J Hum Genet, 2018

Computational generation of HPO terms?

MedLEE



Son et al, Am J Hum Genet, 2018

Go back to case study #1

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