



University of
Zurich^{UZH}

BIO392

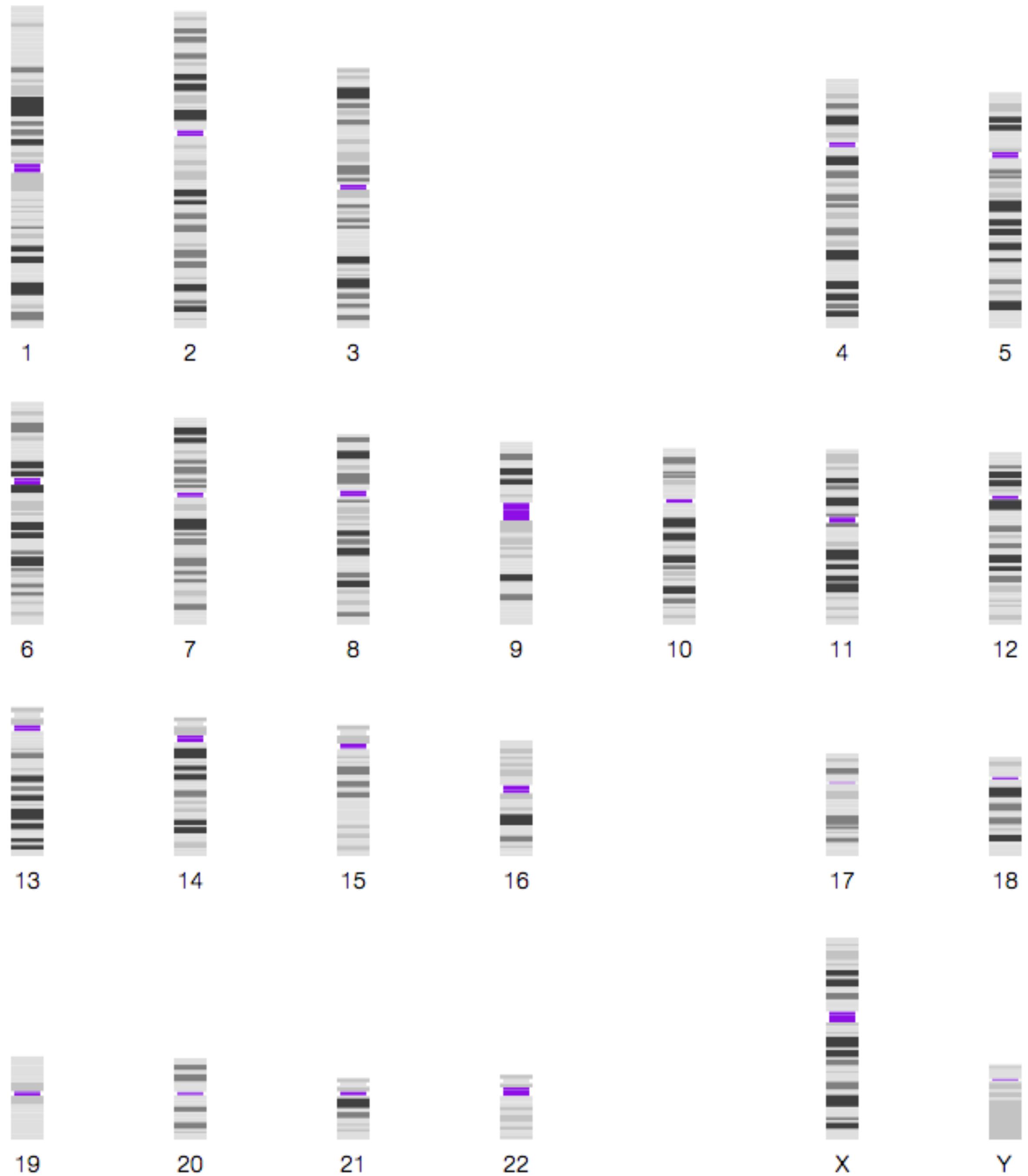
Bioinformatics of Genome Variations

Genome Editions and Genomics Resources

Michael Baudis **UZH SIB**
Computational Oncogenomics

Genome Editions

Sizes | positions | mappings



Chromosome	Basepair length (GRCh38)
1	248'956'422
2	242'193'529
3	198'295'559
4	190'214'555
5	181'538'259
6	170'805'979
7	159'345'973
8	145'138'636
9	138'394'717
10	133'797'422
11	135'086'622
12	133'275'309
13	114'364'328
14	107'043'718
15	101'991'189
16	90'338'345
17	83'257'441
18	80'373'285
19	59'128'983
20	64'444'167
21	46'709'983
22	50'818'468
X	156'040'895
Y	57'227'415
	3'080'419'480



genome.ucsc.edu
cytoBand_UCSC_hg18.txt

chr1	0	2300000	p36.33	gneg
chr1	2300000	5300000	p36.32	gpos25
chr1	5300000	7100000	p36.31	gneg
chr1	7100000	9200000	p36.23	gpos25
chr1	9200000	12600000	p36.22	gneg
chr1	12600000	16100000	p36.21	gpos50
chr1	16100000	20300000	p36.13	gneg
chr1	20300000	23800000	p36.12	gpos25
chr1	23800000	27800000	p36.11	gneg
chr1	27800000	30000000	p35.3	gpos25
chr1	30000000	32200000	p35.2	gneg
chr1	32200000	34400000	p35.1	gpos25
chr1	34400000	39600000	p34.3	gneg
chr1	39600000	43900000	p34.2	gpos25
chr1	43900000	46500000	p34.1	gneg
chr1	46500000	51300000	p33	gpos75
chr1	51300000	56200000	p32.3	gneg
chr1	56200000	58700000	p32.2	gpos50
chr1	58700000	60900000	p32.1	gneg
...
chrX	130300000	133500000	q26.2	gpos25
chrX	133500000	137800000	q26.3	gneg
chrX	137800000	140100000	q27.1	gpos75
chrX	140100000	141900000	q27.2	gneg
chrX	141900000	146900000	q27.3	gpos100
chrX	146900000	154913754	q28	gneg
chrY	0	1700000	p11.32	gneg
chrY	1700000	3300000	p11.31	gpos50
chrY	3300000	11200000	p11.2	gneg
chrY	11200000	11300000	p11.1	acen
chrY	11300000	12500000	q11.1	acen
chrY	12500000	14300000	q11.21	gneg
chrY	14300000	19000000	q11.221	gpos50
chrY	19000000	21300000	q11.222	gneg
chrY	21300000	25400000	q11.223	gpos50
chrY	25400000	27200000	q11.23	gneg
chrY	27200000	57772954	q12	gvar

Cytogenetic band Sizes

chromosome	band start position	band stop position	cytogenetic band	staining intensity	band size
chr6	63400000	63500000	q11.2	gneg	100000
chr15	64900000	65000000	q22.32	gpos25	100000
chr17	22100000	22200000	p11.1	acen	100000
chrX	65000000	65100000	q11.2	gneg	100000
chrY	11200000	11300000	p11.1	acen	100000
chr17	35400000	35600000	q21.1	gneg	200000
chr3	44400000	44700000	p21.32	gpos50	300000
chr3	51400000	51700000	p21.2	gpos25	300000
chr9	132500000	132800000	q34.12	gpos25	300000
chr13	45900000	46200000	q14.13	gneg	300000
chr15	65000000	65300000	q22.33	gneg	300000
chr1	120700000	121100000	p11.2	gneg	400000
chr8	39500000	39900000	p11.22	gpos25	400000
chr9	72700000	73100000	q21.12	gneg	400000
chr16	69400000	69800000	q22.2	gpos50	400000
chr19	43000000	43400000	q13.13	gneg	400000
chr9	70000000	70500000	q13	gneg	500000
chr20	41100000	41600000	q13.11	gneg	500000
...
chr9	51800000	60300000	q11	acen	8500000
chrX	76000000	84500000	q21.1	gpos100	8500000
chr11	76700000	85300000	q14.1	gpos100	8600000
chr13	77800000	86500000	q31.1	gpos100	8700000
chr7	77400000	86200000	q21.11	gpos100	8800000
chr8	29700000	38500000	p12	gpos75	8800000
chr3	14700000	23800000	p24.3	gpos100	9100000
chr5	82800000	91900000	q14.3	gpos100	9100000
chr6	104800000	113900000	q21	gneg	9100000
chrX	120700000	129800000	q25	gpos100	9100000
chr9	60300000	70000000	q12	gvar	9700000
chr1	212100000	222100000	q41	gpos100	10000000
chr1	128000000	142400000	q12	gvar	14400000
chr1	69500000	84700000	p31.1	gpos100	15200000
chrY	27200000	57772954	q12	gvar	30572954

Positional genomic data has to be evaluated
in the context of the correct edition

Chromosome	Basepairs 2003 (HG16)	Basepairs 2006 (HG18)	Basepairs 2009 (HG19)	Basepairs 2013 (GRCh38)	HG16 => HG19
1	246'127'941	247'249'719	249'250'621	248'956'422	2'828'481
2	243'615'958	242'951'149	243'199'373	242'193'529	-1'422'429
3	199'344'050	199'501'827	198'022'430	198'295'559	-1'048'491
4	191'731'959	191'273'063	191'154'276	190'214'555	-1'517'404
5	181'034'922	180'857'866	180'915'260	181'538'259	503'337
6	170'914'576	170'899'992	171'115'067	170'805'979	-108'597
7	158'545'518	158'821'424	159'138'663	159'345'973	800'455
8	146'308'819	146'274'826	146'364'022	145'138'636	-1'170'183
9	136'372'045	140'273'252	141'213'431	138'394'717	2'022'672
10	135'037'215	135'374'737	135'534'747	133'797'422	-1'239'793
11	134'482'954	134'452'384	135'006'516	135'086'622	603'668
12	132'078'379	132'349'534	133'851'895	133'275'309	1'196'930
13	113'042'980	114'142'980	115'169'878	114'364'328	1'321'348
14	105'311'216	106'368'585	107'349'540	107'043'718	1'732'502
15	100'256'656	100'338'915	102'531'392	101'991'189	1'734'533
16	90'041'932	88'827'254	90'354'753	90'338'345	296'413
17	81'860'266	78'774'742	81'195'210	83'257'441	1'397'175
18	76'115'139	76'117'153	78'077'248	80'373'285	4'258'146
19	63'811'651	63'811'651	59'128'983	59'128'983	-4'682'668
20	63'741'868	62'435'964	63'025'520	64'444'167	702'299
21	46'976'097	46'944'323	48'129'895	46'709'983	-266'114
22	49'396'972	49'691'432	51'304'566	50'818'468	1'421'496
X	153'692'391	154'913'754	155'270'560	156'040'895	2'348'504
Y	50'286'555	57'772'954	59'373'566	57'227'415	6'940'860
	3'070'128'059	3'080'419'480	3'095'677'412	3'088'781'199	18'653'140



samtools

<http://samtools.github.io>

Repositories 14 Packages People 14 Projects

Find a repository... Type: All Language: All

samtools

Tools (written in C using htslib) for manipulating next-generation sequencing data

● C 449 ⚡ 979 ⚠ 167 ⚡ 21 Updated 4 hours ago

bcftools

This is the official development repository for BCFtools. To compile, the develop branch of htslib is needed: git clone --branch=develop git://github.com/samtools/htslib.git htslib

● C 165 ⚡ 329 ⚠ 168 ⚡ 3 Updated 4 hours ago

htslib

C library for high-throughput sequencing data formats

● C 348 ⚡ 491 ⚠ 113 ⚡ 20 Updated 4 days ago

hts-specs

Specifications of SAM/BAM and related high-throughput sequencing file formats

● TeX 140 ⚡ 393 ⚠ 114 ⚡ 38 Updated 15 days ago

Report abuse

Top languages

● C ● Java ● Perl ● CSS ● TeX

People 14 >



Genome Liftover

Moving between genome editions

SOFTWARE TOOL ARTICLE

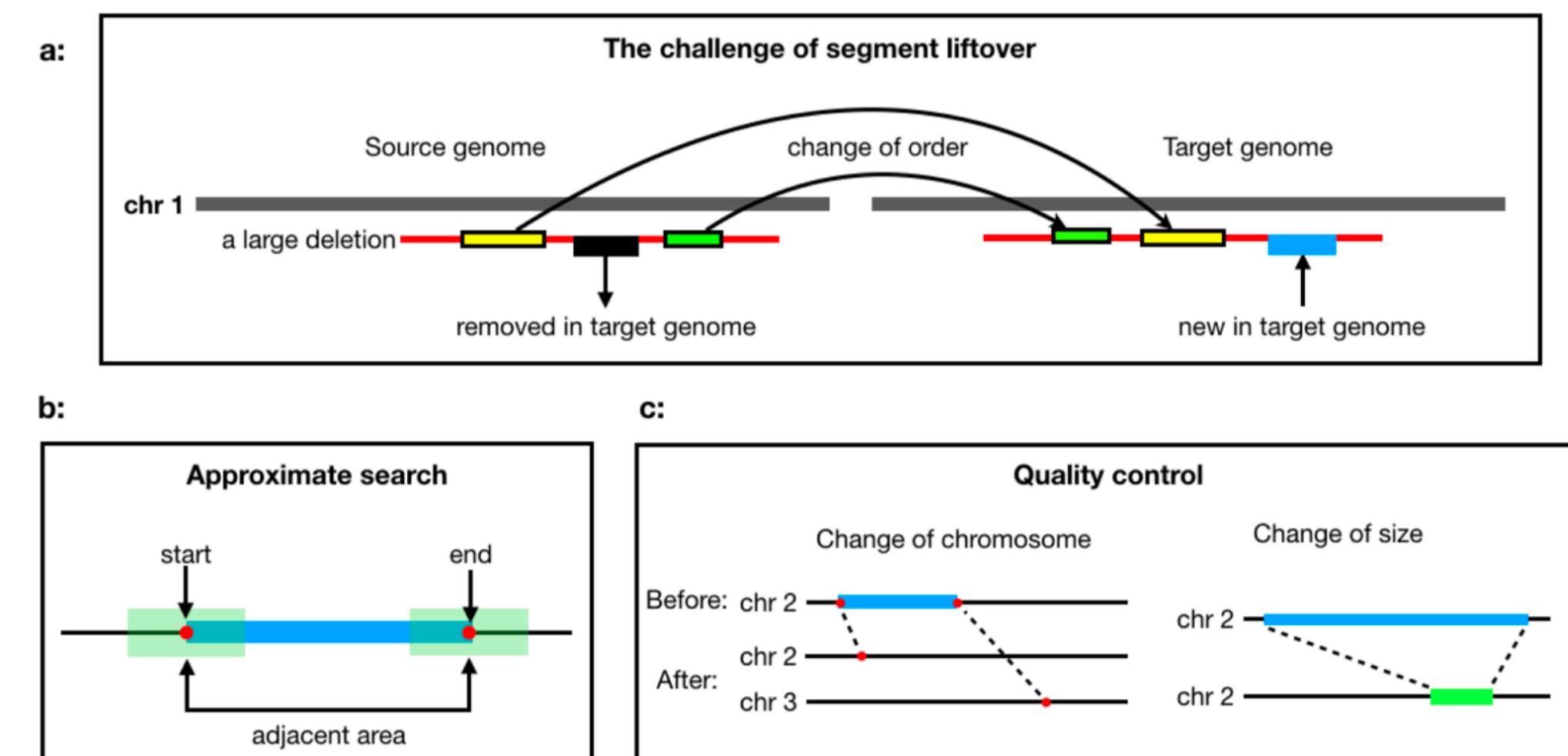
REVISED **segment_liftover** : a Python tool to convert segments between genome assemblies [version 2; referees: 2 approved]

Bo Gao  1,2, Qingyao Huang  1,2, Michael Baudis  1,2

¹Institute of molecular Life Sciences, University of Zürich, Zürich, CH-8057, Switzerland

²Swiss Institute of Bioinformatics, University of Zürich, Zürich, CH-8057, Switzerland

- different genome editions lead to shifting positions of defined elements such as genes
- local regions are frequently stable between editions
- shifts from change in regional lengths are defined in "chain files"
- chain files serve as guides for positional remapping using liftover methods
- Task: Read up on liftover techniques (starting w/ our article) & explore resources and other applications



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

Task: Reading up on Genome Technologies

- General NGS technologies
- count based vs. intensity based as principle
- dig deeper for some (molecular)-cytogenetic techniques:
 - ▶ banding analysis
 - ▶ SNP, aCGH arrays
 - ▶ SKY, M-FISH
 - ▶ chromosomal CGH
- ➔ notes about usage (research, clinical, historical vs. current)

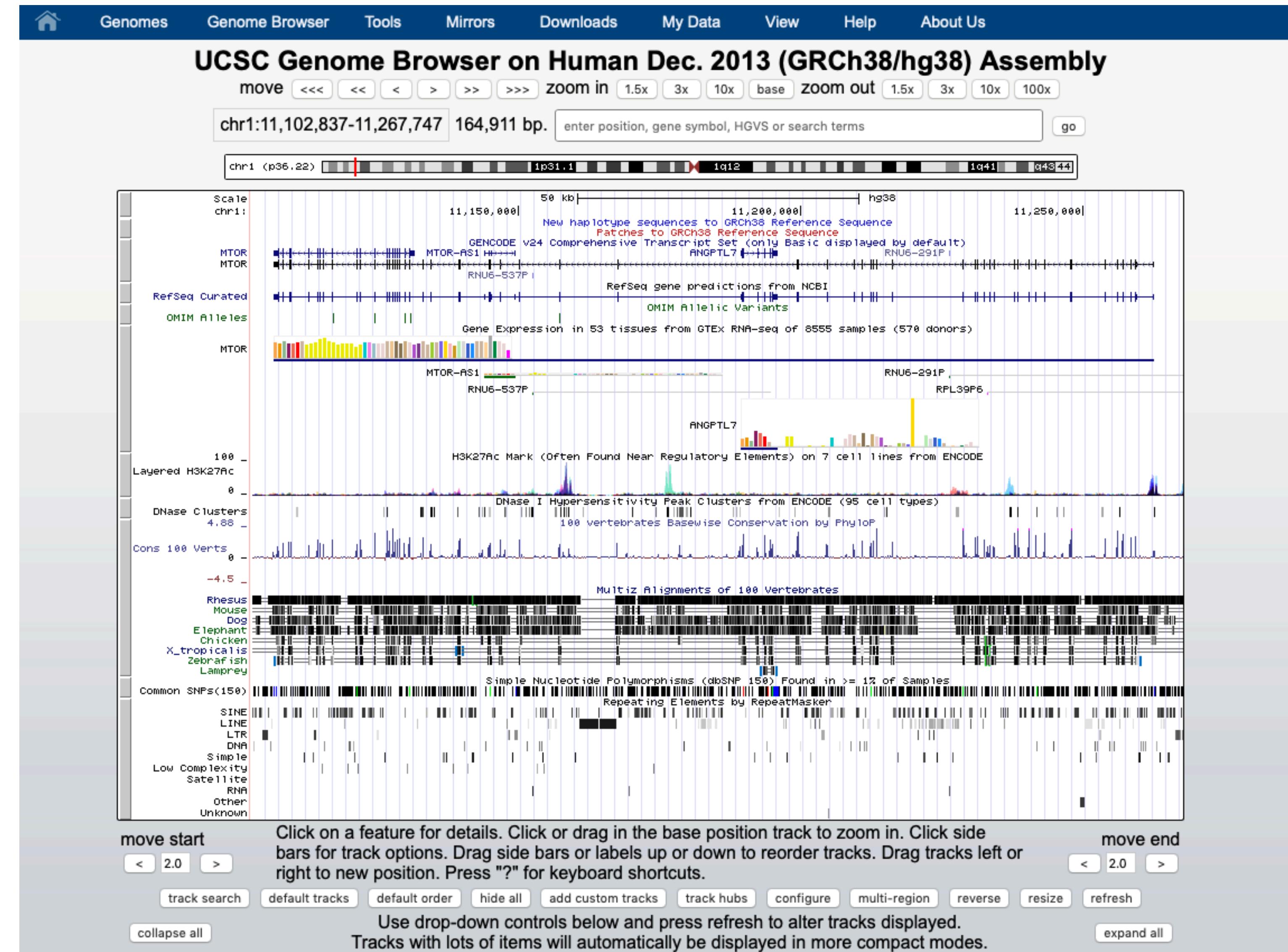
Genome Resources

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/

RESOURCES FOR GENOMICS: ENSEMBL

- ▶ Entry point for many genome data services and collections
- ▶ Downloads ("BioMart"), REST API

[www.ensembl.org/
Homo sapiens/Info/Index](http://www.ensembl.org/Homo_sapiens/Info/Index)

The screenshot shows the Ensembl Human GRCh38.p12 genome browser. At the top, there's a navigation bar with links to BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and a Blog. A search bar is also present. Below the header, it says "Human (GRCh38.p12) ▾". The main content area includes sections for "Search Human (Homo sapiens)", "Genome assembly: GRCh38.p12 (GCA_000001405.27)", "Gene annotation", "Comparative genomics", "Regulation", "Variation", and "ENCODE data in Ensembl". Each section contains links to more information and download options, along with small icons and examples. On the right side, there are boxes for "Example gene" (showing Pax6, INS, FOXP2, BRCA2, DMD, ssh) and "Example transcript" (showing a multi-exon gene structure). The bottom right corner features the "Ve!P" logo.

Search Human (*Homo sapiens*)

Search all categories ▾ Search Human... Go

e.g. **BRCA2** or **17:63992802-64038237** or **rs1333049** or **osteoarthritis**

Genome assembly: GRCh38.p12 (GCA_000001405.27)

- More information and statistics
- Download DNA sequence (FASTA)
- Convert your data to GRCh38 coordinates
- Display your data in Ensembl

Other assemblies
GRCh37 Full Feb 2014 archive with BLAST, VEP and BioMart Go

Gene annotation

What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

- More about this genebuild
- Download genes, cDNAs, ncRNA, proteins (FASTA)
- Update your old Ensembl IDs

Pax6 INS FOXP2 BRCA2 DMD ssh
Example gene

Example transcript

Comparative genomics

What can I find? Homologues, gene trees, and whole genome alignments across multiple species.

- More about comparative analysis
- Download alignments (EMF)

Example gene tree

Regulation

What can I find? DNA methylation, transcription factor binding sites, histone modifications, and regulatory features such as enhancers and repressors, and microarray annotations.

- More about the Ensembl regulatory build and microarray annotation
- Experimental data sources
- Download all regulatory features (GFF)

Example regulatory feature

ENCODE
ENCODE data in Ensembl

Variation

What can I find? Short sequence variants and longer structural variants; disease and other phenotypes

- More about variation in Ensembl
- Download all variants (GVF)
- Variant Effect Predictor

ATCGAGCT ATCCAGCT ATCGAGAT
Example variant

Ve!P

Example phenotype

Example structural variant

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

The screenshot shows the ClinGen Clinical Genome Resource website. At the top right is a search bar with the placeholder "Search our Knowledge Base for genes and diseases..." and a magnifying glass icon. Below the search bar are navigation links: About ClinGen, Working Groups, Resources, GenomeConnect, Share Your Data (highlighted in blue), and Curation Activities. The main banner features a blue background with a blurred image of laboratory glassware and a computer screen displaying genetic data. The text "Defining the clinical relevance of genes & variants for precision medicine and research..." is centered above three large numbers: 1496 (ClinGen Curated Genes), 31 (Expert Groups), and 10446 (Expert Reviewed Variants in ClinVar). To the right of these numbers is a magnifying glass icon labeled "Knowledge Base Search". Below the banner, the tagline "Sharing Data. Building Knowledge. Improving Care." is displayed, followed by a description of ClinGen's mission. Six call-to-action boxes are arranged in a grid at the bottom:

- ClinGen-ClinVar Partnership (Icon: DNA helix inside a circle)
- How to share genomic & health data (Icon: DNA helix inside a circular arrow)
- Learn about ClinGen curation activities (Icon: Computer monitor with DNA helix)
- GenomeConnect Patient Registry (Icon: Three DNA helices)
- View ClinGen's Resources & Tools (Icon: Computer monitor with multiple windows)
- Get Involved (Icon: Computer keyboard, mouse, and notepad)

clinicalgenome.org

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)

- Supporting **sharing** of variants interpretations



- Maintaining a publicly available **database** of:
 - Interpretations of the clinical significance of variants
 - Submitter information
 - Supporting evidence and individual level data, when available

clinicalgenome.org

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 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

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

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

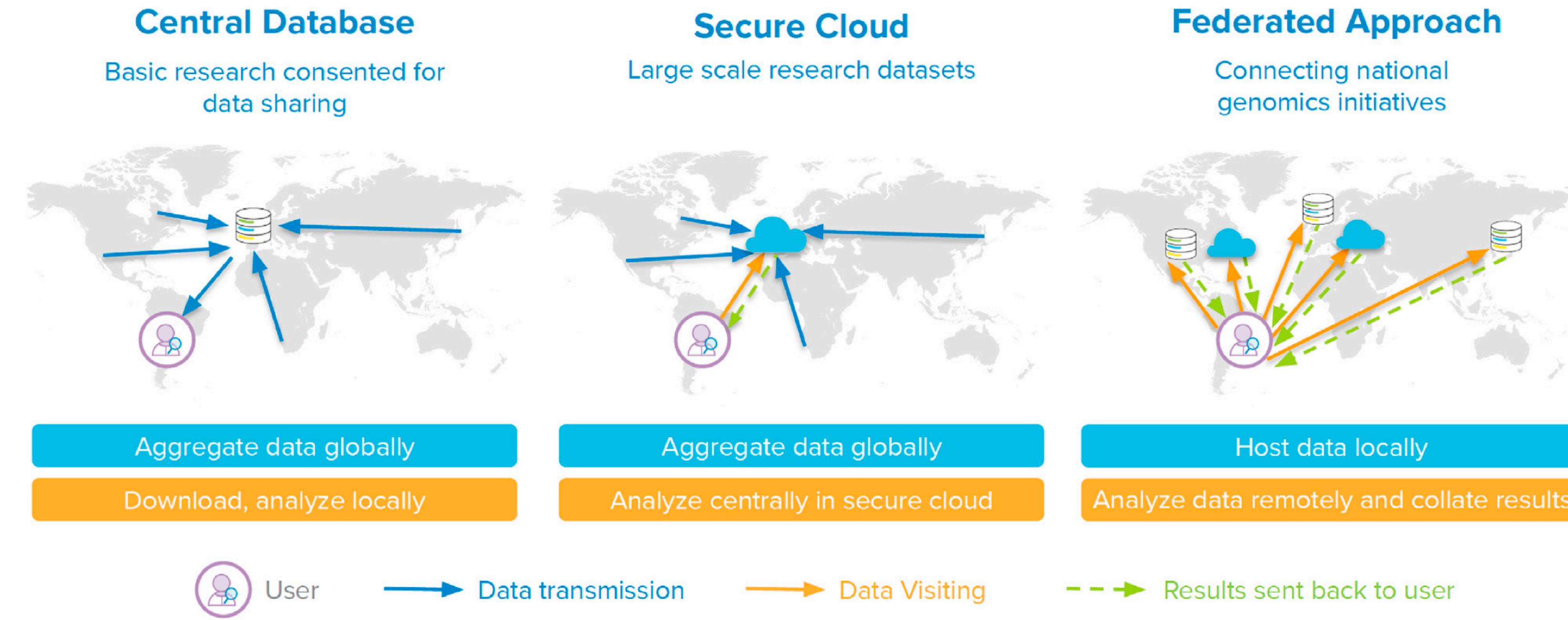


Figure 1. Data sharing approaches: Central database, secure cloud, and federated

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.

Progenetix in 2022

Cancer Genomics Reference Resource

- largest open resource for curated cancer genome profiles
- >116'000 cancer CNV profiles, from >800 NCI codes
- majority of data from genomic arrays with ~50% overall from SNP platforms with original data re-processing
- structured diagnostic encodings for NCI, ICD-O 3, UBERON
- identifier mapping for PMID, GEO, Cellosaurus, TCGA, cBioPortal where appropriate
- core biosample and technical metadata annotations where accessible (TNM, genotypic sex, survival ...)
- publication database and code mapping services

Cancer CNV Profiles

- ICD-O Morphologies
- ICD-O Organ Sites
- Cancer Cell Lines
- Clinical Categories

Search Samples

arrayMap

- TCGA Samples
- 1000 Genomes Reference Samples
- DIPG Samples
- cBioPortal Studies
- Gao & Baudis, 2021

Publication DB

- Genome Profiling
- Progenetix Use

Services

- NCI Mappings
- UBERON Mappings

Upload & Plot

Beacon⁺

Documentation

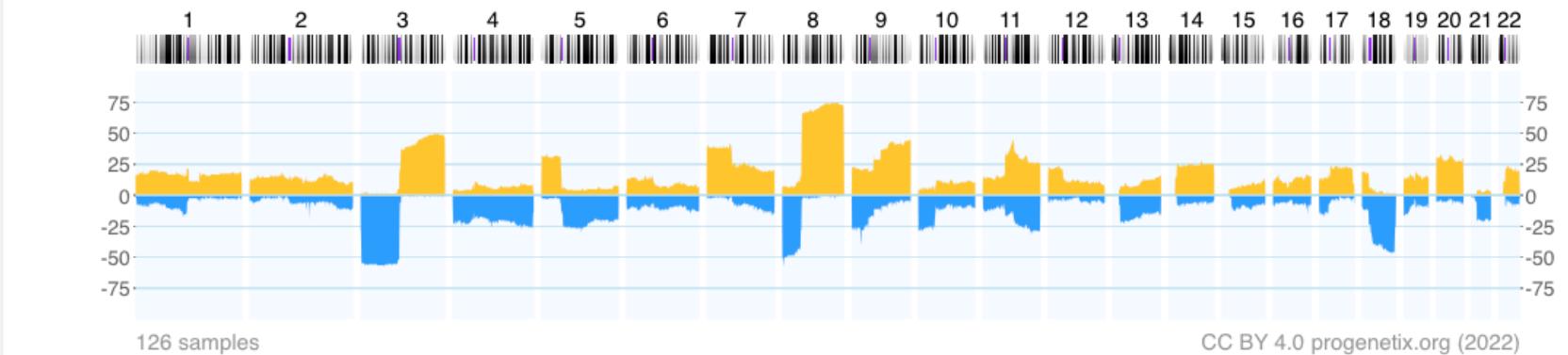
- News
- Downloads & Use Cases
- Sevices & API

Baudisgroup @ UZH

Cancer genome data @ progenetix.org

The Progenetix database provides an overview of mutation data in cancer, with a focus on copy number abnormalities (CNV / CNA), for all types of human malignancies. The data is based on *individual sample data* from currently **142063** samples.

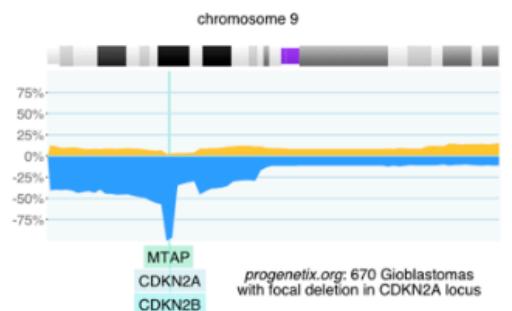
Floor of the Mouth Neoplasm (NCIT:C4401)



[Download SVG](#) | [Go to NCIT:C4401](#) | [Download CNV Frequencies](#)

Example for aggregated CNV data in 126 samples in Floor of the Mouth Neoplasm.
Here the frequency of regional **copy number gains** and **losses** are displayed for all 22 autosomes.

Progenetix Use Cases



Local CNV Frequencies

A typical use case on Progenetix is the search for local copy number aberrations - e.g. involving a gene - and the exploration of cancer types with these CNVs. The [\[Search Page \]](#) provides example use cases for designing queries. Results contain basic statistics as well as visualization and download options.

Cancer CNV Profiles

The progenetix resource contains data of **834** different cancer types (NCI neoplasm classification), mapped to a variety of biological and technical categories. Frequency profiles of regional genomic gains and losses for all categories (diagnostic entity, publication, cohort ...) can be accessed through the [\[Cancer Types \]](#) page with direct visualization and options for sample retrieval and plotting options.

Cancer Genomics Publications

Through the [\[Publications \]](#) page Progenetix provides **4164** annotated references to research articles from cancer genome screening experiments (WGS, WES, aCGH, cCGH). The numbers of analyzed samples and possible availability in the Progenetix sample collection are indicated.

Progenetix in 2022

Cancer Genomics Reference Resource

- largest open resource for curated cancer genome profiles
- >116'000 cancer CNV profiles, from >800 NCIt codes
- majority of data from genomic arrays with ~50% overall from SNP platforms with original data re-processing
- structured diagnostic encodings for NCIt, ICD-O 3, UBERON
- identifier mapping for PMID, GEO, Cellosaurus, TCGA, cBioPortal where appropriate
- core biosample and technical metadata annotations where accessible (TNM, genotypic sex, survival ...)
- publication database and code mapping services

progenetix

Search Samples Modify Query

Assembly: GRCh38 Chro: 9 Start: 21500001-21975098 End: 21967753-22500000
Type: DEL Filters: NCIT:C3058

progenetix

Samples: 668 Variants: 286 Calls: 675

Found Variants (.pgxseg) All Sample Variants (.json) All Sample Variants (.pgxseg) Show Variants in UCSC

UCSC region JSON Response Visualization options

Results Biosamples Biosamples Map Variants

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22

75% 50% 25% 0% -25% -50% -75%

-75% -50% -25% 0% 25% 50% 75%

progenetix: 670 samples

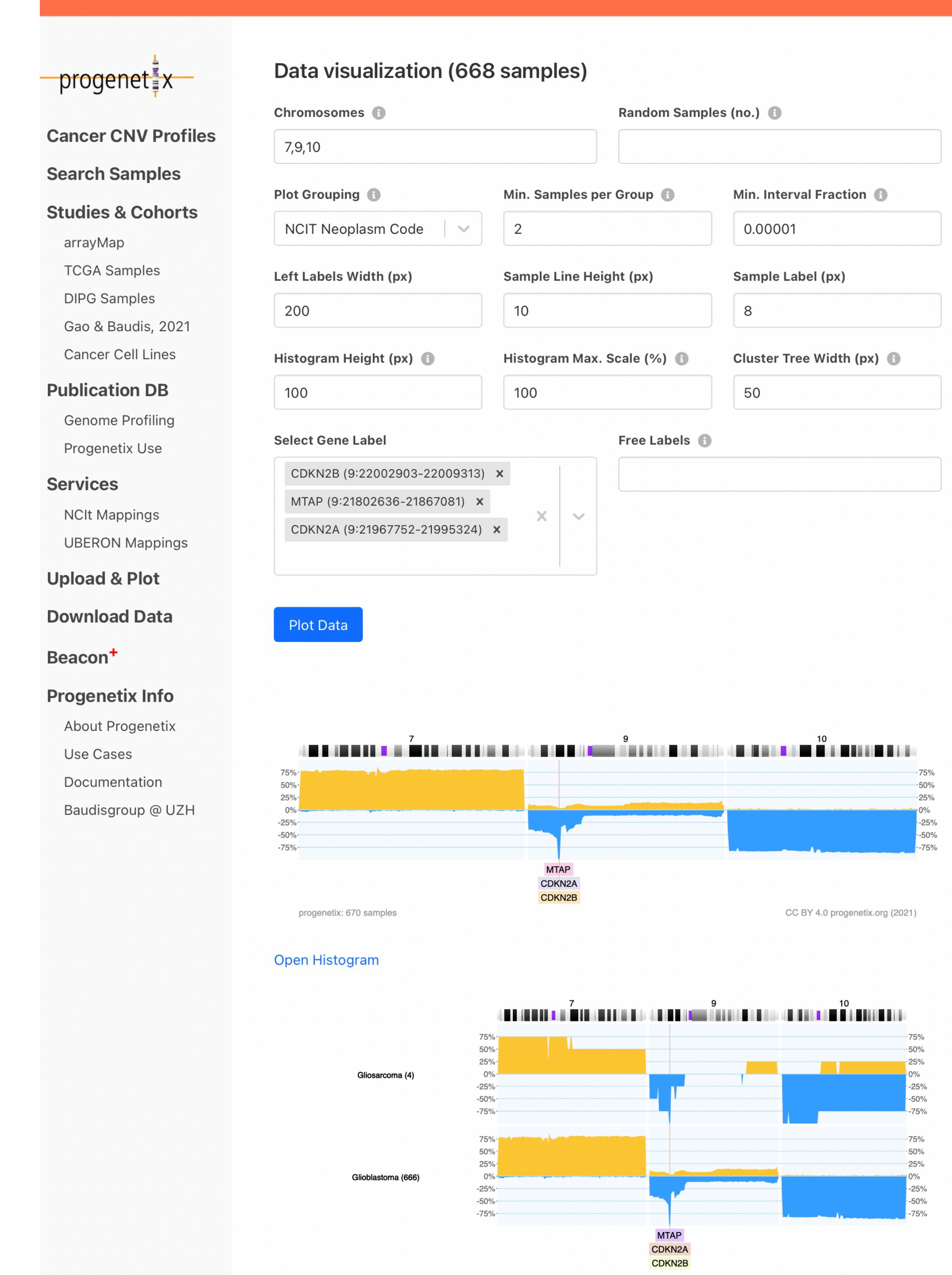
CC BY 4.0 progenetix.org (2021)

Matched Subset Codes	Subset Samples	Matched Samples	Subset Match Frequencies
UBERON:0002021	4	1	0.250
icdot-C71.4	4	1	0.250
icdom-94403	4291	664	0.155
NCIT:C3058	4375	664	0.152
UBERON:0016525	14	2	0.143
icdot-C71.1	14	2	0.143
UBERON:0000955	7068	651	0.092
icdot-C71.9	7066	651	0.092
icdom-94423	84	4	0.048
NCIT:C3796	84	4	0.048
UBERON:0001869	1712	14	0.008
icdot-C71.0	1712	14	0.008

Progenetix in 2022

Cancer Genomics Reference Resource

- largest open resource for curated cancer genome profiles
- >116'000 cancer CNV profiles, from >800 NCIt codes
- majority of data from genomic arrays with ~50% overall from SNP platforms with original data re-processing
- structured diagnostic encodings for NCIt, ICD-O 3, UBERON
- identifier mapping for PMID, GEO, Cellosaurus, TCGA, cBioPortal where appropriate
- core biosample and technical metadata annotations where accessible (TNM, genotypic sex, survival ...)
- publication database and code mapping services



Progenetix in 2022

Cancer Genomics Reference Resource

- contains special data subsets, identified using the "cohorts" concept
 - ▶ TCGA CNV data
 - ▶ 1000Genomes germline CNVs (WGS)
 - ▶ Cancer cell line CNVs with upcoming addition of annotated SNV ... data
 - ▶ cBioPortal studies
 - ▶ ...

TCGA CNV Data

Search Genomic CNV Data from TCGA

This search page accesses the TCGA subset of the Progenetix collection, based on 22142 samples (tumor and references) from The Cancer Genome Atlas project. The results are based upon data generated by the [TCGA Research Network](#). Disease-specific subsets of TCGA data (aka. projects) can be accessed below.

TCGA Cancer samples (pgx:cohort-TCGAcancers)

11090 samples

CC BY 4.0 progenetix.org (2022)

[Download SVG](#) | [Go to pgx:cohort-TCGAcancers](#) | [Download CNV Frequencies](#)

Edit Query

Publication DB

- Genome Profiling
- Progenetix Use

Services

- NCIt Mappings
- UBERON Mappings

Upload & Plot

Beacon⁺

Documentation

- News
- Downloads & Use Cases
- Sevices & API

TCGA Cancer Studies

Filter subsets e.g. by prefix Hierarchy Depth: 2 levels

No Selection

- pgx:TCGA-ACC: TCGA ACC project (180 samples)
- pgx:TCGA-BLCA: TCGA BLCA project (810 samples)
- pgx:TCGA-BRCA: TCGA BRCA project (2219 samples)
- pgx:TCGA-CESC: TCGA CESC project (586 samples)

Progenetix in 2022

Cancer Genomics Reference Resource

- contains special data subsets, identified using the "cohorts" concept
 - ▶ TCGA CNV data
 - ▶ 1000Genomes germline CNVs (WGS)
 - ▶ Cancer cell line CNVs with upcoming addition of annotated SNV ... data
 - ▶ cBioPortal studies
 - ▶ ...



Cancer CNV Profiles
ICD-O Morphologies
ICD-O Organ Sites
Cancer Cell Lines
Clinical Categories

Search Samples

arrayMap
TCGA Samples
1000 Genomes
Reference Samples
DIPG Samples
cBioPortal Studies
Gao & Baudis, 2021

Publication DB
Genome Profiling
Progenetix Use

Services
NCIt Mappings
UBERON Mappings

Upload & Plot

Beacon⁺

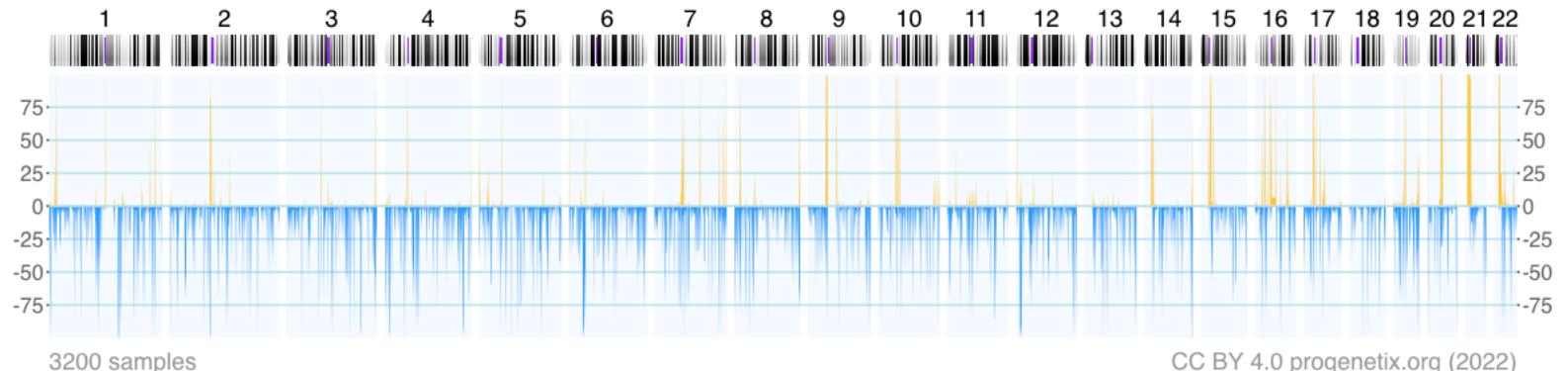
Documentation
News
Downloads & Use Cases
Sevices & API

1000 Genomes Germline CNVs

Search Genomic CNV Data from the Thousand Genomes Project

This search page accesses the reference germline CNV data of 3200 samples from the 1000 Genomes Project. The results are based on the data from the Illumina DRAGEN caller re-analysis of 3200 whole genome sequencing (WGS) samples downloaded from the AWS store of the Illumina-led reanalysis project.

1000 genomes reference samples (pgx:cohort-oneKgenomes)



Download SVG | Go to pgx:cohort-oneKgenomes | Download CNV Frequencies

Please note that the CNV spikes are based on the frequency of occurrence of any CNV in a given 1Mb interval, not on their overlap. Some genome bins may have at least one small CNV in each sample - especially in peri-centromeric regions - and therefore will display with a 100% frequency - although many of those may not overlap.

Search Samples

Range Example

Chromosome (Structural) Variant Type

Start or Position End (Range or Structural Var.)

Reference Base(s) Alternate Base(s)

Progenetix in 2022

Cancer Genomics Reference Resource

- largest open resource for curated cancer genome profiles
- >116'000 cancer CNV profiles, from >800 NCI codes
- majority of data from genomic arrays with ~50% overall from SNP platforms with original data re-processing
- structured diagnostic encodings for NCI, ICD-O 3, UBERON
- identifier mapping for PMID, GEO, Cellosaurus, TCGA, cBioPortal where appropriate
- core biosample and technical metadata annotations where accessible (TNM, genotypic sex, survival ...)
- publication database and code mapping services



[Cancer CNV Profiles](#)

[Search Samples](#)

[Studies & Cohorts](#)

[arrayMap](#)

[TCGA Samples](#)

[DIPG Samples](#)

[Gao & Baudis, 2021](#)

[Cancer Cell Lines](#)

[Publication DB](#)

[Genome Profiling](#)

[Progenetix Use](#)

[Services](#)

[NCI Mappings](#)

[UBERON Mappings](#)

[Upload & Plot](#)

[Download Data](#)

[Beacon⁺](#)

[Progenetix Info](#)

[About Progenetix](#)

Progenetix Publication Collection

The current page lists articles describing whole genome screening (WGS, WES, aCGH, cCGH) experiments in cancer, registered in the Progenetix publication collection. For each publication the table indicates the numbers of samples analysed with a given technology and if sample profiles are available in Progenetix.

Please [contact us](#) to alert us about additional articles you are aware of. The inclusion criteria are described in the documentation [↗](#).

New Oct 2021 You can now directly submit suggestions for matching publications to the [oncopubs](#) repository on [Github](#) [↗](#).

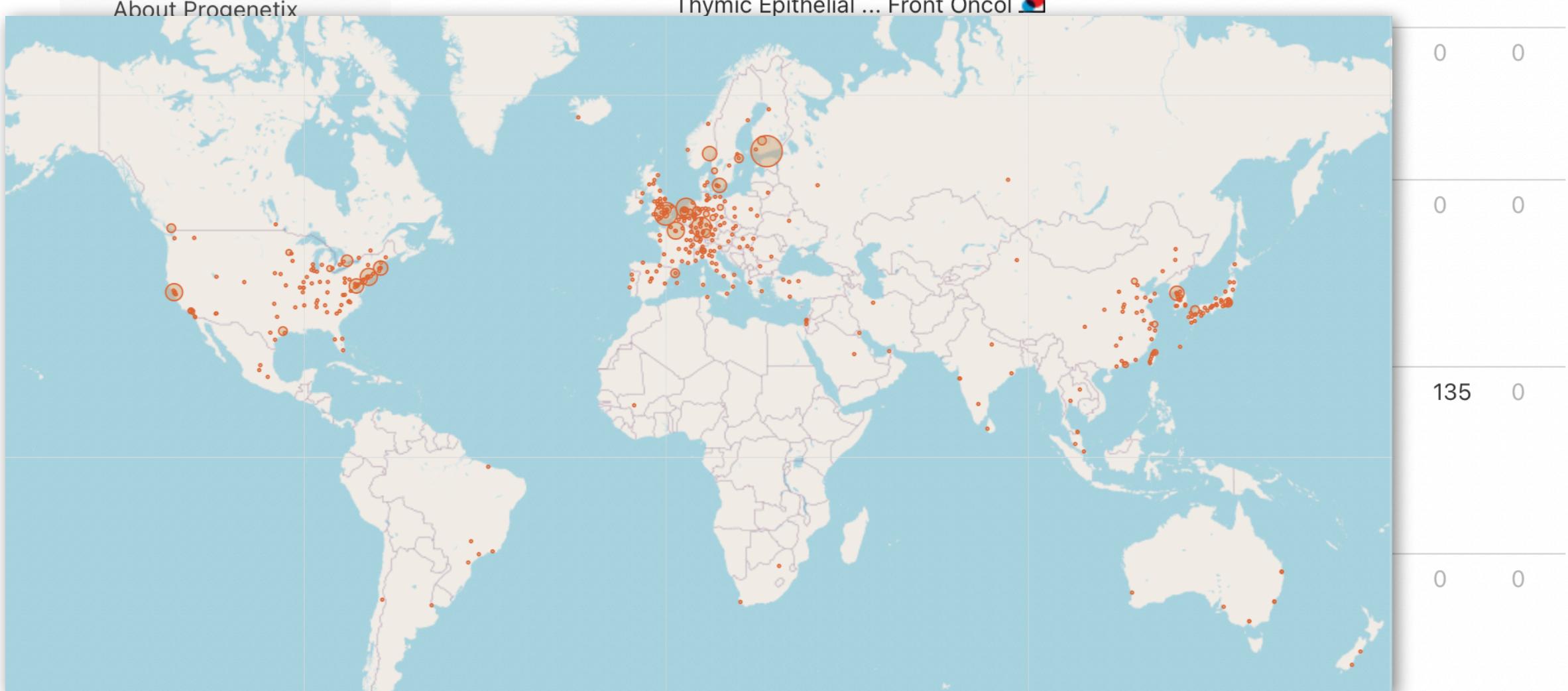
Filter [i](#)

City [i](#)

 Type to search... | [▼](#)

Publications (3349)

id i ▾	Publication	Samples				
		cCGH	aCGH	WES	WGS	pgx
PMID:34604048	Dai J, Jiang M, He K, Wang H, Chen P et al. (2021) DNA Damage Response and Repair Gene Alterations Increase Tumor Mutational Burden and ... <i>Front Oncol</i>	0	0	122	0	0
PMID:34573430	Juhari WKW, Ahmad Amin Noordin KB et al. (2021) Whole-Genome Profiles of Malay Colorectal Cancer Patients with Intact MMR Proteins. ... <i>Genes (Basel)</i>	0	0	0	7	0
PMID:34307137	Xu S, Li X, Zhang H, Zu L, Yang L et al. (2021) Frequent Genetic Alterations and Their Clinical Significance in Patients With Thymic Epithelial ... <i>Front Oncol</i>	0	0	0	123	0



Ontologies and Classifications



Services: Ontologymaps (NCIt)

The **ontologymaps** service provides equivalency mapping between ICD-O and other classification systems, notably NCIt and UBERON. It makes use of the sample-level mappings for NCIT and ICD-O 3 codes developed for the individual samples in the Progenetix collection.

NCIT and ICD-O 3

While NCIT treats diseases as **histologic** and **topographic** described entities (e.g. [NCIT:C7700: Ovarian adenocarcinoma](#)), these two components are represented separately in ICD-O, through the **Morphology** and **Topography** coding arms (e.g. here [8140/3 + C56.9](#)).

More documentation with focus on the API functionality can be found on the [documentation pages](#).

The data of all mappings can be retrieved trough this API call: [{JSON ↗}](#)

Code Selection ⓘ

NCIT:C4337: Mantle Cell Lymphoma X | ▾

Optional: Limit with second selection | ▾

Matching Code Mappings [{JSON ↗}](#)

NCIT:C4337: Mantle Cell Lymphoma	pgx:icdom-96733: Mantle cell lymphoma	pgx:icdot-C77.9: Lymph nodes, NOS
NCIT:C4337: Mantle Cell Lymphoma	pgx:icdom-96733: Mantle cell lymphoma	pgx:icdot-C18.9: large intestine, excl. rectum and rectosigmoid junction
NCIT:C4337: Mantle Cell Lymphoma	pgx:icdom-96733: Mantle cell lymphoma	pgx:icdot-C42.2: Spleen

More than one code groups means that either mappings need refinements (e.g. additional specific NCIT classes for ICD-O T topographies) or you started out with an unspecific ICD-O M class and need to add a second selection.

In Progenetix all cancer diagnoses are coded to both NCIt neoplasm codes and ICD-O 3 Morphology + Topography combinations. The matched mappings are provided as lookup-service since neither an official ICD-O ontology nor such a "disease defined by ICD-O M+T" concept is codified anywhere.

List of filters recognized by different query endpoints

Public Ontologies with CURIE-based syntax

CURIE prefix	Code/Ontology	Examples
NCIT	NCIt Neoplasm ¹	NCIT:C27676
HP	HPO ²	HP:0012209
PMID	NCBI Pubmed ID	PMID:18810378
geo	NCBI Gene Expression Omnibus ³	geo:GPL6801, geo:GSE19399, geo:GSM491153
arrayexpress	EBI ArrayExpress ⁴	arrayexpress:E-MEXP-1008
cellosaurus	Cellosaurus - a knowledge resource on cell lines ⁵	cellosaurus:CVCL_1650
UBERON	Uberon Anatomical Ontology ⁶	UBERON:0000992
cBioPortal	cBioPortal ⁹	cBioPortal:msk_impact_2017

Private filters

Since some classifications cannot directly be referenced, and in accordance with the upcoming Beacon v2 concept of "private filters", Progenetix uses additionally a set of structured non-CURIE identifiers.

For terms with a `pgx` prefix, the [identifiers.org resolver](#) will

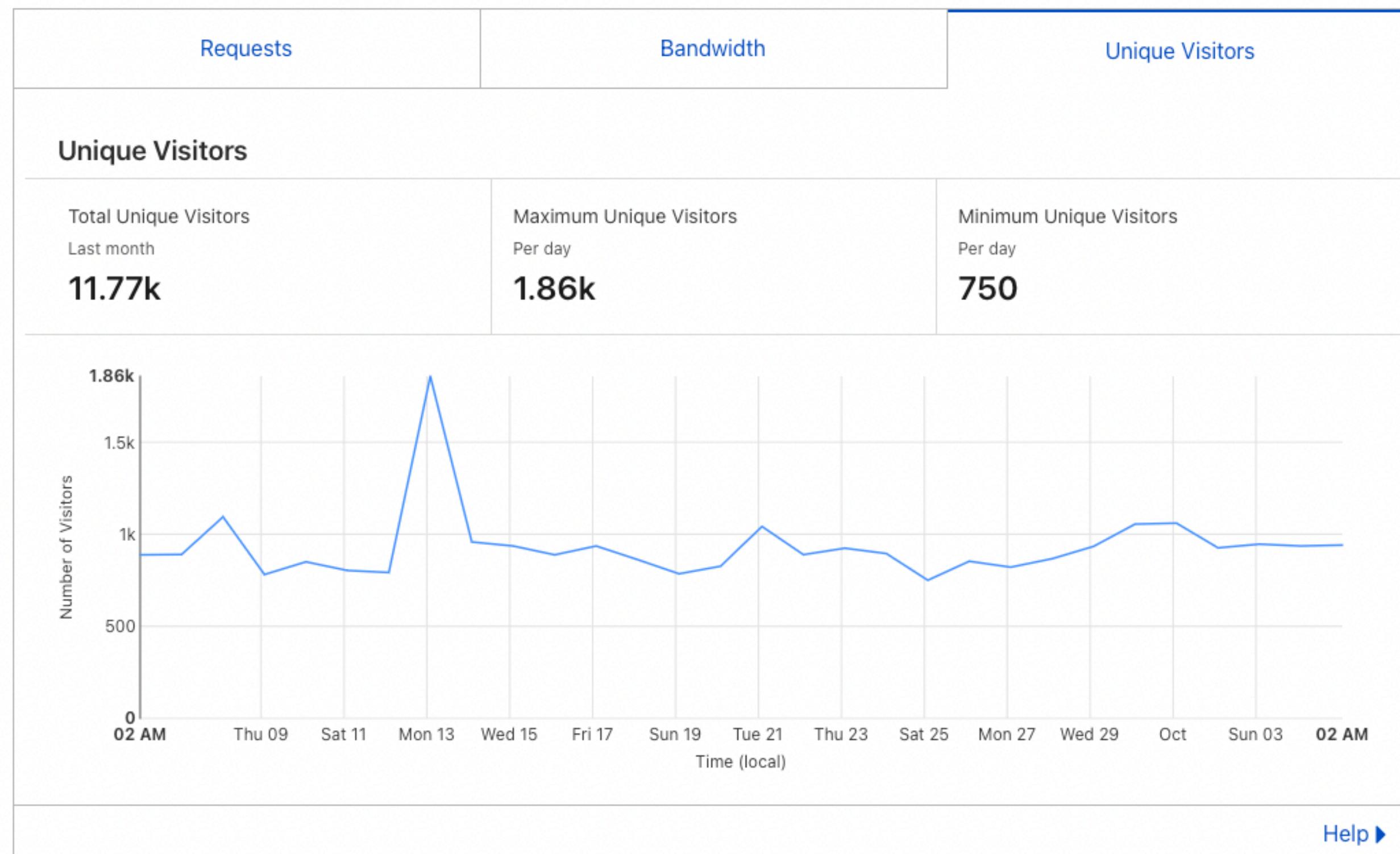
Filter prefix / local part	Code/Ontology	Example
pgx:icdom...	ICD-O 3 ⁷ Morphologies (Progenetix)	pgx:icdom-81703
pgx:icdot...	ICD-O 3 ⁷ Topographies(Progenetix)	pgx:icdot-C04.9
TCGA	The Cancer Genome Atlas (Progenetix) ⁸	TCGA-000002fc-53a0-420e-b2aa-a40a358bba37
pgx:pgxcohort...	Progenetix cohorts ¹⁰	pgx:pgxcohort-arraymap

Progenetix in 2022

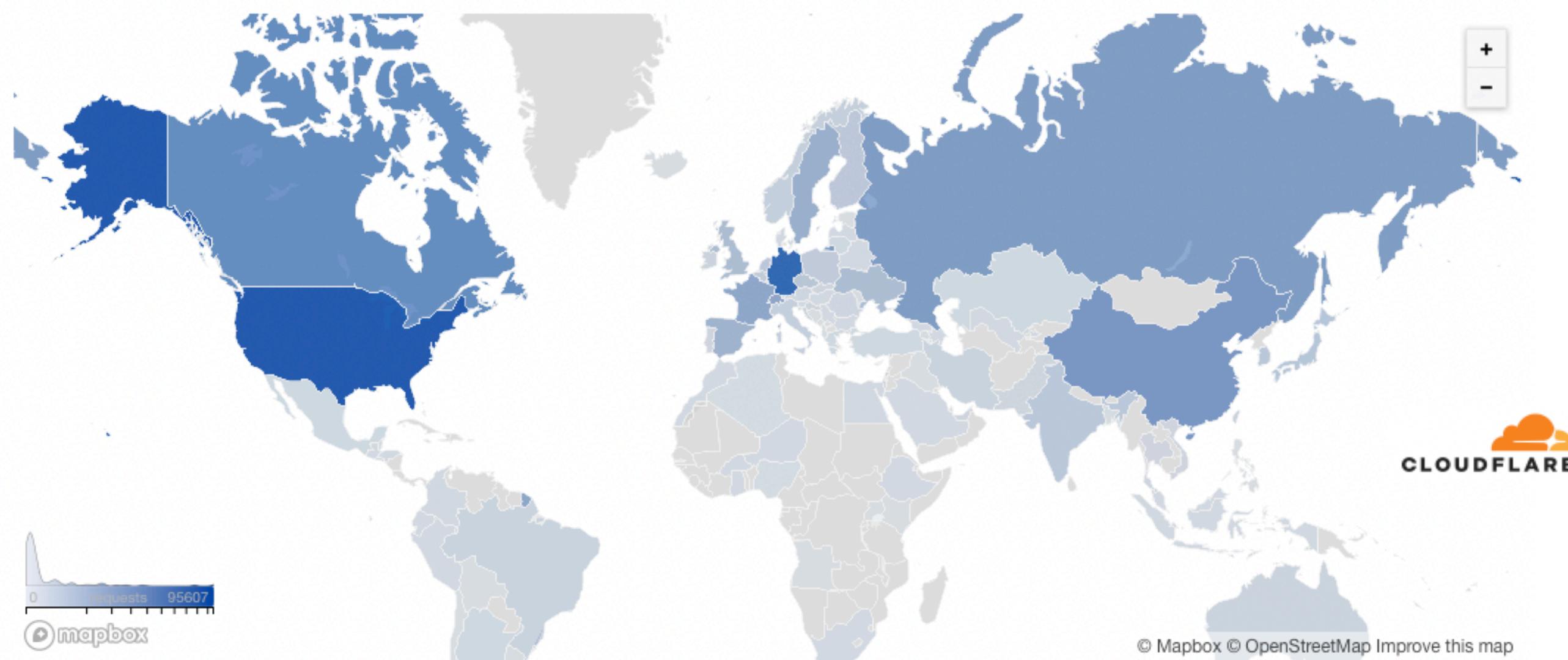
Cancer Genomics Reference Resource

- largest open resource for curated cancer genome profiles
- >116'000 cancer CNV profiles, from >800 NCI codes
- majority of data from genomic arrays with ~50% overall from SNP platforms with original data re-processing
- structured diagnostic encodings for NCI, ICD-O 3, UBERON
- identifier mapping for PMID, GEO, Cellosaurus, TCGA, cBioPortal where appropriate
- core biosample and technical metadata annotations where accessible (TNM, genotypic sex, survival ...)
- publication database and code mapping services

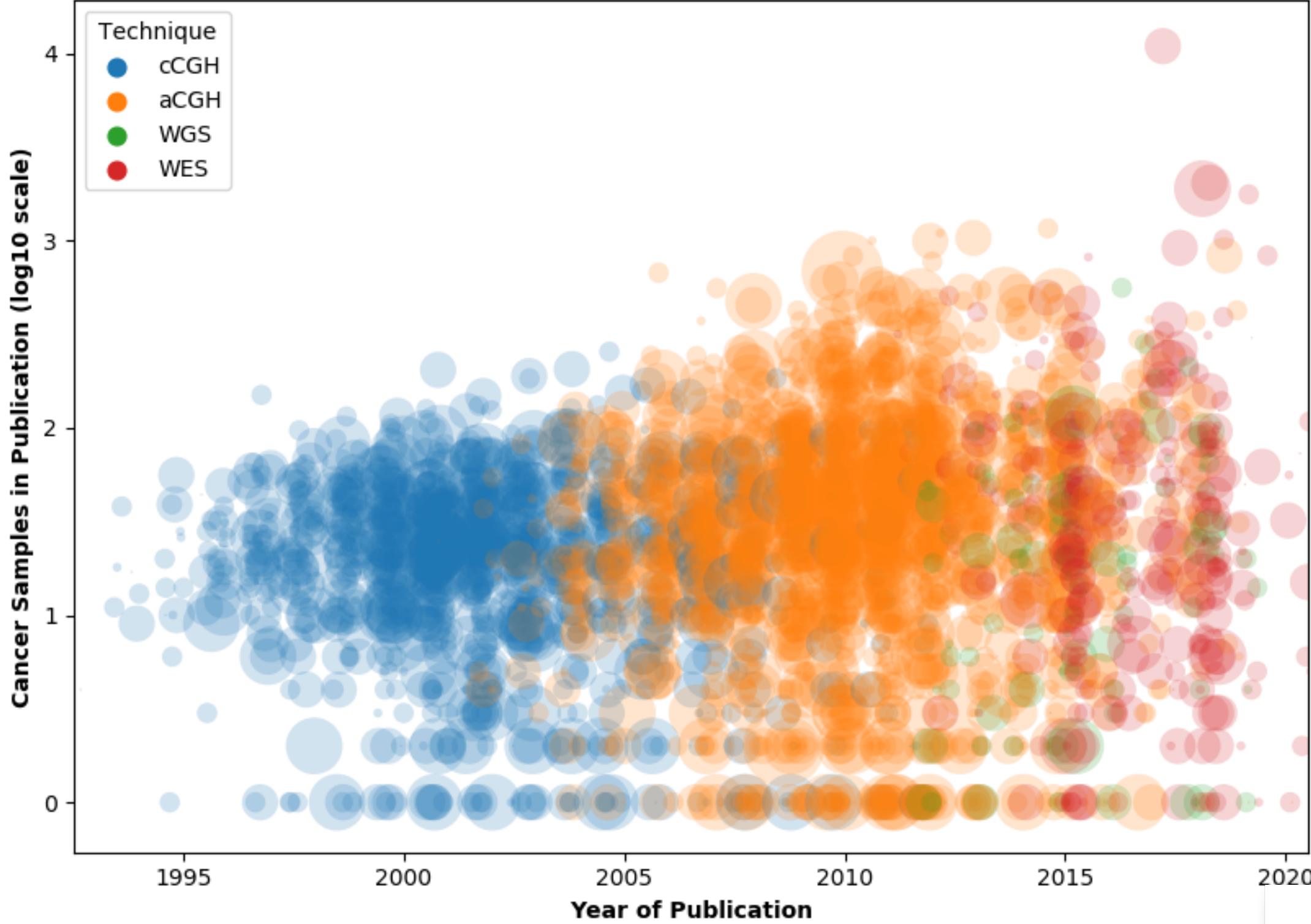
Web Traffic



Web Traffic Requests by Country



Number of tumor samples for each publication across the years



Map of the geographic distribution (by first author affiliation) of the 104'543 genomic array, 36'766 chromosomal CGH and 15'409 whole genome/exome based cancer genome datasets.

The numbers are derived from the 3'240 publications registered in the Progenetix database.



Progenetix Publication Collection

The current page lists articles describing whole genome screening (WGS, WES, aCGH, cCGH) experiments in cancer, registered in the Progenetix publication collection. For each publication the table indicates the numbers of samples analysed with a given technology and if sample profiles are available in Progenetix.

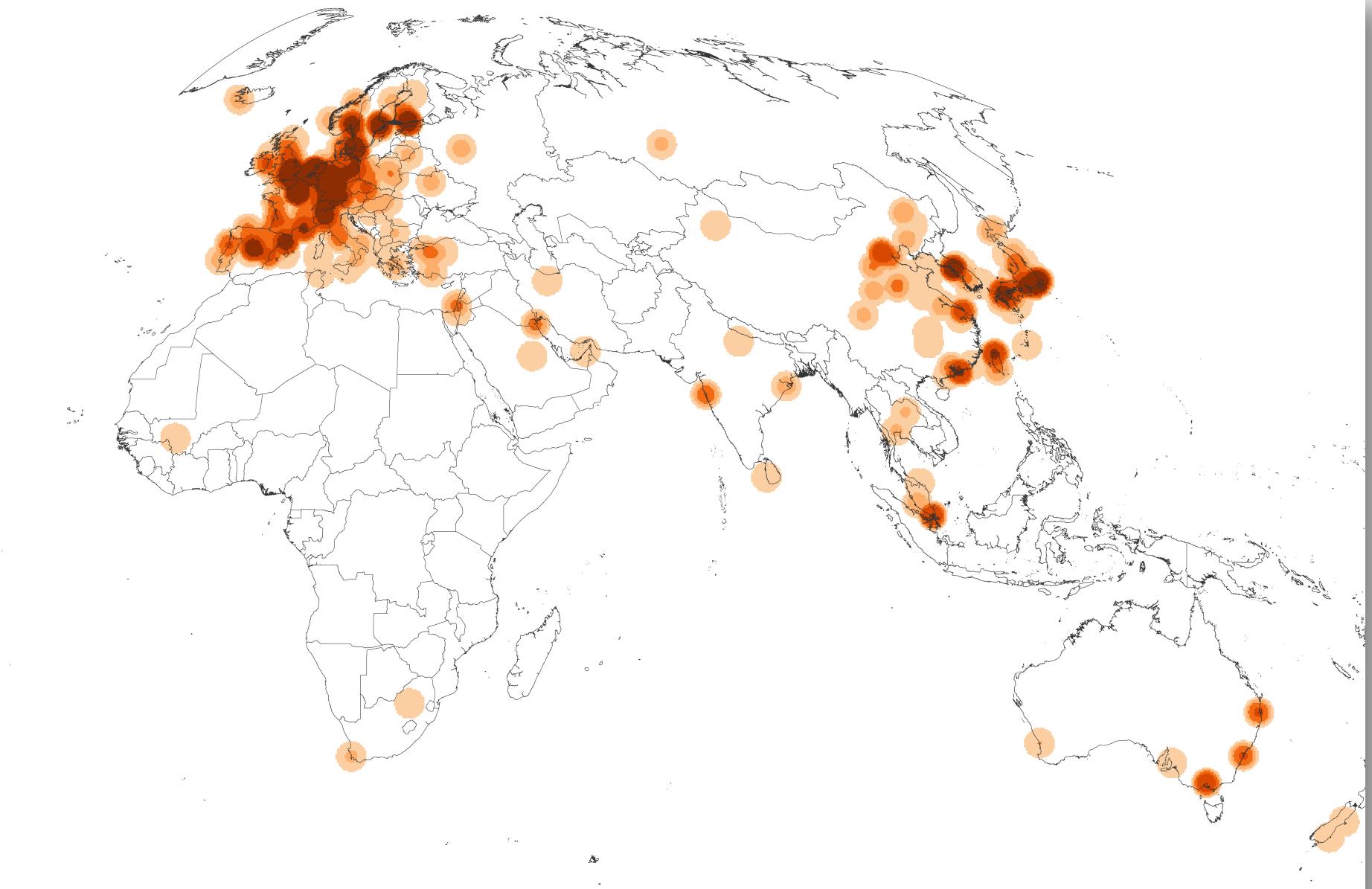
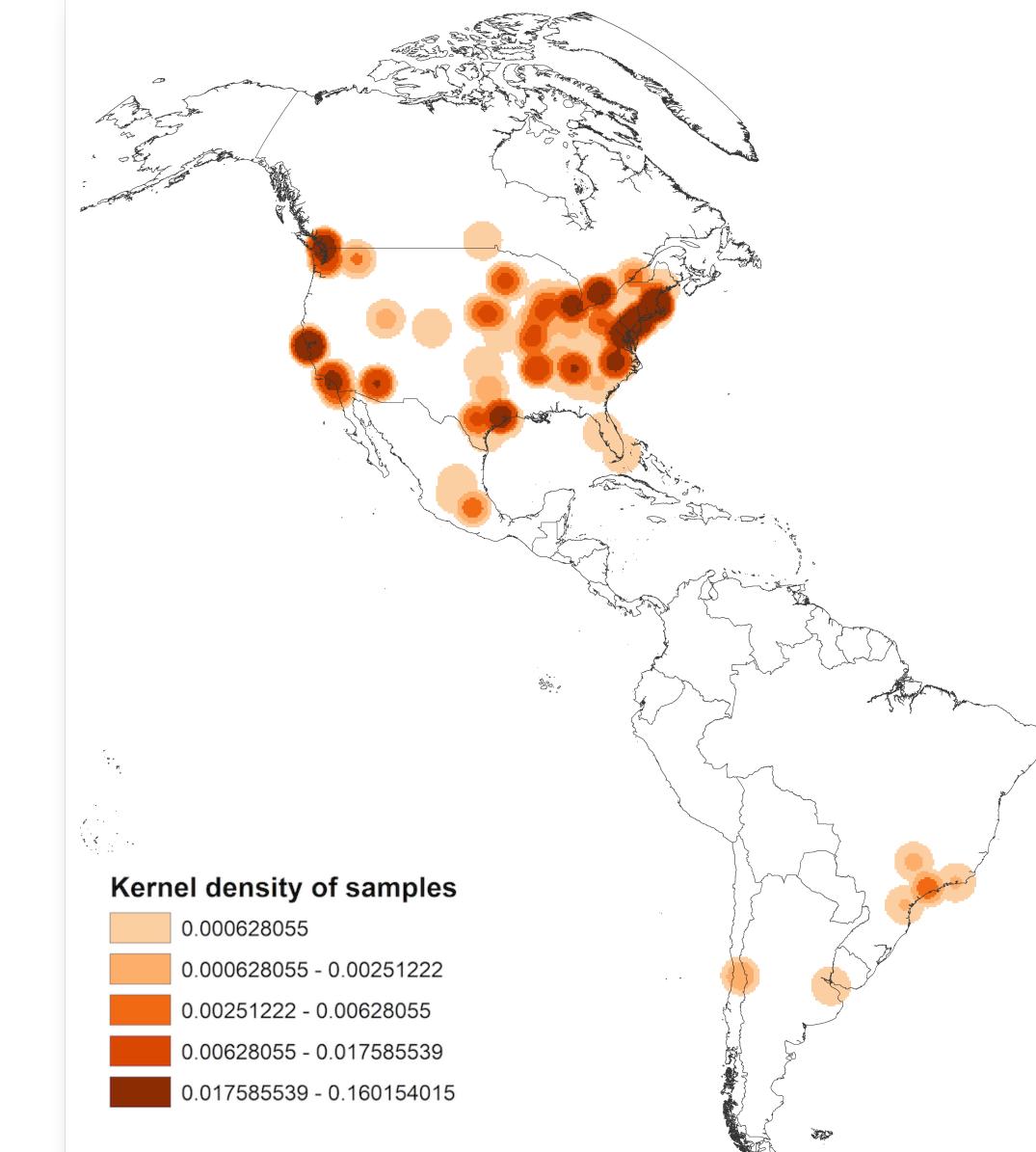
Please [contact us](#) to alert us about additional articles you are aware of. The inclusion criteria are described in the documentation [🔗](#).

Filter [i](#) City [i](#)

 Type to search... [▼](#)

Publications (3324)

id i ▾	Publication	cCGH	aCGH	WES	WGS	pgx
PMID:34103027	Peng G, Chai H, Ji W, Lu Y, Wu S et al. (2021) Correlating genomic copy number alterations with clinicopathologic findings in 75 cases of ... <i>BMC Med Genomics</i>	0	79	0	0	0
PMID:34059130	Tsui DWY, Cheng ML, Shady M, Yang JL et al. (2021) Tumor fraction-guided cell-free DNA profiling in metastatic solid tumor patients. ... <i>bioRxiv</i>	0	0	5	113	0



GA4GH Genome Beacons
A Driver Project of the Global Alliance for Genomics and Health GA4GH and supported through ELIXIR

News
Specification & Roadmap
Beacon Networks
Events
Examples, Guides & FAQ
Contributors & Teams
Contacts
Meeting Minutes

Related Sites
ELIXIR BeaconNetwork
Beacon @ ELIXIR
GA4GH
beacon-network.org
Beacon+
GA4GH::SchemaBlocks
GA4GH::Discovery

Github Projects
Beacon API and Tools
SchemaBlocks

Tags
CNV EB FAQ SV VCF beacon clinical
code compliance contacts definitions
developers development events filters
minutes network press proposal
queries releases roadmap
specification teams v2 versions
website

beacon-project.io



Baudisgroup @ UZH

Ni Ai

Michael Baudis

Haoyang Cai

Paula Carrio Cordo

Bo Gao

Qingyao Huang

Saumya Gupta

Nitin Kumar

Rahel Paloots

Ziying Yang

Hangjia Zhao

Pierre-Henri Toussaint
Sofia Pfund



Beacon Protocol for Genomic Data Sharing
Beacons provide discovery services for genomic data using the Beacon API developed by the Global Alliance for Genomics and Health (GA4GH). The Beacon protocol itself is a standard for genomic data discovery. To provide a framework for publishing genomic data, the Beacon protocol defines a set of rules for publishing genomic data.

Samples
Request Allele Request Example

This example shows the query for CNV deletion variants overlapping the CDKN2A gene's coding region. The query is limited to hits that overlap with at least a single base, but limited to "focal" hits (here i.e. $\leq \sim 2\text{Mbp}$ in size). The query is against the Beacon API and can be modified e.g. through changing the assembly, the gene, or the variant type.

This query type is for copy number queries ("variantCNVRequest"). It uses the "start" and "end" positions to capture a set of similar variants.

Start Position: 21000001-21975098
End Position: 21967753-23000000
Classification(s): C3058: Glioblastoma (2119)

City: Select...
21000001 21975098
21967753 23000000

Query Beacon

beacon.progenetix.org/beaconPlus/

ELIXIR h-CNV
Christophe Béroux
David Salgado
many more ...

Beacon API Leads

Jordi Rambla

Anthony Brooks

Discovery WS

Michael Baudis (Beacon)

Marc Fiume (Networks)

ga4gh-beacon / beacon-v2 Public

Code Issues 12 Pull requests 4 Discussions Actions Security Insights

main Go to file Add file Code

mbaudis Update ComplexValue.md ... 3 days ago 317

.github/workflows remove PDF; update variant table 3 months ago

bin Update README.md 7 days ago

docs Update ComplexValue.md 3 days ago

StringTerms description refinement 6 days ago

commonDefinitions.json 3 days ago

structuring intro pages 3 months ago

Revert "Update .gitignore" 3 days ago

repository changes moved to docs page 3 months ago

chatting 3 months ago

Initial commit 4 months ago

Update README.md 7 days ago

Creating the REST page last month

Switch to mermaid2 plugin 3 months ago

Unified repository for Beacon v2 Code & Documentation

Description

This repository is a unified repository representing the different parts of the Beacon API:

- framework
- models
- Beacon v2 Documentation
 - authoritative source already in this repository [/docs](#)
 - rendered version through [here](#) (alternative address is [docs.genomebeacons.org](#))

github.com/ga4gh-beacon/

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" _(___/__