Variant Representation, Annotation, and Interpretation

Sequencing Technologies and Bioinformatics Analysis 2021
Cold Spring Harbor Laboratories

You should be familiar with...

- NGS File Formats
- Fundamentals of sequence alignment
- Variant calling

Variant Representation

- How a variant is represented varies depending on use case

- Some standard variant representation formats:
  - Variant Call Format (VCF)
  - Human Genome Variation Society (HGVS) Variant Nomenclature
  - National Center for Biotechnology Information (NCBI)
    Sequence Position Deletion Insertion (SPDI) specification
  - Global Alliance for Genomics and Health (GA4GH)
    Variation Representation Specification (VRS)

Variant Call Format

- Compact representation of *many genomic variants* over *many samples*
- Useful for large-scale genomic sequencing projects, such as the 1000 Genomes Project
- Predominant output format for variant calling in bioinformatics pipelines
- Optimized for short / non-complex variants on a genomic reference

Variant Call Format - Header

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS      ID         REF   ALT    QUAL  FILTER   INFO                              FORMAT       NA00001         NA00002          NA00003
20     14370    rs6054257  G     A      29    PASS     NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ  0|0:48:1:51,51  1|0:48:8:51,51   1/1:43:5:.,.
20     17330    .          T     A      3     q10      NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ  0|0:49:3:58,50  0|1:3:5:65,3     0/0:41:3
20     1110696  rs6040355  A     G,T    67    PASS     NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ  1|2:21:6:23,27  2|1:2:0:18,2     2/2:35:4
20     1230237  .          T     .      47    PASS     NS=3;DP=13;AA=T                   GT:GQ:DP:HQ  0|0:54:7:56,60  0|0:48:4:51,51   0/0:61:2
20     1234567  microsat1  GTC   G,GTCT 50    PASS     NS=3;DP=9;AA=G                    GT:GQ:DP     0/1:35:4        0/2:17:2         1/1:40:3
```

Variant Call Format - Columns

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF  ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G    A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T    A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A    G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T    .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC  G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Variant Call Format - Columns

	Name	Brief description (see the specification for details).
1	CHROM	The name of the sequence (typically a chromosome) on which the variation is being called. This sequence is usually known as 'the reference sequence', i.e. the sequence against which the given sample varies.
2	POS	The 1-based position of the variation on the given sequence.
3	ID	The identifier of the variation, e.g. a dbSNP rs identifier, or if unknown a ".". Multiple identifiers should be separated by semi-colons without white-space.
4	REF	The reference base (or bases in the case of an indel) at the given position on the given reference sequence.
5	ALT	The list of alternative alleles at this position.
6	QUAL	A quality score associated with the inference of the given alleles.
7	FILTER	A flag indicating which of a given set of filters the variation has passed.
8	INFO	An extensible list of key-value pairs (fields) describing the variation. See below for some common fields. Multiple fields are separated by semicolons with optional values in the format: `<key>=<data>[,data]` .
9	FORMAT	An (optional) extensible list of fields for describing the samples.
+	SAMPLEs	For each (optional) sample described in the file, values are given for the fields listed in FORMAT

Variant Call Format - Info

	Name	Brief description (see the specification for details).
1	CHROM	The name of the sequence (typically a chromosome) on which the variation is being called. This sequence is usually known as 'the reference sequence', i.e. the sequence against which the given sample varies.
2	POS	The 1-based position of the variation on the given sequence.
3	ID	The identifier of the variation, e.g. a dbSNP rs identifier, or if unknown a ".". Multiple identifiers should be separated by semi-colons without white-space.
4	REF	The reference base (or bases in the case of an indel) at the given position on the given reference sequence.
5	ALT	The list of alternative alleles at this position.
6	QUAL	A quality score associated with the inference of the given alleles.
7	FILTER	A flag indicating which of a given set of filters the variation has passed.
8	INFO	An extensible list of key-value pairs (fields) describing the variation. See below for some common fields. Multiple fields are separated by semicolons with optional values in the format: `<key>=<data>[,data]` .
9	FORMAT	An (optional) extensible list of fields for describing the samples.
+	SAMPLEs	For each (optional) sample described in the file, values are given for the fields listed in FORMAT

VCF - Info fields

Name	Brief description
AA	ancestral allele
AC	allele count in genotypes, for each ALT allele, in the same order as listed
AF	allele frequency for each ALT allele in the same order as listed (use this when estimated from primary data, not called genotypes)
AN	total number of alleles in called genotypes
BQ	RMS base quality at this position
CIGAR	cigar string describing how to align an alternate allele to the reference allele
DB	dbSNP membership
DP	combined depth across samples, e.g. DP=154
END	end position of the variant described in this record (for use with symbolic alleles)
H2	membership in hapmap2
H3	membership in hapmap3
MQ	RMS mapping quality, e.g. MQ=52
MQ0	Number of MAPQ == 0 reads covering this record
NS	Number of samples with data
SB	strand bias at this position
SOMATIC	indicates that the record is a somatic mutation, for cancer genomics
VALIDATED	validated by follow-up experiment
1000G	membership in 1000 Genomes

VCF - Info fields

Name	Brief description
AA	ancestral allele
AC	allele count in genotypes, for each ALT allele, in the same order as listed
AF	allele frequency for each ALT allele in the same order as listed (use this when estimated from primary data, not called genotypes)
AN	total number of alleles in called genotypes
BQ	RMS base quality at this position
CIGAR	cigar string describing how to align an alternate allele to the reference allele
DB	dbSNP membership
DP	combined depth across samples, e.g. DP=154
END	end position of the variant described in this record (for use with symbolic alleles)
H2	membership in hapmap2
H3	membership in hapmap3
MQ	RMS mapping quality, e.g. MQ=52
MQ0	Number of MAPQ == 0 reads covering this record
NS	Number of samples with data
SB	strand bias at this position
SOMATIC	indicates that the record is a somatic mutation, for cancer genomics
VALIDATED	validated by follow-up experiment
1000G	membership in 1000 Genomes

Variant Call Format - Format and Samples

	Name	Brief description (see the specification for details).
1	CHROM	The name of the sequence (typically a chromosome) on which the variation is being called. This sequence is usually known as 'the reference sequence', i.e. the sequence against which the given sample varies.
2	POS	The 1-based position of the variation on the given sequence.
3	ID	The identifier of the variation, e.g. a dbSNP rs identifier, or if unknown a ".". Multiple identifiers should be separated by semi-colons without white-space.
4	REF	The reference base (or bases in the case of an indel) at the given position on the given reference sequence.
5	ALT	The list of alternative alleles at this position.
6	QUAL	A quality score associated with the inference of the given alleles.
7	FILTER	A flag indicating which of a given set of filters the variation has passed.
8	INFO	An extensible list of key-value pairs (fields) describing the variation. See below for some common fields. Multiple fields are separated by semicolons with optional values in the format: `<key>=<data>[,data]` .
9	FORMAT	An (optional) extensible list of fields for describing the samples.
+	SAMPLEs	For each (optional) sample described in the file, values are given for the fields listed in FORMAT

VCF - Format Fields

Name	Brief description
AD	Read depth for each allele
ADF	Read depth for each allele on the forward strand
ADR	Read depth for each allele on the reverse strand
DP	Read depth
EC	Expected alternate allele counts
FT	Filter indicating if this genotype was “called”
GL	Genotype likelihoods
GP	Genotype posterior probabilities
GQ	Conditional genotype quality
GT	Genotype
HQ	Haplotype quality
MQ	RMS mapping quality
PL	Phred-scaled genotype likelihoods rounded to the closest integer
PQ	Phasing quality
PS	Phase set

VCF - Format Fields

Name	Brief description
AD	Read depth for each allele
ADF	Read depth for each allele on the forward strand
ADR	Read depth for each allele on the reverse strand
DP	Read depth
EC	Expected alternate allele counts
FT	Filter indicating if this genotype was “called”
GL	Genotype likelihoods
GP	Genotype posterior probabilities
GQ	Conditional genotype quality
GT	Genotype
HQ	Haplotype quality
MQ	RMS mapping quality
PL	Phred-scaled genotype likelihoods rounded to the closest integer
PQ	Phasing quality
PS	Phase set

VCF - Format Fields

- GT (String): Genotype, encoded as allele values separated by either of / or |. The allele values are 0 for the reference allele (what is in the REF field), 1 for the first allele listed in ALT, 2 for the second allele list in ALT and so on. For diploid calls examples could be 0/1, 1 | 0, or 1/2, etc. Haploid calls, e.g. on Y, male non-pseudoautosomal X, or mitochondrion, are indicated by having only one allele value. A triploid call might look like 0/0/1. If a call cannot be made for a sample at a given locus, '.' must be specified for each missing allele in the GT field (for example './.' for a diploid genotype and '.' for haploid genotype). The meanings of the separators are as follows (see the PS field below for more details on incorporating phasing information into the genotypes):
    - / : genotype unphased
    - | : genotype phased

Variant Call Format - Format and Samples

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
```

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00003
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51	0/0:61:2
20	1234567	microsat1	GTC	G,GTCT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

Variant Call Format - Genotype

Ref (0): A
Alt (1, 2): G, T

NA00001: 1|2 (G phased on haplotype 1, T phased on haplotype 2)
NA00002: 2|1 (T phased on haplotype 1, G phased on haplotype 2)
NA00003: 2/2 (Homozygous T)

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00003
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51	0/0:61:2
20	1234567	microsat1	GTC	G,GTCT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

Variant Call Format - Genotype

Ref (0): A

Alt (1, 2): G, T

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00003
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51	0/0:61:2
20	1234567	microsat1	GTC	G,GTCT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

Variant Call Format - Genotype

Ref (0): A
Alt (1, 2): G, T

NA00001: 1|2 (G phased on haplotype 1, T phased on haplotype 2)
NA00002: 2|1 (T phased on haplotype 1, G phased on haplotype 2)
NA00003: 2/2 (Homozygous T)

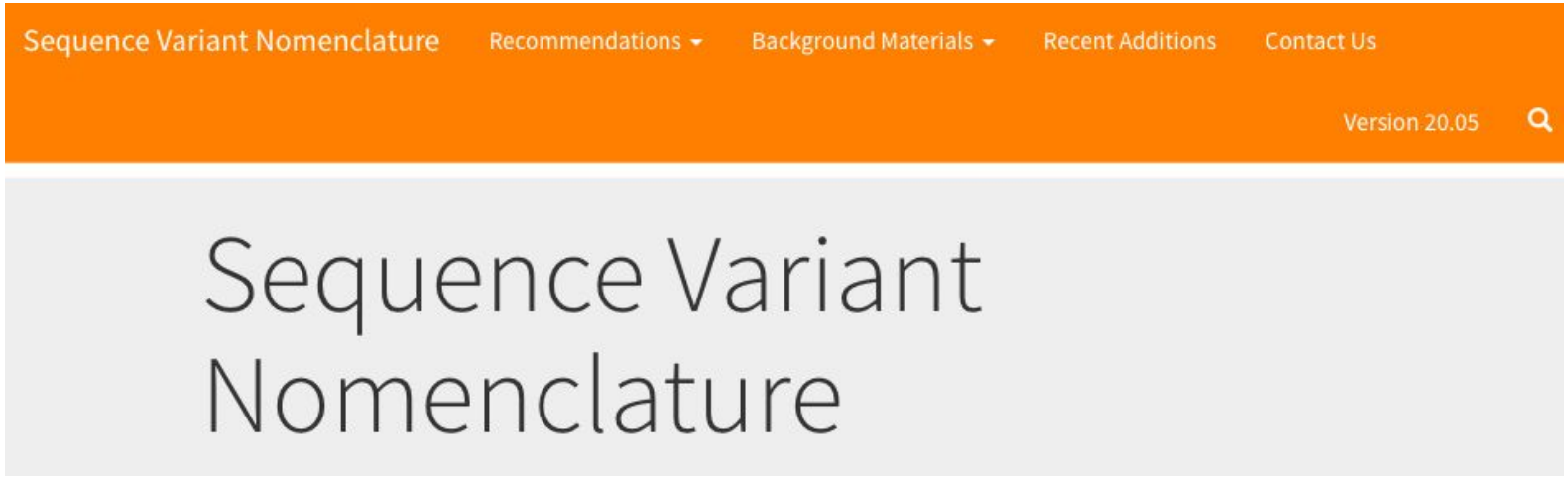
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00003
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51	0/0:61:2
20	1234567	microsat1	GTC	G,GTCT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

Questions about VCF?

HGVS Variant nomenclature

- Compact format for human-parsable variant description
- Useful for variant reporting in documents
- Predominant format in biomedical literature and human-readable UIs
- Emphasis on readability and reference sequence design
- Describes variants on any sequence (i.e. genome, transcript, protein)

HGVS Varnomen Resource



Interactive component. Follow along at: http://varnomen.hgvs.org/

Questions about HGVS?

SPDI variant format

- Simple format for sequence variants
- Useful for variant reporting in documents
- Mostly seen in NCBI resources (e.g. ClinVar and dbSNP)
- Emphasis on readability and computation
- Limited scope to simple variants

SPDI Format

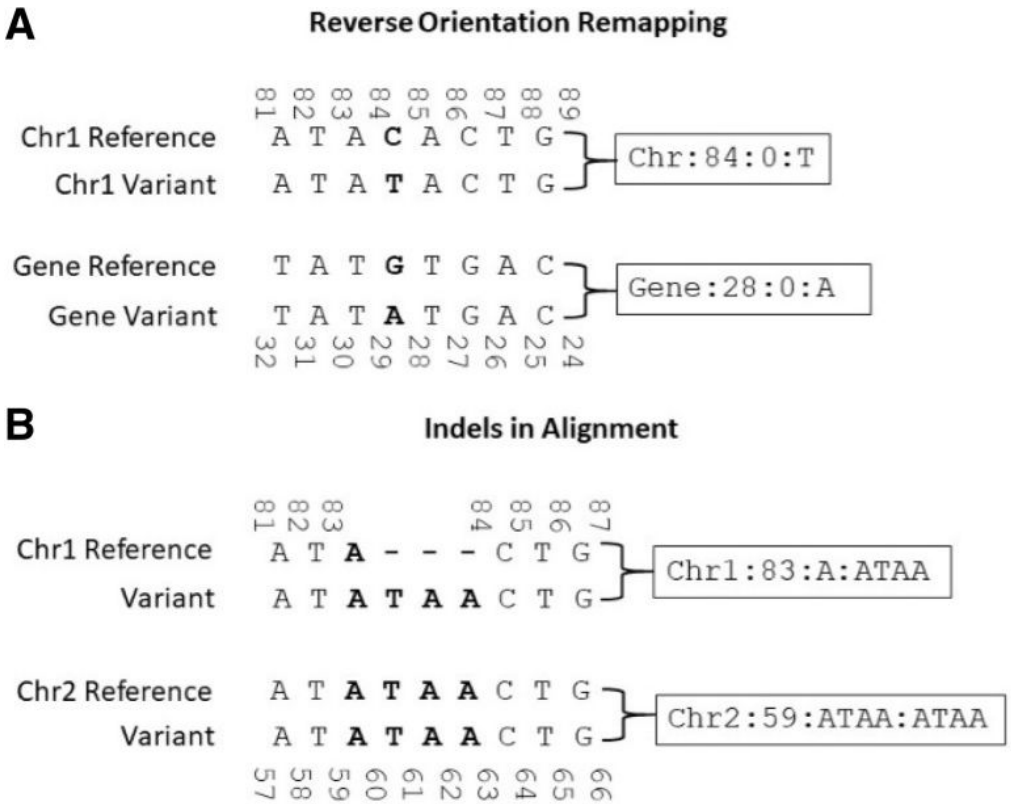
Sequence : **Position** : **Deletion** : **Insertion**

SPDI Format

NC_00001.1:12345:0:ATAA

Sequence : **Position** : **Deletion** : **Insertion**

SPDI - Orientation and Indels



Holmes, et al. Bioinformatics. (2020)
doi: 10.1093/bioinformatics/btz856.

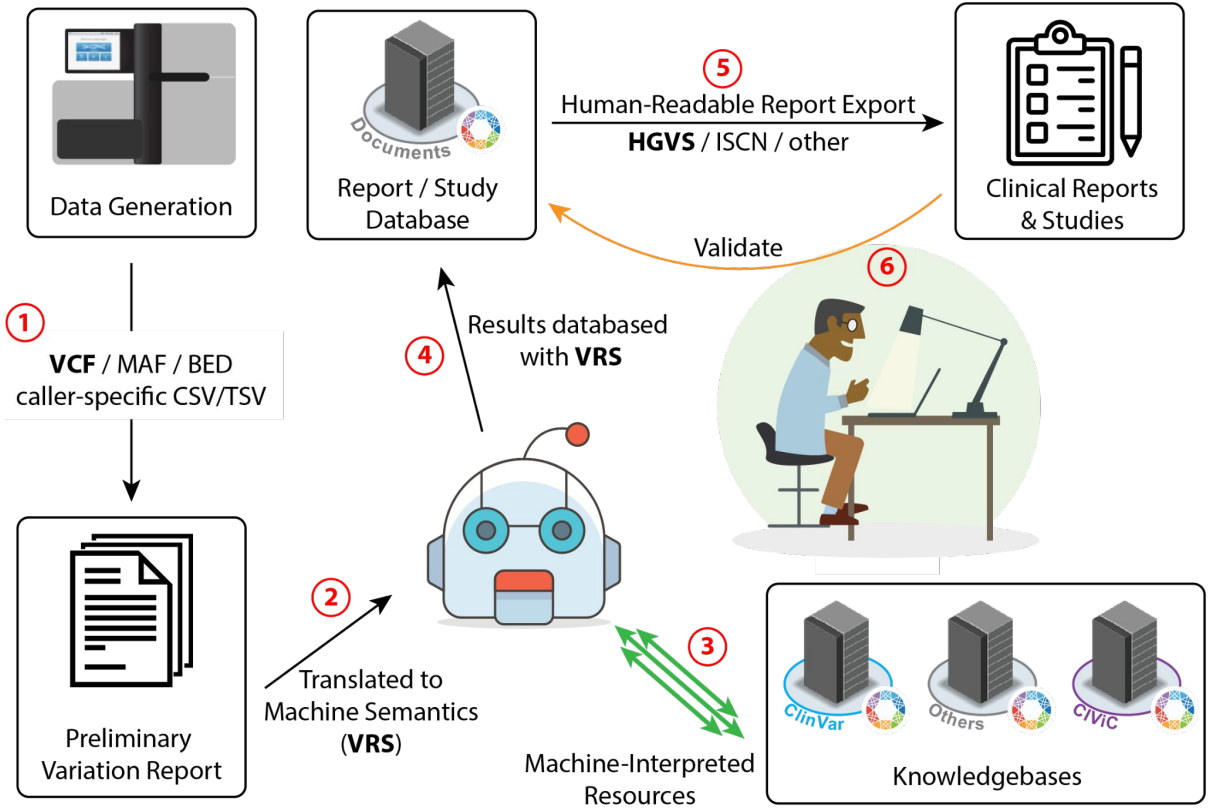
Questions about SPDI?

GA4GH Variation Representation Specification

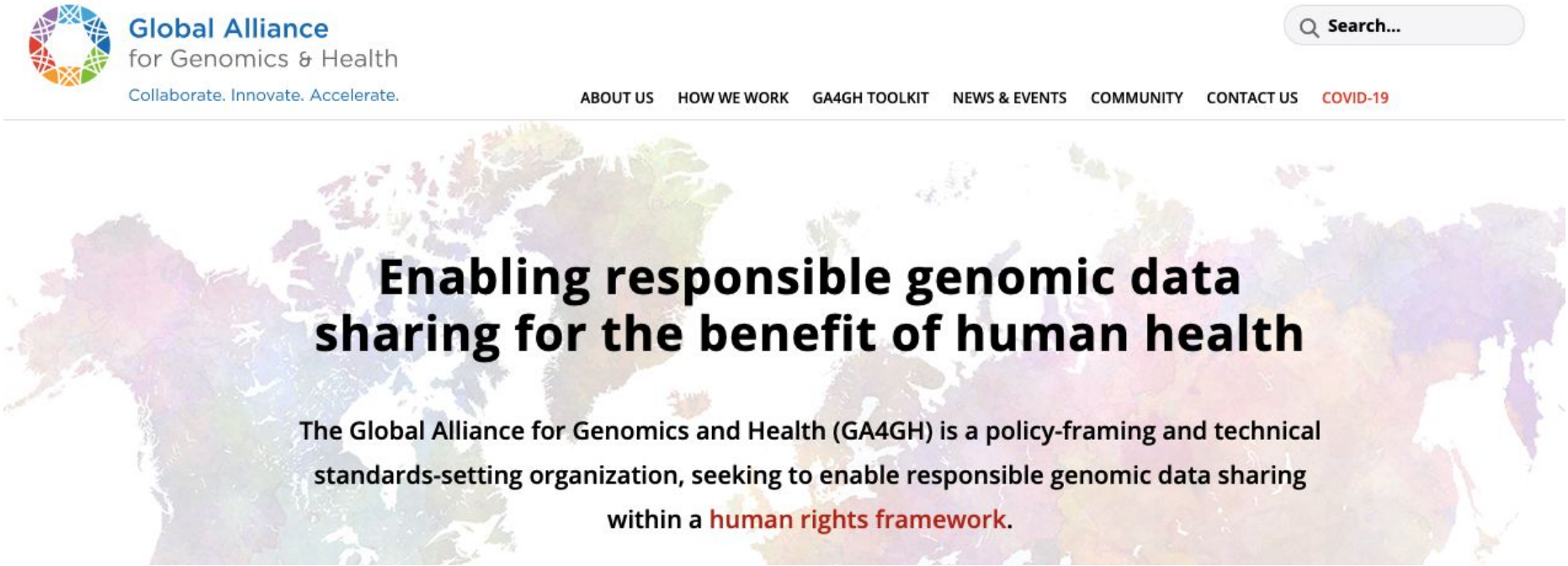
- Computable format for all forms of biomolecular variation
- Useful for association with Real World Evidence
- New format, natively supported in a handful of resources
- Emphasis on computability and value-object design
- Broadest scope; covers variation across multiple coordinate systems

VRS Provides Mechanism for Scalability

Global Alliance for Genomics & Health

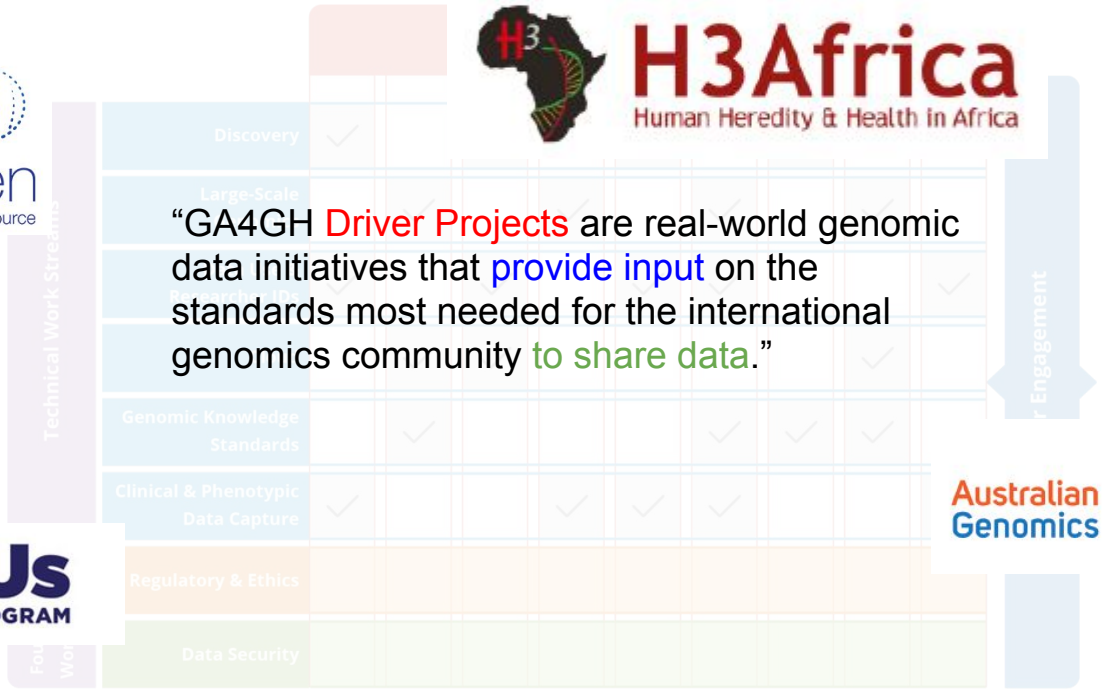


The Global Alliance for Genomics and Health



https://www.ga4gh.org/

The GA4GH Working Model



ClinGen
Clinical Genome Resource

H3Africa
Human Heredity & Health in Africa

GEM Japan

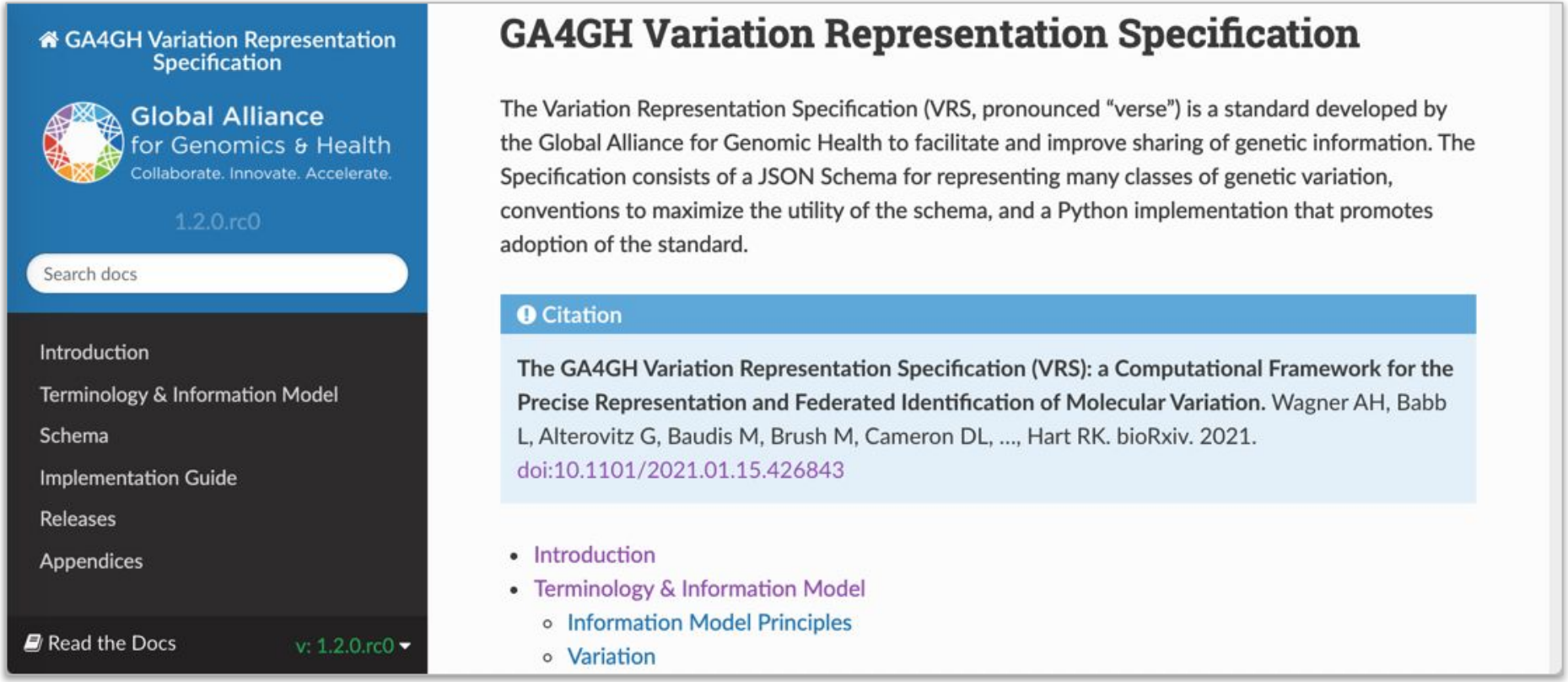
"GA4GH Driver Projects are real-world genomic data initiatives that provide input on the standards most needed for the international genomics community to share data."

VICC

All of Us
RESEARCH PROGRAM

Australian Genomics

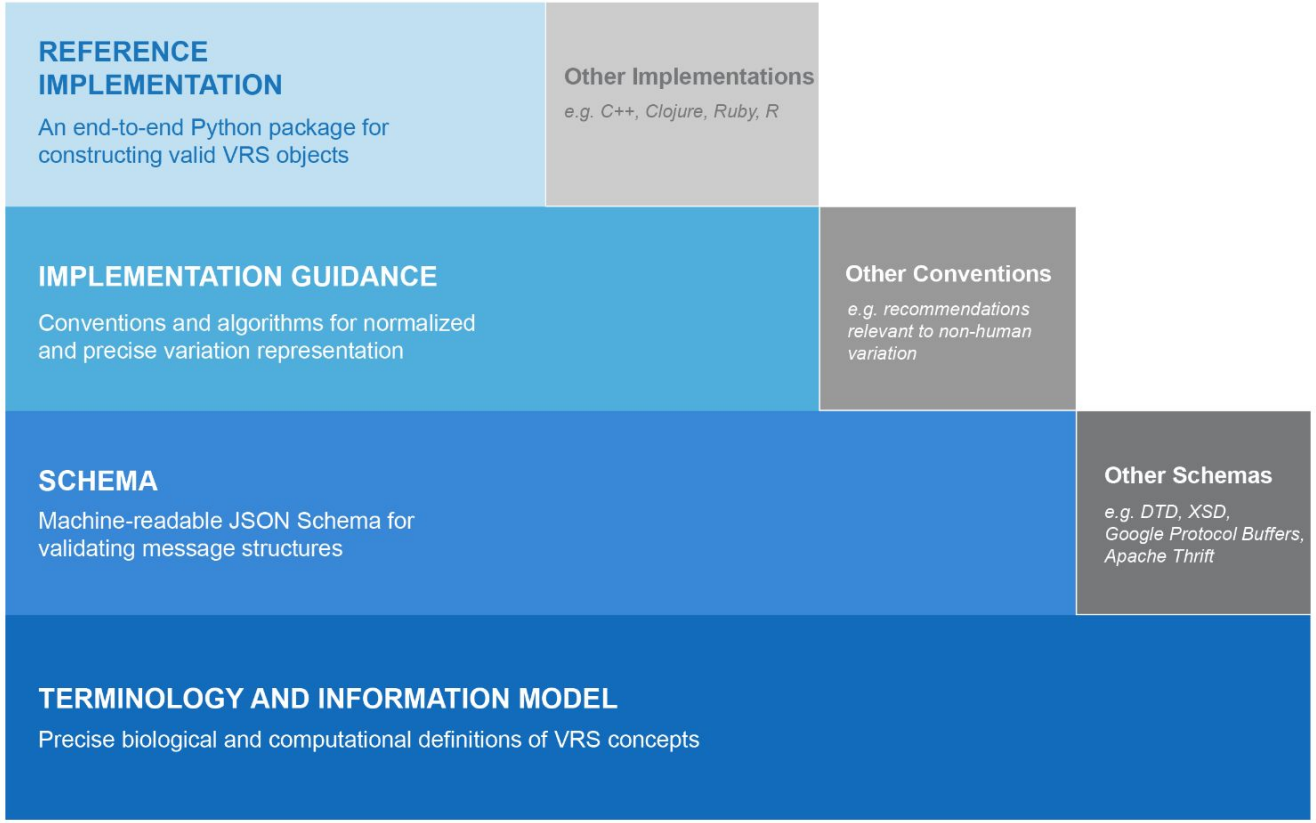
https://www.ga4gh.org/how-we-work/



https://vrs.ga4gh.org

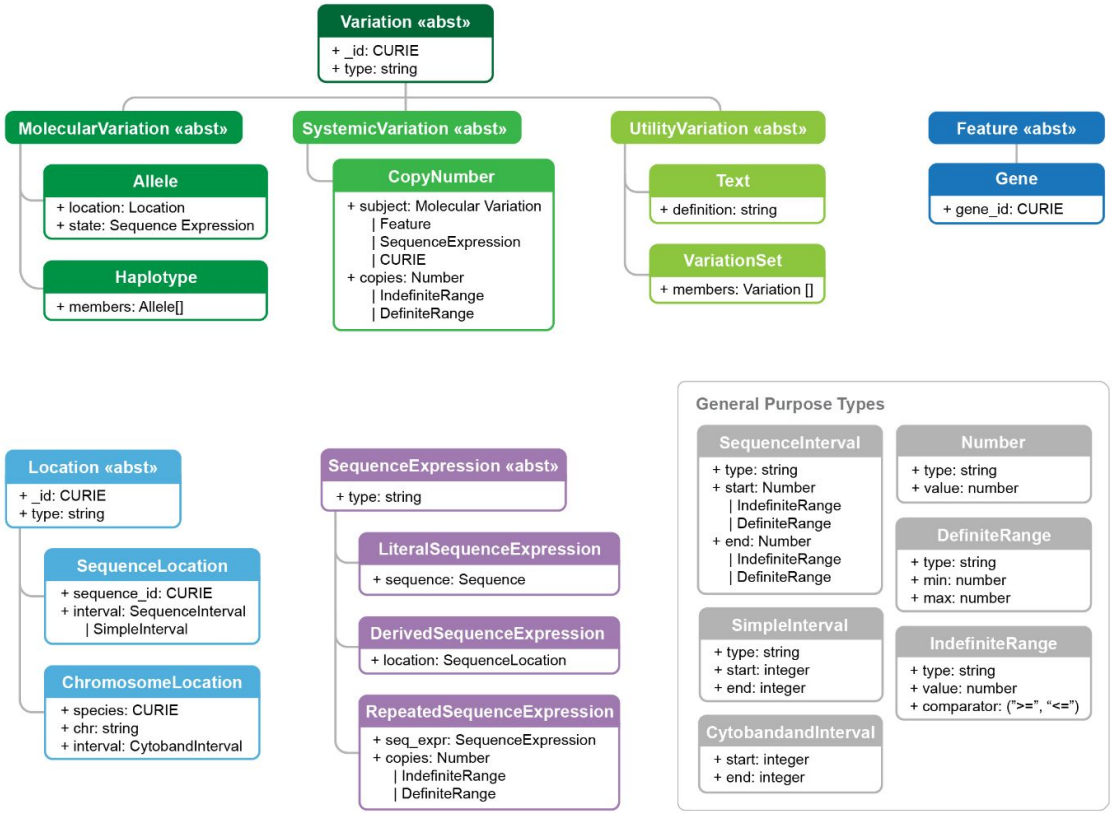
Components of VRS

Global Alliance for Genomics & Health



Extensible Information Model (v 1.2 and going)

Global Alliance for Genomics & Health



VRS objects are value objects

VRS objects are intentionally designed to be **value objects**.

Value objects represent entities whose equality is **based on the values of its attributes, not an identity**[1].

[1] Value Objects, Wikipedia, https://en.wikipedia.org/wiki/Value_object

Value objects are defined by attributes

GA4GH headquarters address as a location **identity**:

661 University Avenue, Suite 510
Toronto, Ontario
Canada

Address is **registered** by city of Toronto

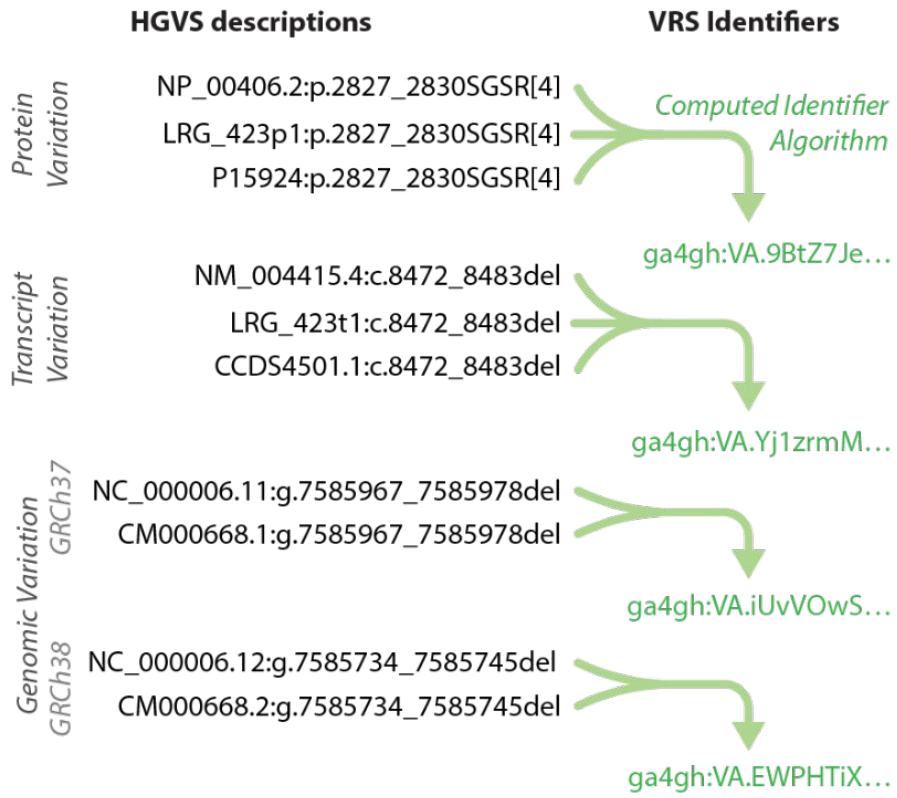
GA4GH headquarters coordinates as a location **value object**:

Latitude: 43.6595
Longitude: -79.3897

Coordinates are a **definitional property**

VRS provides unique variation identifiers

Global Alliance for Genomics & Health



All of these are the same variant. Or not.

Global Alliance for Genomics & Health

```
NC_000001.10:g.103471457_103471459delCAT
= NC_000001.10:g.103471486_103471488delTCA
```
Right shifted per HGVS Nomenclature guidelines (ClinVar Id 93966)

```
NM_001166478.1:c.30_31insT
= NM_001166478.1:c.35dupT
```
Normalized and rewritten

```
NM_080588.2:c.139C>G (rs4073458)
= ENST00000367279:c.139C>G
```
Has identical CDS and exon structure, including UTR

```
NP_003768.2:p.(Arg4412Alafs*2)
= NP_003768.2:p.(Arg4412Alafs)
= NP_003768.2:p.(Arg4412AlaTrpTer)
```
Same protein truncation (rs72658833; + wo/parens and 1-letter forms!)

Fully-Justified Normalization Captures Region of Shuffling Ambiguity

Global Alliance for Genomics & Health

Normalization Example: In sequence TCAGCAGCT, replace CA at bases 5-6 with CAGCA

Actual location of variation is ambiguous due to the sequence context

(HGVS format: S:g.5_6delinsCAGCA)

$$TCAG\left[\frac{CA}{CAGCA}\right]GCT$$

Holmes, J.B., Moyer, E., Phan, L., Maglott, D. & Kattman, B. L.
SPDI: Data Model for Variants and Applications at NCBI.
bioRxiv 537449 (2019). doi:10.1101/537449

Design decision: inter-residue coordinates

Global Alliance for Genomics & Health

Sequence	T	C	A	G	C	A	G	C	A	G	C	T
Residue	1	2	3	4	5	6	7	8	9	10	11	12

Design decision: inter-residue coordinates

Global Alliance for Genomics & Health

Insertion between AG in Sequence																								
Sequence		T		C		A		G		C		A		G		C		A		G		C		T
Residue		1		2		3		4		5		6		7		8		9		10		11		12

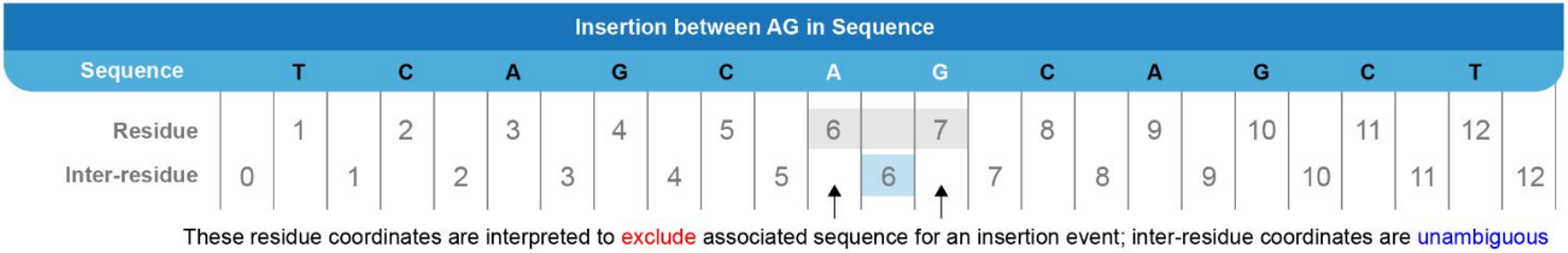
Design decision: inter-residue coordinates

Global Alliance for Genomics & Health

Insertion between AG in Sequence																									
Sequence		T		C		A		G		C		A		G		C		A		G		C		T	
Residue		1		2		3		4		5		6		7		8		9		10		11		12	
Inter-residue	0		1		2		3		4		5		6		7		8		9		10		11		12

Design decision: inter-residue coordinates

Global Alliance for Genomics & Health



These residue coordinates are interpreted to exclude associated sequence for an insertion event; inter-residue coordinates are unambiguous

Design decision: inter-residue coordinates

Global Alliance for Genomics & Health

Insertion between AG in Sequence

Sequence		T		C		A		G		C		A	G		C		A		G		C		T		
Residue		1		2		3		4		5		6		7		8		9		10		11		12	
Inter-residue	0		1		2		3		4		5		6		7		8		9		10		11		12

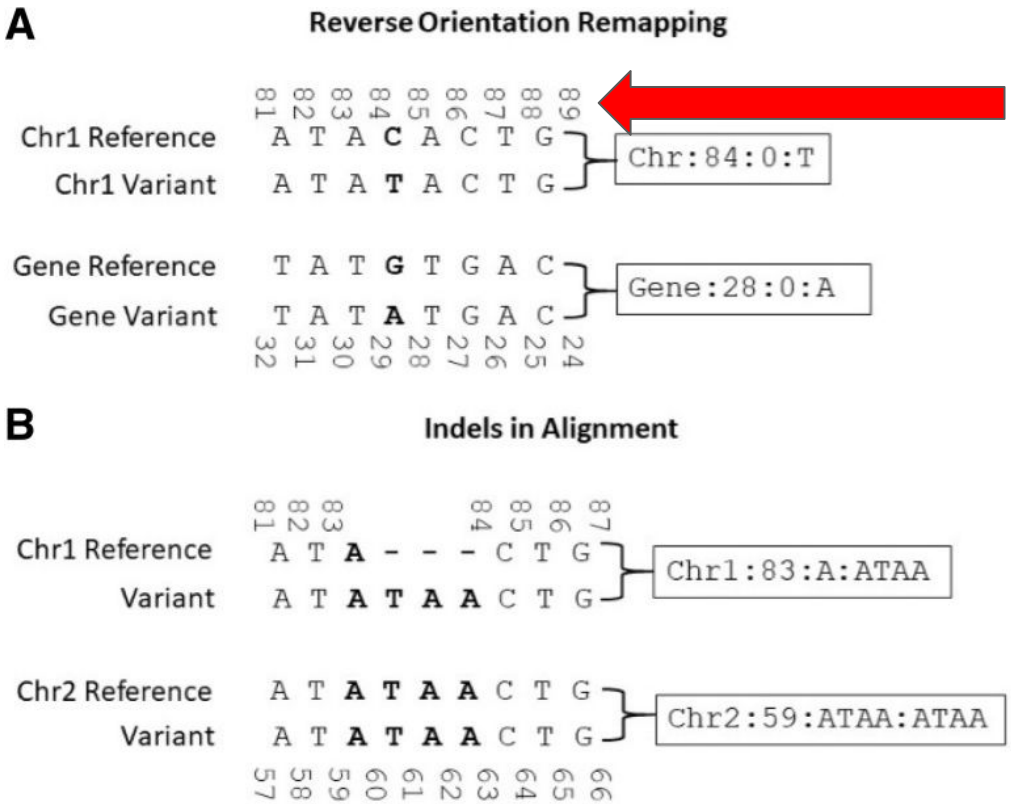
These residue coordinates are interpreted to exclude associated sequence for an insertion event; inter-residue coordinates are unambiguous

Deletion/Substitution of AG in Sequence

Sequence		T		C		A		G		C		A	G		C		A		G		C		T		
Residue		1		2		3		4		5		6		7		8		9		10		11		12	
Inter-residue	0		1		2		3		4		5		6		7		8		9		10		11		12

The same residue coordinates are interpreted to include associated sequence for a deletion or substitution event; inter-residue coordinates remain unambiguous

SPDI - Orientation and Indels



Holmes, et al. Bioinformatics. (2020)
doi: 10.1093/bioinformatics/btz856.

Fully-Justified Normalization Captures Region of Shuffling Ambiguity

Global Alliance for Genomics & Health

Normalization Example: In sequence TCAGCAGCT, replace CA at bases 5-6 with CAGCA

Actual location of variation is ambiguous due to the sequence context

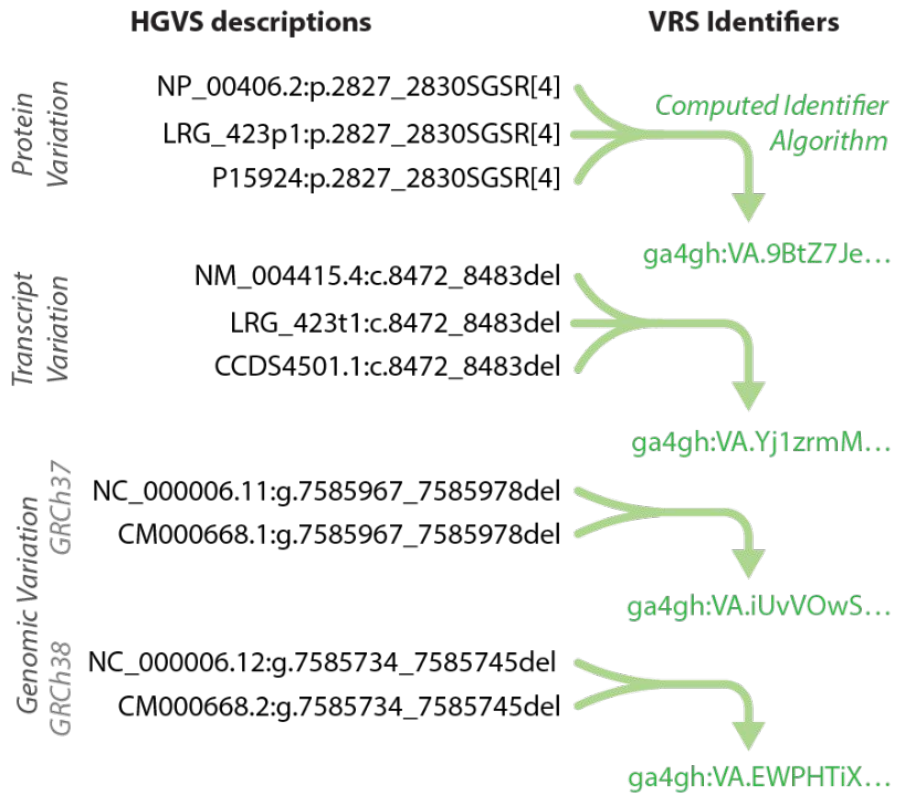
(HGVS format: S:g.5_6delinsCAGCA)

$$TCAG\left[\frac{CA}{CAGCA}\right]GCT$$

Holmes, J.B., Moyer, E., Phan, L., Maglott, D. & Kattman, B. L.
SPDI: Data Model for Variants and Applications at NCBI.
bioRxiv 537449 (2019). doi:10.1101/537449

VRS provides unique variation identifiers

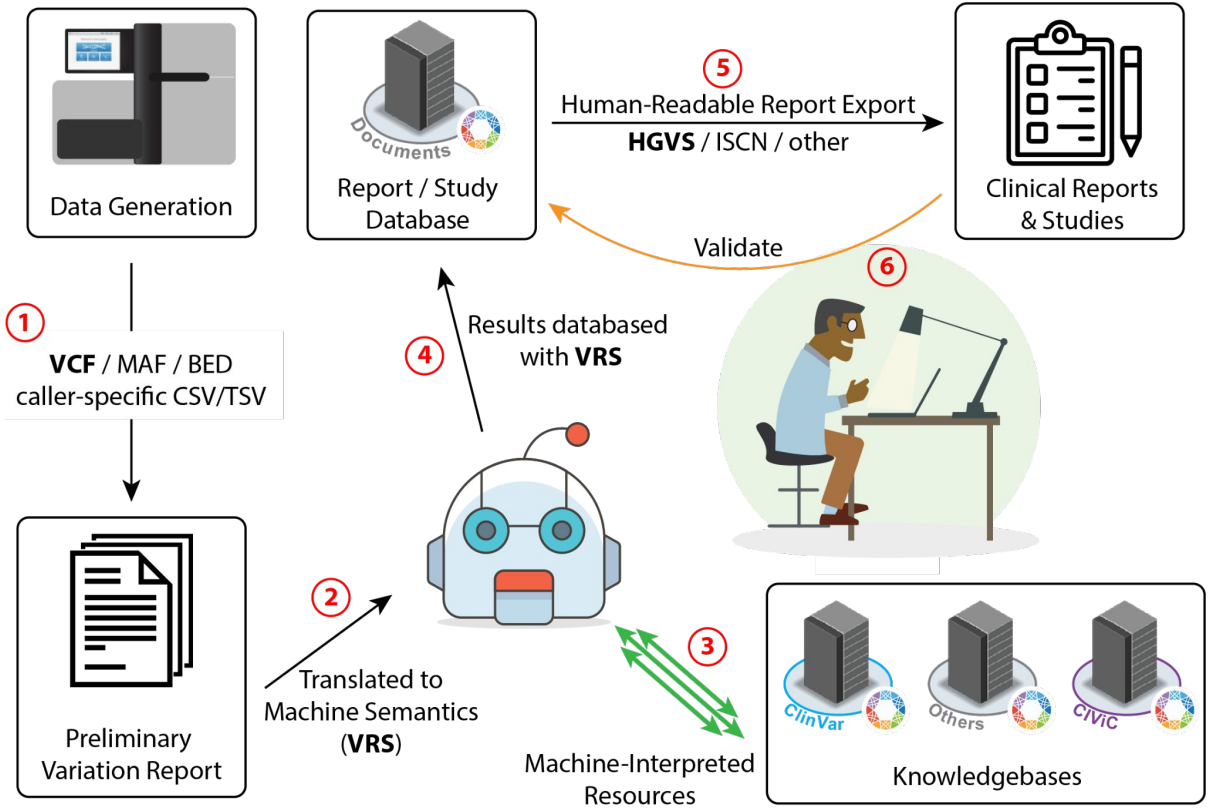
Global Alliance for Genomics & Health



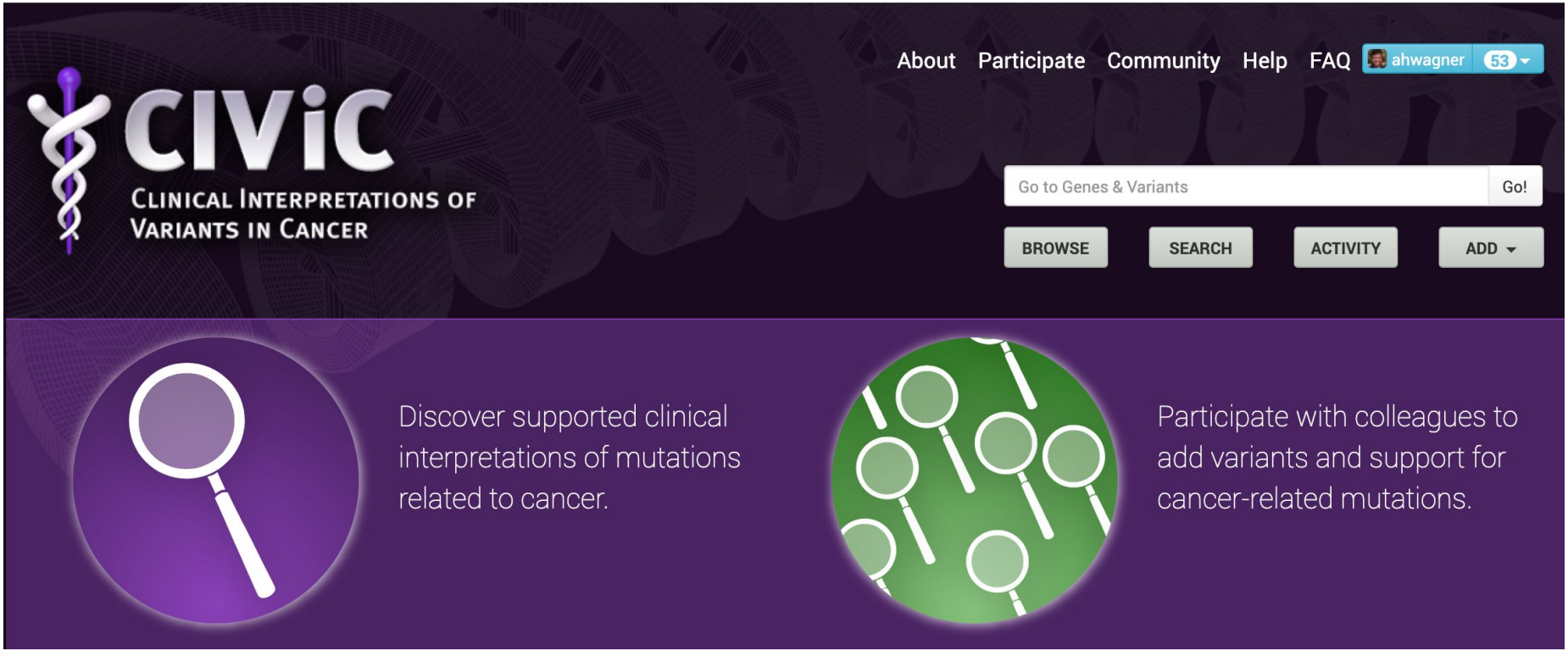
Questions about VRS?

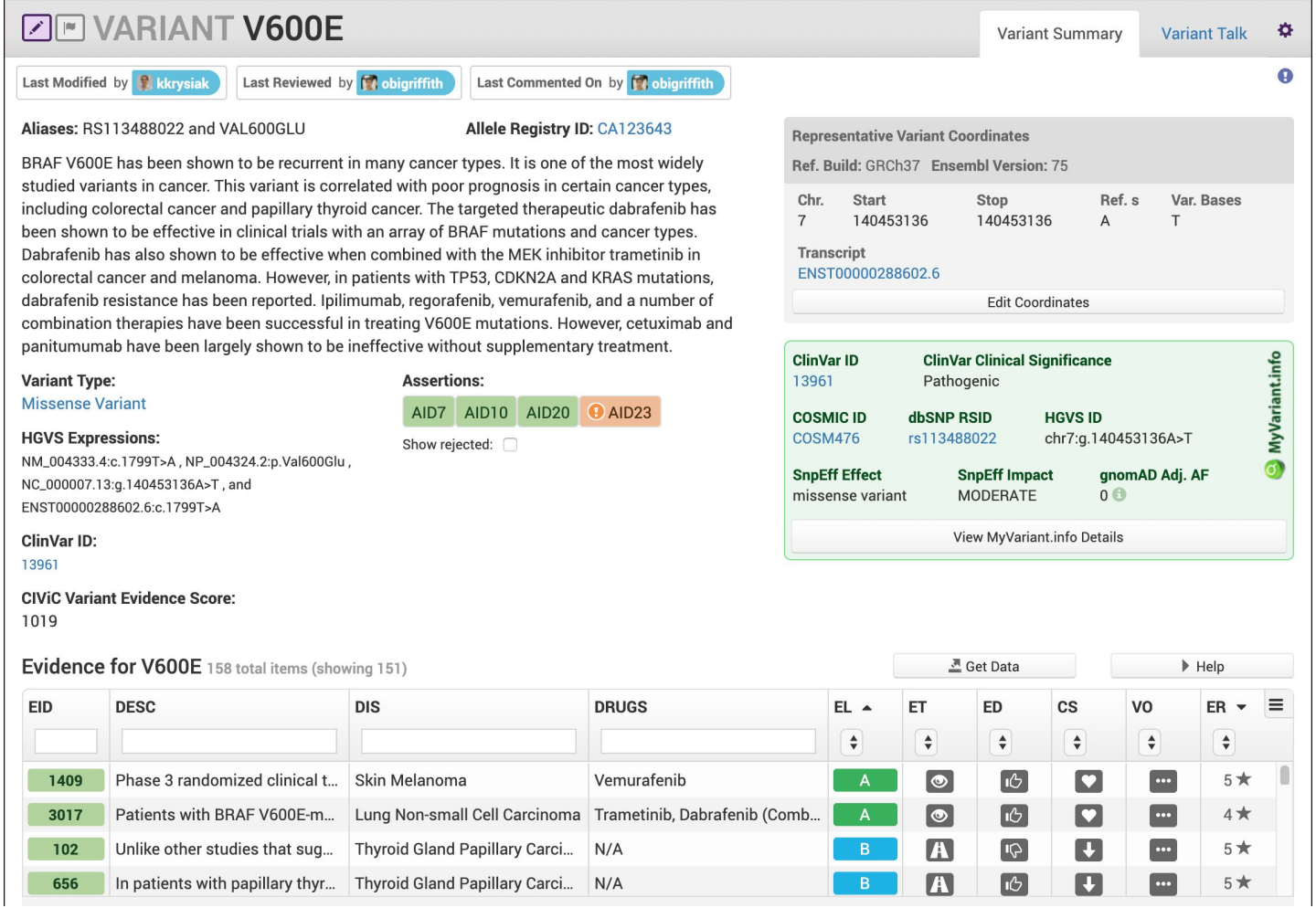
VRS Provides Mechanism for Scalability

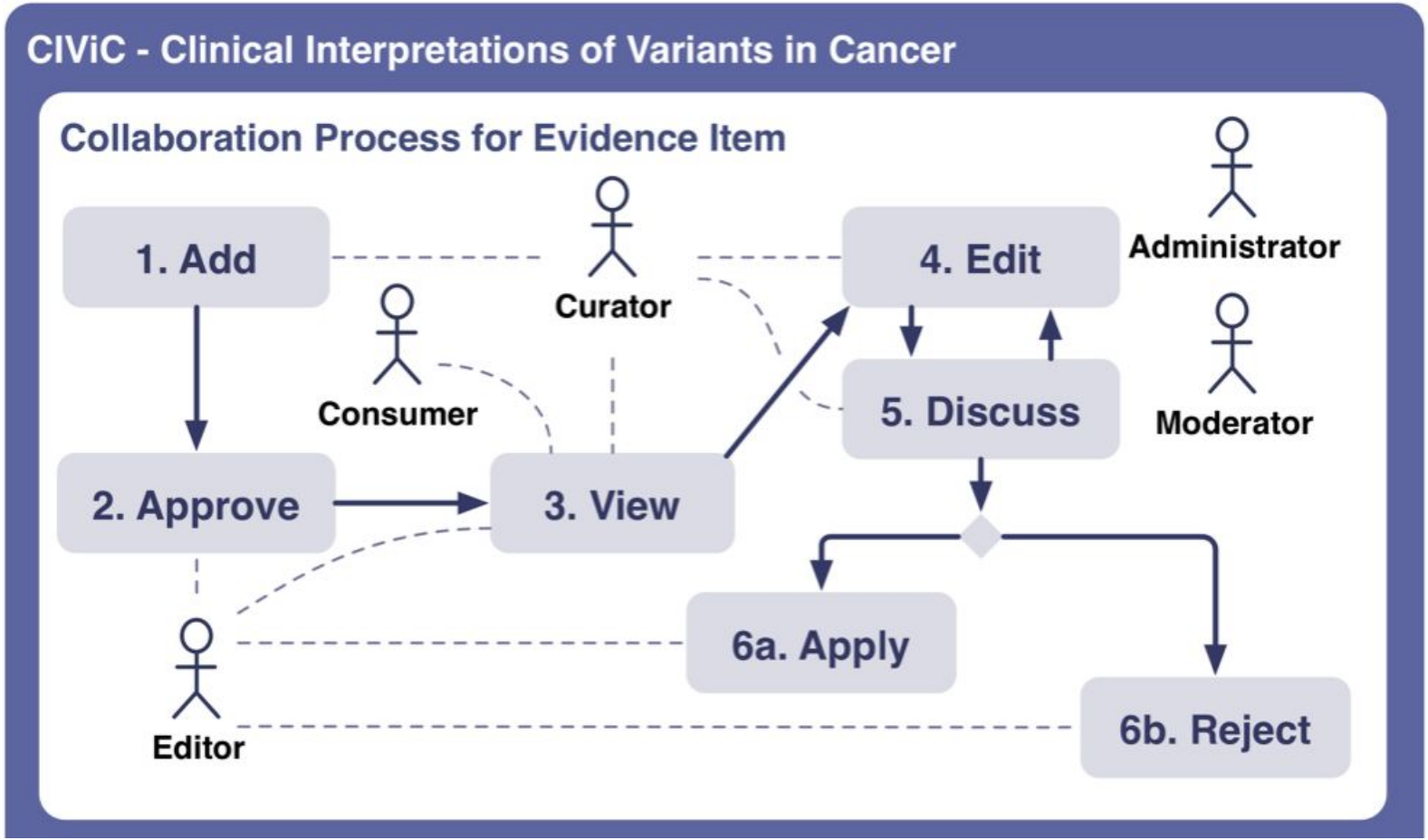
Global Alliance for Genomics & Health



Somatic Clinical Interpretation Resources


About
Participate
Community
Help
FAQ
ahwagner
53
CIViC
Clinical Interpretations of Variants in Cancer
Go to Genes & Variants
Go!
BROWSE
SEARCH
ACTIVITY
ADD
Discover supported clinical interpretations of mutations related to cancer.
Participate with colleagues to add variants and support for cancer-related mutations.


VARIANT V600E
Variant Summary
Variant Talk
Last Modified by kkrysiak
Last Reviewed by obigriffith
Last Commented On by obigriffith
Aliases: RS113488022 and VAL600GLU
Allele Registry ID: CA123643
BRAF V600E has been shown to be recurrent in many cancer types. It is one of the most widely studied variants in cancer. This variant is correlated with poor prognosis in certain cancer types, including colorectal cancer and papillary thyroid cancer. The targeted therapeutic dabrafenib has been shown to be effective in clinical trials with an array of BRAF mutations and cancer types. Dabrafenib has also shown to be effective when combined with the MEK inhibitor trametinib in colorectal cancer and melanoma. However, in patients with TP53, CDKN2A and KRAS mutations, dabrafenib resistance has been reported. Ipilimumab, regorafenib, vemurafenib, and a number of combination therapies have been successful in treating V600E mutations. However, cetuximab and panitumumab have been largely shown to be ineffective without supplementary treatment.
Variant Type:
Missense Variant
HGVS Expressions:
NM_004333.4:c.1799T>A , NP_004324.2:p.Val600Glu , NC_000007.13:g.140453136A>T , and ENST00000288602.6:c.1799T>A
ClinVar ID:
13961
CIViC Variant Evidence Score:
1019
Assertions:
AID7 AID10 AID20 AID23
Show rejected:
Representative Variant Coordinates
Ref. Build: GRCh37 Ensembl Version: 75
Chr. 7
Start 140453136
Stop 140453136
Ref. s A
Var. Bases T
Transcript
ENST00000288602.6
Edit Coordinates
ClinVar ID 13961
ClinVar Clinical Significance Pathogenic
COSMIC ID COSM476
dbSNP RSID rs113488022
HGVS ID chr7:g.140453136A>T
SnpEff Effect missense variant
SnpEff Impact MODERATE
gnomAD Adj. AF 0
MyVariant.info
View MyVariant.info Details
Evidence for V600E 158 total items (showing 151)
Get Data
Help
EID DESC DIS DRUGS EL ET ED CS VO ER
1409 Phase 3 randomized clinical t... Skin Melanoma Vemurafenib A 5★
3017 Patients with BRAF V600E-m... Lung Non-small Cell Carcinoma Trametinib, Dabrafenib (Comb... A 4★
102 Unlike other studies that sug... Thyroid Gland Papillary Carci... N/A B 5★
656 In patients with papillary thyr... Thyroid Gland Papillary Carci... N/A B 5★


CIViC - Clinical Interpretations of Variants in Cancer
Collaboration Process for Evidence Item
1. Add
2. Approve
3. View
4. Edit
5. Discuss
6a. Apply
6b. Reject
Curator
Consumer
Editor
Administrator
Moderator

OncoKB

Precision Oncology Knowledge Base

595	4472	38	79
Genes	Alterations	Tumor Types	Drugs

Search Gene / Alteration

Level 1	Level 2	Level 3	Level 4	Level R1	Level R2
FDA-approved	Standard care	Clinical evidence	Biological evidence	Standard care	Clinical evidence
20 Genes	10 Genes	25 Genes	14 Genes	4 Genes	6 Genes

@handlerwagner

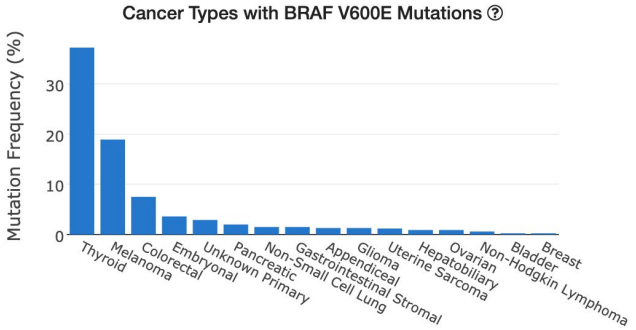
https://oncokb.org/

BRAF V600E

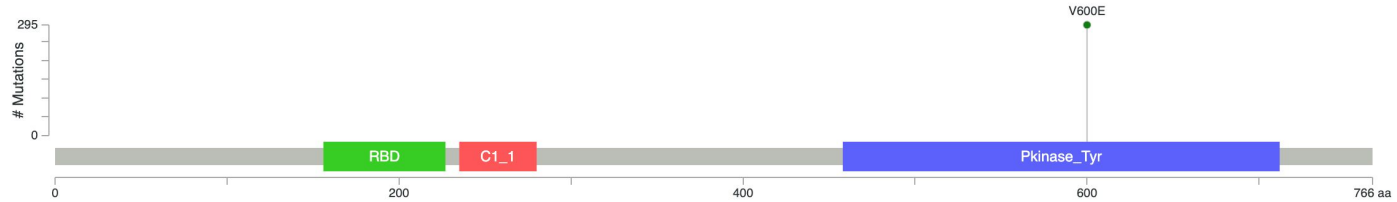
Oncogenic · Gain-of-function , Level 1

BRAF, an intracellular kinase, is frequently mutated in melanoma, thyroid and lung cancers among others. The BRAF V600E mutation is known to be oncogenic.

See additional BRAF information



Annotated Mutation Distribution in MSK-IMPACT Clinical Sequencing Cohort (Zehir et al., Nature Medicine, 2017)



Search:

▲ Alteration	Cancer Type	Drug(s)	▼ Level	Citations
V600E	Anaplastic Thyroid Cancer	Dabrafenib + Trametinib	1	1 reference
V600E	Non-Small Cell Lung Cancer	Dabrafenib + Trametinib	1	2 references

BRAF Oncogenic Mutations

Search:

▲ Alteration	▼ Oncogenic	Mutation Effect	Citations
V600R	Yes	Gain-of-function	12 references
F247L	Likely	Likely Gain-of-function	2 references
T599dup	Yes	Gain-of-function	4 references
R462E	Likely	Likely Gain-of-function	1 reference
K601E	Likely	Gain-of-function	6 references
L597Q	Yes	Gain-of-function	9 references
V459L	Yes	Gain-of-function	2 references
G596C	Likely	Gain-of-function	1 reference
E275K	Likely	Likely Gain-of-function	1 reference
G466V	Yes	Gain-of-function	9 references
A728V	Likely	Gain-of-function	1 reference
PAPSS1-BRAF Fusion	Likely	Gain-of-function	2 references
SND1-BRAF Fusion	Yes	Gain-of-function	4 references
L514V	Likely	Likely Gain-of-function	1 reference

https://oncokb.org/gene/BRAF/Oncogenic%20Mutations

Cancer Biomarkers database

Last update: 2018/01/17

The Cancer Biomarkers database is curated and maintained by several clinical and scientific experts in the field of precision oncology supported by the European Union's Horizon 2020 funded project. This database is currently being integrated with knowledge databases of other institutions in a collaborative effort of the Global Alliance for Genomics and Health. The contribution of the community is encouraged and proposals of edition or comments about the information contained in this database can be given by contacting us here or by using the feedback icon located at the left of each entry of the table. The database follows the data model originally described by Dienstmann et al. This table provides a summary of the content of the database that can be interactively browsed. Additional information, including the genomic coordinates of the variants, can be accessed via the download feature. This database is licensed under a Creative Commons Public Domain Dedication (CC0 1.0 Universal). When referring to this database, please cite: Cancer Genome Interpreter Annotates The Biological And Clinical Relevance Of Tumor Alterations; doi: https://doi.org/10.1101/140475.

Download | Feedback

Biomarker	Drug	Effect	Evidence	Source	Curator	Tumor type
Search here...						
ABL1 (E255K,E255V,Y25...	Nilotinib (BCR-ABL inhibitor 2nd gen)	Resistant	European Leuke...	PMID:21562040	CRubio-Perez	CML
ABL1 (F359V,F359C,F359...	Dasatinib (BCR-ABL inhibitor 2nd gen)	Responsive	NCCN guidelines	PMID:21562040	RDientsmann	CML
ABL1 (I242T,M244V,K24...	Imatinib (BCR-ABL inhibitor 1st gen&K...	Resistant	European Leuke...	PMID:21562040	CRubio-Perez	CML
ABL1 (T315A,F317L,F317...	Bosutinib (BCR-ABL inhibitor 3rd gen)	Responsive	NCCN guidelines	PMID:21562040	RDientsmann	CML
ABL1 (T315A,F317L,F317...	Nilotinib (BCR-ABL inhibitor 2nd gen)	Responsive	NCCN guidelines	PMID:21562040	RDientsmann	CML

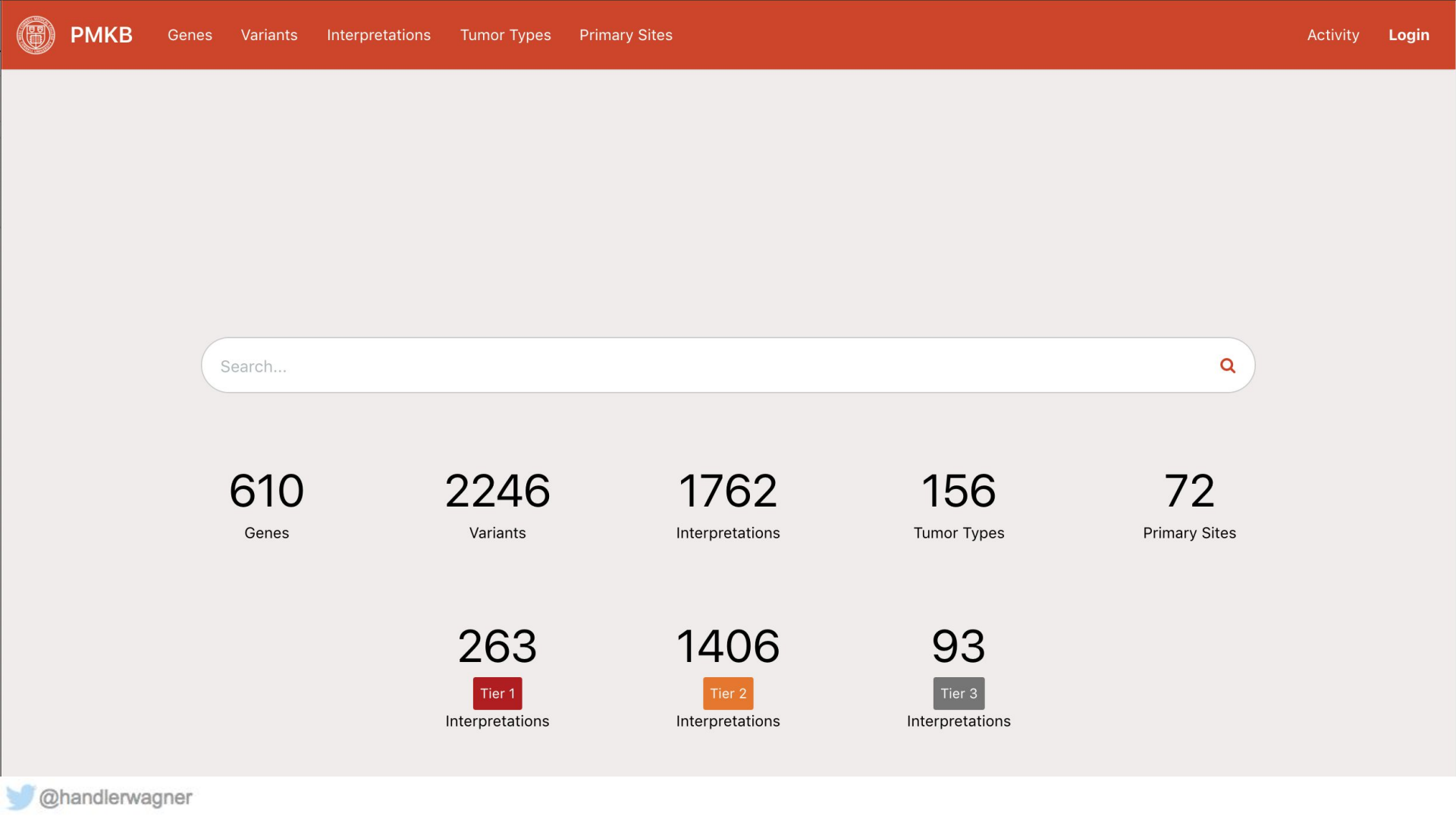
Catalog of Validated Oncogenic Mutations

Last update: 2018/01/17

Compiled inventory of mutations in cancer genes that are demonstrated to drive tumor growth or predispose to cancer. This was retrieved by combining the data contained in the DoCM (PMID:27684579), ClinVar (PMID:26582918) and OncoKB (PMID:28890946) databases as well as the results of several published experimental assays and additional manual curation efforts. We also considered as oncogenic the mutations reported to increase sensitivity to targeted drugs included in the Cancer Biomarkers Database of the Cancer Genome Interpreter. Germline variants found to predispose to cancer, which we retrieved from the ClinVar (PMID:26582918) and IARC (PMID:17311302) resources, were also included. The aggregation of the data includes (among others) the harmonization of the syntax of variants and the cancer type taxonomy (referred as "cancer" when the specific tumor type of the observation is not available) across the different data sources to guarantee the interoperability of all the resources that form the Cancer Genome Interpreter. Contradictory data (i.e. a variant stated as oncogenic and neutral by different resources) was flagged and filtered out. The content of each of these resources is licensed under the following terms: DoCM license, ClinVar license, OncoKB license, IARC license and Cancer Biomarkers database license.

Download

Gene	GDNA	Protein change	Transcript	Context	Tumor type
Search he					
ABCB4	chr7:g.87053221C>T	.	ENST00000265723	germline	Hepatic carcinoma predisposition
ABL1	chr9:g.133738306G>A	p.E236K	ENST00000318560	somatic	Chronic myeloid leukemia
ABL1	chr9:g.133738309A>G	p.M237V	ENST00000318560	somatic	Chronic myeloid leukemia


PMKB
Genes
Variants
Interpretations
Tumor Types
Primary Sites
Activity
Login
Search...
610
Genes
2246
Variants
1762
Interpretations
156
Tumor Types
72
Primary Sites
263
Tier 1
Interpretations
1406
Tier 2
Interpretations
93
Tier 3
Interpretations
@handlerwagner

Search...

Interpretation 2351

Information View History Pending Review

EGFR

Variants
EGFR G796S

Tier 1

Primary Sites
Lung

Tumor Types
Non-Small Cell Lung Carcinoma Adenocarcinoma

Interpretation
Somatic mutations in the tyrosine kinase domain of the epidermal growth factor receptor (EGFR) gene are present in approximately 80% of the lung adenocarcinomas that respond to first and second generation EGFR tyrosine kinase inhibitors (TKIs, eg, gefitinib, erlotinib and afatinib). Two types of mutations account for approximately 80-90% of all EGFR mutations: short in-frame deletions in Exon 19 and a point mutation in exon 21 at codon 858 (L858R). Other less common mutations in exons 18, 20, and 21 are found in 10-20% of EGFR-mutated cases. Exon 20 mutations are more commonly associated with resistance to tyrosine kinase inhibitors (TKIs), but may respond to third generation TKI (eg, osimertinib). This EGFR variant (G796S) lies within the tyrosine kinase domain and has been reported in rare cases of lung adenocarcinomas, squamous cell carcinoma of head and neck and prostate adenocarcinoma. In silico studies suggest G796S mutation may confer resistance to TKIs. However, additional studies are needed to further elucidate the oncogenicity of the mutation and therapeutic implications of this rare variant.

Citations

1. Sequist LV, et al. First-line gefitinib in patients with advanced non-small-cell lung cancer harboring somatic EGFR mutations. J Clin Oncol 2008;26(15):2442-9
2. Pao W, et al. EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib. Proc Natl Acad Sci U S A 2004;101(36):13306-11
3. Pirker R Third-generation epidermal growth factor receptor tyrosine kinase inhibitors in advanced nonsmall cell lung cancer. Curr Opin Oncol 2016;28(2):115-21
4. Ma C, et al. T790M and acquired resistance of EGFR TKI: a literature review of clinical reports. J Thorac Dis 2011;3(1):10-8
5. Schwentner I, et al. Identification of the rare EGFR mutation p.G796S as somatic and germline mutation in white patients with squamous cell carcinoma of the head and neck. Head & neck 2008;30(8):1040-4
6. Ou SI, et al. Emergence of novel and dominant acquired EGFR solvent-front mutations at Gly796 (G796S/R) together with C797S/R and L792F/H mutations in one EGFR (L858R/T790M) NSCLC patient who progressed on osimertinib. Lung cancer (Amsterdam, Netherlands) 2017;108:228-231
7. Douglas DA, et al. Novel mutations of epidermal growth factor receptor in localized prostate cancer. Frontiers in bioscience: a journal and virtual library 2006;11:2518-25
8. Goldberg ME, et al. Multiple configurations of EGFR exon 20 resistance mutations after first- and third-generation EGFR TKI treatment affect treatment options in NSCLC. PLoS One. 2018 Nov 27;13(11):e0208097.

Last updated: 2019-07-15 15:39:26 UTC

The Clinical Knowledgebase (CKB)

Powered by The Jackson Laboratory

CKB is a dynamic digital resource for interpreting complex cancer genomic profiles in the context of protein impact, therapies, and clinical trials. CKB CORE is the public access version we have been providing to the community since 2016. CKB CORE contains all the content associated with 85 genes that are commonly found on cancer hotspot panels. New and updated content is pushed out daily and viewable genes are available on a quarterly rotating schedule.

Not finding the content you need? Need more advanced searching?

Check out the CKB **BOOST** subscription version for content extending to 1,000+ genes.

Basic Search

- Explore by Gene
- Explore by Variant
- Explore by DrugClass - Available in CKB BOOST
- Explore by Drug - Available in CKB BOOST
- Explore by Indication/Tumor Type - Available in CKB BOOST

News

Aug 6, 2019 - Meet the CKB Team and tour a live demo in Nashville!

Jul 1, 2019 - CKB BOOST now has AMP/CAP/ASCO evidence level coding!

Jun 28, 2019 - CKB CORE brings back EGFR, PIK3CA, removes BRCA1, BRCA2, KRAS, and offers new content

Molecular Profile Detail

Profile Name	BRAF V600E
Gene Variant Detail	BRAF V600E (gain of function)
Relevant Treatment Approaches	BRAF Inhibitor; MEK inhibitor (Pan); MEK1 Inhibitor; MEK2 Inhibitor; RAF Inhibitor (Pan)

Variant Level Evidence 232 | Complex Molecular Profile Evidence 200 | Gene Level Evidence 835 | Treatment Approach Evidence 125 | Variant Associated Clinical Trials 49 | Gene Associated Clinical Trials 215

Filtering and Sorting

Filter rows:

Showing 1 to 232 of 232 entries

Molecular Profile	Indication/Tumor Type	Response Type	Relevant Treatment Approaches	Therapy Name	Approval Status	Evidence Type	Efficacy Evidence	References
BRAF V600E	renal cell carcinoma	predicted - sensitive	RAF Inhibitor (Pan)	Vemurafenib	Case Reports/Case Series	Actionable	In a clinical case study, a patient with metastatic renal cell carcinoma harboring BRAF V600E demonstrated a partial response following treatment with Zelboraf (vemurafenib) (PMID: 26918217).	26918217
BRAF V600E	colon neuroendocrine neoplasm	predicted - sensitive	BRAF Inhibitor	Dabrafenib	Case Reports/Case Series	Actionable	In a clinical case study, Tafinlar (dabrafinib) treatment of a patient with recurrent neuroendocrine carcinoma of the colon harboring a BRAF V600E mutation resulted in stable disease for 6 months before disease progression (PMID: 30181415).	30181415

Molecular Profile Detail

Profile Name	BRAF V600E
Gene Variant Detail	BRAF V600E (gain of function)
Relevant Treatment Approaches	BRAF Inhibitor; MEK inhibitor (Pan); MEK1 Inhibitor; MEK2 Inhibitor; RAF Inhibitor (Pan)

Variant Level Evidence 232 | Complex Molecular Profile Evidence 200 | Gene Level Evidence 835 | Treatment Approach Evidence 125 | Variant Associated Clinical Trials 49 | Gene Associated Clinical Trials 215

Filtering and Sorting

Filter rows:

Showing 1 to 200 of 200 entries

Molecular Profile	Indication/Tumor Type	Response Type	Relevant Treatment Approaches	Therapy Name	Approval Status	Evidence Type	Efficacy Evidence	References
BRAF amp BRAF V600E	colorectal cancer	resistant	RAF Inhibitor (Pan)	Cetuximab + Vemurafenib	Case Reports/Case Series	Actionable	In a clinical case study, a patient with BRAF V600E colorectal cancer developed progressive disease after a partial response lasting 16 weeks to Erbitux (cetuximab) and Zelboraf (vemurafenib) combination treatment, amplification of BRAF V600E was identified as an acquired alteration at the time of progression (PMID: 28951457).	28951457
BRAF amp BRAF V600E	colorectal cancer	predicted - resistant	RAF Inhibitor (Pan)	Panitumumab + Vemurafenib	Case Reports/Case Series	Actionable	In a clinical case study, a patient with BRAF V600E colorectal cancer developed progressive disease after a partial response lasting 24 weeks to Vectibix (panitumumab) and Zelboraf (vemurafenib) combination treatment, amplification of BRAF V600E was identified	28951457

https://ckb.jax.org/molecularProfile/showMolecularProfileByVariant?geneVariantId=49

Molecular Profile Detail

Profile Name	BRAF V600E
Gene Variant Detail	BRAF V600E (gain of function)
Relevant Treatment Approaches	BRAF Inhibitor, MEK inhibitor (Pan), MEK1 Inhibitor, MEK2 Inhibitor, RAF Inhibitor (Pan)

Variant Level Evidence 232 | Complex Molecular Profile Evidence 200 | Gene Level Evidence 835 | Treatment Approach Evidence 125 | Variant Associated Clinical Trials 49

Gene Associated Clinical Trials 215

Filtering and Sorting

Filter rows:

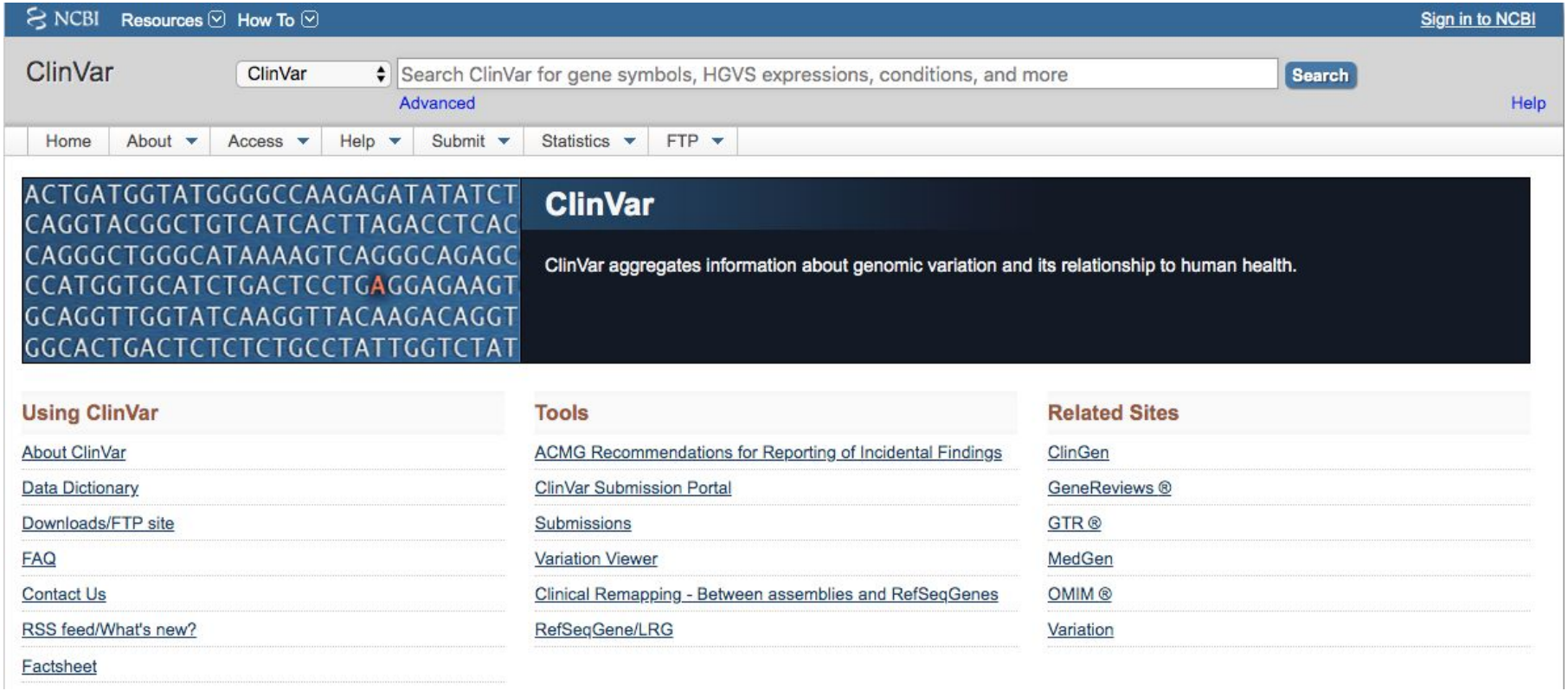
Showing 1 to 49 of 49 entries

Clinical Trial	Phase	Therapies	Title	Recruitment Status
NCT01336634	Phase II	Dabrafenib; Dabrafenib + Trametinib	Study of Selective BRAF Kinase Inhibitor Dabrafenib Monotherapy Twice Daily and in Combination With Dabrafenib Twice Daily and Trametinib Once Daily in Combination Therapy in Subjects With BRAF V600E Mutation Positive Metastatic (Stage IV) Non-small Cell Lung Cancer.	Active, not recruiting
NCT01709292	Phase II	Vemurafenib	Vemurafenib Neoadjuvant Trial in Locally Advanced Thyroid Cancer	Active, not recruiting
NCT01711632	Phase II	Vemurafenib	BRAF Inhibitor, Vemurafenib, in Patients With Relapsed or Refractory Hairy Cell Leukemia	Active, not recruiting
NCT01740648	Phase I	Fluorouracil + Trametinib	Trametinib, Fluorouracil, and Radiation Therapy Before Surgery in Treating Patients With Stage II-III Rectal Cancer	Active, not recruiting

https://ckb.jax.org/molecularProfile/showMolecularProfileByVariant?geneVariantId=49

Germline and Specialized Interpretation Resources

ClinVar



https://www.ncbi.nlm.nih.gov/clinvar/

FDA-Recognized ClinGen Classifications

Search results

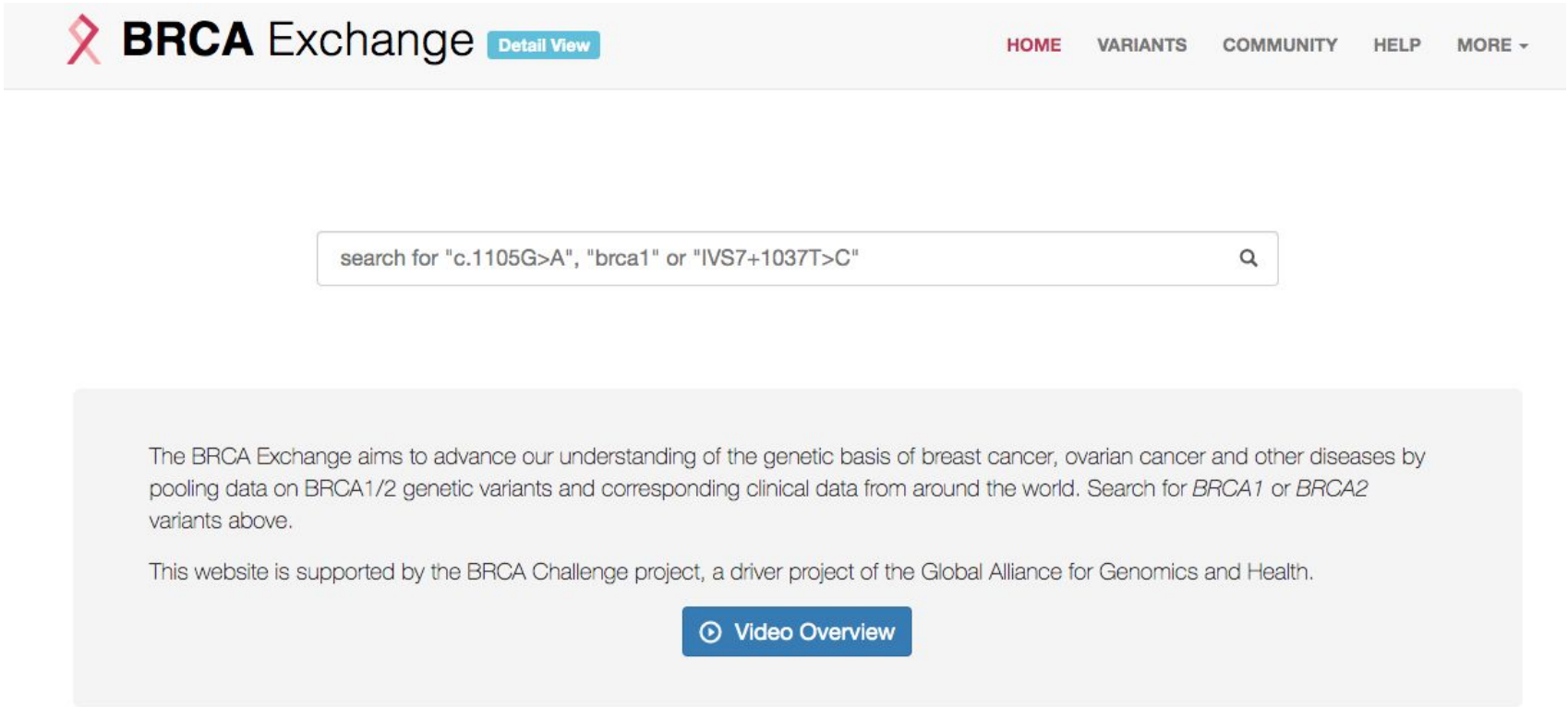
Items: 1 to 100 of 299

<< First < Prev Page 1 of 3 Next > Last >>

Filters activated: Pathogenic, Expert panel. Clear all to show 2796 items.

	Variation *Location*	Gene(s)	Protein change	Condition(s)	Clinical significance (Last reviewed)	Review status	Accession
1.	NM_004700.4(KCNQ4):c.853G>A (p.Gly285Ser) *GRCh37:* Chr1:41285565 *GRCh38:* Chr1:40819893	KCNQ4	G285S	DFNA 2 Nonsyndromic Hearing Loss, Nonsyndromic hearing loss and deafness	Pathogenic (Aug 20, 2015)	**reviewed by expert panel** FDA Recognized Database	VCV000006241
2.	NM_206933.3(USH2A):c.11241C>A (p.Tyr3747Ter) *GRCh37:* Chr1:215932085 *GRCh38:* Chr1:215758743	USH2A	Y3747*	Usher syndrome, Usher syndrome, type 2A	Pathogenic (Jan 30, 2018)	**reviewed by expert panel** FDA Recognized Database	VCV000506273
3.	NM_206933.3(USH2A):c.8682-9A>G *GRCh37:* Chr1:216040521 *GRCh38:* Chr1:215867179	USH2A		Usher syndrome, type 2A, Retinitis pigmentosa 39, not provided, Usher syndrome, type 2A, Usher syndrome	Pathogenic (May 7, 2015)	**reviewed by expert panel** FDA Recognized Database	VCV000197510
4.	NM_206933.3(USH2A):c.8559-2A>G *GRCh37:* Chr1:216051224 *GRCh38:* Chr1:215877882	USH2A		Usher syndrome, Retinitis pigmentosa 39, Usher syndrome, type 2A, not provided, Retinitis pigmentosa, Usher syndrome, type 2A	Pathogenic (Oct 9, 2018)	**reviewed by expert panel** FDA Recognized Database	VCV000048604

Gene Focus: BRCA1 and BRCA2



https://brcaexchange.org/

Variant Details

chr17:g.43094692:G>C

or

NM_007294.3(*BRCA1*):c.839C>G p.(Ala280Gly)

Hide Empty Items

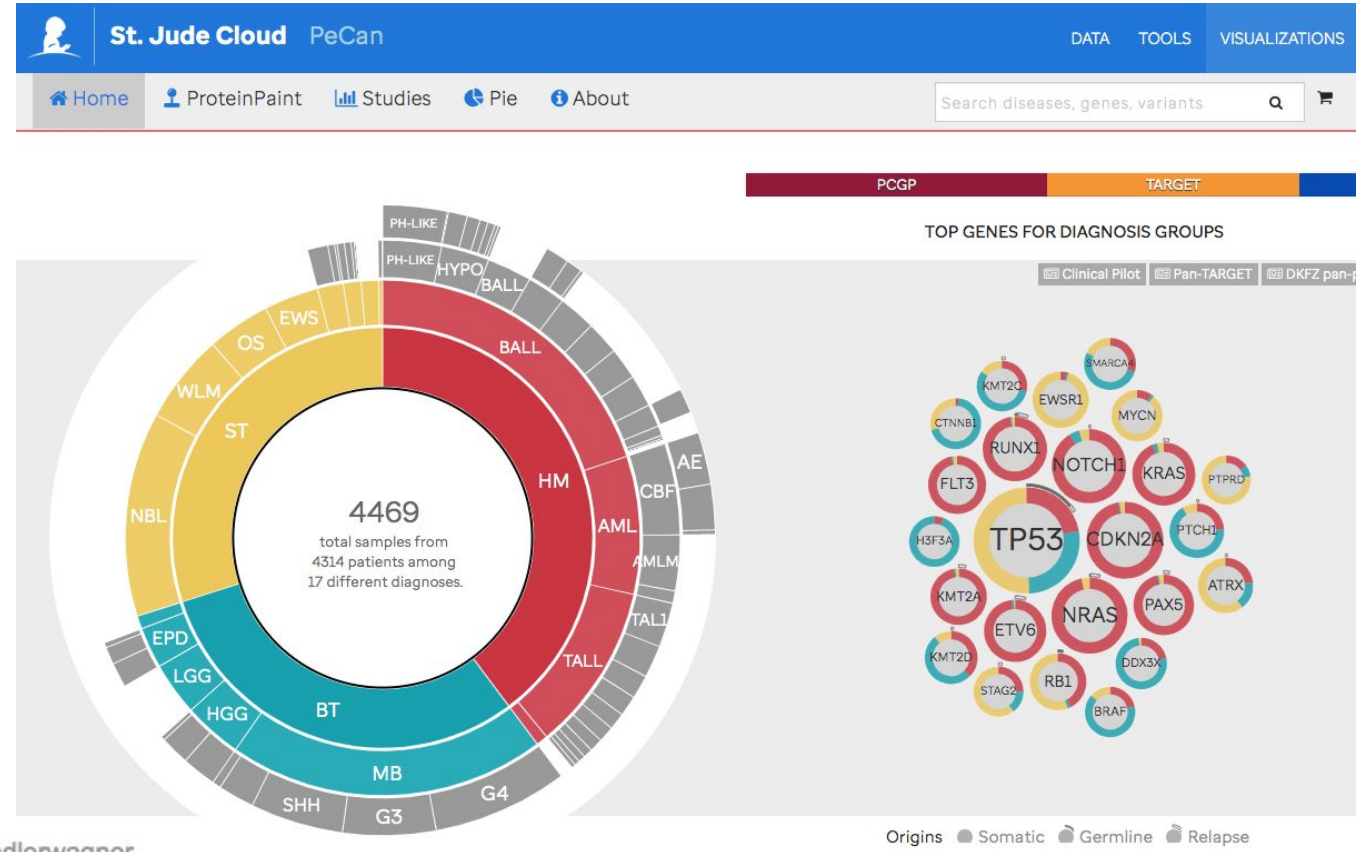
Variant Names

Gene	*BRCA1*
HGVS Nucleotide	c.839C>G
Transcript Identifier	NM 007294.3
HGVS RNA	-
HGVS Protein	p.(Ala280Gly)
Protein Identifier	NP_009225.1
Abbreviated AA Change	A280G
BIC Designation	958C>G
Genomic Nomenclature (GRCh38)	chr17:g.43094692:G>C
Genomic Nomenclature (GRCh37)	chr17:g.41246709:G>C

Clinical Significance (ENIGMA)

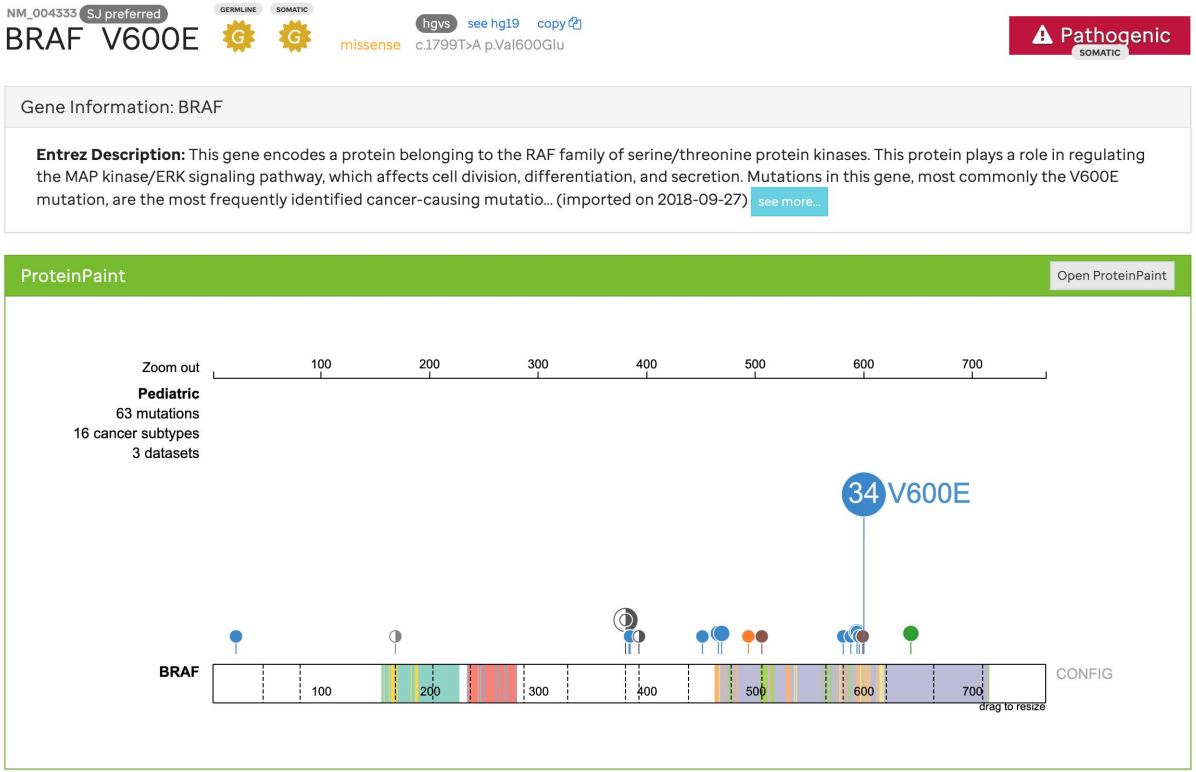
Clinical Significance	Benign / Little Clinical Significance
IARC Class	Benign
Comment on Clinical Significance	IARC class based on posterior probability from multifactorial likelihood analysis, thresholds for class as per Plon et al. 2008 (PMID: 18951446). Class 1 based on posterior probability = 0.0000767
Clinical Significance Citations	PMID: 21990134
Supporting Evidence URL(s)	link to multifactorial analysis
Date Last Evaluated	10 August 2015
Assertion Method	ENIGMA BRCA1/2 Classification Criteria (2015)
Assertion Method Citation	Enigma Rules version Mar 26, 2015
Allele Origin	Germline
ClinVar Accession	SCV000244413.1

Disease Focus: Pediatric Cancers



https://pecan.stjude.cloud

Disease Focus: Pediatric Cancers



https://pecan.stjude.cloud/variant/74

Licensing

BRCA Exchange

Cancer Biomarkers Database

ClinVar

Pecan PIE

CIViC

PMKB

OncoKB

CKB CORE

More Permissive

Less Permissive

PUBLIC DOMAIN

Attribution Only

Non-commercial / Research Only Custom Licensing

Non-commercial and Share-Alike

Paid-Access Only

General Questions