



University of  
Zurich<sup>UZH</sup>

# BIO392 Bioinformatics of Genome Variations

Genomes: Core of "Personalized Health" & "Precision Medicine"

Michael Baudis **UZH SIB**  
Computational Oncogenomics

# BIO392: Course Schedule

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

	Friday 2025-04-04	Tuesday 2025-04-08	Wednesday 2025-04-09	Thursday 2025-04-10	Friday 2025-04-11	Tuesday 2025-04-15	Wednesday 2025-04-16	Thursday 2025-04-17	Tuesday 2025-04-29	Wednesday 2025-04-30	Friday 2025-05-02	Tuesday 2025-05-06	Wednesday 2025-05-07
09:00 - 10:00	* Room information * Administrative - discuss times/days - exam		Jiahui: Terminal / Unix / Files	Hangjia: R environment introduction	Michael: Genomic Resources & Data Sharing		Feifei: Sequence analysis practical. FastQC, trimmomatic, BWA-MEM2, SAMtools, GangSTR, BCFtools	Michael: Genomic Data & Privacy		Feifei & Ziying: survival	Feifei: population structure		
10:00 - 11:00	Tina Siegenthaler: technical introduction (room, computer, accounts)		Jiahui: Terminal / Unix / Files	Hangjia: R exercise	Michael: Genomic Resources & Data Sharing		Feifei:: Sequence analysis practical	Michael: Genomic Data & Privacy		Feifei & Ziying: survival	Feifei: population structure		Discussion
11:00 - 12:00	* explore course site * create Github accounts and forward to bio392@compbiozurich.org *feifei&jiahui: overall schedule of the course		jiahui: SIB online introduction to Unix		Michael: Genomic Resources & Data Sharing		Feifei:: Sequence analysis practical	Michael: Genomic Data & Privacy		Feifei: survival	Feifei: population structure		
13:00 - 14:00	jiahui: Github	Michael: Introduction	Jiahui & Ziying: Python	Hangjia: CNV paper reading	Hangjia: Clinvar and Clingen	Feifei: Sequence analysis introduction. Overview of pipeline from raw reads -> variant calling & interpretation	Feifei:: STR reading up			Feifei:: analysis & interpretation. Parsing VCF (cvcf2), UCSD genome browser, ENSEMBL variant effect predictor	Feifei: population structure	Feifei & others: Presentation & Discussion	Exam revision, Q&A
14:00 - 15:00	jiahui: Github&Git exercise	Michael: Introduction	Exercise	Hangjia: Progenetix as tool for CNV frequencies etc.	Hangjia: blast	Feifei: Sequence analysis introduction	Feifei:: STR reading up			Feifei:: analysis & interpretation.	Feifei: population structure	Feifei & others: Presentation & Discussion	
15:00 - 16:30		Michael: Introduction			Hangjia: Blast exercise					Feifei:: analysis & interpretation.	Feifei: population structure	Feifei & others: Presentation & Discussion	

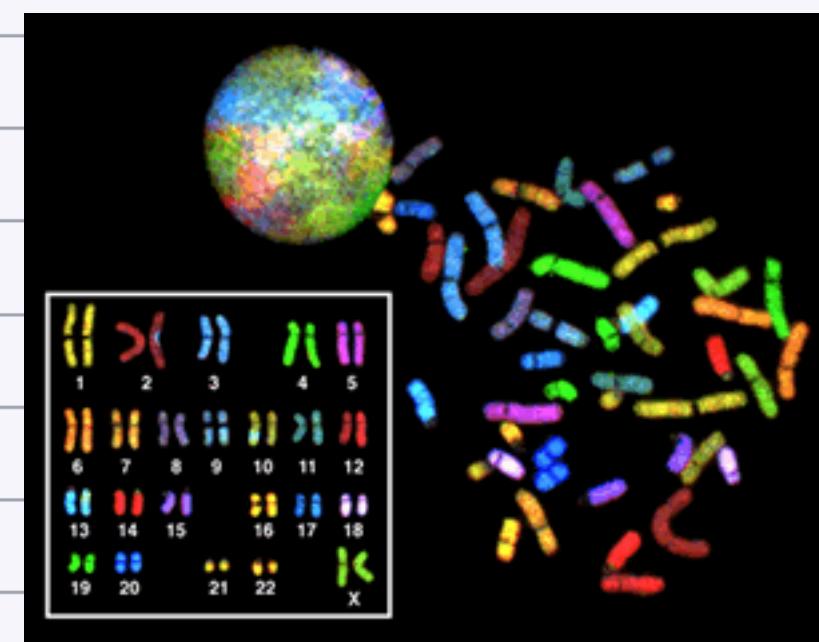
<https://drive.switch.ch/index.php/s/PB1czLjrjAKR6Q2>

**Let's Have Some Standards!**

## 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)
- ▶ Example Burkitt Lymphoma Karyotype...
  - **39-52,XY,+X,-1,t(2;17)(q13;q21), +3,+4,+5,+6,+7,-8,+8,t(8;22) (q24;q11),+dup(12) (q13q24),-13,-13,+14,+der(14)t(14;18)(q32;q21),t(14;18)(q32;q21), +16,+18,+19,+20,+21,+22,-22,+mar**

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. 2pter refers to the end of the long arm of chromosome 2)
tri	Trisomy
trp	Triplication of a portion of a chromosome



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

# Genome Coordinates

## More than one way to get them wrong

### 0-start, half-open genomic coordinate system

#### Definition

Two integers that define the start and end positions of a range of residues, possibly with length zero, and specified using "0-start, half-open" coordinates.

The following also applies to coordinates:

- Coordinates start at 0 and finish at the length of the sequence
- Start must be greater than 0
- End must be greater than the start
- The length of an interval is (end - start)
- The reverse start is (sequence length - end)
- The reverse end is (sequence length - (start-1))
- A zero-length interval (start == end) is a point between two residues
- An interval of length 1 is a residue position
- Two intervals are equal if their start and end are equal
- Two intervals intersect if start or end occurs between the start and end of the other
- Two intervals coincide if they intersect or if they are equal

Product	"0-start, half-open"	"1-start, fully-closed"	Interbase
BAM/CRAM	X		
SAM		X	
VCF		X	
BCF	X		
htsget		X	
refget	X		
Beacon		X	
VMC			X

#### GA4GH Recommendation

- We recommends the use of "0-start, half-open" (interbase) coordinate system in all systems
- This is not a retrospective recommendation for existing standards and products
- "1-start, fully-closed" should be used when displaying coordinates through a GUI or report

#### How '0-start, half-open' works

G A G T G C  
G G T G G A G T G C G C C G C C A T G G  
1 1 1 1 1 1 1 1 1 1 2  
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0

"0-start, half-open" breaks down into two integer positions. The first, "0-start", refers to the start coordinate and uses an indexing scheme starting at 0 to refer to bases within a sequence, similar to array indexes in most C based programming languages. The second, "half-open", refers to the end coordinate and is one higher than the start (effectively using an indexing system starting at 1).

This scheme makes sub-sequences very easy to define. In the above example we have highlighted the subsequence GAGTGC, which starts at position 4 and ends at position 10. Calculating the length of this subsequence is easily done by subtracting start from end e.g. (10-4) = 6. Other transformations are less prone to programming errors than the alternative system "1-start, fully-closed".

This same coordinate system can be used to flag insertions and deletions as a start and an end which equal each other refers to a space between two residues e.g. 4,4 would flag an event occurring between GGTG and GAGTGC .

# Dates and Times

## There is only the ISO way!

The specification of a time point is given through the concatenation of

- a date in YYYY-MM-DD
- the designator "T" indicating a following time description
- the time of day in HH:MM:SS.SSS form, where "SSS" represents a decimal fraction of a second
- a time zone offset in relation to UTC

Examples

- year (YYYY)
  - 2015
  - Time points with year granularity are both common for obfuscated personal data as well as technical metadata (e.g. year of publication of an analysis).
- date (e.g. date of birth) in YYYY-MM-DD
  - 2015-02-10
  - This represents the standard way of representing a specific day, e.g. a date of birth.
- time stamp in milliseconds in YYYY-MM-DDTHH:MM:SS.SSS
  - 2015-02-10T00:03:42.123Z
  - Timepoints with millisecond granularity are typical use cases for timing computer generated entries, e.g. the time of a record's update ("updateTime").

- Y = year
- M = month
- D = day
- H = hour
- M = minute
- S = second
- .S = decimal fraction of a second

### PUBLIC SERVICE ANNOUNCEMENT:

OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.

THIS IS THE CORRECT WAY TO WRITE NUMERIC DATES:

2013-02-27

THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:

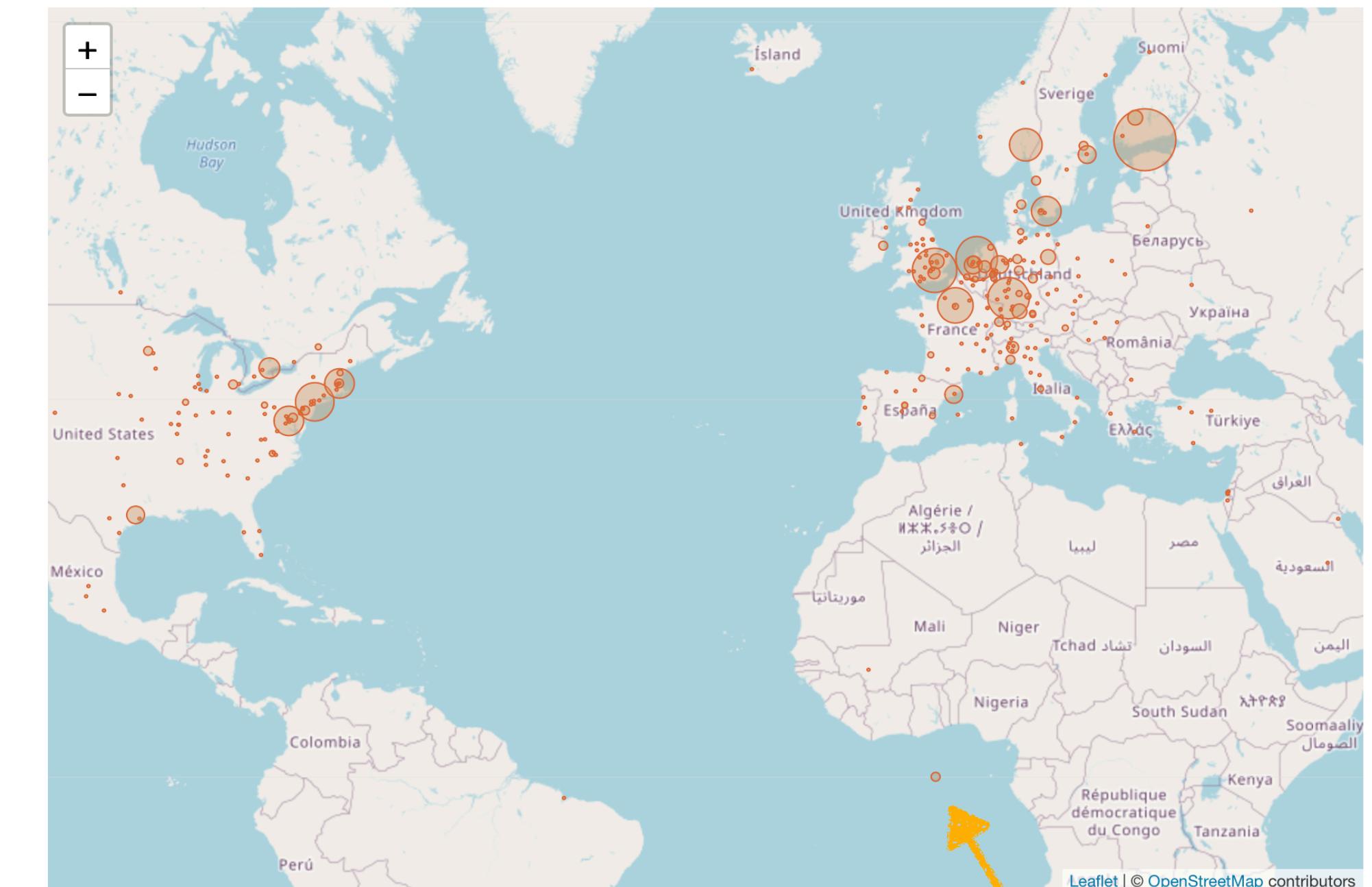
02/27/2013 02/27/13 27/02/2013 27/02/13  
20130227 2013.02.27 27.02.13 27-02-13  
27.2.13 2013. II. 27. 27/2-13 2013.158904109  
MMXIII-II-XXVII MMXIII  $\frac{LVII}{CCCLXV}$  1330300800  
 $((3+3)\times(111+1)-1)\times3/3-1/3^3$  2013   
10/11011/1101 02/27/20/13  $\frac{2}{5} \frac{3}{6} \frac{1}{7} \frac{4}{8}$

<https://xkcd.com/1179/>

# Data Curation

Provide "clean and correct data" - but final verification of data from external resources lies with the user ...

- correct data is important for any type of scientific analysis
  - errors in formats and values can occur during all steps between data acquisition and analysis (numerous "Excelgates"!)
  - "meta"-resources and analyses are prone to erroneous data due to varying input formats and lack of source control
- ➡ always look for batch effects and outliers!



Geographic distribution (by corresponding author) of the 118554 genomic array, 36766 chromosomal CGH and 42105 whole genome/exome based cancer genome datasets from the 3306 listed publications. Area sizes correspond to the sample numbers reported from a given location.

# Data Curation - Geolocations

Provide "clean and correct data" - but final verification of data from external resources lies with the user ...

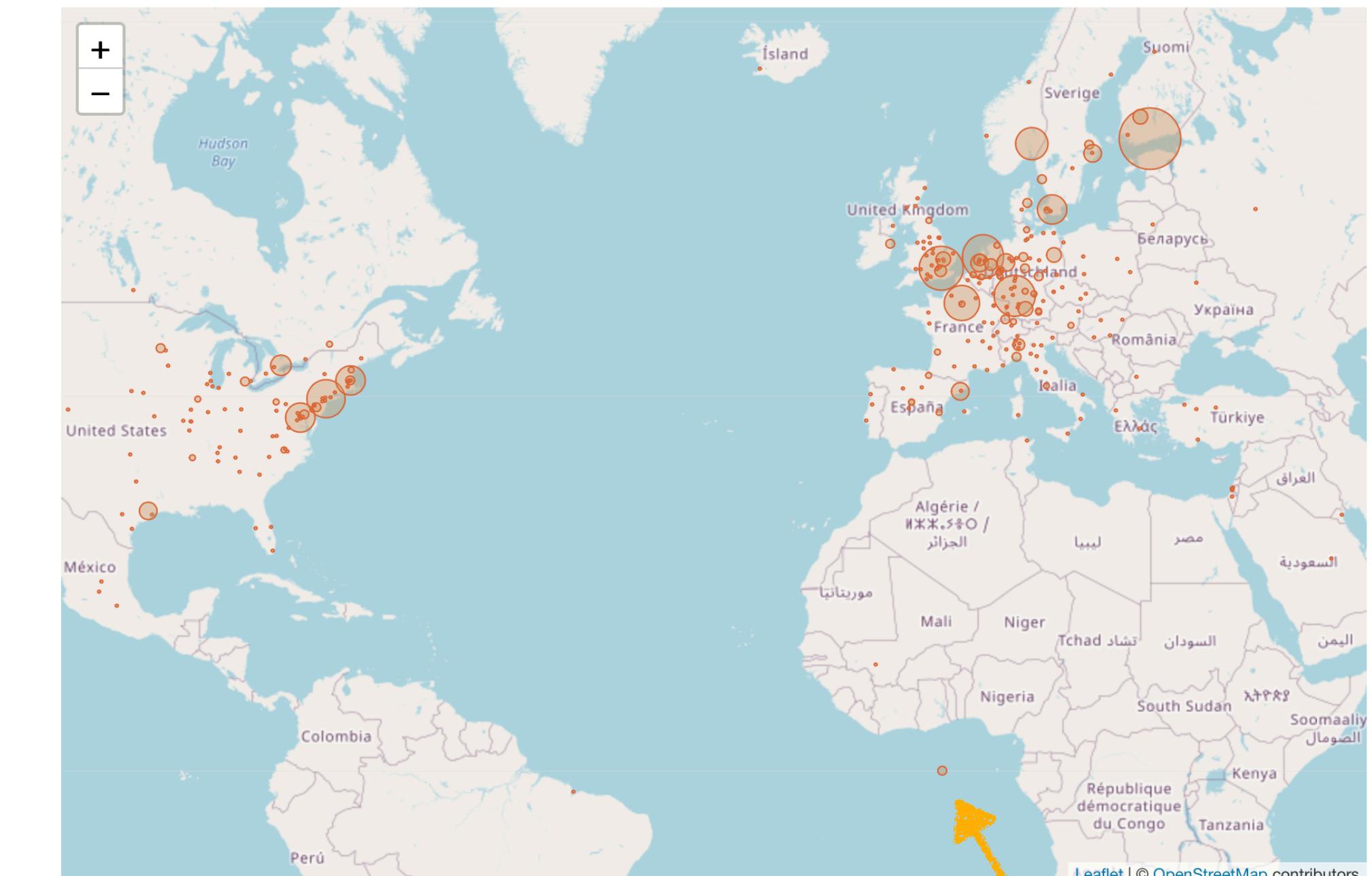
The most geo-tagged place on earth is Null Island



A troubleshooting country has been added with an Indeterminate sovereignty class called **Null Island** ([1](#), [2](#)). It is a fictional, 1 meter square island located off Africa where the equator and prime meridian cross. Being centered at 0,0 (zero latitude, zero longitude) it is useful for flagging geocode failures which are routed to 0,0 by most mapping services. Aside: "Null Islands" exist for all local coordinate reference systems besides WGS84 like State Plane (and global if not using modern [Greenwich prime meridian](#)). Null Island in Natural Earth is scaleRank 100, indicating it should never be shown in mapping. Side note: Rank 30 (zoom 29 in Google speak)

[https://en.wikipedia.org/wiki/Null\\_Island](https://en.wikipedia.org/wiki/Null_Island)

Michael Szell: The Data Science Process 2  
[http://michael.szell.net/downloads/lecture26\\_datasciprocess2.pdf](http://michael.szell.net/downloads/lecture26_datasciprocess2.pdf)  
2020-11-25



Progenetix publication collection  
[progenetix.org/publications/list](http://progenetix.org/publications/list)  
2020-11-28

25 / 3306 publications

# Documentation Strategies

## (Not so) Best Practices

- What is documentation? I'll remember this! \\_(`)\_/
- Just email me if help is needed, unexpectedly
- We had money for a chat bot.
- Clean code documents itself - Just use explicit variable/function names.
- Clean code documents itself - Never use explicit variable/function names.
- Perl POD it is. There is a command to show the notes in your terminal...
- I wrote a paper about the resource. In 2001.
- Haven't you found the GoogleGroups account?
- Documentation? StackOverflow, whelp!

mbaudis@netscape.net

```
normalize_variant_values_for_export(v, byc, drop_fields=None):
```

BIOINFORMATICS APPLICATIONS NOTE Vol. 17 no. 12 2001  
Pages 1228–1229



**Progenetix.net: an online repository for molecular cytogenetic aberration data**

Michael Baudis<sup>1, 2,\*</sup> and Michael L. Cleary<sup>2</sup>

<sup>1</sup>Medizinische Klinik und Poliklinik V der Universität Heidelberg, Germany and

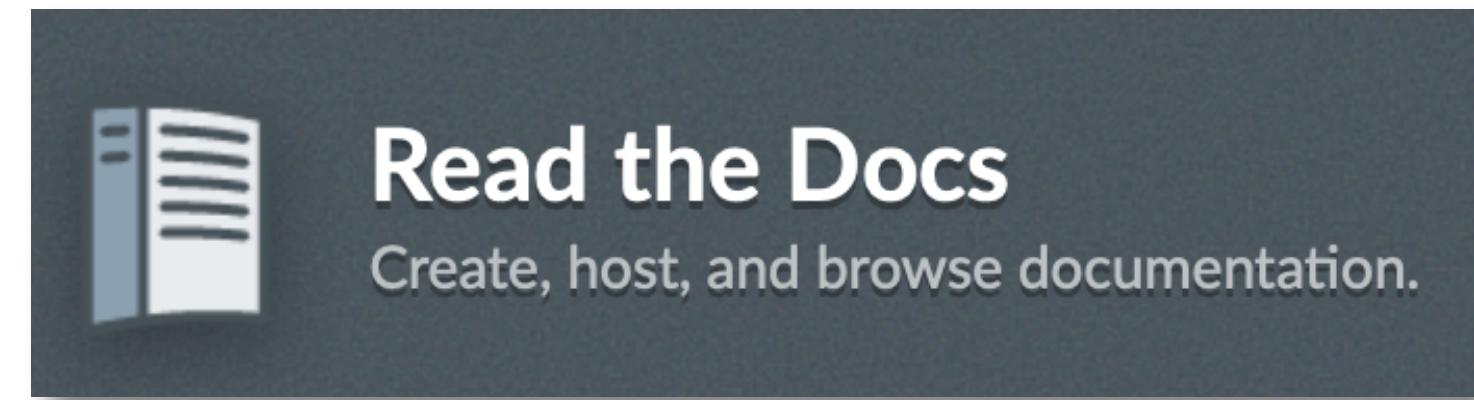
<sup>2</sup>Department of Pathology, Stanford University Medical Center, Stanford, CA 94305, USA

Received on July 5, 2001; revised on July 9, 2001; accepted on July 16, 2001

```
f_d = f_d_s[c_t]
r = {}
for k in res_schema.keys():
    if k in f_d:
        r.update({k:f_d[k]})
```

# Documentation Strategies Currently en Vogue

- Cloud-based documentation systems with online compilation
- written in simplified markup languages
  - Markdown (Yeah!)
  - Restructured Text (Meeh...)
- local and/or service based compilation and hosting
- build systems & output hosting
  - ReadTheDocs
    - ▶ direct building from .rst document tree or MkDocs based
  - Github Pages
    - ▶ direct using Jekyll or over MkDocs through GH actions



**Local Testing**

```

FOLDERS
progenetix-web
  .github
  .next
  docs
    css
mkdocs.yaml
  1 | site_
  2 | site_
  3 | site_
  4 | copyr
  5 | repo_name: 'progenetix-web'
  6 | repo_url: https://github.com/progenetix/progenetix-web

[→ progenetix-web git:(main) mkdocs serve
INFO - Building documentation...
INFO - [macros] - Macros arguments: {'module_name': 'main',
'modules': [], 'include_dir': '', 'include_yaml': [],
'j2_block_start_string': '', 'j2_block_end_string': '',
'j2_variable_start_string': '', 'j2_variable_end_string': '',
'on_undefined': 'keep', 'on_error_fail': False, 'verbose': False}
INFO - [macros] - Extra variables (config file):
['excerpt_separator', 'blog_list_length', 'social']
INFO - [macros] - Extra filters (module): ['pretty']
INFO - MERMAID2 - Initialization arguments: {}
INFO - MERMAID2 - Using javascript library (8.8.0):
  https://unpkg.com/mermaid@8.8.0/dist/mermaid.min.js
INFO - Cleaning site directory
INFO - The following pages exist in the docs directory, but are not
included in the "nav" configuration:
  - beaconplus.md
  - changelog.md
  - classifications-and-ontologies.md
  - progenetix-data-review.md
  - progenetix-website-builds.md
  - publication-collection.md
INFO - MERMAID2 - Found superfences config: {'custom_fences': [{name': 'mermaid', 'class': 'mermaid', 'format': <function fence_mermaid at 0x104075ab0>}]}
INFO - MERMAID2 - Page 'Technical Notes': found 2 diagrams, adding scripts
INFO - Documentation built in 0.83 seconds
INFO - [09:05:32] Watching paths for changes: 'docs', 'mkdocs.yaml'
INFO - [09:05:32] Serving on http://127.0.0.1:8000/
INFO - [09:05:33] Browser connected:
  http://127.0.0.1:8000/classifications-and-ontologies/

```

**Web Deployment (Github)**

the Progenetix oncogenes and their role in cancer development  
Michael Baudis and progenetix.org

```

classifications-and-ontologies.md
# Classification and Ontology
The Progenetix team is moving towards a more modular and flexible architecture. This involves
decentralizing certain components and creating a more dynamic system for managing data and
processes. One key aspect of this transition is the use of GitHub Actions to handle deployment
and automation tasks. In this section, we will discuss the workflow setup and how it enables
efficient and reliable deployment of the Progenetix website and documentation.

## Workflow Setup
The Progenetix GitHub repository contains a workflow named 'mk-progenetix-docs' defined in
'mk-progenetix-docs.yaml'. This workflow is triggered by pushes to the main branch and consists
of several steps:
1. **refseq_ids_in_examples_aggregator_start**: A step that starts the aggregator process for
refseq IDs in examples.
2. **Update_VariantsDataTable_js**: A step that updates the VariantsDataTable.js file.
3. **Update_VariantsDataTable_js**: Another step that updates the VariantsDataTable.js file.

## Workflow Runs
The 'mk-progenetix-docs' workflow has been run 178 times. Some recent runs include:
- A run from 3 days ago that started the aggregator UI.
- A run from 11 days ago that updated the VariantsDataTable.js.
- A run from 16 days ago that also updated the VariantsDataTable.js.

## Contributors
The workflow was last updated by mbaudis, who performed a cleanup task. There is currently 1 contributor listed.

## Repository Structure
The repository structure includes:
- CURIE prefix: A table mapping CURIE prefixes to their corresponding namespaces.
- NCIT: National Cancer Institute Thesaurus.
- HP: Human Phenotype Ontology.
- PMID: PubMed ID.
- progenetix.org/services: A service for interacting with the Progenetix API.
- geo: Geographical entities.
- services/ids/geo: Geographical entity identifiers.
- GSM491153: A specific geographical entity identifier.
- arrayexpress: ArrayExpress dataset.
- cellosaurus: Cellosaurus dataset.
- UBERON: UBERON ontology.
- cbioportal: CBIOPORTAL dataset.
- //progenetix.org/services: A placeholder for the API endpoint.
- Private filters: A section discussing the use of private filters in the API.

## Conclusion
This section provides an overview of the workflow setup and deployment process for the Progenetix website and documentation. By leveraging GitHub Actions, the team can ensure that the website remains up-to-date and functional without manual intervention. The use of CURIE prefixes and structured data formats like JSON-LD and RDFa allows for better integration with external databases and APIs.
```

# Documentation Strategies

## Best Practices

- start early
- update often
- sometimes try to follow your own guide
- balance between inline documentation & doc system
- use Markdown
- plan for contingencies
  - ➡ cloud providers disappear | cancel services | change terms



[https://en.wikipedia.org/wiki/List\\_of\\_defunct\\_social\\_networking\\_services](https://en.wikipedia.org/wiki/List_of_defunct_social_networking_services)

[https://en.wikipedia.org/wiki/List\\_of\\_search\\_engines#Defunct\\_or\\_acquired\\_search\\_engines](https://en.wikipedia.org/wiki/List_of_search_engines#Defunct_or_acquired_search_engines)

# Task: Reading up Some Standards...

- genome coordinates - take your time...
- Don't you want to rename all of your files "ISO style"?
  - write your time-anchored age in ISO ...

# Task: Make some Docs

```
pip3 install mkdocs && mkdocs serve
```

- look up mkdocs
  - extended <https://squidfunk.github.io/mkdocs-material/>

# Genomic File Formats



# Genomic File Formats

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

The image consists of three main parts:

- File Info Dialog:** A screenshot of a Mac OS X file info window for a file named "GSM1904006.CEL". The file is 69.1 MB and was modified on 3 February 2016 at 17:46. The "General" tab shows details like kind (FLC animation), size (69'078'052 bytes), and location (arrayRAID → arraymapln → affyRaw → GSE73822 → GPL6801). The "Preview" tab shows a thumbnail of a video file (FLC) with a large red X over it, accompanied by the text "not a movie...".
- BED File Example:** A screenshot of a BED file content. The file starts with "browser position chr7:127471196-127495720" and "browser hide all". It then lists genomic tracks for chromosome 7, each with a start position, end position, strand (+ or -), and itemRgb values. The last line is "itemRgb='On'".
- List of Genomic File Formats:** A vertical list of 20 genomic file formats, each preceded by a small blue square icon:
  - 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

# SAM/BAM and related specifications

These documents are maintained by the Large Scale Genomics work stream of the Global Alliance for Genomics & Health ([GA4GH](#)). Information on GA4GH procedures and how to get involved is [available here](#). Lists of contributors and acknowledgements can generally be found in each individual specification document.

## Specifications:

- [SAM v1](#)
- [SAM tags](#)
- [CRAM v2.1](#)
- [CRAM v3.x](#)
- [CRAM codecs](#)
- [BCF v1](#)
- [BCF v2.1](#)
- [CSI v1](#)
- [Tabix](#)
- [VCF v4.1](#)
- [VCF v4.2](#)
- [VCF v4.3](#)
- [VCF v4.4](#)
- [BED v1](#)
- [crypt4gh](#)
- [Htsget](#)
- [Refget](#)

## Alignment data files

[SAMv1.tex](#) is the canonical specification for the SAM (Sequence Alignment/Map) format, BAM (its binary equivalent), and the BAI format for indexing BAM files. [SAMtags.tex](#) is a companion specification describing the predefined standard optional fields and tags found in SAM, BAM, and CRAM files. These formats are discussed on the [samtools-devel mailing list](#).

[CRAMv3.tex](#) is the canonical specification for the CRAM format, while [CRAMv2.1.tex](#) describes its now-obsolete predecessor. [CRAMcodecs.tex](#) contains details of the CRAM custom compression codecs. Further details can be found at [ENA's CRAM toolkit page](#) and [GA4GH's CRAM page](#). CRAM discussions can also be found on the [samtools-devel mailing list](#).

The [tabix.tex](#) and [CSIV1.tex](#) quick references summarize more recent index formats: the tabix tool indexes generic textual genome position-sorted files, while CSI is [htslib](#)'s successor to the BAI index format.

## Unaligned sequence data files

We do not define or endorse any dedicated unaligned sequence data format. Instead we recommend storing such data in one of the alignment formats (SAM, BAM, or CRAM) with the unmapped flag set. However for completeness, we list the commonest formats below with external links.

[FASTA](#) is an early sequence-only format originally defined by William Pearson's tool of the same name.

[FASTQ](#) was designed as a replacement for FASTA, combining the sequence and quality information in the same file. It has no formal definition and several incompatible variants, but is described in a paper by Cock et al.

## Variant calling data files

[VCFv4.4.tex](#) is the canonical specification for the Variant Call Format and its textual (VCF) and binary (BCF) encodings, while [VCFv4.1.tex](#), [VCFv4.2.tex](#) and [VCFv4.3.tex](#) describe their predecessors. These formats are discussed on the [vcftools-spec mailing list](#).

[BCFv1\\_qref.tex](#) summarizes the obsolete BCF1 format historically produced by [samtools](#). This format is no longer recommended for use, as it has been superseded by the more widely-implemented BCF2.

[BCFv2\\_qref.tex](#) is a quick reference describing just the layout of data within BCF2 files.

## Discrete genomic feature data files

[BEDv1.tex](#) is the canonical specification for the GA4GH Browser Extensible Data (BED) format.

## File encryption

[crypt4gh.tex](#) is the canonical specification of the crypt4gh format which can be used to wrap existing file formats in an encryption layer.

## Transfer protocols

[Htsget.md](#) describes the *hts-get* retrieval protocol, which enables parallel streaming access to data sharded across multiple URLs or files.

[Refget.md](#) enables access to reference sequences using an identifier derived from the sequence itself.

## 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
ACTACCCCCGGAAGTAACCTGTCTTAAGTTGTGTCCCTCCTGAATTGTTAAGGCATAAGTTCTGCT
TTGACTTTAGGTTGGTTTGTGGTAGACACAGGGACAAGAGACAGTGAGGGATGTGCCATTGAC
TGATTGGGTGGAAAAGCTGTACTCTGTTAGAGAGTTCCCACCTCTGCTGCTGCCATTGAAAT
TGACTGGAAACCAGGAGGTCCCTGTCCATGATTCACCTGGTGGCCTAGCCAACTTCAAAGTAAAAGT
TTGCATTCTTGAACCTTCTAAATTGGAGTTGTTACAAACCCAGGAAAGGGCAATACAGTAGGTAAAAG
GATTAGGTATTCACTGGAAAAAAATTAAATCCATATTAAAGAACATGGTCAAATCAAACACAG
ATACACATGATTAGAACATGAAATGATTCCGTATTATGTTGTCAGCAATATAGTTTACAAATAAC
CCATATGAAAATGTAACAAAGCATATTACATCTCACATGCCATCTGTATTGACTGAATAAGCTTAGTG
ACATTATTGCAAATCTGTAGTTAATTGTACATAGACATTGCGTTAAAAGGAAATGTACATAATG
TAAAATAAATTACATTACGCAATTACAAAGTAATATTAAACAAAATTCTAGACAGCTGCCTCCTATT
TAAACAAAATAAATTACAGGTAGTTAAATTACATAAAACACATTAGGAAATAATAATTATAGAA
AGACAGATTGCAAATTAAAGTTATATTACAATGATAGACTGATCTCTCAAATCTGTGTGA
TAGAAATGGAGAAAAAAAGTACCAAGAAAAGGAATCTAAATGTTACTTCTAAAATAACACAAACAGA
TTCTGAAAAAATAGGGAAAAGTTACTTGAGGGTTAAAGTAGGTAATCTAGAAACTATGGCTAAAAC
AATAAAATCTACAAAACACAAGACTGACAATTATATTCTAAATAATAGAGATTGATCACTGAA
AACATGACTCCCACAAACTAAAGCTCTCATATACTGCCATTAAAGATCTGACTTGGTAGAAACACA
GAAAAATAAAATGCAAAATTAACTGTTAGCATTAGTTCAATTAAATAATGTAGACATAACCATT
TTTCATTGTCCCTGCTAGATATAAAATTATTAAACACACTGCAAACCCATCCTTTATAATGGATAAC
TATTGCTGGCTCACACACCAAGTTCTGATAACCTGAAATCCTGCTGCAGCCAGGGCACCTGAGGGC
AGGACCTGGGGAGACCCTTATTCCCAGAACAGCAGATGTAGTTCTCACAACTAAACTAGTCCCAGGAAA
GATCACATTCTGACAAGATTCCCTCACAGATTGCTCAAGGACTACTGTTTTCAACACCCTCAATCTTA
CAGTGGAAATAGAAGAAGAACCCACACTTGAATTGGTAATATATTATAAACAGGAGATCCCAGATCAT
TTGGGAATTGTGCTTCTCATGTACTATTGAGACCCACGTCAAGCTTGAACAGGCTCTCCCTGTATG
GTACTGAAAAGTACAGTCCCTCACTGTCTGTGGTGGAAATGTGAACAAAGACTCGAGATGGAGGC
AGGAGGATATGGGATGGTCTAAAGCAAGTGTAGGCATGGACATTTCAGGAAAGGGCTTTTTTTTT
TTTTTCTGCATGCCCTCACATTTCCTTATTCTTGAACAGTGGATTGGTG...
```

Homo sapiens chromosome 7, GRCh38.p12 Primary Assembly  
NCBI Reference Sequence: NC\_000007.14

# The VCF file format

## Standard for genomic variant representation

### Example

VCF header

```

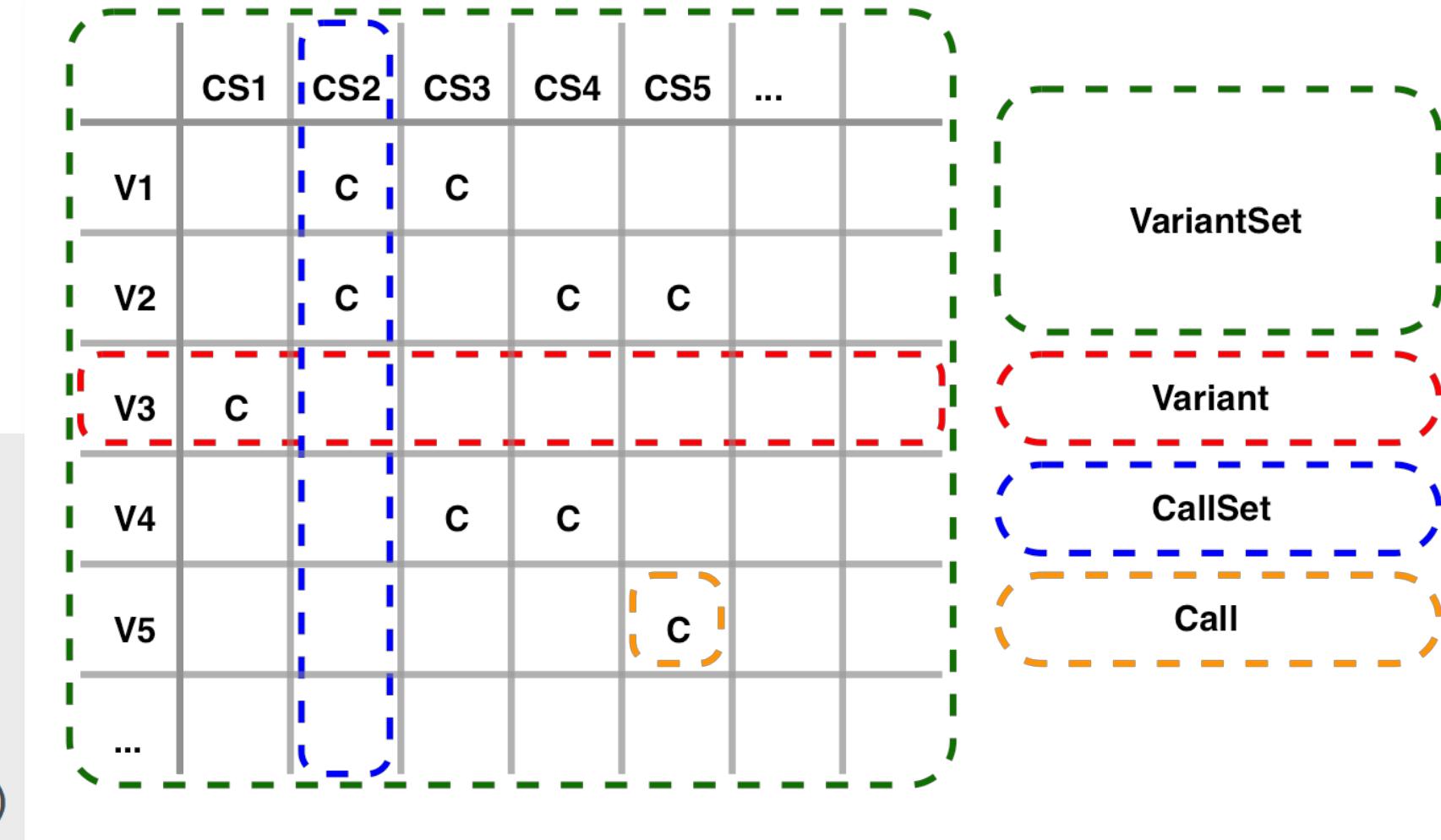
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String>Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag>Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String>Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer>Description="Genotype 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">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2
1 1 . ACG A,AT .
1 2 rs1 C T,CT .
1 5 . G <DEL> .
1 100 . T .

```

Body

Annotations:

- Deletion**: Variant at position 5 is a Deletion (<DEL>).
- SNP**: Variant at position 2 is a SNP (rs1).
- Large SV**: Variant at position 5 is a Large SV (SVTYPE=DEL; END=300).
- Insertion**: Variant at position 2 is an Insertion (T).
- Other event**: Variant at position 5 is an Other event (G).
- Mandatory header lines**: Lines starting with ##.
- Optional header lines (meta-data)**: Lines starting with ##INFO or ##FORMAT.
- Reference alleles (GT=0)**: Reference alleles for the variants.
- Alternate alleles (GT>0 is an index to the ALT column)**: Alternate alleles for the variants.
- Phased data**: Phased data for the variants across samples (SAMPLE1 and SAMPLE2).



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

# Storage



# 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 2021) ~35'000CHF (65x16TB á CHF550)
- 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)



# Task: Estimate Storage Requirements for 1000 Genomes

## How much computer storage is required for 1000 Genomes

- WES & WGS
- Different file formats
  - SAM
  - BAM
  - VCF
  - FASTA
- Associated costs
  - Cost factors
  - Raw Storage costs
- Familiarize with VCF format  
→specification in article collection



IBM-storage-unit-3500-Schiphol-1957

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.

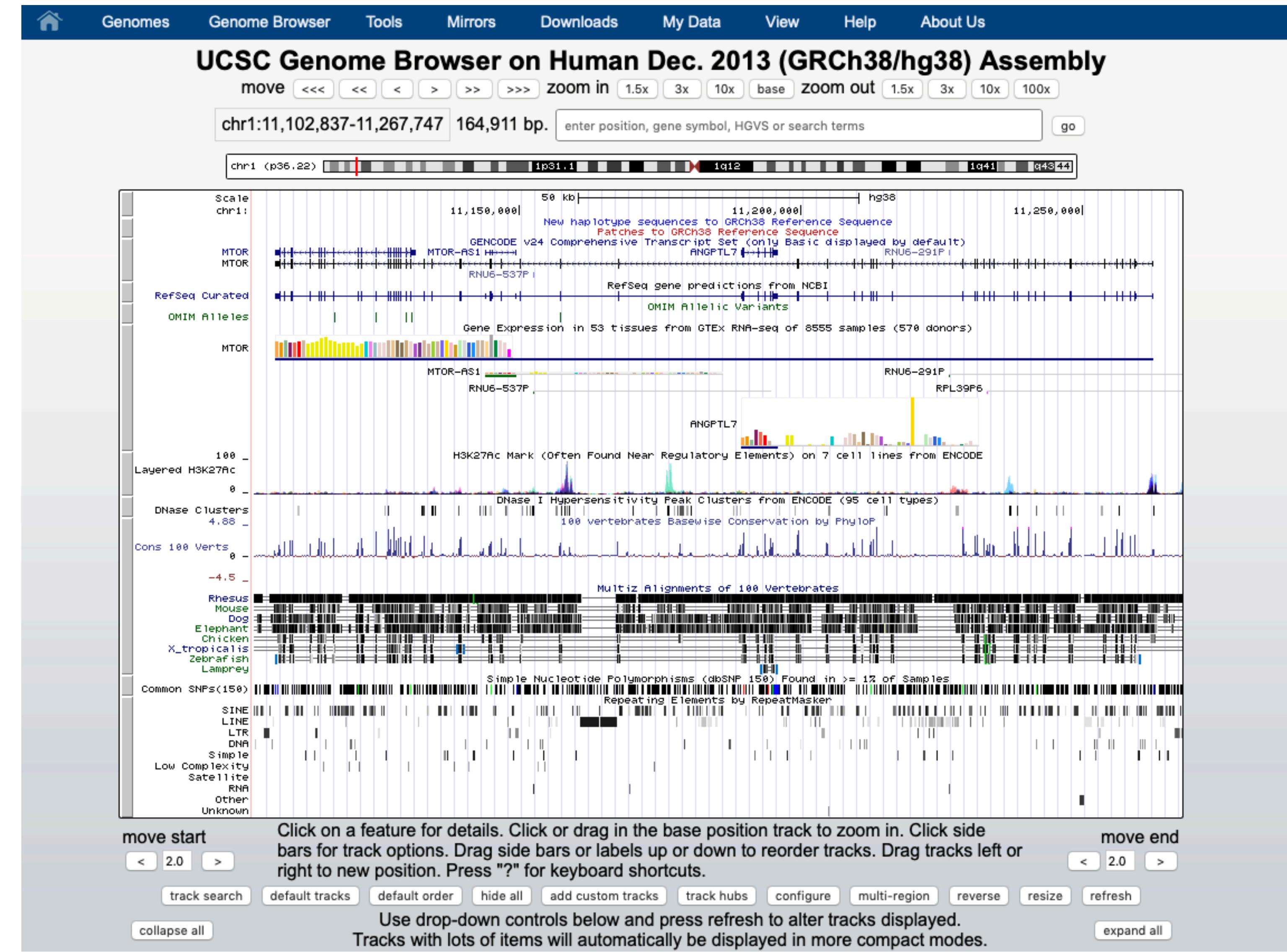
# Genome Resources

Sequences | Variants | Interpretations

## RESOURCES FOR GENOMICS: UCSC GENOME BROWSER

- ▶ Originated from the Human Genome Project
- ▶ Most widely used general genome browser
- ▶ many default tracks
- ▶ many species
- ▶ customization with "BED" files

[genome.ucsc.edu](http://genome.ucsc.edu)



## RESOURCES FOR GENOMICS: HUMAN GENOME RESOURCES AT NCBI

NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information Log in

### Human Genome Resources at NCBI

Download Browse View Learn

Search for Human Genes

Select a chromosome to access the [Genome Data Viewer](#)

Download

	GRCh38	GRCh37
Reference Genome Sequence	<a href="#">Fasta</a>	<a href="#">Fasta</a>
RefSeq Reference Genome Annotation	<a href="#">gff3</a>	<a href="#">gff3</a>
RefSeq Transcripts	<a href="#">Fasta</a>	<a href="#">Fasta</a>
RefSeq Proteins	<a href="#">Fasta</a>	<a href="#">Fasta</a>
ClinVar	<a href="#">vcf</a>	<a href="#">vcf</a>
dbSNP	<a href="#">vcf</a>	<a href="#">vcf</a>
dbVar	<a href="#">vcf</a>	<a href="#">vcf</a>

[www.ncbi.nlm.nih.gov/projects/genome/guide/human/](http://www.ncbi.nlm.nih.gov/projects/genome/guide/human/)

- ▶ Entry point for genome reference data
- ▶ Human genome assemblies
- ▶ Human variant collections (dbVar, ClinVar, dbSNP) for download

**Where to find genome *variant* data ...**

# Reference Resources for Human Genome Variants

## NCBI:dbSNP



- single nucleotide polymorphisms (SNPs) and multiple small-scale variations
- including insertions/deletions, microsatellites, non-polymorphic variants

## NCBI:dbVAR



- genomic structural variation
- insertions, deletions, duplications, inversions, multinucleotide substitutions, mobile element insertions, translocations, complex chromosomal rearrangements

## NCBI:ClinVar



- aggregates information about genomic variation and its relationship to human health

## EMBL-EBI:EVA



- open-access database of all types of genetic variation data from all species

## Ensembl



- portal for many things genomic...

# RESOURCES FOR CANCER GENOMICS

**COSMIC**  
Catalogue of somatic mutations in cancer

Home ▾ Resources ▾ Curation ▾ Tools ▾ Data ▾ News ▾ Help ▾ About ▾ Search COSMIC... Login ▾

**COSMIC v79, released 14-NOV-16**

COSMIC, the Catalogue Of Somatic Mutations In Cancer, is the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer.

Start using COSMIC by searching for a gene, cancer type, mutation, etc. below, or by browsing a region of the human genome using the map to the right.

eg: *Braf, COLO-829, Carcinoma, V600E, BRCA-UK, Campbell* **SEARCH**

**R Resources**

*Key COSMIC resources*

- Cell Lines Project
- COSMIC
- Whole Genomes
- Cancer Gene Census
- Drug Sensitivity
- Mutational Signatures
- GRCh37 Cancer Archive

**T Tools**

*Additional tools to explore COSMIC*

- Cancer Browser
- Genome Browser
- GA4GH Beacon
- CONAN

**C Expert Curation**

*High quality curation by expert postdoctoral scientists*

- Drug Resistance
- Cancer Gene Census
- Curated Genes
- Gene Fusions
- Genome-Wide Screens

**D Data**

*Further details on using COSMIC's content*

- Downloads
- License
- Submission
- Genome Annotation
- Datasheets
- Help
- FAQ

Browse the [genomic landscape](#) of cancer

**Cancer Gene Census Update**

7 genes have been added to the [Cancer Gene Census](#) -

- EPAS1 - Endothelial PAS domain protein 1.
- PTPRT - Protein tyrosine phosphatase, receptor type T.
- PPM1D - Protein phosphatase, Mg<sup>2+</sup>/Mn<sup>2+</sup> dependent 1D.
- BTK - Bruton tyrosine kinase.
- PREX2 - Phosphatidylinositol-3,4,5-trisphosphate dependent Rac exchange factor 2.
- TP63 - Tumour protein p63.
- QKI - QKI, KH domain containing RNA binding.

For full details, see the [Datasheet](#).

# RESOURCES FOR GENOMICS: CLINGEN

- ▶ "The Genomic Variant WG brings together representatives from the Sequence and Structural Variant communities for focused discussions on resolving discrepancies in variant interpretation and creating consistent curation guidelines."
- ▶ Interpreted genome variants with disease association

[clinicalgenome.org](https://clinicalgenome.org)



Get Started About Us Curation Activities Working Groups Expert Panels Documents & Announcements Tools 🔍

## Explore the clinical relevance of genes & variants

ClinGen is a National Institutes of Health (NIH)-funded resource dedicated to building a central resource that defines the clinical relevance of genes and variants for use in precision medicine and research.

🔍 Gene

Search

All Curated Genes Gene-Disease Validity Dosage Sensitivity Clinical Actionability Curated Variants Statistics More ?

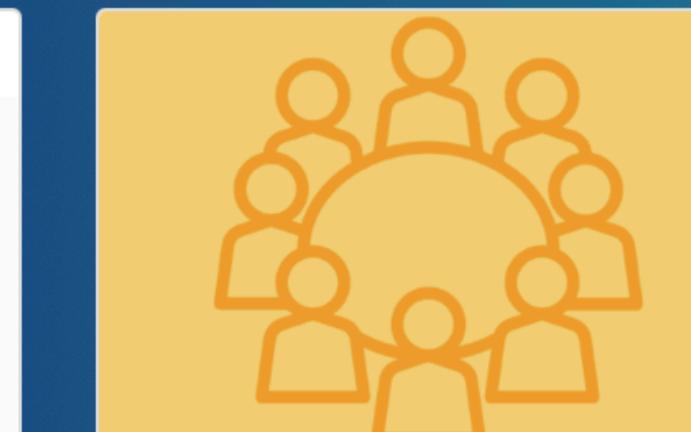
### ClinGen is defining the clinical relevance of genes and variants

Founded in 2013 by the National Human Genome Research Institute, ClinGen is a growing collaborative effort, involving three grants, nine principal investigators and over 2,700 contributors from more than 70 countries. Below are a series of recent updates that ClinGen has been working on.



#### Advancing genomic knowledge through global curation

In Genetics in Medicine, a special article from ClinGen describing the methods of genomic curation and development of software and infrastructure needed to support a global consortium capable of large-scale evidence-based curation.



#### Volunteer as an Expert

Interested in contributing your expert knowledge and/or data to a ClinGen GCEP or VCEP? Learn more and take our volunteer expert survey.



#### ClinGen Downloads and APIs

Visit our File Downloads and APIs page for a summary of available ClinGen curation files and API resources.



#### Volunteer to Curate

Please take a brief survey to tell us more about your interests and desired level of involvement so we can pair you with an appropriate curation activity and/or Expert Panel.

# The ClinGen and ClinVar Partnership

Both provide resources to support genomic interpretation

- ▶ ClinVar (an NCBI database/resource) is used as basis for curated variant <-> disease associations in ClinGen
- ▶ ClinGen - a funded project (application/funding limited)
- ▶ ClinVar - an internal NIH resource (dependent on political "goodwill")

### ClinGen - A Program

An NIH funded project

Building a central resource that defines the clinical relevance of genes and variants

ClinGen is addressing the following critical questions:

- Is the gene associated with disease?
- Is the variant pathogenic?
- Is the variant/gene information actionable?

Encouraging data sharing

- Promote lab submissions to ClinVar
- Facilitate patient data sharing through GenomeConnect



Assessing the clinical **validity** and **actionability** of genes and their relationship to diseases

### ClinVar- A Database

Funded by intramural NIH funding

Freely accessible and downloadable public archive of reports of the relationship between variants and conditions

Maintained by the National Center for Biotechnology Information (NCBI)



Maintaining a publicly available **database** of:

- Interpretations of the clinical significance of variants
- Submitter information
- Supporting evidence and individual level data, when available

**ClinGen**

Find out more online...

**ClinVar**

# RESOURCES FOR CANCER GENOMICS

National Cancer Institute U.S. National Institutes of Health | www.cancer.gov

## CANCER GENOME ANATOMY PROJECT

**CGAP How To**

**Tools**

**CGAP Info**

- Educational Resources
- Slide Tour
- Team Members
- References

**CGAP Data**

**Quick Links:**

- ICG
- NCI Home
- NCICB Home
- NCBI Home
- OCG

**Genes** **Chromosomes** **Tissues** **SAGE Genie** **RNAi** **Pathways**

### Cancer Genome Anatomy Project (CGAP)

The NCI's Cancer Genome Anatomy Project sought to determine the gene expression profiles of normal, precancer, and cancer cells, leading eventually to improved detection, diagnosis, and treatment for the patient. Resources generated by the CGAP initiative are available to the broad cancer community. Interconnected modules provide access to all CGAP data, bioinformatic analysis tools, and biological resources allowing the user to find "in silico" answers to biological questions in a fraction of the time it once took in the laboratory.

### The CGAP Website

Interconnected modules provide access to all CGAP data, bioinformatic analysis tools, and biological resources allowing the user to find "in silico" answers to biological questions in a fraction of the time it once took in the laboratory.

**Genes** Gene information, clone resources, SNP500Cancer, GAI, and transcriptome analysis.

**Chromosomes** FISH-mapped BAC clones, SNP500Cancer, and the Mitelman database of chromosome aberrations.

**Tissues** cDNA library information, methods, and EST-based gene expression analysis.

**Pathways** Diagrams of biological pathways and protein complexes, with links to genetic resources for each known protein.

**RNAi** RNA-interference constructs, targeted specifically against cancer relevant genes. New addition: Validated set of shRNAs.

International Cancer Genome Consortium

**Home** **Cancer Genome Projects** **Committees and Working Groups** **Policies and Guidelines** **Media**

### ICGC Cancer Genome Projects

Committed projects to date: 89

Sort by: Project

Biliary Tract Cancer Japan	Biliary Tract Cancer Singapore	Bladder Cancer China
Bladder Cancer United States	Blood Cancer China	Blood Cancer Singapore
Blood Cancer South Korea	Blood Cancer United States	Blood Cancer United States
Blood Cancer United States	Blood Cancer United States	Bone Cancer France
Bone Cancer United Kingdom	Bone Cancer United States	Brain Cancer Canada
Brain Cancer China	Brain Cancer United States	Brain Cancer United States
Breast Cancer China	Breast Cancer European Union / United Kingdom	Breast Cancer France
Breast Cancer Mexico	Breast Cancer South Korea	Breast Cancer South Korea

**ICGC Goal:** To obtain a comprehensive description of genomic, transcriptomic and epigenomic changes in 50 different tumor types and/or subtypes which are of clinical and societal importance across the globe.

[Read more »](#)

**Launch Data Portal »**

**Apply for Access to Controlled Data »**

**Announcements**

**23/August/2016** - The ICGC Data Coordination Center (DCC) is pleased to announce ICGC data portal data release 22 (<http://dcc.icgc.org>).

ICGC data release 22 in total comprises data from more than 16,000 cancer donors spanning 70 projects and 21 tumour sites.

**17/April/2016** - ICGCmed is pleased to announce the release of its white paper (<http://icgcmed.org>).

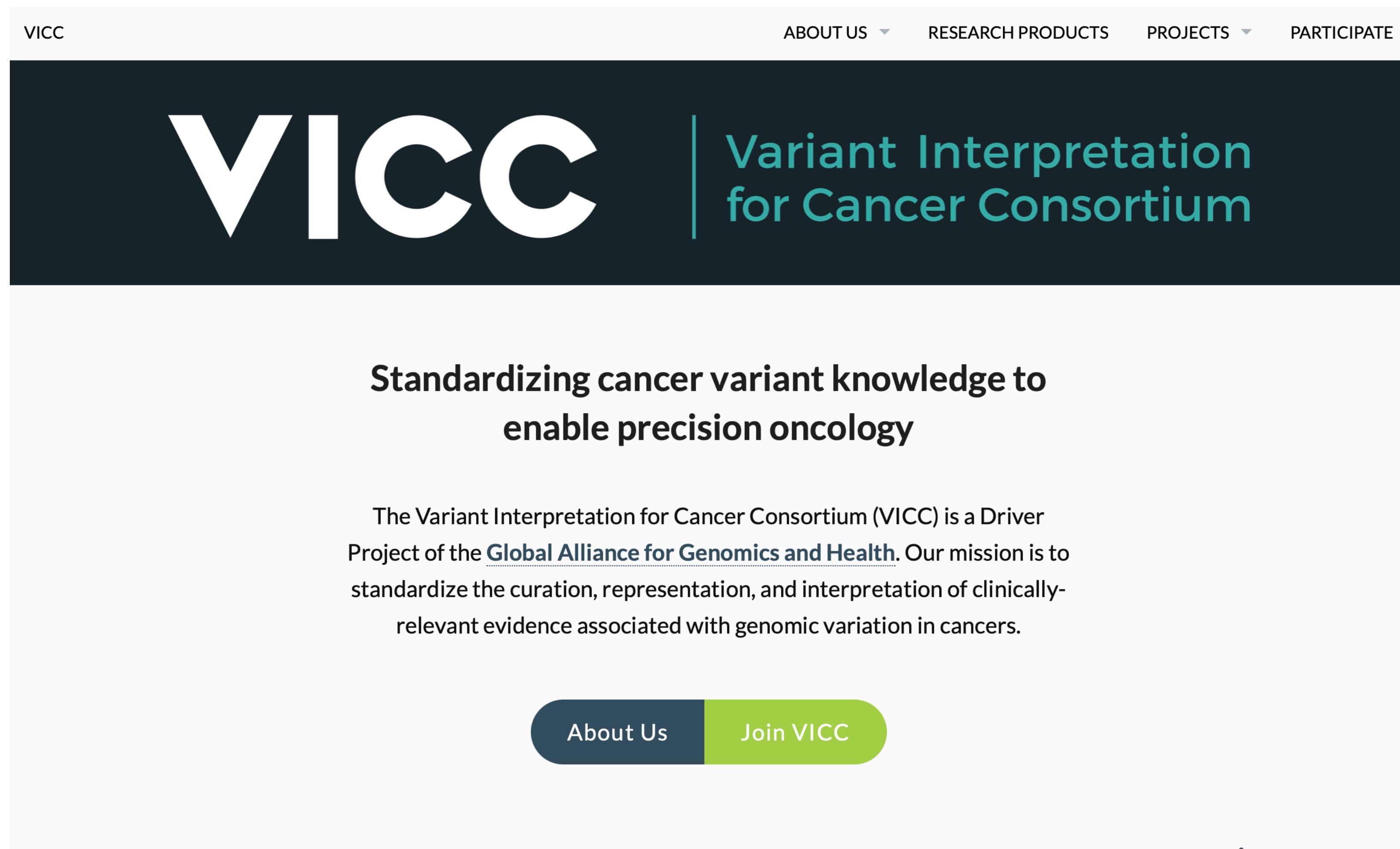
The International Cancer Genome Consortium for Medicine (ICGCmed) will link genomics data to clinical information, health and responses to therapies.

**18/November/2015** - The International Cancer Genome Consortium (ICGC) PanCancer dataset generated by the PanCancer Analysis of Whole Genomes (PCAWG) study is now available on Amazon Web Services (AWS), giving cancer researchers access to over 2,400 consistently analyzed genomes corresponding to over 1,100 unique ICGC donors (<https://icgc.org/icgc-in-the-cloud>).

## RESOURCES FOR CANCER GENOMICS

---

- ▶ The field of precision medicine aspires to a future in which a cancer patient's molecular information can be used to inform diagnosis, prognosis and treatment options most likely to benefit that individual patient. Many groups have created knowledgebases to annotate cancer genomic mutations associated with evidence of pathogenicity or relevant treatment options. However, clinicians and researchers are unable to fully utilize the accumulated knowledge derived from such efforts. Integration of the available knowledge is currently infeasible because each group curates their own knowledgebase without adherence to any interoperability standards. Therefore, there is a clear need to standardize and coordinate clinical-genomics curation efforts, to facilitate integration of the knowledge and evidence provided by institutions in academia, government, and industry alike.



The image shows the homepage of the Variant Interpretation for Cancer Consortium (VICC) website. The header features the VICC logo and navigation links for About Us, Research Products, Projects, and Participate. The main title "Variant Interpretation for Cancer Consortium" is displayed prominently in large teal letters. Below the title, a subtitle reads "Standardizing cancer variant knowledge to enable precision oncology". A detailed description of the consortium's mission follows, mentioning it is a Driver Project of the Global Alliance for Genomics and Health. At the bottom, there are two buttons: "About Us" (dark blue) and "Join VICC" (green).

<https://cancervariants.org/>

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

# RESOURCES FOR GENOMICS - THEY MAY BREAK SOMETIMES ...

NCBI Resources How To Sign in to NCBI

We are sorry, but the page you requested is no longer available.

NCBI's SKY-CGH site has been retired.

The public data from this resource can be downloaded from our [FTP server](#) and will soon be available in the [dbVar database \(SKY-CGH\)](#).

You are here: NCBI > National Center for Biotechnology Information Write to the Help Desk

Skip Navigation

GETTING STARTED RESOURCES POPULAR FEATURED NCBI INFORMATION

NCBI Education	Chemicals & Bioassays	PubMed	Genetic Testing Registry	About NCBI
NCBI Help Manual	Data & Software	Bookshelf	PubMed Health	Research at NCBI
NCBI Handbook	DNA & RNA	PubMed Central	GenBank	NCBI News
Training & Tutorials	Domains & Structures	PubMed Health	Reference Sequences	NCBI FTP Site
Submit Data	Genes & Expression	BLAST	Gene Expression Omnibus	NCBI on Facebook
	Genetics & Medicine	Nucleotide	Map Viewer	NCBI on Twitter
	Genomes & Maps	Genome	Human Genome	NCBI on YouTube
	Homology	SNP	Mouse Genome	
	Literature	Gene	Influenza Virus	
	Proteins	Protein	Primer-BLAST	
	Sequence Analysis	PubChem	Sequence Read Archive	
	Taxonomy			
	Variation			

Cancer Genome Anatomy Project (CGAP)

The NCI's [Cancer Genome Anatomy Project](#) sought to determine the gene expression profiles of normal, precancer, and cancer cells for diagnosis, and treatment for the patient. Resources generated by the CGAP initiative are available to the broad cancer community. Data, bioinformatic analysis tools, and biological resources allowing the user to find "in silico" answers to biological questions in a timely manner.

[Read more about CGAP](#) and access the many valuable resources.

Cancer Genome Characterization Initiative (CGCI)

The [Cancer Genome Characterization \(CGC\) Initiative](#): Assessing the use of new genomics technologies to strategically characterize tumors. Groups involved with the CGCI Initiative make all of their data available through a publicly accessible database. Cancer CGCI incorporates genomic characterization methods including exome and transcriptome analysis using second generation sequencing to identify genetic changes leading to cancer.

[Read more about the CGC Initiative](#) and how the project is enabling the next generation of discovery through rapid data release and analysis.

Download Plugin: [Windows](#) [Mac OS X](#) [Linux](#)

National Center for Biotechnology Information, U.S. National Library of Medicine  
8600 Rockville Pike, Bethesda MD, 20894 USA  
[Policies and Guidelines](#) | [Contact](#)

NATIONAL LIBRARY OF MEDICINE NATIONAL INSTITUTES OF HEALTH USA.gov

A Service of the National Cancer Institute

as of 2018-09-19

# Beyond a Single Resource: Federation

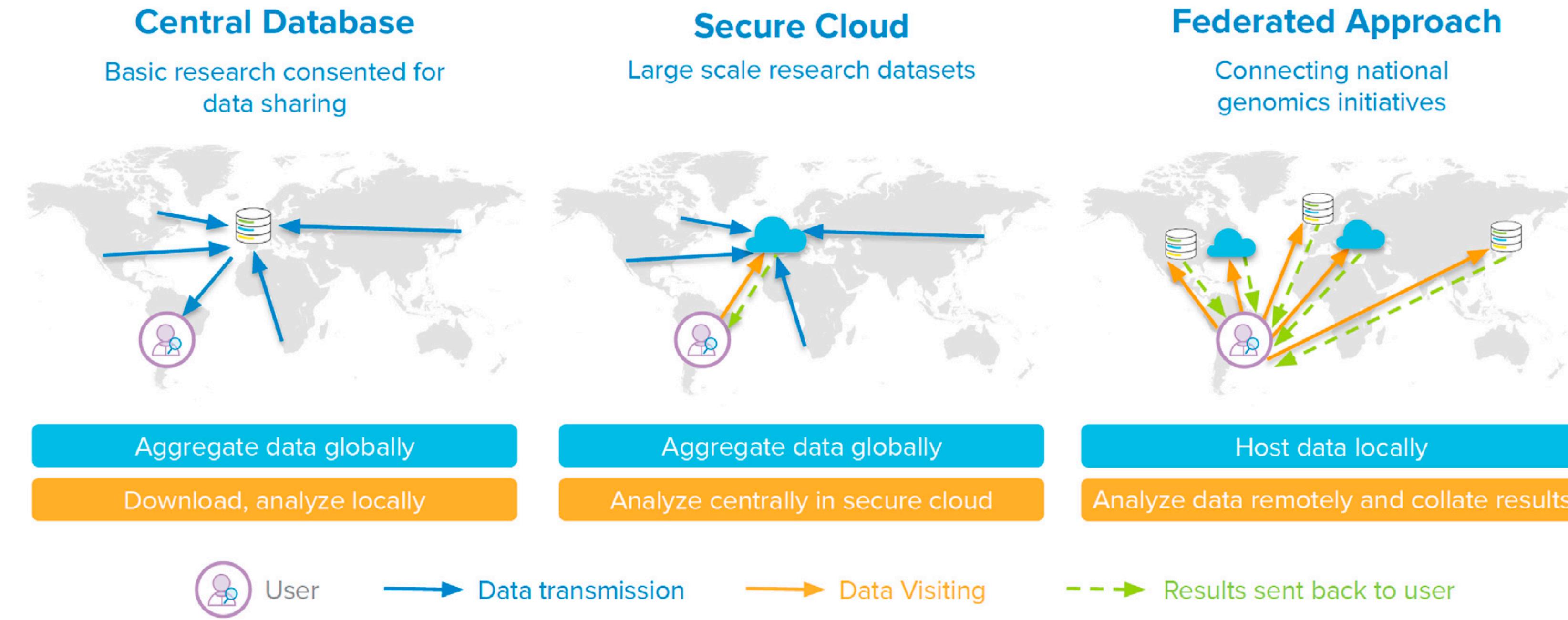
Cell Genomics

CellPress  
OPEN ACCESS

Commentary

International federation of genomic medicine databases using GA4GH standards

Adrian Thorogood,<sup>1,2,\*</sup> Heidi L. Rehm,<sup>3,4</sup> Peter Goodhand,<sup>5,6</sup> Angela J.H. Page,<sup>4,5</sup> Yann Joly,<sup>2</sup> Michael Baudis,<sup>7</sup> Jordi Rambla,<sup>8,9</sup> Arcadi Navarro,<sup>8,10,11,12</sup> Tommi H. Nyronen,<sup>13,14</sup> Mikael Linden,<sup>13,14</sup> Edward S. Dove,<sup>15</sup> Marc Fiume,<sup>16</sup> Michael Brudno,<sup>17</sup> Melissa S. Cline,<sup>18</sup> and Ewan Birney<sup>19</sup>



**Figure 1. Data sharing approaches: Central database, secure cloud, and federated**

Central database: Data from multiple sources are pooled in a central database. Researchers download copies of data and analyze them in their own computing environment.

Secure cloud: Data from multiple sources are pooled in a central cloud environment. Researchers remotely visit data and run their analyses in the cloud and download the result.

Federation: Data remain within locally controlled databases and computing environments, which may be cloud environments. Researchers remotely visit data, run their analyses at each site, and receive a local result, which can then be aggregated.

# Task: Exploring Genome Resources

- primary deposition databases
- interpreted databases (e.g. variant annotations...)
- suggestion: VICC paper (Wagner et al.)
  - Wagner et al (2020): A harmonized meta-knowledgebase of clinical interpretations of somatic genomic variants in cancer
- make some notes about different genome resources and their primary use
  - ➡ Don't think only "human" \\_(\_\\_) /

# BIO392: Course Schedule

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

	Friday 2025-04-04	Tuesday 2025-04-08	Wednesday 2025-04-09	Thursday 2025-04-10	Friday 2025-04-11	Tuesday 2025-04-15	Wednesday 2025-04-16	Thursday 2025-04-17	Tuesday 2025-04-29	Wednesday 2025-04-30	Friday 2025-05-02	Tuesday 2025-05-06	Wednesday 2025-05-07
09:00 - 10:00	* Room information * Administrative - discuss times/days - exam		Jiahui: Terminal / Unix / Files	Hangjia: R environment introduction	Michael: Genomic Resources & Data Sharing		Feifei: Sequence analysis practical. FastQC, trimmomatic, BWA-MEM2, SAMtools, GangSTR, BCFtools	Michael: Genomic Data & Privacy		Feifei & Ziying: survival	Feifei: population structure		
10:00 - 11:00	Tina Siegenthaler: technical introduction (room, computer, accounts)		Jiahui: Terminal / Unix / Files	Hangjia: R exercise	Michael: Genomic Resources & Data Sharing		Feifei:: Sequence analysis practical	Michael: Genomic Data & Privacy		Feifei & Ziying: survival	Feifei: population structure		Discussion
11:00 - 12:00	* explore course site * create Github accounts and forward to bio392@compbiozurich.org *feifei&jiahui: overall schedule of the course		jiahui: SIB online introduction to Unix		Michael: Genomic Resources & Data Sharing		Feifei:: Sequence analysis practical	Michael: Genomic Data & Privacy		Feifei: survival	Feifei: population structure		
13:00 - 14:00	jiahui: Github	Michael: Introduction	Jiahui & Ziying: Python	Hangjia: CNV paper reading	Hangjia: Clinvar and Clingen	Feifei: Sequence analysis introduction. Overview of pipeline from raw reads -> variant calling & interpretation	Feifei:: STR reading up			Feifei:: analysis & interpretation. Parsing VCF (cvcf2), UCSD genome browser, ENSEMBL variant effect predictor	Feifei: population structure	Feifei & others: Presentation & Discussion	Exam revision, Q&A
14:00 - 15:00	jiahui: Github&Git exercise	Michael: Introduction	Exercise	Hangjia: Progenetix as tool for CNV frequencies etc.	Hangjia: blast	Feifei: Sequence analysis introduction	Feifei:: STR reading up			Feifei:: analysis & interpretation.	Feifei: population structure	Feifei & others: Presentation & Discussion	
15:00 - 16:30		Michael: Introduction			Hangjia: Blast exercise					Feifei:: analysis & interpretation.	Feifei: population structure	Feifei & others: Presentation & Discussion	