



University of
Zurich^{UZH}

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>

Genomic File Formats

Types | Sizes | Use Cases

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)



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)

not a movie...

genome.ucsc.edu/FAQ/FAQformat.html

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

itemRgb="On"

BED file example

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.

The VCF file format

Standard for genomic variant representation

Example

VCF header

```

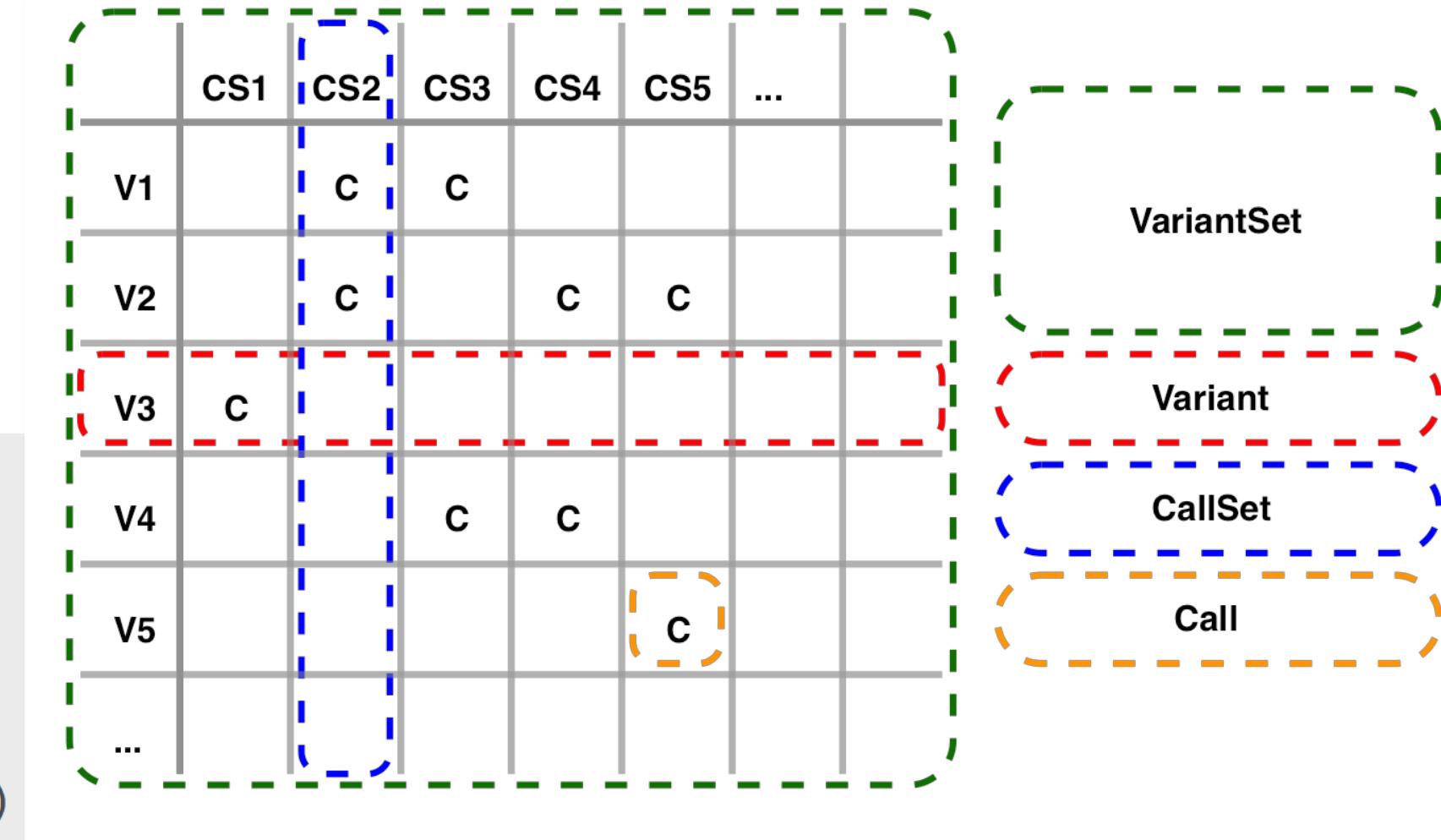
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype 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 ().
- SNP**: Variant at position 2 is a SNP (rs1).
- Large SV**: Variant at position 5 is a large structural variant (SVTYPE=DEL; END=300).
- Insertion**: Variant at position 2 has an insertion (T).
- Other event**: Variant at position 5 is an other event ().
- 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

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.

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.

Task: Reading up on Genome Technologies

- General NGS technologies
- count based vs. intensity based as principle
- long and short read technologies
 - ▶ advantages/applications for either
- dig deeper for some (molecular)-cytogenetic techniques:
 - ▶ banding analysis, SKY, M-FISH
 - ▶ SNP, aCGH arrays
 - ▶ chromosomal CGH
- ➔ notes about usage (research, clinical, historical vs. current)
- "T2T genome"
 - ▶ What technologies enabled this?
- Graph Genomes
 - ▶ Principles?

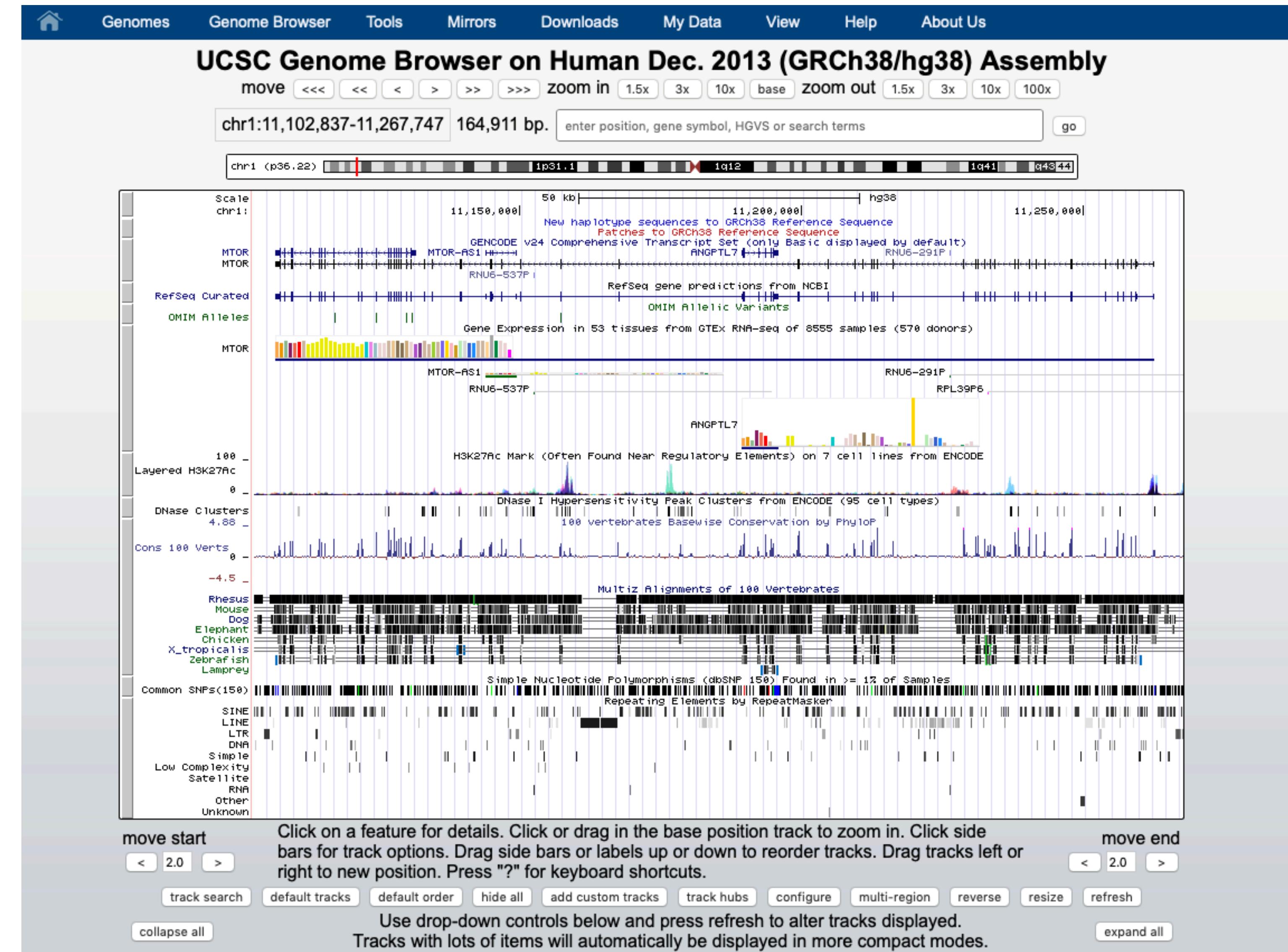
Genome Resources

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



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 | Fasta | Fasta |
| RefSeq Reference Genome Annotation | gff3 | gff3 |
| RefSeq Transcripts | Fasta | Fasta |
| RefSeq Proteins | Fasta | Fasta |
| ClinVar | vcf | vcf |
| dbSNP | vcf | vcf |
| dbVar | vcf | vcf |

www.ncbi.nlm.nih.gov/projects/genome/guide/human/

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



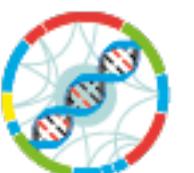
- genomic structural variation
- insertions, deletions, duplications, inversions, multinucleotide substitutions, mobile element insertions, translocations, complex chromosomal rearrangements

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, Mg²⁺/Mn²⁺ 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



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 ?

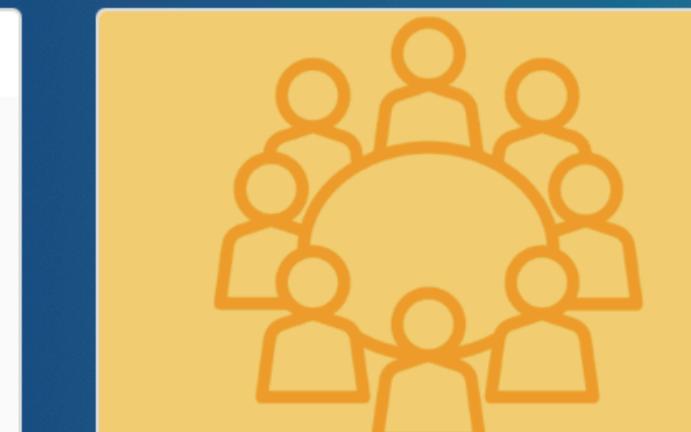
ClinGen is defining the clinical relevance of genes and variants

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.

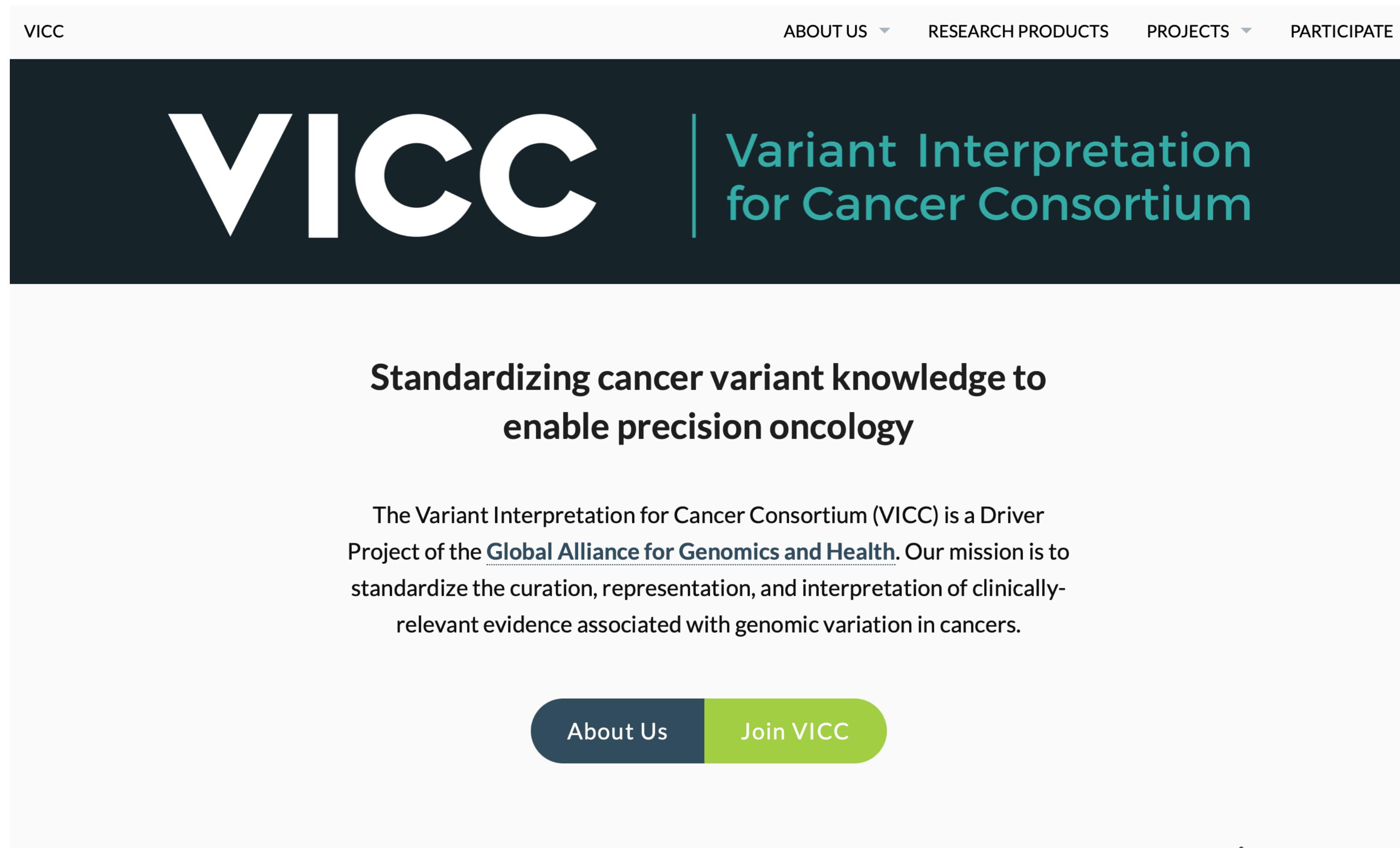
17/April/2016 - ICGCmed is pleased to announce the release of its white paper (<http://icgcmed.org>).

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 | 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 |

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,^{1,2,*} Heidi L. Rehm,^{3,4} Peter Goodhand,^{5,6} Angela J.H. Page,^{4,5} Yann Joly,² Michael Baudis,⁷ Jordi Rambla,^{8,9} Arcadi Navarro,^{8,10,11,12} Tommi H. Nyronen,^{13,14} Mikael Linden,^{13,14} Edward S. Dove,¹⁵ Marc Fiume,¹⁶ Michael Brudno,¹⁷ Melissa S. Cline,¹⁸ and Ewan Birney¹⁹

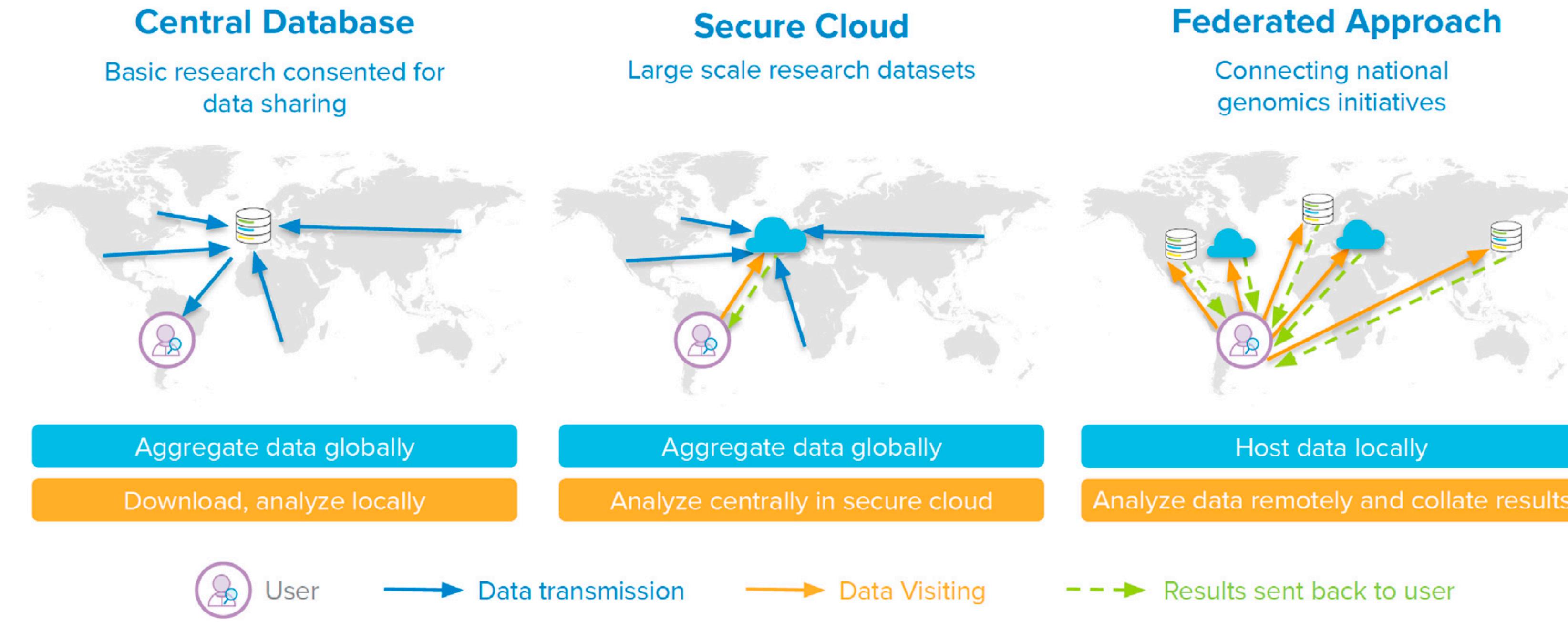


Figure 1. Data sharing approaches: Central database, secure cloud, and federated

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 | |