

Variant Representation and Interpretation

Sequencing Technologies and Bioinformatics Analysis 2022
Cold Spring Harbor Laboratories

You should be familiar with...

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

Wait a minute!

Our schedule didn't mention Variation ***Representation***.
Why do we care about that?

- 6:00pm - 7:00pm: DINNER
- 7:00pm - 8:00pm: “Variant interpretation” lecture (Alex Wagner)
- 8:00pm - 9:00pm: Variant interpretation lab (Alex Wagner)



Look! Right there!

Domain-Specific Applications of Genomic Data Standards

High-throughput computing

Clinical / biomedical reporting

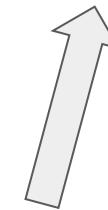
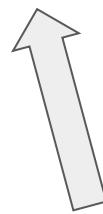
Computer-driven discovery

Domain-Specific Applications of Genomic Data Standards

High-throughput computing

Clinical / biomedical reporting

Computer-driven discovery



These different applications **play complementary roles** in variant interpretation!

Domain-Specific Applications of Genomic Data Standards

High-throughput computing

*How do we represent sequencing
instrument data for a sample?*

*How do we represent all of the
data across multiple samples?*

Fig. 1. (a) Example of valid VCF.

(a) VCF example

Header	##fileformat=VCFv4.1 ##fileDate=20110413 ##source=VCFtools ##reference=file:///refs/human_NCBI36.fasta ##contig=<ID=1,length=249250621,md5=1b22b98cdeb4a9304cb5d48026a85128,species="Homo Sapiens"> ##contig=<ID=X,length=155270560,md5=7e0e2e580297b7764e31dbc80c2540dd,species="Homo Sapiens"> ##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele"> ##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership"> ##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"> ##ALT=<ID=DEL,Description="Deletion"> ##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant"> ##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
Body	#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2
	1 1 . ACG A,AT 40 PASS . GT:DP 1/1:13 2/2:29
	1 2 . C T,CT . PASS H2:AA=T GT 0 1 2/2
	1 5 rs12 A G 67 PASS . GT:DP 1 0:16 2/2:20
	X 100 . T . PASS SVTYPE=DEL;END=299 GT:GQ:DP 1:12:. 0/0:20:36

(b) SNP

Alignment	VCF representation
1234	POS REF ALT
ACGT	2 C T
ATGT	ACGT
^	^

(c) Insertion

12345	POS	REF	ALT
AC-GT	2	C	CT
ACTGT	^		

(d) Deletion

1234	POS	REF	ALT
ACGT	1	ACG	A
A--T	^		

(e) Replacement

1234	POS	REF	ALT
ACGT	1	ACG	AT
A-TT	^		

(f) Large structural variant

Alignment	VCF representation
100 110 120 290 300	POS REF ALT INFO
ACGTACGTACGTACGTACGT[...]ACGTACGTACGTAC	100 T SVTYPE=DEL;END=299
ACGT-----[...]-----GTAC	

Variant Call Format - Header

(a) VCF example

```
##fileformat=VCFv4.1
##fileDate=20110413
##source=VCFtools
##reference=file:///refs/human_NCBI36.fasta
##contig=<ID=1,length=249250621,md5=1b22b98cdeb4a9304cb5d48026a85128,species="Homo Sapiens">
##contig=<ID=X,length=155270560,md5=7e0e2e580297b7764e31dbc80c2540dd,species="Homo Sapiens">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##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">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2
1 1 . ACG A,AT 40 PASS .
1 2 . C T,CT . PASS H2:AA=T GT 0|1 2/2
1 5 rs12 A G 67 PASS .
X 100 . T <DEL> . PASS SVTYPE=DEL;END=299 GT:GQ:DP 1:12:. 0/0:20:36
```

(b) SNP

Alignment	VCF representation
1234	POS REF ALT
ACGT	2 C T
ATGT	

(c) Insertion

12345	POS	REF	ALT
AC-GT	2	C	CT
ACTGT	^		

(d) Deletion

1234	POS	REF	ALT
ACGT	1	ACG	A
A--T	^		

(e) Replacement

1234	POS	REF	ALT
ACGT	1	ACG	AT
A-TT	^		

(f) Large structural variant

Alignment	VCF representation
100 110 120 290 300	POS REF ALT INFO
ACGTACGTACGTACGTACGT[...]ACGTACGTACGTAC	100 T SVTYPE=DEL;END=299
ACGT-----[...]-GTAC	



Variant Call Format - Columns

(a) VCF example

Header {
##fileformat=VCFv4.1
##fileDate=20110413
##source=VCFtools
##reference=file:///refs/human_NCBI36.fasta
##contig=<ID=1,length=249250621,md5=1b22b98cdeb4a9304cb5d48026a85128,species="Homo Sapiens">
##contig=<ID=X,length=155270560,md5=7e0e2e580297b7764e31dbc80c2540dd,species="Homo Sapiens">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##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">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2
1 1 . ACG A,AT 40 PASS . GT:DP 1/1:13 2/2:29
1 2 . C T,CT . PASS H2:AA=T GT 0|1 2/2
1 5 rs12 A G 67 PASS . GT:DP 1|0:16 2/2:20
X 100 . T . PASS SVTYPE=DEL;END=299 GT:GQ:DP 1:12:. 0/0:20:36

(b) SNP

Alignment VCF representation
1234 POS REF ALT
ACGT 2 C T
ATGT ^

(c) Insertion

12345 POS REF ALT
AC-GT 2 C CT
ACTGT ^

(d) Deletion

1234 POS REF ALT
ACGT 1 ACG A
A--T ^

(e) Replacement

1234 POS REF ALT
ACGT 1 ACG AT
A-TT ^

(f) Large structural variant

Alignment VCF representation
100 110 120 290 300
ACGTACGTACGTACGTACGT[...]ACGTACGTACGTAC
ACGT-----[...]-----GTAC

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 semicolons 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 semicolons 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 - Info Field: END

(a) VCF example

```
##fileformat=VCFv4.1
##fileDate=20110413
##source=VCFtools
##reference=file:///refs/human_NCBI36.fasta
##contig=<ID=1,length=249250621,md5=1b22b98cdeb4a9304cb5d48026a85128,species="Homo Sapiens">
##contig=<ID=X,length=155270560,md5=7e0e2e580297b7764e31dbc80c2540dd,species="Homo Sapiens">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##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">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2
1 1 . ACG A,AT 40 PASS .
1 2 . C T,CT . PASS H2;AA=T GT:DP 1/1:13 2/2:29
1 5 rs12 A G 67 PASS .
X 100 . T <DEL> . PASS SVTYPE=DEL;END=299 GT:GQ:DP 1:12:. 0/0:20:36
```

(b) SNP

Alignment	VCF representation
1234	POS REF ALT
ACGT	2 C T
ATGT	
^	

(c) Insertion

12345	POS	REF	ALT
AC-GT	2	C	CT
ACTGT	^		

(d) Deletion

1234	POS	REF	ALT
ACGT	1	ACG	A
A--T	^		

(e) Replacement

1234	POS	REF	ALT
ACGT	1	ACG	AT
A-TT	^		

(f) Large structural variant

Alignment	VCF representation
100 110 120 290 300	POS REF ALT INFO
ACGTACGTACGTACGTACGT[...]ACGTACGTACGTAC	100 T SVTYPE=DEL;END=299
ACGT-----[...]-----GTAC	

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 semicolons 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 Field: END

(a) VCF example

Header {
##fileformat=VCFv4.1
##fileDate=20110413
##source=VCFtools
##reference=file:///refs/human_NCBI36.fasta
##contig=<ID=1,length=249250621,md5=1b22b98cdeb4a9304cb5d48026a85128,species="Homo Sapiens">
##contig=<ID=X,length=155270560,md5=7e0e2e580297b7764e31dbc80c2540dd,species="Homo Sapiens">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele"
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership"
##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">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant"
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant"
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2
1 1 . ACG A,AT 40 PASS . GT:DP 1/1:13 2/2:29
1 2 . C T,CT . PASS H2:AA=T GT 0|1 2/2
1 5 rs12 A G 67 PASS . GT:DP 1|0:16 2/2:20
X 100 . T . PASS SVTYPE=DEL;END=299 GT:GQ:DP 1:12:. 0/0:20:36

(b) SNP

Alignment VCF representation
1234 POS REF ALT
ACGT 2 C T
ATGT ^

(c) Insertion

12345 POS REF ALT
AC-GT 2 C CT
ACTGT ^

(d) Deletion

1234 POS REF ALT
ACGT 1 ACG A
A--T ^

(e) Replacement

1234 POS REF ALT
ACGT 1 ACG AT
A-TT ^

(f) Large structural variant

Alignment
100 110 120 290 300
ACGTACGTACGTACGTACGT[...]ACGTACGTACGTAC
ACGT-----[...]-----GTAC

VCF representation
POS REF ALT INFO
100 T SVTYPE=DEL;END=299

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

Variant Call Format - Format Field: GT

(a) VCF example

Header								
	##fileformat=VCFv4.1							
	##fileDate=20110413							
	##source=VCFtools							
	##reference=file:///refs/human_NCBI36.fasta							
	##contig=<ID=1,length=249250621,md5=1b22b98cdeb4a9304cb5d48026a85128,species="Homo Sapiens">							
	##contig=<ID=X,length=155270560,md5=7e0e2e580297b7764e31dbc80c2540dd,species="Homo Sapiens">							
	##INFO=<ID=AA,Number=1>Type=String,Description="Ancestral Allele">							
	##INFO=<ID=H2,Number=0>Type=Flag,Description="HapMap2 membership">							
	##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">							
	##ALT=<ID=DEL>Description="Deletion">							
	##INFO=<ID=SVTYPE,Number=1>Type=String,Description="Type of structural variant">							
	##INFO=<ID=END,Number=1>Type=Integer,Description="End position of the variant">							
Body								
	CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
1	1	.		ACG	A,AT	40	PASS	.
1	2	.		C	T,CT	.	PASS	H2:AA=T
1	5	rs12		A	G	67	PASS	.
X	100	.	T		.	PASS	SVTYPE=DEL;END=299	
							FORMAT	SAMPLE1 SAMPLE2
							GT:DP	1/1:13 2/2:29
							GT	0 1 2/2
							GT:DP	1 0:16 2/2:20
							GT:GQ:DP	1 12:.. 0/0:20:36

(b) SNP

Alignment	VCF representation		
1234	POS	REF	ALT
ACGT	2	C	T
ATGT			

(c) Insertion

12345	POS	REF	ALT
AC-GT	2	C	CT

(d) Deletion

1234	POS	REF	ALT
ACGT	1	ACG	A

(e) Replacement

1234	POS	REF	ALT
ACGT	1	ACG	AT

(f) Large structural variant

Alignment	VCF representation			
100	POS	REF	ALT	INFO
ACGTACGTACGTACGTACGTACGT[...]		T		SVTYPE=DEL;END=299
ACGT-----[...]-----GTAC				



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 Field: GT

(a) VCF example

Header									
##fileformat=VCFv4.1									
##fileDate=20110413									
##source=VCFTools									
##reference=file:///refs/human_NCBI36.fasta									
##contig=<ID=1,length=249250621,md5=1b22b98cdeb4a9304cb5d48026a85128,species="Homo Sapiens">									
##contig=<ID=X,length=155270560,md5=7e0e2e580297b7764e31dbc80c2540dd,species="Homo Sapiens">									
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">									
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">									
##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">									
##ALT=<ID=DEL,Description="Deletion">									
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">									
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">									
Body									
CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1
1	1	.	ACG	A,AT	40	PASS	.	GT:DP	1/1:13
1	2	.	C	T,CT	.	PASS	H2:AA=T	GT	0 1:29
1	5	rs12	A	G	67	PASS	.	GT:DP	1 0:16
X	100	.	T		.	PASS	SVTYPE=DEL;END=299	GT:GQ:DP	2/2:20
									0/0:20:36

Ref = C
Alt1 = T
Alt2 = CT

Sample 1: C | T
Sample 2: CT / CT

(b) SNP

Alignment	VCF representation		
	POS	REF	ALT
1234	2	C	T
ACGT			
ATGT			
^			

(c) Insertion

12345	POS	REF	ALT
AC-GT	2	C	CT
ACTGT	^		

(d) Deletion

1234	POS	REF	ALT
ACGT	1	ACG	A
A--T	^		

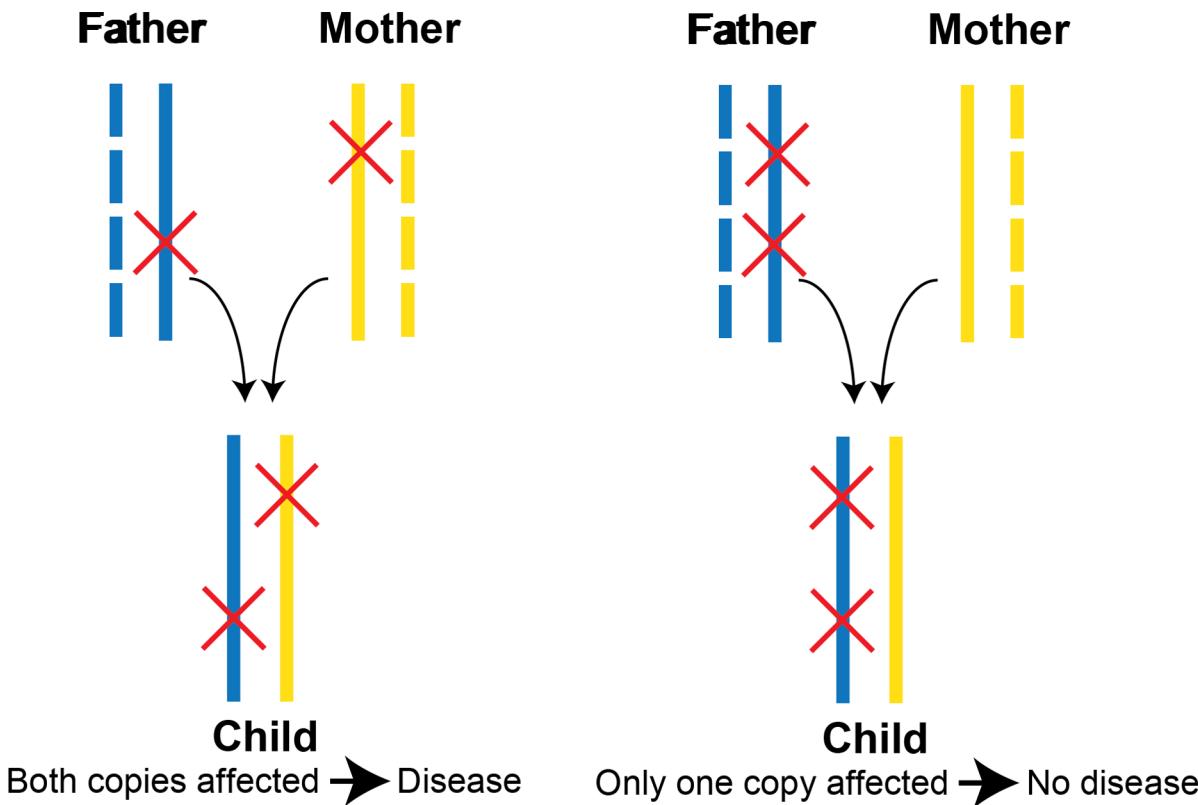
(e) Replacement

1234	POS	REF	ALT
ACGT	1	ACG	AT
A-TT	^		

(f) Large structural variant

Alignment	VCF representation			
	POS	REF	ALT	INFO
100	110	120	290	300
ACGTACGTACGTACGTACGTACGT[...]				
ACGT-----[...]				SVTYPE=DEL;END=299

Why is phasing important for variant interpretation?



Variant Call Format - Format Field: GT

(a) VCF example

Header	##fileformat=VCFv4.1 ##fileDate=20110413 ##source=VCFtools ##reference=file:///refs/human_NCBI36.fasta ##contig=<ID=1,length=249250621,md5=1b22b98cdeb4a9304cb5d48026a85128,species="Homo Sapiens"> ##contig=<ID=X,length=155270560,md5=7e0e2e580297b7764e31dbc80c2540dd,species="Homo Sapiens"> ##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele"> ##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership"> ##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"> ##ALT=<ID=DEL,Description="Deletion"> ##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant"> ##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
Body	#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2 1 1 . ACG A,AT 40 PASS . GT:DP 1/1:13 2/2:29 1 2 . C T,CT . PASS H2;AA=T GT 0 1 2/2 1 5 rs12 A G 67 PASS . GT:DP 1 0 16 2/2:20 X 100 . T . PASS SVTYPE=DEL;END=299 GT;GQ:DP 1:12:.. 0/0:20:30

Sample 1:
2 C phased with 5 G
2 T phased with 5 A

(b) SNP

<i>Alignment</i>	<i>VCF representation</i>		
1234	POS	REF	ALT
ACGT	2	C	T
ATGT			
^			

(c) Insertion

12345	POS	REF	AL
AC-GT	2	C	CT
ACTGT			
	^		

(d) Deletion

1234	POS	REF	ALT
ACGT	1	ACG	A
A-T			
^			

(e) Replacement

1234	POS	REF	ALT
ACGT	1	ACG	AT
A-TT			
^			

(f) Large structural variant

```

Alignment
 100      110      120      290      300
ACGTACGTACGTACGTACGTACGT[...]ACGTACGTACGTAC
ACGT-----[...]-----GTAC

```

VCF representation
POS REF ALT INFO
100 T SVTYPE=DEL;END=299



Domain-Specific Applications of Genomic Data Standards

High-throughput computing

*How do we represent sequencing
instrument data for a sample?*

*How do we represent all of the
data across multiple samples?*

- “Foundational standards”
- Focus on file formats
- Most widespread adoption
- Computational emphasis

Domain-Specific Applications of Genomic Data Standards

Clinical / biomedical reporting

How do we compactly describe genomic variation in a way that humans readily understand it?

How do we encode sufficient information to do this accurately?

Variant Nomenclature

Variant Nomenclature

...in the news!

Genomic Data Cracks Cold Case

WORLD EXCLUSIVE: Jack the Ripper unmasked: How amateur sleuth used DNA breakthrough to identify Britain's most notorious criminal 126 years after string of terrible murders

- DNA evidence on a shawl found at Ripper murder scene nails killer
- By testing descendants of victim and suspect, identifications were made
- Jack the Ripper has been identified as Polish-born Aaron Kosminski

[Daily Mail, September 6, 2014](#)

**WORLD EXCLUSIVE: Jack the Ripper
unmasked: How amateur sleuth used
DNA breakthrough to identify Britain's
most notorious criminal 126 years
after string of murders**

six weeks later...

- DNA evidence on a shawl found at Ripper murder scene nails killer
- By testing descendants of victim and suspect, identifications were made
- Jack the Ripper has been identified as Polish-born Aaron Kosminski

Daily Mail, September 9, 2014

Genomic Data Cracks Cold Case... or does it?

Jack the Ripper: Scientist who claims to have identified notorious killer has 'made serious DNA error'

'Error of nomenclature' undermines case against Polish immigrant barber accused of carrying out the atrocities in 1888

[The Independent, October 20, 2014](#)

Genomic Data Cracks Cold Case... or does it?

Jack the Ripper: Scientist who claims
to have identified serial killer has
**What happened?
'made serious DNA error'**

'Error of nomenclature' undermines case against Polish immigrant barber accused of carrying out the atrocities in 1888

The Independent, October 20, 2014

Methodology

1. DNA sample including killer's blood includes mitochondrial variation
2. An insertion in the sample not found in large mitochondrial variation database;
a private familial variant!
3. Descendant of suspect Kosminski's sister also has familial variation
4. Conclusion: must be Kosminski!

Methodology

1. DNA sample including killer's blood includes mitochondrial variation
2. An insertion in the sample not found in large mitochondrial variation database;
a private familial variant!
3. Descendant of suspect Kosminski's sister also has familial variation
4. Conclusion: must be Kosminski!

There are several shortfalls in the conclusions drawn here.

Methodology

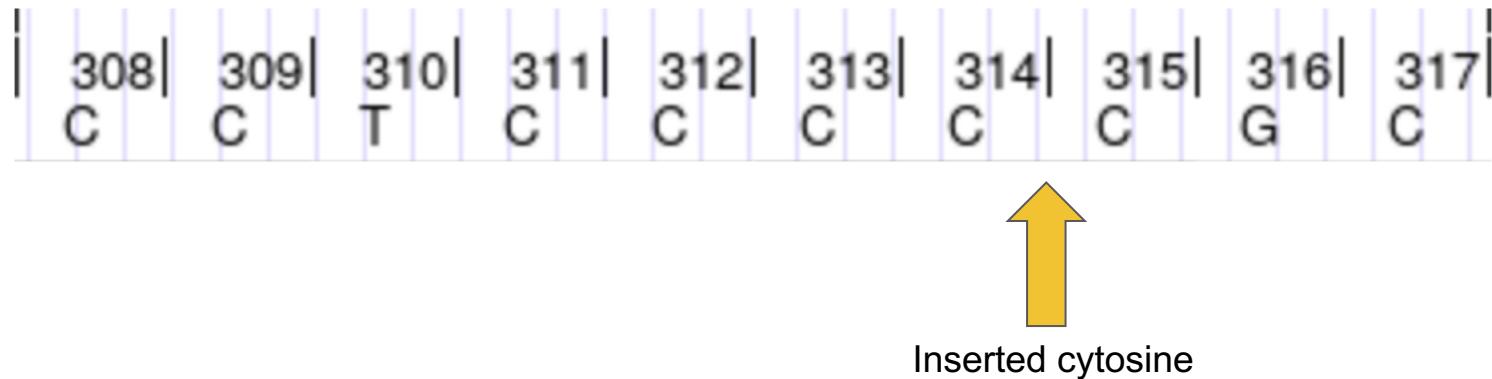
1. DNA sample including killer's blood includes mitochondrial variation
2. An insertion in the sample not found in large mitochondrial variation database;
a private familial variant!
3. Descendant of suspect Kosminski's sister also has familial variation
4. Conclusion: must be Kosminski!

*There are several shortfalls in the conclusions drawn here.
But the key piece of evidence was misidentified!*

An error of nomenclature!

That private, familial variation: **314.1C**

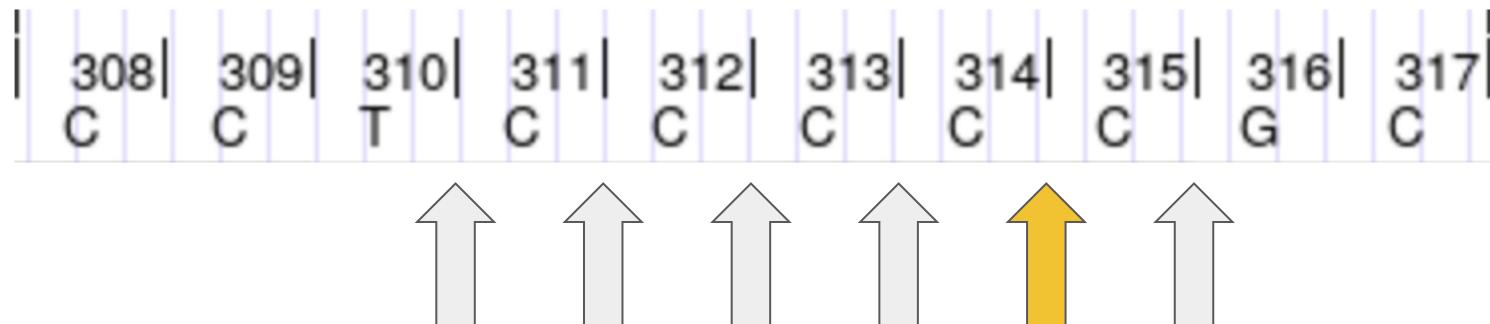
Reported frequency of 314.1C in forensics database: **absent**



An error of nomenclature!

That private, familial variation: **314.1C**

Reported frequency of 314.1C in forensics database: **absent**



Inserted cytosine... somewhere

ISFG Mitochondrial Variant Nomenclature (ca. 2000)

Insertions are described by first noting the site immediately 5' to the insertion followed by a decimal point and a '1' (for the first insertion), a '2' (if there is a second insertion), and so on, and then by the nucleotide that is inserted. In the case of homopolymeric tracts, where the exact position at which the insertion has occurred is unknown, the assumption is always made that the insertion has occurred at the highest numbered end of the homopolymeric region. For example, a homopolymeric region, at which insertions are common, occurs between nucleotide positions 311 and 315 (inclusive). The polymorphism, a C insertion, is assumed to occur after site 315, so the nomenclature used is 315.1C.

DNA Commission of the International Society for Forensic Genetics: guidelines for mitochondrial DNA typing. *Forensic Science International*. (2000)

ISFG Mitochondrial Variant Nomenclature (ca. 2000)

Insertions are described by first noting the site immediately 5' to the insertion followed by a decimal point and a '1' (for the first insertion), a '2' (if there is a second insertion), and so on, and then by the nucleotide that is inserted. In the case of homopolymeric tracts, where the exact position at which the insertion has occurred is unknown, the assumption is always made that the insertion has occurred at the highest numbered end of the homopolymeric region. For example, a homopolymeric region, at which insertions are common, occurs between nucleotide positions 311 and 315 (inclusive). The polymorphism, a C insertion, is assumed to occur after site 315, so the nomenclature used is 315.1C.

DNA Commission of the International Society for Forensic Genetics: guidelines for mitochondrial DNA typing. *Forensic Science International*. (2000)

ISFG Mitochondrial Variant Nomenclature (ca. 2000)

Insertions are described by first noting the site immediately 5' to the insertion followed by a decimal point and a '1' (for the first insertion), a '2' (if there is a second insertion), and so on, and then by the nucleotide that is inserted. In the case of homopolymeric tracts, where the exact position at which the insertion has occurred is unknown, the assumption is always made that the insertion has occurred at the highest numbered end of the homopolymeric region. For example, a homopolymeric region, at which insertions are common, occurs between nucleotide positions 311 and 315 (inclusive). The polymorphism, a C insertion, is assumed to occur after site 315, so the nomenclature used is 315.1C.

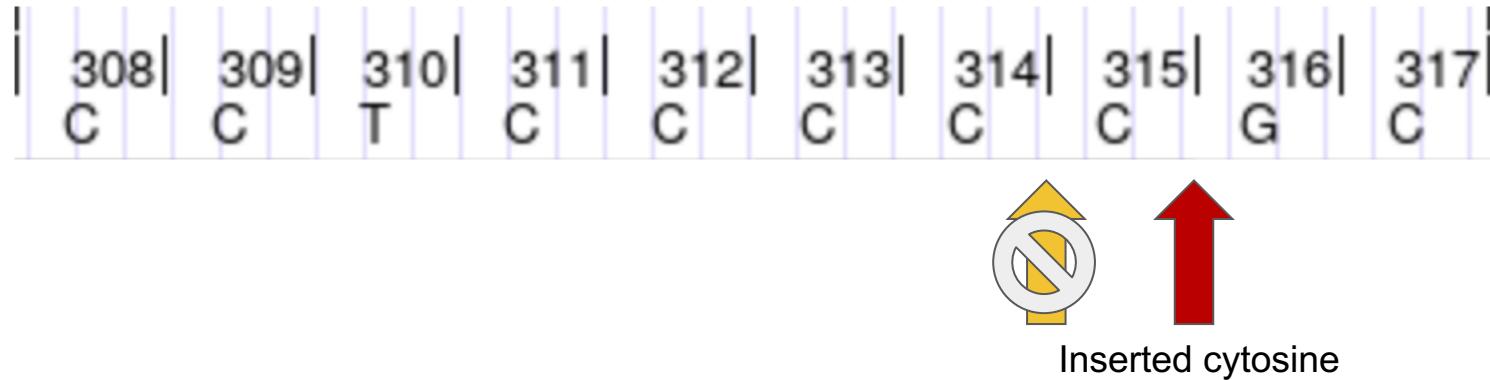
DNA Commission of the International Society for Forensic Genetics: guidelines for mitochondrial DNA typing. *Forensic Science International*. (2000)

The (Nomenclature) Culprit

That private, familial variation: **315.1C**

Reported frequency of 314.1C in forensics database: **absent**

Reported frequency of 315.1C: **>99%** among European Descent



The (Nomenclature) Culprit

That private, familial variation: **315.1C**

Reported frequency of 314.1C in forensics database: **absent**

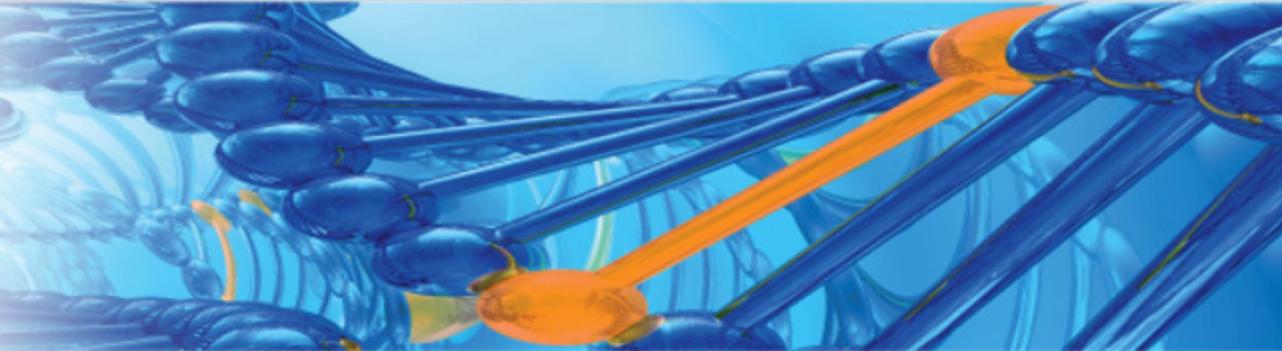
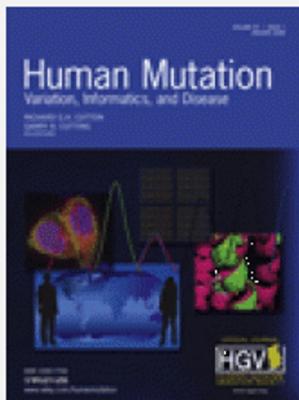
Reported frequency of 315.1C: **>99%** among European Descent



This type of error is an example consequence of the
Variant Overprecision problem (we will get back to this)



Inserted cytosine

[HOME](#)[ABOUT HGVS ▾](#)[GUIDELINES](#)[MEMBERSHIP ▾](#)[DATABASES ▾](#)[MEETINGS ▾](#)[CONTACT US](#)

ABOUT THE SOCIETY

The Society aims to foster discovery and characterization of genomic variations including population distribution and phenotypic associations. Promote collection, documentation and free distribution of genomic variation information and associated clinical variations. Endeavor to foster the development of the necessary methodology and informatics.

The Society is an Affiliate of the International Federation of Human Genetics Societies ([IFHGS](#)) and also the Human Genome Organisation ([HUGO](#)).

FORTHCOMING EVENTS

EXCEPTIONAL CASES

Berlin, Germany
(A Satellite of the [ESHG Meeting](#))

NEWS & EVENTS

hgvs.org landing page



@handlerwagne

go.osu.edu/wagner-review 43

HGVS Variant Nomenclature

Table 1. Nomenclature Definitions with Example Variant Descriptions

Substitution (>)	g.1318G>T	A change where one nucleotide is replaced by one other nucleotide
Deletion (del)	g.3661_3706del	A change where one or more nucleotides are not present (deleted)
Inversion (inv)	g.495_499inv	A change where more than one nucleotide replaces the original sequence and is the reverse-complement of the original sequence (e.g., CTCGA to TCGAG)
Duplication (dup)	g.3661_3706dup	A change where a copy of one or more nucleotides are inserted directly 3' of the original copy of that sequence
Insertion (ins)	g.7339_7340insTAGG	A change where one or more nucleotides are inserted in a sequence and where the insertion is not a copy of a sequence immediately 5'
Conversion (con)	g.333_590con1844_2101	A specific type of deletion-insertion where a range of nucleotides replacing the original sequence are a copy of a sequence from another site in the genome
Deletion-insertion (delins/indel)	g.112_117delinsTG	A change where one or more nucleotides are replaced by one or more other nucleotides and which is not a substitution, inversion, or conversion

Read “*a change where*” as “*a change where in a specific sequence compared to the reference sequence . . .*”

Sequence Variant Nomenclature

This site covers **HGVS nomenclature**, the recommendations for the description of sequence variants in DNA, RNA and protein sequences. It is used to report and exchange information of such variants and serves as an international standard. When using the recommendations please cite: *Den Dunnen et al. 2016, Hum.Mutat. 37:564-569.* HGVS-nomenclature is authorised by the Human Genome Organization (HUGO), under the responsibility of the HGVS Variant Nomenclature Committee (HVNC).

Current Recommendations

[General](#)[DNA](#)[RNA](#)[Protein](#)[Uncertain](#)[Checklist](#)[Open Issues](#)

<https://varnomen.hgvs.org/>

ClinVar

NCBI Resources How To Sign in to NCBI

ClinVar Search ClinVar for gene symbols, HGVS expressions, conditions, and more Search Advanced

Home About Access Help Submit Statistics FTP

ACTGATGGTATGGGCCAAGAGATATCT
CAGGTACGGCTGTCACTTAGACCTCAC
CAGGGCTGGGCATAAAAGTCAGGGCAGAGC
CCATGGTGATCTGACTCCTGAGGAGAACT
GCAGGTTGGTATCAAGGTTACAAGACAGGT
GGCACTGACTCTCTGCCTATTGGTCTAT

ClinVar

ClinVar aggregates information about genomic variation and its relationship to human health.

Using ClinVar

[About ClinVar](#)
[Data Dictionary](#)
[Downloads/FTP site](#)
[FAQ](#)
[Contact Us](#)
[RSS feed/What's new?](#)
[Factsheet](#)

Tools

[ACMG Recommendations for Reporting of Incidental Findings](#)
[ClinVar Submission Portal](#)
[Submissions](#)
[Variation Viewer](#)
[Clinical Remapping - Between assemblies and RefSeqGenes](#)
[RefSeqGene/LRG](#)

Related Sites

[ClinGen](#)
[GeneReviews ®](#)
[GTR ®](#)
[MedGen](#)
[OMIM ®](#)
[Variation](#)



@handlerwagner

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

ClinVar



NCBI Resources How To Sign in to NCBI

ClinVar Search ClinVar for gene symbols, HGVS expressions, conditions, and more Search Advanced

Home About Access Help Submit Statistics FTP

ACTGATGGTATGGGCCAAGAGATATCT
CAGGTACGGCTGTCACTTAGACCTCAC
CAGGGCTGGGCATAAAAGTCAGGGCAGAGC
CCATGGTGATCTGACTCCTGAGGAGAACT
GCAGGTTGGTATCAAGGTTACAAGACAGGT
GGCACTGACTCTCTGCCTATTGGTCTAT

ClinVar

ClinVar aggregates information about genomic variation and its relationship to human health.

Using ClinVar

- [About ClinVar](#)
- [Data Dictionary](#)
- [Downloads/FTP site](#)
- [FAQ](#)
- [Contact Us](#)
- [RSS feed/What's new?](#)
- [Factsheet](#)

Tools

- [ACMG Recommendations for Reporting of Incidental Findings](#)
- [ClinVar Submission Portal](#)
- [Submissions](#)
- [Variation Viewer](#)
- [Clinical Remapping - Between assemblies and RefSeqGenes](#)
- [RefSeqGene/LRG](#)

Related Sites

- [ClinGen](#)
- [GeneReviews ®](#)
- [GTR ®](#)
- [MedGen](#)
- [OMIM ®](#)
- [Variation](#)



@handlerwagner

<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) <i>GRCh37:</i> Chr1:41285565 <i>GRCh38:</i> 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) <i>GRCh37:</i> Chr1:215932085 <i>GRCh38:</i> 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 <i>GRCh37:</i> Chr1:216040521 <i>GRCh38:</i> 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 <i>GRCh37:</i> Chr1:216051224 <i>GRCh38:</i> 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



The following is a list of all names submitted to ClinVar for a single variant:

607008.0001
985A>G
985A>G (K304E)
985A>G (K329E)
A985G
ACADM, LYS304GLU
K304E
K304E (985 A->G)
K304E (K329E)

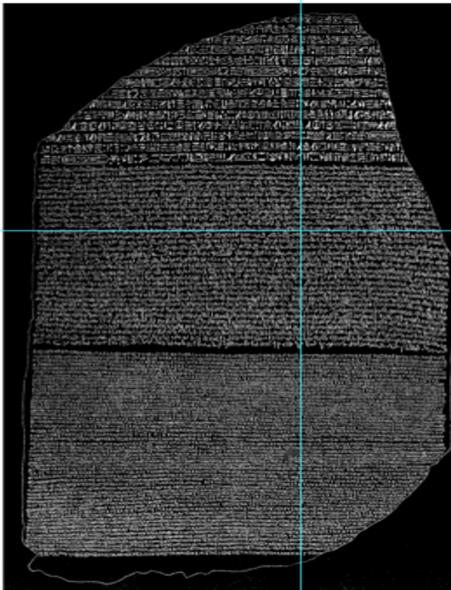
K304E only
K329E
K329E(985A>G)
LYS304GLU
Mutation c.985A>G (p.K304E)

c.985A>G
c.985A>G (p.K304E)
c.985A>G (p.Lys304Glu)
c985A>G

includes: K304E (985A>G)
p.K304E

p.Lys329Glu
previously known as p.Lys329Glu

Analysis of ACADM 985A>G mutation

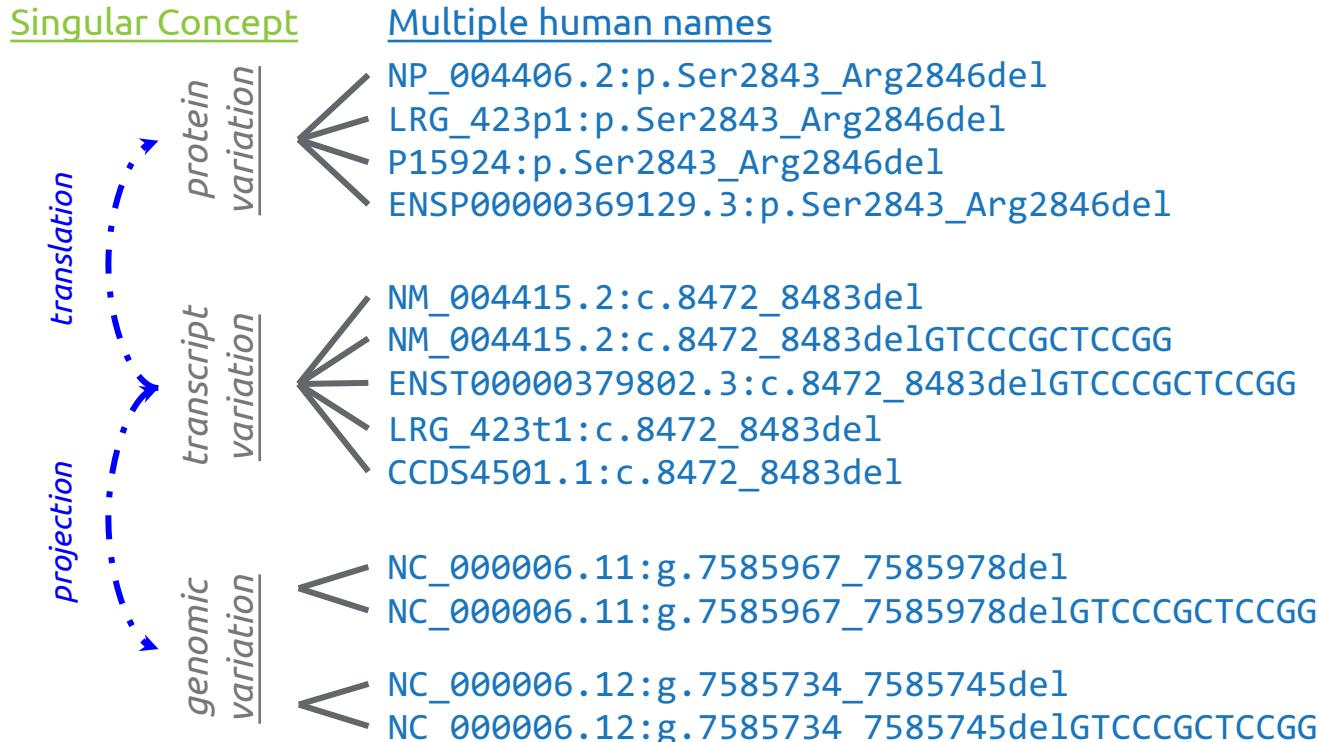


NC_000001.11:g.75761161A>G
NC_000001.10:g.76226846A>G
NG_007045.1:g.41804A>G
NM_000016.4:c.985A>G
NP_000007.1:p.Lys329Glu
NM_000016.5(ACADM):c.985A>G
(p.Lys329Glu)

Multiplicity in assemblies,
transcripts, legacy
conventions for
numbering systems,
abbreviations for amino
acids, formats

Source: Maglott, D. "The Variant Rosetta Stone [Powerpoint slides]" NCBI, 08 May 2015.

Variant De-duplication Problem



Single concept, multiple *non-overlapping* representations

ERBB2 (NP_004439.2) reference protein sequence



Non-standard HGVS: ERBB2 p.E770delinsEAYVM



Standard HGVS: NP_004439.2:p.Y772_A775dup



Wagner AH, et al. *Nat Genet*. 2020

Single concept, multiple *non-overlapping* representations

ERBB2 (NP_004439.2) reference protein sequence

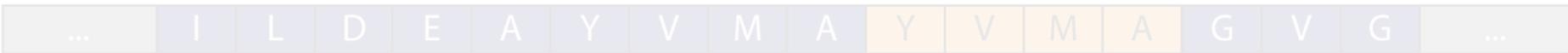


This is another instance of the **Variant Overprecision problem**, a key challenge for interoperability between standards.

Non-standard HGVS: ERBB2 p.E770delinsEAYVM



Standard HGVS: NP_004439.2:p.Y772_A775dup



Wagner AH, et al. *Nat Genet*. 2020

Fully-Justified Normalization Captures Region of Shuffling Ambiguity

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)

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

left shuffle ↲
(à la VCF)

↓ fully-justified
(à la SPDI & GA4GH VRS)

↘ right shuffle
(à la HGVS)

$$T \left[\frac{CA}{CAG} \right] CAGCAGCT$$



over-precise

$$T \left[\frac{CAGCAGC}{CAGCAGCAGC} \right] T$$



precise region of ambiguity

$$TCAGCAGC \left[\frac{AGC}{AGC} \right] T$$



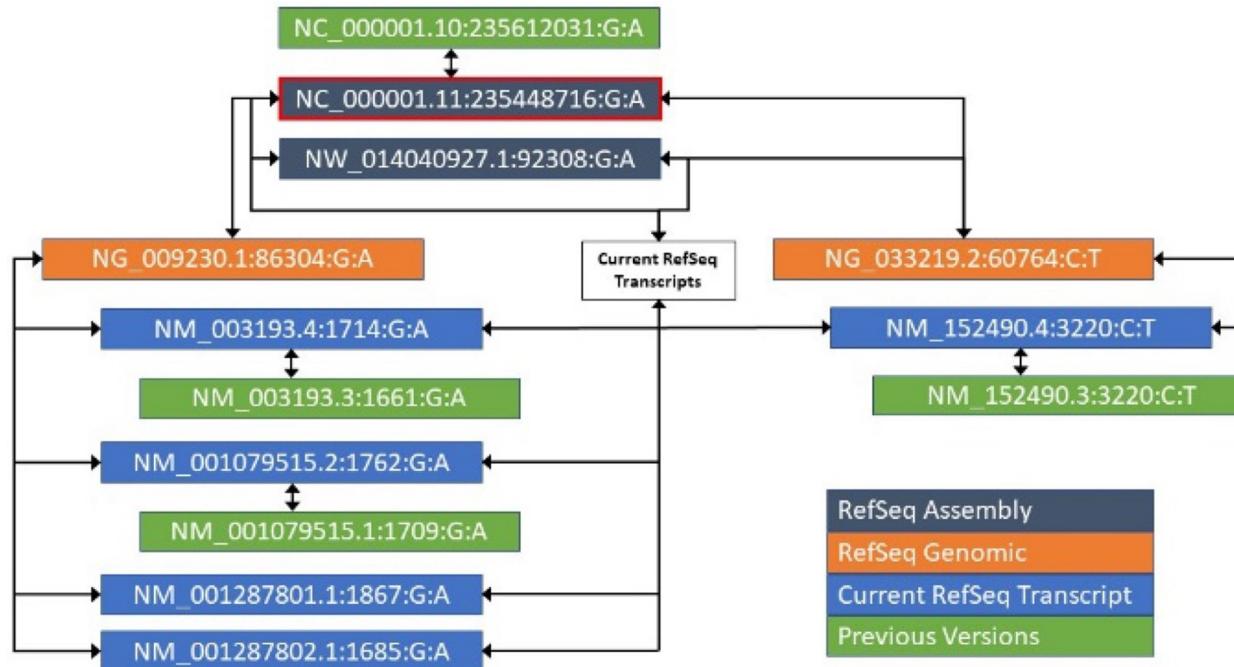
over-precise

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



The NCBI Sequence Position Deletion Insertion (SPDI) format

Fig. 1. For rs756655831, a representation of the alignments between various sequences, and the resulting SPDIs.



The SPDI Canonical Allele

The Canonical Allele extends identification across related, or congruent sequences, taking into account sequence changes (see Section 2.5). For the purposes of producing a reference catalog, all Contextual Alleles that are placed together in a canonical set are considered the same allele because they result in the same local sequence in a congruent region by alignment. That is, the Canonical Allele represents a set of congruent Contextual Alleles. One contextual representation is chosen as a Canonical Allele Representative and we use its Contextual SPDI as the identifier for the Canonical Allele.

Holmes JB, et al. *Bioinformatics* 2019



HGVS and the SPDI Canonical Allele in ClinVar

NM_004333.6(BRAF):c.2125C>G (p.Gln709Glu)

Allele ID: 1200728
Variant type: single nucleotide variant
Variant length: 1 bp
Cytogenetic location: 7q34
Genomic location: 7: 140739814 (GRCh38) GRCh38 UCSC
7: 140439614 (GRCh37) GRCh37 UCSC

HGVS:

Nucleotide	Protein	Molecular consequence
NM_004333.6:c.2125C>G MANE SELECT ⓘ	NP_004324.2:p.Gln709Glu	missense
NM_001354609.2:c.2125C>G	NP_001341538.1:p.Gln709Glu	missense
NM_001374244.1:c.2245C>G	NP_001361173.1:p.Gln749Glu	missense

... more HGVS

Protein change: Q621E, Q657E, Q672E, Q675E, Q687E, Q709E, Q712E, Q749E

Other names: -

Canonical SPDI: ⓘ NC_000007.14:140739813:G:C



Functional consequence: -

Global minor allele frequency (GMAF): -

Allele frequency: -

Links: [VarSome](#)

Coordinate Systems: Residue Coordinates

What is the meaning of coordinates **Sequence : 6–7**?

Insertion between AG in Sequence														
Sequence	T	C	A	G	C	A	G	C	A	G	C	A	G	T
Residue	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

Deletion/Substitution of AG in Sequence														
Sequence	T	C	A	G	C	A	G	C	A	G	C	A	G	T
Residue	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



Coordinate Systems: Inter-residue Coordinates

How can coordinate concepts be described unambiguously? (SPDI, VRS)

Sequence : 6-6

Insertion between AG in Sequence																
Sequence	T	C	A	G	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**

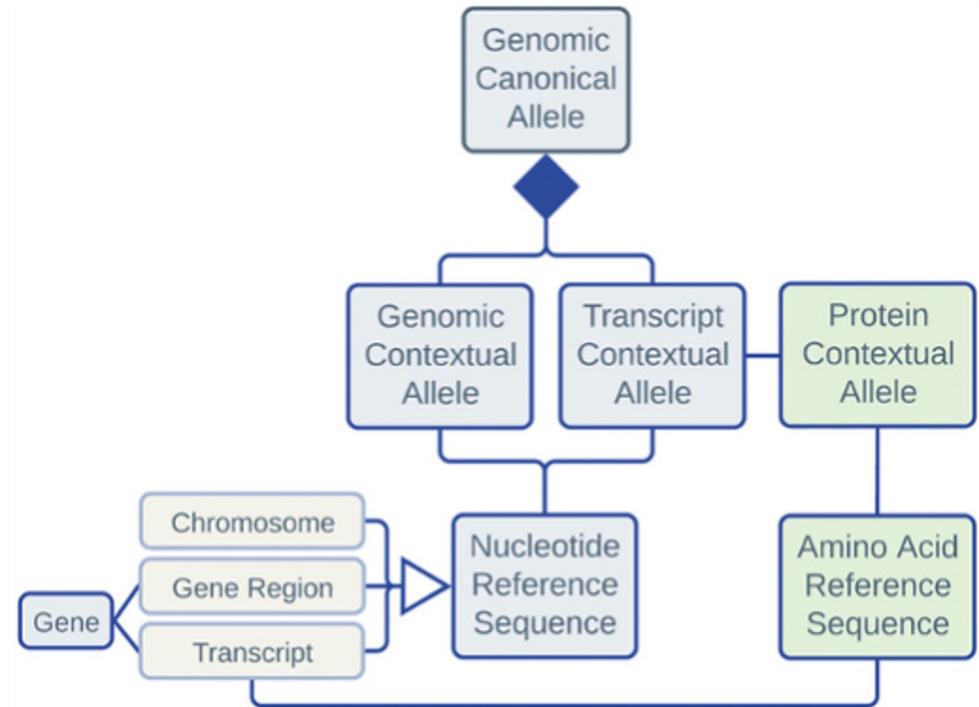
Deletion/Substitution of AG in Sequence																
Sequence	T	C	A	G	C	A	G	C	A	G	C	A	G	C	T	
Residue	1	2	3	4	5	6	7	8	9	10	11	12				
Inter-residue	0	1	2	3	4	5	6	7	8	9	10	11	12			

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

Sequence : 5-7

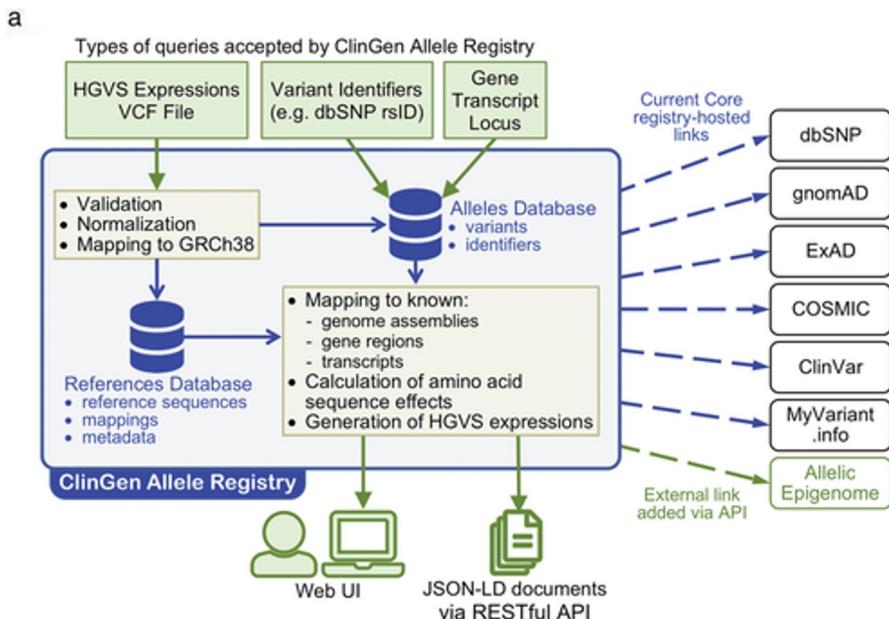


ClinGen Allele Registry



Figures from Pawliczek P, et al. *Hum. Mut.* 2018

WILEY



b

Canonical Allele Identifier: CA321211

Gene: NDUFS8 [HGNC] [NCBI]

Identifiers and link-outs to other resources

ClinVar Variation Id: 214835 ⓘ

ClinVar RCV Id: RCV000196794 ⓘ RCV000276295 ⓘ

dbSNP Id: rs369602258 ⓘ

ExAC: 11:67799758 C / T ⓘ

gnomAD: 11:67799758 C / T ⓘ

MyVariant Identifiers: chr11:g.67799758C>T (hg19) ⓘ

chr11:g.68032291C>T (hg38) ⓘ

JSON-LD ⓘ Calculator ⓘ

Canonical Allele Identifier: CA16602730

Gene: EGFR

[HGNC](#)

[NCBI](#)

Linked Data

ClinVar Variation Id: 376282 [↗](#)

COSMIC: COSM12429 [↗](#)

ClinVar RCV Id: RCV000425876 [↗](#)

PubMed: PMID:25157968 [↗](#)

dbSNP Id: rs1057519848 [↗](#)

[JSON-LD](#)

Genomic Alleles

HGVS

NC_000007.14:g.55191822_55191823delinsGT , CM000669.2:g.55191822_55191823delinsGT

Genome Assembly

[GRCh38](#)

NC_000007.13:g.55259515_55259516delinsGT , CM000669.1:g.55259515_55259516delinsGT

[GRCh37](#)

NC_000007.12:g.55227009_55227010delinsGT

[NCBI36](#)

NG_007726.3:g.177791_177792delinsGT , LRG_304:g.177791_177792delinsGT

Transcript Alleles

HGVS

ENST00000275493.7:c.2573_2574delinsGT [MANE Select](#)



Amino-acid change

[ENSP00000275493.2:p.Leu858Arg](#) [↗](#)



ENST00000275493.6:c.2573_2574delinsGT

[ENSP00000275493.2:p.Leu858Arg](#) [↗](#)

Domain-Specific Applications of Genomic Data Standards

Clinical / biomedical reporting

How do we compactly describe genomic variation in a way that humans readily understand it?

How do we encode sufficient information to do this accurately?

- “Nomenclature standards”
- Focus on text representation
- Many community-specific standards
- Readability emphasis

Domain-Specific Applications of Genomic Data Standards

Computer-driven discovery

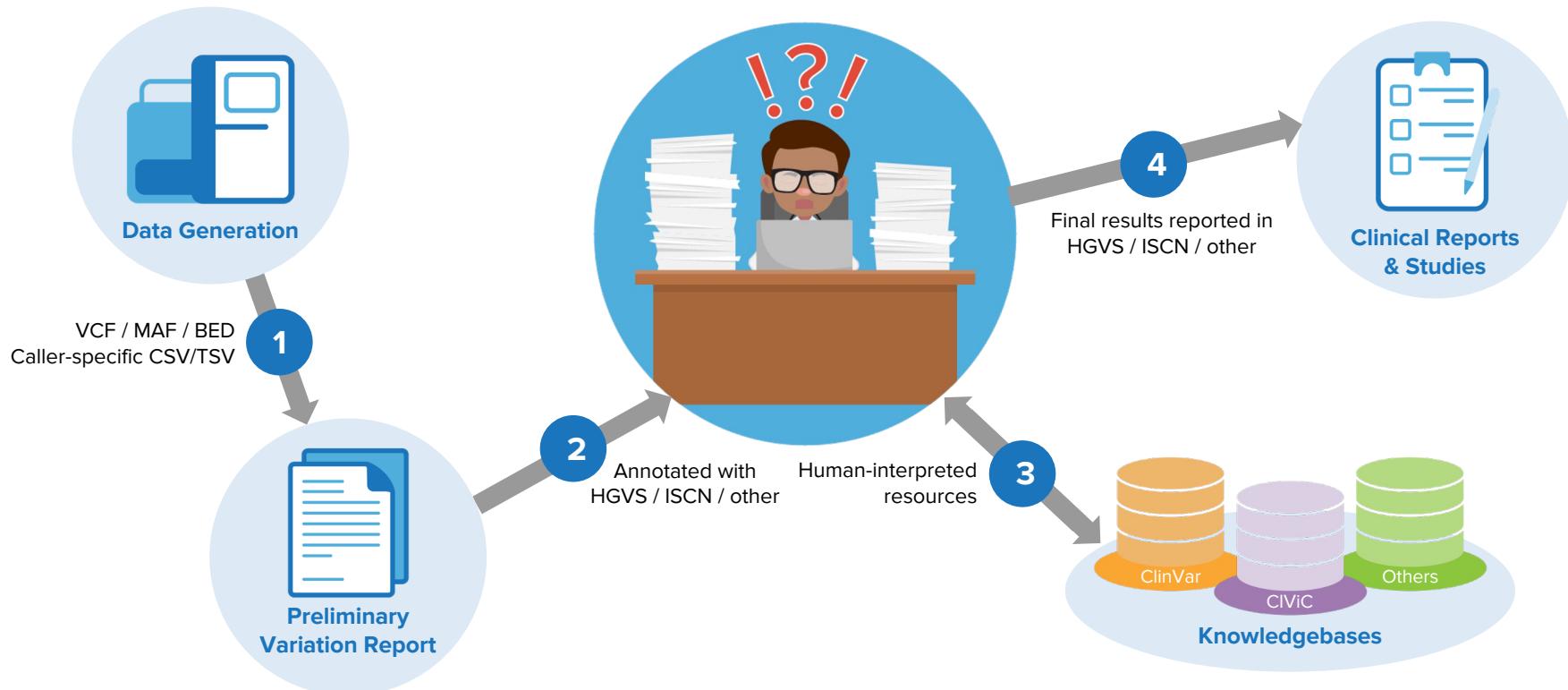
How do we unambiguously and computationally define genomic data concepts for AI-readiness?

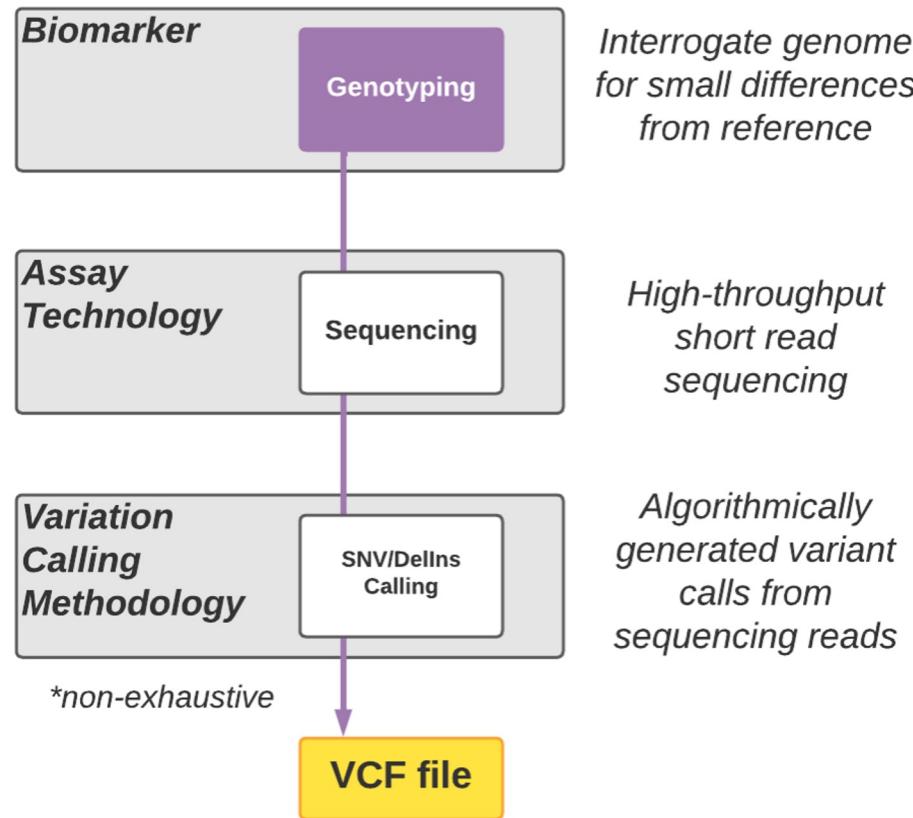
How do we unify diverse variation concepts and embed them in complex computable documents?

Human Evaluation Creates Bottleneck



Global Alliance
for Genomics & Health





Biomarkers*

Genotyping

Gene Fusions

RNA Expression

DNA Methylation

Structural Variation

Copy Number

Assay Technologies*

Histochemistry

PCR

Sequencing

Microarrays

Variation Calling Methodologies*

Human Expert Classification and Scoring

Expression Estimation

Sequence Junction Detection

SNV/DellIns Calling

Methylated DNA Calling

Copy Number Estimation

*non-exhaustive

Observed Variation

Introduction

Terminology & Information Model

Schema

Implementation Guide

Releases

Appendices

GA4GH Variation Representation Specification

The Variation Representation Specification (VRS, pronounced “verse”) is a standard developed by the Global Alliance for Genomic Health to facilitate and improve sharing of genetic information. The Specification consists of a JSON Schema for representing many classes of genetic variation, conventions to maximize the utility of the schema, and a Python implementation that promotes adoption of the standard.

Citation

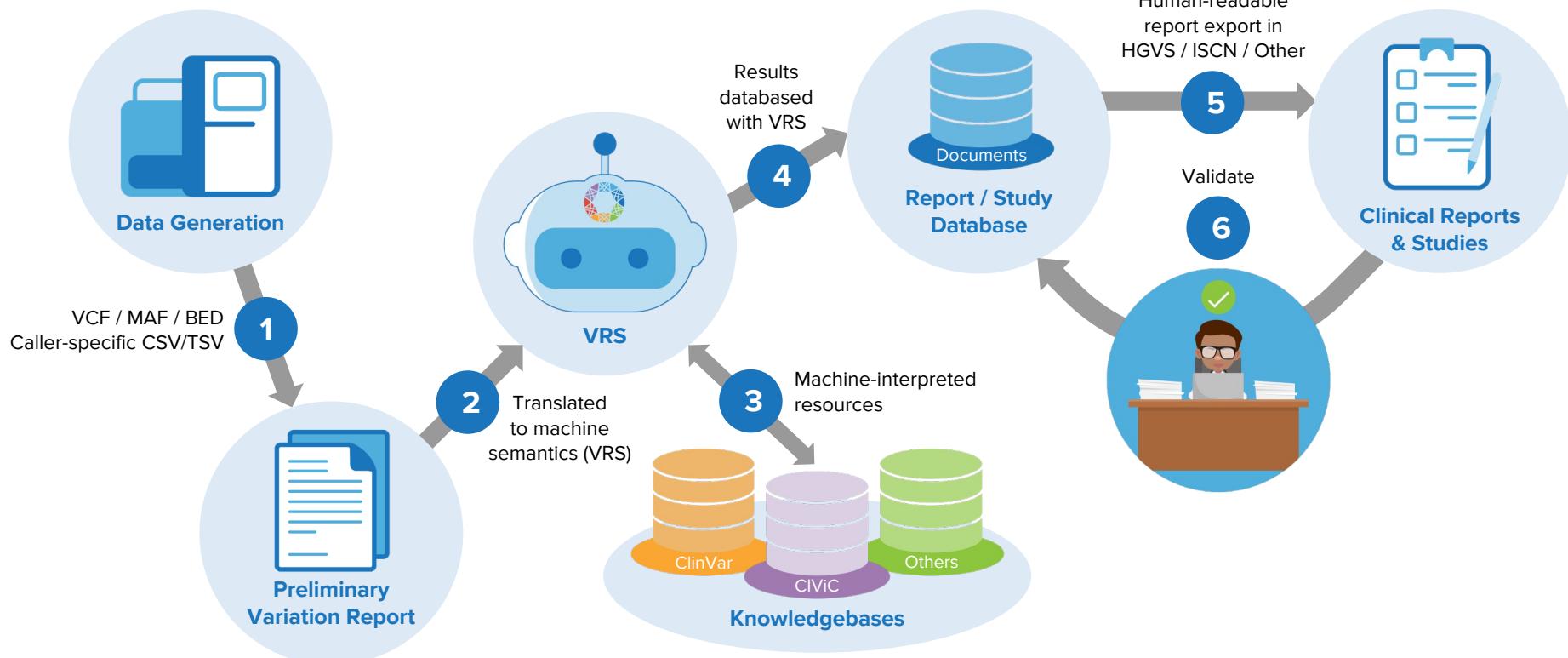
The GA4GH Variation Representation Specification (VRS): a computational framework for variation representation and federated identification. Wagner AH, Babb L, Alterovitz G, Baudis M, Brush M, Cameron DL, ..., Hart RK. *Cell Genomics*. Volume 1 (2021).
[doi:10.1016/j.xgen.2021.100027](https://doi.org/10.1016/j.xgen.2021.100027)

<https://vrs.ga4gh.org>

Computable Standards to Alleviate Bottleneck



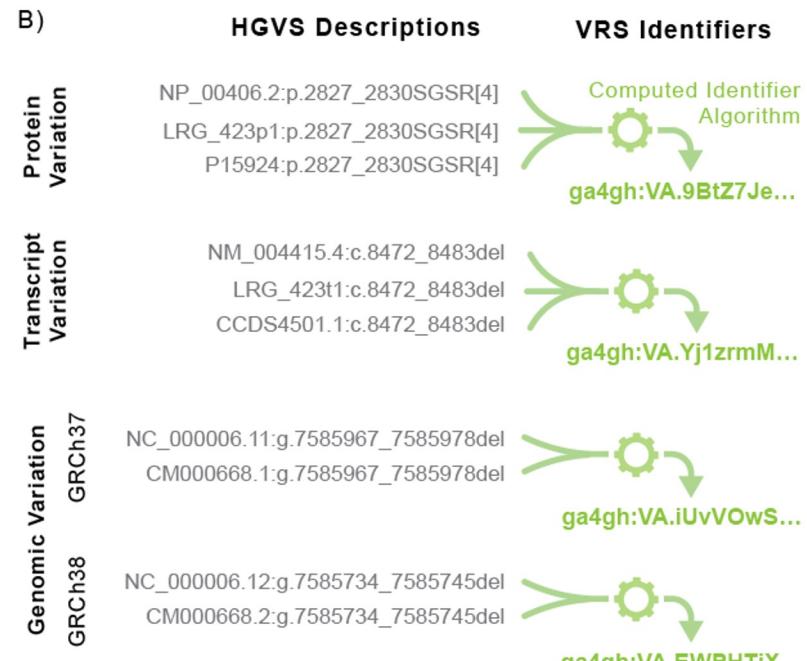
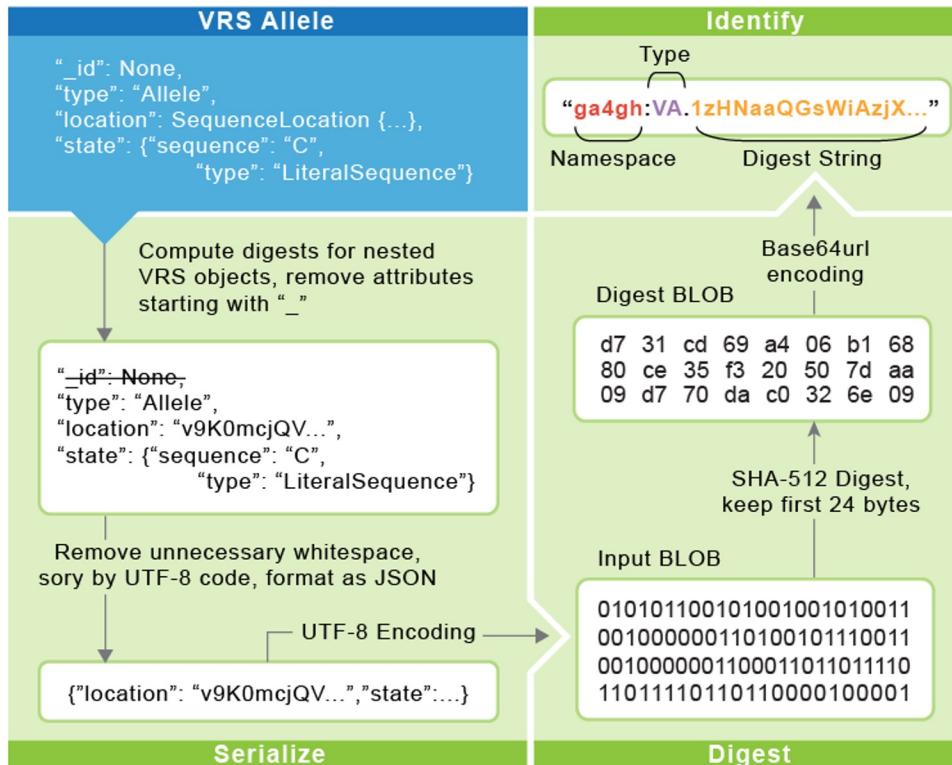
Global Alliance
for Genomics & Health



Converting Value Objects into Identifiers



Global Alliance
for Genomics & Health



Questions?



Practical: Be the Curator!



Bonus Content

Somatic Clinical Interpretation Resources



About Participate Community Help FAQ

 ahwagner 53 ▾

Go to Genes & Variants

Go!

BROWSE

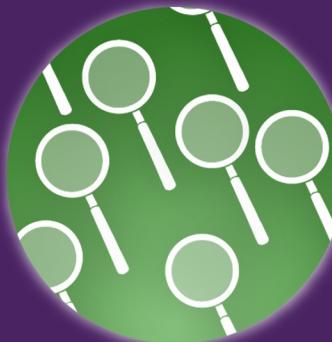
SEARCH

ACTIVITY

ADD ▾



Discover supported clinical interpretations of mutations related to cancer.



Participate with colleagues to add variants and support for cancer-related mutations.

 **V600E** RS113488022, VAL600GLU, V640E, VAL640GLU
[Summary](#)[Comments](#)[Revisions](#)[Flags](#)[Events](#)Curators: Editors: **Description**

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.

Sources

None specified

Aliases

RS113488022, VAL600GLU, V640E, VAL640GLU

Variant Type Missense Variant

NM_004333.4:c.1799T>A

HGVS

NP_004324.2:p.Val600Glu

Descriptions

NC_000007.13:g.140453136A>T

Gene BRAF**Allele Registry ID** CA123643**CIViC Variant Evidence Score**

1,353.5

ClinVar IDs 13961 376069**Representative Variant Coordinates**

Ref. Build	GRCH37	Ensembl Version	75
------------	--------	-----------------	----

Coordinates

Chr.	7
------	---

Start	140453136
-------	-----------

Stop	140453136
------	-----------

Ref. Bases	A
------------	---

Var. Bases	T
------------	---

Transcript	 ENST00000288602.6
------------	---

MyVariantInfo[Overview](#)[ClinVar](#)[gnomAD \(2.1.1\)](#)[EXAC \(0.3.1\)](#)[CADD](#) ...

Data and Knowledge Production

Millions of raw sequence reads are produced for a patient tumor.



TAACCAAGAT GATAAACGCCA TAACCCAGGAT
CCAGATTATCAGA CATACTTACAG CTCCTGGACGAC
CGAGATTATCAGA CATAGCTTACAG CACATTACCTTA

Sequences are aligned to the reference genome and tumor-specific events predicted.



Data are reviewed and validation experiments performed to identify high quality events.



Events are annotated and scored in an effort to predict events of functional significance.

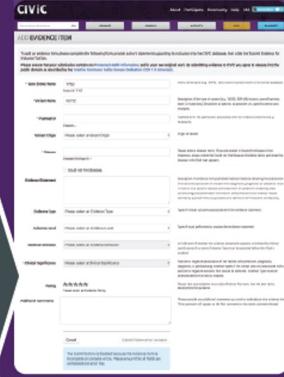


CIViC Curation

Crowdsourced curation efforts, moderated by experts in oncology and bioinformatics, help to build a knowledge-base of clinical interpretations of variants in cancer, describing the therapeutic, prognostic, diagnostic, and predisposing relevance of inherited and somatic variants of all types. Anyone may sign up to be a curator, add evidence, suggest changes to records, and discuss ongoing curation efforts.

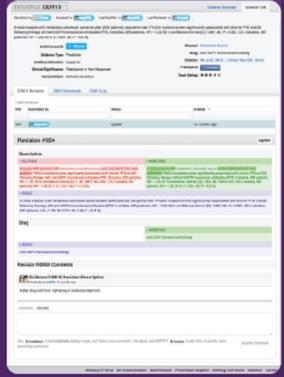


Add New Evidence



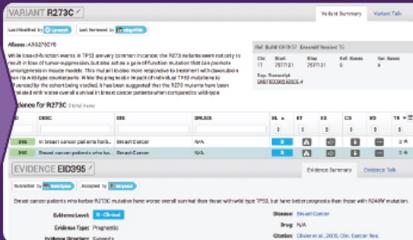
This screenshot shows the "Add New Evidence" form. It includes fields for "Evidence Name" (e.g., TP53), "Evidence Type" (e.g., TP53), "Evidence Source" (e.g., TP53), "Evidence Score" (e.g., 100), "Evidence Description" (e.g., TP53), "Evidence Summary" (e.g., TP53), "Evidence References" (e.g., TP53), "Evidence Prognostic" (e.g., TP53), and "Evidence Therapeutic" (e.g., TP53).

Review and Discuss Edits



This screenshot shows the "Review and Discuss Edits" form. It includes sections for "Review and Discuss Edits", "Review and Discuss Edits", "Review and Discuss Edits", and "Review and Discuss Edits". The "Review and Discuss Edits" section contains a table with columns for "Event ID", "Event Name", "Event Type", "Event Score", and "Event Description".

Research Gene, Variant, & Evidence Summaries



This screenshot shows the "Research Gene, Variant, & Evidence Summaries" page for variant R273C. It includes sections for "Variant R273C", "Gene Summary", "Variant Tab", "RefSeq Build 37/37 Exonized RefSeq ID: R273C", "Clinical Trials", "Evidence ID: EID995", "Evidence Summary", and "Evidence Details".

A genome analyst uses CIViC's summaries to interpret and prioritize functionally significant events in the context of published literature, clinical trials, and linked knowledgebases.

Pathologists and oncologists review analysts' reports to help evaluate the significance of potentially clinically actionable events and incorporate into patient care.





Precision Oncology Knowledge Base

595

Genes

4472

Alterations

38

Tumor Types

79

Drugs

Search Gene / Alteration

Level 1

FDA-approved

20 Genes

Level 2

Standard care

10 Genes

Level 3

Clinical evidence

25 Genes

Level 4

Biological evidence

14 Genes

Level R1

Standard care

4 Genes

Level R2

Clinical evidence

6 Genes



@handlerwagner

<https://oncokb.org/>

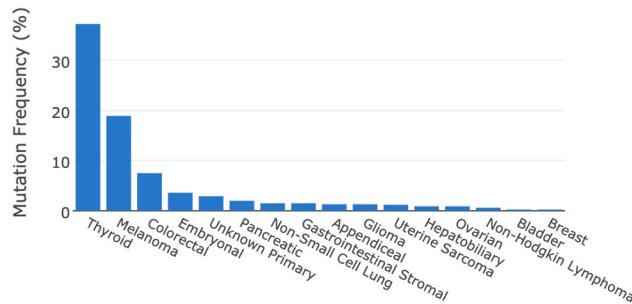
BRAF V600E

Oncogenic • Gain-of-function Level 1

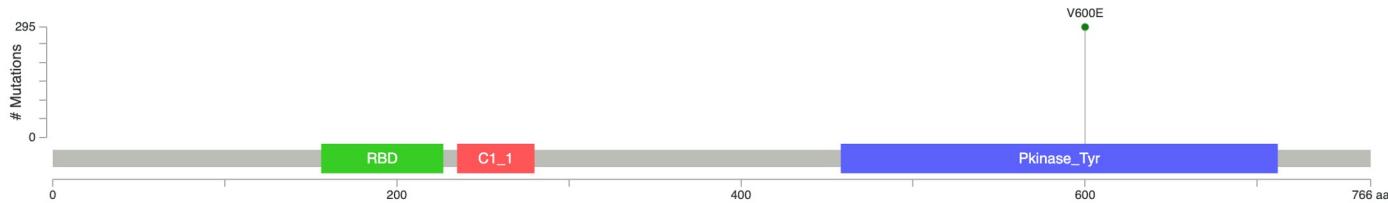
BRAF, an intracellular kinase, is frequently mutated in melanoma, thyroid and lung cancers among others. The BRAF V600E mutation is known to be oncogenic.

See additional BRAF information 

Cancer Types with BRAF V600E Mutations



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
<u>V600R</u>	Yes	Gain-of-function	12 references
<u>F247L</u>	Likely	Likely Gain-of-function	2 references
<u>T599dup</u>	Yes	Gain-of-function	4 references
<u>R462E</u>	Likely	Likely Gain-of-function	1 reference
<u>K601E</u>	Likely	Gain-of-function	6 references
<u>L597Q</u>	Yes	Gain-of-function	9 references
<u>V459L</u>	Yes	Gain-of-function	2 references
<u>G596C</u>	Likely	Gain-of-function	1 reference
<u>E275K</u>	Likely	Likely Gain-of-function	1 reference
<u>G466V</u>	Yes	Gain-of-function	9 references
<u>A728V</u>	Likely	Gain-of-function	1 reference
<u>PAPSS1-BRAF Fusion</u>	Likely	Gain-of-function	2 references
<u>SND1-BRAF Fusion</u>	Yes	Gain-of-function	4 references
<u>I511V</u>	Likely	Likely Gain-of-function	1 reference

Molecular Oncology Almanac

Browser

A collection of putative alteration/action relationships identified in clinical, preclinical, and inferential studies.

Search 145 molecular features associated with 865 assertions.

Multiple search terms may be combined. [Click here for search help.](#)

Or browse alterations by:

Cancer Type

Select from 64 cancer types:

Predictive Implication Level

Select from 6 predictive implication levels

Therapy

Select from 149 therapies:

Molecular Oncology Almanac Search Results

x Gene:ALK

Search

Multiple search terms may be combined. [Click here for search help.](#)

Click on any alteration below to view more details about the alteration-actionability relationship.

Show 10 ▾ entries

Feature type	Feature	Therapy	Response	Cancer Type	Predictive Level	▼
Rearrangement	ALK Fusion	Alectinib	Sensitivity	Non-Small Cell Lung Cancer	FDA-Approved	
Rearrangement	ALK Fusion	Crizotinib	Sensitivity	Non-Small Cell Lung Cancer	FDA-Approved	
Rearrangement	ALK Fusion	Lorlatinib	Sensitivity	Non-Small Cell Lung Cancer	FDA-Approved	
Rearrangement	ALK	Ceritinib	Sensitivity	Non-Small Cell Lung Cancer	FDA-Approved	
Rearrangement	ALK--EML4 Fusion	Crizotinib	Sensitivity	Inflammatory Myofibroblastic Tumor	FDA-Approved	
Rearrangement	ALK Fusion	Alectinib	Sensitivity	Non-Small Cell Lung Cancer	Guideline	
Rearrangement	ALK Translocation	Crizotinib	Sensitivity	Inflammatory Myofibroblastic Tumor	Guideline	
Rearrangement	ALK Translocation	Ceritinib	Sensitivity	Inflammatory Myofibroblastic Tumor	Guideline	
Rearrangement	EML4--ALK Fusion	Crizotinib	Sensitivity	Non-Small Cell Lung Cancer	Guideline	
Rearrangement	ALK	Ceritinib	Sensitivity	Non-Small Cell Lung Cancer	Clinical trial	

Showing 1 to 10 of 17 entries

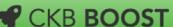
Previous 1 2 Next

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 (dabrafenib) 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: <input type="text"/>	

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

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 				
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

Germline and Specialized Interpretation Resources

Gene Focus: BRCA1 and BRCA2



HOME VARIANTS COMMUNITY HELP MORE ▾

search for "c.1105G>A", "brca1" or "IVS7+1037T>C"



The BRCA Exchange aims to advance our understanding of the genetic basis of breast cancer, ovarian cancer and other diseases by pooling data on BRCA1/2 genetic variants and corresponding clinical data from around the world. Search for *BRCA1* or *BRCA2* variants above.

This website is supported by the BRCA Challenge project, a driver project of the Global Alliance for Genomics and Health.

▶ Video Overview

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

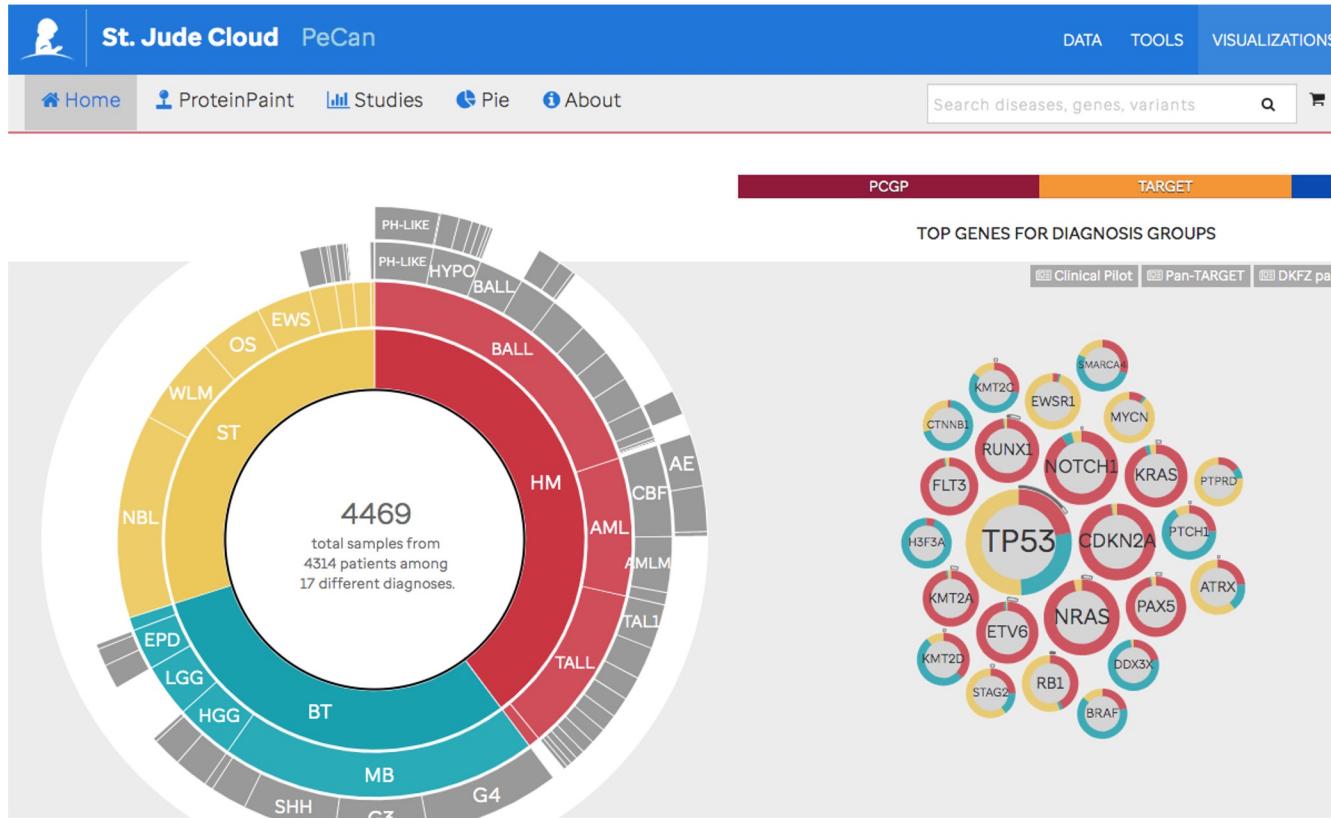
Clinical Significance (ENIGMA)	
Clinical Significance	Benign / Little Clinical Significance
IARC Class	Benign
Comment on Clinical Significance	IARC class based on posterior probability from multifactorial likelihood analysis, thresholds for class as per Plon et al. 2008 (PMID: 18951446). Class 1 based 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



@handlerwagner

<https://brcaexchange.org/variant/230931>

Disease Focus: Pediatric Cancers



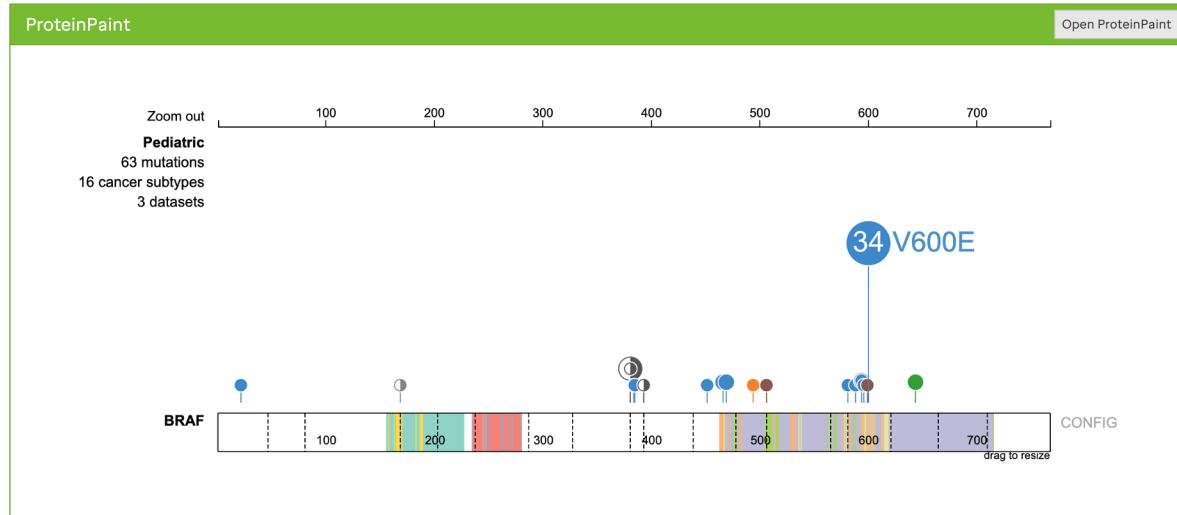
Disease Focus: Pediatric Cancers

NM_004333 SJ preferred
BRAF V600E GERMLINE SOMATIC
missense hgvs see hg19 copy
c.1799T>A p.Val600Glu

 Pathogenic
SOMATIC

Gene Information: BRAF

Entrez Description: This gene encodes a protein belonging to the RAF family of serine/threonine protein kinases. This protein plays a role in regulating the MAP kinase/ERK signaling pathway, which affects cell division, differentiation, and secretion. Mutations in this gene, most commonly the V600E mutation, are the most frequently identified cancer-causing mutatio... (imported on 2018-09-27) [see more...](#)





Licensing

BRCA Exchange

Cancer Biomarkers Database

ClinVar



Pecan PIE



OncoKB



More
Permissive

Less
Permissive



Attribution Only



Non-commercial /
Research Only
Custom Licensing

Non-commercial
and Share-Alike



Paid-Access Only

General Questions