VARIANT IDENTIFICATION AND INTERPRETATION FROM NEXT GENERATION SEQUENCING

Bioinformatics Unit Seminar Series
Tuesday, May 28th

Seth I. Berger, MD, PhD
Assistant Professor
Center for Genetic Medicine Research, Children's National Medical Center
Rare Disease Institute, Division of Genetics and Metabolism, Children's National Medical Center
Department of Genomics and Precision Medicine; and of Pediatrics, GWU SMHS

Objectives

- Overview of Next Generation Sequencing
- Variant Identification Tools and Pipeline
- Variant Interpretation

Next Generation Sequencing

- **Next-generation sequencing** refers to non-Sanger-based high-throughput DNA sequencing technologies.
- Millions or billions of DNA strands can be sequenced in parallel

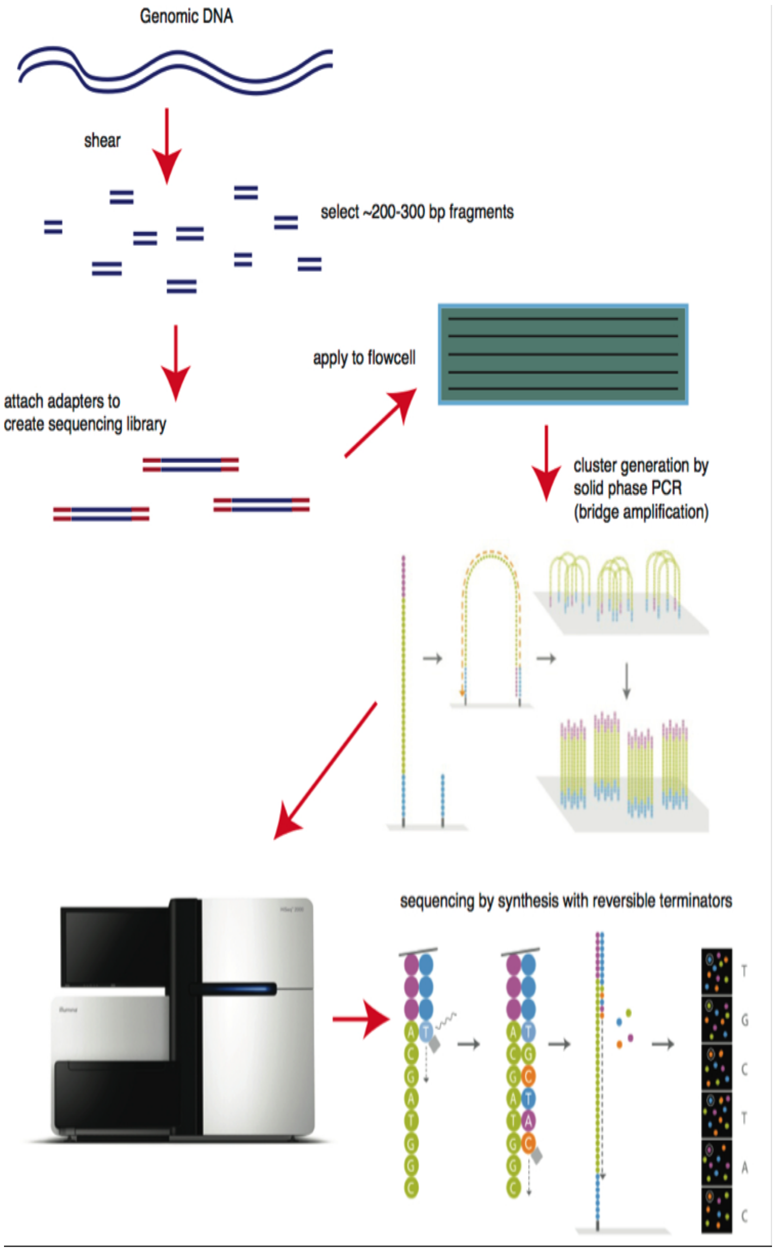
- More throughput and minimizing the need for the fragment-cloning methods that are often used in Sanger sequencing of genomes

Sequencing

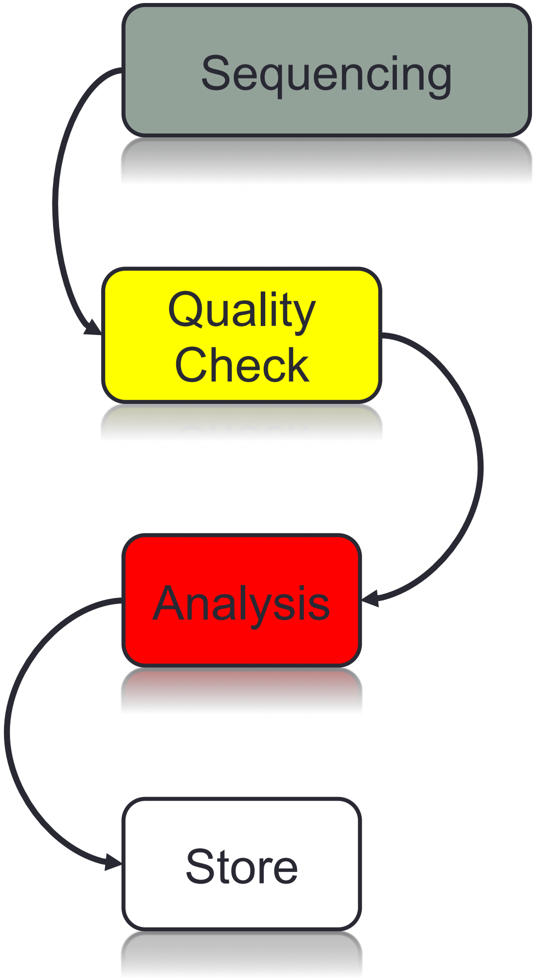
Purpose: Sequence bases identification

4 Steps:

- ✓ Library preparation
- ✓ Cluster generation
- ✓ Sequencing
- ✓ Data analysis


Genomic DNA
shear
select ~200-300 bp fragments
attach adapters to create sequencing library
apply to flowcell
cluster generation by solid phase PCR (bridge amplification)
sequencing by synthesis with reversible terminators
T
G
C
T
A
C

Bioinformatic Workflow



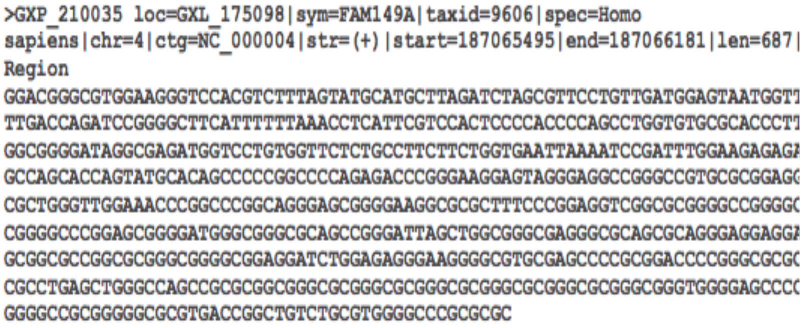
Definitions

- Reads: inferred sequence of base pairs corresponding to a part or a whole DNA
- Raw data: Formatted Sequence of DNA/RNA/AA
- Paired ends: Forward and reverse read of same cluster
- Pipeline: structured execution commands

Files description

FastA format

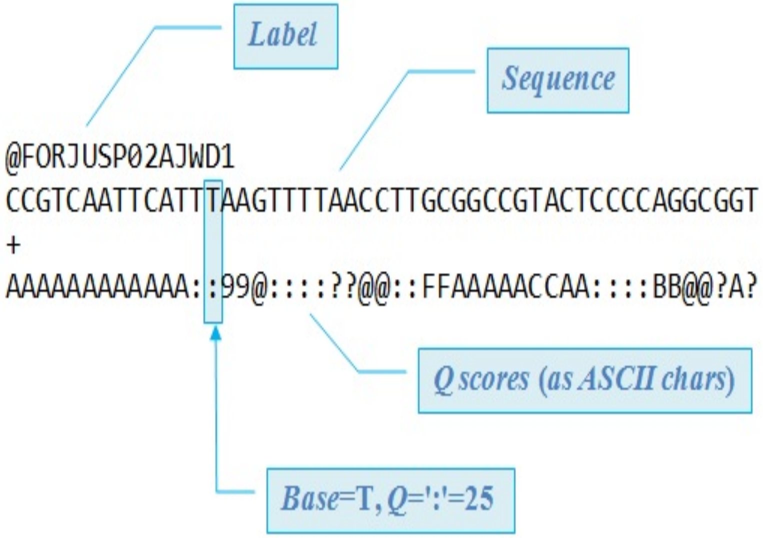
- Header
- Sequence



```
>GXP_210035 loc=GXL_175098|sym=FAM149A|taxid=9606|spec=Homo
sapiens|chr=4|ctg=NC_000004|str=(+)|start=187065495|end=187066181|len=687|
Region
GGACGGGCGTGGAAGGGTCCACGTCTTTAGTATGCATGCTTAGATCTAGCGTTCCTGTTGATGGAGTAATGGTT
TTGACCAGATCCGGGGCTTCATTTTTTAAACCTCATTCGTCCACTCCCCACCCCAGCCTGGTGTGCGCACCCTT
GGCGGGGATAGGCGAGATGGTCCTGTGGTTCTCTGCCTTCTTCTGGTGAATTAAAATCCGATTTGGAAGAGAGA
GCCAGCACCAGTATGCACAGCCCCCGGCCCCAGAGACCCGGGAAGGAGTAGGGAGGCCGGGCCGTGCGCGGAGG
CGCTGGGTTGGAAACCCGGCCCGGCAGGGAGCGGGGAAGGCGCGCTTTCCCGGAGGTCGGCGCGGGGCCGGGGC
CGGGGCCCGGAGCGGGGATGGGCGGGCGCAGCCGGGATTAGCTGGCGGGCGAGGGCGCAGCGCAGGGAGGAGGA
GCGGCGCCGGCGCGGGCGGGGCGGAGGATCTGGAGAGGGAAGGGGCGTGCGAGCCCCGCGGACCCCGGGCGCGC
CGCCTGAGCTGGGCCAGCCGCGCGGCGGGCGCGGGCGCGGGCGCGGGCGCGGGCGCGGGCGGGTGGGGAGCCCC
GGGGCCGCGGGGGCGCGTGACCGGCTGTCTGCGTGGGGCCCGCGCGC
```

FastQ format

- Header
- Quality score
- Sequence



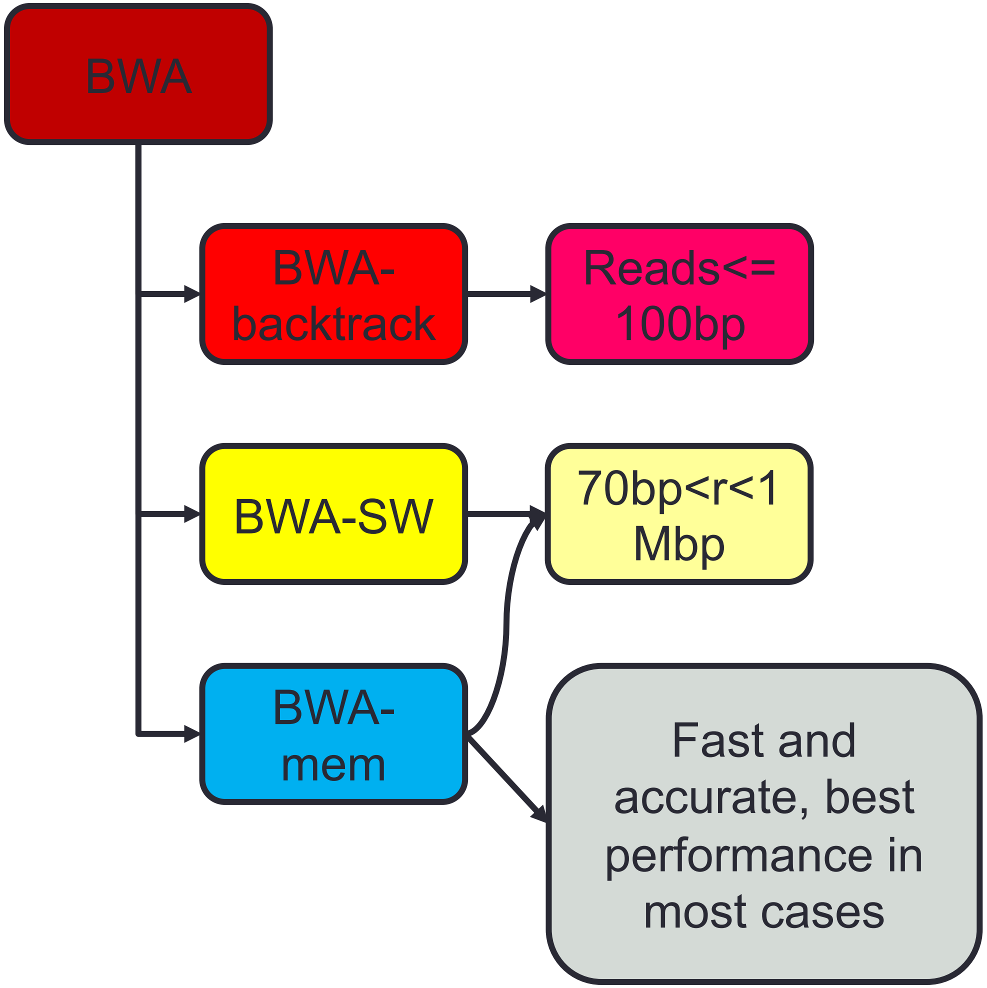
```
@FORJUSP02AJWD1
CCGTCAATTCATTTAAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA::99@::::??@@::FFAAAAACCAA::::BB@@?A?
```

Alignment

Burrows-Wheeler Aligner (BWA)

Map low-divergent sequences against a large reference genome.

- ✓ SOLiD
- ✓ Illumina
- ✓ IonTorrent
- ✓ Sanger
- ✓ Assembly contigs
- ✓ PacBio
- ✓ BACSeq
- ✓ 454



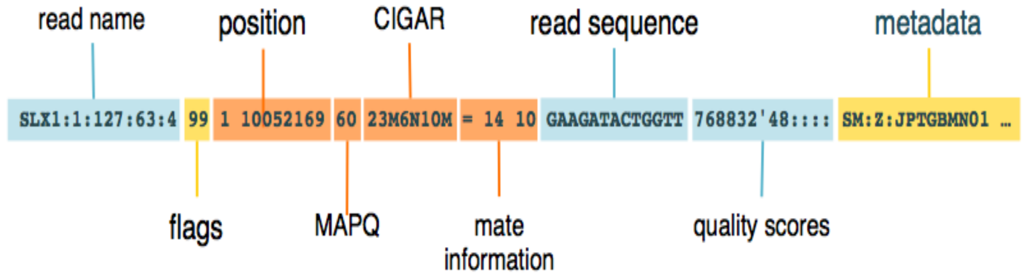
Files description

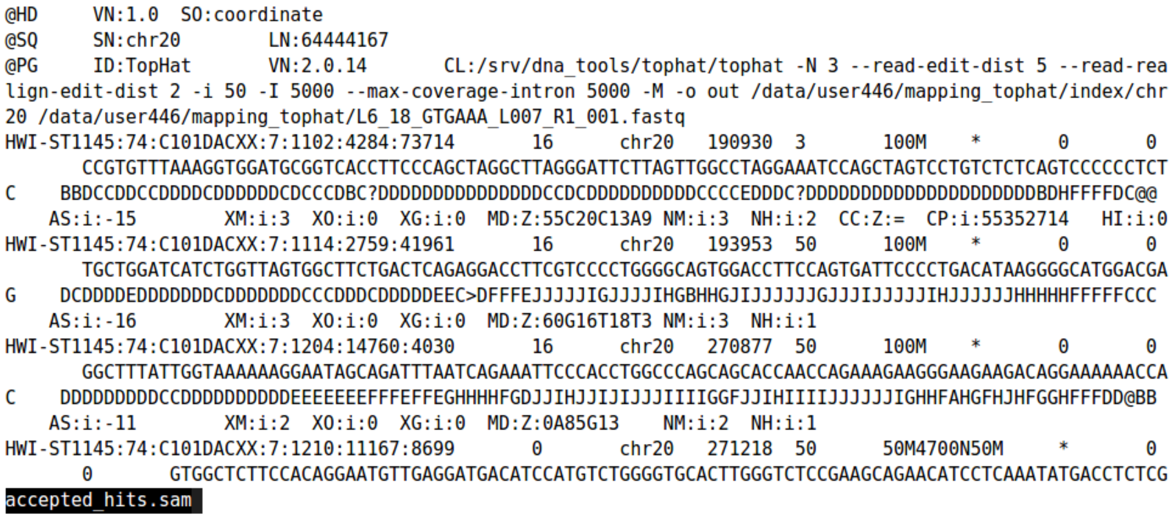
Sequence Alignment Map format (SAM)

- Alignment
- Quality Scores
- Read
- Sample information
- → Data compression
  - BAM
  - CRAM

HEADER containing metadata (sequence dictionary, read group definitions etc)

RECORDS containing structured read information (1 line per read record)



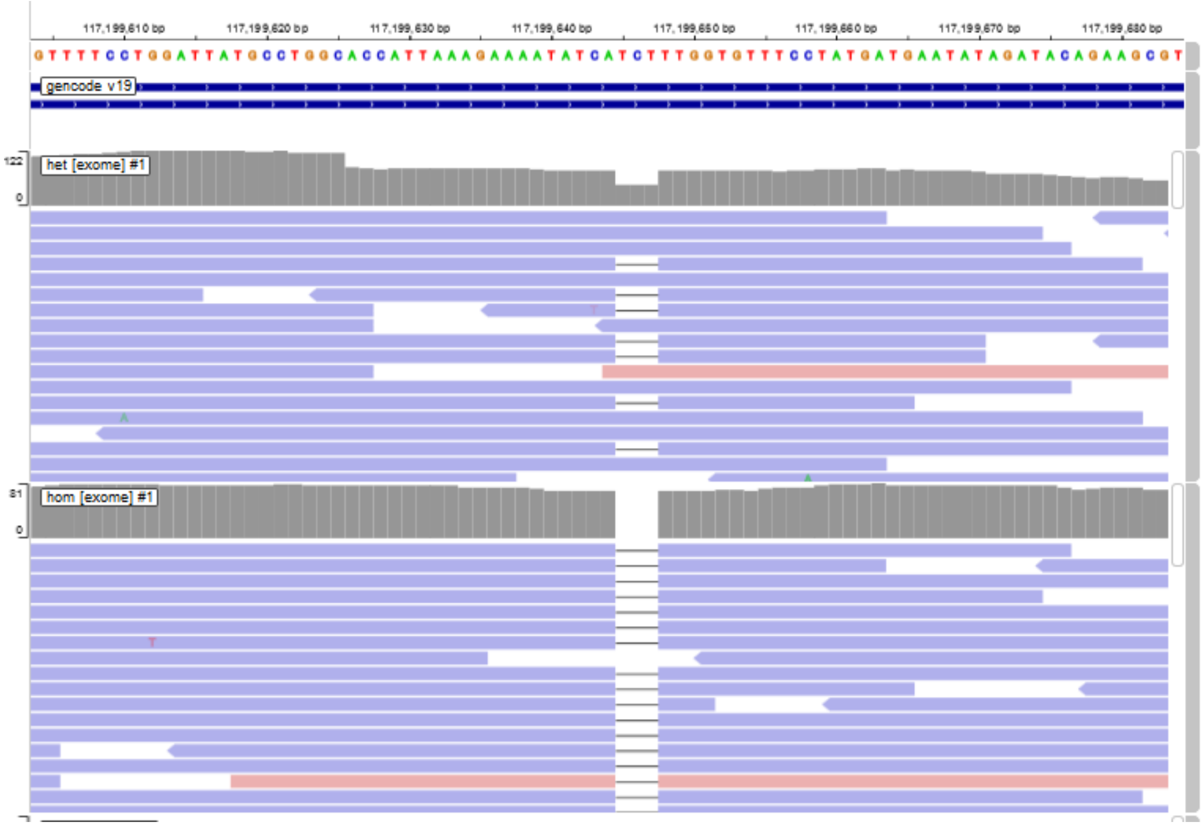


```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714       16      chr20   190930  3       100M    *       0       0
       CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C     BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961       16      chr20   193953  50      100M    *       0       0
       TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G     DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030       16      chr20   270877  50      100M    *       0       0
       GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C     DDDDDDDDDCCDDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
    AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699       0       chr20   271218  50      50M4700N50M     *       0
       0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
accepted_hits.sam
```

Variant Pileups

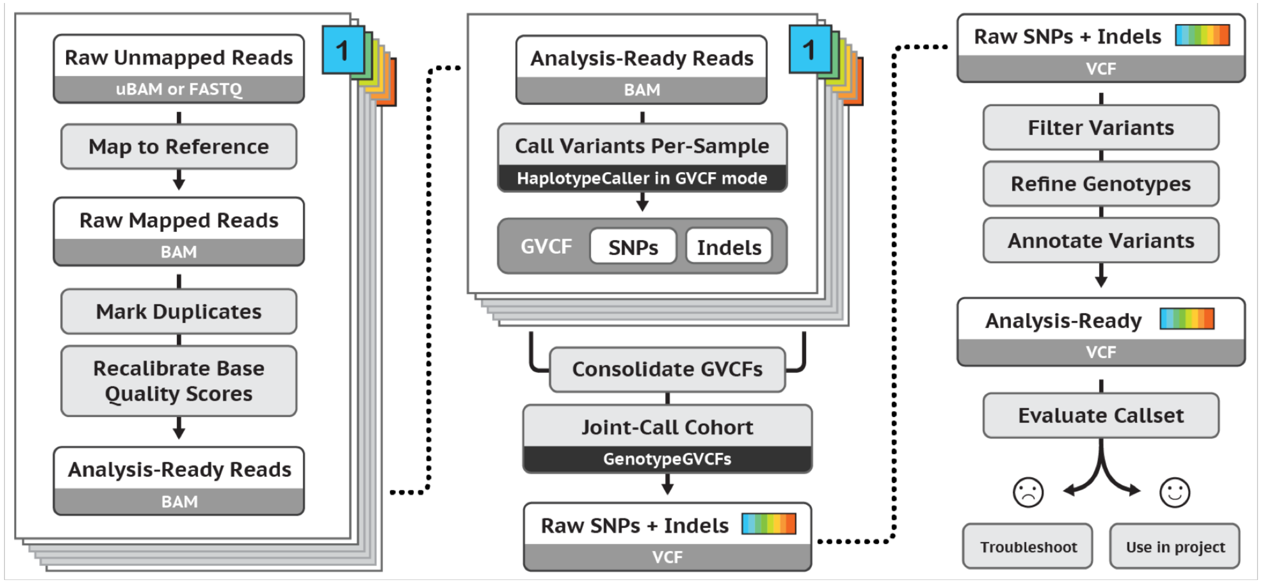
Viewed in IVG

Example heterozygous and homozygous 3 base pair deletion



Variant Calling

- Statistically call variants from aligned sequences
- GATK Haplotype Caller, Platypus, etc
- Calls SNVs and Indels



Other Types of Variant Calling

- Most clinical pipelines are optimized for germline variants
- Other specific variant callers can be integrated into pipeline

- Copy Number Variants
  - Depth of coverage (XHMM, etc)
  - Aberrant Mate Pair Insertion size (Mseq)
- Translocation
  - Aberrant Mate pair distribution (Breakdancer, etc)
- Mosaic Variants
  - LoFreq, Mutect2, etc

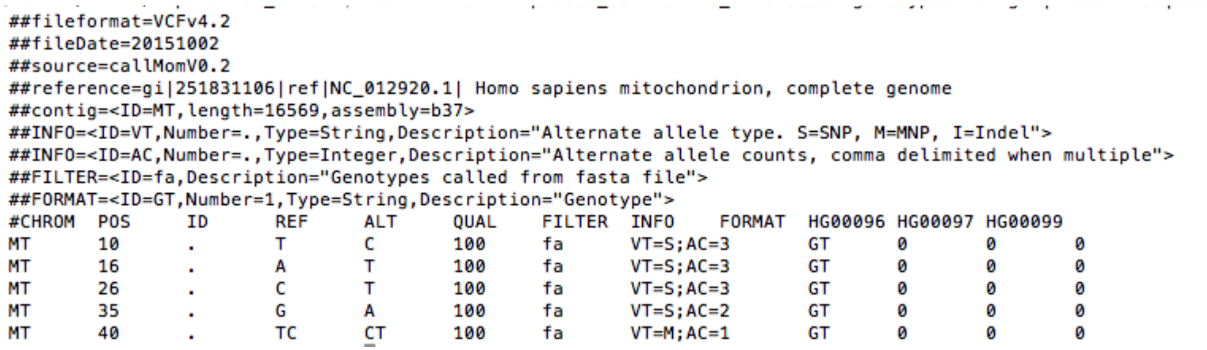
File format description: Summary

Formats	Input for	Output from
FastA	BLAST Multiple Sequence Alignment (MSA) Database query tools	Older Sequencers Sequence Database store Converters
FastQ	FasQC Aligners Assemblers Variants detection/SNP callers	Sequencers (default format) Format Converters
SAM/BAM/CRAM	Alignment algorythms Some Assemblers Alignment viewers Variants detection	Aligners Assemblers
VCF	VCF tools SNP Annotation	SNP caller Haplotyping Software Variant Information Database
BED/BEDGrap	Alignment viewers Bed tools Some annotation	Features detection

Files description

Variant Calling Format (VCF)

- Shows variants, not genetic features.
- Can have additional per variant and per sample annotations



```
##fileformat=VCFv4.2
##fileDate=20151002
##source=callMomV0.2
##reference=gi|251831106|ref|NC_012920.1| Homo sapiens mitochondrion, complete genome
##contig=<ID=MT,length=16569,assembly=b37>
##INFO=<ID=VT,Number=.,Type=String,Description="Alternate allele type. S=SNP, M=MNP, I=Indel">
##INFO=<ID=AC,Number=.,Type=Integer,Description="Alternate allele counts, comma delimited when multiple">
##FILTER=<ID=fa,Description="Genotypes called from fasta file">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  HG00096 HG00097 HG00099
MT      10      .       T       C       100     fa      VT=S;AC=3       GT      0       0       0
MT      16      .       A       T       100     fa      VT=S;AC=3       GT      0       0       0
MT      26      .       C       T       100     fa      VT=S;AC=3       GT      0       0       0
MT      35      .       G       A       100     fa      VT=S;AC=2       GT      0       0       0
MT      40      .       TC      CT      100     fa      VT=M;AC=1       GT      0       0       0
```

Annotation

ANNOtate VARiants (ANNOVAR)

- ✓Program for functional annotation.
- ✓Can identify Structure variants and examine functional consequences on gene.
- ✓Disease identification.
- ✓All species.

SNP Effect (SNPEff)

- ✓Program for annotating and predicting the effects of SNP.
- ✓Annotate thousands variants/sec.
- ✓Based on the genomic position.
- ✓Many species.

Objectives

- Overview of Next Generation Sequencing
- Variant Identification Tools and Pipeline
- Variant Interpretation

Clinical Variant Interpretation

- Clinical versus Research Variant Interpretation
- Cautions about variant interpretation
- ACMG Reportable Secondary Finding Genes
- ACMG Variant Classification Guidelines

Clinical vs Research Variant Interpretation

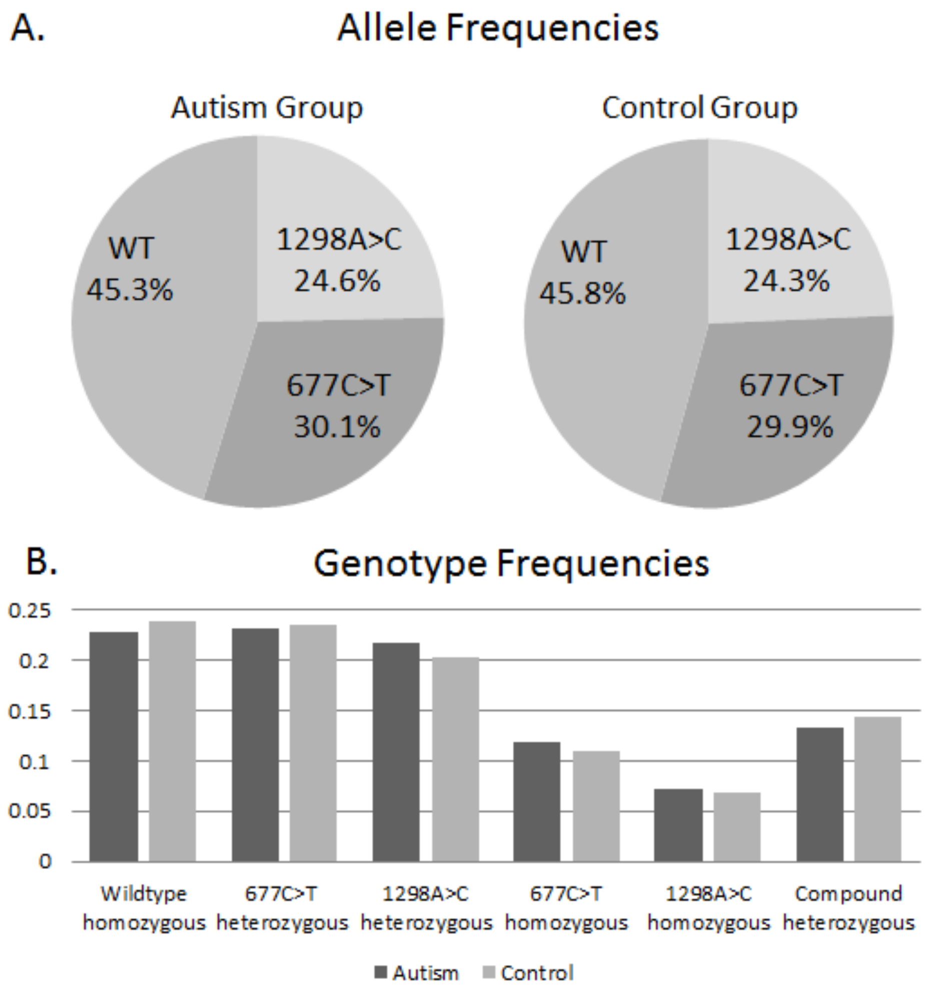
- Research Variant Interpretation
  - Association of variants to Phenotype
    - Note: GWAS variants are not necessarily disease causing, typically linked to a truly pathogenic variant
  - Functional effects of variants
  - Molecular function, pathways, structure, localization, etc…

- Clinical Variant Interpretation
  - Does this variant cause a phenotype?
  - Does this variant change what the doctor does clinically?
  - Is more testing needed for this patient or does variant provide an answer?
  - Can family make reproductive decisions based on variant?

Cautions about variant interpretation

- Proving a variant causes functional changes in a protein does not mean it is clinically relevant.
  - Models may not always be able to recapitulate a phenotype.
  - Example: MTHFR: 1298A>C
    - Originally studied as possible contribution to thrombosis
    - Functional studies showed variants with somewhat decreases/altered activity
    - Multiple conflicting studies linking to many different things from autism to cancer.
    - Found to be very very common in population at large
    - Some families refusing vaccination based on this or spending money on expensive supplements without known clinical benefit…

MTHFR Common Polymorphism Frequencies at CNMC



What variants get reported clinically?

- Varies from lab to lab

- Depends on test ordered:
  - Targeted gene tests/panel
  - Untargeted tests (exomes, genomes)
  - Prenatal / Postnatal
  - Screening vs. Testing

- Carrier status?

- Secondary findings?

ACMG Reportable Secondary Findings

- ACMG curated list of genes where pathogenic variants in a healthy individual would change medical management
    - Cancer susceptibility genes with early screening options
    - Cardiac arrhythmia genes with medication options/possible defibrillators
    - Serious potential drugs side effects where medication choices may prevent life threatening effects
    - Treatable inborn errors of metabolism with late onset

- Only Pathogenic / Likely Pathogenic variants returnable
- May not report uncertain variants

- Recommends people be offered option of receiving these results even if these are not included on indication for sequencing.

- Some research studies will offer return of these results.

How Clinicians Talk about Variants

- Pathogenic → Most likely causes a disease
  - Describes variant (not genotype)

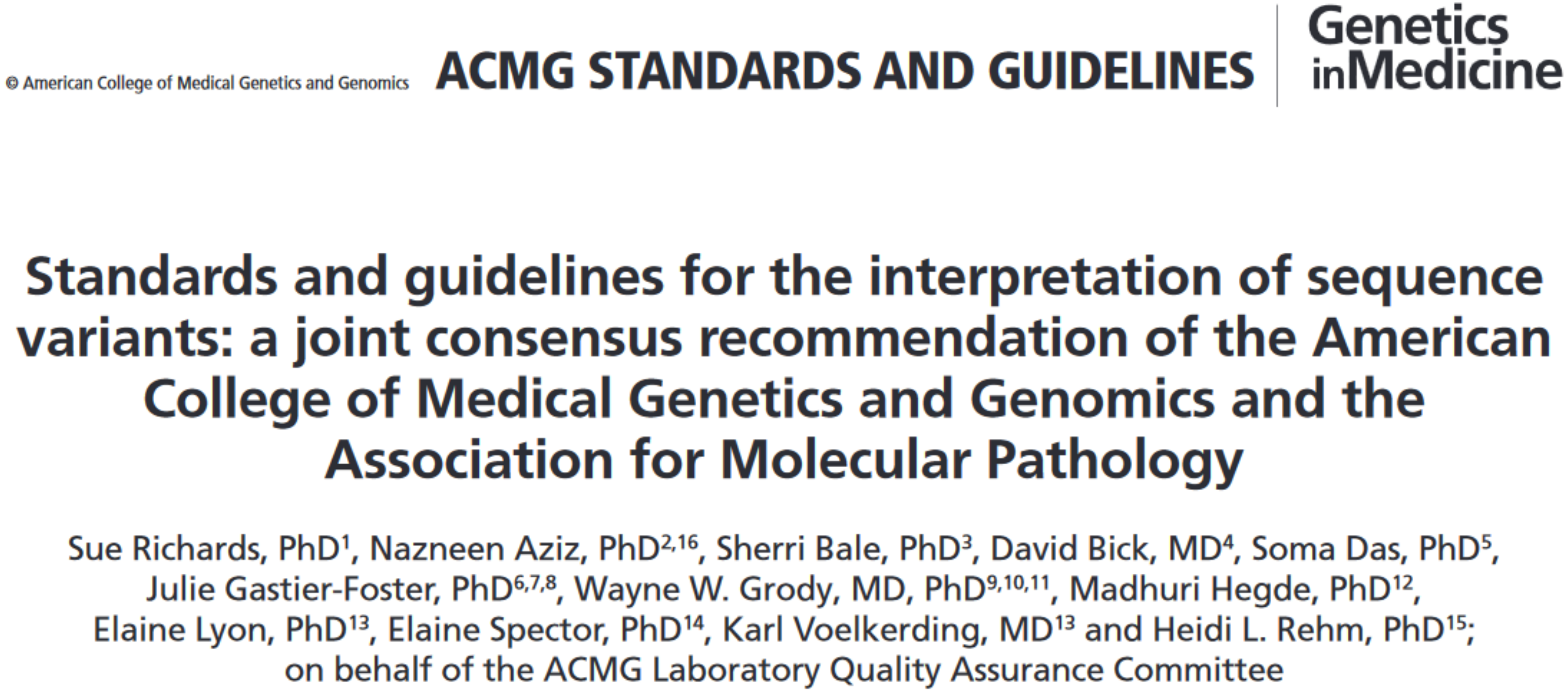
- Likely Pathogenic → Probably related to the disease, but can’t be certain
  - In right clinical setting, can be considered the answer

- Variant of uncertain significance → Not enough evidence to judge
  - Challenging discussions and clinical reasoning

- Likely Benign → No reason to believe it causes a disease, but can’t be sure

- Benign → Too Common to cause disease

ACMG Variant Classification Guidelines

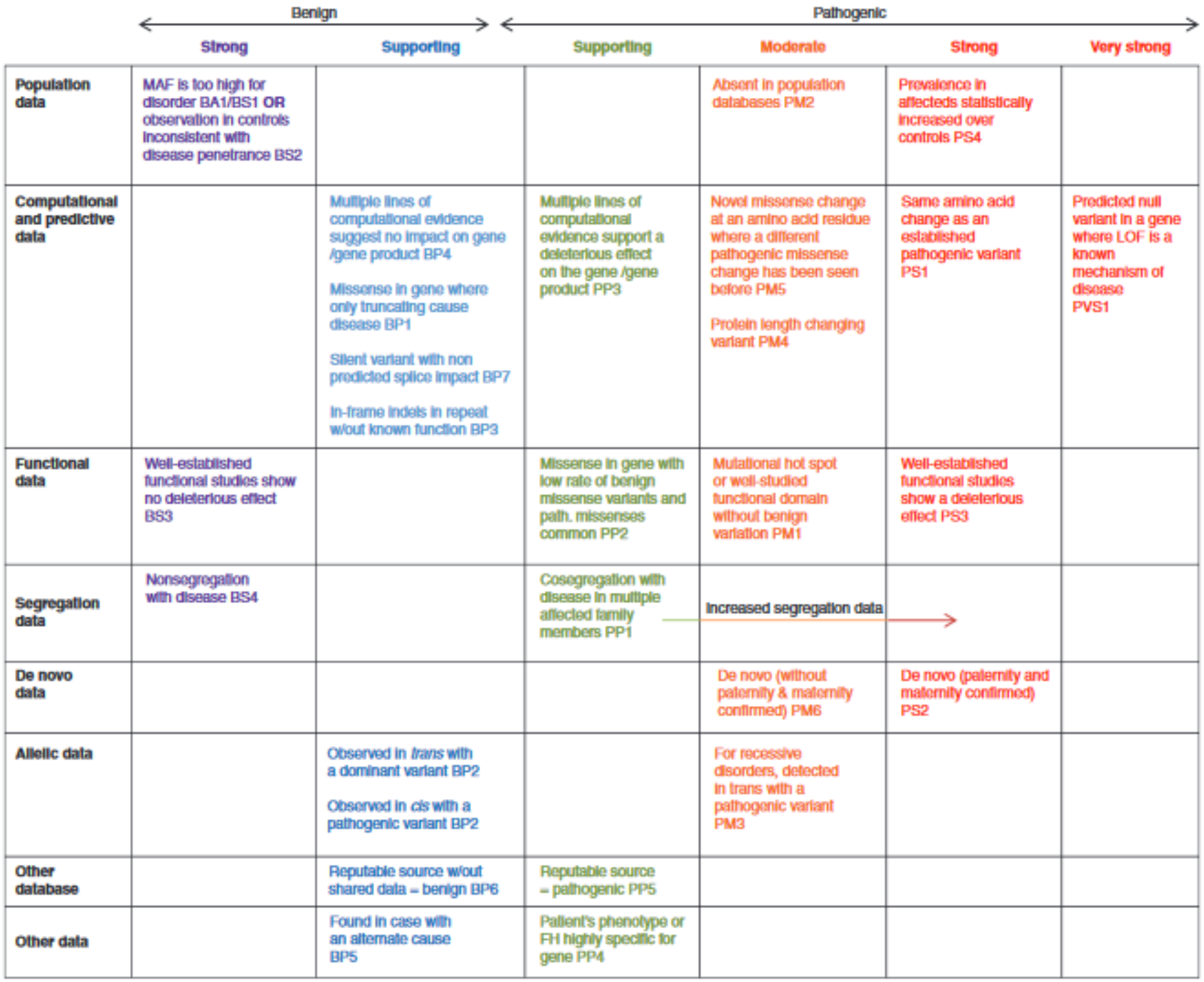


ACMG STANDARDS AND GUIDELINES | Genetics inMedicine

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

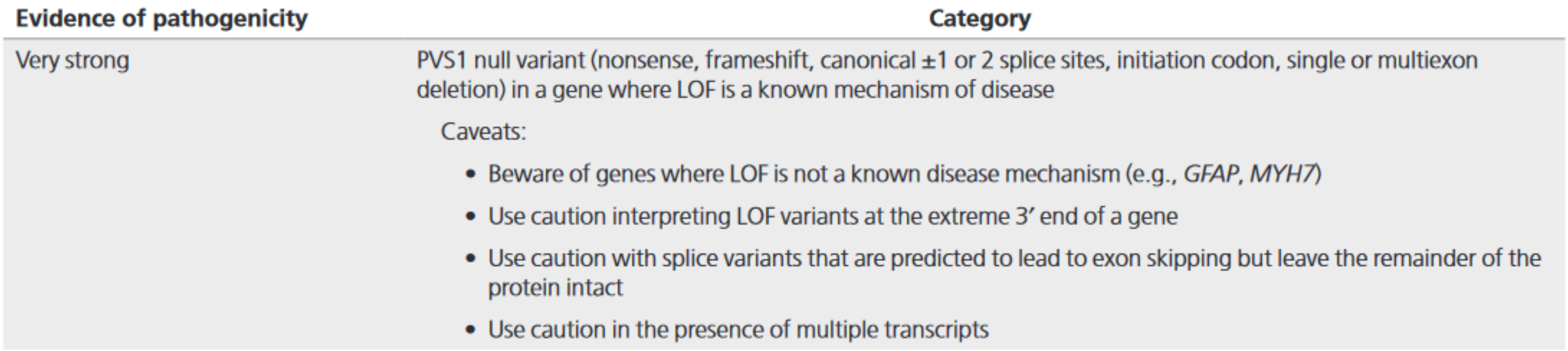
Sue Richards, PhD[1], Nazneen Aziz, PhD[2,16], Sherri Bale, PhD[3], David Bick, MD[4], Soma Das, PhD[5], Julie Gastier-Foster, PhD[6,7,8], Wayne W. Grody, MD, PhD[9,10,11], Madhuri Hegde, PhD[12], Elaine Lyon, PhD[13], Elaine Spector, PhD[14], Karl Voelkerding, MD[13] and Heidi L. Rehm, PhD[15]; on behalf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

Criteria Overview



	Benign		Pathogenic			
	Strong	Supporting	Supporting	Moderate	Strong	Very strong
Population data	MAF is too high for disorder BA1/BS1 **OR** observation in controls inconsistent with disease penetrance BS2			Absent in population databases PM2	Prevalence in affecteds statistically increased over controls PS4	
Computational and predictive data		Multiple lines of computational evidence suggest no impact on gene /gene product BP4 Missense in gene where only truncating cause disease BP1 Silent variant with non predicted splice impact BP7 In-frame indels in repeat w/out known function BP3	Multiple lines of computational evidence support a deleterious effect on the gene /gene product PP3	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before PM5 Protein length changing variant PM4	Same amino acid change as an established pathogenic variant PS1	Predicted null variant in a gene where LOF is a known mechanism of disease PVS1
Functional data	Well-established functional studies show no deleterious effect BS3		Missense in gene with low rate of benign missense variants and path. missenses common PP2	Mutational hot spot or well-studied functional domain without benign variation PM1	Well-established functional studies show a deleterious effect PS3	
Segregation data	Nonsegregation with disease BS4		Cosegregation with disease in multiple affected family members PP1	Increased segregation data →		
De novo data				De novo (without paternity & maternity confirmed) PM6	De novo (paternity and maternity confirmed) PS2	
Allelic data		Observed in *trans* with a dominant variant BP2 Observed in *cis* with a pathogenic variant BP2		For recessive disorders, detected in trans with a pathogenic variant PM3		
Other database		Reputable source w/out shared data = benign BP6	Reputable source = pathogenic PP5			
Other data		Found in case with an alternate cause BP5	Patient's phenotype or FH highly specific for gene PP4			

Very Strong Criteria

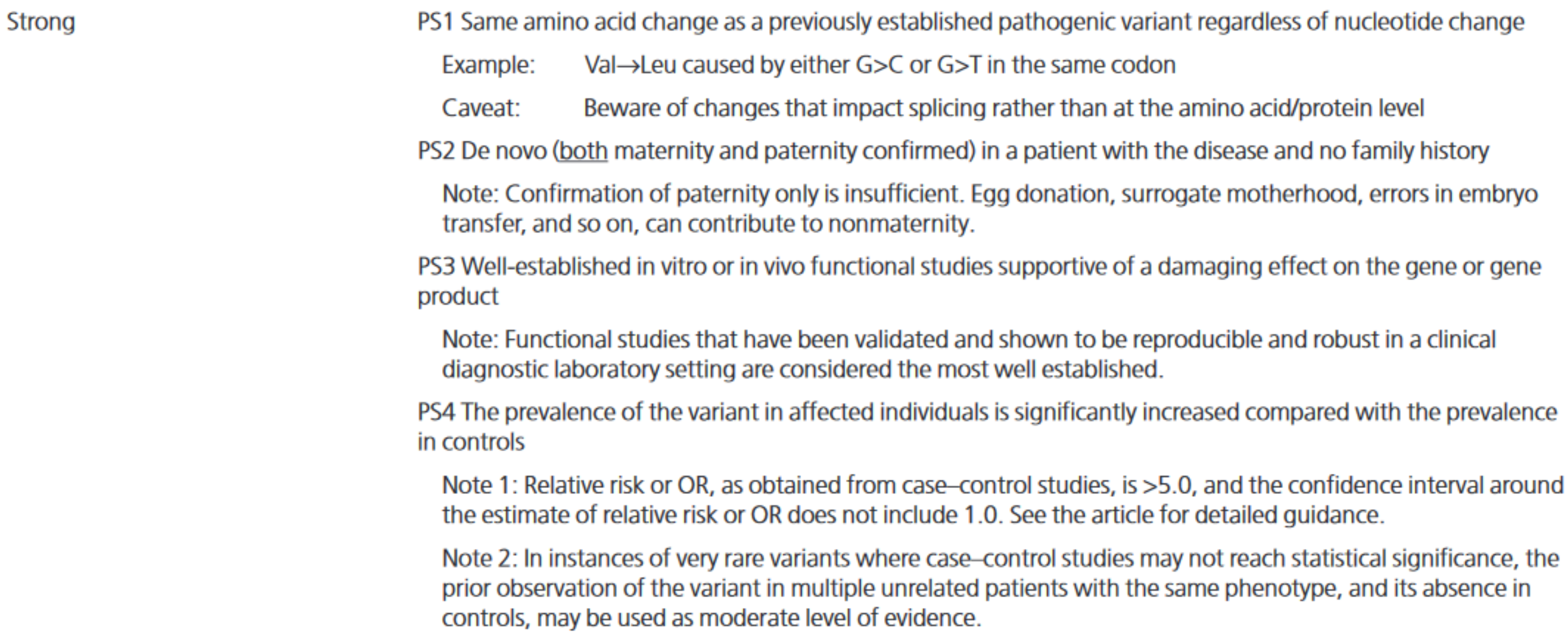


Evidence of pathogenicity	Category
Very strong	PVS1 null variant (nonsense, frameshift, canonical ±1 or 2 splice sites, initiation codon, single or multiexon deletion) in a gene where LOF is a known mechanism of disease Caveats: • Beware of genes where LOF is not a known disease mechanism (e.g., *GFAP, MYH7*) • Use caution interpreting LOF variants at the extreme 3′ end of a gene • Use caution with splice variants that are predicted to lead to exon skipping but leave the remainder of the protein intact • Use caution in the presence of multiple transcripts

Note the caveats

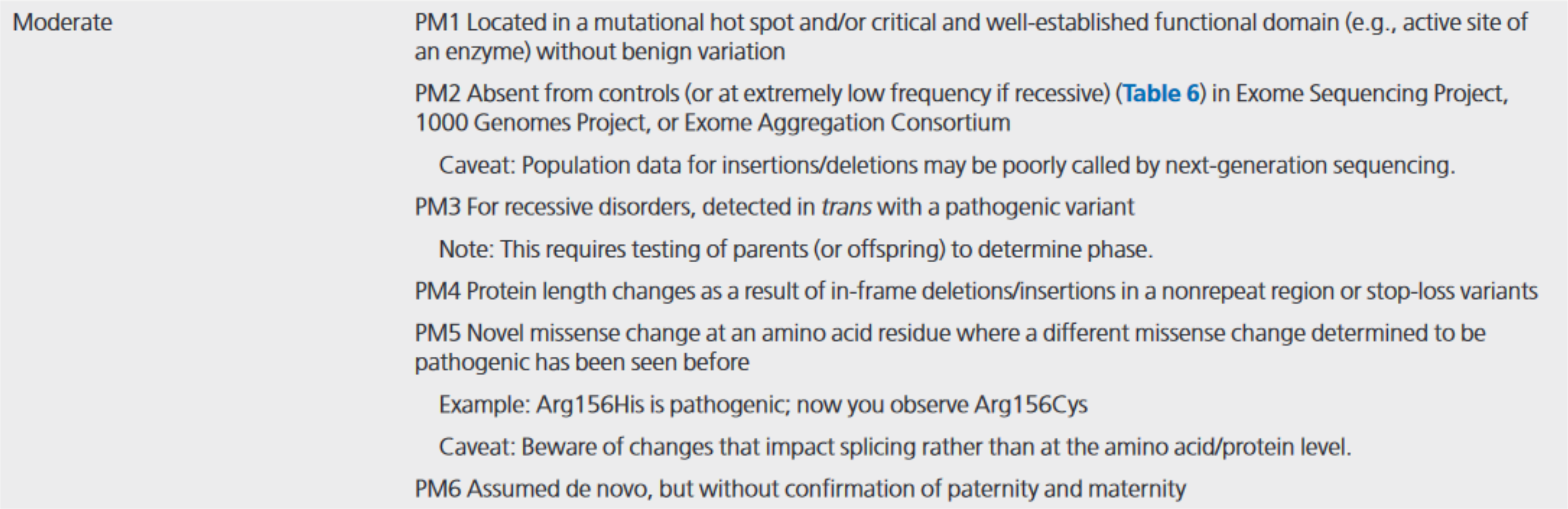
- **Loss of function must be a known mechanism**
- Variants that escape nonsense mediated decay
  - Certain disease genes for example WHIM syndrome and FAM83H Amelogenisis imperfecta, truncation mutations lead to gain of function effect which are pathogenic while loss of function mutations are tolerable (at least in a heterozygous state)
- In-frame exon skipping
- Minor Transcripts

Strong Criteria



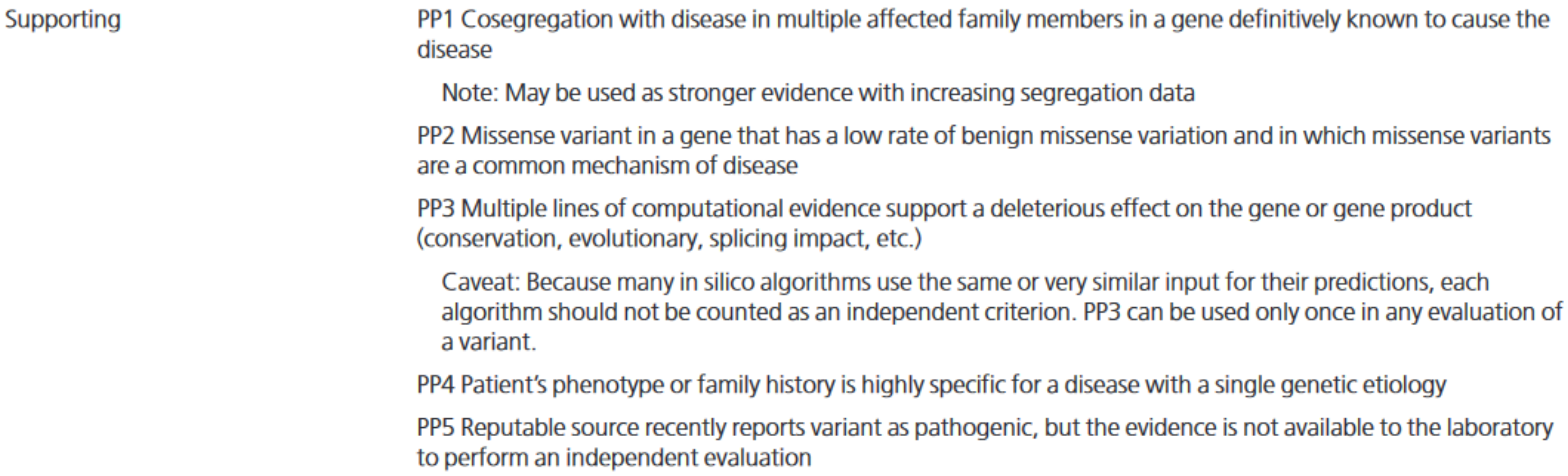
Strong	PS1 Same amino acid change as a previously established pathogenic variant regardless of nucleotide change
	Example: Val→Leu caused by either G>C or G>T in the same codon
	Caveat: Beware of changes that impact splicing rather than at the amino acid/protein level
	PS2 De novo (both maternity and paternity confirmed) in a patient with the disease and no family history
	Note: Confirmation of paternity only is insufficient. Egg donation, surrogate motherhood, errors in embryo transfer, and so on, can contribute to nonmaternity.
	PS3 Well-established in vitro or in vivo functional studies supportive of a damaging effect on the gene or gene product
	Note: Functional studies that have been validated and shown to be reproducible and robust in a clinical diagnostic laboratory setting are considered the most well established.
	PS4 The prevalence of the variant in affected individuals is significantly increased compared with the prevalence in controls
	Note 1: Relative risk or OR, as obtained from case–control studies, is >5.0, and the confidence interval around the estimate of relative risk or OR does not include 1.0. See the article for detailed guidance.
	Note 2: In instances of very rare variants where case–control studies may not reach statistical significance, the prior observation of the variant in multiple unrelated patients with the same phenotype, and its absence in controls, may be used as moderate level of evidence.

Moderate Criteria



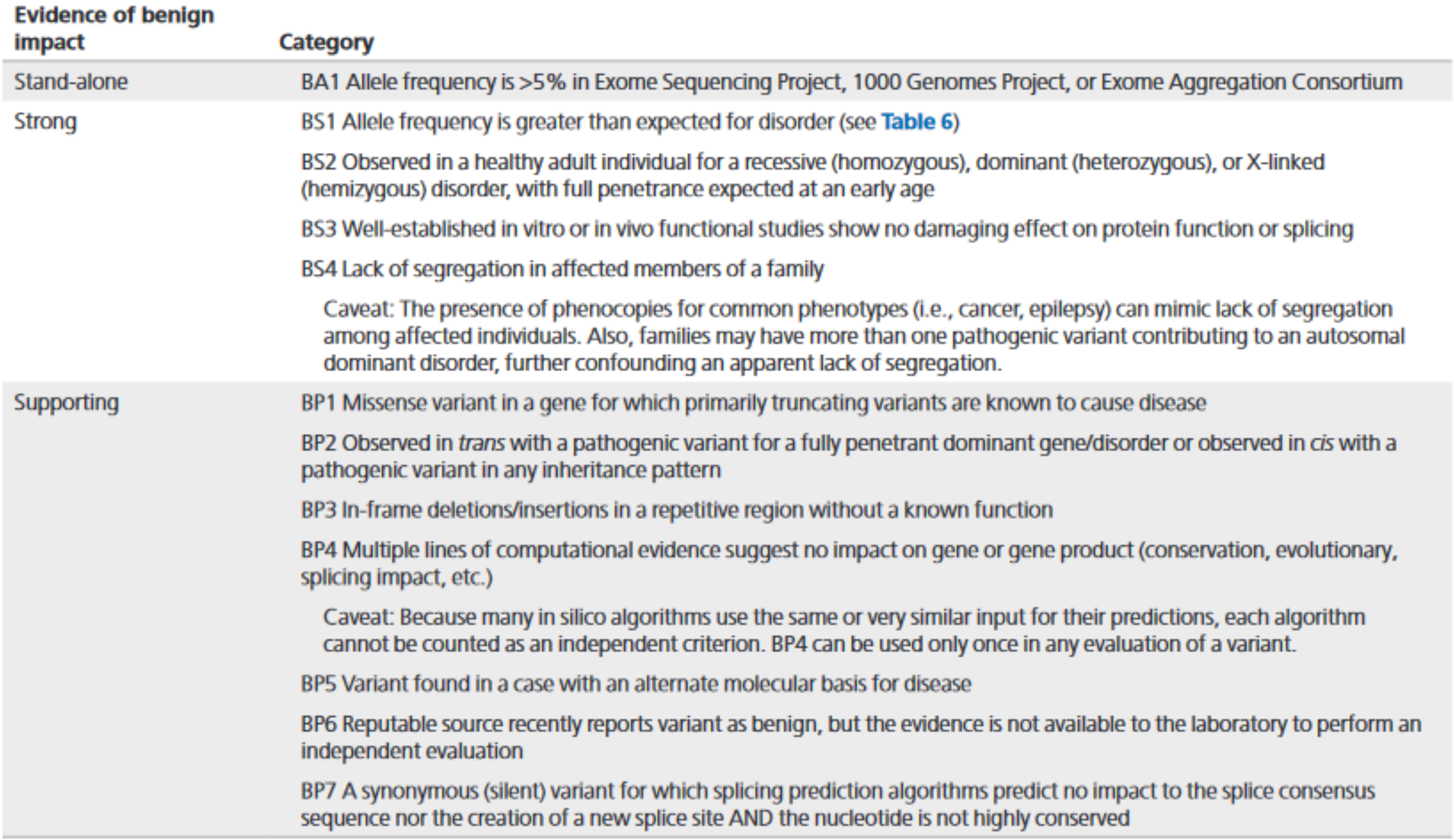
Moderate	PM1 Located in a mutational hot spot and/or critical and well-established functional domain (e.g., active site of an enzyme) without benign variation PM2 Absent from controls (or at extremely low frequency if recessive) (**Table 6**) in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium Caveat: Population data for insertions/deletions may be poorly called by next-generation sequencing. PM3 For recessive disorders, detected in *trans* with a pathogenic variant Note: This requires testing of parents (or offspring) to determine phase. PM4 Protein length changes as a result of in-frame deletions/insertions in a nonrepeat region or stop-loss variants PM5 Novel missense change at an amino acid residue where a different missense change determined to be pathogenic has been seen before Example: Arg156His is pathogenic; now you observe Arg156Cys Caveat: Beware of changes that impact splicing rather than at the amino acid/protein level. PM6 Assumed de novo, but without confirmation of paternity and maternity

Supporting Criteria



Supporting	PP1 Cosegregation with disease in multiple affected family members in a gene definitively known to cause the disease Note: May be used as stronger evidence with increasing segregation data 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 impact, etc.) Caveat: Because many in silico algorithms use the same or very similar input for their predictions, each algorithm should not be counted as an independent criterion. PP3 can be used only once in any evaluation of a variant. PP4 Patient's phenotype or family history is highly specific for a disease with a single genetic etiology PP5 Reputable source recently reports variant as pathogenic, but the evidence is not available to the laboratory to perform an independent evaluation

Benign Criteria



Evidence of benign impact	Category
Stand-alone	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 (see Table 6)
	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
	BS3 Well-established in vitro or in vivo functional studies show no damaging effect on protein function or splicing
	BS4 Lack of segregation in affected members of a family
	Caveat: The presence of phenocopies for common phenotypes (i.e., cancer, epilepsy) can mimic lack of segregation among affected individuals. Also, families may have more than one pathogenic variant contributing to an autosomal dominant disorder, further confounding an apparent lack of segregation.
Supporting	BP1 Missense variant in a gene for which primarily truncating variants are known to cause disease
	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
	BP3 In-frame deletions/insertions in a repetitive region without a known function
	BP4 Multiple lines of computational evidence suggest no impact on gene or gene product (conservation, evolutionary, splicing impact, etc.)
	Caveat: Because many in silico algorithms use the same or very similar input for their predictions, each algorithm cannot be counted as an independent criterion. BP4 can be used only once in any evaluation of a variant.
	BP5 Variant found in a case with an alternate molecular basis for disease
	BP6 Reputable source recently reports variant as benign, but the evidence is not available to the laboratory to perform an independent evaluation
	BP7 A synonymous (silent) variant for which splicing prediction algorithms predict no impact to the splice consensus sequence nor the creation of a new splice site AND the nucleotide is not highly conserved

Combining Criteria

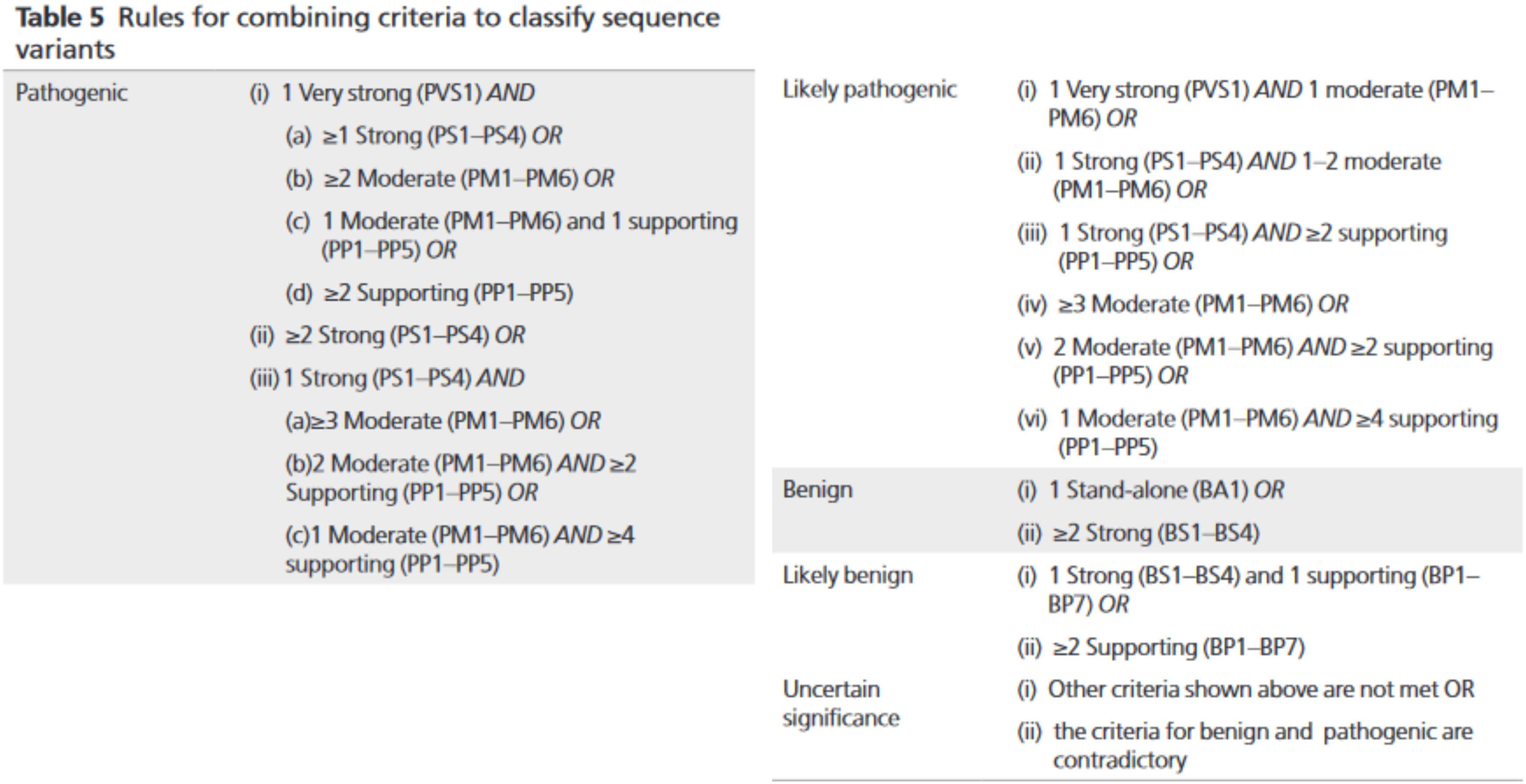


Table 5 Rules for combining criteria to classify sequence variants

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

Source of Information

- Annovar / VEP / SnpEff/ etc
  - Variant annotations (Computational predictors, population frequencies)
  - SIFT, Polyphen, CADD, etc…
  - Splice prediction
- Exac / Gnomad
  - Population Frequencies
- Varsome, etc
  - Variant annotations in web interface
- UCSC genome browser
- Uniprot
- Clinvar
  - Previously classified variants

Mutation and Variant Databases

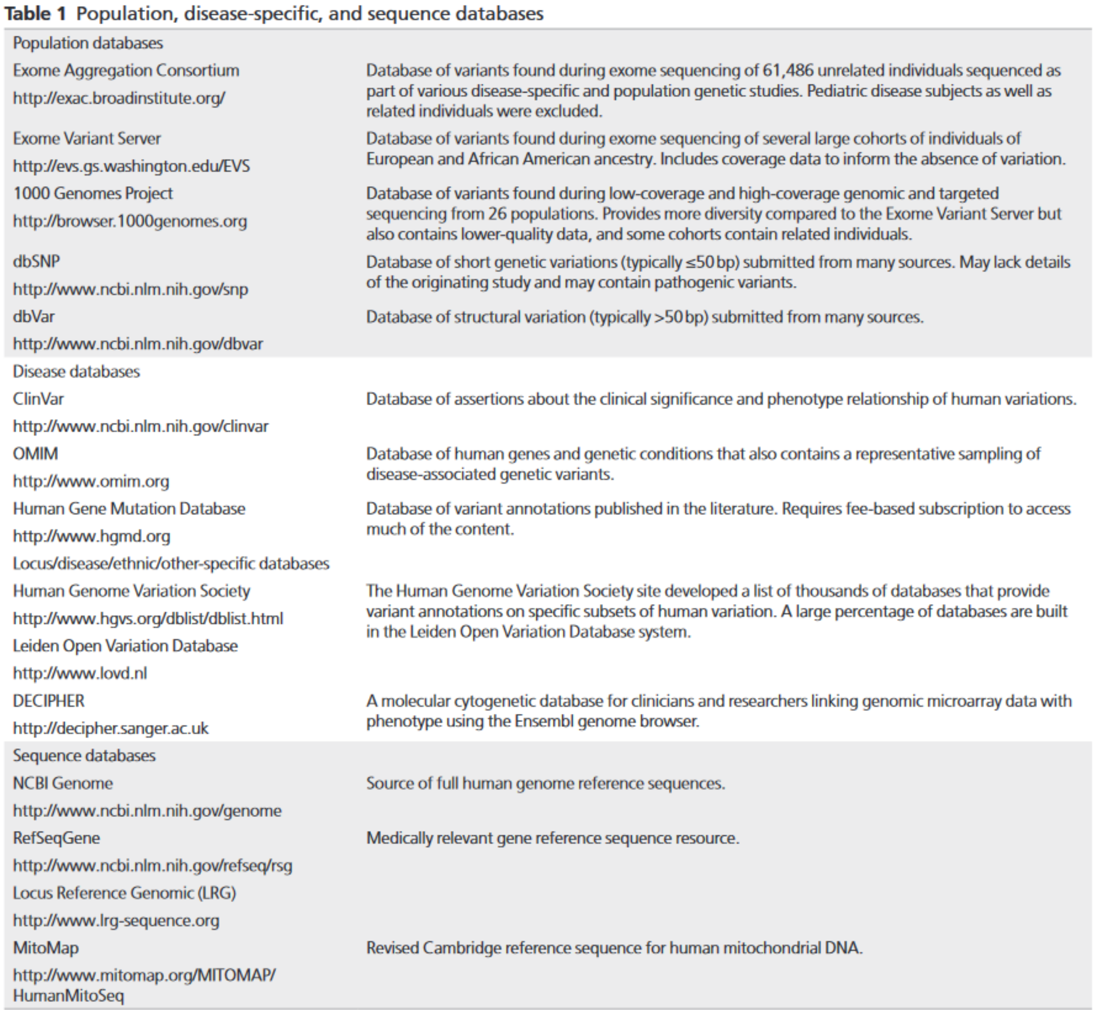


Table 1 Population, disease-specific, and sequence databases

Population databases	
Exome Aggregation Consortium http://exac.broadinstitute.org/	Database of variants found during exome sequencing of 61,486 unrelated individuals sequenced as part of various disease-specific and population genetic studies. Pediatric disease subjects as well as related individuals were excluded.
Exome Variant Server http://evs.gs.washington.edu/EVS	Database of variants found during exome sequencing of several large cohorts of individuals of European and African American ancestry. Includes coverage data to inform the absence of variation.
1000 Genomes Project http://browser.1000genomes.org	Database of variants found during low-coverage and high-coverage genomic and targeted sequencing from 26 populations. Provides more diversity compared to the Exome Variant Server but also contains lower-quality data, and some cohorts contain related individuals.
dbSNP http://www.ncbi.nlm.nih.gov/snp	Database of short genetic variations (typically ≤50 bp) submitted from many sources. May lack details of the originating study and may contain pathogenic variants.
dbVar http://www.ncbi.nlm.nih.gov/dbvar	Database of structural variation (typically >50 bp) submitted from many sources.
Disease databases	
ClinVar http://www.ncbi.nlm.nih.gov/clinvar	Database of assertions about the clinical significance and phenotype relationship of human variations.
OMIM http://www.omim.org	Database of human genes and genetic conditions that also contains a representative sampling of disease-associated genetic variants.
Human Gene Mutation Database http://www.hgmd.org	Database of variant annotations published in the literature. Requires fee-based subscription to access much of the content.
Locus/disease/ethnic/other-specific databases	
Human Genome Variation Society http://www.hgvs.org/dblist/dblist.html	The Human Genome Variation Society site developed a list of thousands of databases that provide variant annotations on specific subsets of human variation. A large percentage of databases are built in the Leiden Open Variation Database system.
Leiden Open Variation Database http://www.lovd.nl	
DECIPHER http://decipher.sanger.ac.uk	A molecular cytogenetic database for clinicians and researchers linking genomic microarray data with phenotype using the Ensembl genome browser.
Sequence databases	
NCBI Genome http://www.ncbi.nlm.nih.gov/genome	Source of full human genome reference sequences.
RefSeqGene http://www.ncbi.nlm.nih.gov/refseq/rsg	Medically relevant gene reference sequence resource.
Locus Reference Genomic (LRG) http://www.lrg-sequence.org	
MitoMap http://www.mitomap.org/MITOMAP/HumanMitoSeq	Revised Cambridge reference sequence for human mitochondrial DNA.

In Silico Prediction tools

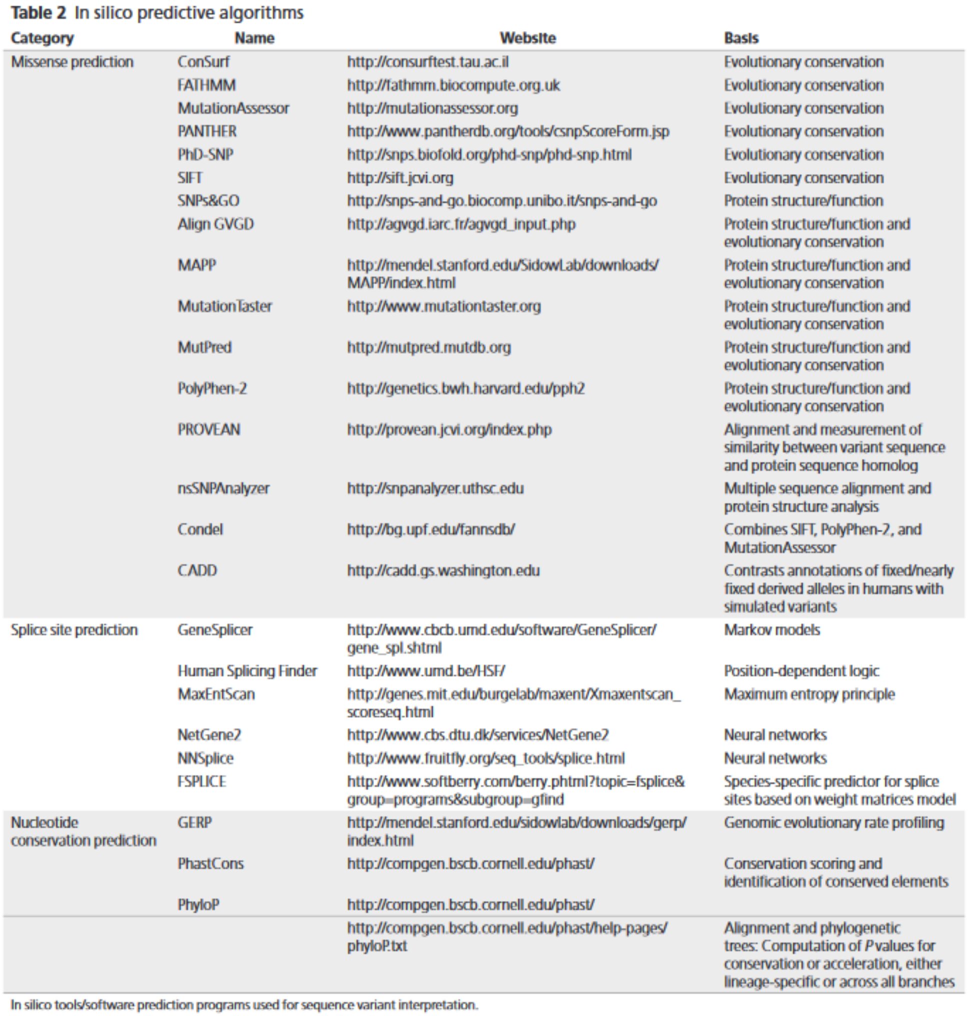
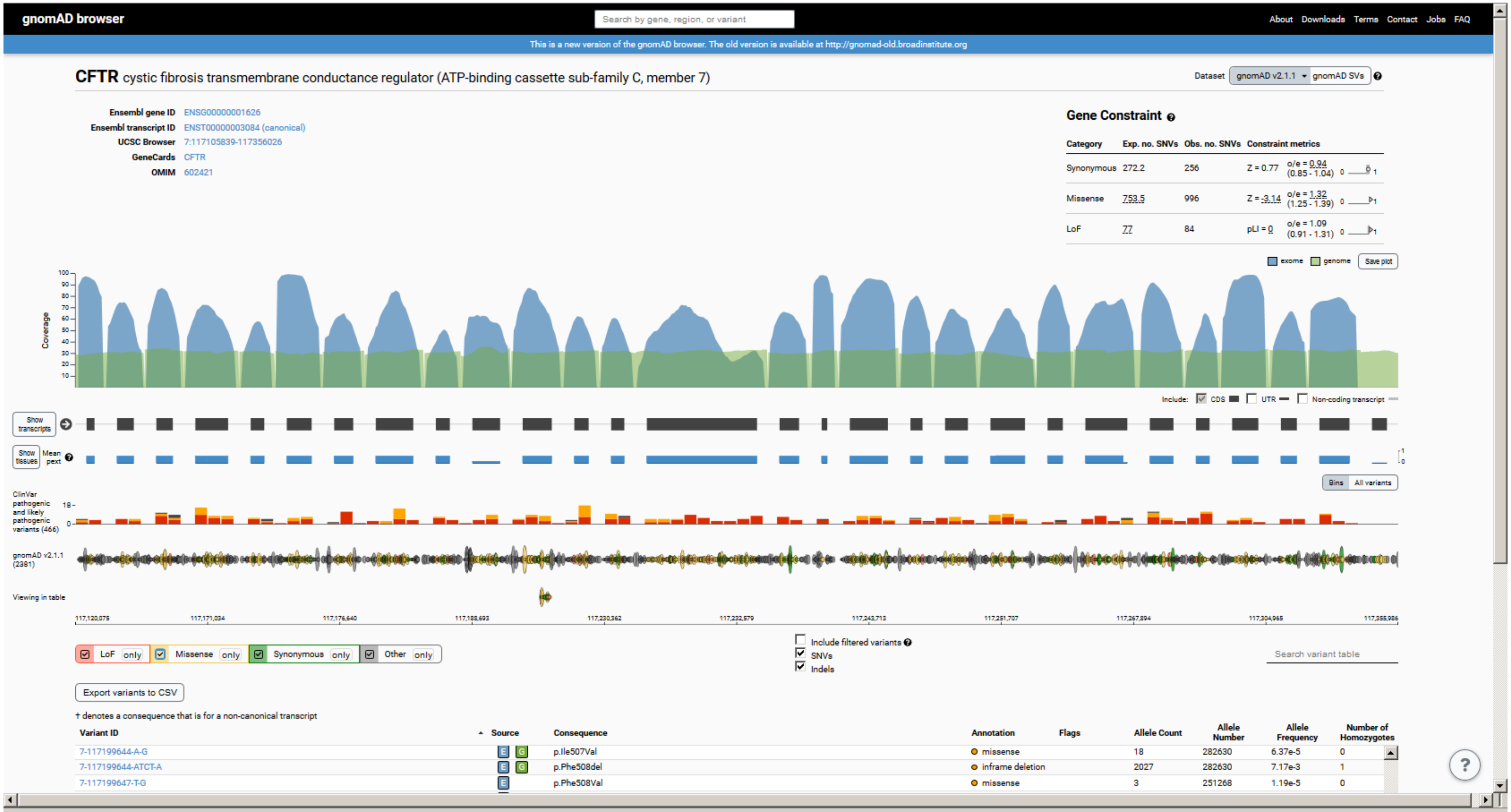


Table 2 In silico predictive algorithms

Category	Name	Website	Basis
Missense prediction	ConSurf	http://consurftest.tau.ac.il	Evolutionary conservation
	FATHMM	http://fathmm.biocompute.org.uk	Evolutionary conservation
	MutationAssessor	http://mutationassessor.org	Evolutionary conservation
	PANTHER	http://www.pantherdb.org/tools/csnpScoreForm.jsp	Evolutionary conservation
	PhD-SNP	http://snps.biofold.org/phd-snp/phd-snp.html	Evolutionary conservation
	SIFT	http://sift.jcvi.org	Evolutionary conservation
	SNPs&GO	http://snps-and-go.biocomp.unibo.it/snps-and-go	Protein structure/function
	Align GVGD	http://agvgd.iarc.fr/agvgd_input.php	Protein structure/function and evolutionary conservation
	MAPP	http://mendel.stanford.edu/SidowLab/downloads/MAPP/index.html	Protein structure/function and evolutionary conservation
	MutationTaster	http://www.mutationtaster.org	Protein structure/function and evolutionary conservation
	MutPred	http://mutpred.mutdb.org	Protein structure/function and evolutionary conservation
	PolyPhen-2	http://genetics.bwh.harvard.edu/pph2	Protein structure/function and evolutionary conservation
	PROVEAN	http://provean.jcvi.org/index.php	Alignment and measurement of similarity between variant sequence and protein sequence homolog
	nsSNPAnalyzer	http://snpanalyzer.uthsc.edu	Multiple sequence alignment and protein structure analysis
	Condel	http://bg.upf.edu/fannsdb/	Combines SIFT, PolyPhen-2, and MutationAssessor
	CADD	http://cadd.gs.washington.edu	Contrasts annotations of fixed/nearly fixed derived alleles in humans with simulated variants
Splice site prediction	GeneSplicer	http://www.cbcb.umd.edu/software/GeneSplicer/gene_spl.shtml	Markov models
	Human Splicing Finder	http://www.umd.be/HSF/	Position-dependent logic
	MaxEntScan	http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html	Maximum entropy principle
	NetGene2	http://www.cbs.dtu.dk/services/NetGene2	Neural networks
	NNSplice	http://www.fruitfly.org/seq_tools/splice.html	Neural networks
	FSPLICE	http://www.softberry.com/berry.phtml?topic=fsplice&group=programs&subgroup=gfind	Species-specific predictor for splice sites based on weight matrices model
Nucleotide conservation prediction	GERP	http://mendel.stanford.edu/sidowlab/downloads/gerp/index.html	Genomic evolutionary rate profiling
	PhastCons	http://compgen.bscb.cornell.edu/phast/	Conservation scoring and identification of conserved elements
	PhyloP	http://compgen.bscb.cornell.edu/phast/	
		http://compgen.bscb.cornell.edu/phast/help-pages/phyloP.txt	Alignment and phylogenetic trees: Computation of *P* values for conservation or acceleration, either lineage-specific or across all branches

In silico tools/software prediction programs used for sequence variant interpretation.

Example GNOMAD View:CFTR



gnomAD browser   Search by gene, region, or variant   About Downloads Terms Contact Jobs FAQ

This is a new version of the gnomAD browser. The old version is available at http://gnomad-old.broadinstitute.org

CFTR cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C, member 7)

Dataset: gnomAD v2.1.1 ▾ | gnomAD SVs

Ensembl gene ID	ENSG00000001626
Ensembl transcript ID	ENST00000003084 (canonical)
UCSC Browser	7:117105839-117356026
GeneCards	CFTR
OMIM	602421

Gene Constraint

Category	Exp. no. SNVs	Obs. no. SNVs	Constraint metrics	
Synonymous	272.2	256	Z = 0.77	o/e = 0.94 (0.85 - 1.04)
Missense	753.5	996	Z = -3.14	o/e = 1.32 (1.25 - 1.39)
LoF	77	84	pLI = 0	o/e = 1.09 (0.91 - 1.31)

☑ LoF only | ☑ Missense only | ☑ Synonymous only | ☑ Other only

□ Include filtered variants
☑ SNVs
☑ Indels

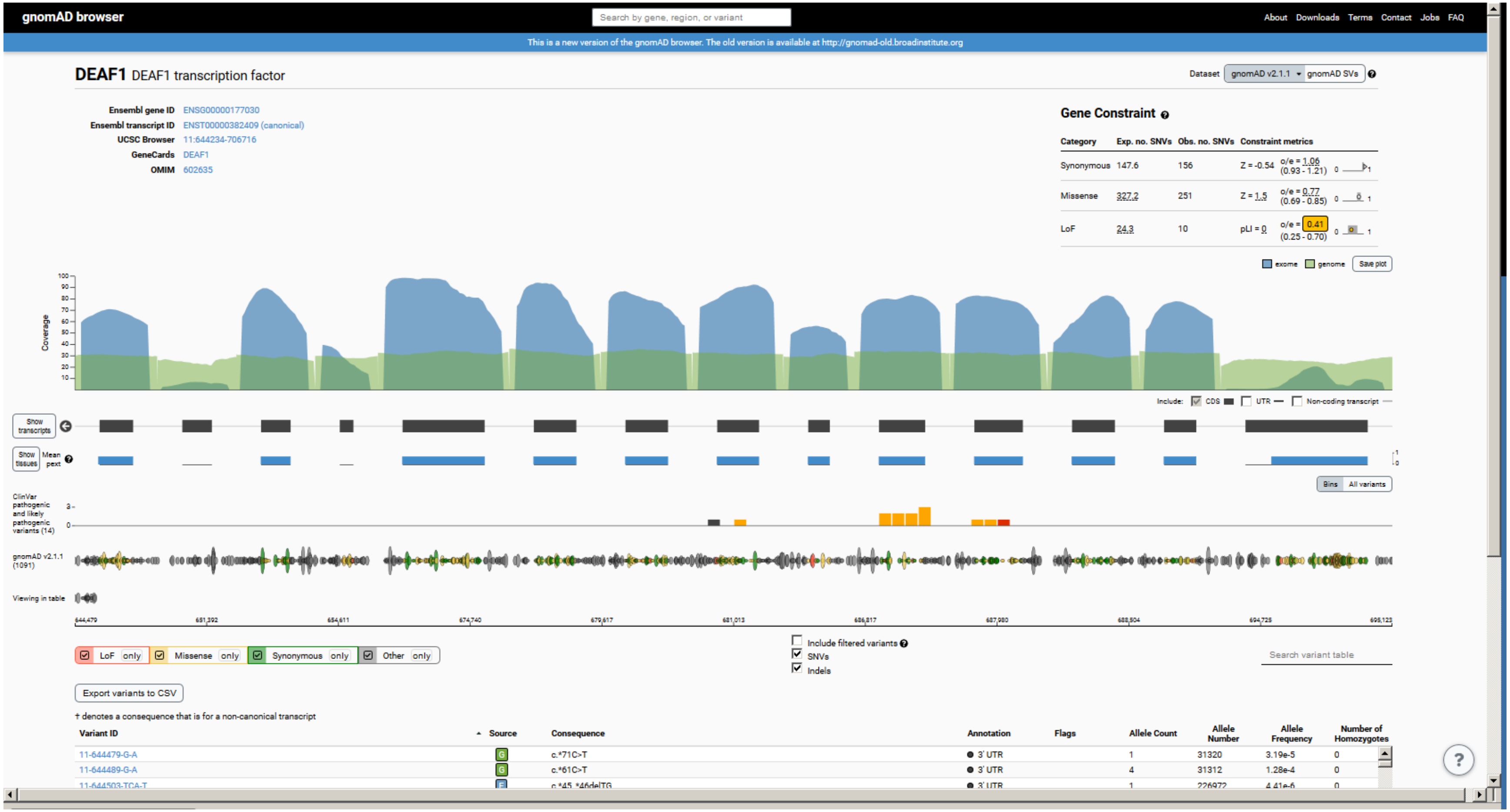
Search variant table

Export variants to CSV

† denotes a consequence that is for a non-canonical transcript

Variant ID	Source	Consequence	Annotation	Flags	Allele Count	Allele Number	Allele Frequency	Number of Homozygotes
7-117199644-A-G	E G	p.Ile507Val	● missense		18	282630	6.37e-5	0
7-117199644-ATCT-A	E G	p.Phe508del	● inframe deletion		2027	282630	7.17e-3	1
7-117199647-T-G	E	p.Phe508Val	● missense		3	251268	1.19e-5	0

Example Gnomad View:DEAF1



gnomAD browser
Search by gene, region, or variant
About Downloads Terms Contact Jobs FAQ
This is a new version of the gnomAD browser. The old version is available at http://gnomad-old.broadinstitute.org
Dataset gnomAD v2.1.1 gnomAD SVs
Coverage
100 90 80 70 60 50 40 30 20 10
exome genome Save plot
Include: ☑ CDS □ UTR □ Non-coding transcript
Show transcripts
Show tissues Mean pext
Bins All variants
ClinVar pathogenic and likely pathogenic variants (14)
3 0
gnomAD v2.1.1 (1091)
Viewing in table
644,479 651,392 654,611 674,740 679,617 681,013 686,817 687,980 688,504 694,725 695,123
☑ LoF only ☑ Missense only ☑ Synonymous only ☑ Other only
Search variant table
Export variants to CSV

DEAF1 DEAF1 transcription factor

Ensembl gene ID	ENSG00000177030
Ensembl transcript ID	ENST00000382409 (canonical)
UCSC Browser	11:644234-706716
GeneCards	DEAF1
OMIM	602635

Gene Constraint

Category	Exp. no. SNVs	Obs. no. SNVs	Constraint metrics	
Synonymous	147.6	156	Z = -0.54	o/e = 1.06 (0.93 - 1.21)
Missense	327.2	251	Z = 1.5	o/e = 0.77 (0.69 - 0.85)
LoF	24.3	10	pLI = 0	o/e = 0.41 (0.25 - 0.70)

□ Include filtered variants
☑ SNVs
☑ Indels

† denotes a consequence that is for a non-canonical transcript

Variant ID	Source	Consequence	Annotation	Flags	Allele Count	Allele Number	Allele Frequency	Number of Homozygotes
11-644479-G-A	G	c.*71C>T	● 3' UTR		1	31320	3.19e-5	0
11-644489-G-A	G	c.*61C>T	● 3' UTR		4	31312	1.28e-4	0
11-644503-TCA-T	E	c.*45_*46delTG	● 3' UTR		1	226972	4.41e-6	0

Example Case:M2922

Clinical Features	Previous Genetic Testing	Exome *de novo* variants (filtering CADD > 20, ExAC freq < 0.001)
▪ DD/ID ▪ Obesity ▪ Upslanting palpebral fissures ▪ Epicanthal folds ▪ Broad nasal bridge ▪ Micrognathia ▪ Hypotonia ▪ Sleep concerns ▪ Sleep disordered breathing ▪ Behavioral issues ▪ Volatile mood ▪ Sensory issues	▪ Normal chromosomal microarray ▪ Normal methylation studies ▪ Negative FISH for VCF and SMS ▪ Research analysis: ○ *RAI1* sequence normal	Gene: *NDN* c.838G>C:p.A280P (ex 1) Variant: chr15:23931527 C>G Type: Missense CADD Phred: 25.8 Exac Frequency: 0 Comment: *NDN* is hemizygous in PWS, imprinted gene, mono-allelic variant expression Gene: *MAPK8IP3* c.1364A>G:p.E455G (ex 11) Variant: chr16:1810461 A>G Type: Missense CADD Phred: 28.4 ExAC Frequency: 0 Comment: Not reported in human disease

NDN (necdin-like protein)

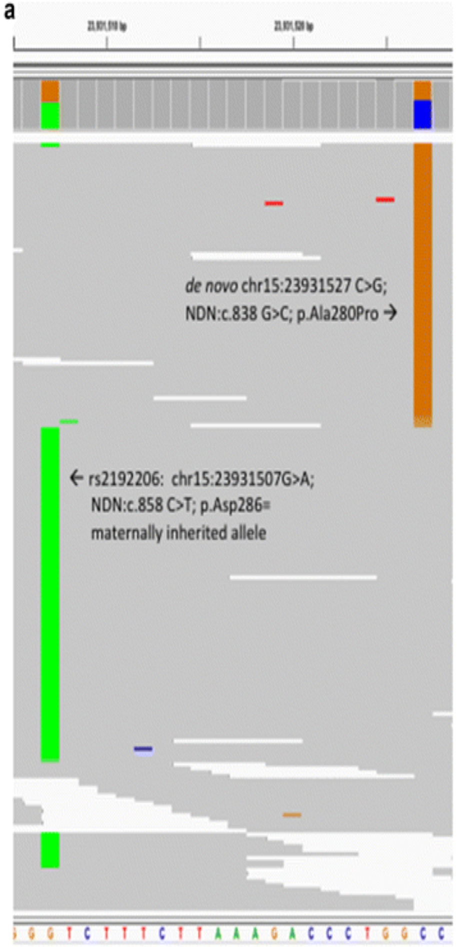
- **One of 7 genes deleted in the Prader-Willi Syndrome contiguous gene deletion syndrome** (15q11.2)
  - PWS is characterized by hypogonadism, hypotonia, cognitive disability, behavior problems, and hyperphagia

- ***NDN* is an imprinted gene**
  - Only paternal allele is expressed

- Not previously described in monogenic human disease but relatively constrained per ExAC

- Identified variant (A280P) results in and Alanine to Proline change in a predicted **alpha helix of the protein structure**
  - Likely to disrupt protein secondary structure

Phasing the Variant



de novo chr15:23931527 C>G; NDN:c.838 G>C; p.Ala280Pro →

← rs2192206: chr15:23931507G>A; NDN:c.858 C>T; p.Asp286= maternally inherited allele

Nearby maternal SNP allows phasing of de novo variant to paternally inherited chromosome

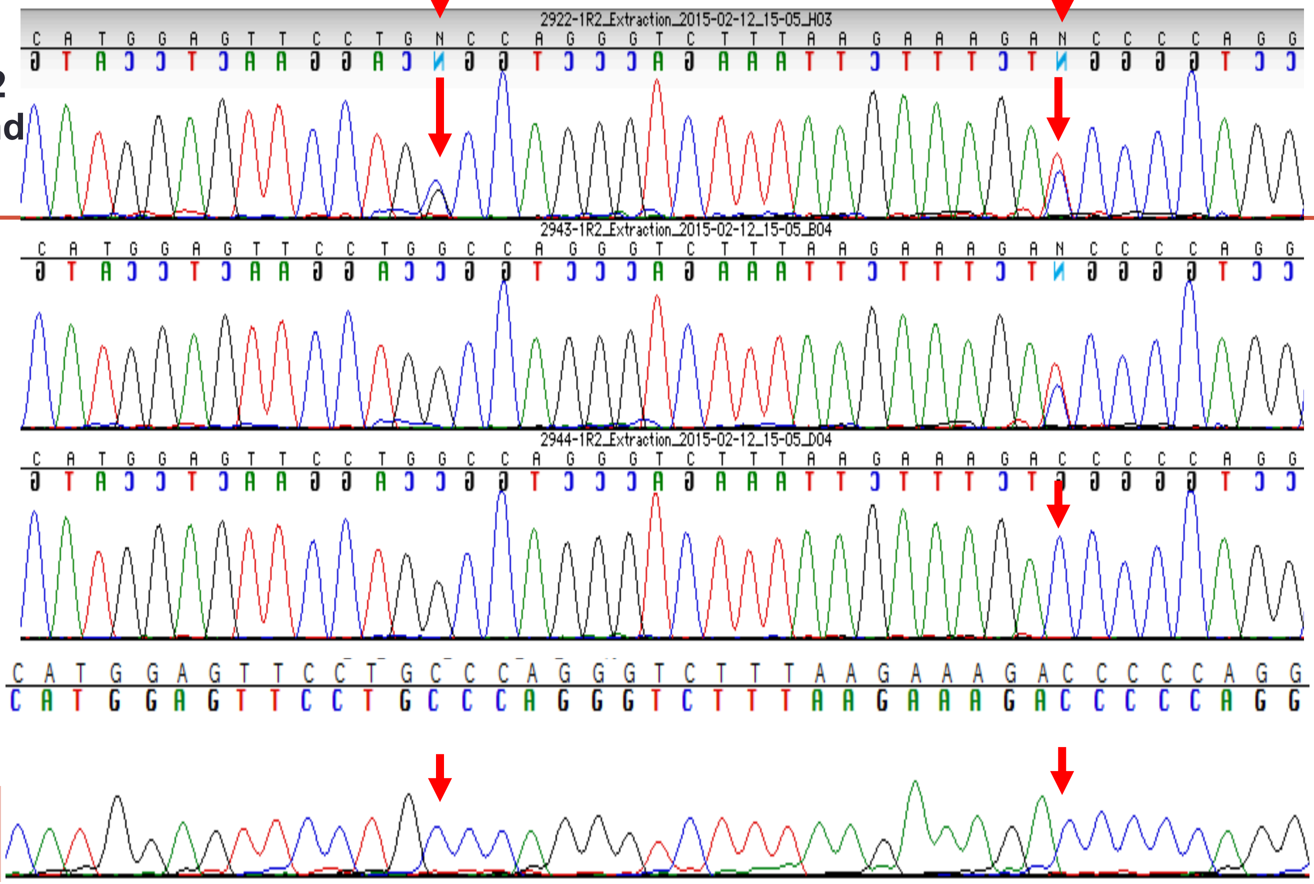
Berger SI, et al. Human Genetics. 2017 Apr;136(4):409-420.

NDN Sanger Validation

(variant on paternal allele)

de novo missense
c.838G>C

Maternal
inherited
silent SNP
c.858C>T



M2922
Proband
gDNA

M294
3
Mom
gDN
A

M29
43
Dad
gDN
A

Prob
and
cDN
A

NDN Discussion

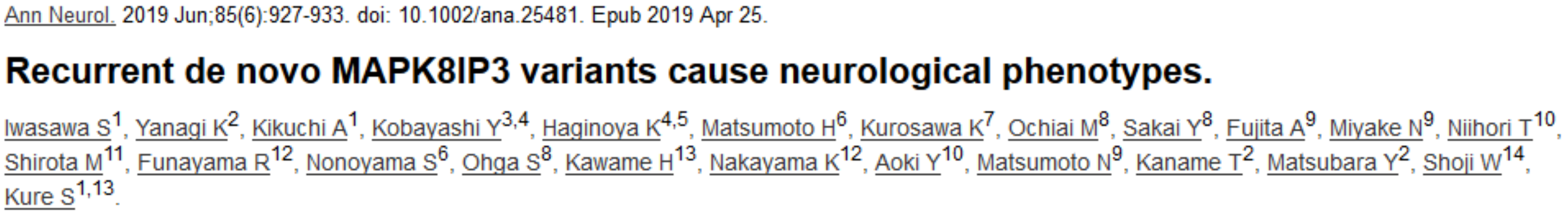
- **First de novo missense variant in this PWS region gene**

- **Our patient’s phenotype may help explain:**
  - Contribution of NDN deficiency to the PWS cognitive phenotype
  - Respiratory features in NDN knockout mouse
  - Exclude contribution of deficient NDN to other PWS phenotypic features

- **Novel c.838G>C:p.A280P missense variant requires further functional evaluation to effects on NDN gene/protein function**

- **Can not exclude role of MAPK8IP3 variant**
  - **(May also interact with NDN… Possible digenic effect)**

Then this happened… MAPK8IP3



Ann Neurol. 2019 Jun;85(6):927-933. doi: 10.1002/ana.25481. Epub 2019 Apr 25.

Recurrent de novo MAPK8IP3 variants cause neurological phenotypes.

Iwasawa S[1], Yanagi K[2], Kikuchi A[1], Kobayashi Y[3,4], Haginoya K[4,5], Matsumoto H[6], Kurosawa K[7], Ochiai M[8], Sakai Y[8], Fujita A[9], Miyake N[9], Niihori T[10], Shirota M[11], Funayama R[12], Nonoyama S[6], Ohga S[8], Kawame H[13], Nakayama K[12], Aoki Y[10], Matsumoto N[9], Kaname T[2], Matsubara Y[2], Shoji W[14], Kure S[1,13].

Summary

- Next generation sequencing pipelines can be tailored for germline variation, mosaic variation, small variants (SNV, indels), Copy number variants, etc.
- Clinical variant interpretation Guidelines
- Secondary findings

Thank you

- Children's Reasearch Institute Computational Biology Unit
    - Dr. Eric Vilain - Director of Center for Genetic Medicine Research
    - Dr. Hiroki Morizono - Director of CBU
    - Dr. Susan Knoblach
    - Payal Banerjee
    - Dr. Surajit Bhattacharya
    - Dr. Hayk Barseghyan