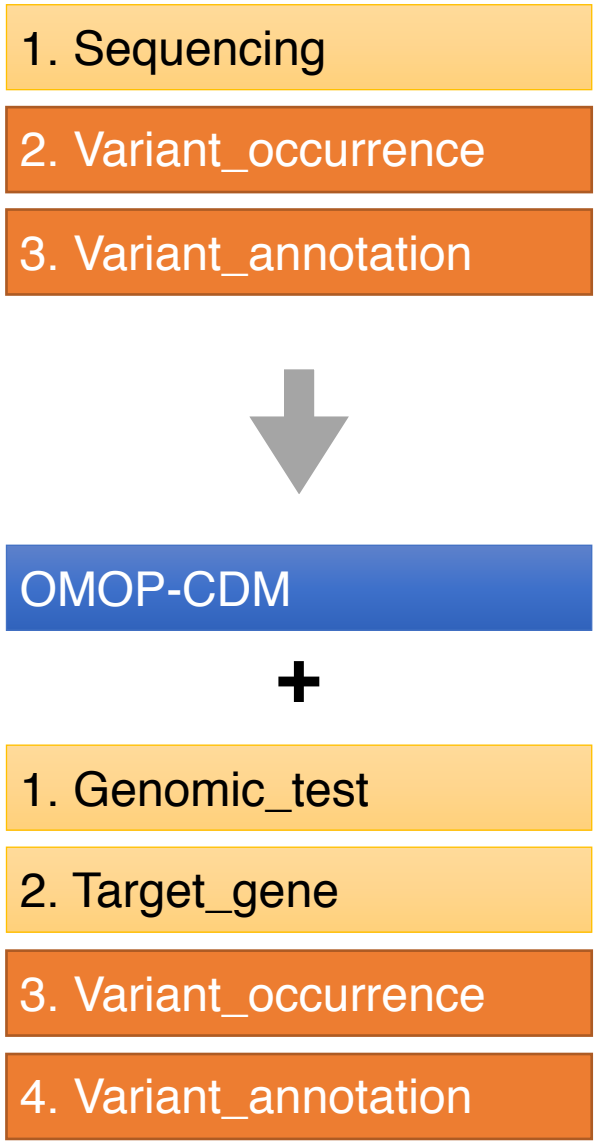
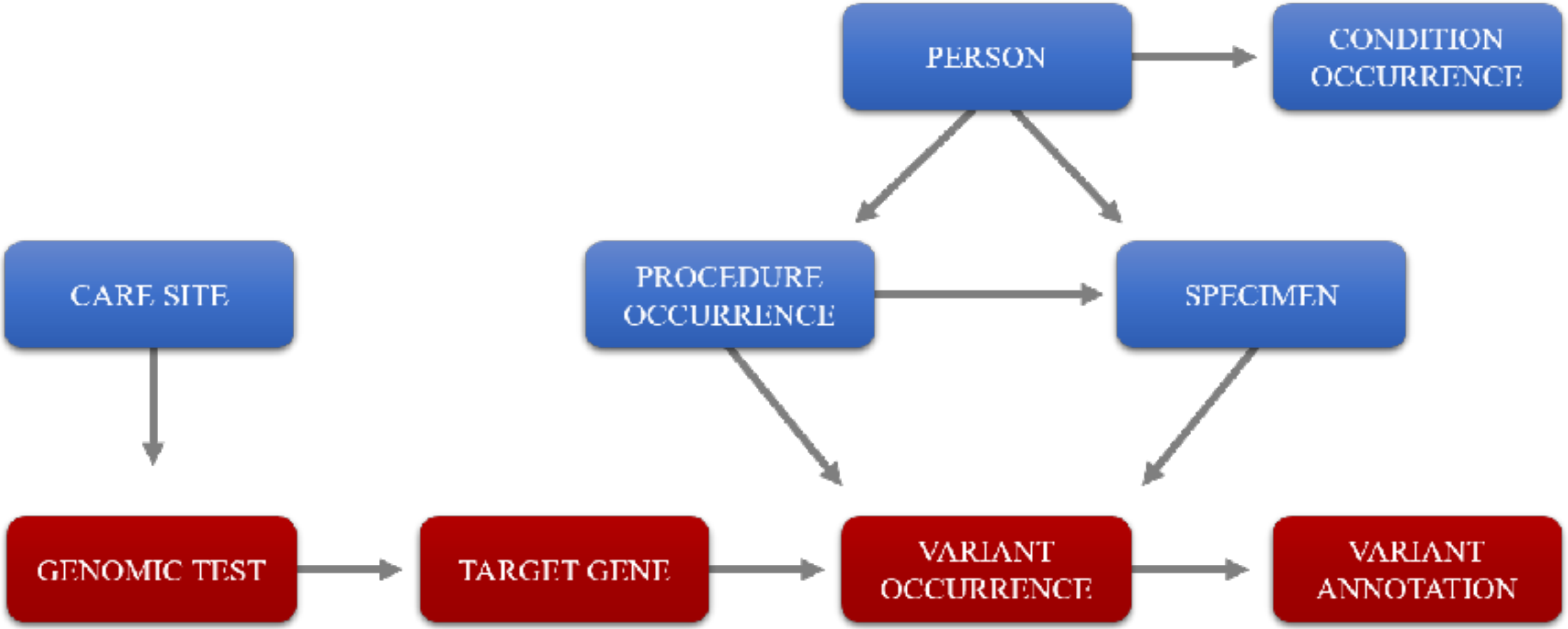
G-CDM Structure

- Beginning version
  - In the OHDSI Symposium in May, 2018
- Upgrade version
  - Take full utilize of the existing OMOP-CDM tables
  - Adapt a standard vocabulary system



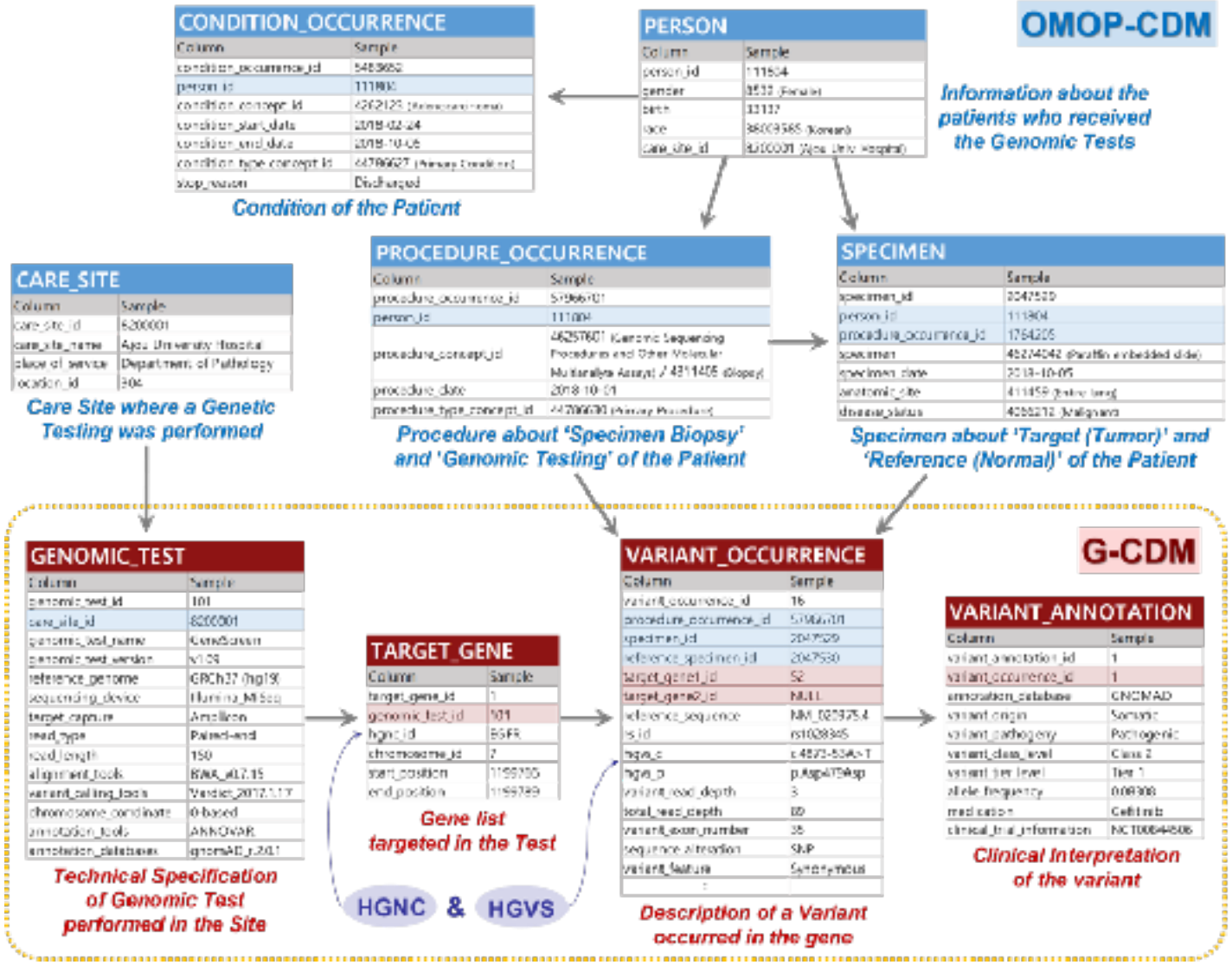
G-CDM Structure



Schematic diagram of the relationship between the tables that make up the GCDM.

G-CDM Structure

Entity-relationship diagram (ERD) of G-CDM as an extension to the OMOP-CDM.



CARE_SITE

Column	Sample
care_site_id	8200001
care_site_name	Ajou University Hospitel
place_of_service	Department of Pathology
location_id	304

Care Site where a Genetic Testing was performed

Column	Sample
procedure_occurrence_id	57966701
person_id	111804
procedure_concept_id	46257501 (Genomic Sequencing Procedures and Other Molecular Multianalyte Assays) / 4311405 (Biopsy)
procedure_date	2018-10-01
procedure_type_concept_id	44786630 (Primary Procedure)

Procedure about 'Specimen Biopsy' and 'Genomic Testing' of the Patient

Column	Sample
specimen_id	2047529
person_id	111804
procedure_occurrence_id	1764205
specimen	46274042 (Paraffin embedded slide)
specimen_date	2018-10-05
anatomic_site	411459 (Entire lung)
disease_status	4066212 (Malignant)

Specimen about 'Target (Tumor)' and 'Reference (Normal)' of the Patient

G-CDM

GENOMIC_TEST

Column	Sample
genomic_test_id	101
care_site_id	8200001
genomic_test_name	GeneScreen
genomic_test_version	v1.09
reference_genome	GRCh37 (hg19)
sequencing_device	Illumina_MiSeq
target_capture	Amplicon
read_type	Paired-end
read_length	150
alignment_tools	BWA_v0.7.15
variant_calling_tools	Vardict_2017.1.17
chromosome_corrdinate	0-based
annotation_tools	ANNOVAR
annotation_databases	gnomAD_r.2.0.1

Technical Specification of Genomic Test performed in the Site

TARGET_GENE

Column	Sample
target_gene_id	1
genomic_test_id	101
hgnc_id	EGFR
chromosome_id	7
start_position	1199766
end_position	1199789

Gene list targeted in the Test

HGNC & HGVS

VARIANT_OCCURRENCE

Column	Sample
variant_occurrence_id	16
procedure_occurrence_id	57966701
specimen_id	2047529
reference_specimen_id	2047530
target_gene1_id	52
target_gene2_id	NULL
reference_sequence	NM_020975.4
rs_id	rs1028345
hgvs_c	c.4873-53A>T
hgvs_p	p.Asp479Asp
variant_read_depth	3
total_read_depth	89
variant_exon_number	35
sequence_alteration	SNP
variant_feature	Synonymous
:	

Description of a Variant occurred in the gene

VARIANT_ANNOTATION

Column	Sample
variant_annotation_id	1
variant_occurrence_id	1
annotation_database	GNOMAD
variant_origin	Somatic
variant_pathogeny	Pathogenic
variant_class_level	Class 2
variant_tier_level	Tier 1
allele_frequency	0.08308
medication	Gefitinib
clinical_trial_information	NCT00844506

Clinical Interpretation of the variant

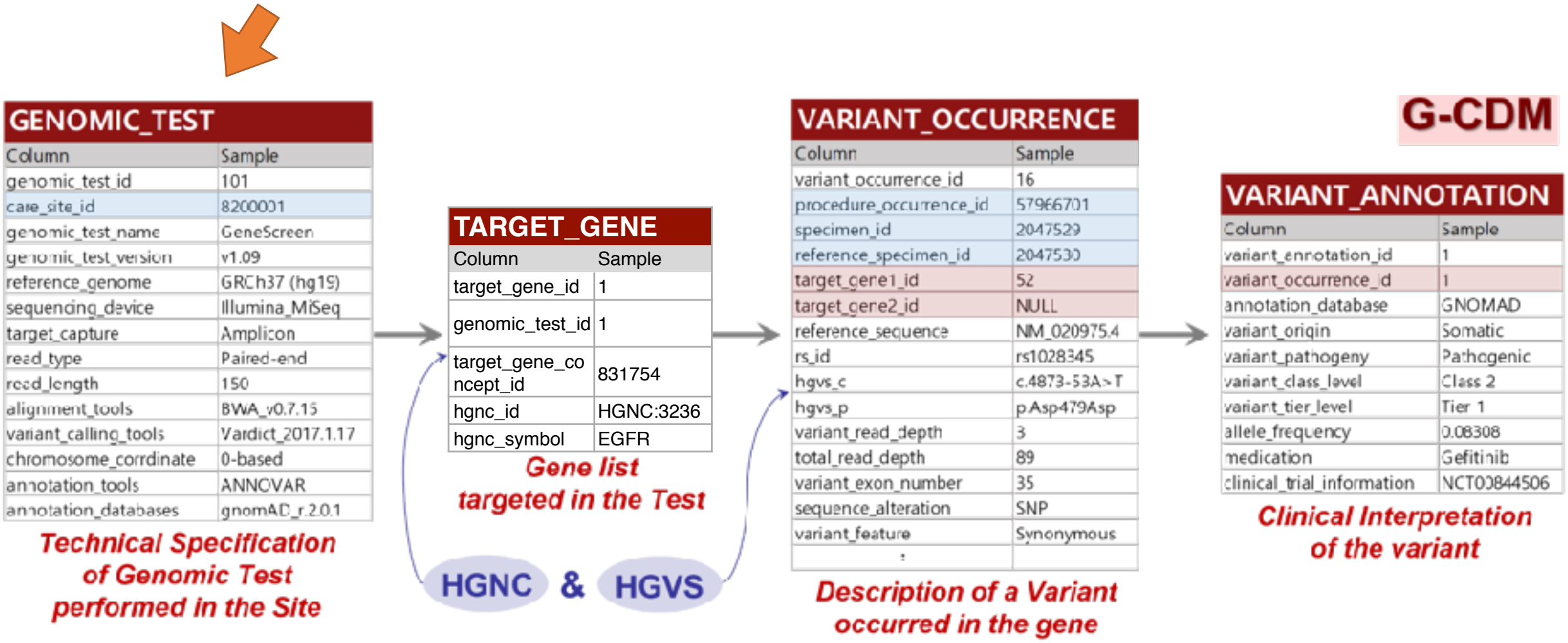
Concept ID Request

Table	Column	Concept_name	Number of Concept_id
Target_gene	target_gene_concept_id	Approved Gene Symbols of HGNC Database	41503
Variant_occurrence	sequence_alteration	MNP	1
	variant_feature	Upstream, Downstream, Stop-loss, Inframe, 5_prime_UTR, 3_prime_UTR, Intron, Splice_donor, Splice_acceptor	9
Variant_annotation	annotation_database	Clinvar, PolyPhen, SIFT, SnpEff, 1000G, GNOMAD, ExAC	7
	variant_pathogeny	Low, Modifier, Moderate, High	4
		Benign, Benign/Likely benign, Likely benign, Unknown significance, Likely pathogenic, Likely pathogenic/ Pathogenic, Conflict pathogenic, Pathogenic, Drug response	9
		(Benign), Possibly damaging, Probably damaging	2
		Tolerated, Tolerated (low confidence), Deleterious, Deleterious (low confidence)	4
	variant_class_level	Class 1~5	5
	variant_tier_level	Tier 1~4	4
			41548

45

Concept ID Request

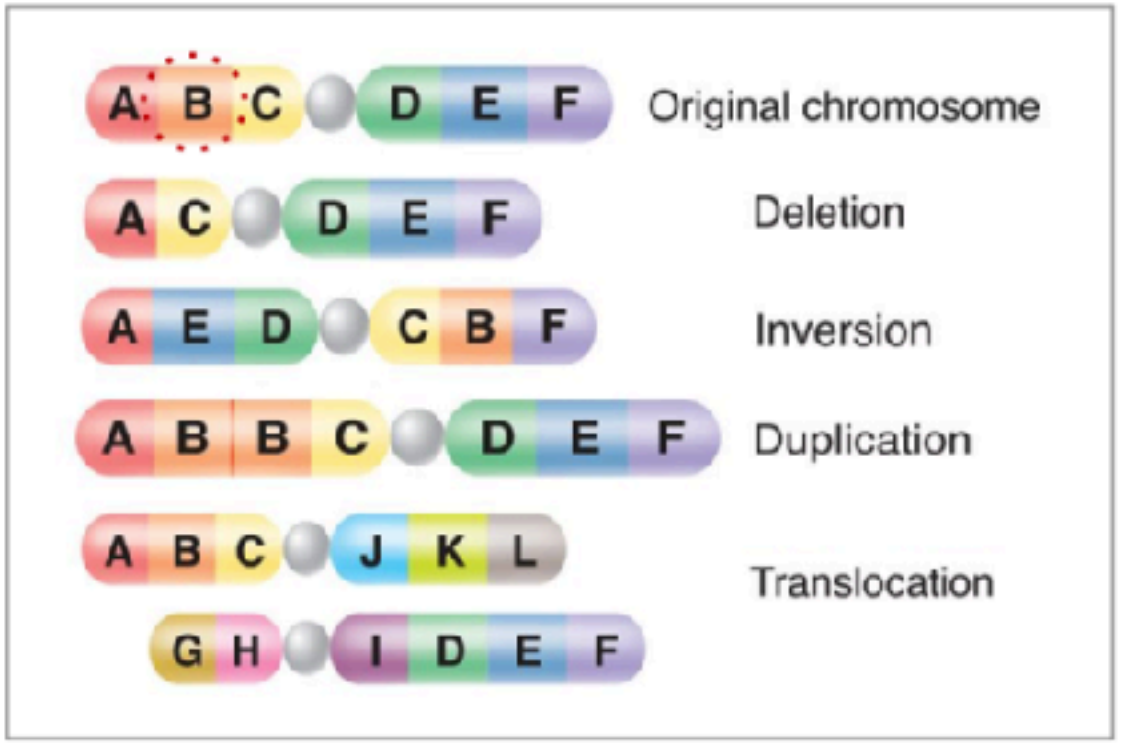
No concept ID is needed for '**Genomic_test**' table.



Concept ID Request

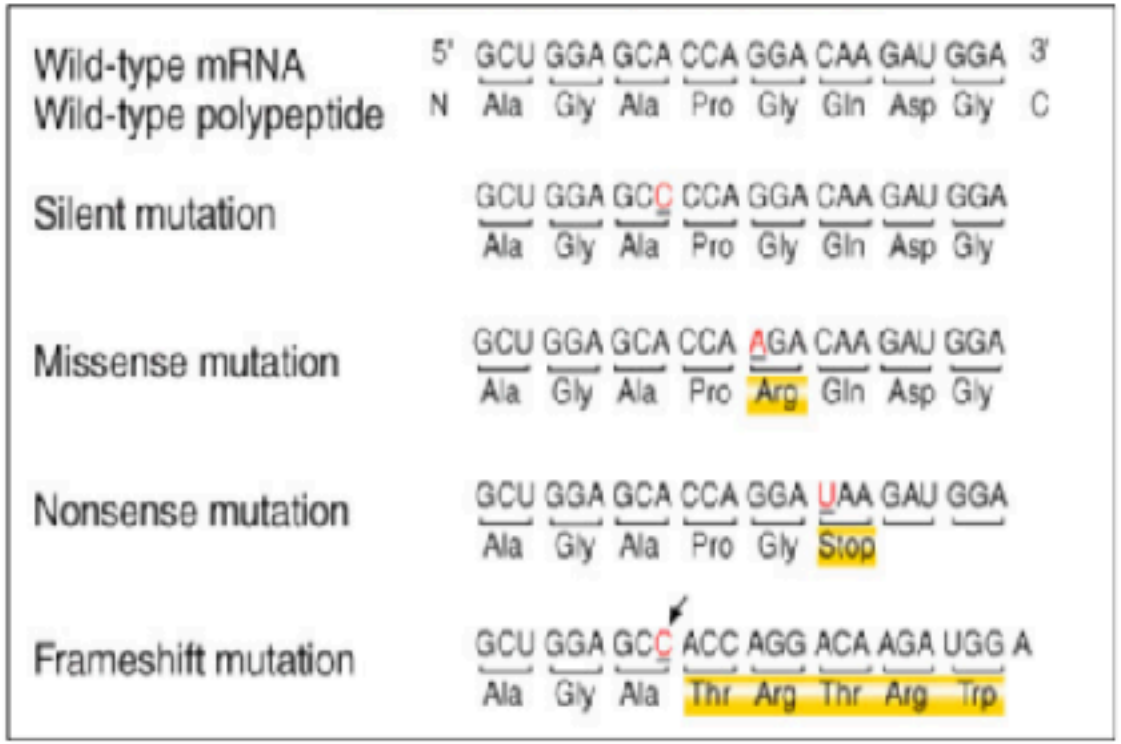
Concept IDs are needed for '**Variant_occurrence**' table.

1. Sequence Alteration



DNA Level **Structural** Variant Types

2. Feature Variant



Protein Level **Functional** Variant Types

Concept ID Request

10 more concept_id are needed to create for variant types.

#	Sequence_alteration (Structural Types)			Concept ID
1	SNP (Single Nucleotide Polymorphism)	SNV (Single Nucleotide Variation)	Substitution	**45880312 (Substitution)**
2	DIP (Deletion Insertion Polymorphism)	INS	Insertion	**45878601 (Insertion)**
3		DEL	Deletion	**45879448 (Deletion)**
4	MNP (Multiple nucleotide polymorphism)	MNP		**MNP**
5	CNV (Copy Number Variation)	Copy number gain	Amplification	**45878168 (Copy number gain)**
6		Copy number loss	Deletion	**45880603 (Copy number loss)**
7	Translocation	Translocation	Fusion	**21499257 (Translocation)**
8	MIXED	MIXED	Mixed / Complex	**21498573 (MIXED)**

#	Variant_feature (Functional Types)			Concept ID
1	Locus Region	Upstream		**Upstream**
2		Downstream		**Downstream**
3	Coding Region	Synonymous	Silent	**45879450 (Silent)**
4		Missense		**45881183 (Missense)**
5		Nonsense	Stop-gained / Stop-codon-mutation	**45879449 (Nonsense),** **45884101 (Stop Codon Mutation)**
6		Stop-loss	Stop-lost	**Stop-loss**
7		Frameshift		**45878252 (Frameshift)**
8		Inframe		**Inframe**
9	Untranslated Region	UTR	5_prime_UTR	**5_prime_UTR**
10			3_prime_UTR	**3_prime_UTR**
11	Intron Region	Intron		**Intron**
12	Spliece Site	Splice	Splice_donor_variant	**Splice_donor**
13			Splice_acceptor_variant	**Splice_acceptor**

HGNC 41,503 approved genes (ID and Symbol) —> Concept ID

Locus Group	Total by Locus Group
protein-coding gene	19197
non-coding RNA	7373
phenotype	571
pseudogene	13188
other	1174
Total Approved Symbols	41503

hgnc_id	symbol	name	locus_group	locus_type	status	location	location_so	alias_symbol	alias_name	prev_symbol	prev_name
HGNC:5	A1BG	alpha-1-B glycoprotein	protein-coding gene	gene with protein produ	Approved	19q13.43	19q13.43				
HGNC:37133	A1BG-AS1	A1BG antisense RNA 1	non-coding RNA	RNA, long non-coding	Approved	19q13.43	19q13.43	FLJ23569		NCRNA00181|A1	non-protein co
HGNC:24086	A1CF	APOBEC1 complementatio	protein-coding gene	gene with protein produ	Approved	10q11.23	10q11.23	ACF|ASP|ACF64|ACF65|A		APOBEC1CF	
HGNC:7	A2M	alpha-2-macroglobulin	protein-coding gene	gene with protein produ	Approved	12p13.31	12p13.31	FWP007|S863-7|CPAMD5			
HGNC:27057	A2M-AS1	A2M antisense RNA 1	non-coding RNA	RNA, long non-coding	Approved	12p13.31	12p13.31				A2M antisense
HGNC:23336	A2ML1	alpha-2-macroglobulin like	protein-coding gene	gene with protein produ	Approved	12p13.31	12p13.31	FLJ25179|p170		CPAMD9	C3 and PZP-l
HGNC:41022	A2ML1-AS1	A2ML1 antisense RNA 1	non-coding RNA	RNA, long non-coding	Approved	12p13.31	12p13.31				A2ML1 antisen
HGNC:41523	A2ML1-AS2	A2ML1 antisense RNA 2	non-coding RNA	RNA, long non-coding	Approved	12p13.31	12p13.31				A2ML1 antisen
HGNC:8	A2MP1	alpha-2-macroglobulin pse	pseudogene	pseudogene	Approved	12p13.31	12p13.31			A2MP	alpha-2-macro
HGNC:30005	A3GALT2	alpha 1,3-galactosyltransfe	protein-coding gene	gene with protein produ	Approved	1p35.1	01p35.1	IGBS3S|IGB3	iGb3 synthas	A3GALT2P	alpha 1,3-gala
HGNC:18149	A4GALT	alpha 1,4-galactosyltransfe	protein-coding gene	gene with protein produ	Approved	22q13.2	22q13.2	A14GALT|Gb	Gb3 synthas	P1	alpha 1,4-gala
HGNC:17968	A4GNT	alpha-1,4-N-acetylglucosan	protein-coding gene	gene with protein produ	Approved	3q22.3	03q22.3	alpha4GnT			
HGNC:13666	AAAS	aladin WD repeat nucleopo	protein-coding gene	gene with protein produ	Approved	12q13.13	12q13.13		aladin|Allgrove, triple-A|adraca		achalasia, adi
HGNC:21298	AACS	acetoacetyl-CoA synthetase	protein-coding gene	gene with protein produ	Approved	12q24.31	12q24.31	FLJ12389|SU	acyl-CoA synthetase family member 1		
HGNC:18226	AACSP1	acetoacetyl-CoA synthetase	pseudogene	pseudogene	Approved	5q35.3	05q35.3			AACSL	acetoacetyl-Co
HGNC:17	AADAC	arylacetamide deacetylase	protein-coding gene	gene with protein produ	Approved	3q25.1	03q25.1	DAC|CES5A1			arylacetamide
HGNC:24427	AADACL2	arylacetamide deacetylase	protein-coding gene	gene with protein produ	Approved	3q25.1	03q25.1	MGC72001			
HGNC:50301	AADACL2-AS1	AADACL2 antisense RNA 1	non-coding RNA	RNA, long non-coding	Approved	3q25.1	03q25.1				
HGNC:32037	AADACL3	arylacetamide deacetylase	protein-coding gene	gene with protein produ	Approved	1p36.21	01p36.21	OTTHUMG00000001887			
HGNC:32038	AADACL4	arylacetamide deacetylase	protein-coding gene	gene with protein produ	Approved	1p36.21	01p36.21	OTTHUMG00000001889			
HGNC:50305	AADACP1	arylacetamide deacetylase	pseudogene	pseudogene	Approved	3q25.1	03q25.1				
HGNC:17929	AADAT	aminoadipate aminotransfe	protein-coding gene	gene with protein produ	Approved	4q33	04q33	KAT||KAT2|K	kynurenine aminotransferase |||L	kynurenine/	
HGNC:25662	AAGAB	alpha and gamma adaptin b	protein-coding gene	gene with protein produ	Approved	15q23	15q23	FLJ11506|p34			
HGNC:19679	AAK1	AP2 associated kinase 1	protein-coding gene	gene with protein produ	Approved	2p13.3	02p13.3	KIAA1048|DKFZp686K16132			
HGNC:30205	AAMDC	adipogenesis associated M	protein-coding gene	gene with protein produ	Approved	11q14.1	11q14.1	PTD015|FLJ21035|CK067		C11orf67	chromosome
HGNC:18	AAMP	angio associated migratory	protein-coding gene	gene with protein produ	Approved	2q35	02q35				
HGNC:19	AANAT	aralkylamine N-acetyltransf	protein-coding gene	gene with protein produ	Approved	17q25.1	17q25.1	SNAT	serotonin N-acetyltransferase		arylalkylamine
HGNC:15886	AAR2	AAR2 splicing factor homol	protein-coding gene	gene with protein produ	Approved	20q11.23	20q11.23	bA234K24.2		C20orf4	chromosome 2
HGNC:33842	AARD	alanine and arginine rich do	protein-coding gene	gene with protein produ	Approved	8q24.11	08q24.11	LOC441376	Alanine and	C8orf85	chromosome 8
HGNC:20	AARS	alanyl-tRNA synthetase	protein-coding gene	gene with protein produ	Approved	16q22.1	16q22.1	CMT2N|AlaR	alanine tRNA ligase 1, cytoplasmic		
HGNC:21022	AARS2	alanyl-tRNA synthetase 2, r	protein-coding gene	gene with protein produ	Approved	6p21.1	06p21.1	KIAA1270|bA	alanine tRNA	AARSL	alanyl-tRNA s
HGNC:28417	AARSD1	alanyl-tRNA synthetase dor	protein-coding gene	gene with protein produ	Approved	17q21.31	17q21.31	MGC2744			

Concept ID Request

Concept IDs are needed for '**Variant_annotation**' table.

Tier Level

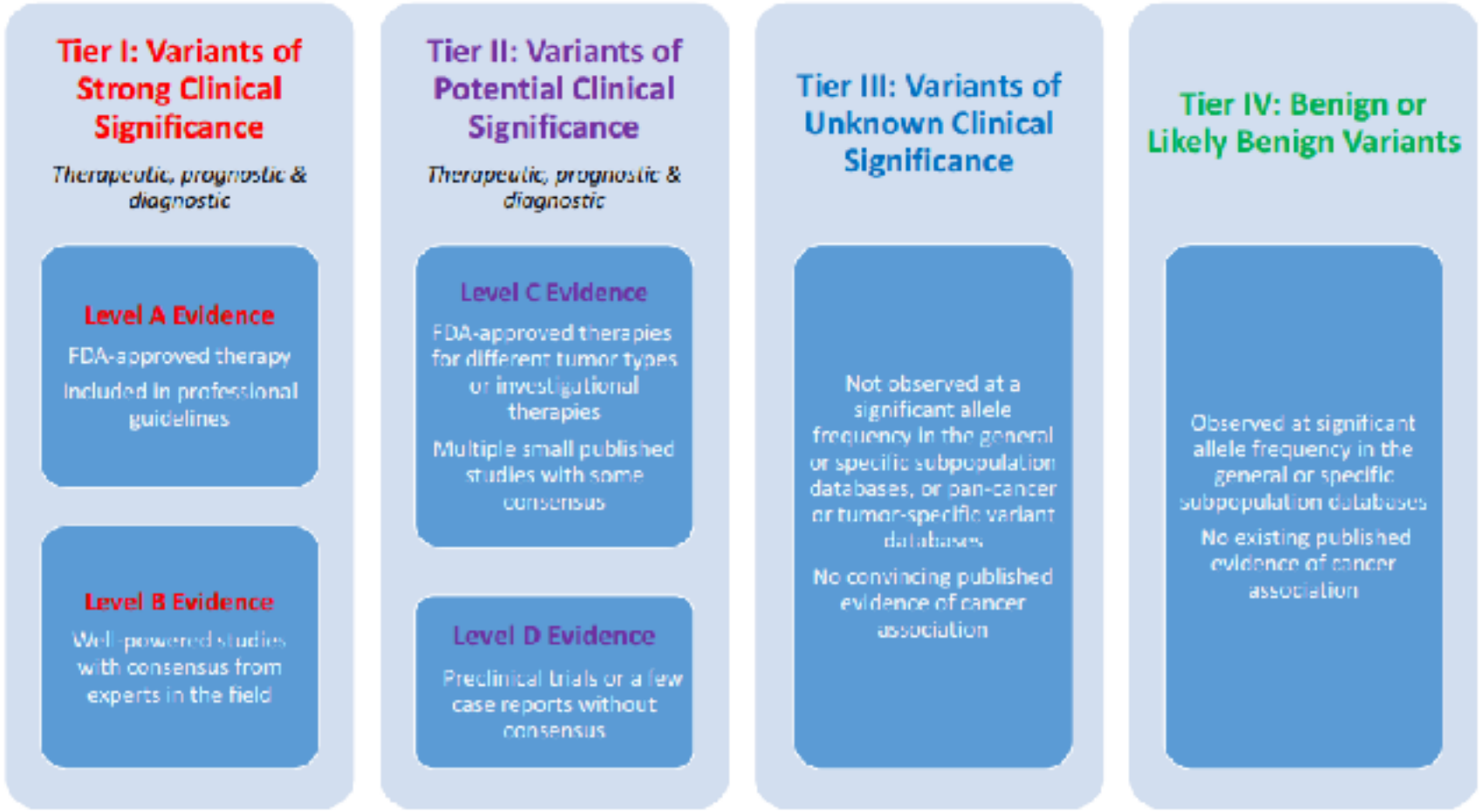
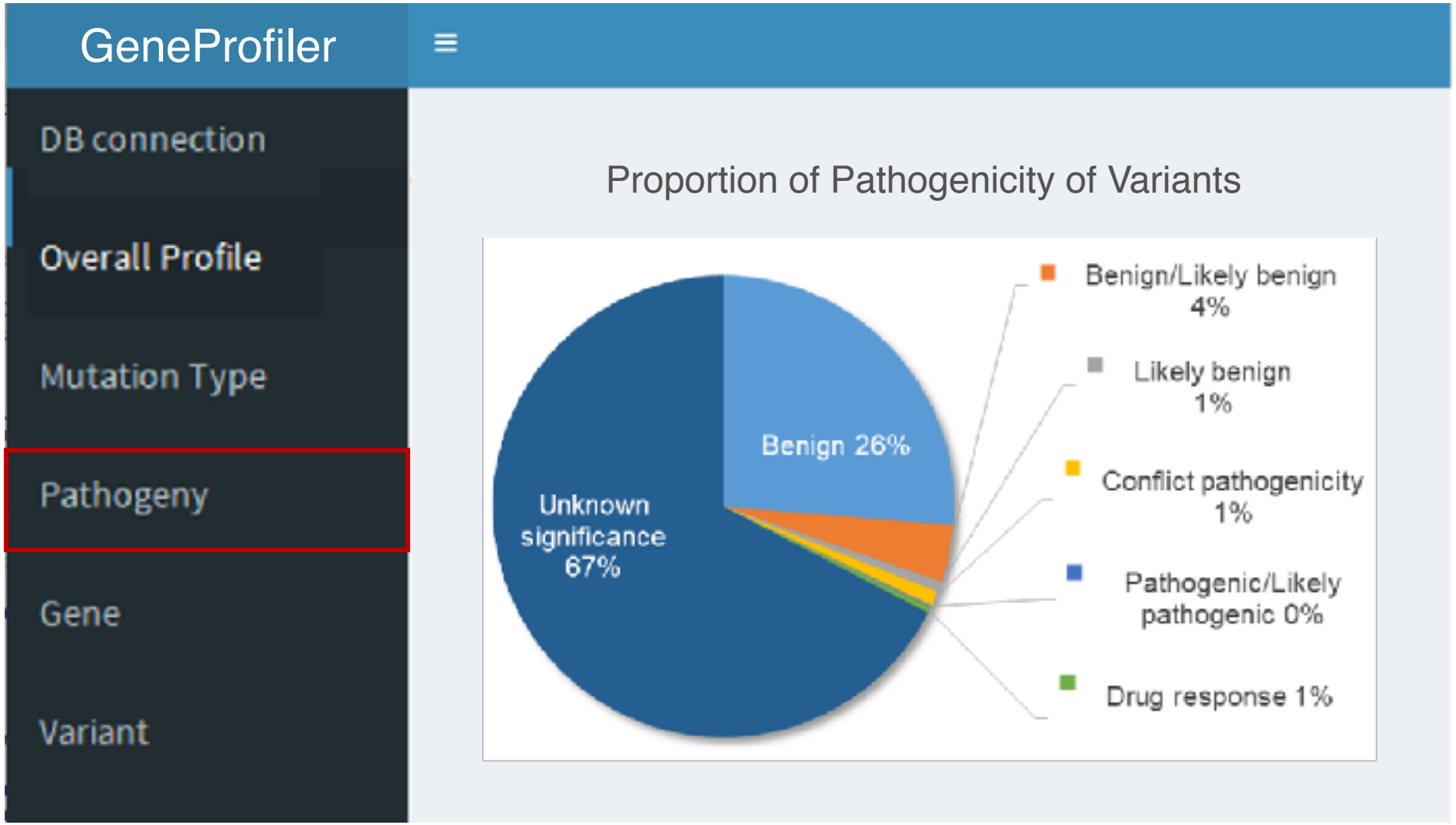


Figure 2 Evidence-based variant categorization. Somatic variants are classified into four tiers based on their level of clinical significance in cancer diagnosis, prognosis, and/or therapeutics. Variants in tier I are of strongest clinical significance, and variants in tier IV are benign or likely benign variants. FDA, Food and Drug Administration.

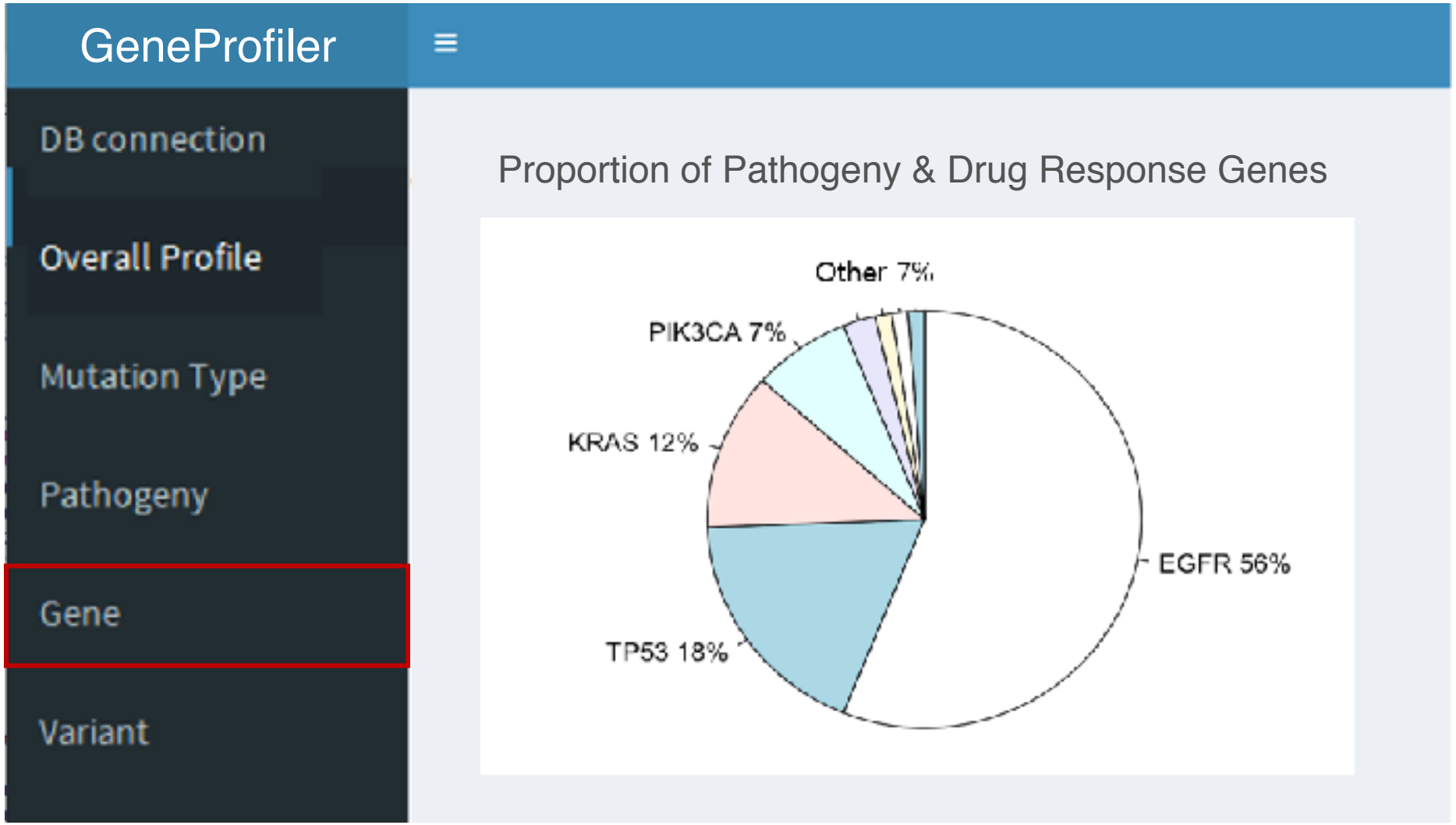
Use-case of the Concept ID

Data Profiling Tool for Genomic Data



Use-case of the Concept ID

Data Profiling Tool for Genomic Data



Use-case of the Concept ID

Data Profiling Tool for Genomic Data