

BIO392

Bioinformatics of Genome Variations

Genome Variation | Function | Data Formats | Resources | Privacy



BIO392: Course Resources



- Course repository & website on Github
 - links to articles and information resources
 - downloads

<https://compbiozurich.org/UZH-BIO392/>

<https://github.com/compbiozurich/UZH-BIO392/>

UZH BIO392

Bioinformatics of Sequence Variation

[Course Info](#)

[Course Days](#)

[Teachers](#)

[Examples, Guides & FAQ](#)

[Related Sites](#)

[CompbioZurich](#)

[UZH390 lectures](#)

[Baudisgroup at UZH](#)

[Github Projects](#)

[compbiozurich](#)

[progenetix](#)

[Tags](#)

FAQ Jekyll Markdown code
documentation exam feedback
teachers website

UZH BIO392 - Bioinformatics of Sequence Variation

This is a repository for materials related to the BIO392 *Bioinformatics of Sequence Variation* introductory course at the University of Zürich.

Summary

One of the fastest growing areas of bioinformatics is in the analysis, warehousing and representation of genomic and protein sequence variants, particularly with view on the use of molecular data in personalised health and biomedical applications in general. This course will engage participants to explore common data formats, online resources and analysis techniques, with a focus on human genome variation data.

Learning Goals

- Core [Learning Goals](#), relevant for passing the test...

Links

- BIO392 HS 2019 in the [UZH OLAT](#) system
- BIO392 HS 2019 in the [UZH directory](#)

Literature and Resources

- [Literature links](#) and recommendations
- [Resource links](#) (browsers and online repositories)

Schedule

Course feedback pages

Location

- Room info

The map shows the Irchel building complex and the Stockwerk F building. Key rooms labeled include Y11, Y13, Y15, Y17, Y21, Y22, Y23, Y25, Y27, Y31, Y32, Y34, Y36, Y52, and Y55. A green box highlights room Y-01F-50, which contains computer workstations on the mezzanine floor. An inset map shows the location of Y11, Y12, Y10, and Y01 on a larger campus map.

Irchel

Stockwerk F

1 Zugang zu Y-01F-50
Computerarbeitsplätze
Zwischengeschoss

9 Hörsaal Y22-F-68

10 Seminarraum/Sitzungs-
zimmer Y35-F-08A

University of
Zurich

[Display a menu](#)

Genomes Everywhere

Organization / Initiative: Name	Organization / Initiative: Category	Cohort
100K Wellness Project	Research Project	107 unaffected individuals (scaling up to 100,000)
23andMe	Organization	>1 million customers (>80% consented to research)
Actionable Cancer Genome Initiative (ACGI)	Data-Sharing Project	Goal: 100,000 individuals
Ancestry.com	Organization	1.4 million customer DNA samples (what % consented to research?)
BioBank Japan	Repository	Specimens from >200,000 patients and unaffected controls
Cancer Moonshot2020	Consortium	Phase 1: 20,000 cancer patients
Children's Hospital of Philadelphia Biorepository	Repository	Capacity for 8.6 million samples
China Kadoorie Biobank	Repository	>512,000 participants (general population, China). Genotyping data available for ~100,000.
CIMBA	Consortium	>15,000 BRCA1 carriers, >8,000 BRCA2 carriers
Clinical Sequencing Exploratory Research (CSER)	Consortium	~4,000 patients and healthy controls
DECIPHER	Repository	19,014 patients (international)
deCode Genetics	Organization	500,000 participants (international)
East London Genes & Health	Research Project	100,000 unaffected individuals (East London, Pakistani or Bangladeshi heritage)
Electronic Medical Records and Genomics (eMERGE) Network	Repository, Consortium, Research Project	55,028 patients
European Network for Genetic and Genomic Epidemiology (ENGAGE)	Research Project	80,000 GWAS scans, and DNA and serum/plasma from >600,000 individuals
Exome Aggregation Consortium (ExAC)	Consortium	60,706 individuals
GENIE/AACR	Data-Sharing Project	>17,000 cancer patients (international)
Genome Asia 100K	Consortium	Goal: 100,000 individuals (Asia)
Genomics England	Organization	Goal: 100,000 genomes from 70,000 individuals (rare disease & cancer patients, and their relatives)
GoT2D	Consortium, Data-Sharing Project	Multiple case-control cohorts
International Cancer Genome Consortium (ICGC)	Consortium	currently data from >16'000 samples
International Genomics of Alzheimer's Project (IGAP)	Consortium	40,000 patients with Alzheimer's disease
International Multiple Sclerosis Genetics (IMSG) Consortium	Consortium	Goal: >50,000 patients with MS
Kaiser Permanente: Genes, Environment, and Health (RPGEH)	Repository, Research Project	200,000 DNA samples (scaling up to 500,000)
Leiden Open Variation Database (LOVD)	Repository	>170,000 individuals
Million Veteran Program	Research Project	Goal: 1 million individuals; first 200,000 is complete.
MyCode® Community Health Initiative	Repository, Research Project	Goal: >250,000 patients
Precision Medicine Initiative	Research Project	Goal: >1 million participants, starting in 2016 (US)
Psychiatric Genomics Consortium (PGC)	Consortium	>170,000 subjects
Resilience Project	Research Project	589,306 individuals
Saudi Human Genome Program	Research Project	Goal: ~100,000 patients and controls (Saudi Arabia)
Scottish Genomes Partnership (SGP)	Research Project	>3,000 individuals (Scotland)
T2D-GENES	Consortium, Data-Sharing Project	10,000 patients and controls (five ethnicities); 600 individuals (Mexican American)
TBResist	Consortium	>2,600 samples
UK Biobank	Repository, Consortium, Research Project	500,000 individuals (age 40-69 years; UK)
UK10K	Research Project	10,000 participants (6,000 patients and 4,000 controls)
Vanderbilt's BioVU	Repository	>215,000 samples

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}
 - $2 * 3 * 10^9 b = 6,000,000,000 b$
 - perfect genome (no overhead): ~715 MB
 - 1PB => ~1'400'000 genomes
- according to Swiss online store (Sep 2019) ~35'000CHF (65x16TB disks)
- this is less than a PhD position per year in Switzerland ...
- (real costs are 2x that, + duplication, facilities, service ... => ~500'000CHF)
- **However: A single 30x BAM file => 100GB**
- Still: 500'000CHF => 1PB => 10'000 genomes => 50CHF/genome (BAM format)



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://cndl.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 (NCBI)			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

CANCER VARIANT KNOWLEDGE BASES

- ▶ cancer variant knowledge databases report evidence 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

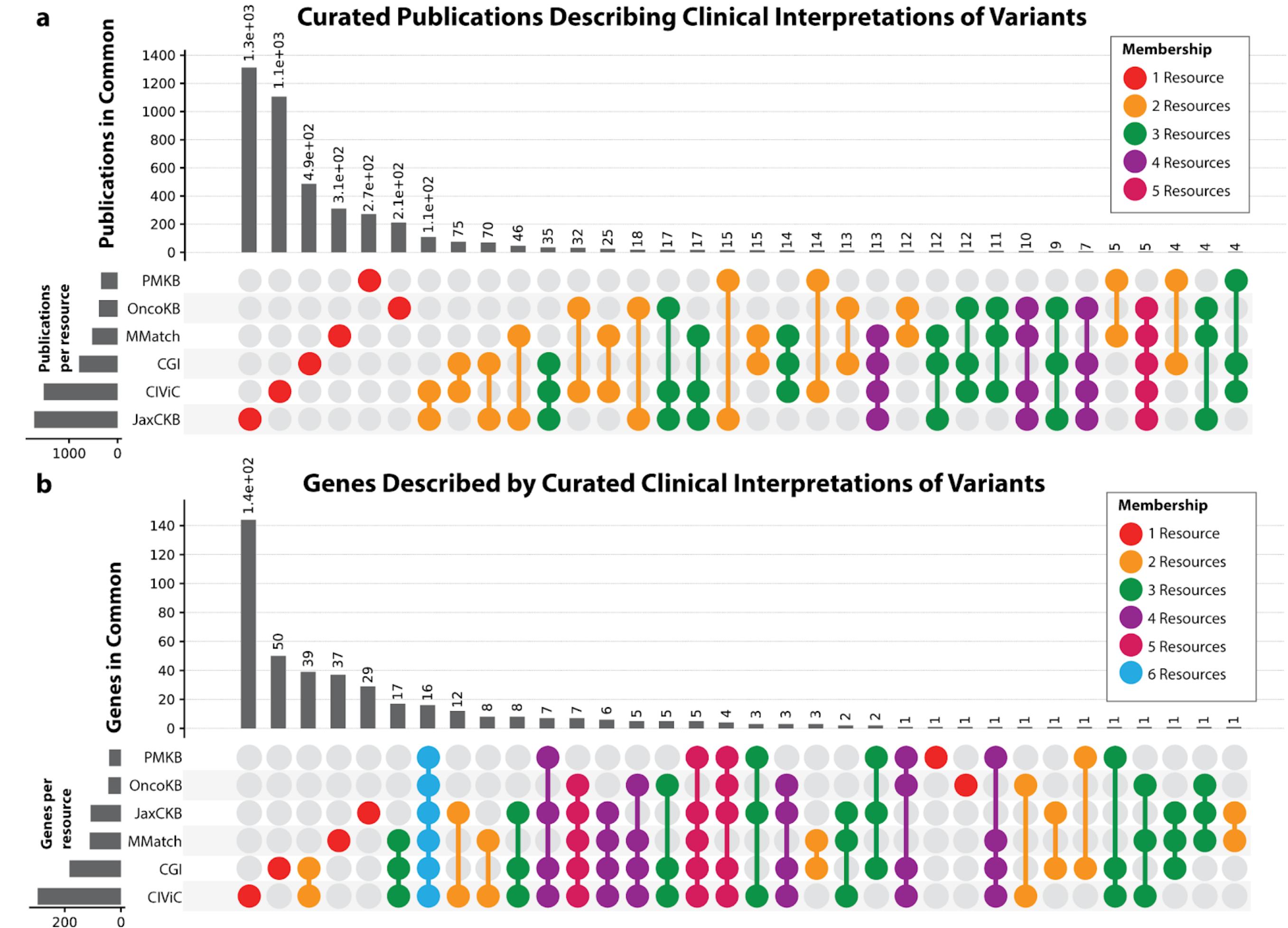


Figure S2 - Knowledgebase overlap

(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

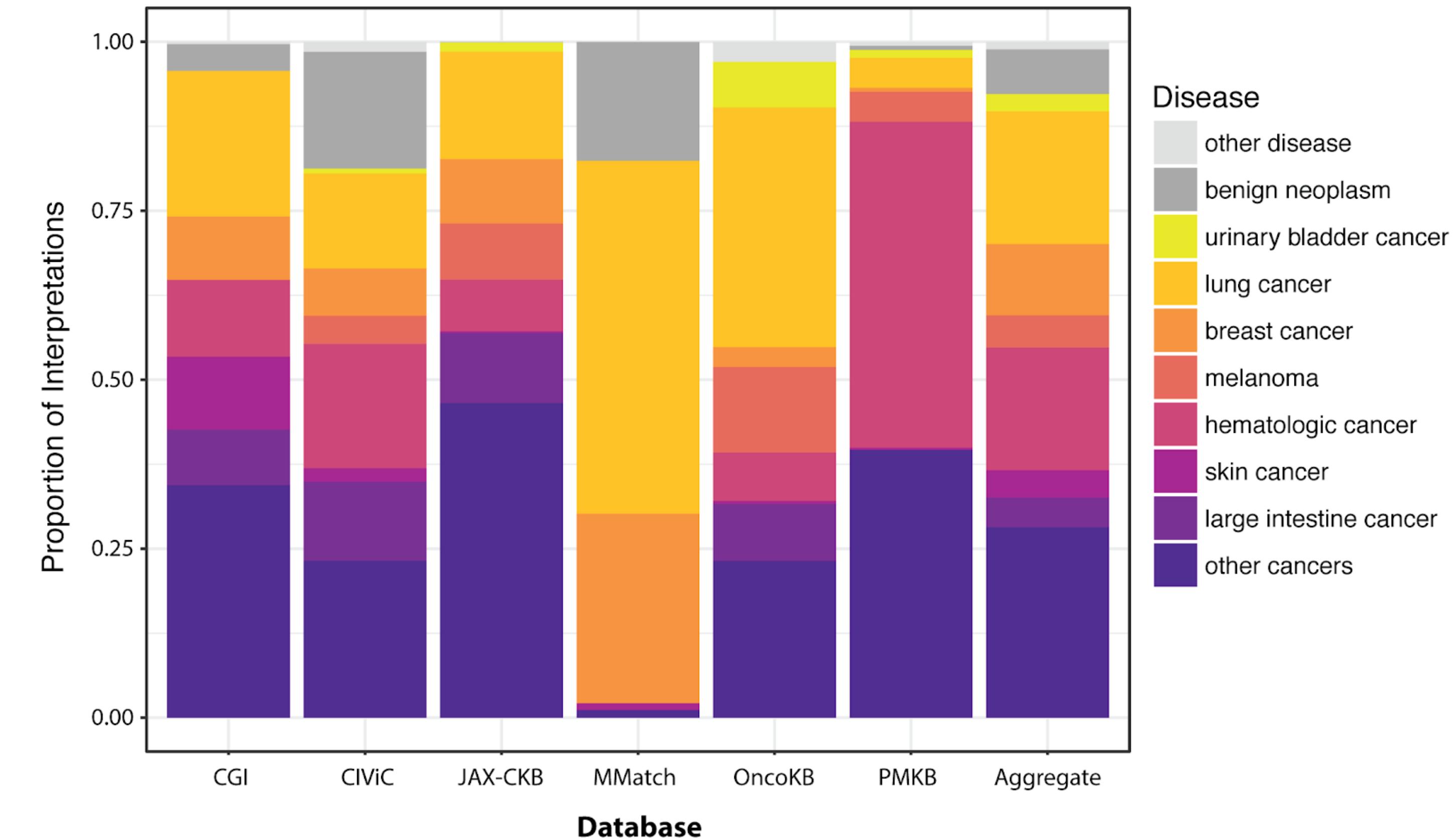


Figure S3 - Knowledgebase disease enrichment

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."

The screenshot shows the CIViC website interface. At the top, there is a navigation bar with links for About, Participate, Community, Help, FAQ, and Sign In/Sign Up. Below the navigation bar, the main content area has tabs for Go to Genes & Variants, BROWSE, SEARCH, ACTIVITY, and ADD. The current tab is 'GENE BRAF'. On the left, there is a sidebar with a pencil icon and a 'Gene Summary' button. The main content area contains a detailed description of BRAF mutations, sources (Li et al., 2009; Oncol. Rep. and Pakneshan et al., 2013; Pathology), and a large blue box containing detailed information about the BRAF gene, including its name, Entrez symbol, aliases, chromosome location, protein domains, and pathways. A 'View MyGene.info Details' button is located at the bottom right of this box. At the bottom of the page, there is a section titled 'BRAF Variants & Variant Groups' with a grid of variant categories.

BRAF mutations are found to be recurrent in many cancer types. Of these, the mutation of valine 600 to glutamic acid (V600E) is the most prevalent. V600E has been determined to be an activating mutation, and cells that harbor it, along with other V600 mutations are sensitive to the BRAF inhibitor dabrafenib. It is also common to use MEK inhibition as a substitute for BRAF inhibitors, and the MEK inhibitor trametinib has seen some success in BRAF mutant melanomas. BRAF mutations have also been correlated with poor prognosis in many cancer types, although there is at least one study that questions this conclusion in papillary thyroid cancer.

Sources:

Li et al., 2009, Oncol. Rep.
Pakneshan et al., 2013, Pathology

Name: B-Raf proto-oncogene, serine/threonine kinase
Entrez Symbol: BRAF Entrez ID: 673
Aliases: B-RAF1, B-raf, BRAF1, NS7, RAFB1
Chromosome: 7 Start: 140419127 End: 140624564 Strand: -1 (GRCh37)
Protein Domains: Diacylglycerol/phorbol-ester binding, Protein kinase C-like, phorbol ester/diacylglycerol-binding domain, Protein kinase domain, Protein kinase, ATP binding site, Protein kinase-like domain... (5 more)
Pathways: Intracellular Signalling Through Adenosine Receptor A2a and Adenosine, Intracellular Signalling Through Adenosine Receptor A2b and Adenosine, EGFR1, MAPK signaling pathway - Homo sapiens (human), ErbB signaling pathway - Homo sapiens (human)... (139 more)

View MyGene.info Details

BRAF Variants & Variant Groups

Show all: filter variants...

AMPLIFICATION

599INST AGK-BRAF AKAP9-BRAF BRAF-CUL1 CUX1-BRAF D594A D594G D594K D594N D594V DEL 485-490 DELNVTAP F595L G464V G466A G466V G469A G469E G469R G469V G496A G596C G596R G596V G606E intron 10 rearrangement intron 9 rearrangement K483M K601E KIAA1549-BRAF L505H L597Q L597R L597S L597V MACF1-BRAF Fusion MUTATION N581S Non-V600 P731T PAPSS1-BRAF PPFIBP2-BRAF TRIM24-BRAF V600 V600_K601DELNSD V600D V600E V600E AMPLIFICATION V600E+V600M V600E/K AMPLIFICATION V600K V600R WASFL-BRAF Fusion WILD TYPE ZKSCAN1-BRAF

Variant Annotation Formats

GENOME DATA FORMATS: HGVS

- ▶ 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](#)

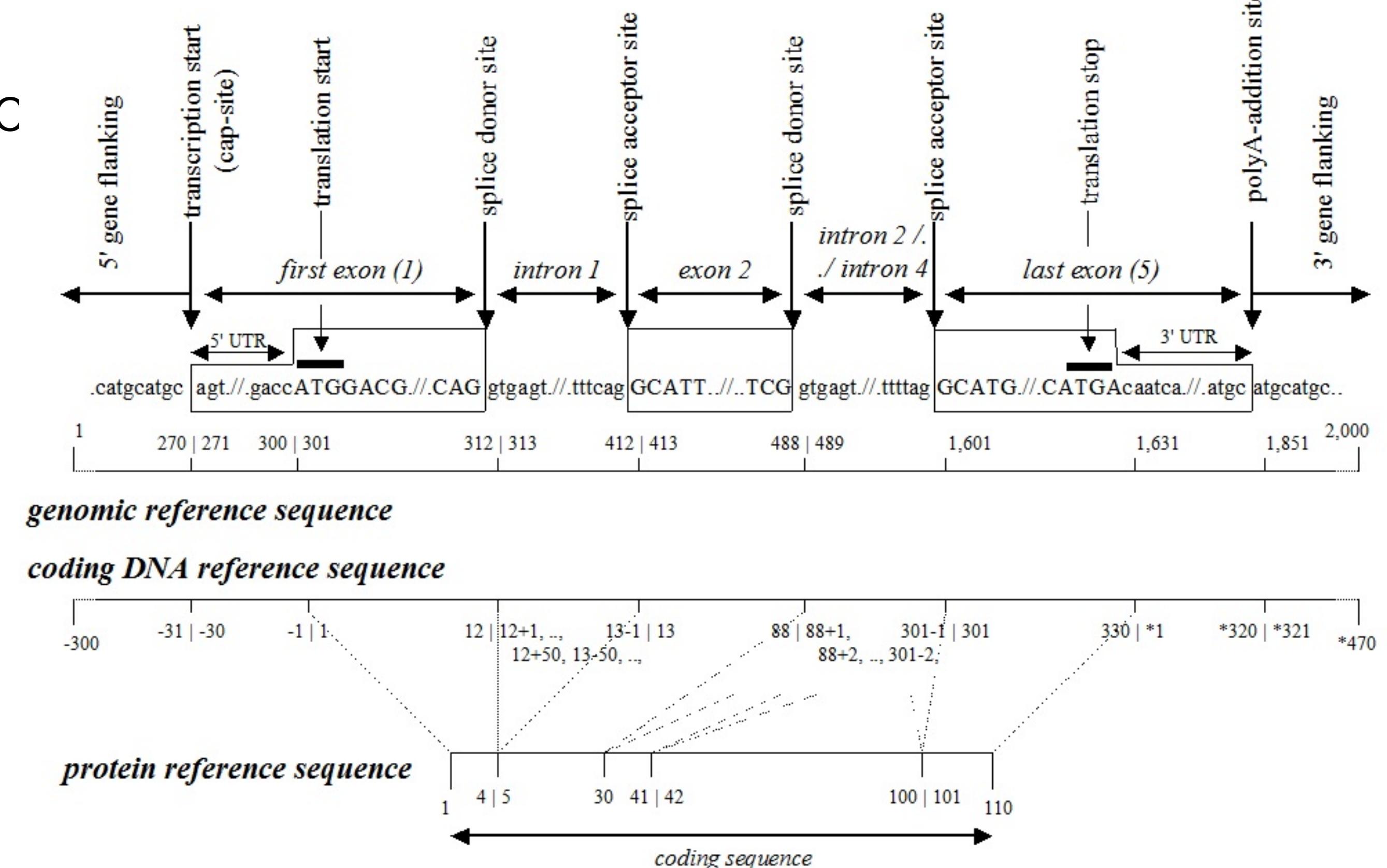
- NC_000008.10:g.19813529A>G
- NG_008855.1:g.21948A>G
- NM_000237.2:c.953A>G
- NP_000228.1:p.Asn318Ser

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

GENOME DATA FORMATS: FASTA

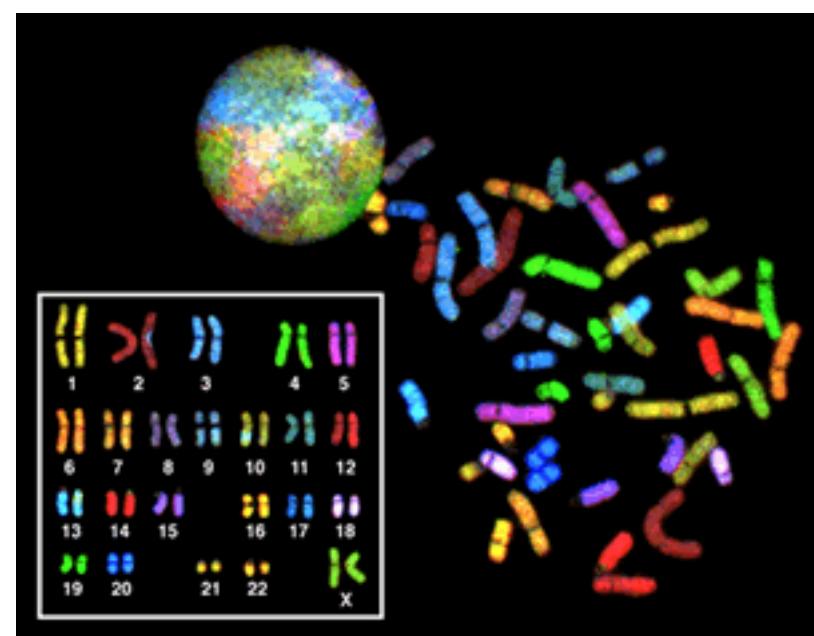
- ▶ Linear annotation of single-letter **nucleotides** or amino acid codes
- ▶ leading information line, usually with unique SeqID
- ▶ text format
 - ▶ "readable"
 - ▶ not optimised for size
- ▶ representation of a sequence without ambiguities or QC data
- ▶ extended as "FASTQ" (Sanger Centre)

```
>NC_000007.14:11369935-11832697 Homo sapiens chromosome 7, GRCh38.p12 Primary Assembly
AGGGCTTAAATGGTCCCTACTTACATTAGCAAATAGCTATTCAGAAAATGTTTAAGTGCAA
ACTACCCCGGAAGTAACCTGTCTTAAGTTGTGTCCCTCCTGAATTGTTAAGGCATAAGTTCTGCT
TTGACTTTAGGTTGGTTGGGTAGACACAGGGACAAGAGACAGTGAGGGATGTGCCTTATGAC
TGATTGGGTGGGAAAAGCTGTACTCTGTTAGAGAGTTCCCCATCCTCTGCTGCTGCCATTGAAAT
TGACTGGAAACCAGGAGGTCCCTGTCCATGATTCACCTGGTGGCCTAGCCAACTTCAAAGTAAAAGT
TTGCATTTCTTGAACCTTCTAAATTGGAGTTGTTATACAACCCAGGAAAGGGCAATACAGTAGGTAAAAG
GATTTAGGTATTCACTGGAAAAAAATTAAATCCATATTAAAGAAGCAATTGGTCAAATCAAACACAG
ATACACATGATTAGAATGAAAATGATTCCGTATTATGTTGTCAGCAATATAGTTATTACAATAAC
CCATATGAAAATGTAAAAAAGCATATTACATCTTCACATGCCATCTGTATTGACTGAATAAGCTTAGTG
ACATTATTGCAAATCTGTAGTTAATTGTACATAGACATTGTGCGTAAAAAGGAAATGTACATAATG
TAAAATAAATTACATTACGCAATTACAAAGTAATATTAACAAAATTCTTAGACAGCTGCCCTTATT
TAAACAAAATAAATTACAGGTAGTTAAATTAAACATAAAACACATTAGGAATAATAATTAGAA
AGACAGATTGCAAATTAAAGTTATATTACAATGATAGATACTGATCTCTCAAATCTGTGTA
TAGAAATGGGAGAAAAAAAGTACCAAGAAAAGGAATCTAAATGTTACTTCTAAAATAACACAAACAGA
TTCTGAAAAATAGGGAAAAGTTACTGAGGGTAAAGTAGGTAATCTAGAAACTATGGCTAAAAC
AATAAATCTACAAAACACAAGACTGACAATTATATTCTAAATAATAGAGATTGATCACTGAA
AACATGACTCCCACAAACTAAAGCTTCTCATATACTGCCCTTAAAGATCTGACTTGGTAGAACACA
GAAAATAAAATGCAAATTAACTGTTAGCATTAGTTCTTAAATGAGACATAACCATT
TTTCATTGTCCCTGCTAGATATAAAATTATAACACACTGCAAACACCATTCTTTATAAATGGATAAC
TATTTGCTGGCTCACACACCAGTTCTGATACCTGAAATCCTGCTGCAGCCAGGGCACCTGAGGGC
AGGACCTGGGAGACCCTTATTCCAGAACAGCAGATGTAGTTCTCACAACTAAACTAGTCCCAGGAAA
GATCACATTCTGACAAGATTCTCACAGATTGCTCAAGGACTACTGTTTTCAACACCCTCAATTAA
CAGTGGAAATAGAAGAAGAACCCACACTTGAATTGGAATATATTATAAACAGGGAGATCCCAGATCAT
TTGGGAATTGTGCTTCTCATGTACTATTGAGACCCACGTCAGCTTAGAACAGGCTCTCCCTCTGTATG
GTACTGAAAGTACAGTCCTCCCTCACTGTCTGTGGATGAAACAAAGACTCGAGATGGAGGC
AGGAGGATATGGGATGGTCTAAAGCAAGTGTAGGCATGGACATTTCAGAGAAAGGGCTTTTTTTT
TTTTTTCTGCATGCCTCCACATTTCCTTATTCAATTCTTGTGACCAGTGGATTGGT
...
```

Homo sapiens chromosome 7, GRCh38.p12 Primary Assembly
NCBI Reference Sequence: NC_000007.14

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	<i>de novo</i> (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	<i>pseudo dicentric</i> - 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

GENOMIC VARIANT FORMATS: DBVAR

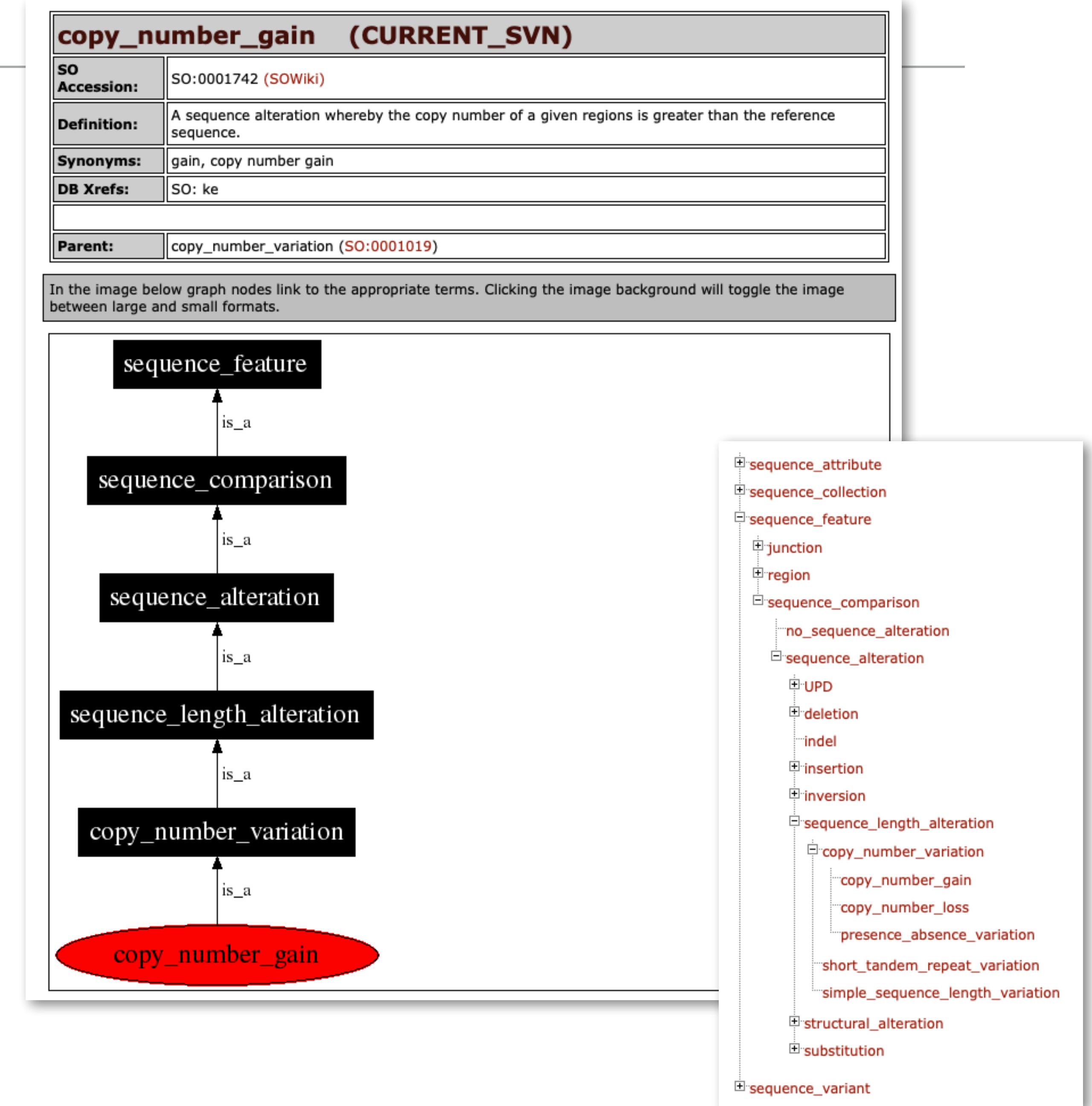
- ▶ dbVar is "NCBI's database of human genomic structural variation – insertions, deletions, duplications, inversions, mobile elements, and translocations"
- ▶ structural genome variations are still not completely solved with respect to unambiguous annotation

[ncbi.nlm.nih.gov/dbvar/content/
overview/](https://www.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

GENOMIC VARIANT FORMATS: SO

- ▶ 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)



GENOME DATA FORMATS: HGVS

- ▶ 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](#)

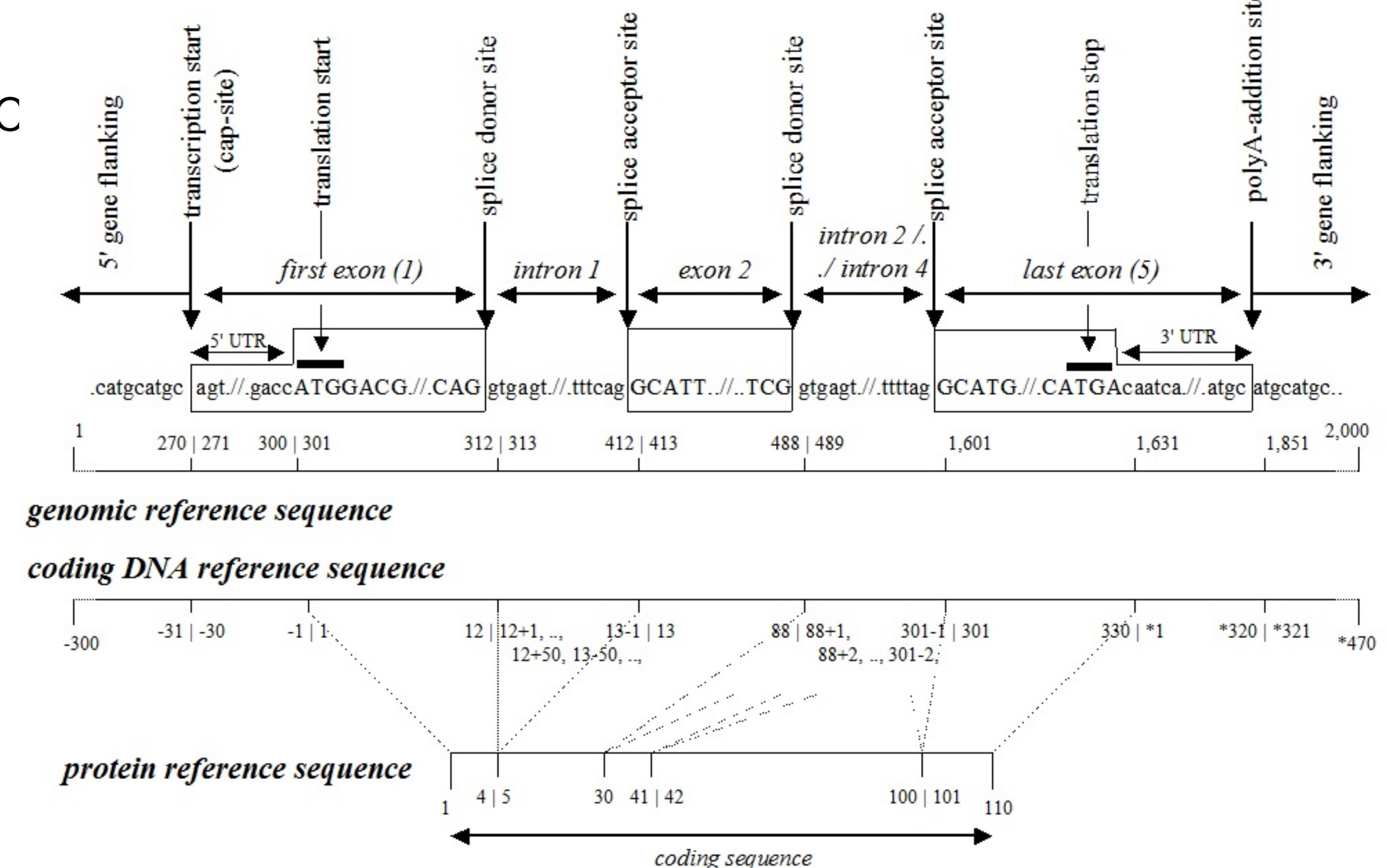
- NC_000008.10:g.19813529A>G
- NG_008855.1:g.21948A>G
- NM_000237.2:c.953A>G
- NP_000228.1:p.Asn318Ser

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

► "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."

Sequence Variant Nomenclature

What is the sequence variant nomenclature?

These pages summarise HGVS-nomenclature: the recommendations for the description of sequence variants. HGVS-nomenclature is used to report and exchange information regarding variants found in DNA, RNA and protein sequences and serves as an international standard. When using the recommendations please cite: [HGVS recommendations for the description of sequence variants - 2016 update, Den Dunnen et al. 2016, Hum.Mutat. 37:564-569](#). HGVS-nomenclature is authorised by the Human Genome Variation Society (HGVS), the Human Variome Project (HVP) and the HUMAN Genome Organization (HUGO).

... .

Current Recommendations

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

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!)

"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)  
    ]  
}
```



VARIANT ANNOTATION FORMAT: THE "GA4GH TRANSITIONAL" MODEL

- ▶ object model for the representation of genomic variants in the context of an experimental read-out ("callset")
- ▶ based on VCF principles (e.g. notation for structural variants such as "DUP", "DEL" ...)
- ▶ allows to place variants into intervals with "fuzzy ends"
- ▶ not yet suitable for genotype reconstruction (e.g. connecting translocations, other out-of-place events)

*[https://schemablocks.org/
schemas/blocks/Variant.html](https://schemablocks.org/schemas/blocks/Variant.html)*

```
{  
  "biosample_id" : "structdb-bs-nhl-0009876",  
  "callset_id" : "structdb-CS-nhl-0009876",  
  "created" : "2019-01-22T03:06:45Z",  
  "digest" : "6:63450000,63550000-63450000,63550000:DEL",  
  "end" : [  
    "63450000",  
    "63550000"  
>,  
  "id" : "structdb-var-123456790",  
  "info" : {  
    "cnv_length" : 85500000,  
    "cnv_value" : "-0.294"  
>},  
  "reference_bases" : "N",  
  "reference_name" : 6,  
  "start" : [  
    "63450000",  
    "63550000"  
>],  
  "updated" : "2019-02-01T12:40:21Z",  
  "variant_type" : "DEL"  
}
```

```
{  
  "alternate_bases" : "AC",  
  "callset_id" : "DIPG_CS_0290",  
  "created" : "2018-11-06T11:46:30.028Z",  
  "digest" : "2:203420136:A>AC",  
  "genotype" : [  
    "1",  
    ".  
>],  
  "id" : "5be1840772798347f0ed9e8b",  
  "reference_bases" : "A",  
  "reference_name" : 2,  
  "start" : [  
    "203420136"  
>],  
  "updated" : "2018-11-06T11:46:30.028Z"  
}
```

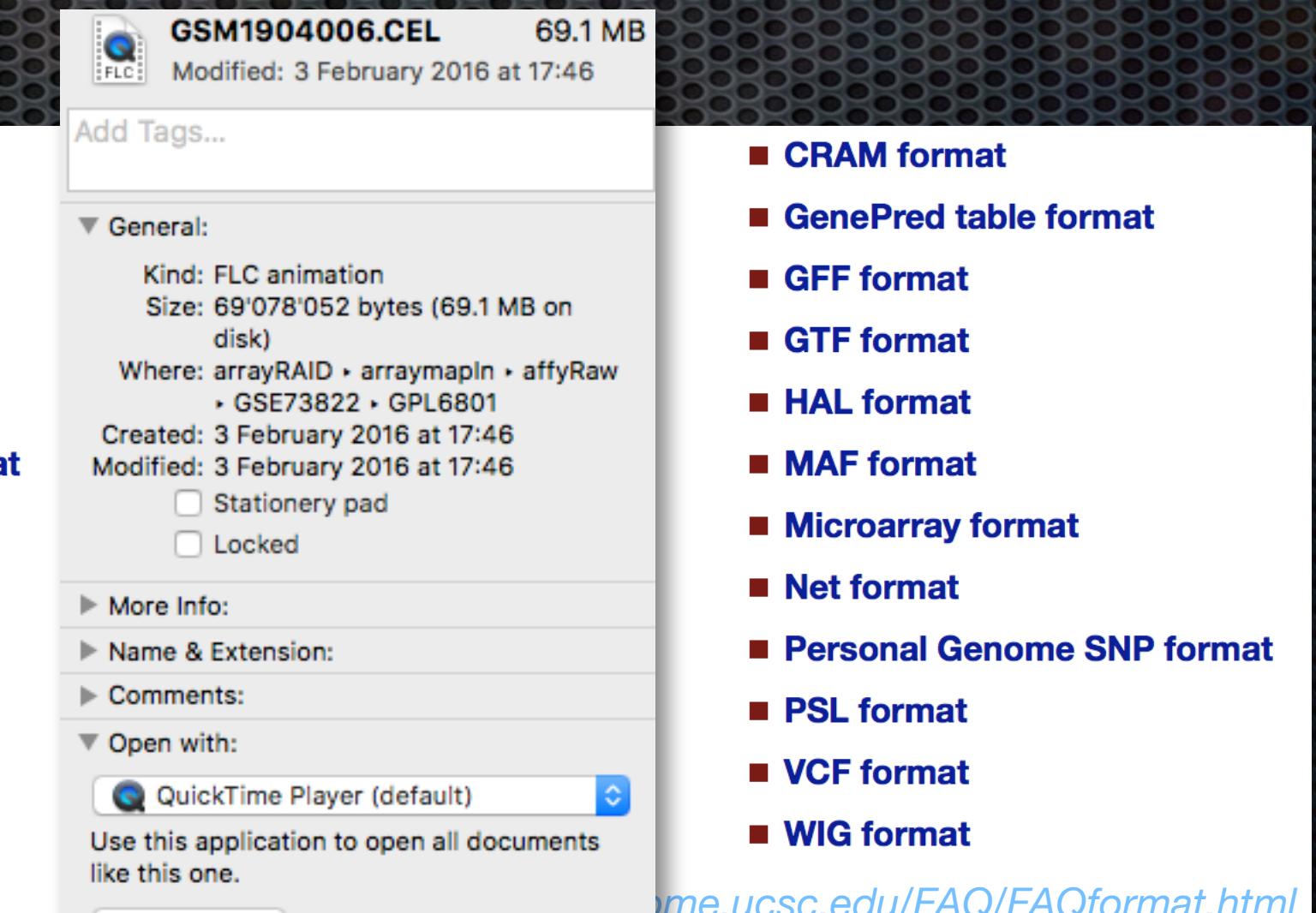
Bioinformatics: File Formats

- **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](#)

```
browser position chr7:127471196-127495720
browser hide all
track name="ItemRGBDemo" description="Item RGB Demo"
chr7 127471196 127472363 Pos1 0 + 127472363
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



The VCF file format

Standard for variant representation

Example

VCF header

```

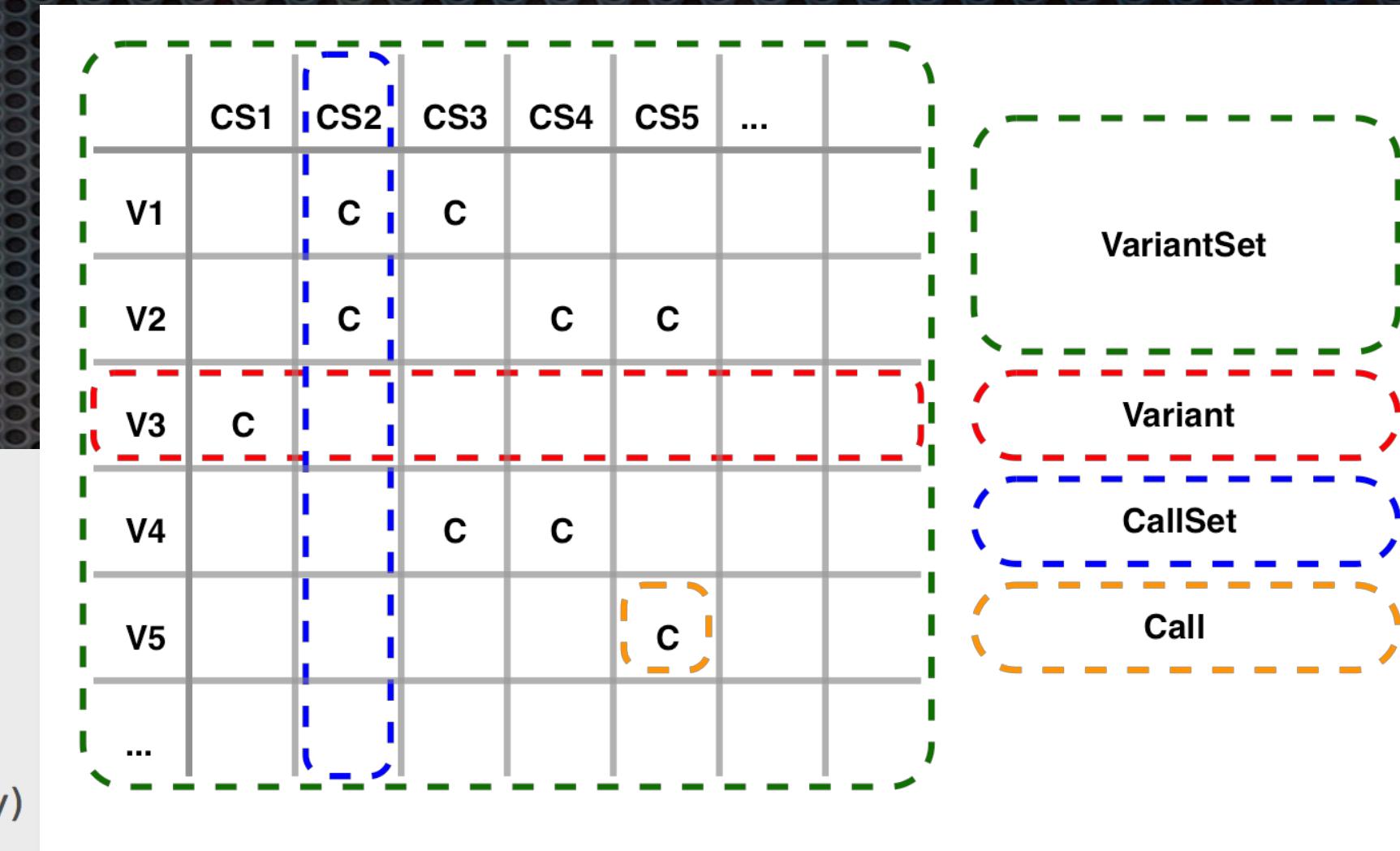
##fileformat=VCFv4.0 ← Mandatory header lines
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<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">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2
1 1 . ACG A,AT . PASS . .
1 2 rs1 C T,CT . PASS H2 ; AA=T 0|1:100 2/2:70
1 5 . G <DEL> . PASS .
1 100 . T . PASS SVTYPE=DEL ; END=300 GT:DP 1/2:13 0/0:29

```

Body

Annotations:

- Deletion**: Row 1, Column 5
- SNP**: Row 2, Column 5
- Large SV**: Row 5, Column 5
- Insertion**: Row 5, Column 5
- Other event**: Row 5, Column 5
- Reference alleles (GT=0)**: SAMPLE1 and SAMPLE2 columns for rows 1, 2, 3, 4
- Alternate alleles (GT>0 is an index to the ALT column)**: SAMPLE1 and SAMPLE2 columns for row 5
- 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



Task: Reading up on Variant Formats

- VCF
 - precise variants
 - structural variants
 - types
 - how does one describe imprecise positional knowledge (start, end)
- HGVS
- cytogenetic annotation basics
 - Estimated resolution of cytogenetic banding?

Task: Estimate Storage Requirements for 1000 Genomes Project

- How much computer storage is required for the 1000 Genomes project
 - WES & WGS
 - Different file formats
 - SAM
 - BAM
 - CRAM
 - VCF
 - FASTA
 - BED
 - 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.



University of
Zurich UZH



Prof. Dr. Michael Baudis
Institute of Molecular Life Sciences
University of Zurich
SIB | Swiss Institute of Bioinformatics
Winterthurerstrasse 190
CH-8057 Zurich
Switzerland

arraymap.org

progenetix.org

info.baudisgroup.org

sib.swiss/baudis-michael

imls.uzh.ch/en/research/baudis



Global Alliance
for Genomics & Health

