BIO392
Bioinformatics of Genome Variations

Genome Variation | Function | Data Formats | Resources | Privacy

Michael Baudis **UZH SIB**
Computational Oncogenomics

University of Zurich[UZH]

Variant - Disease Knowledge Bases

VARIANT RESOURCES FOR CANCER GENOMICS

Resource name	Primary institute	Constituent Knowledge base	Cancer focused	Therapeutic evidence	Predisp. evidence	Diagnostic evidence	Prognostic evidence	Variant emphasis	URL
Cancer Genome Interpreter (CGI)	Institute for Research in Biomedicine, Barcelona, Spain	x	x	x				Somatic	https://www.cancergenomeinterpreter.org/home
Clinical Interpretation of Variants in Cancer (CIViC)	Washington University School of Medicine (WashU)	x	x	x	x	x	x	All variants	www.civicdb.org
JAX Clinical Knowledgebase (CKB)	The Jackson Laboratory	x	x	x	x	x	x	Somatic	https://ckb.jax.org/
Molecular Match	Molecular Match	x	x	x			x	Somatic	https://app.molecularmatch.com/
OncoKB	Memorial Sloan Kettering Cancer Center	x	x	x				Somatic	http://oncokb.org/#/
Precision Medicine Knowledgebase (PMKB)	Weill Cornell Medical College	x	x	x	x	x	x	Somatic	https://pmkb.weill.cornell.edu/
BRCA exchange	GA4GH	x	x		x			Germline	http://brcaexchange.org/
Cancer Driver Log (CanDL)	Ohio State University (OSU) / James Cancer Hospital		x	x				Somatic	https://candl.osu.edu/
Gene Drug Knowledge Database	Synapse		x	x		x	x	Somatic	https://www.synapse.org/#!Synapse:syn2370773/wiki/62707
MatchMiner	Dana-Farber Cancer Institute		x					Somatic	http://bcb.dfci.harvard.edu/knowledge-systems/
COSMIC Drug Resistance Curation	Wellcome Trust Sanger Institute		x	x				Somatic	http://cancer.sanger.ac.uk/cosmic/drug_resistance
My Cancer Genome	Vanderbilt University		x	x		x	x	Somatic	https://www.mycancergenome.org/
NCI Clinical Trials	National Cancer Institute of the National Institutes of Health		x					Somatic	www.cancer.gov/about-cancer/treatment/clinical-trials
Personalized Cancer Therapy Database	MD Anderson Cancer Center		x	x	x	x	x	Somatic	https://pct.mdanderson.org/#/home
ClinGen Knowledge Base	ClinGen				x			Germline	https://www.clinicalgenome.org/resources-tools/
ClinVar	National Center for Biotechnology Information			x	x			All variants	http://www.ncbi.nlm.nih.gov/clinvar/
Pharmacogenomics Knowledgebase (PharmGKB)	Stanford University			x				Germline	https://www.pharmgkb.org/
The Human Gene Mutation Database (HGMD)	Institute of Medical Genetics in Cardiff				x			Germline	http://www.hgmd.cf.ac.uk

Wagner et al. A harmonized meta-knowledgebase of clinical interpretations of cancer genomic variants. Nature Genetics (under review 2018)

CANCER VARIANT KNOWLEDGE BASES

- cancer variant knowledge databases report evidences for disease association (causative, therapeutic targets...)
- data selection is driven by arbitrary observations and sample selections
- limited overlap of reported variant associations is evidence for large gaps in knowledge

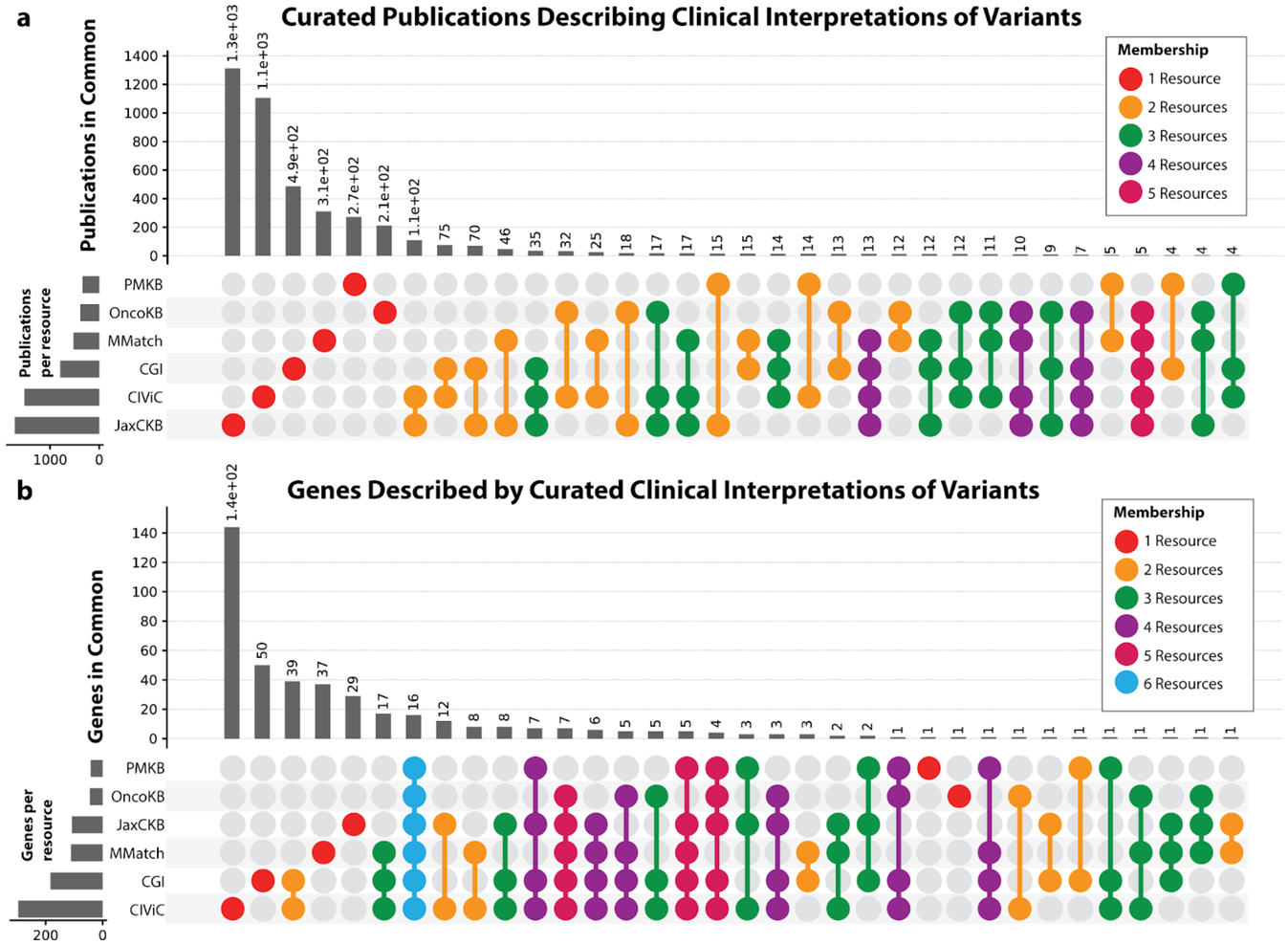


Figure S2 - Knowledgebase overlap

(a) Upset plot of publications supporting clinical interpretations of variants. the overwhelming majority of publications are observed in only 1 of 6 resources. **(b)** Upset plot of genes described by clinical interpretations of variants. Compared to other interpretation elements, genes are much more commonly shared between resources.

CANCER VARIANT KNOWLEDGE BASES

- cancer variant knowledge databases report evidences for disease association (causative, therapeutic targets...)
- data selection is driven by arbitrary observations and sample selections
- limited overlap of reported variant associations is evidence for large gaps in knowledge

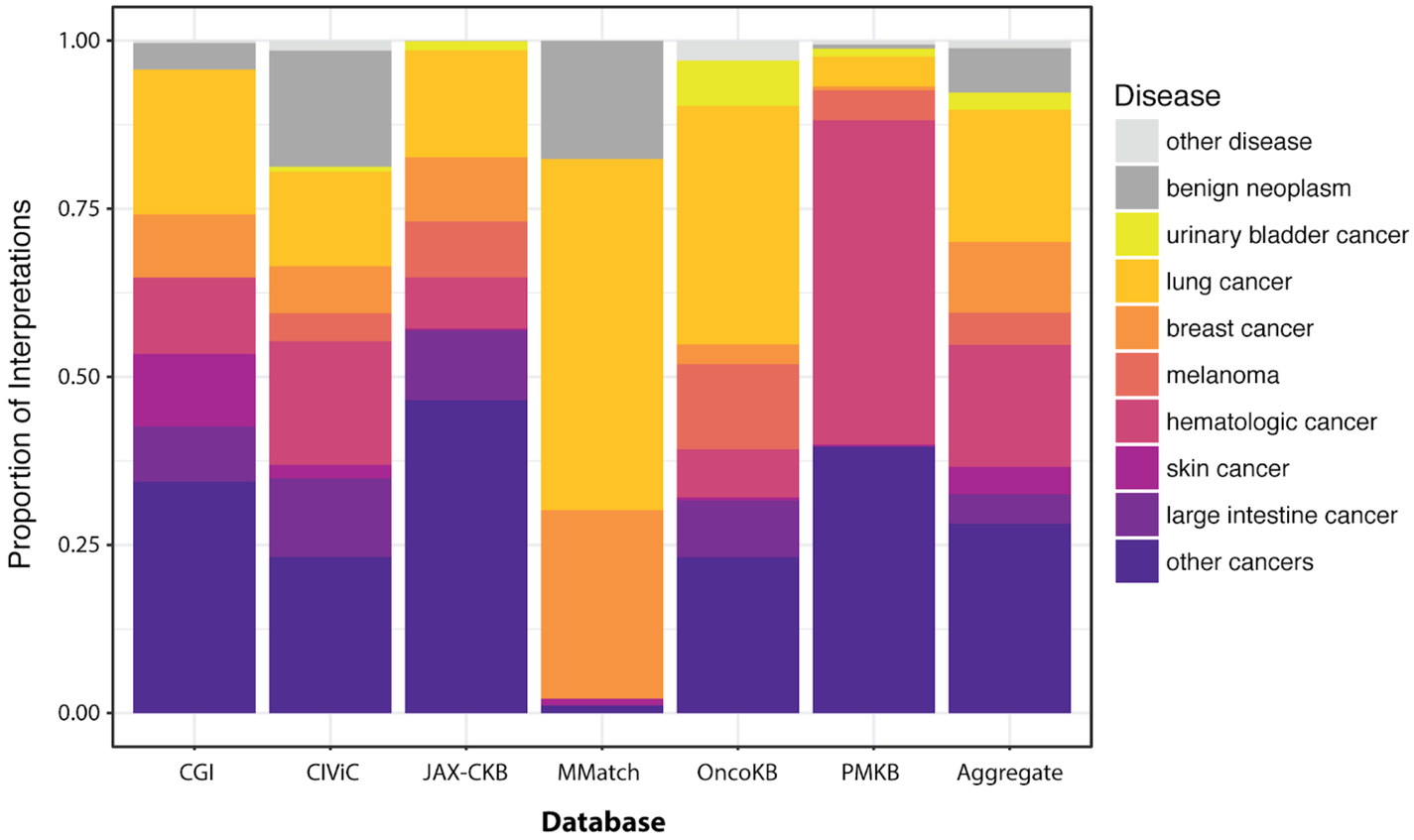
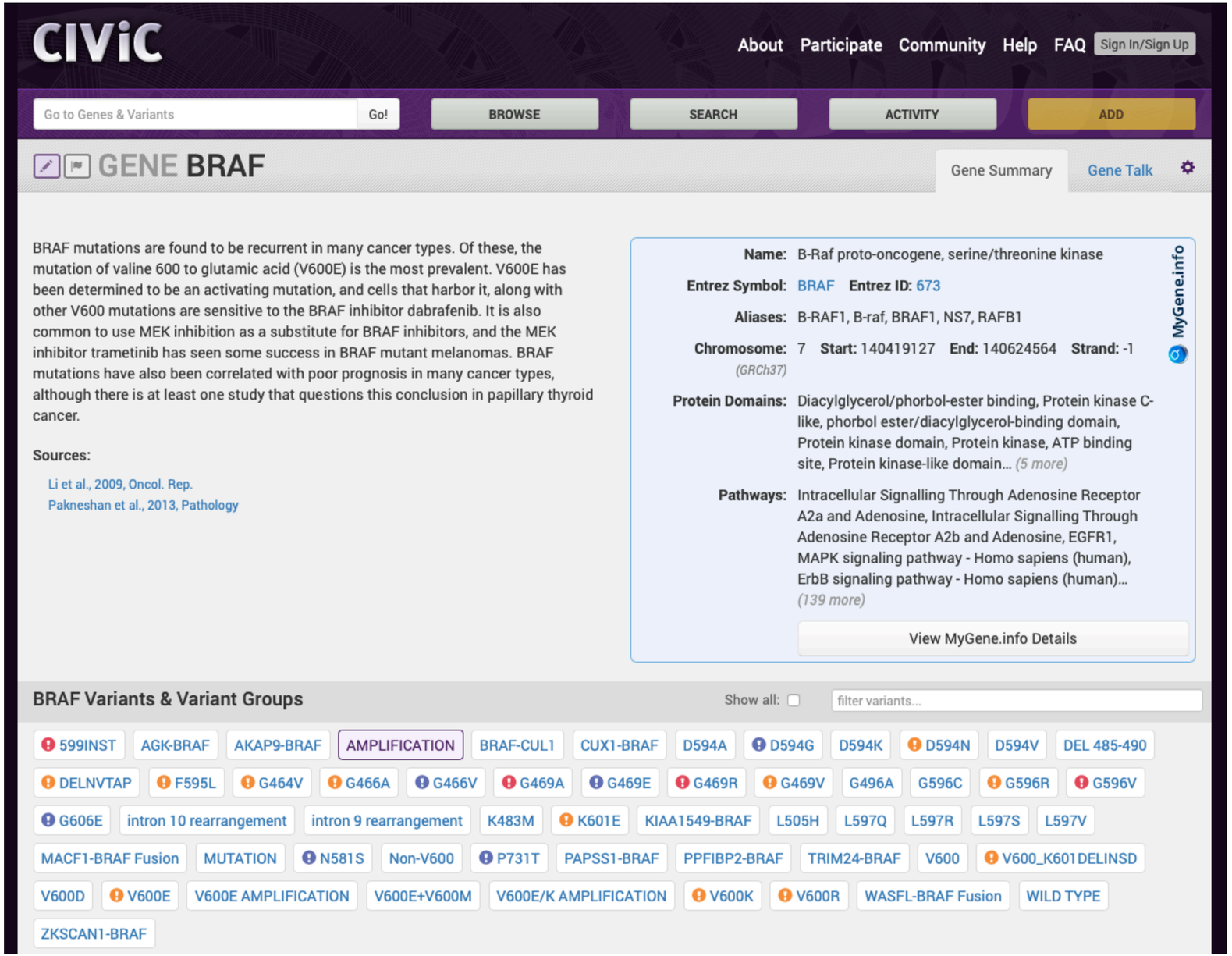


Figure S3 - Knowledgebase disease enrichment
Relative distribution of interpretations describing diseases across the VICC resources. Several resources are strongly enriched for one or more diseases compared to the entire dataset (see related **Table S8**).

CANCER VARIANT KNOWLEDGE BASES: CIVIC

- "CIViC is a community-edited forum for discussion and interpretation of peer-reviewed publications pertaining to the clinical relevance of variants (or biomarker alterations) in cancer."

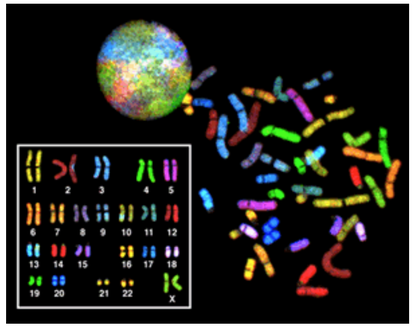


CIViC
Clinical Interpretations of Variants in Cancer

Variant Annotation Formats

GENOMIC VARIANT FORMATS: ISCN

- ISCN - "International System for Human Cytogenetic Nomenclature"
- Annotation format for chromosomal aberrations, i.e. traditional microscopically visible structural and quantitative abnormalities in karyotypes
- extensions for "molecular cytogenetics" (e.g. M-FISH, SKY, genomic arrays)



SKY - Spectral Karyotyping of tumour metaphase (source: https://www.genome.gov)

Symbol	Description
,	Separates modal number (total number of chromosomes), sex chromosomes, and chromosome abnormalities
-	Loss of a chromosome
( )	Grouping for breakpoints and structurally altered chromosomes
+	Gain of a chromosome
;	Separates rearranged chromosomes and breakpoints involving more than one chromosome
/	Separates cell lines or clones
//	Separates recipient and donor cell lines in bone marrow transplants
del	Deletion
der	Derivative chromosome
dic	Dicentric chromosome
dn	*de novo* (not inherited) chromosomal abnormality
dup	Duplication of a portion of a chromosome
fra	Fragile site (usually used with Fragile X syndrome)
h	Heterochromatic region of chromosome
i	Isochromosome
ins	Insertion
inv	Inversion
.ish	Precedes karyotype results from FISH analysis
mar	Marker chromosome
mat	Maternally-derived chromosome rearrangement
p	Short arm of a chromosome
pat	Paternally-derived chromosome rearrangement
psu dic	*pseudo dicentric* - only one centromere in a Dicentric chromosome is active
q	Long arm of a chromosome
r	Ring chromosome
t	Translocation
ter	Terminal end of arm (e.g. 2qter refers to the end of the long arm of chromosome 2)
tri	Trisomy
trp	Triplication of a portion of a chromosome

Source: Wikipedia (https://en.wikipedia.org/wiki/Cytogenetic_notation)

- Linear annotation of single-letter **nucleotides** or amino acid codes
- leading information line, usually with unique SeqID
- text format
  - "readable"
  - not optimised for size
- representation of a sequence without ambiguities or QC data
  - extended as "FASTQ" (Sanger Centre)

```
>NC_000007.14:11369935-11832697 Homo sapiens chromosome 7, GRCh38.p12 Primary Assembly
AGGGCTTAAATGGTTCCCTACTTACTTACATTTTAGCAAATAGCTATTTCAGAAAATGTTTTAACTGCAA
ACTACCCCGGAAGTAACTTGTCTTAAGTTGTGTGTCCCTTCCTGAATTTGTTAAGGCATAAGTTTCTGCT
TTGACTTTTAGGTTTGGTTTTTTGTGGGTAGACACAGGGACAAGAGACAGTGAGGGATGTGCCTTATGAC
TGATTGGGTGGGAAAAGCTGTACTCTGTTGTAGAGAGTTTCCCCATCCTCTGCTGCTGCCATTTTTGAAT
TGACTGGAAACCAGGAGGTCCCTTGTCCATGATTCACCTGGTGGCCTCAGCCAACTTTCAAAGTAAAAGT
TTGCATTTCTTGAACTTTCTAAATTGGAGTTGTTATACAACCCAGGAAAGGGCAATACAGTAGGTAAAAG
GATTTAGGTATTCACTGGAAAAAAATTTTAATCCATATTTAAGAAGCAATTGGTGCAAATCAAAACACAG
ATACACATGATTAGAATGAAAATGATTTCCGTATTTATGTTTGTCAGCAATATAGTTATTTACAAATAAC
CCATATGAAAATGTAAAAAAGCATATTACATCTTCACATGCCATCTGTATTGACTGAATAAAGCTTAGTG
ACATTATTTGCAAATCTGTAGTTAATTGTACATAGACATTTTGTGCGTTAAAAAGGAAATGTACATAATG
TAAAATAAATTACATTACGCAATTTACAAAGTAATATTAACAAAAATTCTTAGACAGCTGCCTCCTTATT
TAAACAAAATAAATATTCAGGTAGTTAAATTAACATAAAATACACATTTAGGGAATAATAATTTATAGAA
AGACAGATTGCAAATTTTAATAAAGTTATATTTTACAATGATAGATACTGATCTCTTCAAATCTGTGTGA
TAGAAATGGAGAAAAAAAGTACCAGAAAAGGAATCTAAATGTTTACTTTCTAAAAATAAACACAAACAGA
TTTCTGAAAAATAGGGAAAAGTTTACTTGAGGGTTAAAAGTAGGTAATCTAGAAACTATGGCTTAAAAAC
AATAAATCTACAAAACACAACAAGACTGACAATTATATTCTAAATAATATAGAGATTTTTGATCACTGAA
AACATGACTCCCACAAACTAAAGCTCTTCATATACTGCCCATTTTTAAGATCTGACTTGGTAGAAACACA
GAAAAATAAAAATGCAAAATTAACTGTTTAGCATTAGTTTCATTCATTTAATAATGTAGACATAACCATT
TTTCATTGTCCCTGCTAGATATAAAATTATTATAACACACTGCAAACCATCCTTTTTATAAATGGATAAC
TATTTGCTTGGCTCACACACCAGTTTCTGATACCTGAAATCCTGTCCTGCAGCCAGGGCCACCTGAGGGC
AGGACCTGGGGAGACCCTTATTCCCAGAACAAGCAGATGTAGTTCTCACAACTAAACTAGTCCCAGGAAA
GATCACATTCTGACAAGATTCCTCACAGATTTGCTCAAGGACTACTGTTTTTTCAACACCCTCAATCTTA
CAGTGGAAATAGAAGAAGAACCCACACTTGAATTTGGTAATATATTTATAAACAGGAGATCCCAGATCAT
TTGGGAATTGTGCTTCTCATGTACTATTGAGACCCACGTCAGCTTTAGAACAGGCTCTCCCTTCTGTATG
GTACTGAAAAGTACAGTCCTCCCTCACTGTCTTGTGGTGTGGAATGTGAACAAAGACTCGAGATGGAGGC
AGGAGGATATGGGATGGTCTAAAGCAAGTGTAGGCATGGACATTTTTACAGAAAAGGGCTTTTTTTTTTT
TTTTTTTCCTGCATGCCTTCCACATTTTTCCTTTTCCCTTTATTCATTTCTTTGACCAGTGGATTTGGTG
...
```

Homo sapiens chromosome 7, GRCh38.p12 Primary Assembly

NCBI Reference Sequence: NC_000007.14

GENOMIC VARIANT FORMATS: DBVAR

- dbVar is "NCBI's database of human genomic structural variation — insertions, deletions, duplications, inversions, mobile elements, and translocations"
- structural genome variations are still not completely solved with respect to unambiguous annotation

ncbi.nlm.nih.gov/dbvar/content/overview/

Variant Call Type	Sequence Ontology ID	Variant Region Type
copy number gain	**SO:0001742** A sequence alteration whereby the copy number of a given region is greater than the reference sequence.	copy number variation
copy number loss	**SO:0001743** A sequence alteration whereby the copy number of a given region is less than the reference sequence.	copy number variation
duplication	**SO:0001742** (copy number gain) A sequence alteration whereby the copy number of a given region is greater than the reference sequence.	copy number variation
deletion	**SO:0000159** The point at which one or more contiguous nucleotides were excised.	copy number variation
insertion	**SO:0000667** The sequence of one or more nucleotides added between two adjacent nucleotides in the sequence.	insertion
mobile element insertion	**SO:0001837** A kind of insertion where the inserted sequence is a mobile element.	mobile element insertion
novel sequence insertion	**SO:0001838** An insertion the sequence of which cannot be mapped to the reference genome.	novel sequence insertion
tandem duplication	**SO:1000173** A duplication consisting of 2 identical adjacent regions.	tandem duplication
inversion	**SO:1000036** A continuous nucleotide sequence is inverted in the same position.	inversion
intrachromosomal breakpoint	**SO:0001874** A rearrangement breakpoint within the same chromosome.	translocation or complex chromosomal mutation
interchromosomal breakpoint	**SO:0001873** A rearrangement breakpoint between two different chromosomes.	translocation or complex chromosomal mutation
translocation	**SO:0000199** A region of nucleotide sequence that has translocated to a new position.	translocation
complex	**SO:0001784** A structural sequence alteration or rearrangement encompassing one or more genome fragments.	complex
sequence alteration	**SO:0001059** A sequence_alteration is a sequence_feature whose extent is the deviation from another sequence.	sequence alteration
short tandem repeat variation	**SO:0002096** A kind of sequence variant whereby a tandem repeat is expanded or contracted with regard to a reference.	short tandem repeat variation

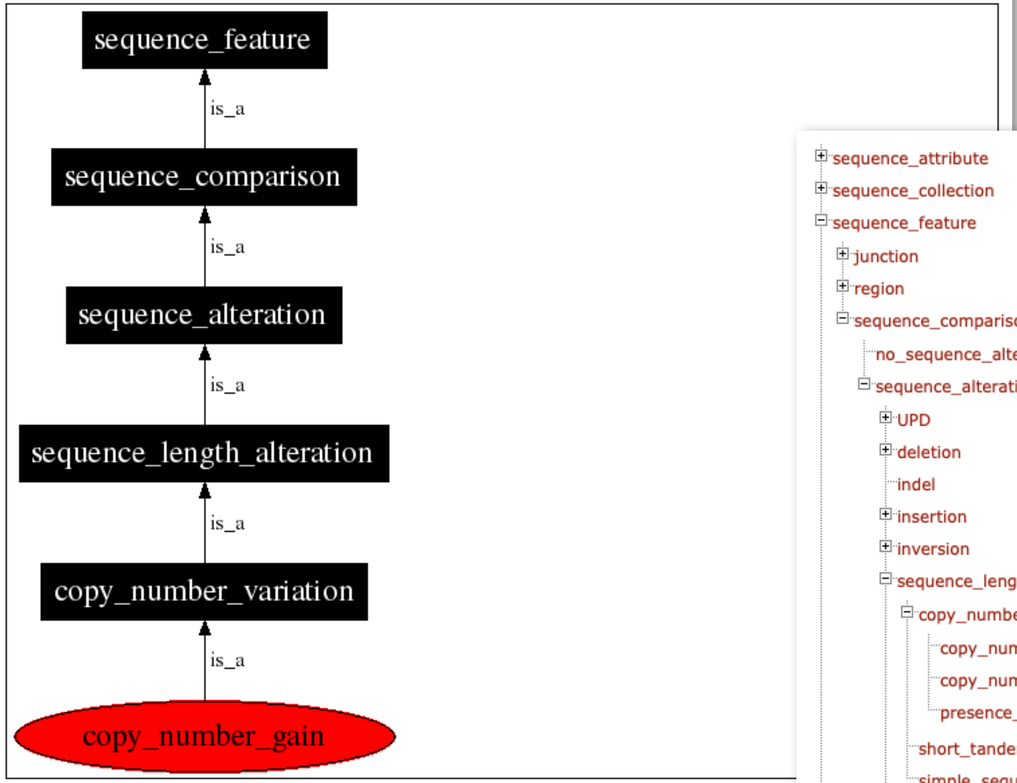
Source: Wikipedia (https://en.wikipedia.org/wiki/Cytogenetic_notation)

- Sequence Ontology describes types of biological sequence alterations (or normal status)
- It is by itself not suitable for complete variant description (e.g. lacking the localisation; has to be attached to a sequence or functional element)

copy_number_gain (CURRENT_SVN)

SO Accession:	SO:0001742 (SOWiki)
Definition:	A sequence alteration whereby the copy number of a given regions is greater than the reference sequence.
Synonyms:	gain, copy number gain
DB Xrefs:	SO: ke
Parent:	copy_number_variation (SO:0001019)

In the image below graph nodes link to the appropriate terms. Clicking the image background will toggle the image between large and small formats.



- sequence_attribute
- sequence_collection
- sequence_feature
  - junction
  - region
  - sequence_comparison
    - no_sequence_alteration
    - sequence_alteration
      - UPD
      - deletion
      - indel
      - insertion
      - inversion
      - sequence_length_alteration
        - copy_number_variation
          - copy_number_gain
          - copy_number_loss
          - presence_absence_variation
          - short_tandem_repeat_variation
          - simple_sequence_length_variation
      - structural_alteration
      - substitution
- sequence_variant

The Sequence Ontology

- HGVS allows the annotation of sequence variants (DNA, RNA, protein) with relation to a genomic ("g") or protein ("c") reference

HGVS Variation Examples

A Single Nucleotide Variant : rs268

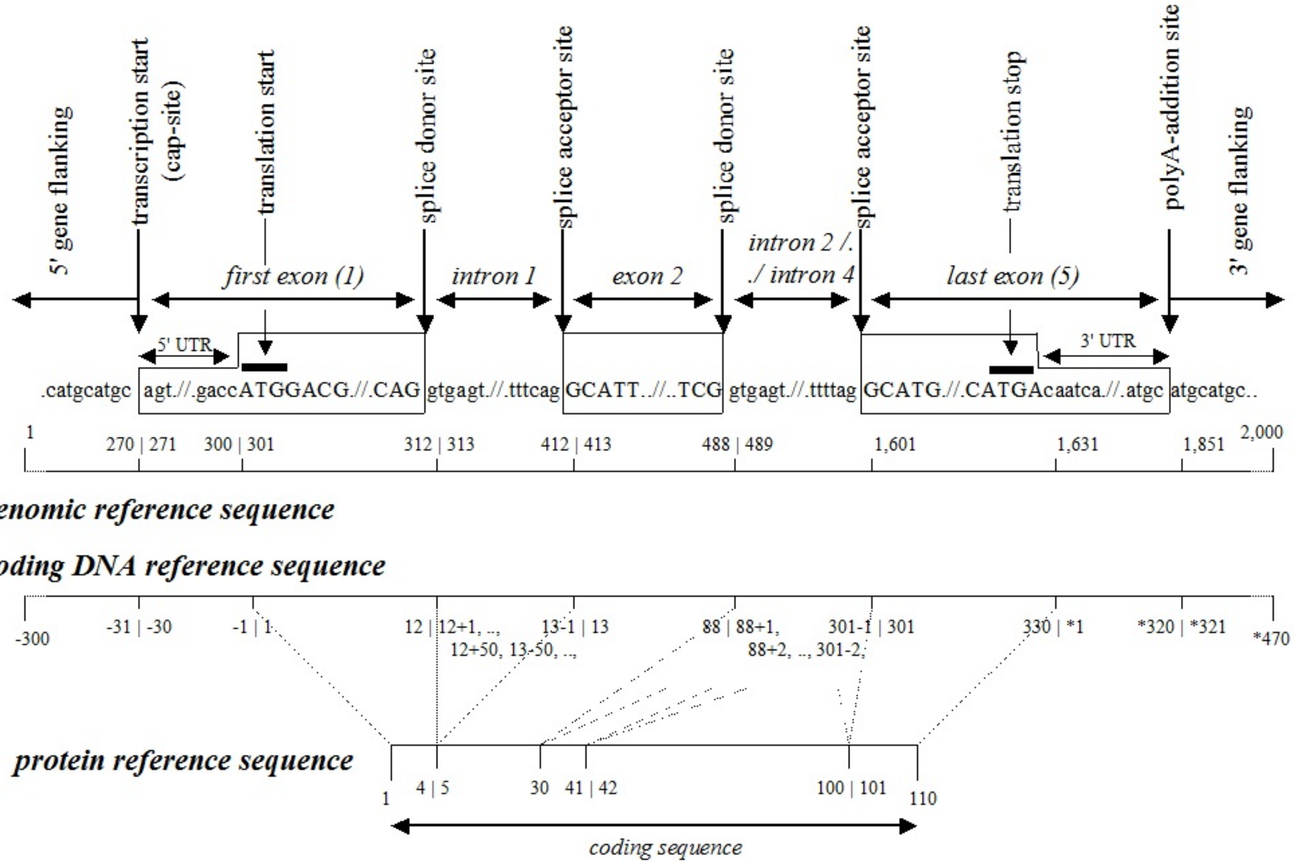
- NC_000008.10:g.19813529A>G
- NG_008855.1:g.21948A>G
- NM_000237.2:c.953A>G
- NP_000228.1:p.Asn318Ser

An Insertion Variant : rs9281300

- NC_000006.11:g.31239170_31239171insA
- NG_029422.2:g.5738_5739insT
- NM_001243042.1:c.344-46_344-45insT
- NM_002117.5:c.344-46_344-45insT

A Deletion Variant : rs1799758

- NC_000016.9:g.2138200_2138203delTGAG
- NG_005895.1:g.43894_43897delTGAG
- NM_000548.3:c.5161-28_5161-25del

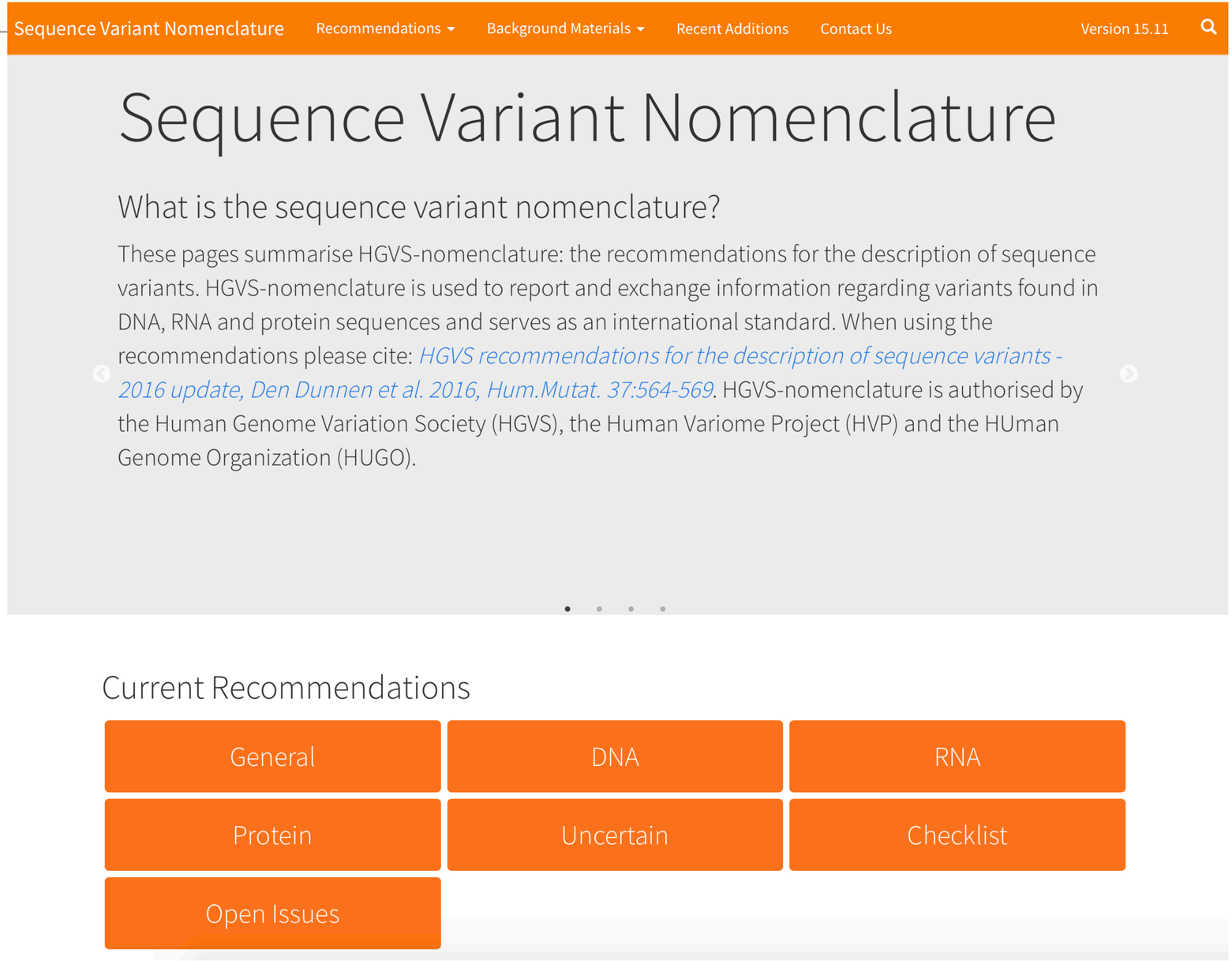


Source NCBI:
https://www.ncbi.nlm.nih.gov/variation/tools/reporter/docs/examples

http://www.hgvs.org/mutnomen/refseq_figure.html

GENOMIC VARIANT FORMATS: HGVS

- "The consistent and unambiguous description of sequence variants is essential to report and exchange information on the analysis of a genome. In particular, DNA diagnostics critically depends on accurate and standardized description and sharing of the variants detected. The sequence variant nomenclature system proposed in 2000 by the Human Genome Variation Society has been widely adopted and has developed into an internationally accepted standard."



All of these are the same variant. Or not.

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

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

Reece Hart, 2016

"The simplest thing that might work."

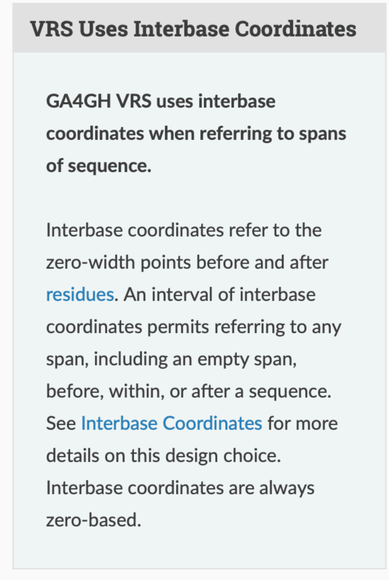
```
"vmc:allele": {
    "reference_sequence_id": "NCBI:NM000059.3",
    "interval": {"start": 50, "end": 51},
    "edit": "A"
},

"vmc:genotype": {
    "alleles": [
        {
            "reference_sequence_id": "NCBI:NM000059.3",
            "interval": {"start": 50, "end": 51},
            "edit": "A",
        },
        {
            "reference_sequence_id": "NCBI:NM000059.3",
            "interval": {"start": 50, "end": 51},
            "edit": "T",
        }
]},
```

Or…


```
{
   "vmc:alleles": [{
      “id”: “VA_5e632de6e7280769”,
      "reference_sequence_id": "VS_451ec666acc937f1",
      "interval": {"start": 50, "end": 51},
      "alternate": "A"
      }, (more alleles)
        ],

   "vmc:genotypes": [{
      “id”: “VG_5e632de6e7280769”,
      "allele_ids": [“VA_5e632de6e7280769”, “VA_72802de6e7695e63”]
      }, (more genotypes)
      ]

   "vmc:haplotypes": [{
      “id”: “VH_de8d7b851fb84223”,
      "allele_ids": [“VA_5e632de6e7280769”, “VA_d7b851fb84223de8”]
      }, (more haplotypes)
   ]

   "vmc:diplotype": [{
      “id”: “VD6fd159c94192f252”,
      "haplotype_ids": [“VH_de8d7b851fb84223”, “VH_b851fb8de8d74223”],
      }, (more diplotypes)
   ]
}
```

Task: Exploration of variant annotation formats

- Which "genomic" variant formats exist & what are their use cases?
    - ISCN
    - HGVS
    - VCF
    - GA4GH Variation Representation Specification
- Genomic coordinate systems
    - 0 or 1-based
    - "interbase"



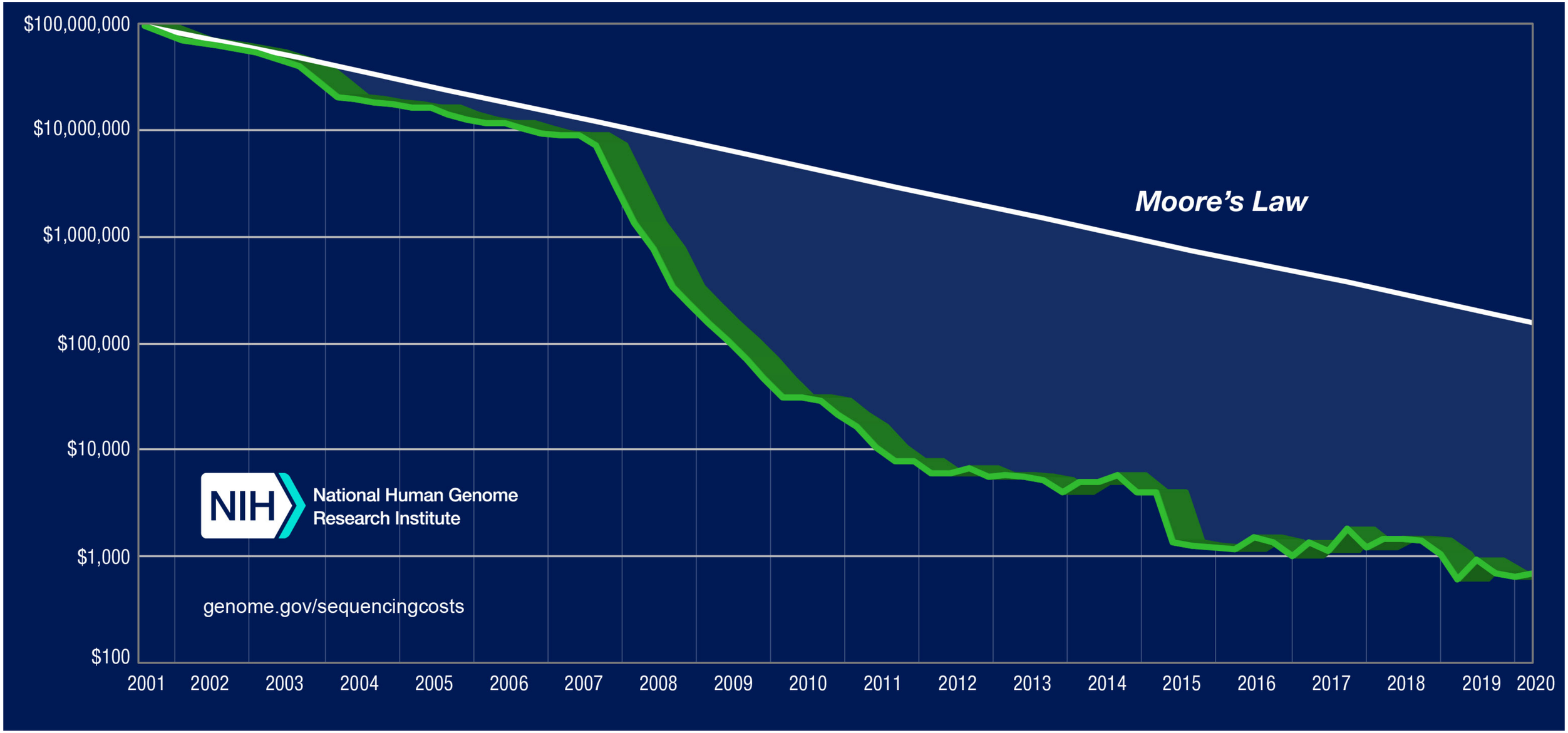
VRS Uses Interbase Coordinates

GA4GH VRS uses interbase coordinates when referring to spans of sequence.

Interbase coordinates refer to the zero-width points before and after residues. An interval of interbase coordinates permits referring to any span, including an empty span, before, within, or after a sequence. See Interbase Coordinates for more details on this design choice. Interbase coordinates are always zero-based.

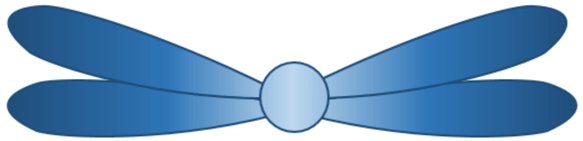
Genomes and Files

The Cost of Sequencing a Human Genome

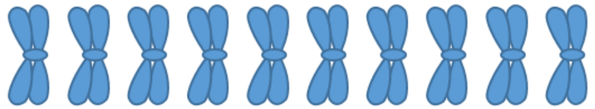


Genome Sequencing

whole genome sequencing (WGS)

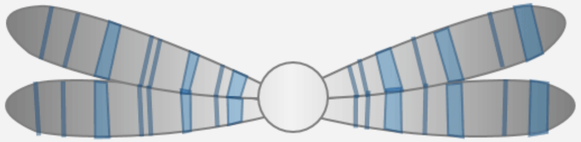


100%



all DNA
(3.1 billion base pairs)

exome sequencing



~1%

protein-coding DNA only
(~31 million base pairs)

What does it cost to sequence a genome?

	Human Genome Project (HGP): 1991-2003	today: 2017
cost:	$2.7 billion	~$1,500
time:	12+ years	< 2 days

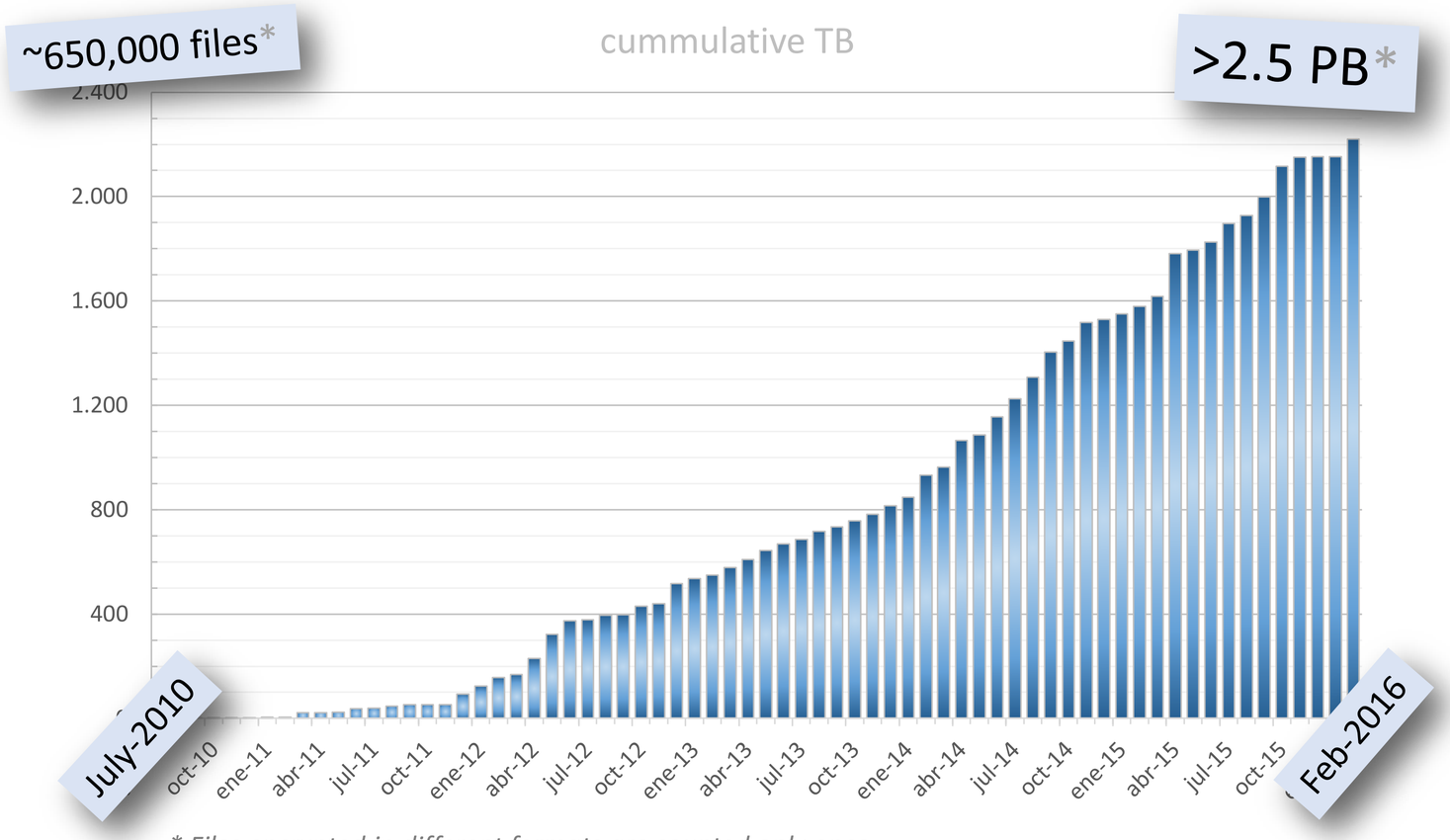
today:
2017

~$530
~3 days

Ryan L. Collins & Brad Wierbowski (2017)

Growth of Genome Data Repositories: Example EGA

The EGA contains a growing amount of data



International Cancer Genome Consortium

IRDiRC INTERNATIONAL RARE DISEASES RESEARCH CONSORTIUM

UK 10K

wellcome trust sanger institute

** Files encrypted in different formats are counted only once*

Source: Jordi Rambla (CRG Barcelona) 2016

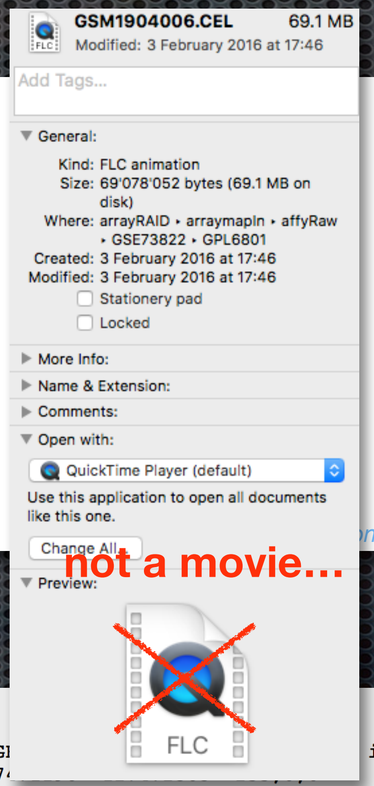
WHAT IS A PB, FOR HUMAN GENOMES? IT DEPENDS.

- 2 bits per base are sufficient to encode TCGA
  - using 00, 01, 10, 11
  - [TCGA]{3’000’000’000}
  - 2 * 3 * 10^9b = 6, 000, 000, 000 b
  - perfect genome (no overhead): ~715 MB
  - 1PB => ~1’400’000 genomes
- according to Swiss online store (Sep 2020) ~30'000CHF (65x16TB disks)
- this is less than a PhD position per year in Switzerland …
- (real costs are 2x that, + duplication, facilities, service … => ~400’000CHF)
- **However: A single 30x BAM file => 100GB**
- Still: 400’000CHF => 1PB => 10’000 genomes => 40CHF/genome (BAM format)

Bioinformatics: **File Formats**

- **text** or **binary** file formats, optimised for specific types of biological data
- examples from genomics:
  - **BAM** - compressed binary version of Sequence Alignment/ Map (SAM)
  - **BED** (Browser Extensible Data) -flexible way to define the data lines in an genome browser annotation tracks
  - **VCF** (Variant Call Format)

- Axt format
- BAM format
- BED format
- BED detail format
- bedGraph format
- barChart and bigBarChart format
- bigBed format
- bigGenePred table format
- bigPsl table format
- bigMaf table format
- bigChain table format
- bigWig format
- Chain format



- CRAM format
- GenePred table format
- GFF format
- GTF format
- HAL format
- MAF format
- Microarray format
- Net format
- Personal Genome SNP format
- PSL format
- VCF format
- WIG format

me.ucsc.edu/FAQ/FAQformat.html

```
browser position chr7:127471196-127495720
browser hide all
track name="ItemRGBDemo" description="Item RGB                                    itemRgb="On"
chr7    127471196  127472363  Pos1  0  +  127                                    
chr7    127472363  127473530  Pos2  0  +  127472363  127473530  255,0,0
chr7    127473530  127474697  Pos3  0  +  127473530  127474697  255,0,0
chr7    127474697  127475864  Pos4  0  +  127474697  127475864  255,0,0
chr7    127475864  127477031  Neg1  0  -  127475864  127477031  0,0,255
chr7    127477031  127478198  Neg2  0  -  127477031  127478198  0,0,255
chr7    127478198  127479365  Neg3  0  -  127478198  127479365  0,0,255
chr7    127479365  127480532  Pos5  0  +  127479365  127480532  255,0,0
chr7    127480532  127481699  Neg4  0  -  127480532  127481699  0,0,255
```

BED file example

File Formats: VCF

Genomic variant storage standard

- The VCF Variant Call Format is an example for a widely used file format with "built-in logic"
- has been essential to master the "genomics data deluge" through providing "logic compression" for genomic annotations which rely on the notion of "assessed variant in a population"
- very expressive, but complex interpretation
- mix of "observed" and "population" variant concepts confusing for some use cases
- no replacement in sight (but new versions)

The Variant Call Format (VCF) Version 4.2 Specification

25 Jun 2020

The master version of this document can be found at `https://github.com/samtools/hts-specs`.
This printing is version 09fbcec from that repository, last modified on the date shown above.

1 The VCF specification

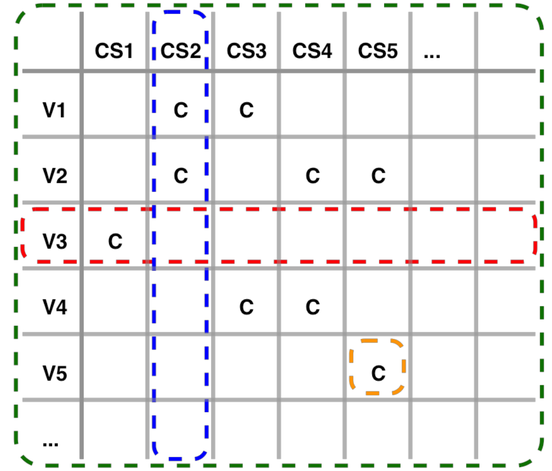
VCF is a text file format (most likely stored in a compressed manner). It contains meta-information lines, a header line, and then data lines each containing information about a position in the genome. The format also has the ability to contain genotype information on samples for each position.

1.1 An example

```
##fileformat=VCFv4.2
##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
```

The VCF file format

Standard for variant representation



Example

VCF header

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##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 (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##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    .   PASS    .                   GT:DP      1/2:13    0/0:29
1       2  rs1     C    T,CT    .   PASS    H2;AA=T             GT:GQ      0|1:100   2/2:70
1       5  .       A    G       .   PASS    .                   GT:GQ      1|0:77    1/1:95
1     100  .       T    <DEL>   .   PASS    SVTYPE=DEL;END=300  GT:GQ:DP   1/1:12:3  0/0:20
```

Mandatory header lines

Optional header lines (meta-data about the annotations in the VCF body)

Reference alleles (GT=0)

Alternate alleles (GT>0 is an index to the ALT column)

Deletion

SNP

Large SV

Insertion

Other event

Phased data (G and C above are on the same chromosome)

Variant
Call
Format

stores the results of a single or multiple interpretations of genome sequencing datasets, in comparison to a reference genome

standard format for file-based storage of human genome variants

Global Alliance for Genomics & Health

Task: Exploration of different file formats

- Which genomic file formats exist & what are their use cases?
    - SAM
    - BAM
    - CRAM
    - VCF
    - FASTA
    - MPEG-G

Task: Estimate Storage Requirements for 1000 Genomes

- How much computer storage is required for 1000 Genomes
  - WES & WGS
  - Different file formats
    - SAM
    - BAM
    - CRAM
    - VCF
    - FASTA
  - Associated costs
    - Cost factors
    - Raw Storage costs

Please provide 1-page size estimates and reasoning for the use of the different file types (i.e. which would you use for storing called variants, which for full archival purposes, browser visualisation), for 3-5 formats.
Submit your files (.md) per pull request to your Github directory.