

BIO390: Introduction to Bioinformatics

Lecture II: What is Bioinformatics?



Course Information BIO390

- Tuesdays at 08:00; 2x45min
- 13 presentations by different lecturers
- (unchecked) homework / preparation exercises w/ focus on test topics
- course language is English
- course slides may/should be made available through the website
- written exam at end of course (== 14th course - December 17)
- Organizer:

Prof. Dr. Michael Baudis

Department of Molecular Life Sciences (IMLS)

University of Zurich Campus Irchel, Y-11F-13

CH-8057 Zurich

email michael@baud.is

web info.baudisgroup.org

**Please use the website for additional
course information**

<https://compbiozurich.org/courses/UZH-BIO390/>



BIO390 - Introduction to Bioinformatics

Summary

The handling and analysis of biological data using computational methods has become an essential part in most areas of biology. In this lecture, students will be introduced to the use of bioinformatics tools and methods in different topics, such as molecular resources and databases, standards and ontologies, sequence and high performance genome analysis, biological networks, molecular dynamics, proteomics, evolutionary biology and gene regulation. Additionally, the use of low level tools (e.g. Programming and scripting languages) and specialized applications will be demonstrated. Another topic will be the visualization of quantitative and qualitative biological data and analysis results.

Practical Information

Requirements

The *introduction to Bioinformatics* is a series of lectures aimed at students w/ a medium to advanced undergraduate level in Life Sciences. Participants are expected to be *knowledgeable in the basic concepts of molecular biology and genetics*, but also to have some *basic understanding in statistics and concepts of programming*, if not practical experience (*i.e.* have attended introductory courses, done some data analyses in R or Python etc.). Experience with common platforms used for shared code/document management (e.g. Gitlab/Github...) is helpful but not strictly required.

Schedule & Notes

- Autumn semesters
- 1 x 2h / week
- Tue 08:00-09:45
- UZH Irchel campus, **Y-03G-85**
- **OLAT** - but not much there...
- No lecture recordings - we do **not record** the lectures since HS23 (regular attendance is expected) *but* there might be still 2022 [lecture recordings](#) available
- Course language is English

Syllabus

Next: What is Bioinformatics? Introduction and Resources

BIO390 UZH HS24 - INTRODUCTION TO BIOINFORMATICS
08:00-09:45 @ UZH IRCHEL Y03-G-85

📅 September 17, 2024

Michael Baudis

This year happening at the second lecture day, the "What is Bioinformatics? Introduction and Resources" provides a general introduction into the field and a description of the lecture topics, timeline and procedures.

Topics covered in the lecture are e.g.:

[→ Continue reading](#)

Upcoming: Statistical Bioinformatics

BIO390 UZH HS24 - INTRODUCTION TO BIOINFORMATICS
08:00-09:45 @ UZH IRCHEL Y03-G-85

📅 September 24, 2024

Mark Robinson

Today's topic is the use of statistical methods in the analysis of biological datasets, with examples from high-throughput (sequencing and array) technologies and single cell analyses.

[→ Continue reading](#)

Upcoming: Biological Sequence Informatics

BIO390 UZH HS24 - INTRODUCTION TO BIOINFORMATICS
08:00-09:45 @ UZH IRCHEL Y03-G-85

📅 October 01, 2024

Christian von Mering

The analysis of biological sequences - primarily DNA, RNA and protein sequences - constitutes one of earliest and core areas of bioinformatics. This lecture introduces principles and examples of bioinformatic sequence analyses and inter-sequence comparisons. [→ Continue reading](#)

BIO390: Course Schedule

- 2024-09-17: Michael Baudis - What is Bioinformatics? Introduction and Resources
- 2024-09-24: Mark Robinson - Statistical Bioinformatics
- 2024-10-01: Christian von Mering - Sequence Bioinformatics
- 2024-10-08: Valentina Boeva (ETHZ) - Machine Learning for Biological Use Cases
- 2024-10-15: Izaskun Mallona - Regulatory Genomics and Epigenomics
- 2024-10-22: Shinichi Sunagawa (ETHZ) - Metagenomics
- 2024-10-29: Katja Baerenfaller (SIAF) - Proteomics
- 2024-11-05: Patrick Ruch - Text mining & Search Tools
- 2024-11-07: Andreas Wagner - Biological Networks
- 2024-11-19: Ahmad Aghaebrahimian (ZHAW) - Semantic Web
- 2024-11-26: Qingyao Huang - Building Biological Information Resources
- 2024-12-03: Valérie Barbie (SIB) - Clinical Bioinformatics
- 2024-12-10: Michael Baudis - Genome Data & Privacy | Feedback
- 2024-12-17: Exam (Multiple Choice)



Some Recommended Books

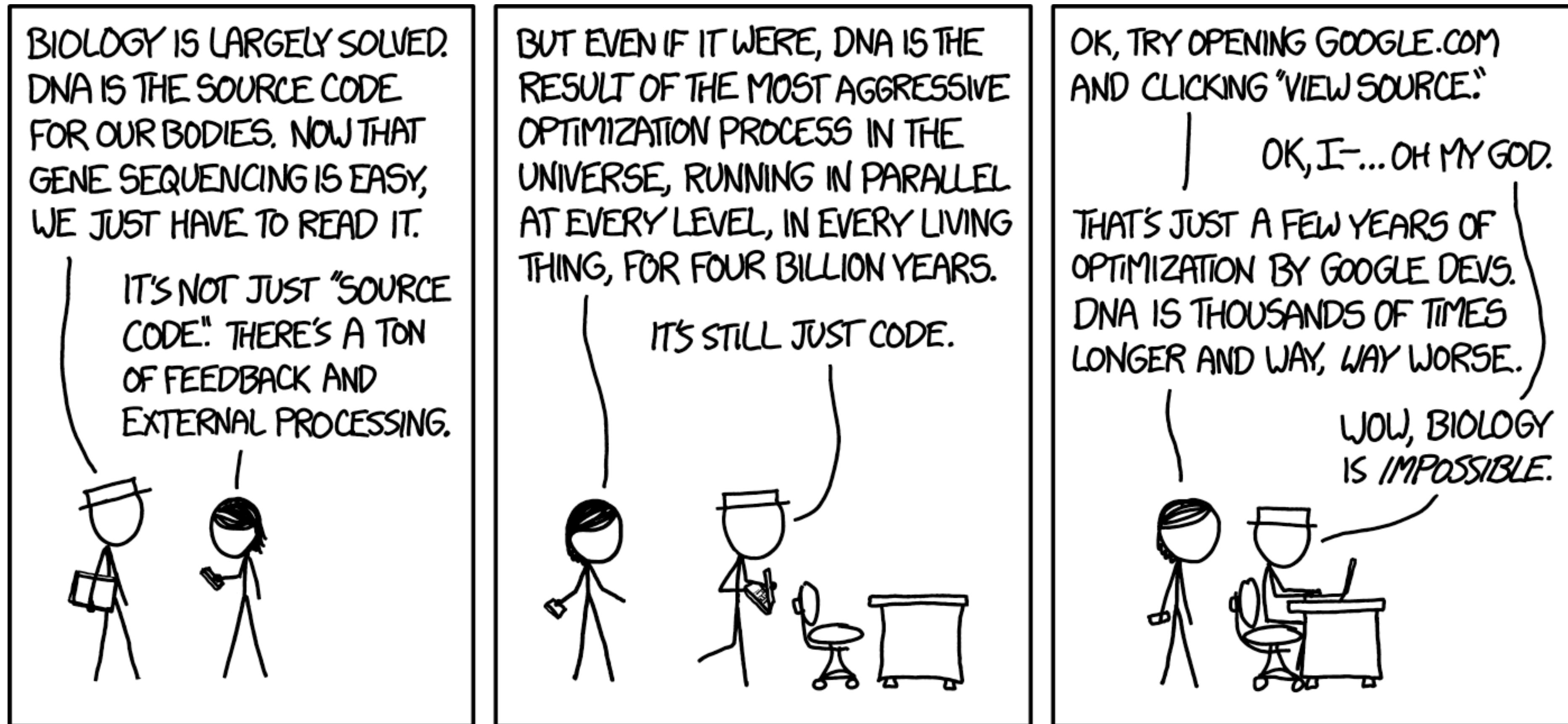
- Anna Tramontano: Introduction to Bioinformatics
- Susan Holmes and Wolfgang Huber: Statistics for Biology
- Robert Gentleman: R Programming for Bioinformatics
- John Maindonald & W. John Braun: Data Analysis and Graphics Using R
- Andy Hector: The New Statistics with R
- Neil C. Jones & Pavel A. Pevzner: Bioinformatics Algorithms
- Edward Tufte: The Visual Display of Quantitative Information (& other works by Tufte)



Why Bioinformatics?

- **hypotheses** are the basis of biological experiments
- biological experiments produce **data**, the quantitative and/or qualitative read-outs of experiments
- both quantitative as well as qualitative data need to be **processed** for
 - **statistical significance**
 - **categorisation**
 - **communication**
- many datatypes are **beyond** the proverbial "**intuitive** understanding"
- analysis of data **confirms** or **refutes** initial **hypotheses** - or requires new hypotheses and new data

Biology is *impossibly complex* - But bioinformatics might help



So, What is Bioinformatics?

- Bioinformatics is "the science that uses the instruments of informatics to analyze biological data in order to formulate hypotheses about life."
(Anna Tramontano)

What is Bioinformatics?



- Bioinformatics is "the **science** that uses the instruments of informatics to analyze biological data in order to formulate hypotheses about life."
(Anna Tramontano)

a : knowledge or a **system of knowledge** covering general truths or the operation of general laws especially as obtained and tested through **scientific method**

b : such knowledge or such a system of knowledge concerned with the physical world and its **phenomena** : **NATURAL SCIENCE**



What is Bioinformatics?

Bioinformatics **uses** informatics tools for analyses

- ✦ Bioinformatics is "the science that **uses** the instruments of informatics to analyze biological data in order to formulate hypotheses about life."
(Anna Tramontano)
- ➔ **software** (programming languages, statistics & visualisation, program and web APIs, databases, hardware drivers)
- ➔ **hardware** (HPC, data storage, signal measurement & processing)
- ➔ **algorithms** (modeling, encryption...)

What is Bioinformatics?

Bioinformatics **develops** informatics tools for analyses

- ✦ Bioinformatics is "the science that uses the **instruments of informatics** to analyze biological data in order to formulate hypotheses about life."
(Anna Tramontano)
- ➔ **software** (statistics & visualisation packages, program and web APIs, file formats)
- ➔ **hardware** (drivers and procedures...)
- ➔ **algorithms** (modeling, encryption...)

What is Bioinformatics?

biological data

- Bioinformatics is "the science that uses the instruments of informatics to analyze **biological data** in order to formulate hypotheses about life." (Anna Tramontano)

sequences, graphs, high-dimensional data, spatial/geometric information, scalar and vector fields, patterns, constraints, images, models, prose, declarative knowledge ... *

What is Bioinformatics?



Bioinformatics **analyzes**

- Bioinformatics is "the science that uses the instruments of informatics to **analyze** biological data in order to formulate hypotheses about life."
(Anna Tramontano)

1 : to study or determine the nature and relationship of the parts of (something) by **analysis**

What is Bioinformatics?




- Bioinformatics is "the science that uses the instruments of informatics to analyze biological data in order to **formulate hypotheses** about life." (Anna Tramontano)

b : an interpretation of a practical situation or condition taken as the ground for action

What is Bioinformatics?

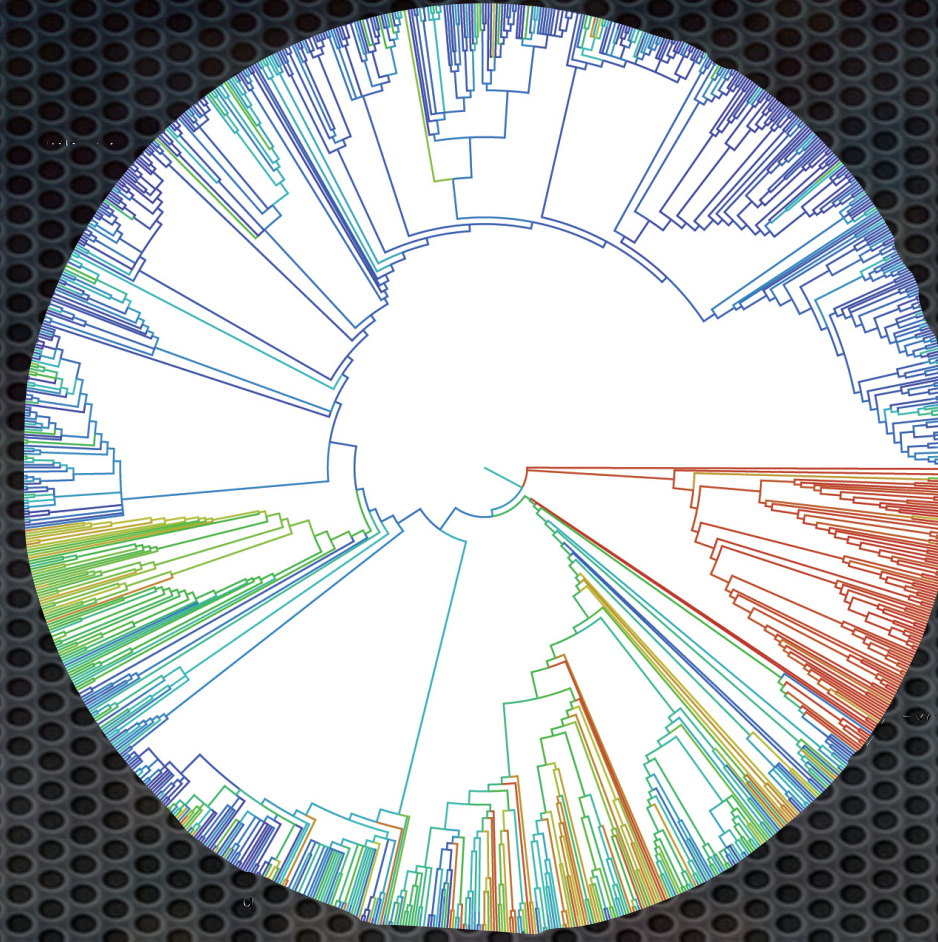


hypothesis 
noun | hy·poth·e·sis | \hī-'pä-thē-sēs\
Popularity: Top 1% of lookups

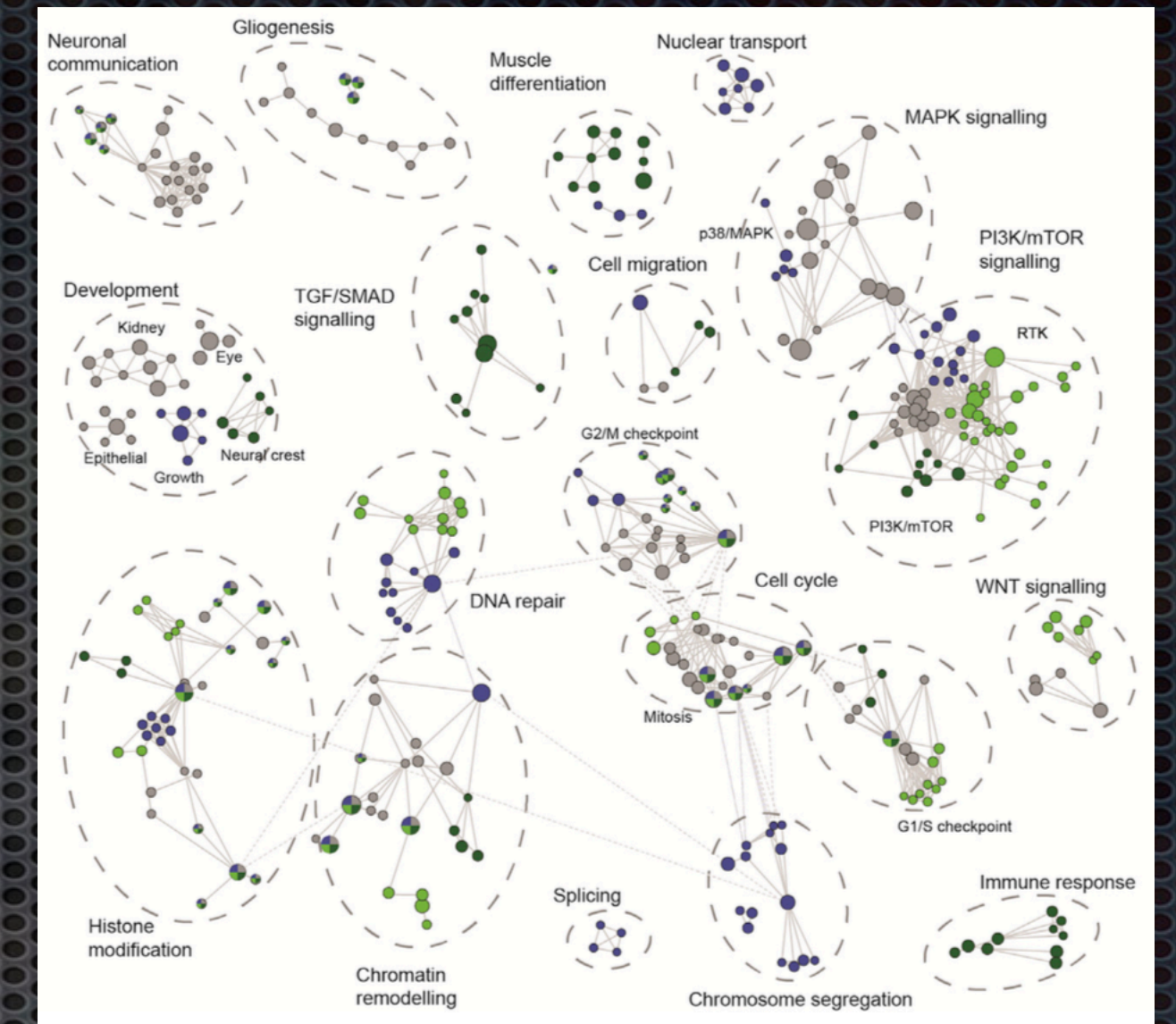


- Bioinformatics is "the science that uses the instruments of informatics to analyze biological data in order to **formulate hypotheses** about life." (Anna Tramontano)

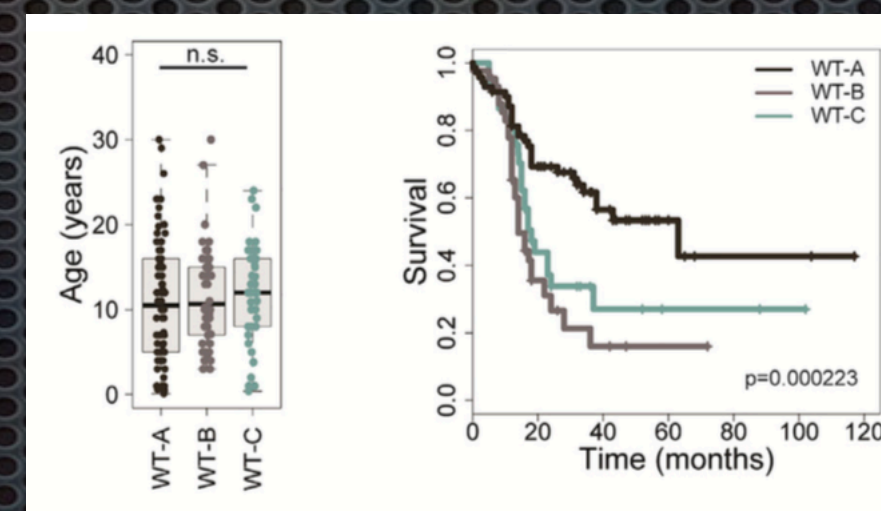
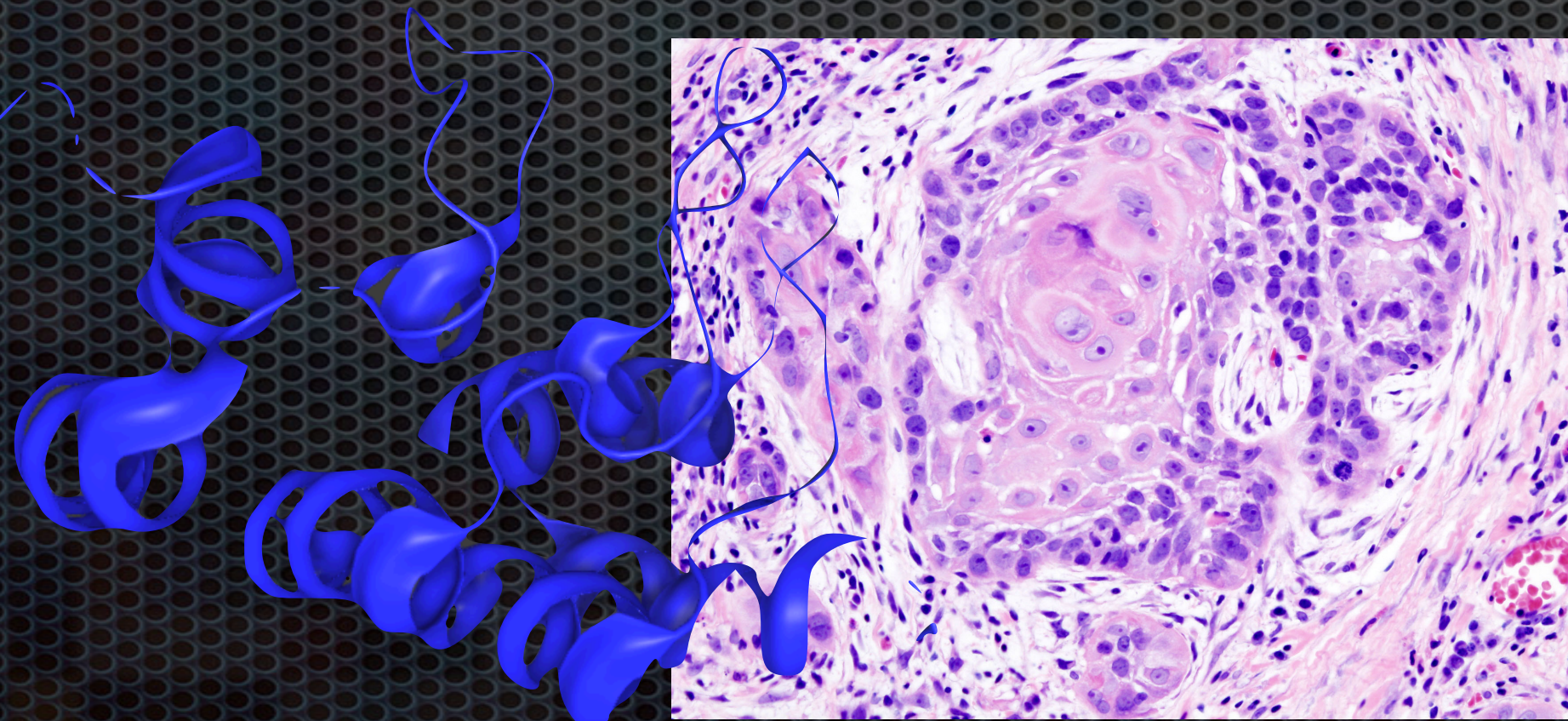
b : an interpretation of a practical situation or condition taken as the ground for action



42



- Bioinformatics is "the science that uses the instruments of informatics to analyze biological data in order to formulate hypotheses about **life**." (Anna Tramontano)

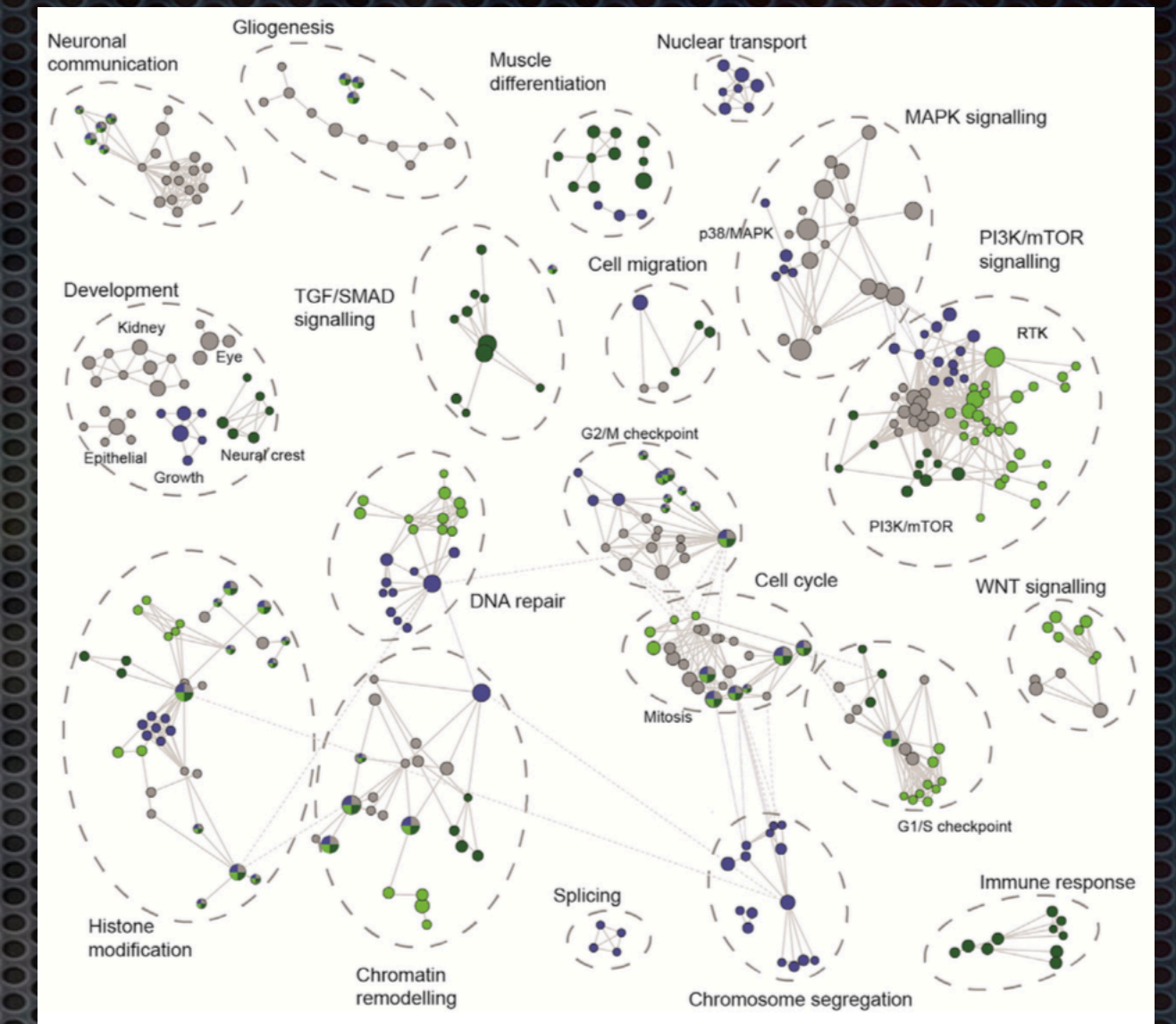
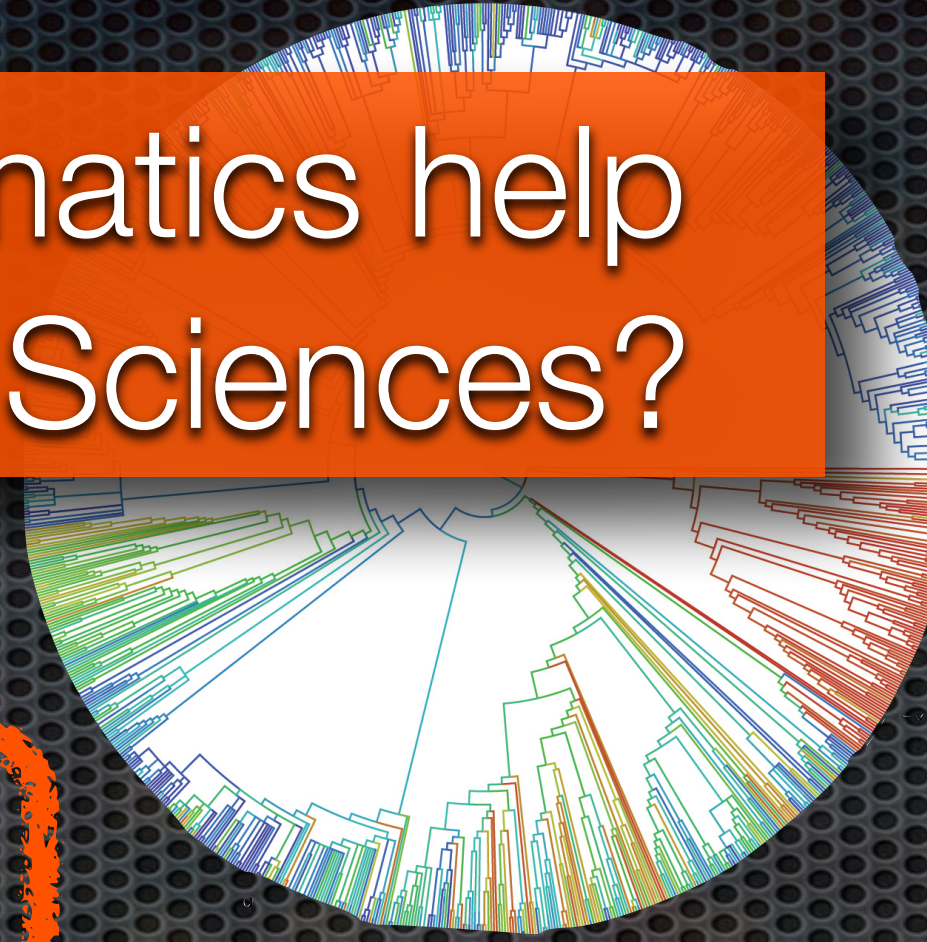


Sources: nextprot | opentreeoflife | wikipedia | MacKay et al., Cancer Cell (2017) | original photos

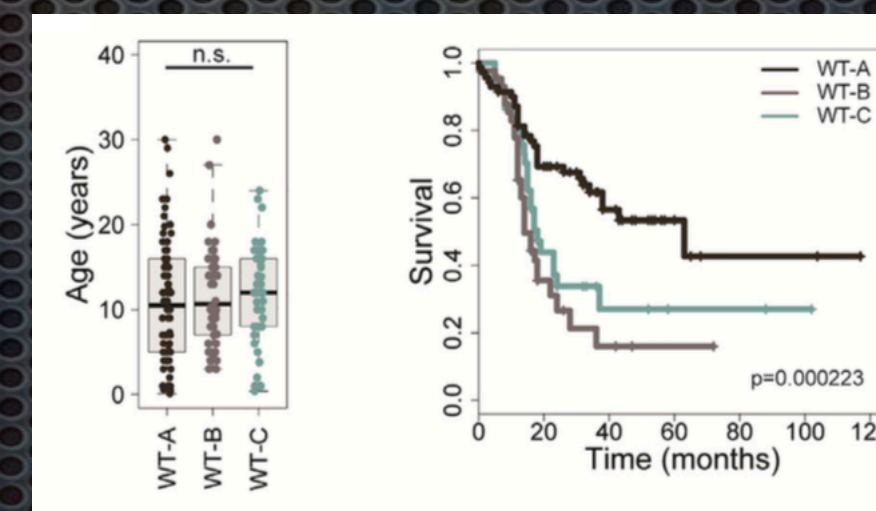
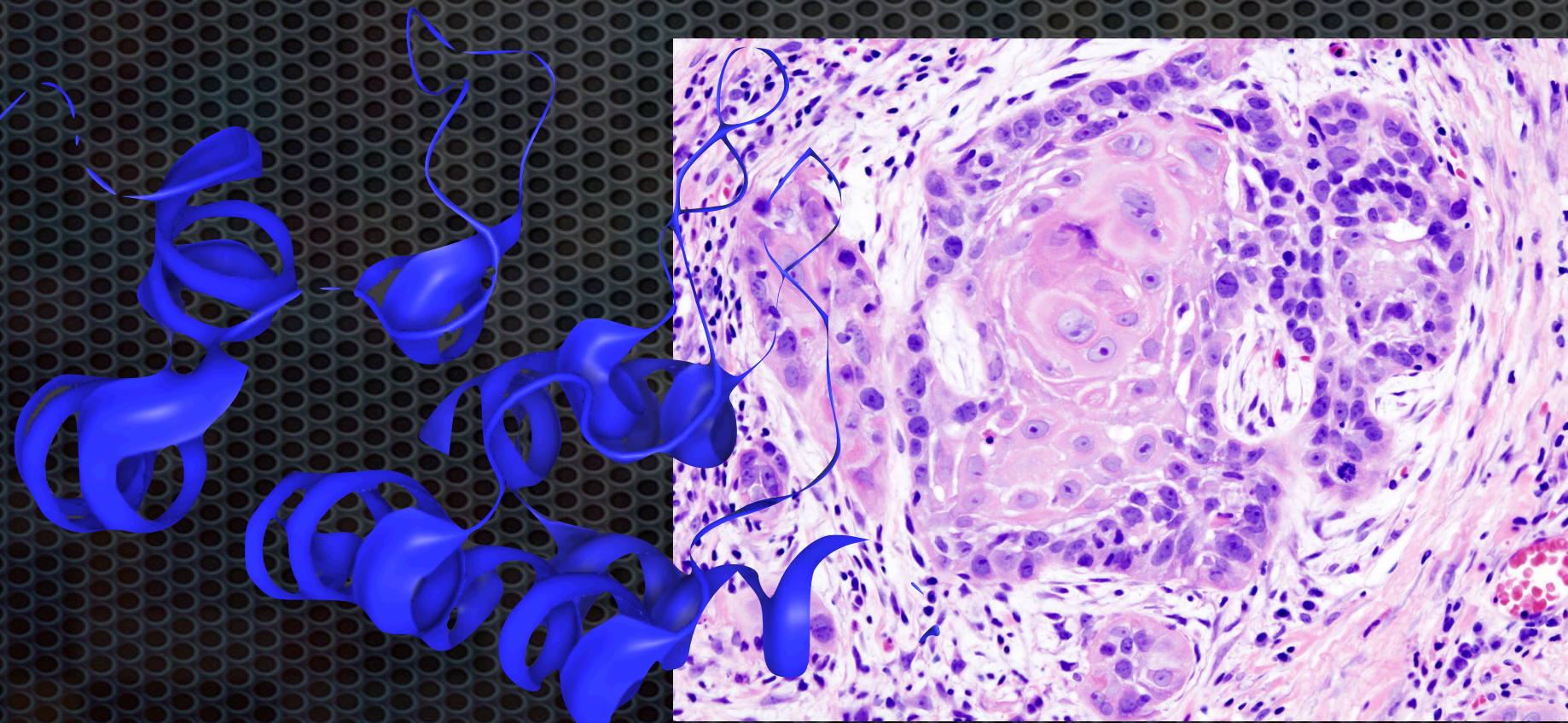


How can Bioinformatics help with the **42** of Life Sciences?

42

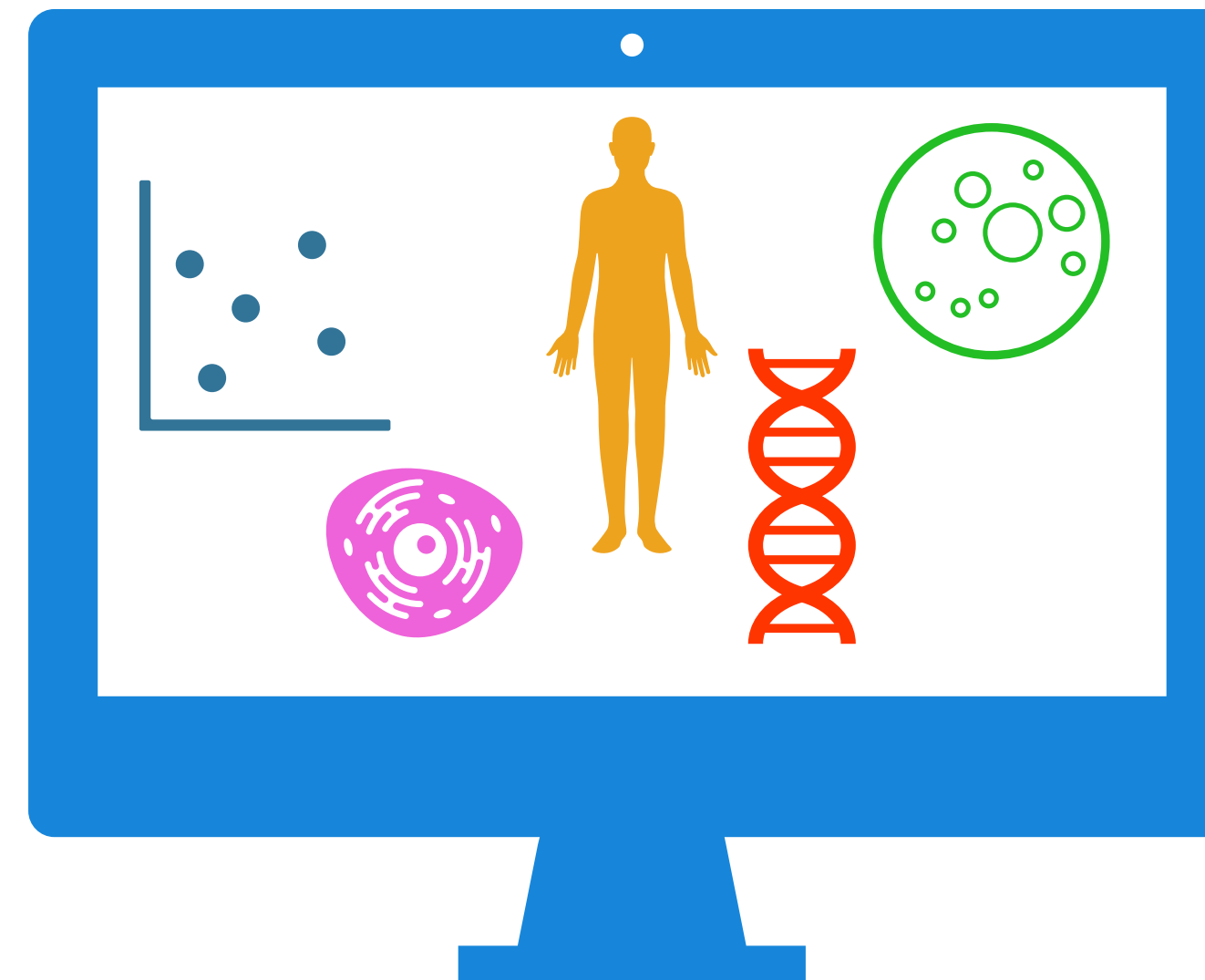


- Bioinformatics is "the science that uses the instruments of informatics to analyze biological data in order to formulate hypotheses about **life**." (Anna Tramontano)

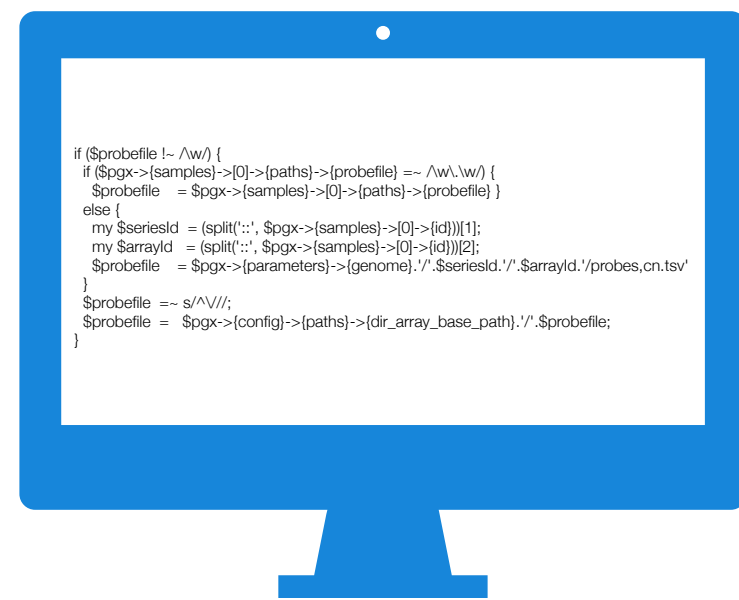


Sources: nextprot | opentreeoflife | wikipedia | MacKay et al., Cancer Cell (2017) | original photos

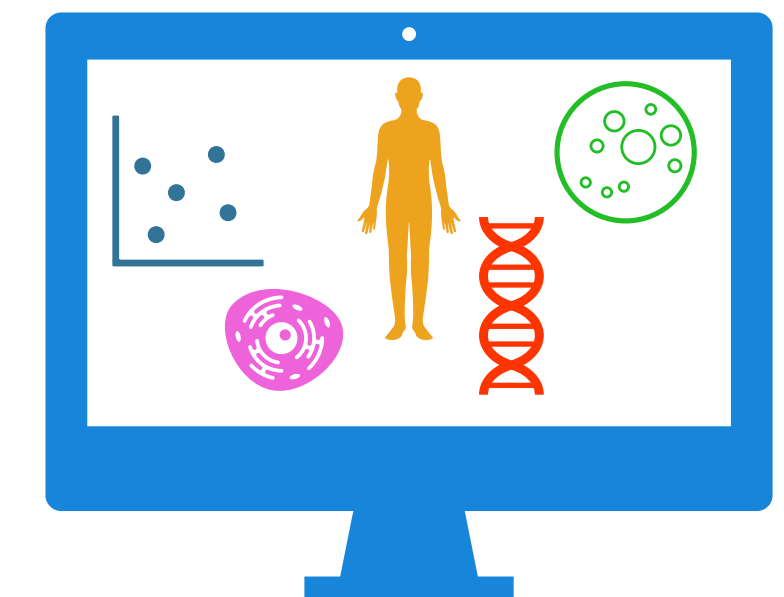
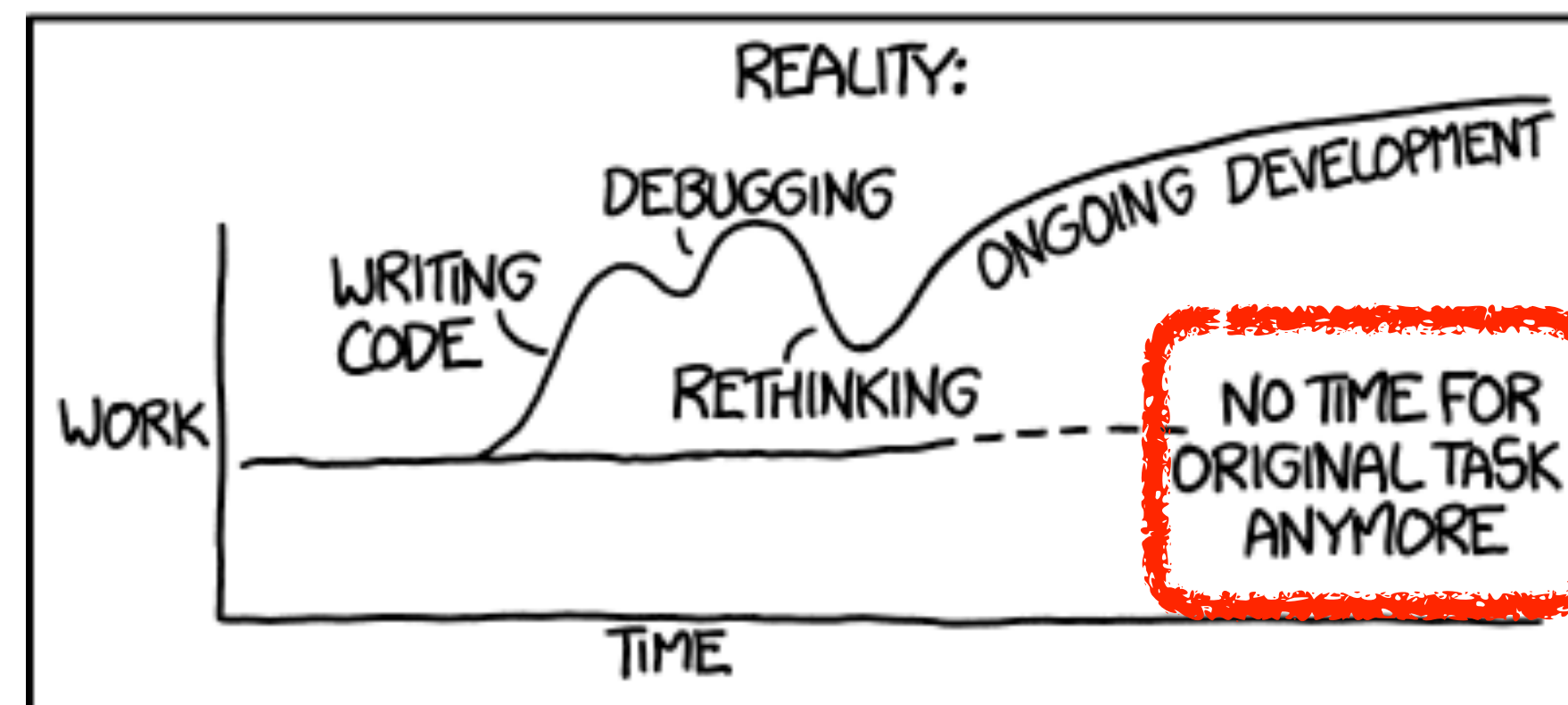
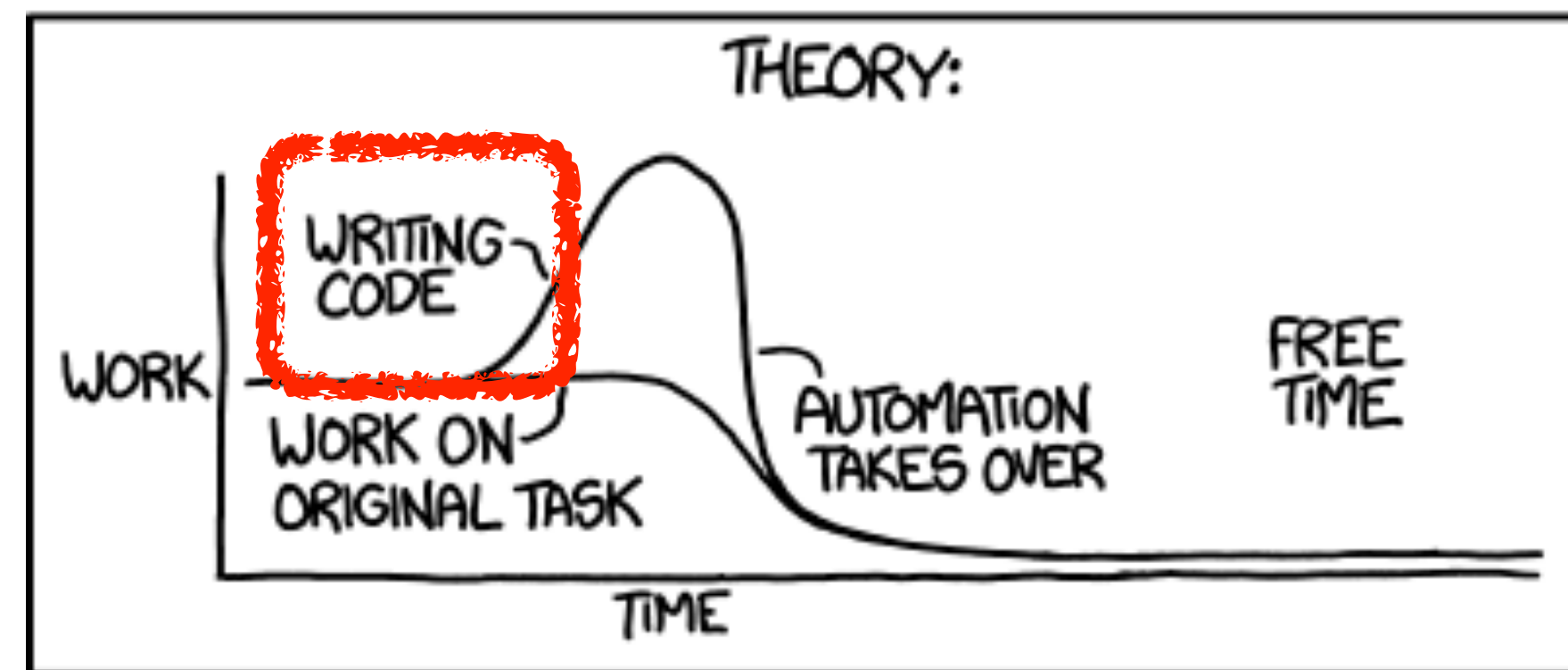
{bio_informatics_science}



{bio_informatics_science}



"I SPEND A LOT OF TIME ON THIS TASK.
I SHOULD WRITE A PROGRAM AUTOMATING IT!"



Bioinformatician

strong biological knowledge

provides hypothesis and / or dataset

sufficient statistical and **computational** expertise to correctly use bioinformatics tools & develop workflows (scripting ...)

expert **user** of informatics tools

may get a Nobel

Bioinformatician

sufficient biological background

provides statistical, analysis methods

sufficient biological or **medical** background to understand problems presented and identify pitfalls and hidden biases arising from data generation

developer of informatics tools

may get rich

Bioinformatician

strong biological knowledge

provides hypothesis and / or dataset

sufficient statistical and **computational** expertise to correctly use bioinformatics tools & develop workflows (scripting ...)

expert **user** of informatics tools

may get a Nobel

Bioinformatician

sufficient biological background

provides statistical, analysis methods

sufficient biological or **medical** background to understand problems presented and identify pitfalls and hidden biases arising from data generation

developer of informatics tools

may get rich

flux

What do Bioinformaticians work on?

Hypothesis & Data Driven Approaches to Biological Topics

- protein **structure** definition
- DNA/RNA/protein **sequence** analysis
- **quantitative** analysis of "-omics" and cytometry data
- **functional** enrichment of target data (e.g. genes, sequence elements)
- **evolutionary** reconstruction and "tree of life" questions
- **image processing** for feature identification and spatial mapping
- **statistical** analysis of measurements and observations
- **protocols** for efficient storage, annotation and retrieval of biomedical data
- **information extraction** from prose & declarative knowledge resources (think publications & data tables)
- **clinical** bioinformatics - risk assessment and therapeutic target identification
- ...

**FITTING
THE MODEL**

**CLEANING
THE DATA**

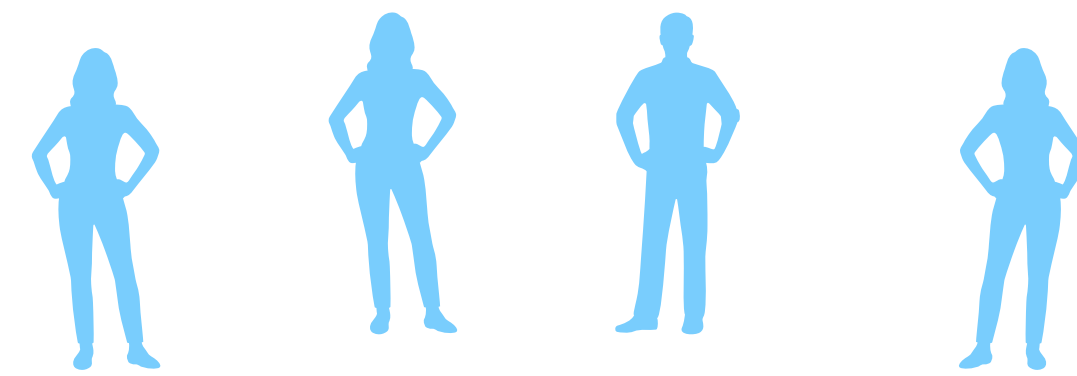
Data sets in tutorials



Data sets in the wild



DATA PIPELINE



BIOCURATION

BIOINFORMATICS



arrayMap

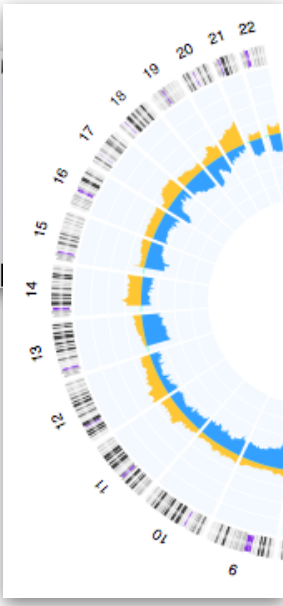
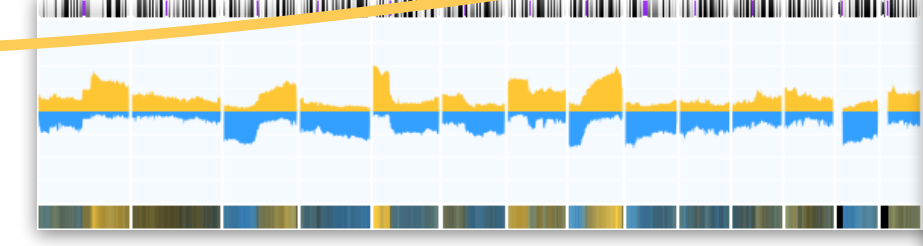
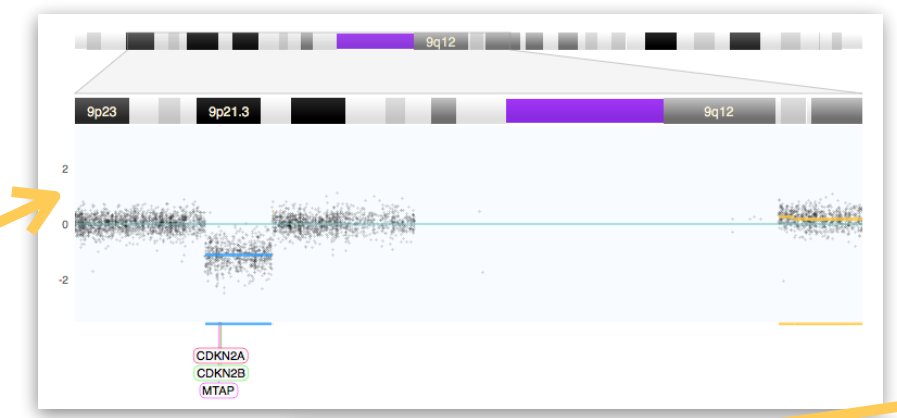


progenetix



GEO
GSE102668
Chronic lymphocytic leukemia, Dataset 1, Specimen 1049

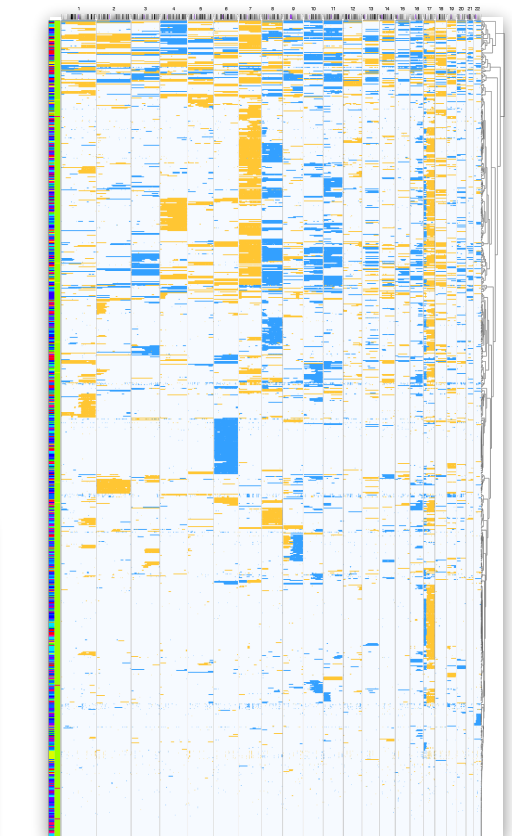
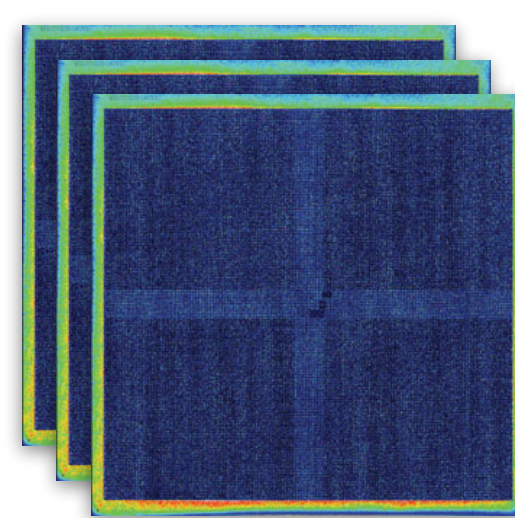
arrayMap
visualizing cancer genome array data @ arraymap.org



ICD Code	ICD Morphology	arraymap	progenetix
00000	not classified in icd-o (e.g. non-neoplastic or benign)	8814	370
80000	neoplasm, malignant	11	1
80100	apoptosis tumor, benign	15	1
80102	carcinoma in situ, nos	20	11
80103	carcinoma, nos	1430	258
80120	large cell carcinoma, nos	45	54
80130	large cell neuroendocrine carcinoma	3	60
80200	carcinoma, undifferentiated type, nos	4	41
80210	carcinoma, anaplastic type, nos	4	1
80220	pleomorphic carcinoma	4	3
80300	apical cell carcinoma	1	1
80303	seromucoid carcinoma	2	7
80410	small cell carcinoma, nos	132	148
80460	non-small cell carcinoma	1195	184
80500	papillary carcinoma, nos	16	14
80503	colloid carcinoma	1	132
80701	premalignant squamous epithelium, nos	45	152
80702	squamous cell carcinoma in situ, nos	65	16
80703	squamous cell carcinoma, nos	2443	2087
80710	squamous cell keratosis, nos	11	1
80750	squamous cell carcinoma, acantholytic	2	2
80772	squamous intraepithelial neoplasia, grade II	136	22
80800	undifferentiated neoplasmy/nerve carcinoma	52	200
80803	basal cell carcinoma, nos	29	15
81000	transitional cell carcinoma in situ	318	10
81001	transitional cell carcinoma, nos	210	423
81301	urothelial papilloma, nos	184	39
81302	papillary transitional cell carcinoma, non-invasive	2	56
81303	papillary transitional cell carcinoma	2	8
81400	adenoma, nos	365	361
81401	atypical adenoma	1	88
81402	adenocarcinoma in situ	149	11
81403	adenocarcinoma, nos	9457	3248
81440	adenocarcinoma, intestinal type	167	206
81450	carcinoma, diffuse type	7	36
81480	granular intrapapillary neoplasia, grade II	1	15
81501	ser cell adenoma	1	18
81502	ser cell carcinoma	1	28
81503	ser cell carcinoma	1	18
81510	neurofibroma, nos	1	28

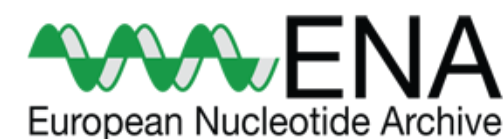
informa
ORIGINAL ARTICLE: RESEARCH
Genomic imbalance defines three prognostic groups for risk stratification of patients with chronic lymphocytic leukemia

ArrayExpress
E-MTAB-98 - Comparative genomic hybridization by array of human peripheral T-cell lymphoma clinical samples to study their genomic aberration profiles



Bioinformatics: Data Categories & Databases

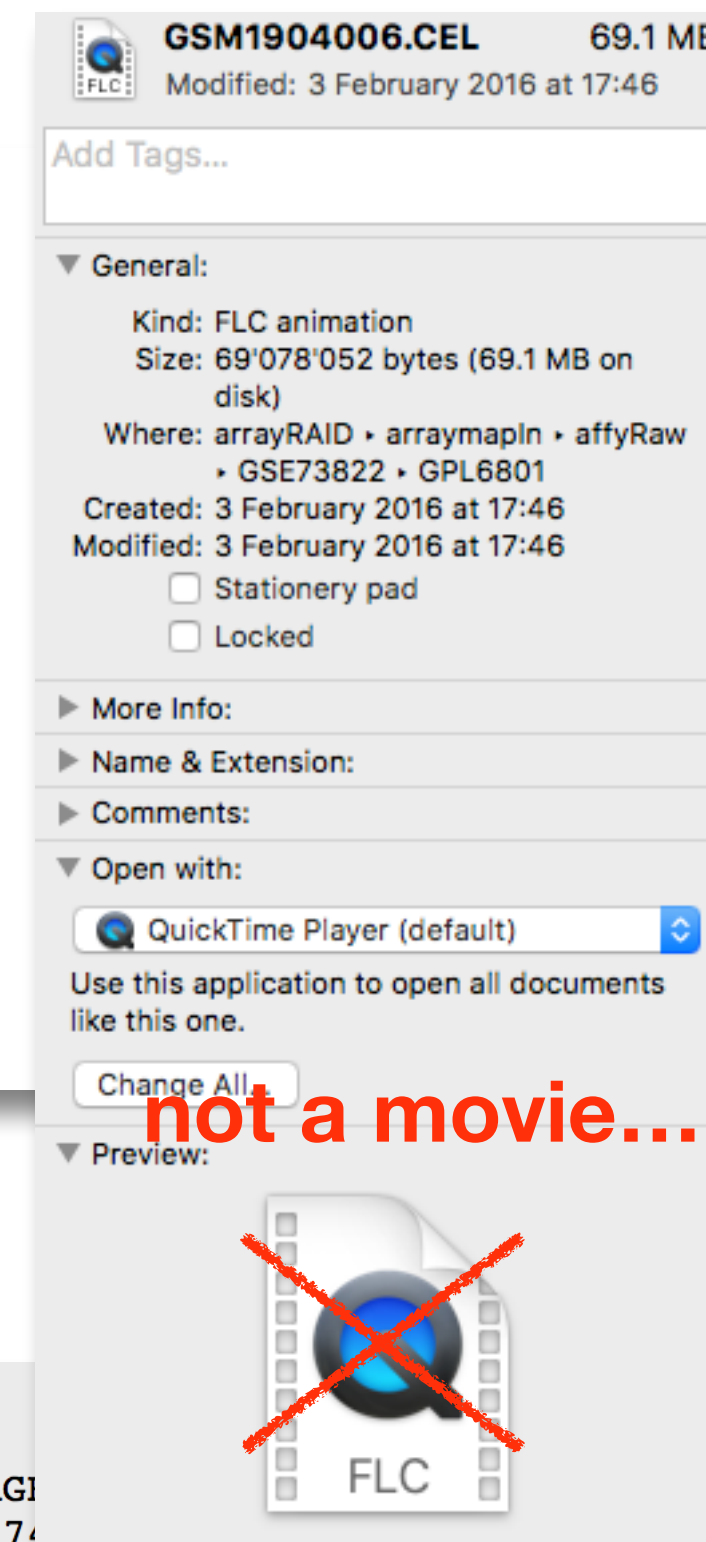
- biological data comes in **3 main categories**:
 - **sequence** data (nucleic acids, aminoacids)
 - **structural** data (DNA, RNA, proteins; intracellular organisation, tissues ...)
 - **functional** data (interactions in time and space)
- data storage & retrieval: importance of local and connected **databases**
 - **primary databases** - for deposition of original, raw data (e.g. SRA - sequence read archive; ENA - European Nucleotide Archive; GEO - NCBI Gene Expression Omnibus; EBI arrayExpress...)
 - **derived databases / resources** - information resources providing agglomerated & **curated** data derived from primary sources (e.g. UniprotKB, nextProt, String, KEGG, Progenetix...)



Bioinformatics: File Formats, Ontologies & APIs

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

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



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

genome.ucsc.edu/FAQ/FAQformat.html

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

BED file example

File Formats: VCF

Genomic variant storage standard

The Variant Call Format (VCF) Version 4.2 Specification

25 Jun 2020

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

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

1 The VCF specification

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

1.1 An example

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001 NA00002 NA00003
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2 2/2:35:4
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20 1234567 microsat1 GTC G,GTCT 50 PASS NS=3;DP=9;AA=G GT:GQ:DP 0/1:35:4 0/2:17:2 1/1:40:3
```

Bioinformatics: File Formats, Ontologies & APIs

- ontologies in information sciences describe concrete and abstract **objects**, there precisely defined **hierarchies** and **relationships**
- ontologies in bioinformatics support the move from a descriptive towards an **analytical science** in describing biological data and relations among it

"The widest use of ontologies within biology is for conceptual annotation – a representation of stored knowledge **more computationally amenable than natural language.**"*

- Gene ontology (GO)
- NCIt Neoplasm Core
- UBERON anatomical structures
- Experimental Factor Ontology (EFO)
- Disease Ontology (DO)



```
id: GO:0000118
name: histone deacetylase complex
namespace: cellular_component
def: "A protein complex that possesses histone deacetylase activity." [GOC:mah]
comment: Note that this term represents a location, not a function; the activities possessed by this complex is mentioned in the
definition for the purpose of de
molecular function term 'histone
synonym: "HDAC complex" EXACT [
is_a: GO:0044451 ! nucleoplasm ]
is_a: GO:1902494 ! catalytic co
```

- Neoplasm by Morphology
 - Epithelial Neoplasm [C3709](#)
 - Germ Cell Tumor [C3708](#)
 - Giant Cell Neoplasm [C7069](#)
 - Hematopoietic and Lymphoid Cell Neoplasm [C27134](#)
 - Melanocytic Neoplasm [C7058](#)
 - Benign Melanocytic Skin Nevus [C7571](#)
 - Dysplastic Nevus [C3694](#)
 - Melanoma [C3224](#)
 - Amelanotic Melanoma [C3802](#)
 - Cutaneous Melanoma [C3510](#)
 - Epithelioid Cell Melanoma [C4236](#)
 - Mixed Epithelioid and Spindle Cell Melanoma [C66756](#)
 - Non-Cutaneous Melanoma [C8711](#)
 - Spindle Cell Melanoma [C4237](#)
 - Meningothelial Cell Neoplasm [C6971](#)



Standardized Data

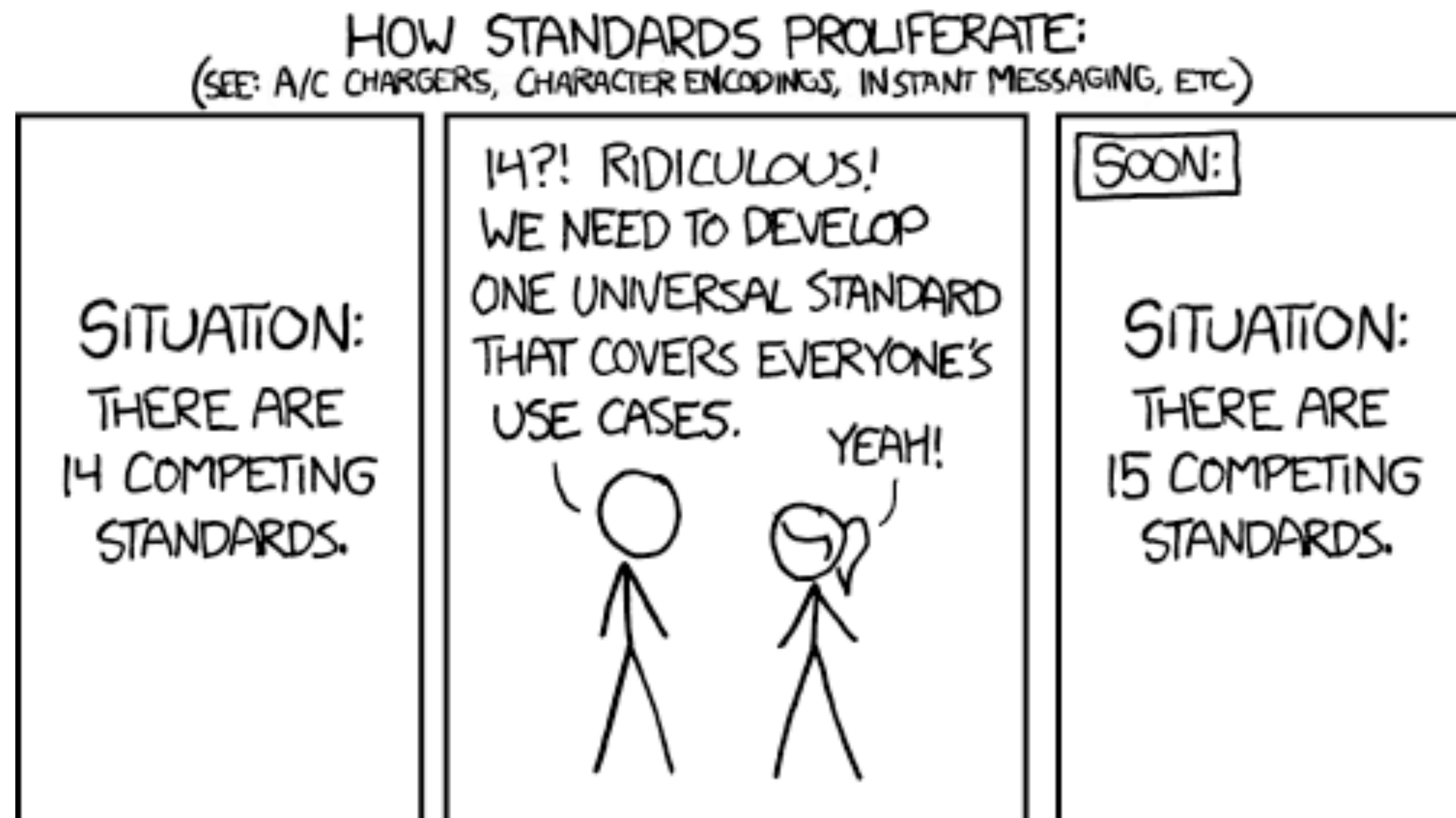
Data re-use depends on standardized, machine-readable metadata

- Multiple international initiatives (ELIXIR, GA4GH, MONARCH...) and resource providers (EBI, NCBI ...) work on the generation and implementation of data annotation standards
- emerging / established principles are the use of **hierarchical** coding systems where individual codes are represented as CURIEs
- other formats for non-categorical annotations based on international standards, e.g.
 - ISO (ISO 8601 time & period, ISO 3166 country codes ...)
 - IETF (GeoJSON ...)
 - W3C (CURIE ...)
- these standards become pervasive throughout GA4GH's ecosystem (e.g. Phenopackets ...)

```
"label" : "no restriction",
"id" : "DUO:0000004"
},
"provenance" : {
  "material" : {
    "type" : {
      "id" : "EFO:0009656",
      "label" : "neoplastic sample"
    }
  },
  "geo" : {
    "label" : "Zurich, Switzerland",
    "precision" : "city",
    "city" : "Zurich",
    "country" : "Switzerland",
    "latitude" : 47.37,
    "longitude" : 8.55,
    "geojson" : {
      "type" : "Point",
      "coordinates" : [
        8.55,
        47.37
      ]
    },
    "ISO-3166-alpha3" : "CHE"
  }
},
{
  "age" : "P25Y3M2D"
```

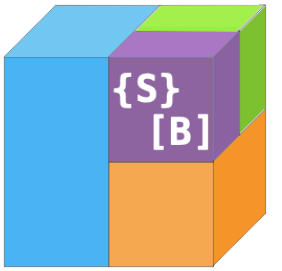
Standardized Data

Data re-use depends on standardized, machine-readable metadata



xkcd

```
"label" : "no restriction",
"id" : "DUO:0000004"
},
"provenance" : {
  "material" : {
    "type" : {
      "id" : "EFO:0009656",
      "label" : "neoplastic sample"
    }
  },
"geo" : {
  "label" : "Zurich, Switzerland",
  "precision" : "city",
  "city" : "Zurich",
  "country" : "Switzerland",
  "latitude" : 47.37,
  "longitude" : 8.55,
  "geojson" : {
    "type" : "Point",
    "coordinates" : [
      8.55,
      47.37
    ]
  },
  "ISO-3166-alpha3" : "CHE"
}
},
{
"age" : "P25Y3M2D"
```



Schemas for Data & APIs - Standardization & Documentation!

BeaconAlleleRequest beacon ↗

{S}[B] Status [i]	implemented
Provenance	<ul style="list-style-type: none"> Beacon API
Used by	<ul style="list-style-type: none"> Beacon Progenetix database schema (Beacon+ backend)
Contributors	<ul style="list-style-type: none"> Marc Fiume Michael Baudis Sabela de la Torre Pernas Jordi Rambla Beacon developers...
Source (v1.1.0)	<ul style="list-style-type: none"> raw source [JSON] Github

Attributes

Type: object

Description: Allele request as interpreted by the beacon.

Properties

Property	Type
alternateBases	string
assemblyId	string
datasetIds	array of string
end	integer
endMax	integer
endMin	integer
mateName	https://schemablocks.org/schemas/beacon/v1.1.0/Chromosome.json [SRC] [HTML]
referenceBases	string
referenceName	https://schemablocks.org/schemas/beacon/v1.1.0/Chromosome.json [SRC] [HTML]
start	integer (int64)
startMax	integer
startMin	integer
variantType	string

alternateBases

- type: string

The bases that appear instead of the reference bases. Accepted values: [ACGTN]*. N is a wildcard, that denotes the position of any base, and can be used as a standalone base of any type or within a partially known sequence. For example a sequence where the first and last bases are known, but the middle portion can exhibit countless variations of [ACGT], or the bases are unknown: ANNT the Ns can take any form of [ACGT], which makes both ACCT and ATGT (or any other combination) viable sequences.

Symbolic ALT alleles (DEL, INS, DUP, INV, CNV, DUP:TANDEM, DEL:ME, INS:ME) will be represented in [variantType](#).

Optional: either [alternateBases](#) or [variantType](#) is required.

alternateBases Value Example

assemblyId

- type: string

Assembly identifier (GRC notation, e.g. [GRCh37](#)).

assemblyId Value Example

Curie sb-vr-spec ↗

{S}[B] Status [i]	implemented
Provenance	<ul style="list-style-type: none"> Curie
Used by	<ul style="list-style-type: none"> Curie
Contributors	<ul style="list-style-type: none"> Curie
Source (v1.0)	<ul style="list-style-type: none"> Curie

Attributes

Type: string

Pattern: ^\w[^\+:\.]+\$

Description: A string that refers sender.

VR does not impose any constraint on the data, the VR Specification RECOM string CURIEs are represented as `namespace:accession` or `name`. The VR specification also RECOM string CURIEs are represented as `namespace:accession` or `name`. The [reference](#) component is a CURIE. A CURIE is a URI. URIs may *locate* a resource. VR uses CURIEs primarily as a name. Implementations MAY provide CURIEs. Using internal ids in public mess

Curie Value Examples

"ga4gh:GA.01234abcde"
"DUO:0000004"
"orcid:0000-0003-3463-0775"
"PMID:15254584"

Biosample sb-phenopackets ↗

{S}[B] Status [i]	implemented
Provenance	<ul style="list-style-type: none"> Phenopackets
Used by	<ul style="list-style-type: none"> Phenopackets
Contributors	<ul style="list-style-type: none"> GA4GH Data Working Group Jules Jacobsen Peter Robinson Michael Baudis Melanie Courtot Isuru Liyanage
Source (v1.0.0)	<ul style="list-style-type: none"> raw source [JSON] Github

Attributes

Type: object

Description: A Biosample refers to a unit of biological material from which the substrate molecules (e.g. genomic DNA, RNA, proteins) for molecular analyses (e.g. sequencing, array hybridisation, mass-spectrometry) are extracted.

Examples would be a tissue biopsy, a single cell from a culture for single cell genome sequencing or a protein fraction from a gradient centrifugation.

Several instances (e.g. technical replicates) or types of experiments (e.g. genomic array as well as RNA-seq experiments) may refer to the same Biosample.

FHIR mapping: [Specimen](#).

Properties

Property	Type
ageOfIndividualAtCollection	https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/Age.json [SRC] [HTML]
ageRangeOfIndividualAtCollection	https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/AgeRange.json [SRC] [HTML]
description	string
diagnosticMarkers	array of https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/OntologyClass.json [SRC] [HTML]
histologicalDiagnosis	https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/OntologyClass.json [SRC] [HTML]
htsFiles	array of https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/HtsFile.json [SRC] [HTML]
id	string
individualId	string
isControlSample	boolean
phenotypicFeature	array of https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/PhenotypicFeature.json [SRC] [HTML]
procedure	https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/Procedure.json [SRC] [HTML]
sampledTissue	https://schemablocks.org/schemas/sb-

Checksum sb-checksum ↗

{S}[B] Status [i]	proposed
Provenance	<ul style="list-style-type: none"> GA4GH DRS (`develop` branch)
Used by	<ul style="list-style-type: none"> GA4GH DRS GA4GH TRS
Contributors	<ul style="list-style-type: none"> Susheel Varma
Source (v0.0.1)	<ul style="list-style-type: none"> raw source [JSON] Github

Attributes

Type: object

Description: Checksum

Properties

Property	Type
checksum	string
type	string

checksum

- type: string

The hexadecimal encoded ([Base16](#)) checksum for the data

checksum Value Example

"77af4d6b9913e693e8d0b4b294fa62ade6054e6b2f1ffb617ac955dd63fb0182"
--

type

- type: string

The digest method used to create the checksum. The value (e.g. [sha-256](#)) SHOULD be listed as [Hash Name String](#) in the [GA4GH Hash Algorithm Registry](#). Other values MAY be used, as long as implementors are aware of the issues discussed in [RFC6920](#).

GA4GH may provide more explicit guidance for use of non-IANA-registered algorithms in the future.

type Value Example

"sha-256"

Bioinformatics: File Formats, Ontologies & APIs

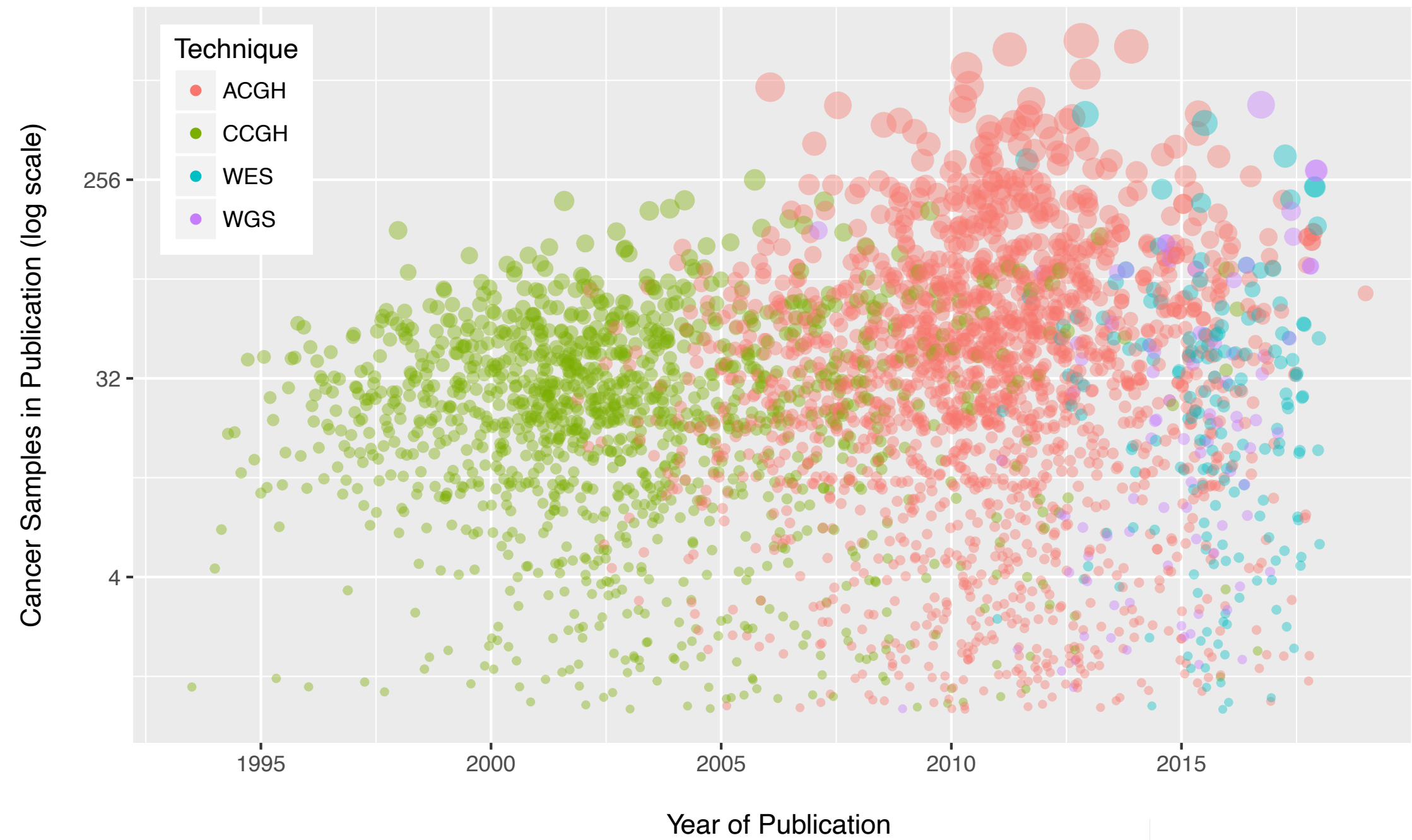
- databases can be accessed through **A**pplication **P**rogramming **I**nterfaces
- *API : set of routines, protocols, and tools that specifies how software components interact, to exchange data and processing capabilities*
- web API example: implementing geographic maps, with parameters provided by the client (e.g. location coordinates, quantitative payload)
- web APIs provide a *machine readable* response to queries over HTTP
- bioinformatic applications frequently make use of web APIs for **data retrieval** or genome browser APIs for **data display**
- bioinformatics software libraries for API functionality are usually implemented in **Perl, Python** and/or **R**

Bioinformatics: File Formats, Ontologies & APIs

```
{
  "$schema": "https://raw.githubusercontent.com/ga4gh-beacon/
beacon-v2/main/framework/json/requests/beaconRequestBody.json",
  "meta": {
    "apiVersion": "2.0",
    "requestedSchemas": [
      {
        "entityType": "genomicVariation",
        "schema": "https://raw.githubusercontent.com/
ga4gh-beacon/beacon-v2/main/models/json/beacon-v2-default-
model/genomicVariations/defaultSchema.json"
      }
    ]
  },
  "query": {
    "requestParameters": {
      "g_variant": {
        "referenceName": "NC_000017.11",
        "start": [ 5000000, 7676592 ],
        "end": [ 7669607, 10000000 ],
        "variantType": "DEL"
      }
    }
  },
  "requestedGranularity": "record",
  "pagination": {
    "skip": 0,
    "limit": 5
  }
}
```

```
{
  "meta": {
    "apiVersion": "v2.0.0",
    "beaconId": "org.progenetix.beacon",
    "createDateTime": "2015-11-13 00:00:00",
    "receivedRequestSummary": {
      "apiVersion": "v2.0.0",
      "response": {
        "resultSets": [
          {
            "exists": true,
            "id": "progenetix",
            "info": {
              "counts": {
                "callCount": 525,
                "sampleCount": 515,
                "variantCount": 247
              }
            },
            "paginatedResultsCount": 247,
            "results": [
              {
                "caseLevelData": [
                  {
                    "analysisId": "pgxcs-kftwbzza",
                    "biosampleId": "pgxbs-kftviv0x",
                    "id": "pgxvar-5c86619f09d374f2dc3bbfcd"
                  }
                ]
              }
            ]
          }
        ]
      }
    }
  }
}
```

<http://docs.genomebeacons.org/variant-queries/>



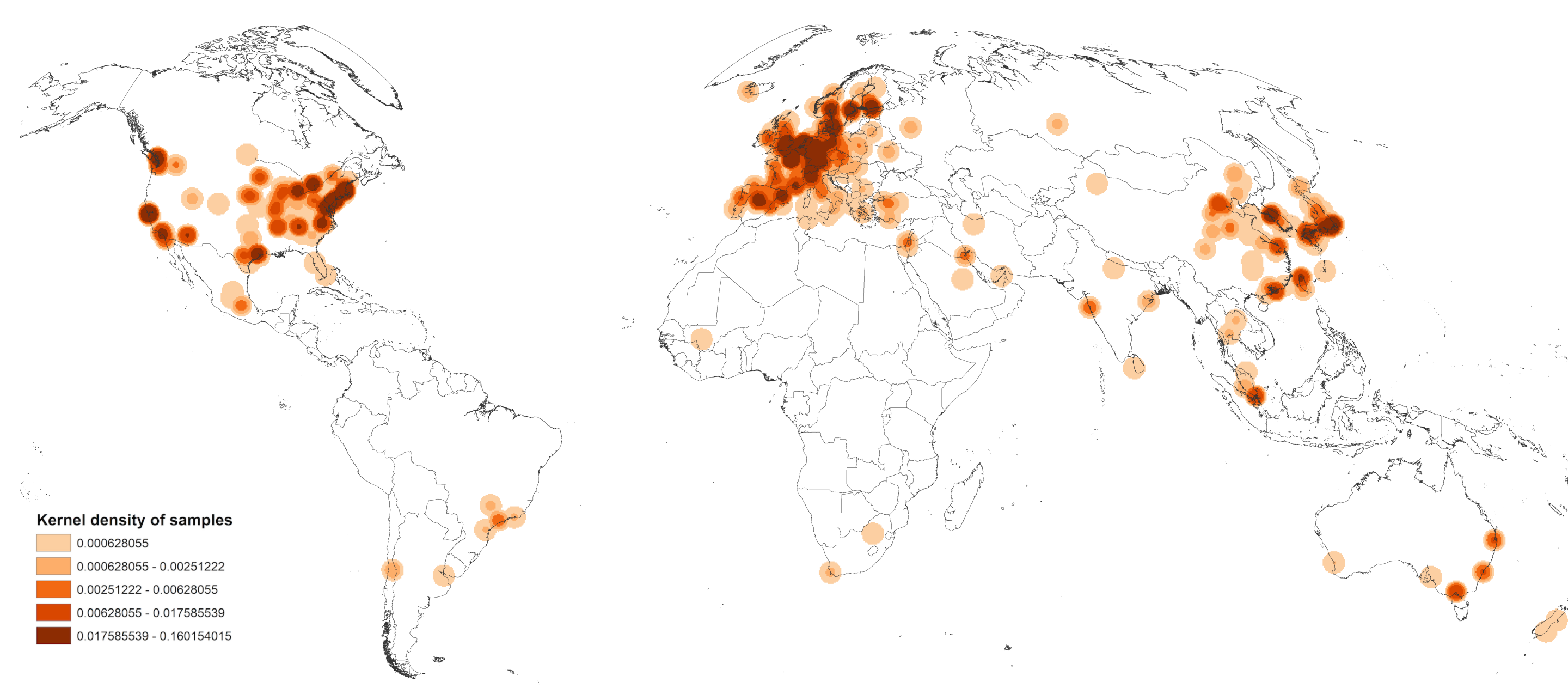
Publication Landscape of Cancer CNV Profiling

Publication statistics for cancer genome screening studies. The graphic shows our assessment of publications reporting whole-genome screening of cancer samples, using molecular detection methods (chromosomal CGH, genomic array technologies, whole exome and genome sequencing).

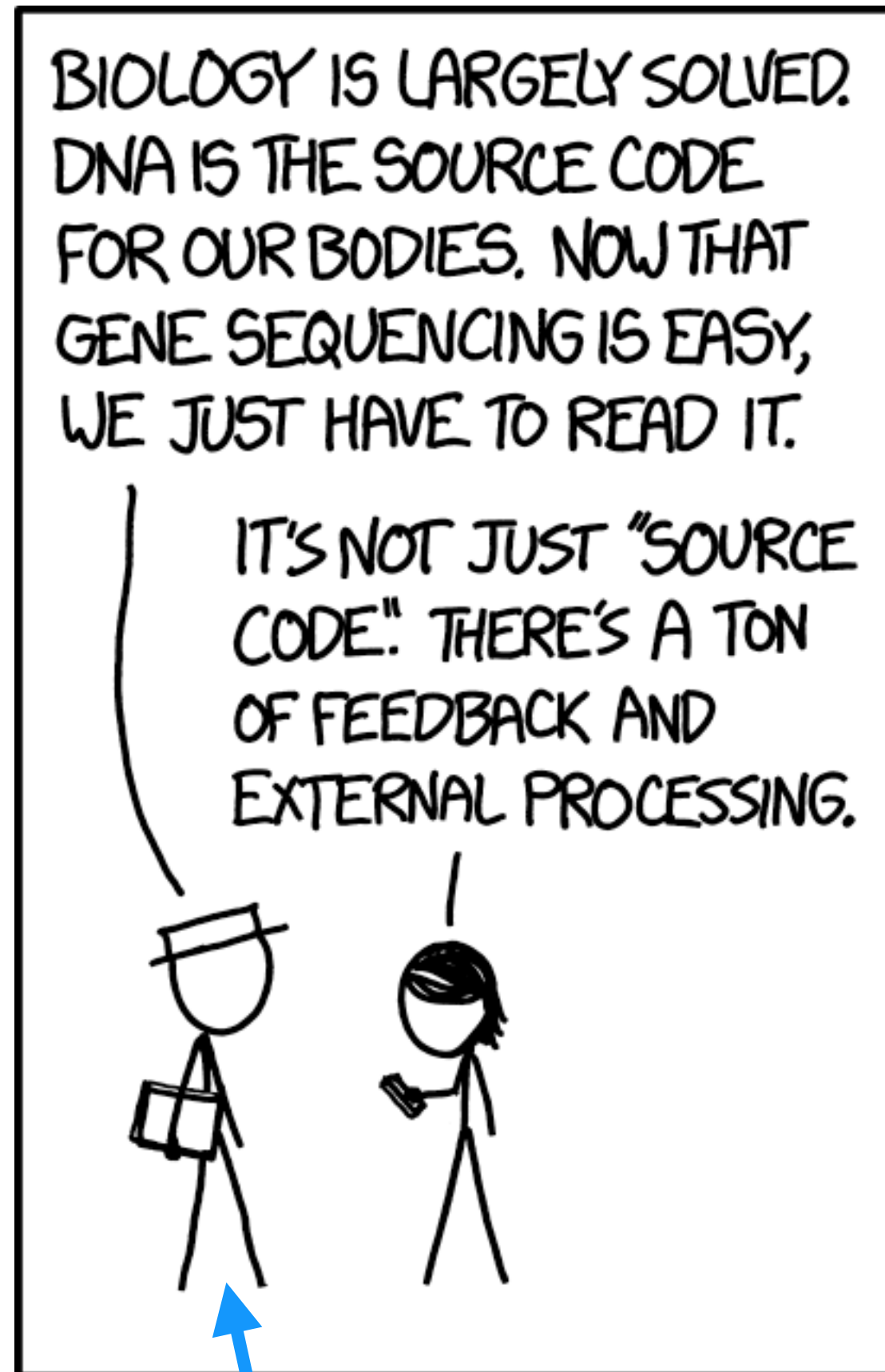
For the years 1993-2018, we found 3'229 publications reporting 174'530 individual samples in single series from 1 to more than 1000 samples. Y-axis and size of the dots correspond to the sample number; the color codes indicate the technology used.

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.

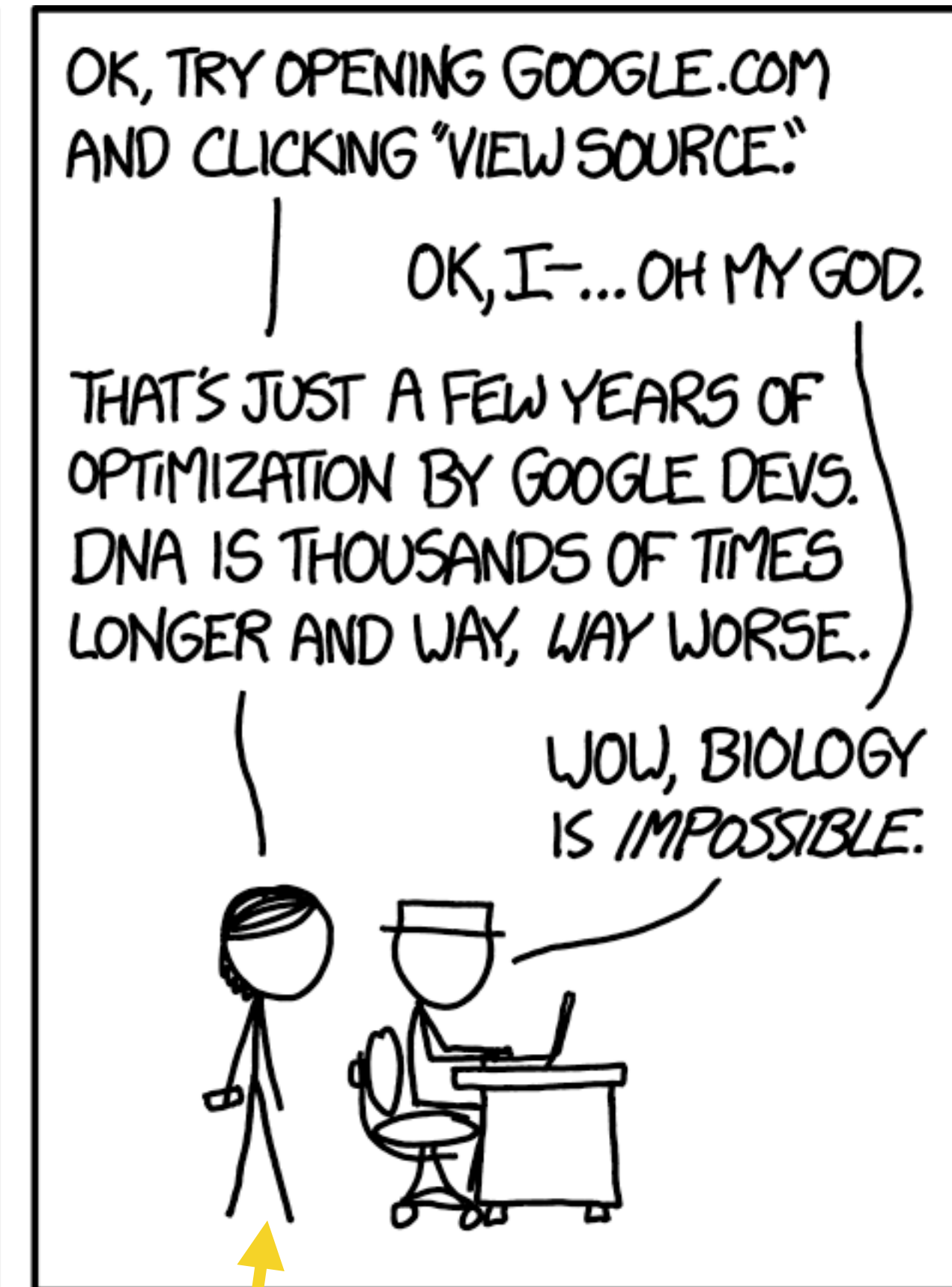
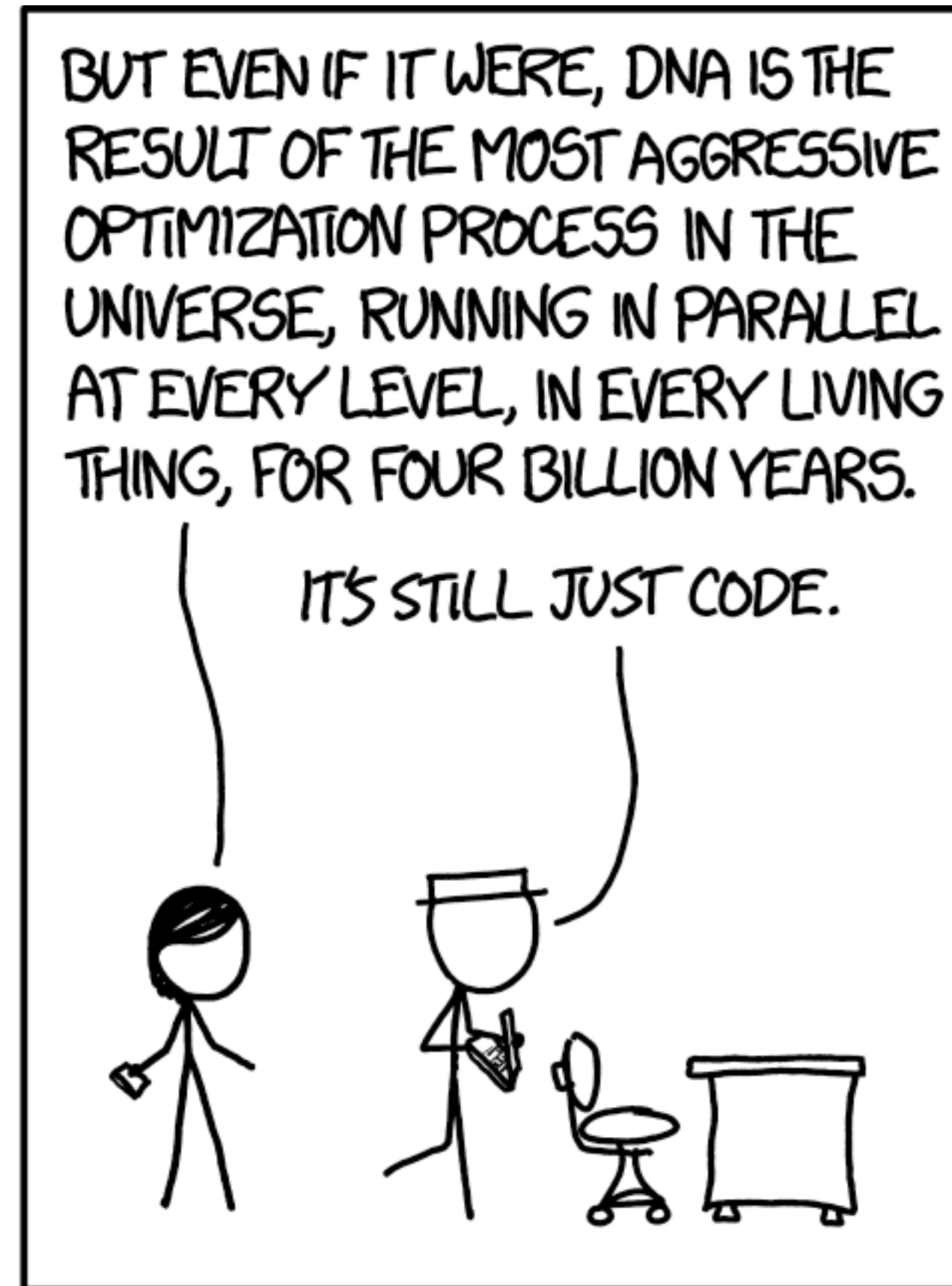
The numbers are derived from the 3'240 publications registered in the Progenetix database.



Who is a Bioinformatician?



bioinformatician



bioinformatician

But: What is not bioinformatics, though being "bio" and using computers?

- *"I do not think all biological computing is bioinformatics, e.g. **mathematical modelling** is not bioinformatics, even when connected with biology-related problems. In my opinion, bioinformatics has to do with management and the subsequent use of biological information, particular genetic information."*
(Richard Durbin)
- **biologically-inspired computation** (neural networks etc.) - though their application may be part of bioinformatics
- **computational & systems biology**, where the emphasis is on **modelling** rather than on **data interpretation**

Bioinformatics OR Computational / Systems Biology?

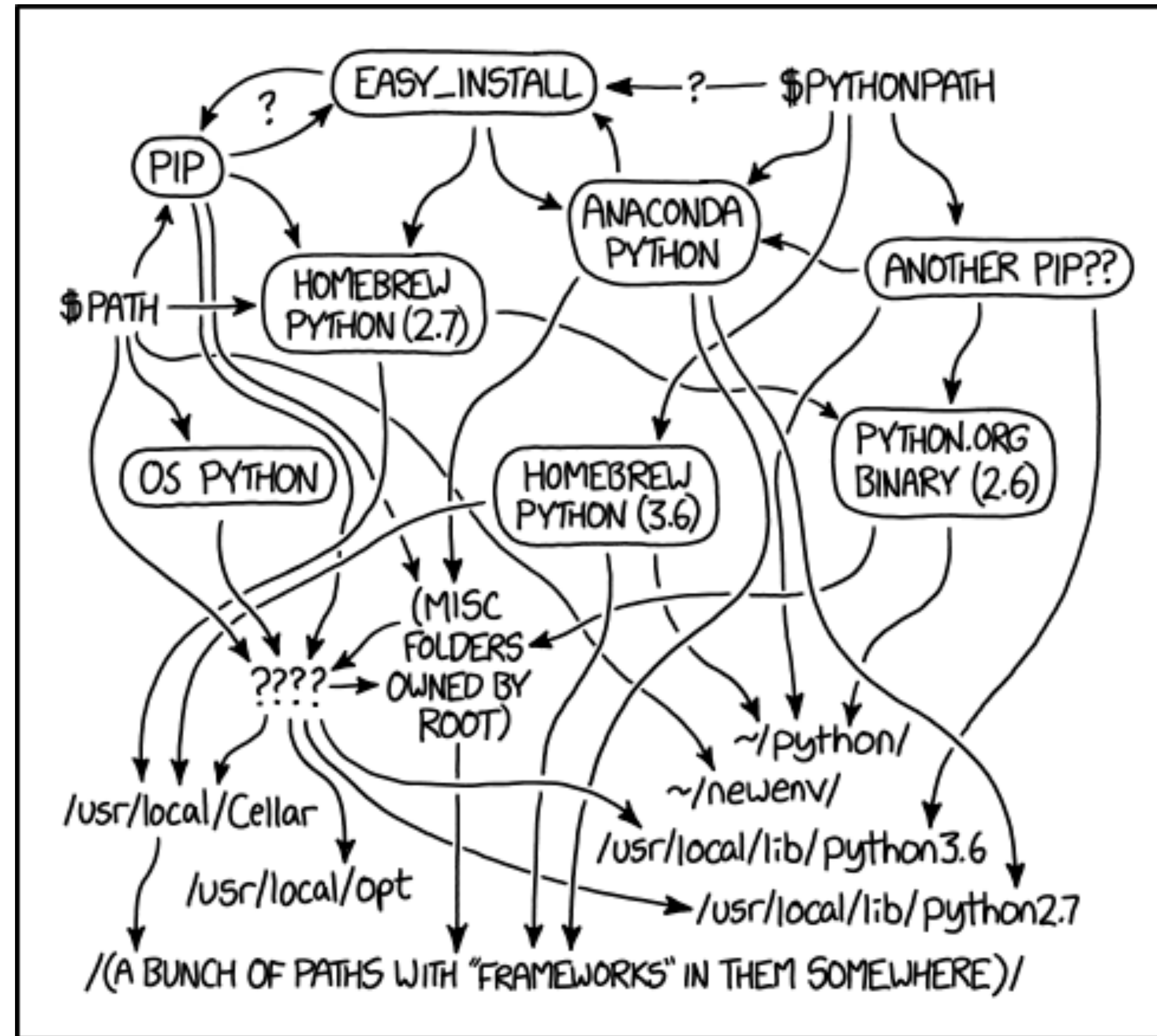
- **Bioinformatics**

Research, development, or **application** of computational **tools** and approaches to make the vast, diverse and complex **life sciences data** more understandable and useful

- **Computational Biology**

The development and application of **mathematical** and computational **approaches** to address **theoretical** and experimental questions in biology

But in reality that is what bioinformaticians do...



MY PYTHON ENVIRONMENT HAS BECOME SO DEGRADED THAT MY LAPTOP HAS BEEN DECLARED A SUPERFUND SITE.

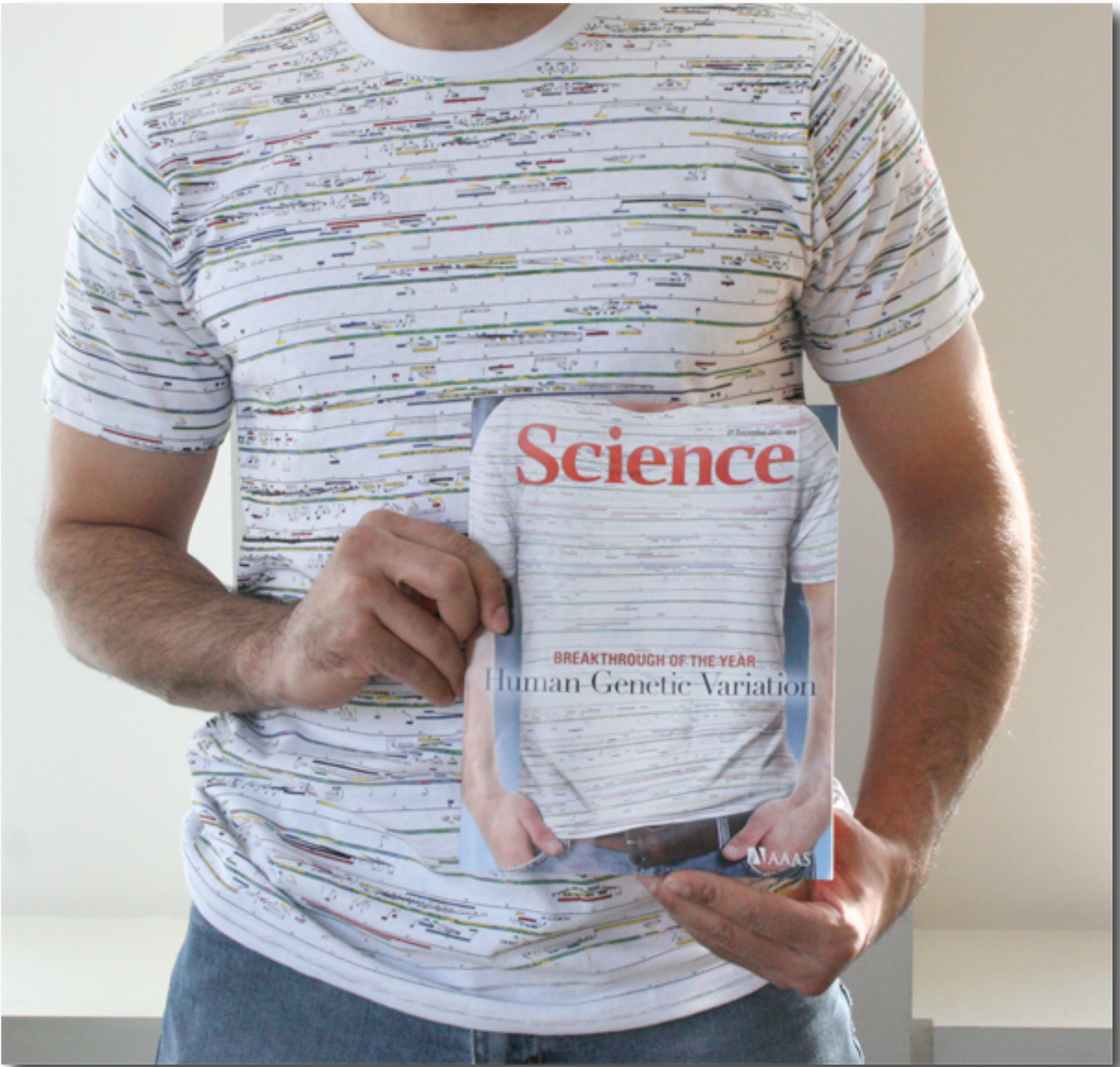


BIO390: Introduction to Bioinformatics

Lecture I: What are Bioinformaticians doing? Example from Theoretical Oncogenomics and Federated Human Data



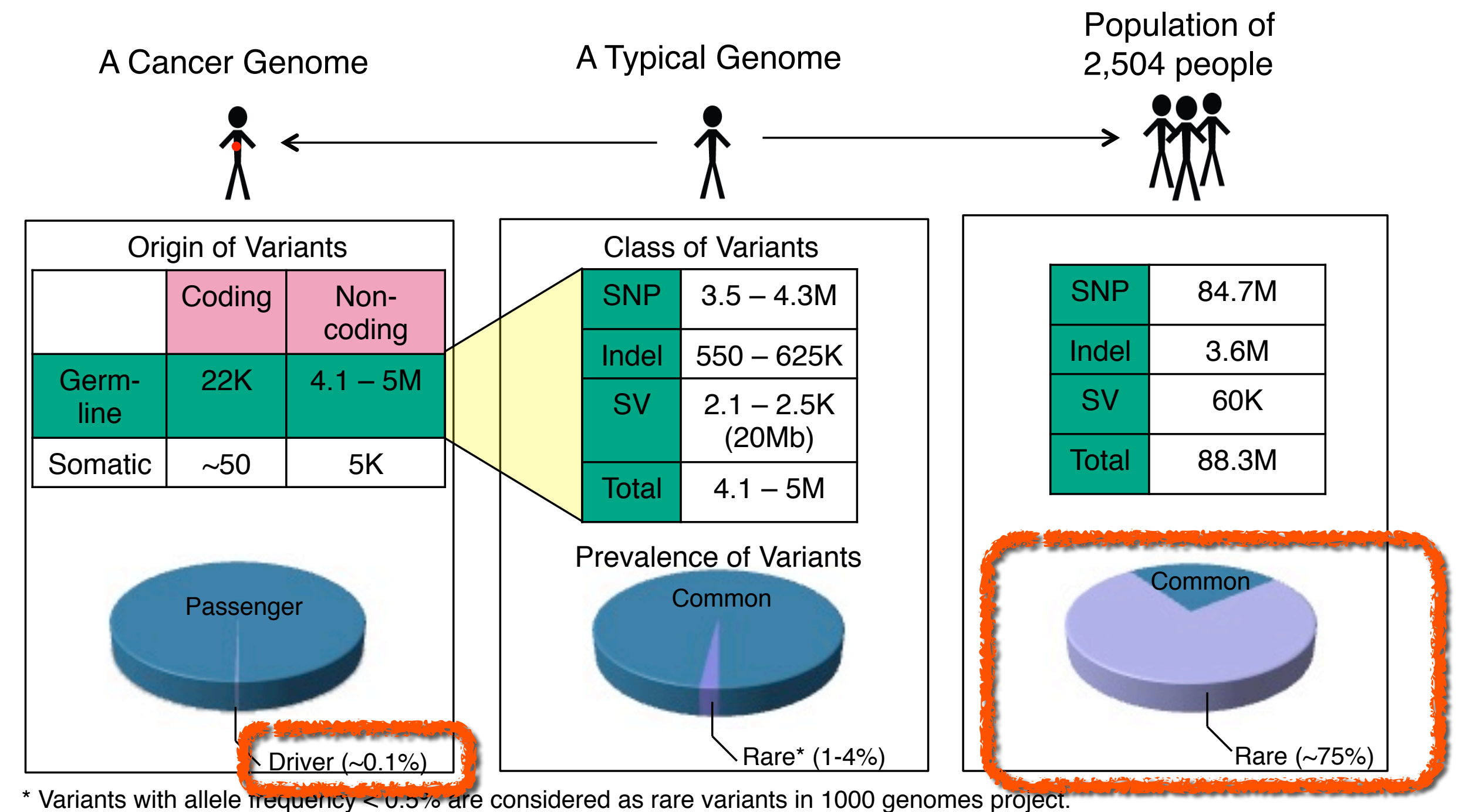
The trouble with human genome variation



Finding Somatic Mutations In Cancer

Many Needles in a Large Haystack

- a typical human genome (~3 billion base pairs) has ~5 million variants
- most of them are "**rare**"; i.e. can only be identified as recurring when sequencing thousands of people
- cancer cells accumulate additional variants, only **few** of which ("**drivers**") are relevant for the disease

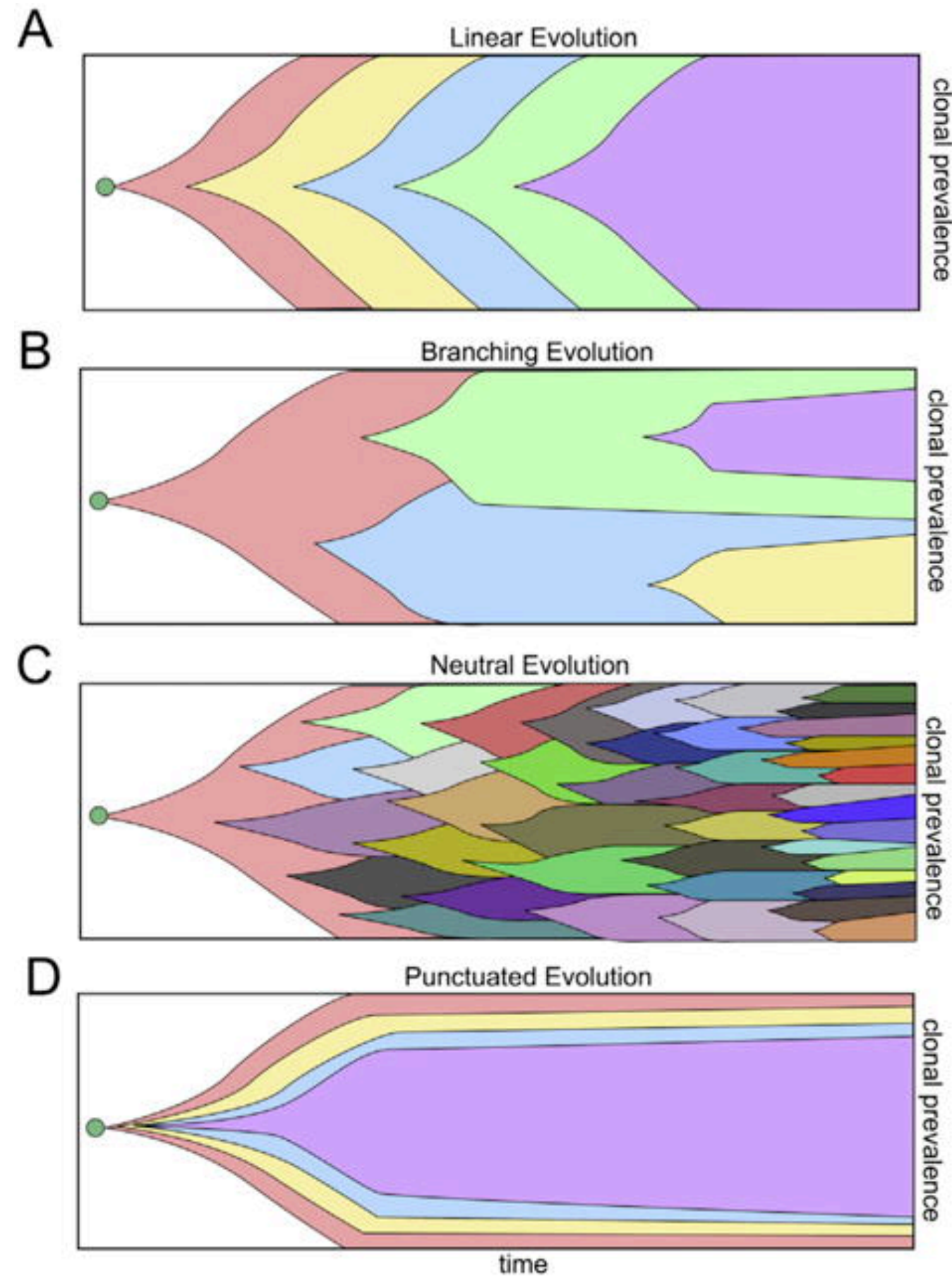


* Variants with allele frequency < 0.5% are considered as rare variants in 1000 genomes project.

The 1000 Genomes Project Consortium, Nature. 2015. 526:68-74
Khurana E. et al. Nat. Rev. Genet. 2016. 17:93-108

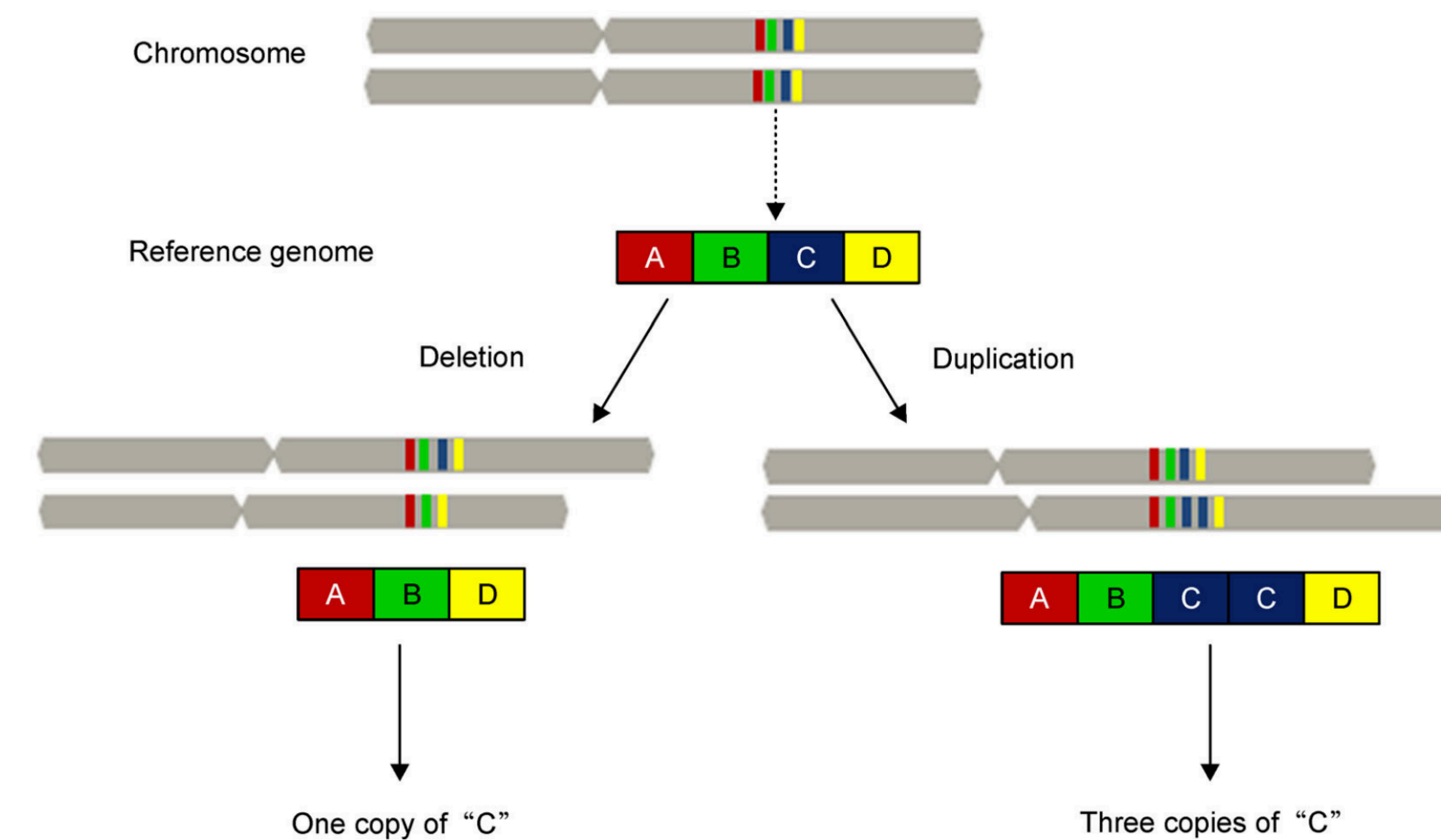
Graphic adapted from Mark Gerstein (GersteinLab.org; @markgerstein)

Somatic CNV in cancer



Davis et al 2017 Biochim Biophys Acta Rev Cancer

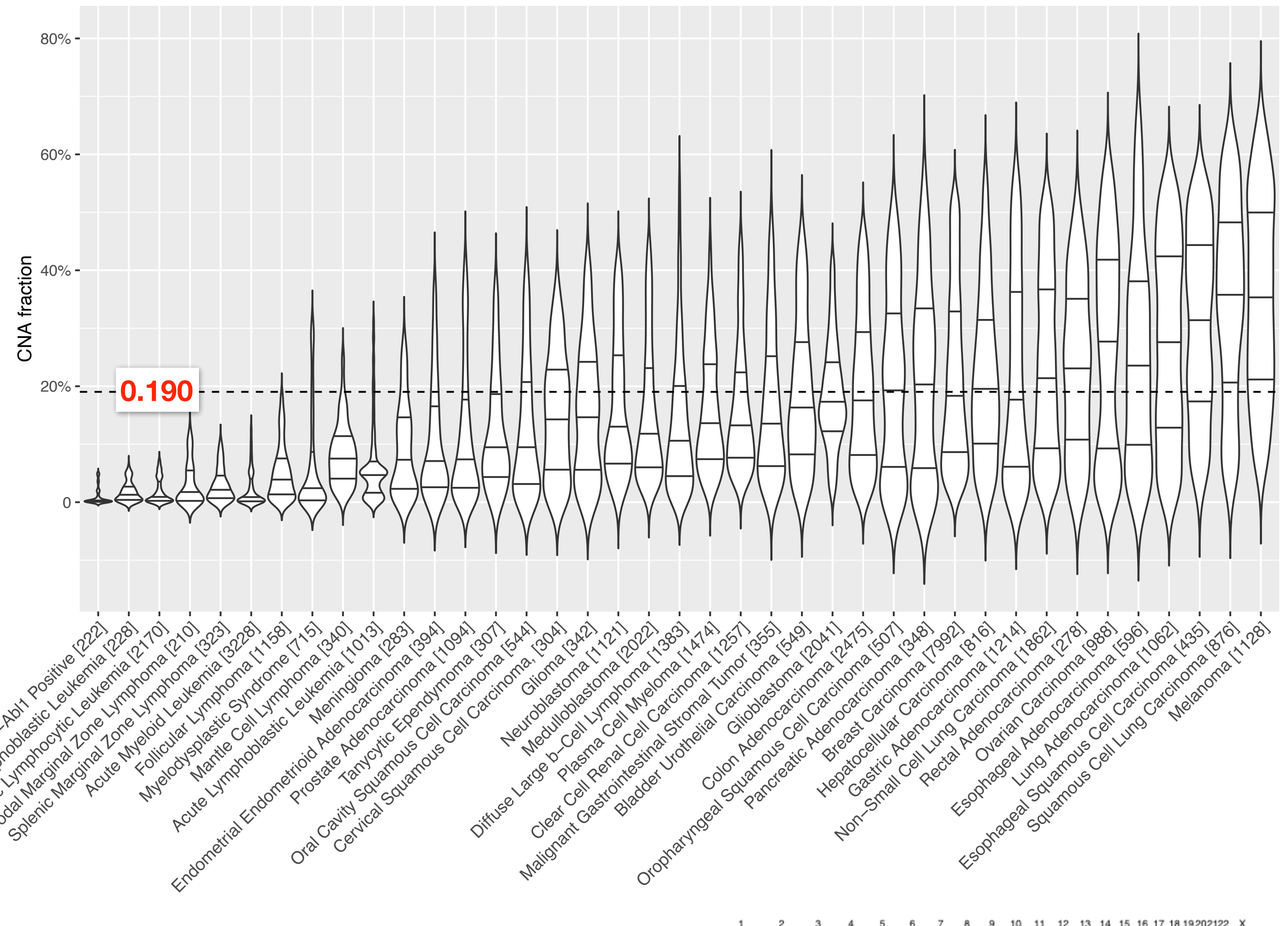
- Point mutations (insertions, deletions, substitutions)
- Structural chromosomal aberrations
- ➔ **Regional Copy Number Variations** (losses, gains)
- Epigenetic changes (e.g. DNA methylation abnormalities)



Aouiche et al 2018 Quant Biol

Genome CNV coverage in Cancer Classes

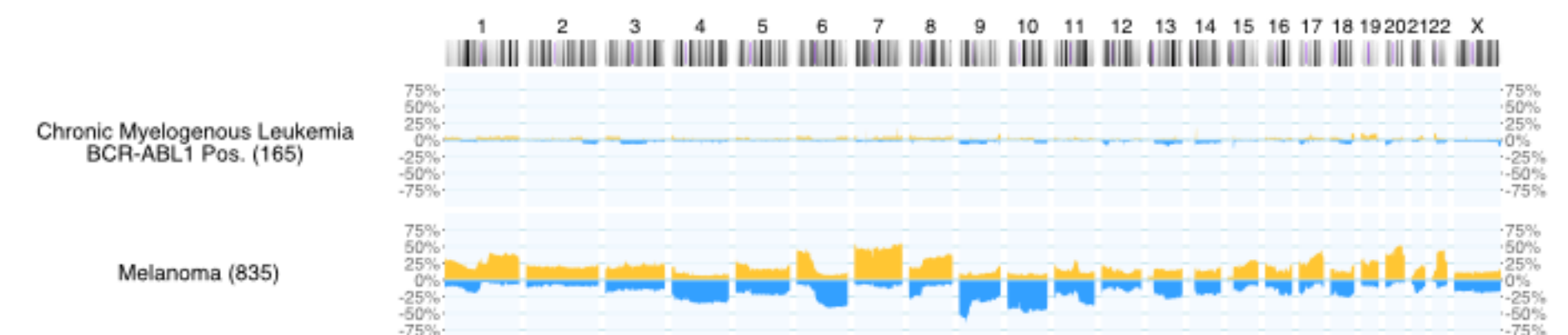
- 43654 out of 93640 CNV profiles; filtered for entities w/ >200 samples (removed some entities w/ high CNV rate, e.g. sarcoma subtypes)
- Single-sample CNV profiles were assessed for the fraction of the genome showing CNVs (relative gains, losses)
- range of medians 0.001 (CML) - 0.358 (malignant melanomas)



Chronic Myelogenous Leukemia Bcr-Abl1 Positive [222]
 T Acute Lymphoblastic Leukemia [228]
 Chronic Lymphocytic Leukemia [2170]
 Nodal Marginal Zone Lymphoma [210]
 Splenic Marginal Zone Lymphoma [323]
 Acute Myeloid Leukemia [3228]
 Follicular Lymphoma [1158]
 Myelodysplastic Syndrome [715]
 Mantle Cell Lymphoma [340]
 Acute Lymphoblastic Leukemia [1013]
 Endometrial Endometrioid Adenocarcinoma [283]
 Prostate Adenocarcinoma [394]
 Tanyocytic Ependymoma [1094]
 Oral Cavity Squamous Cell Carcinoma [307]
 Cervical Squamous Cell Carcinoma [544]
 Diffuse Large b-Cell Lymphoma [304]
 Medulloblastoma [342]
 Plasma Cell Myeloma [1121]
 Clear Cell Renal Cell Carcinoma [2022]
 Malignant Gastrointestinal Stromal Tumor [1383]
 Bladder Urothelial Carcinoma [1474]
 Oropharyngeal Carcinoma [1257]
 Colon Adenocarcinoma [355]
 Glioblastoma [549]
 Pancreatic Adenocarcinoma [2041]
 Squamous Cell Carcinoma [2475]
 Breast Adenocarcinoma [507]
 Hepatocellular Carcinoma [348]
 Gastric Adenocarcinoma [7992]
 Non-Small Cell Lung Carcinoma [816]
 Rectal Adenocarcinoma [1214]
 Esophageal Adenocarcinoma [1862]
 Ovarian Carcinoma [278]
 Lung Adenocarcinoma [988]
 Esophageal Squamous Cell Carcinoma [596]
 Squamous Cell Lung Carcinoma [1062]
 Squamous Cell Lung Carcinoma [435]
 Melanoma [876]
 Melanoma [1128]



Lowest / Highest CNV fractions =>



Somatic Mutations In Cancer: Patterns

Making the case for genomic classifications

Some related cancer entities show similar copy number profiles

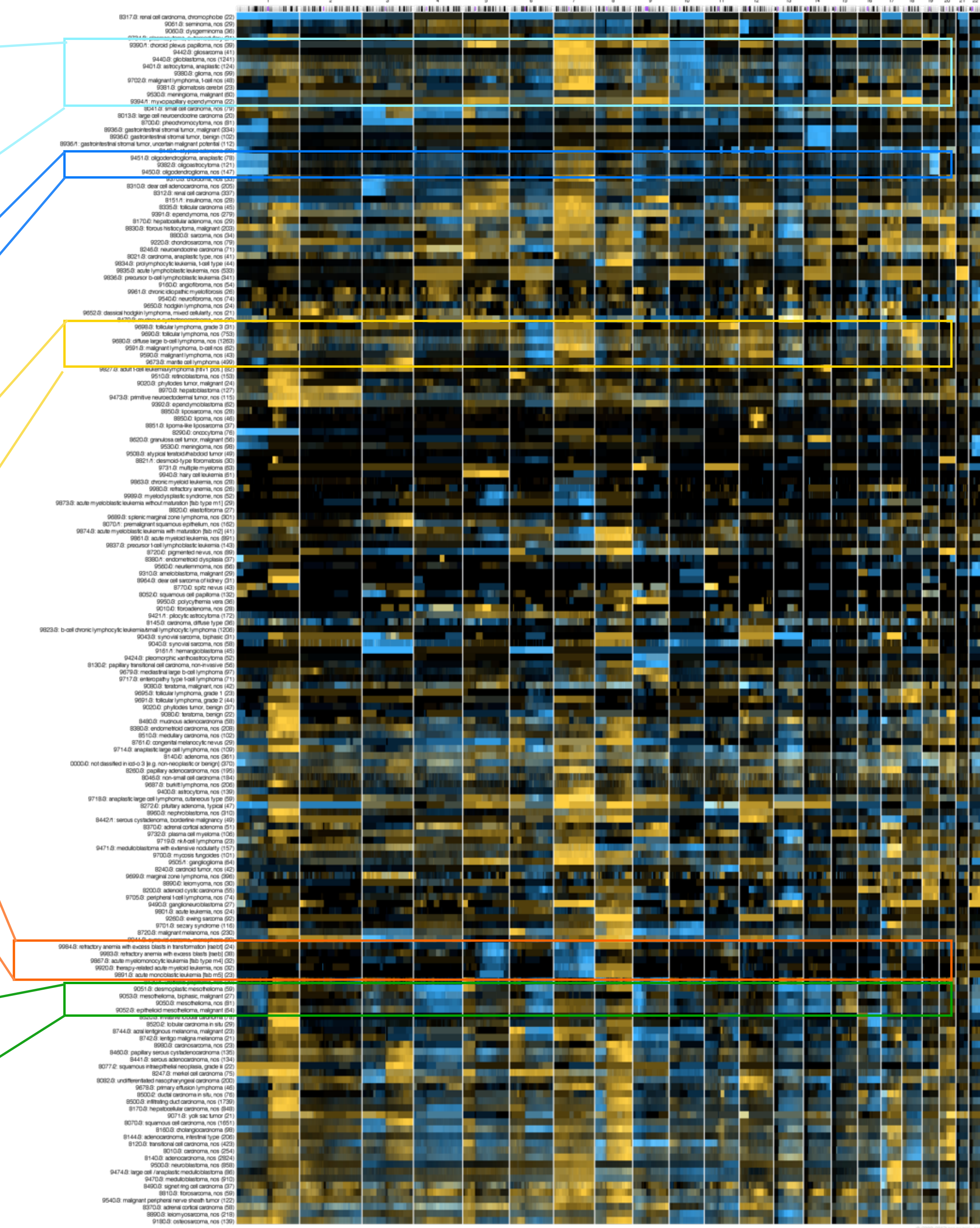
- 9390/1: choroid plexus papilloma, nos (39)
- 9442/3: gliosarcoma (41)
- 9440/3: glioblastoma, nos (1241)
- 9401/3: astrocytoma, anaplastic (124)
- 9380/3: glioma, nos (99)
- 9702/3: malignant lymphoma, t-cell nos (48)
- 9381/3: gliomatosis cerebri (23)
- 9530/3: meningioma, malignant (60)
- 9394/1: myxopapillary ependymoma (22)

- 9451/3: oligodendroglioma, anaplastic (78)
- 9382/3: oligoastrocytoma (121)
- 9450/3: oligodendroglioma, nos (147)

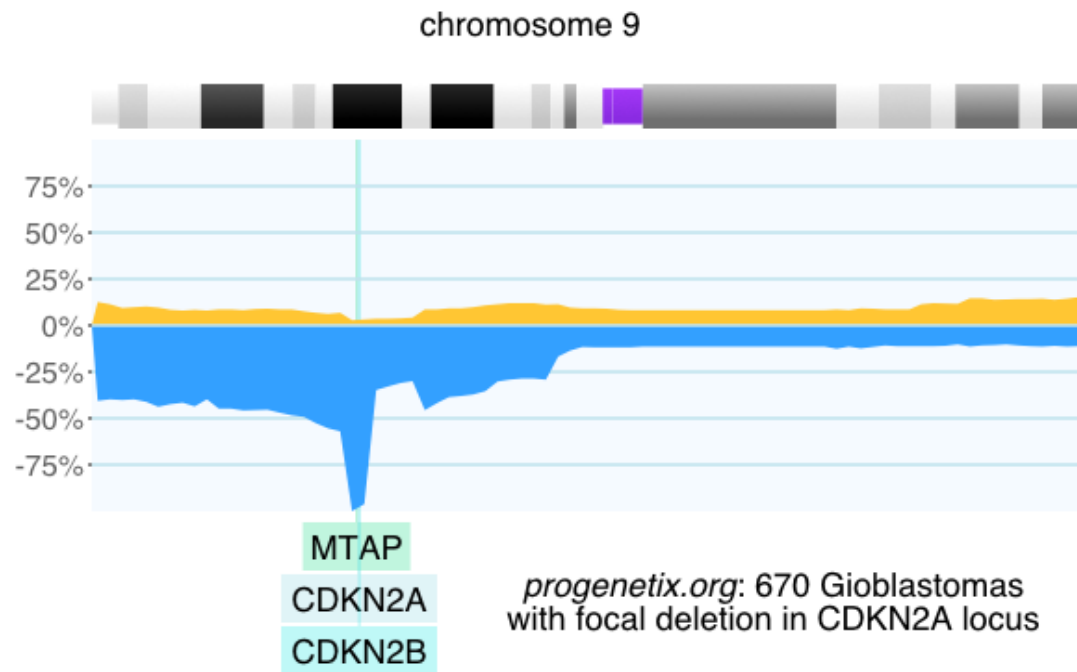
- 9698/3: follicular lymphoma, grade 3 (31)
- 9690/3: follicular lymphoma, nos (753)
- 9680/3: diffuse large b-cell lymphoma, nos (1263)
- 9591/3: malignant lymphoma, b-cell nos (62)
- 9590/3: malignant lymphoma, nos (43)
- 9673/3: mantle cell lymphoma (499)

- 9984/3: refractory anemia with excess blasts in transformation [raebt] (24)
- 9983/3: refractory anemia with excess blasts [raeb] (38)
- 9867/3: acute myelomonocytic leukemia [fab type m4] (32)
- 9920/3: therapy-related acute myeloid leukemia, nos (32)
- 9891/3: acute monoblastic leukemia [fab m5] (23)

- 9051/3: desmoplastic mesothelioma (59)
- 9053/3: mesothelioma, biphasic, malignant (27)
- 9050/3: mesothelioma, nos (81)
- 9052/3: epithelioid mesothelioma, malignant (64)

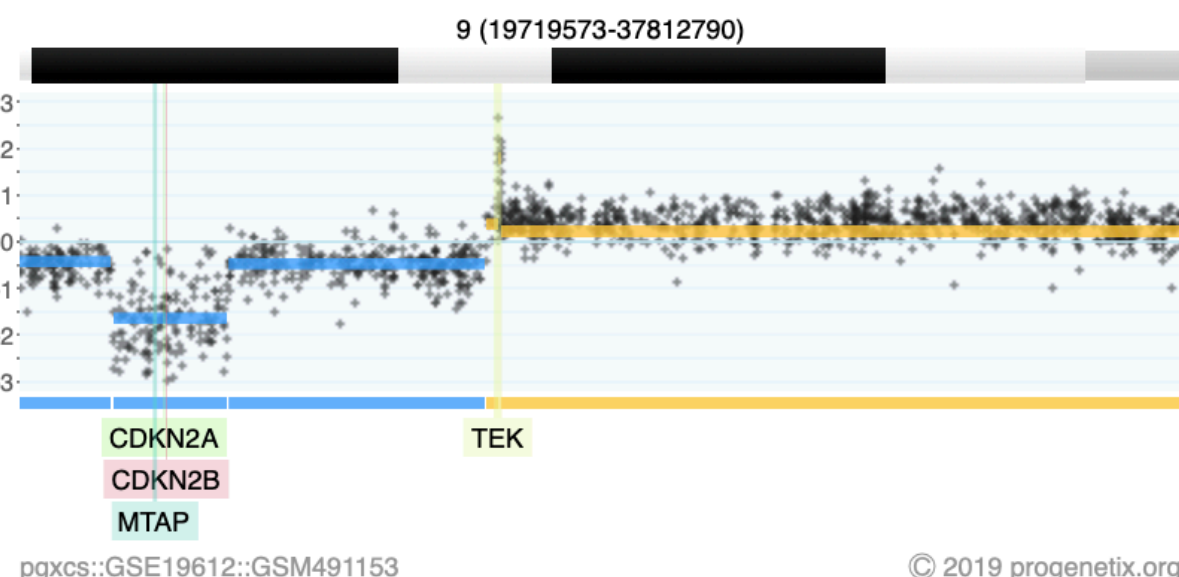
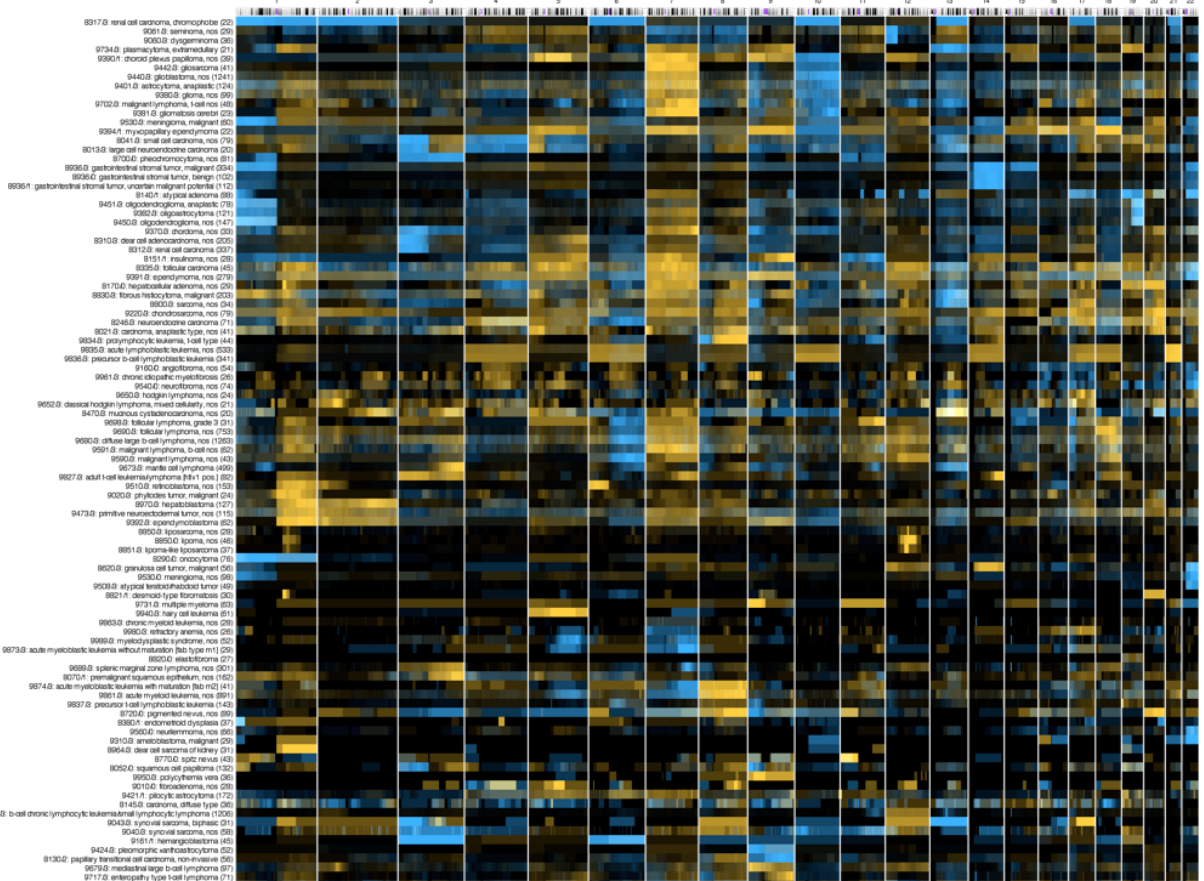


Theoretical Cytogenetics and Oncogenomics Research | Methods | Standards

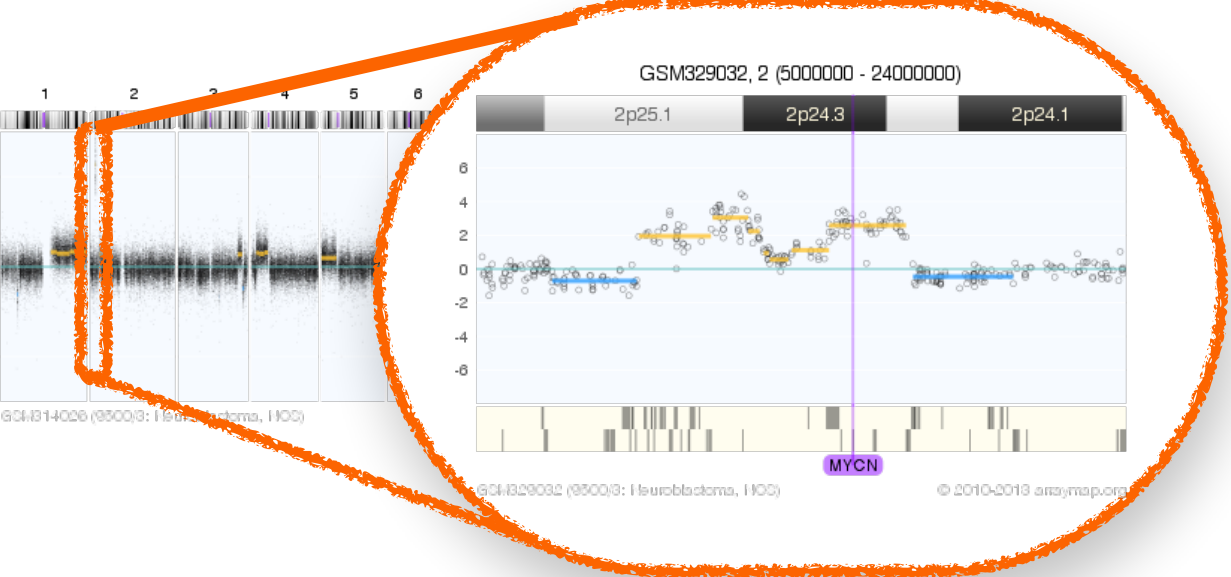


Curators
~~Data Parasites~~

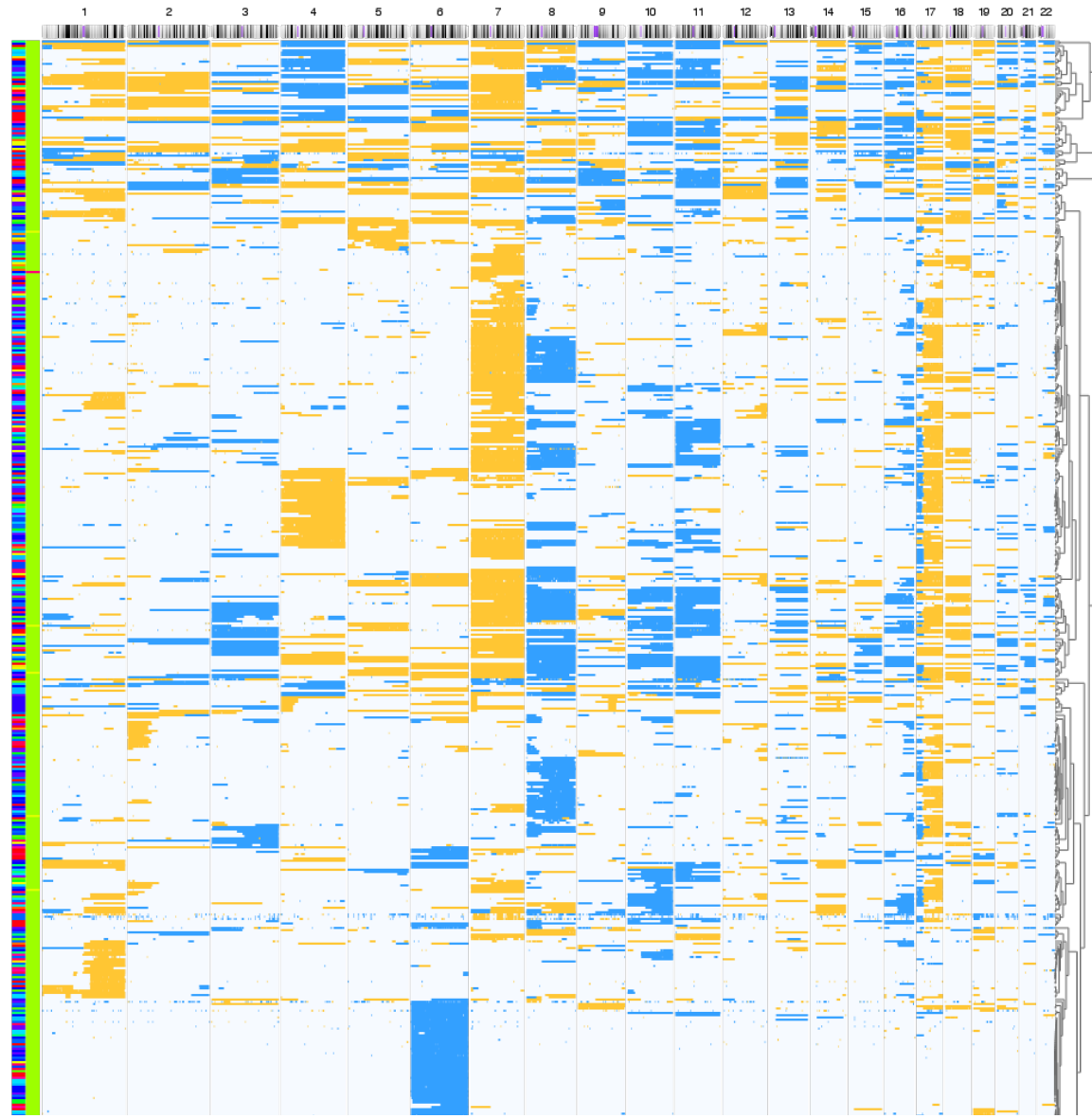
Genomic Imbalances in Cancer Copy Number Variations (CNV)



2-event, homozygous deletion in a Glioblastoma



MYCN amplification in neuroblastoma (GSM314026, SJNB8_N cell line)



Cancer Genomics Reference Resource

- **open** resource for oncogenomic profiles
- over **116'000** cancer **CNV** profiles
- more than 800 diagnostic types
- inclusion of reference datasets (e.g. TCGA)
- standardized encodings (e.g. NCIt, ICD-O 3)
- identifier mapping for PMID, GEO, Cellosaurus, TCGA, cBioPortal where appropriate
- core clinical data (TNM, sex, survival ...)
- data mapping services
- recent addition of SNV data for some series

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⁺

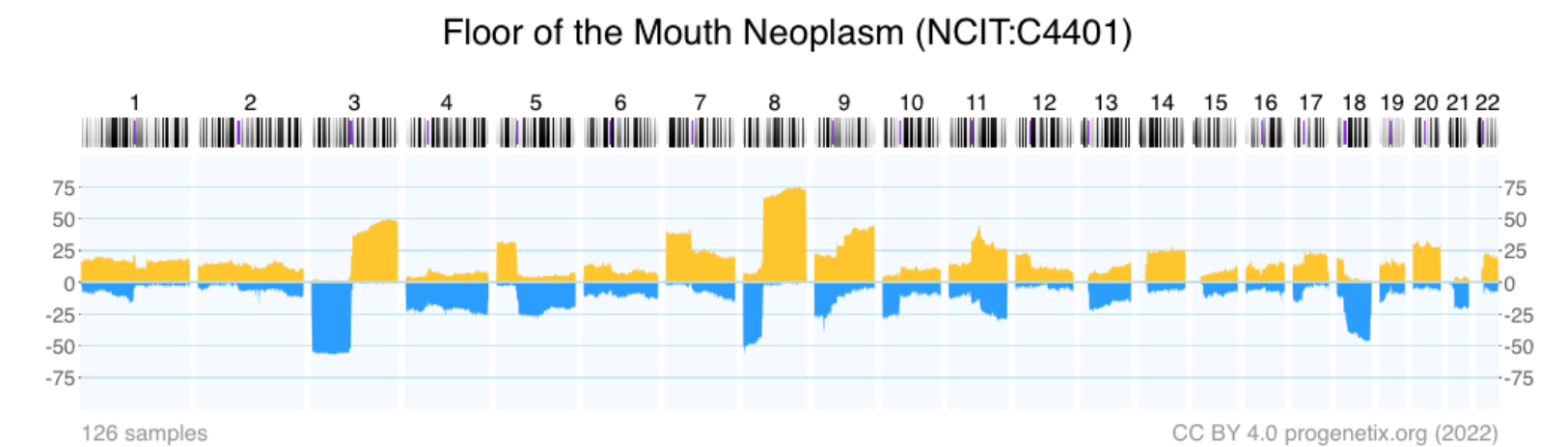
Documentation

News
Downloads & Use
Cases
Services & 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.



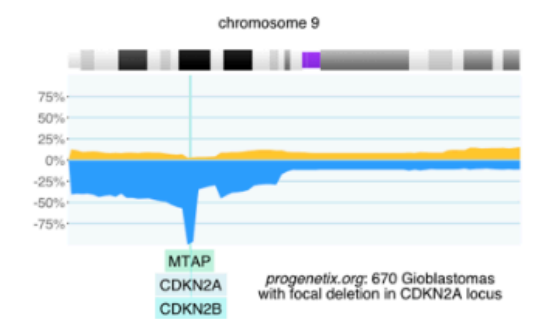
[Download SVG](#) | [Go to NCIT:C4401](#) | [Download CNV Frequencies](#)

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 (NCIt 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.

Cancer Genomics Reference Resource

- **open** resource for oncogenomic profiles
- over 116'000 cancer CNV profiles
- more than **800 diagnostic types**
- inclusion of reference datasets (e.g. TCGA)
- standardized encodings (e.g. NCIt, ICD-O 3)
- identifier mapping for PMID, GEO, Cellosaurus, TCGA, cBioPortal where appropriate
- core clinical data (TNM, sex, survival ...)
- data mapping services
- recent addition of SNV data for some series

Cancer Types by National Cancer Institute NCIt Code

The cancer samples in Progenetix are mapped to several classification systems. For each of the classes, aggregated data is available by clicking the code. Additionally, a selection of the corresponding samples can be initiated by clicking the sample number or selecting one or more classes through the checkboxes.

Sample selection follows a hierarchical system in which samples matching the child terms of a selected class are included in the response.

Filter subsets e.g. by prefix

Hierarchy Depth: 4 levels

No Selection

- ▼ **NCIT:C3262: Neoplasm** (144956 samples, 118106 CNV profiles)
 - ▶ **NCIT:C3263: Neoplasm by Site** (112295 samples, 111637 CNV profiles)
 - ▶ **NCIT:C000000: Unplaced Entities** (27417 samples, 1219 CNV profiles)
 - ▼ **NCIT:C4741: Neoplasm by Morphology** (110745 samples, 110092 CNV profiles)
 - ▶ **NCIT:C27134: Hematopoietic and Lymphoid C...** (26137 samples, 26137 CNV profiles)

Head and Neck Squamous Cell Carcinoma (NCIT:C34447)

Subset Type

- NCI Thesaurus OBO Edition [NCIT:C34447](#)

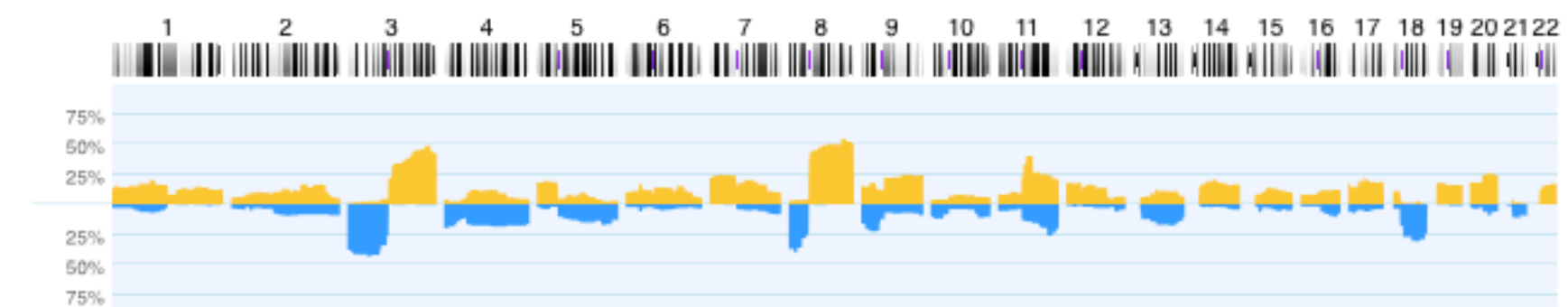
Sample Counts

- 2061 samples
- 57 direct *NCIT:C34447* code matches
- 200 CNV analyses
 - [Download CNV frequencies](#)

Search Samples

Select *NCIT:C34447* samples in the [Search Form](#)

Raw Data (click to show/hide)



progenetix.org

Cancer Genomics Reference Resource

- **open** resource for oncogenomic profiles
- over 116'000 cancer CNV profiles
- more than 800 diagnostic types
- inclusion of reference datasets (e.g. TCGA)
- standardized encodings (e.g. NCIt, ICD-O 3)
- identifier mapping for PMID, GEO, Cellosaurus, TCGA, cBioPortal where appropriate
- core clinical data (TNM, sex, survival ...)
- data mapping services
- recent addition of SNV data for some series



Edit Query

Assembly: GRCh38 Chro: refseq:NC_000009.12 Start: 21500001-21975098
End: 21967753-22500000 Type: EFO:0030067 Filters: NCIT:C3058

progenetix

Matched Samples: 657
Retrieved Samples:
Variants: 276
Calls: 659

[UCSC region](#)

[Variants in UCSC](#)

[Dataset Responses \(JSON\)](#)

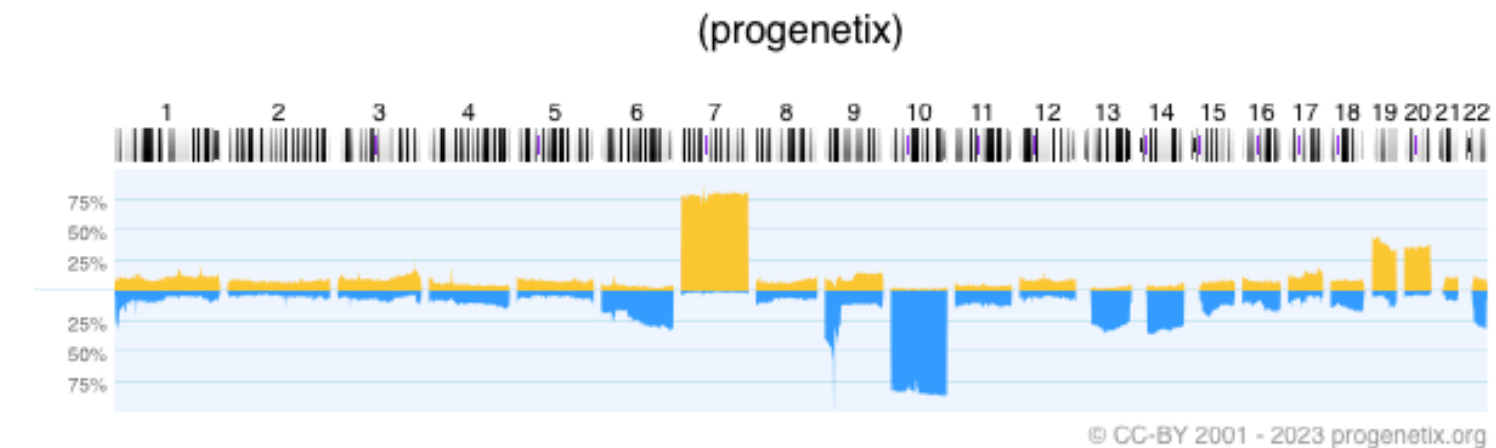
Visualization options

Results

Biosamples

Biosamples Map

Variants



[Reload histogram in new window](#)

Matched Subset Codes	Subset Samples	Matched Samples	Subset Match Frequencies
pgx:icdot-C71.4	4	1	0.250
pgx:icdom-94403	4286	653	0.152
NCIT:C3058	4370	653	0.149
pgx:icdot-C71.1	14	2	0.143
pgx:icdot-C71.9	7204	640	0.089
NCIT:C3796	84	4	0.048
pgx:icdom-94423	84	4	0.048
pgx:icdot-C71.0	1714	14	0.008

Download Sample Data (TSV)

1-657

Download Sample Data (JSON)

1-657

Progenetix Use

- CNV data is used e.g. as reference data in cancer genomics studies
- diagnosis specific CNV profiles serve as "fast look-up" in clinical genomics laboratories
- we loosely track publications in in our literature database but there is no systematic check-back mechanism...

Example: 2024 article using Progenetix' *pgxRpi* Beacon/R interface to retrieve & visualize 117'587 cancer CNV profiles for a study into pluripotent stem cells' genomics

Progenetix References

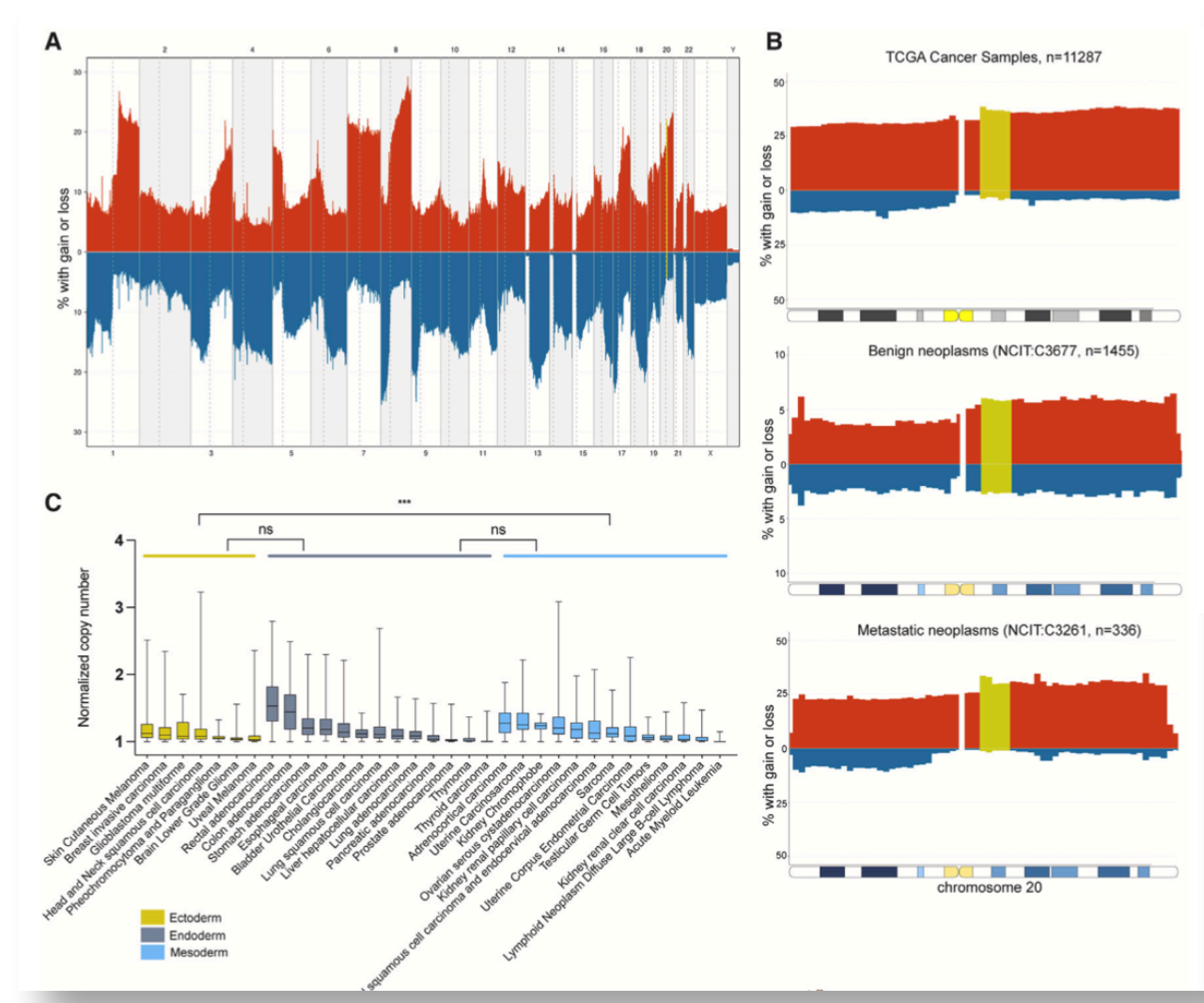
arrayMap | progenetix | cancerCellLines

Articles Citing - or Using - Progenetix

This page lists articles which we found to have made use of, or referred to, the Progenetix resource ecosystem. These articles may not necessarily contain original case profiles themselves. Please [contact us](#) to alert us about additional articles you are aware of. Also, you can now directly submit suggestions for matching publications to the [oncopubs repository on Github](#).

Filter ⓘ

Publications (121)		Samples	
id ⓘ	Publication	Genomes	pgx
PMID:38157850	Krivec N, Ghosh MS et al. (2024) Gains of 20q11.21 in human pluripotent stem cells: Insights from cancer research. ... Stem Cell Reports	0	0
PMID:37627037	Austin BK, Firooz A, Valafar H et al. (2023) An Updated Overview of Existing Cancer Databases and Identified Needs. Biology (Basel)	0	0
PMID:37393410	Liu SC, Wang CI, Liu TT, Tsang NM et al. (2023) A 3-gene signature comprising CDH4, STAT4 and EBV-encoded LMP1 for early diagnosis ... Discov Oncol	0	0



Stem Cell Reports Review



OPEN ACCESS

Gains of 20q11.21 in human pluripotent stem cells: Insights from cancer research

Nuša Krivec,^{1,2} Manjusha S. Ghosh,^{1,2} and Claudia Spits^{1,2,*}
¹Research Group Reproduction and Genetics, Faculty of Medicine and Pharmacy, Vrije Universiteit Brussel, Brussels, Laarbeeklaan 103, 1090 Brussels, Belgium
²These authors contributed equally
 *Correspondence: claudia.spits@vub.be
<https://doi.org/10.1016/j.stemcr.2023.11.013>

Figure 2. Copy-number alterations of human chromosome 20q11.21 in cancers
 (A) Aggregated copy-number variation (CNV) data of 117,587 neoplasms (NCIT: C3262) from the Progenetix database (Huang et al., 2021) were plotted using R library *pgxRpi*. The percentage of samples with aberrations (red, gain; blue, loss) for the whole chromosome are indicated on the y axis. Chromosomal regions are depicted on the x axis; the minimal region of interest at chr20:31216079-35871578 is marked in moss green.
 (B) Top to bottom: Aggregated CNV data of 11,287 TCGA cancer samples, 336 metastatic neoplasms (NCIT: C3261), and 1,455 benign neoplasms (NCIT: C3677) from the Progenetix database (Huang et al., 2021), respectively, were plotted using R library *pgxRpi*. The percentage of samples with aberrations (red, gain; blue, loss) for the whole chromosome are indicated on the y axis. Chromosomal regions are depicted on the x axis; the minimal region of interest at chr20:31216079-35871578 is marked in moss green.

pgxRpi

An interface API for analyzing Progenetix CNV data in R using the Beacon+ API

GitHub: <https://github.com/progenetix/pgxRpi>

Bioconductor

README.md

pgxRpi

Welcome to our R wrapper package for Progenetix REST API. Please note that a stable internal version is aimed to simplify the process of accessing or analyzing data.

You can install this package from GitHub using the following code:

```
install.packages("devtools")
devtools::install_github("progenetix/pgxRpi")
```

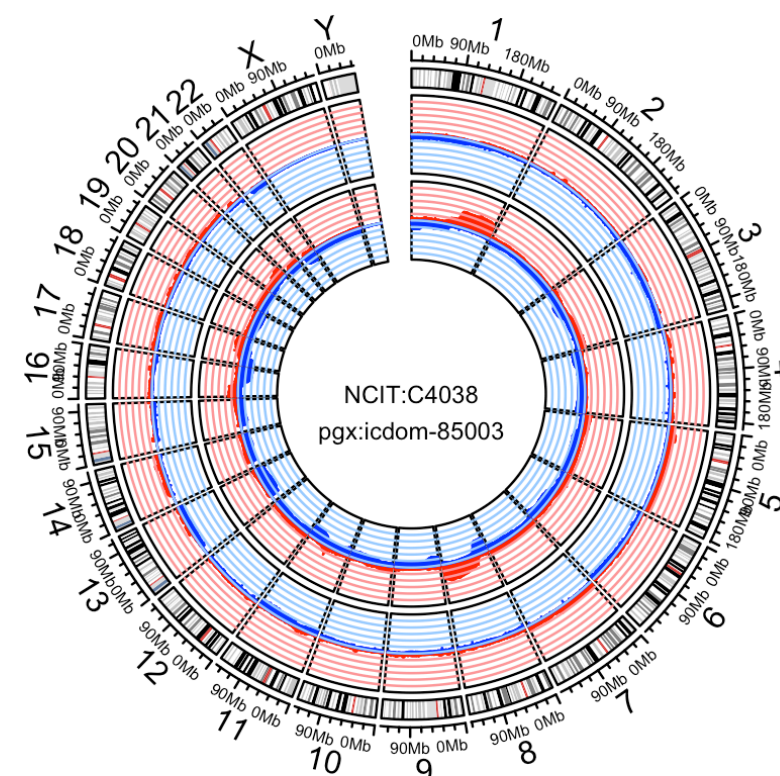
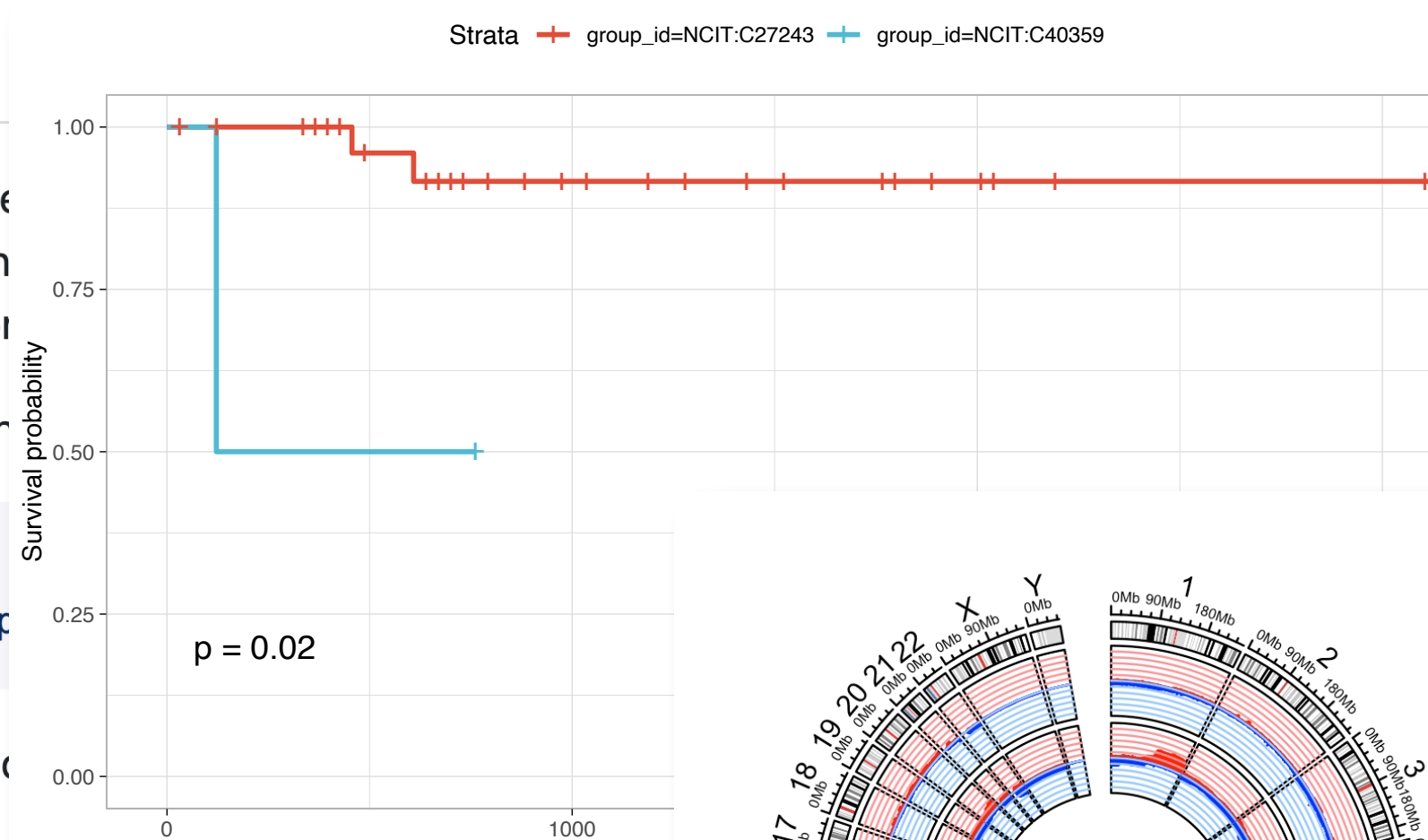
For accessing metadata of biosamples/individuals, get started from this vignette [Introduction_1_loadmetadata](#).

For accessing CNV variant data, get started from this vignette [Introduction_2_loadcnv](#).

For accessing CNV frequency data, get started from this vignette [Introduction_3_loadfreq](#).

For processing local pgxseg files, get started from this vignette [Introduction_4_loadseg](#).

If you encounter problems, try to reinstall the latest version. If reinstallation does not work, please contact the maintainer.



pgxRpi

platforms all rank 2218 / 2221 support 0 / 0 in Bioc level only
build ok updated < 1 month dependencies 144

DOI: [10.18129/B9.bioc.pgxRpi](https://doi.org/10.18129/B9.bioc.pgxRpi)

This is the **development** version of pgxRpi; to use it, please install the [development version](#) of Bioconductor.

R wrapper for Progenetix

Bioconductor version: Development (3.19)

The package is an R wrapper for Progenetix REST API built upon the Beacon v2 protocol. Its purpose is to provide a seamless way for retrieving genomic data from Progenetix database—an open resource dedicated to curated oncogenomic profiles. Empowered by this package, users can effortlessly access and visualize data from Progenetix.

Author: [Hangjia Zhao \[aut, cre\]](#) , [Michael Baudis \[aut\]](#) 

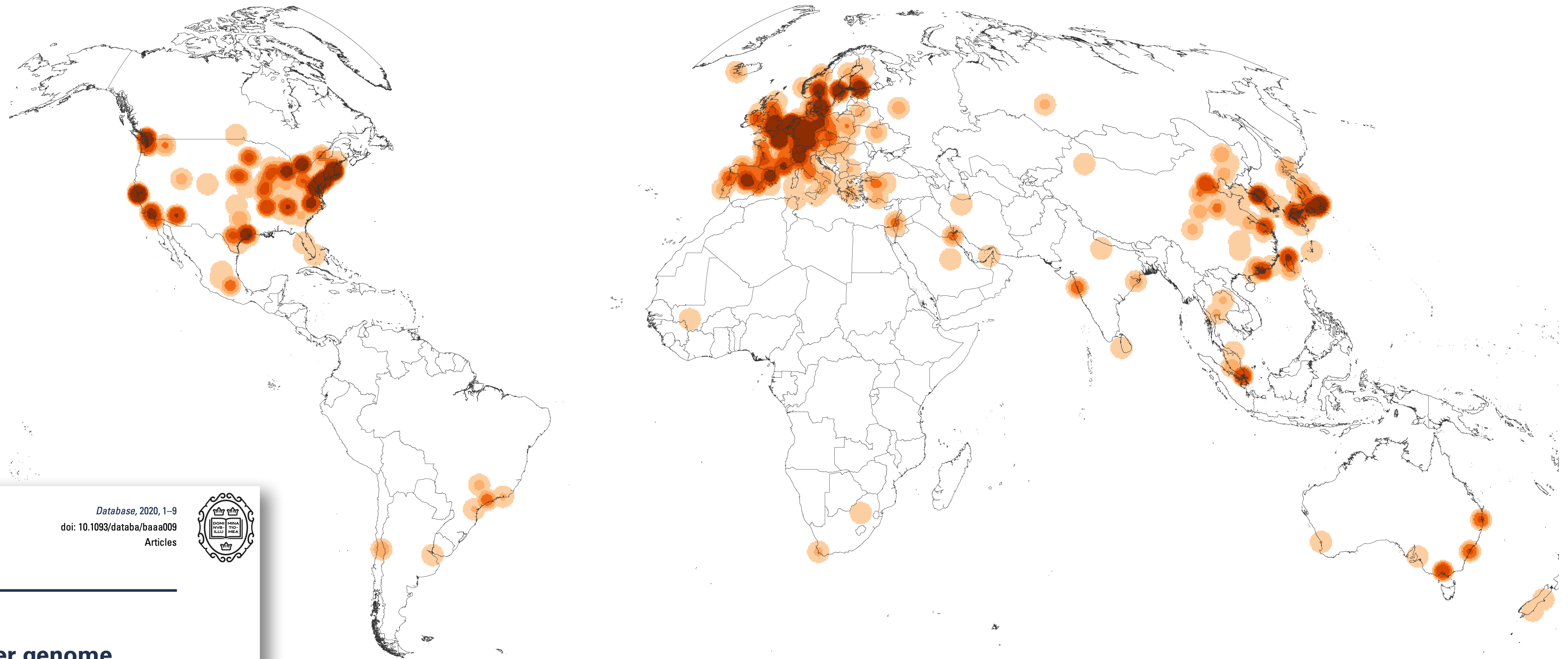
Maintainer: Hangjia Zhao <hangjia.zhao at uzh.ch>

Citation (from within R, enter `citation("pgxRpi")`):

Zhao H, Baudis M (2023). *pgxRpi: R wrapper for Progenetix*. doi:10.18129/B9.bioc.pgxRpi, R package version 0.99.9, <https://bioconductor.org/packages/pgxRpi>.


Where does Genomic Data Come From?

Geographic bias in published cancer genome profiling studies



DATABASE
The Journal of Biological Databases and Curators

Database, 2020, 1–9
doi: 10.1093/databa/baaa009
Articles



Articles

Geographic assessment of cancer genome profiling studies

Paula Carrio-Cordo^{1,2}, Elise Acheson³, Qingyao Huang^{1,2} and Michael Baudis^{1,*}

¹Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland ²Swiss Institute of Bioinformatics, Zurich, Switzerland ³Department of Geography, University of Zurich, Zurich, Switzerland

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.



Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.

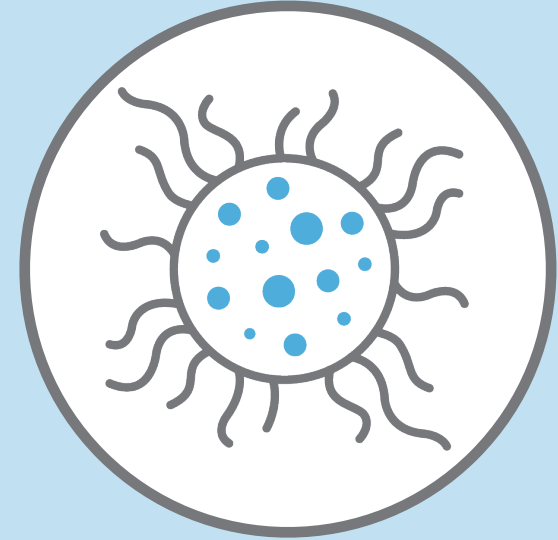
GENOMICS

A federated ecosystem for sharing genomic, clinical data

Silos of genome data collection are being transformed into seamlessly connected, independent systems



Global Genomic Data Sharing Can...



Demonstrate
patterns in health
& disease



Increase statistical
significance of
analyses



Lead to
“stronger” variant
interpretations



Increase
accurate
diagnosis



Advance
precision
medicine

Different Approaches to Data Sharing



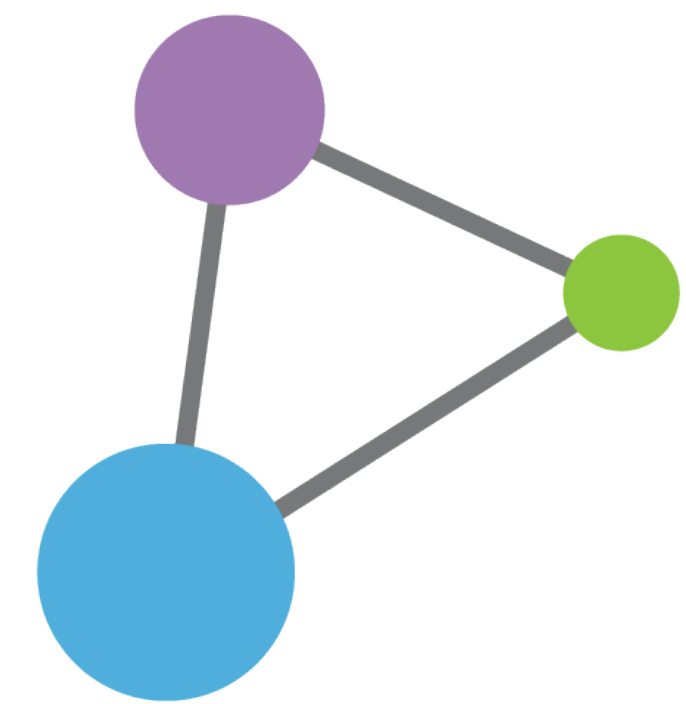
Centralized Genomic Knowledge Bases



Data Commons
Trusted, controlled repository of multiple datasets



Hub and Spoke
Common data elements, access, and usage rules



Linkage of distributed and disparate datasets

Different Approaches to Data Sharing

progenetix



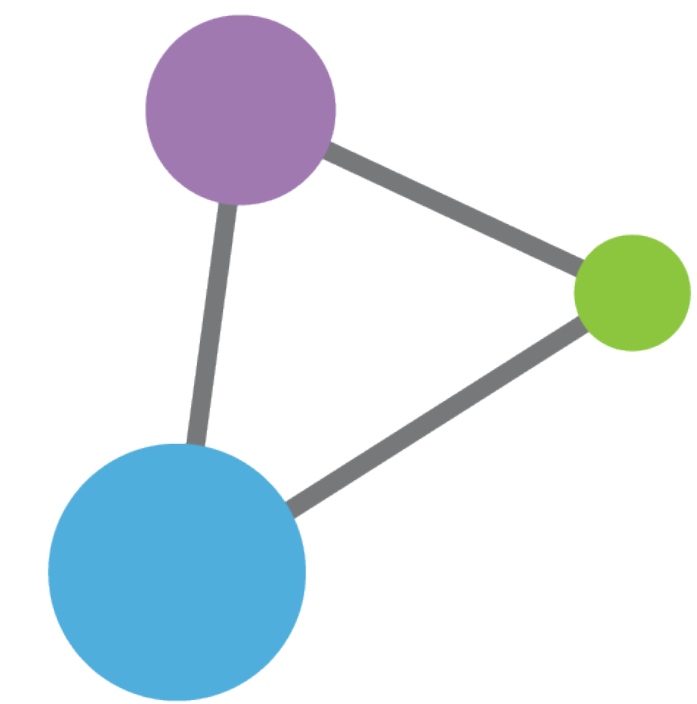
Centralized Genomic Knowledge Bases



Data Commons
Trusted, controlled repository of multiple datasets



Hub and Spoke
Common data elements, access, and usage rules



Linkage of distributed and disparate datasets

Different Approaches to Data Sharing



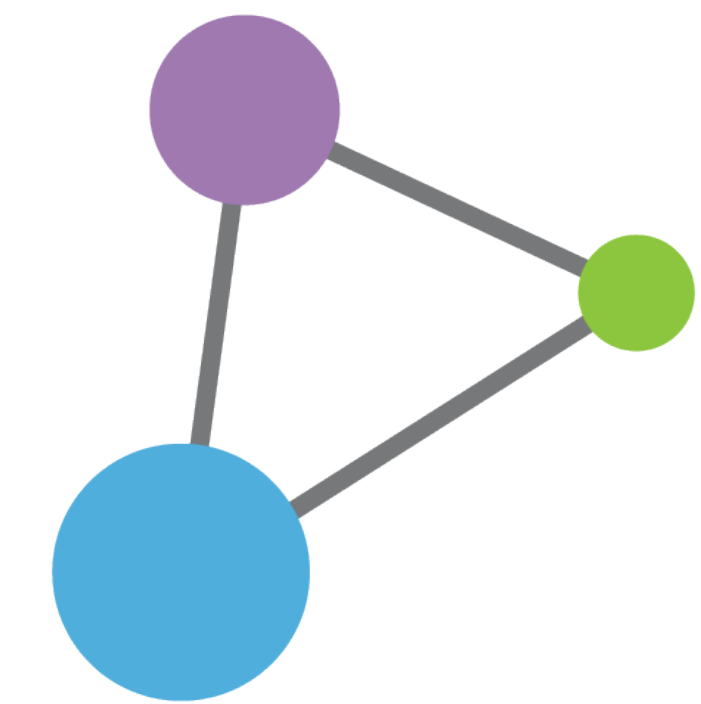
Centralized Genomic Knowledge Bases



Data Commons
Trusted, controlled repository of multiple datasets



Hub and Spoke
Common data elements, access, and usage rules



Linkage of distributed and disparate datasets

The EGA



Long term secure archive for human biomedical research sensitive data, with focus on reuse of the data for further research (or “*broad and responsible use of genomic data*”)



Different Approaches to Data Sharing



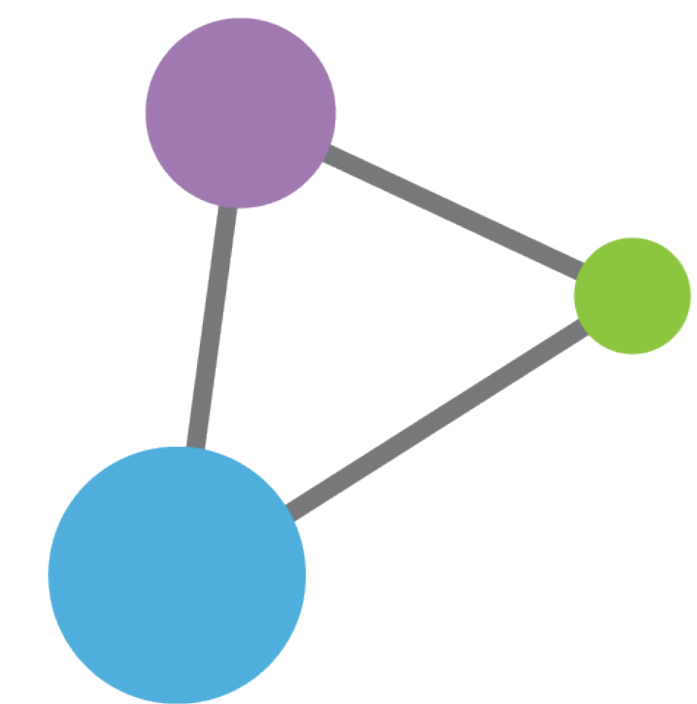
Centralized Genomic Knowledge Bases



Data Commons
Trusted, controlled repository of multiple datasets



Hub and Spoke
Common data elements, access, and usage rules

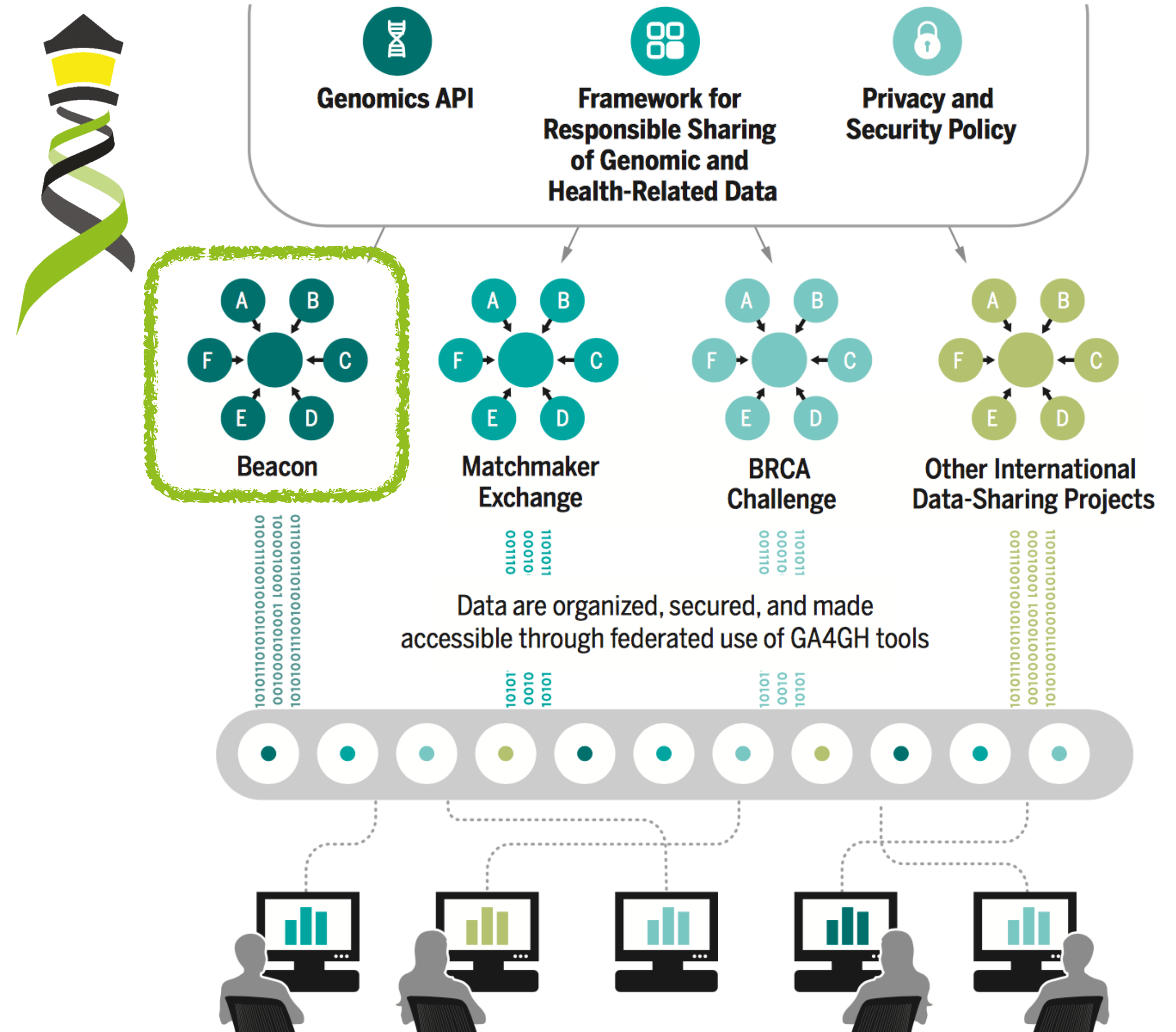


Linkage of distributed and disparate datasets

Federation



A federated data ecosystem. To share genomic data globally, this approach furthers medical research without requiring compatible data sets or compromising patient identity.



GENOMICS

A federated ecosystem for sharing genomic, clinical data

Silos of genome data collection are being transformed into seamlessly connected, independent systems



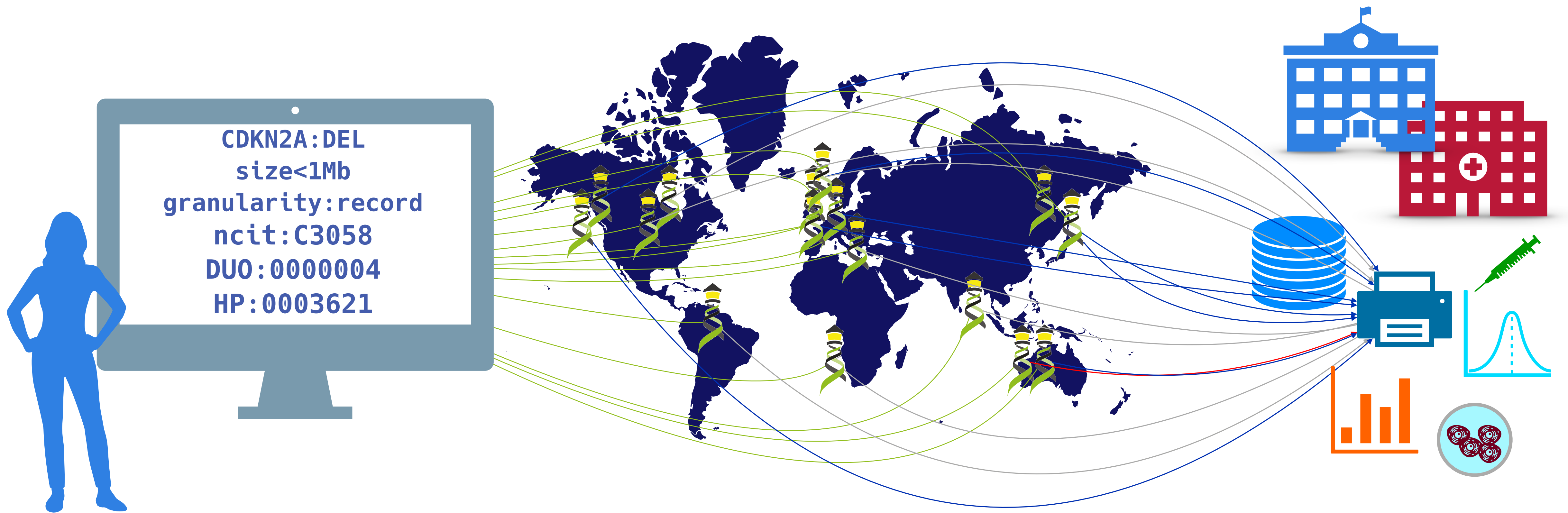


Beacon

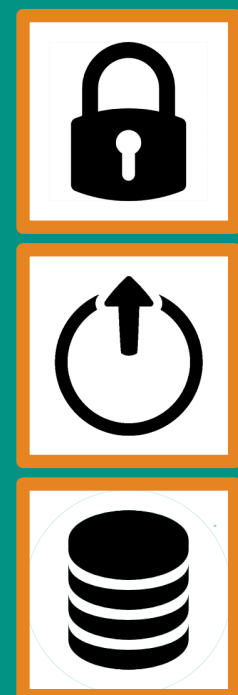


A **Beacon** answers a query for a specific genome variant against individual or aggregate genome collections

YES | **NO** | \0



Can you provide data about focal deletions in CDKN2A in Glioblastomas from juvenile patients with unrestricted access?



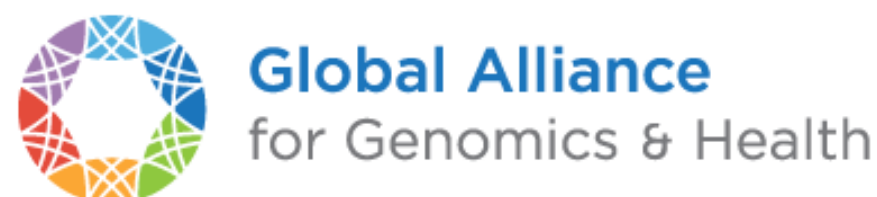
Beacon v2 API

The Beacon API v2 represents a simple but powerful **genomics API** for **federated** data discovery and retrieval

Progenetix & Beacon

Implementation driven standards development

- Progenetix Beacon+ has served as implementation driver since 2016
- prototyping of advanced Beacon features such as
 - ➔ structural variant queries
 - ➔ data handovers
 - ➔ Phenopackets integration
- leading contributor to ELIXIR Beacon network development



The screenshot shows the Progenetix Beacon network resource interface. At the top, there are logos for ELIXIR, the European Union flag, and Fundació "la Caixa", along with the version number V0.5.1. Below the logos is a search bar containing the text "filtering term comma-separated, ID>=<value" and a magnifying glass icon. Underneath the search bar are two tabs: "QUERY EXAMPLES" and "FILTERING TERMS". The main content area features two highlighted boxes. The first box is for the "European Genome-Phenome Archive (EGA) BEACON", with the EGA logo and text: "This beacon is based on the European Genome-Phenome Archive datasets. Now it only contains a public dataset coming from tcga coadread, with samples for patients with colorectal adenocarcinoma disease." Below this box are links for "Beacon API", "Visit us", and "Contact us". The second box is for the "PROGENETIX CANCER GENOMICS BEACON+", with the Progenetix logo and text: "Theoretical Cytogenetics and Oncogenomics group at UZH and SIB". Below this box are also links for "Beacon API", "Visit us", and "Contact us".



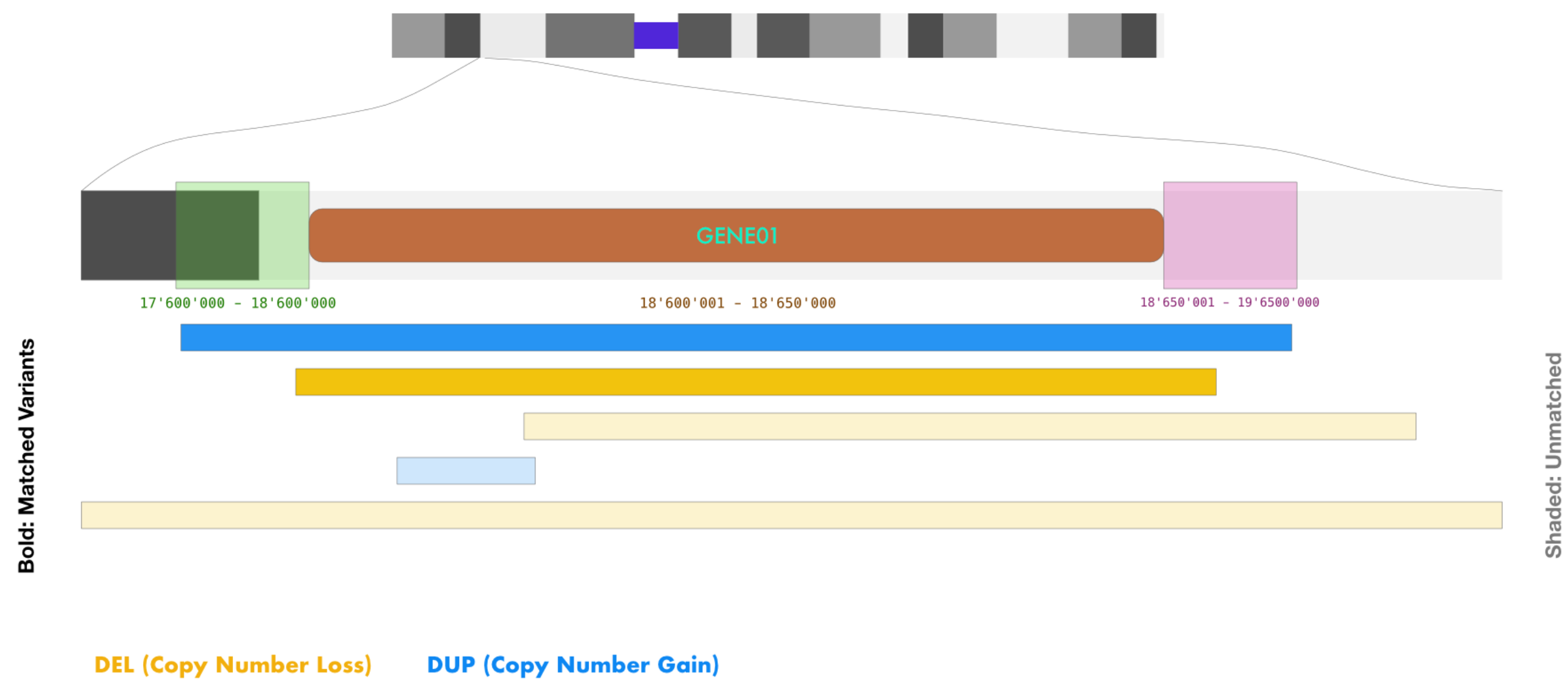
Variation Queries

Bracket ("CNV") Query

- defined through the use of 2 start, 2 end
- any contiguous variant...

Beacon Bracket Query

Example for complete regional match



Beacon Query Types

Sequence / Allele **CNV (Bracket)** Genomic Range Aminoacid Gene ID HGVS Sarr

Dataset: Test Database - examplez x | v

Chromosome: 9 (NC_000009.12) | v Variant Type: EFO:0030067 (copy number deletion) | v

Start or Position: 21000001-21975098 End (Range or Structural Var.): 21967753-23000000

Select Filters: NCIT:C3058: Glioblastoma (100) x | v

Chromosome 9: 21000001, 21975098
21967753 23000000

Query Database

Form Utilities: Gene Spans Cytoband(s)

Query Examples: CNV Example SNV Example Range Example Gene Match
Aminoacid Example Identifier - HeLa

This example shows the query for CNV deletion variants overlapping the CDKN2A gene's coding region with at least a single base, but limited to "focal" hits (here i.e. <= ~2Mbp in size). The query is against the examplez collection and can be modified e.g. through changing the position parameters or data source.

CNV Term Use in Computational (File/Schema) Formats



- Consistent terminologies are essential for cross-resource analyses
- Based on our experience w/ Progenetix together w/ the ELIXIR hCNV community a CNV classes tree was developed (for EFO)
- Terms were adopted by the GA4GH VRS standard
- Consecutive tool development for concordant variant level calling

EFO	Beacon v2	VCF	SO	GA4GH VRS1.3
EFO:0030070 copy number gain	DUP or EFO:0030070	DUP SVCLAIM=D	SO:0001742 copy_number_gain	EFO:0030070 copy number gain
EFO:0030071 low-level copy number gain	DUP or EFO:0030071	DUP SVCLAIM=D	SO:0001742 copy_number_gain	EFO:0030071 low-level gain
EFO:0030072 high-level copy number gain	DUP or EFO:0030072	DUP SVCLAIM=D	SO:0001742 copy_number_gain	EFO:0030072 high-level gain
EFO:0030067 copy number loss	DEL or EFO:0030067	DEL SVCLAIM=D	SO:0001743 copy_number_loss	EFO:0030067 copy number loss
EFO:0030068 low-level copy number loss	DEL or EFO:0030068	DEL SVCLAIM=D	SO:0001743 copy_number_loss	EFO:0030068 low-level loss
EFO:0020073 high-level copy number loss	DEL or EFO:0020073	DEL SVCLAIM=D	SO:0001743 copy_number_loss	EFO:0020073 high-level loss
EFO:0030069 complete genomic loss	DEL or EFO:0030069	DEL SVCLAIM=D	SO:0001743 copy_number_loss	EFO:0030069 complete genomic loss



Briefings in Bioinformatics, 2024, 25(2), 1–12
<https://doi.org/10.1093/bib/bbad541>
 Problem Solving Protocol

labelSeg: segment annotation for tumor copy number alteration profiles

Hangjia Zhao and Michael Baudis
 Corresponding author: Michael Baudis, Department of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland.
 Tel.: (+41) 44 635 34 86; E-mail: michael.baudis@mls.uzh.ch

Progenetix Stack

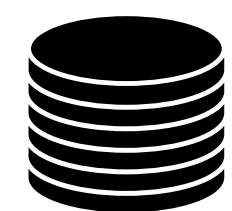


- JavaScript front-end is populated for query results using asynchronous access to multiple handover objects
 - biosamples and variants tables, CNV histogram, UCSC .bed loader, .pgxseg variant downloads...
- the complete middleware / CGI stack is provided through the **bycon** package
 - schemas, query stack, data transformation (e.g. Phenopackets generation)...
- data collections mostly correspond to the main Beacon default model entities
 - no separate *runs* collection; integrated w/ analyses
 - *variants* are stored per observation instance

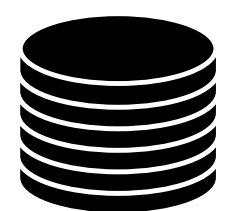


- *collations* contain pre-computed data (e.g. CNV frequencies, statistics) and information for all grouping entity instances and correspond to **filter values**
 - PMID:10027410, NCIT:C3222, pgx:cohort-TCGA, pgx:icdom-94703...
- *querybuffer* stores id values of all entities matched by a query and provides the corresponding access handle for **handover** generation

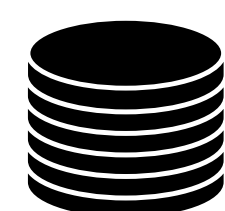
```
_id: ObjectId("6249bb654f8f8d67eb94953b"),
id: '0765ee26-5029-4f28-b01d-9759abf5bf14',
source_collection: 'variants',
source_db: 'progenetix',
source_key: '_id',
target_collection: 'variants',
target_count: 667,
target_key: '_id',
target_values: [
  ObjectId("5bab578b727983b2e0ca99e"),
  ObjectId("5bab578d727983b2e0cb505")
]
```



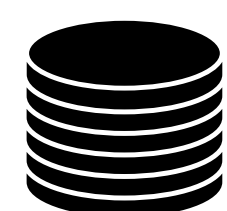
variants



analyses

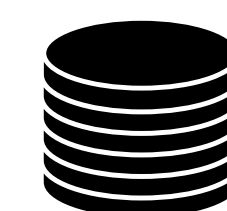


biosamples

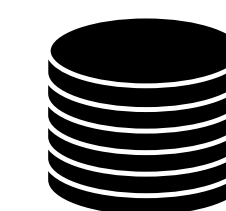


individuals

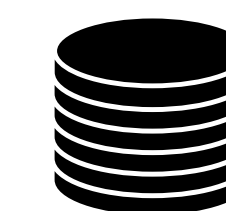
Entity collections



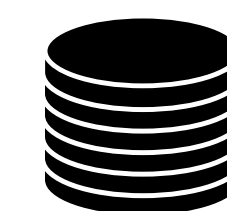
collations



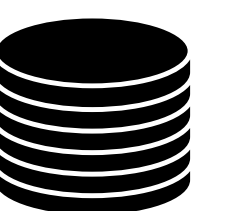
geolocs



genespans



publications



qBuffer

Utility collections

progenetix / byconaut

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

bycon.progenetix.org
github.com/progenetix/bycon/

byconaut Public

main 2 branches

mbaudis get_plot_parameters

- bin
- docs
- exports
- imports
- local
- rsrc
- services
- tmp
- .gitignore
- LICENSE
- README.md
- __init__.py
- install.py
- install.yaml
- mkdocs.yaml

progenetix / beaconplus-web

Code Pull requests Actions Projects Security Insights Settings

beaconplus-web Public

forked from progenetix/progenetix-web

main 1 branch 0 tags

This branch is 44 commits ahead, 24 commits behind progenetix:main.

mbaudis code cleaning, no feature changes

- .github/workflows cleanup
- docs still first implementation clean-up
- extra documentation
- public graphic refinement
- src code cleaning, no feature changes
- .babelrc Simplify query generation and add
- .env.development first working version
- .env.local first working version
- .env.production env
- .env.staging env
- .eslintrc.json BioSubsetsPage perf optimisations

progenetix / bycon

Code Issues Pull requests 1 Actions Projects Wiki Security 3 Insights Settings

bycon Public

main 4 branches 25 tags

mbaudis 1.3.6 be19a12 3 days ago 852 commits

.github/workflows	Create mk-bycon-docs.yaml	8 months ago
bycon	1.3.6	3 days ago
docs	1.3.6	3 days ago
local	1.3.5 preparation	2 weeks ago
.gitignore	Update .gitignore	3 months ago
LICENSE	Create LICENSE	3 years ago
MANIFEST.in	major library & install disentanglement	9 months ago
README.md	#### 2023-07-23 (v1.0.68)	4 months ago
install.py	1.3.6	3 days ago
install.yaml	v1.0.57	5 months ago
mkdocs.yaml	1.1.6	3 months ago
requirements.txt	1.3.6	3 days ago
setup.cfg	...	10 months ago
setup.py	1.3.6	3 days ago
updev.sh	1.3.6	3 days ago

beaconplus.progenetix.org
[.../progenetix/beaconplus-web/](https://github.com/progenetix/beaconplus-web/)

bycon.progenetix.org
github.com/progenetix/bycon/

About

Bycon - A Python Based Beacon API (beacon-project.io) implementation leveraging the Progenetix (progenetix.org) data model

- Readme
- CC0-1.0 license
- Activity
- 5 stars
- 4 watching
- 6 forks
- Report repository

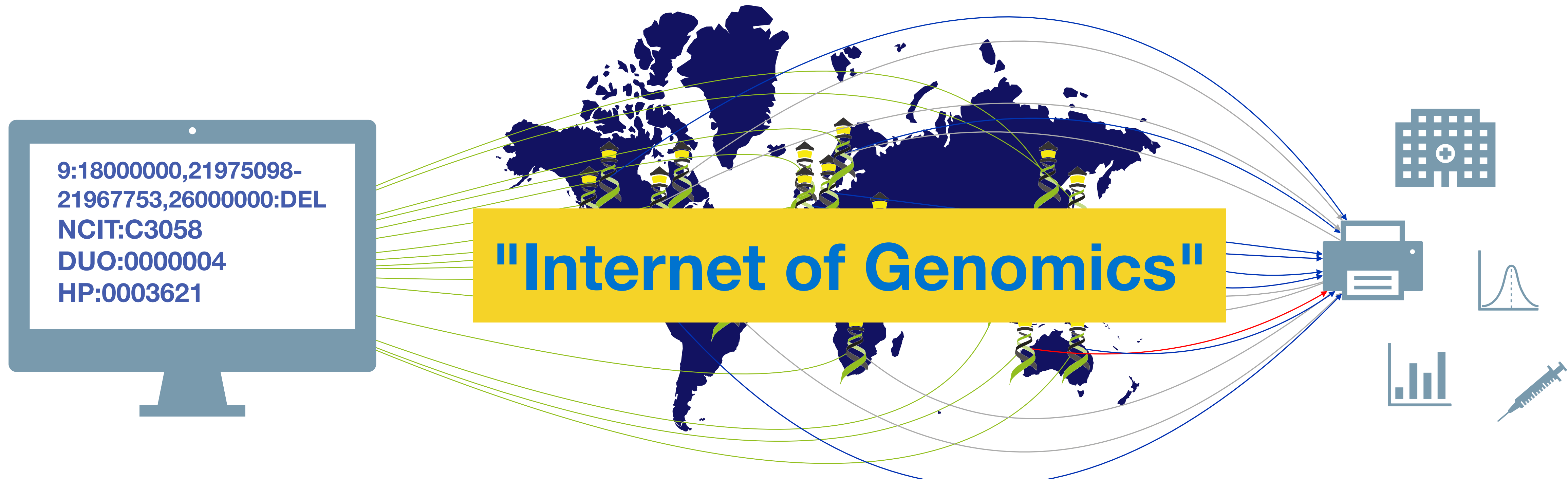
Releases

25 tags

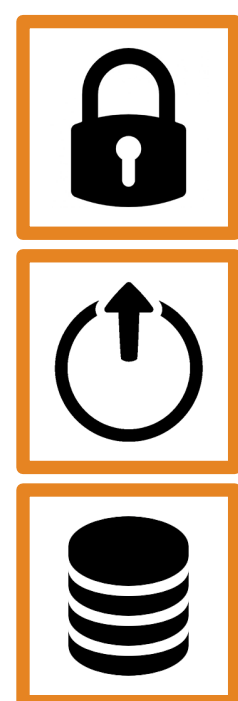
[Create a new release](#)

Packages

No packages published
[Publish your first package](#)



Have you seen deletions in this region on chromosome 9 in Glioblastomas from a juvenile patient, in a dataset with unrestricted access?



Beacon v2 API

The Beacon API v2 proposal opens the way for the design of a simple but powerful **"genomics API"**.

Vague but exciting ...

CERN DD/OC

Tim Berners-Lee, CERN/DD

Information Management: A Proposal

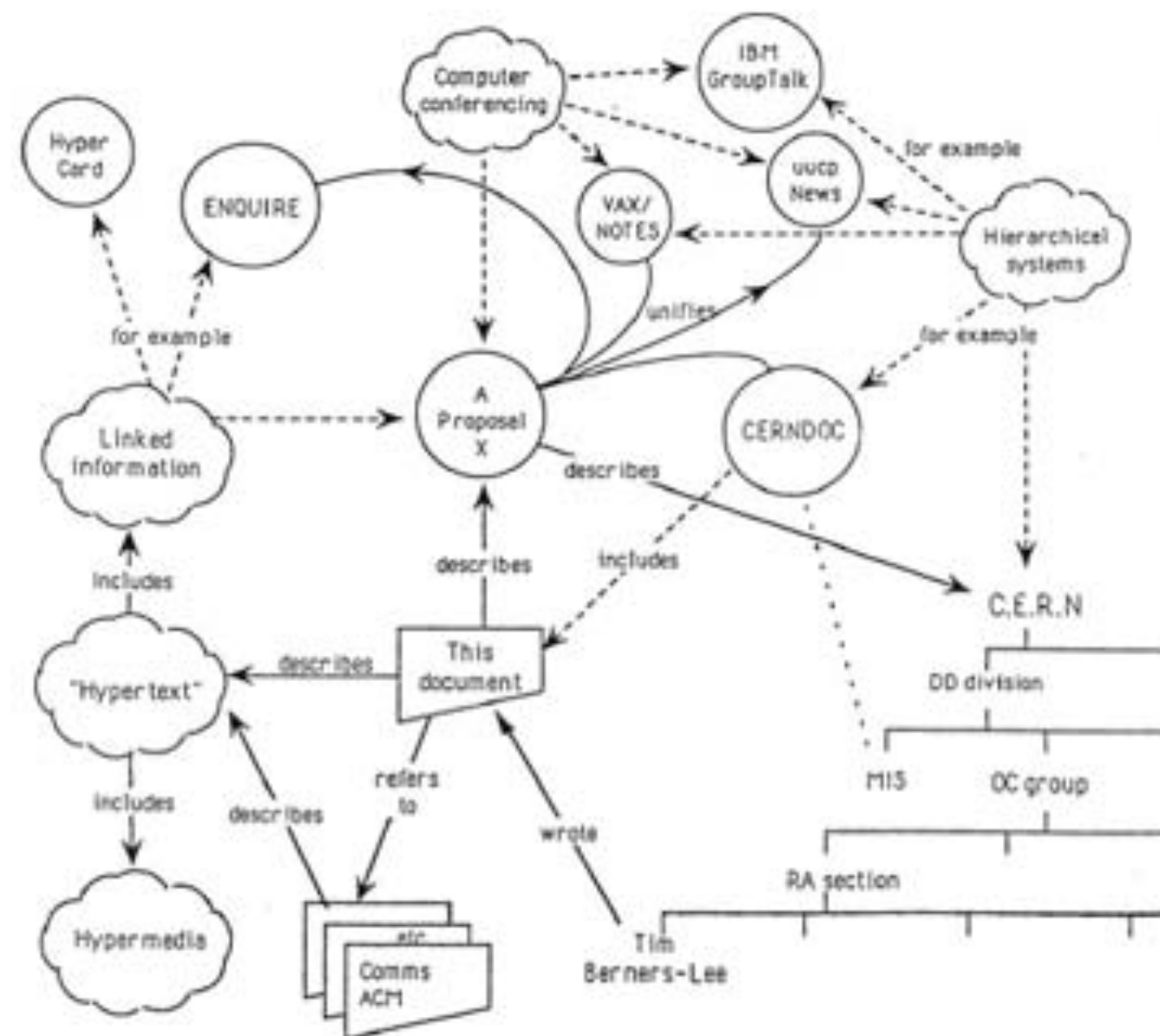
March 1989

Information Management: A Proposal

Abstract

This proposal concerns the management of general information about accelerators and experiments at CERN. It discusses the problems of loss of information about complex evolving systems and derives a solution based on a distributed hypertext system.

Keywords: Hypertext, Computer conferencing, Document retrieval, Information management, Project control



Tim Berners-Lee: Information Management: A Proposal (CERN 1989) & WWW: First Page (1990)

World Wide Web

The WorldWideWeb (W3) is a wide-area [hypermedia](#) information retrieval initiative aiming to give universal access to a large universe of documents.

Everything there is online about W3 is linked directly or indirectly to this document, including an [executive summary](#) of the project, [Mailing lists](#), [Policy](#), November's [W3 news](#), [Frequently Asked Questions](#).

What's out there?

Pointers to the world's online information, [subjects](#), [W3 servers](#), etc.

Help

on the browser you are using

Software Products

A list of W3 project components and their current state. (e.g. [Line Mode](#), [X11 Viola](#), [NeXTStep](#), [Servers](#), [Tools](#), [Mail robot](#), [Library](#))

Technical

Details of protocols, formats, program internals etc

Bibliography

Paper documentation on W3 and references.

People

A list of some people involved in the project.

History

A summary of the history of the project.

How can I help?

If you would like to support the web..

Getting code

Getting the code by [anonymous FTP](#), etc.

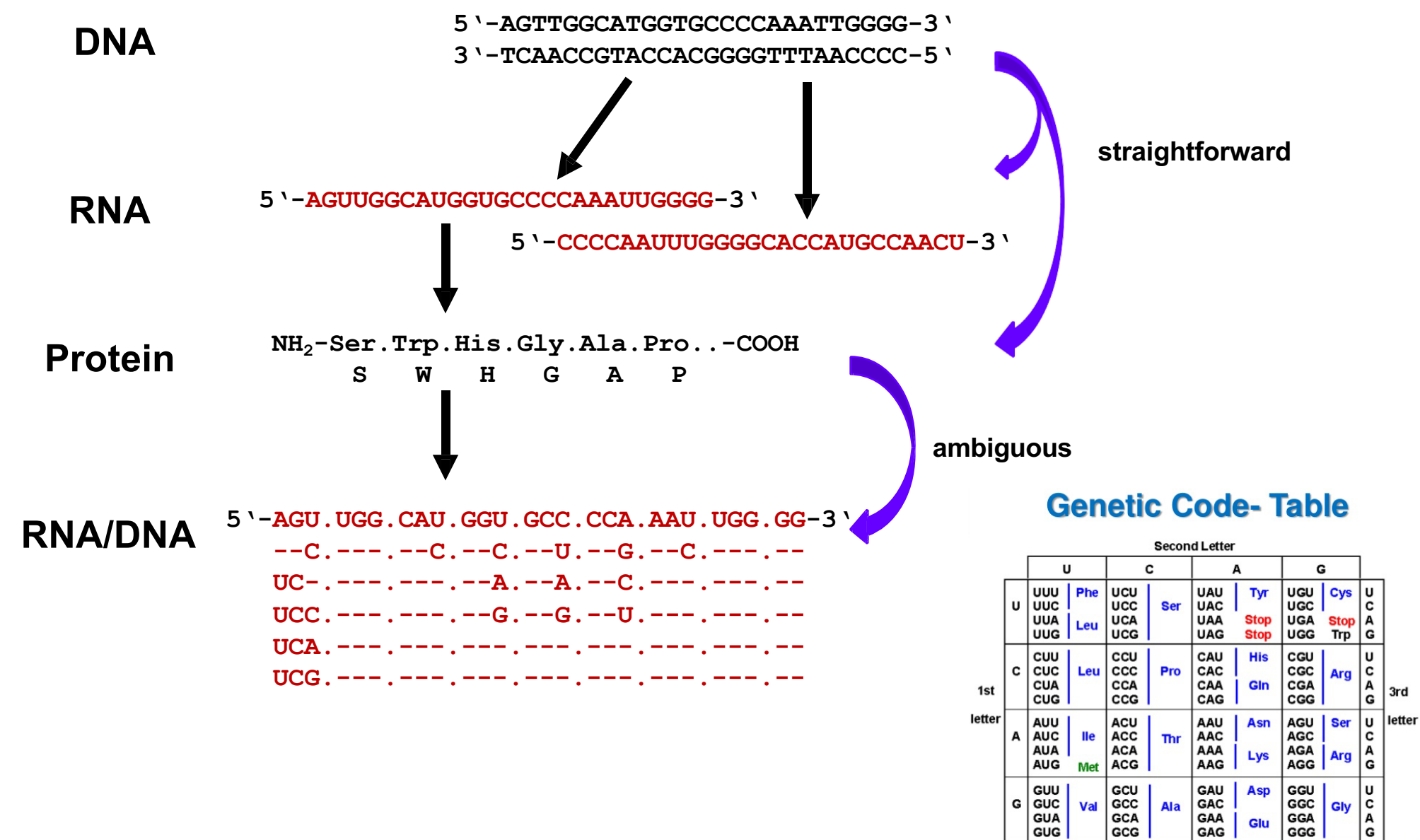
BIO390: Course Schedule

- 2024-09-17: Michael Baudis - What is Bioinformatics? Introduction and Resources
- 2024-09-24: Mark Robinson - Statistical Bioinformatics
- 2024-10-01: Christian von Mering - Sequence Bioinformatics
- 2024-10-08: Valentina Boeva (ETHZ) - Machine Learning for Biological Use Cases
- 2024-10-15: Izaskun Mallona - Regulatory Genomics and Epigenomics
- 2024-10-22: Shinichi Sunagawa (ETHZ) - Metagenomics
- 2024-10-29: Katja Baerenfaller (SIAF) - Proteomics
- 2024-11-05: Patrick Ruch - Text mining & Search Tools
- 2024-11-07: Andreas Wagner - Biological Networks
- 2024-11-19: Ahmad Aghaebrahimian (ZHAW) - Semantic Web
- 2024-11-26: Qingyao Huang - Building Biological Information Resources
- 2024-12-03: Valérie Barbie (SIB) - Clinical Bioinformatics
- 2024-12-10: Michael Baudis - Genome Data & Privacy | Feedback
- 2024-12-17: Exam (Multiple Choice)

Biological Sequence Informatics

Christian von Mering

Sequences can be interconverted computationally



Sequence Similarity

Many possible definitions of "similarity": length, character content, character distribution,.....

Biological definition: (interrupted) stretches of **identical** or **similar** characters

E.g. search **identical sequence segments** for assembly of long sequences from short, overlapping fragments

AAGCTTACC AAAATTGAAGGGACGTTGACGTAGGGG**GACGCTTTAG**
GACGCTTTAGTTTAGCCACCGGTATTTAGC

Similar characters: physico-chemical characteristics, functional characteristics, evolutionary relation.....

Comparison of two (or more) sequences: **Alignment** of **identical** and **similar** sequence segments

AAGCTTACC AAAATTGAAGGGACGTTGACGTAGGGG**GACGCTTTAG**
AATCTAGCAATTAT**TGAAGGGACGTTGACGAAGGGGTT**CGCTACCG

Challenge: Find the best possible alignment **(and do it fast)**

AAGCT**T**ACC AAAATTGAAGGGACGTTGACGTAGGGG**GACGCTTTAG**
AATCTAG**C**AAATTAT**TGAAGGGACGTTGACGAAGGGGTT**CGCTACCG

Statistical Bioinformatics

Mark Robinson



University of Zurich

Statistical Bioinformatics // Institute of Molecular Life Sciences

False positives / false negatives

Most statistical testing regimes set an error rate (5%)

Type I error = false positive
Type II error = false negative

Arthur Charpentier @freakonometrics

Statistical Errors

	$\hat{Y} = 0$ NEGATIVE	$\hat{Y} = 1$ POSITIVE
$Y = 0$ NOT PREGNANT	TRUE NEGATIVE You're not pregnant	FALSE POSITIVE You're pregnant TYPE 1 ERROR
$Y = 1$ PREGNANT	FALSE NEGATIVE You're not pregnant TYPE 2 ERROR	TRUE POSITIVE You're pregnant

<https://twitter.com/freakonometrics/status/779060142239260672>

40



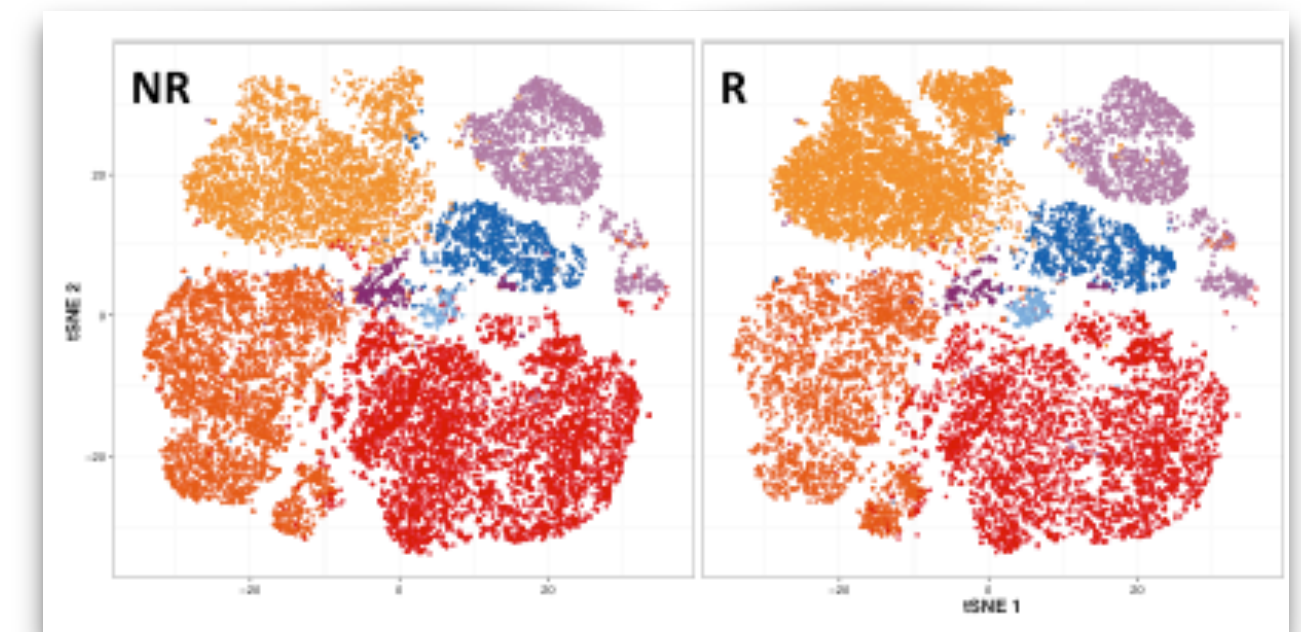
University of Zurich

Statistical Bioinformatics // Institute of Molecular Life Sciences

Differential abundance of cell populations

tSNE projection
(each dot = cell, cells from multiple patients)

NR: non-responders
R: responders



Under the hood: Generalized linear mixed model to assess the change in relative abundance of subpopulations.

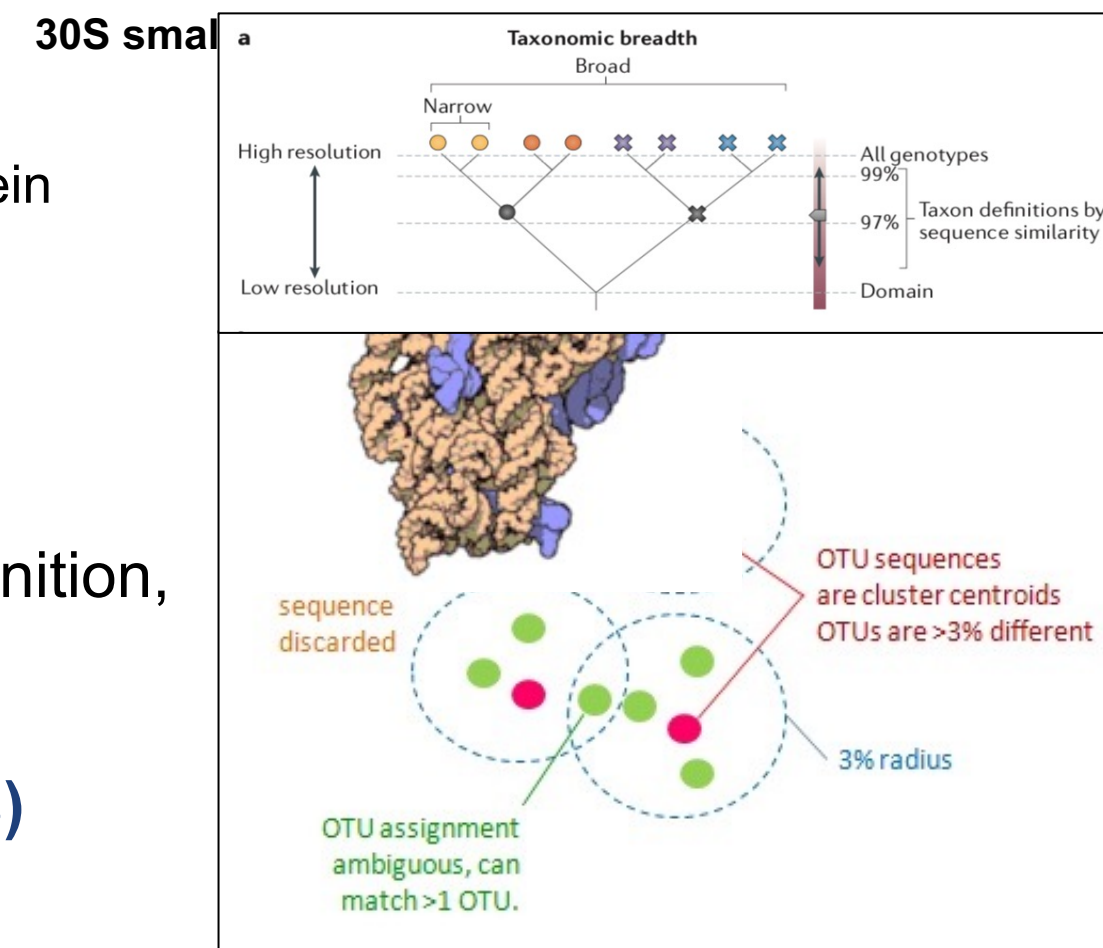
30

Metagenomics

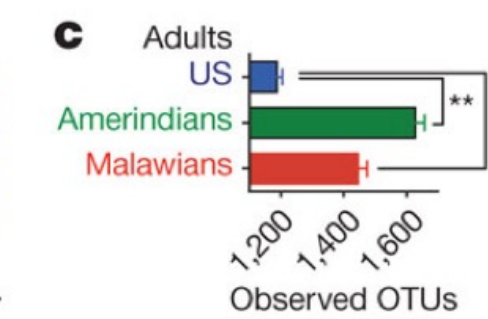
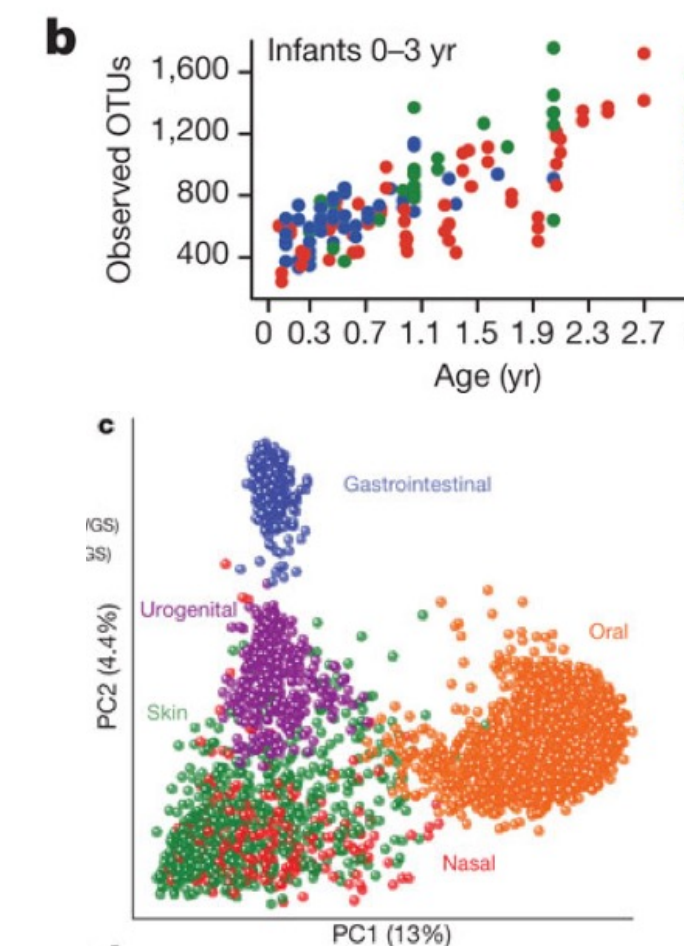
Shinichi Sunagawa (ETHZ)

Review: 16S rRNA-based Operational Taxonomic Units (OTUs)

- 16S rRNA
 - present in all prokaryotes
 - conserved function as integral part of the protein synthesis machinery
 - similar mutation rate: → molecular clock
- Proxy for phylogenetic relatedness of organisms
- Owing to lack of prokaryotic species definition, 97% sequence similarity is often used to define ‘species’-like:
 - “Operational Taxonomic Units” (OTUs)**



Applied examples I

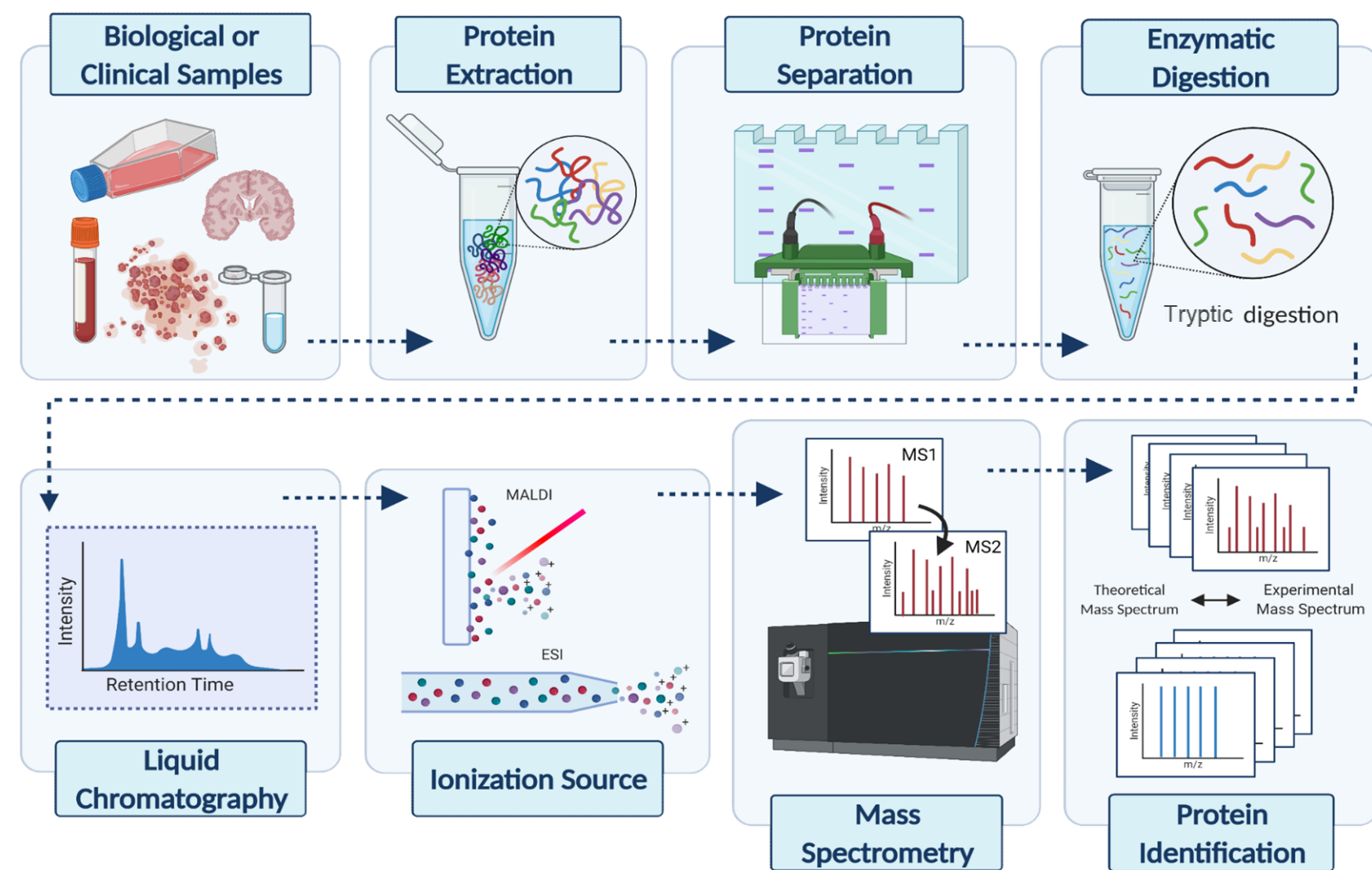


- Microbial diversity in human gut increases with age
- US citizens harbor less diverse gut microbiota relative to other populations
- Microbial communities cluster by human body site rather than by individual

Proteomics

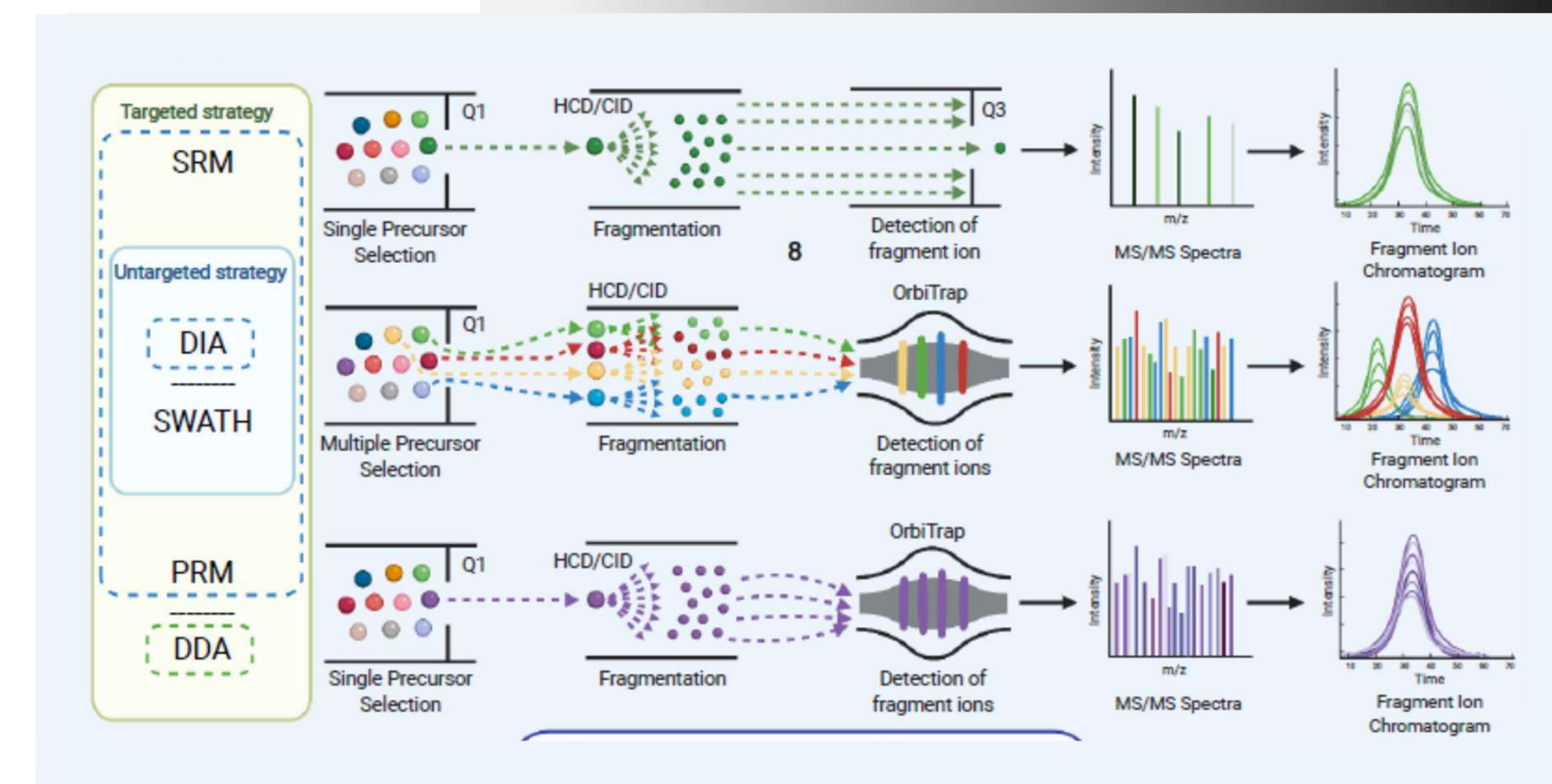
Katja Bärenfaller (SIAF)

Generic mass spectrometry-based proteomics experiment



Mass spectrometry

Hypothesis-driven, targeted bottom-up proteomics approaches



Radzikowska et al., *EAACI Position Paper*, in revision

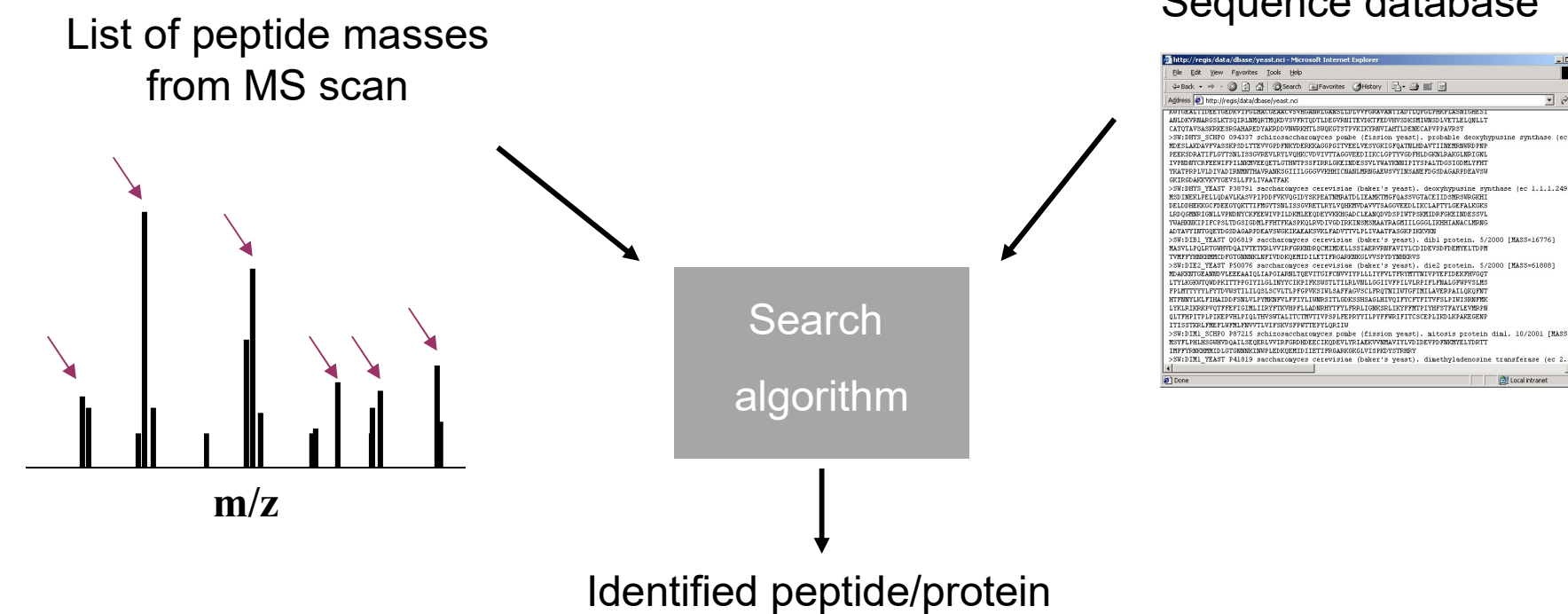
S/MRM: Selected/Multiple Reaction Monitoring; the proteins are pre-selected and provide information on the characteristic peptide precursor and fragment ion signals (transitions)

DIA/SWATH: Data Independent Acquisition/Sequential Windowed Acquisition of All Theoretical Mass Spectra

Peptide Mass Fingerprint

fragment ion signals from a precursor ion

Identifying peptides using an MS spectrum:



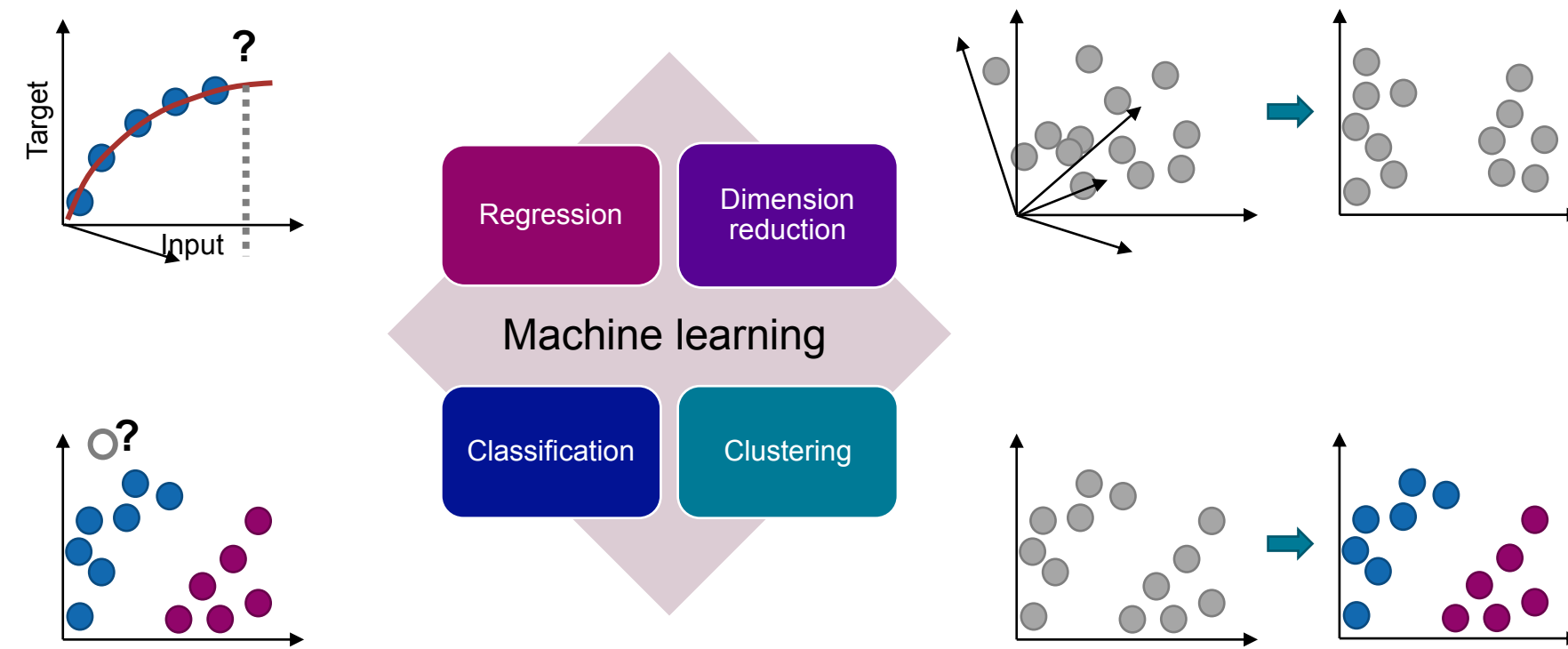
- Peptide spectrum assignment with Peptide Mass Fingerprinting is only advisable with samples of low complexity and small sequence databases, as the number of all possible peptides with a given mass over charge is huge in large sequence databases.

Mass spectrometry

Machine Learning for Biological Use Cases

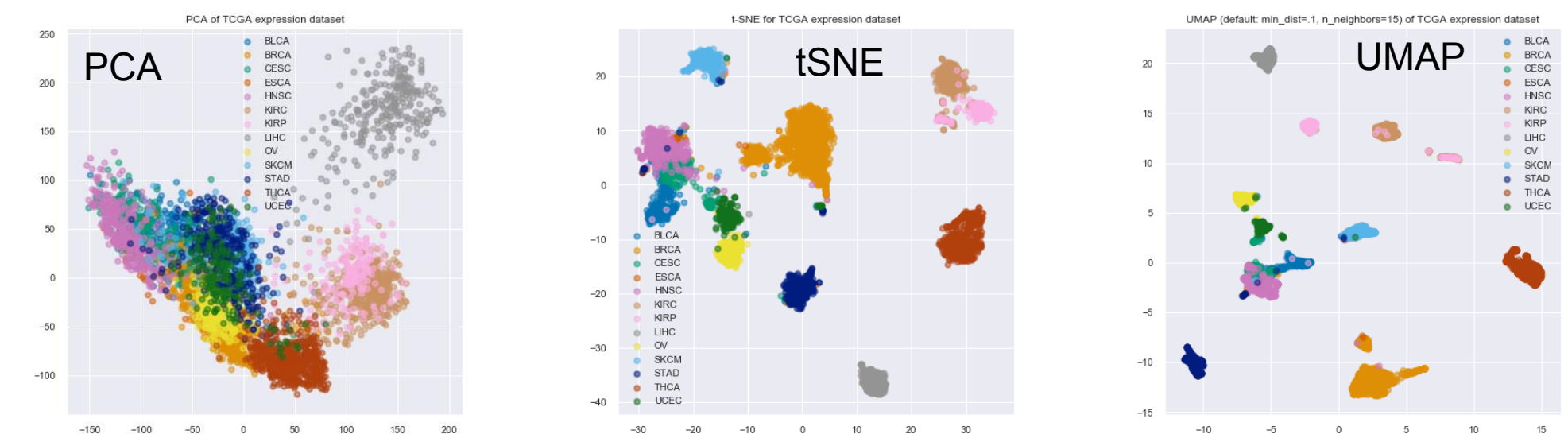
Valentina Boeva (ETHZ)

Map of classical machine learning methods



Uniform Manifold Approximation and Projection (UMAP)

- UMAP: nonlinear dimensionality reduction technique. Idea is similar to tSNE, but
 - Much faster
 - Not limited to the first 2-3 dimensions
 - Uses binary cross-entropy as a cost function instead of the KL-divergence
 - Preserves global structure
 - Uses the number of nearest neighbors instead of perplexity



First introduced by [McInnes, L., Healy, J, ArXiv e-prints 1802.03426, 2018](https://arxiv.org/abs/1802.03426)

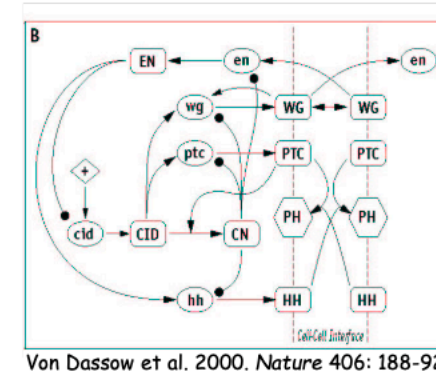
Biological Networks

Pouria Dasmeh / Andreas Wagner

Cell-biological networks

1. Small networks dedicated to a specific task (up to dozens of gene products)

Chemotaxis
Cell-cycle regulation
Fruit fly segmentation
Flower development
...

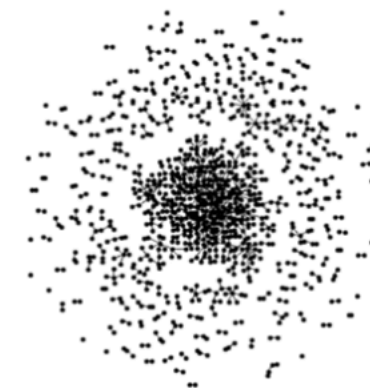


Mathematical characterization based on detailed, quantitative biochemical information

Cell-biological networks

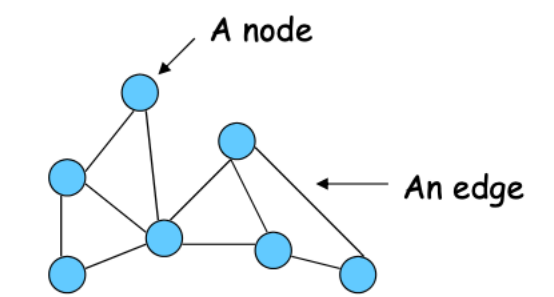
2. Genome-scale networks (hundreds to thousands of gene products)

Protein interaction networks
Metabolic networks
Transcriptional regulation networks
Genetic interaction networks
...



Mathematical characterization based on qualitative understanding of network topology

Graphs

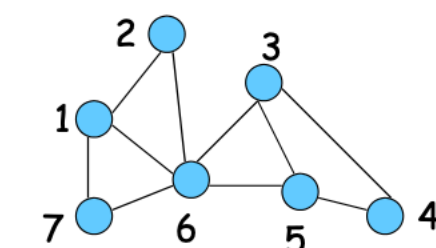


A graph $G=(V,E)$ comprises
a set V of nodes (vertices)
a set E of edges

$$V = \{V_1, \dots, V_n\}$$

$$E = \{(V_i, V_j), \dots, (V_k, V_l)\}$$

Protein interaction networks are undirected graphs
(Individual node pairs in E are unordered.)



The degree (connectivity) k_i of a node V_i is the number of edges incident with the node (e.g., $k_1=3$, $k_6=5$).

$$k_i = \sum_j a_{ij}$$

Graphs can be characterized according to their degree distribution $P(k)$, the fraction of nodes having degree k .

Text Mining

Patrick Ruch (HES-SO Genève)

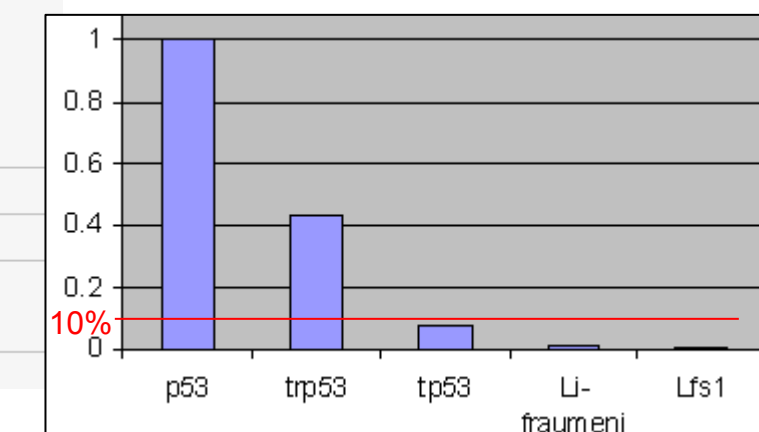
Features

- Words
- Subwords (character N-grams)
- Stems
- Word N-grams
- Syntactic entities (noun phrases, verb phrases, ...),
- Semantic entities (gene names, chem. compounds, diseases, ...)

Term normalization: database & ontology vs. reality !

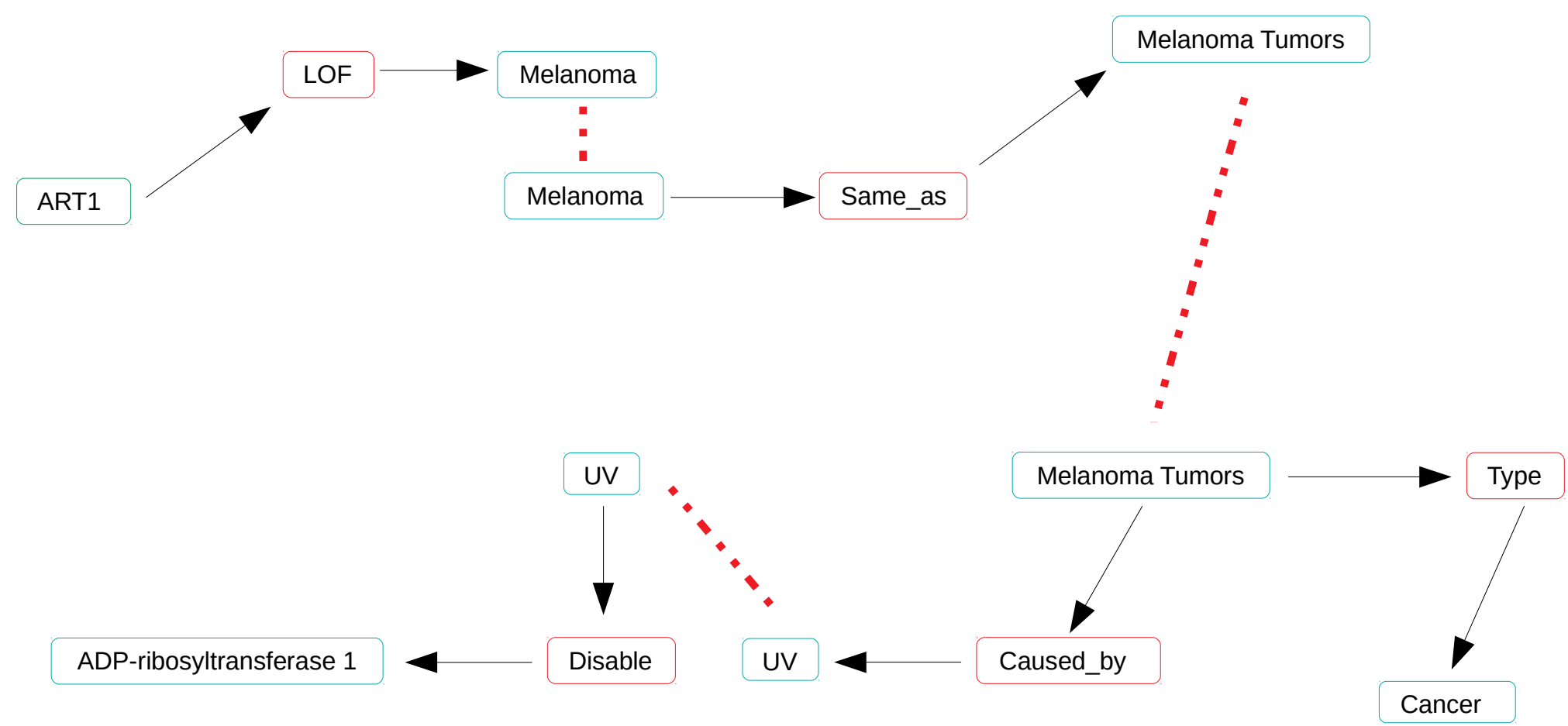
<input type="checkbox"/> Antigen NY-CO-13	Protein	SwissProt:P04637
<input type="checkbox"/> Cellular tumor antigen p53	Protein [preferred]	SwissProt:P04637
<input type="checkbox"/> FLJ92943	Gene	EntrezGene:7157
<input type="checkbox"/> LFS1	Gene	EntrezGene:7157 HGNC:11998
<input type="checkbox"/> Li-Fraumeni syndrome	Gene	HGNC:11998
<input type="checkbox"/> p53	Gene	EntrezGene:7157 HGNC:11998
<input type="checkbox"/> P53	Gene	OMIM:191170 SwissProt:P04637
<input type="checkbox"/> p53 antigen	Gene	EntrezGene:7157
<input type="checkbox"/> p53 transformation suppressor	Gene	EntrezGene:7157
<input type="checkbox"/> p53 tumor suppressor	Gene	EntrezGene:7157
<input type="checkbox"/> phosphoprotein p53	Gene	EntrezGene:7157
<input type="checkbox"/> Phosphoprotein p53	Protein	SwissProt:P04637
<input type="checkbox"/> TP53	Gene [preferred]	HGNC:11998 SwissProt:P04637 EntrezGene:7157 OMIM:191170
<input type="checkbox"/> transformation-related protein 53	Gene	EntrezGene:7157
<input type="checkbox"/> TRANSFORMATION-RELATED PROTEIN 53	Gene	OMIM:191170
<input type="checkbox"/> TRP53	Gene	EntrezGene:7157 OMIM:191170
<input type="checkbox"/> tumor protein p53	Gene [preferred]	HGNC:11998

Synonyms	#
p53	53362
trp53	23364
tp53	4156
li-fraumeni	775
lfs1	431



Semantic Web

Ahmad Aghaebrahimian (ZHAW)



Semantic Web Standards

RDF:

RDF is a **graph-based data model** and the set of **syntax** that allows us to write **description** about the resources on the web and to exchange them. It presents data in the **triple format** and gives it structures and unique identifiers so that data can be easily linked.

Principles:

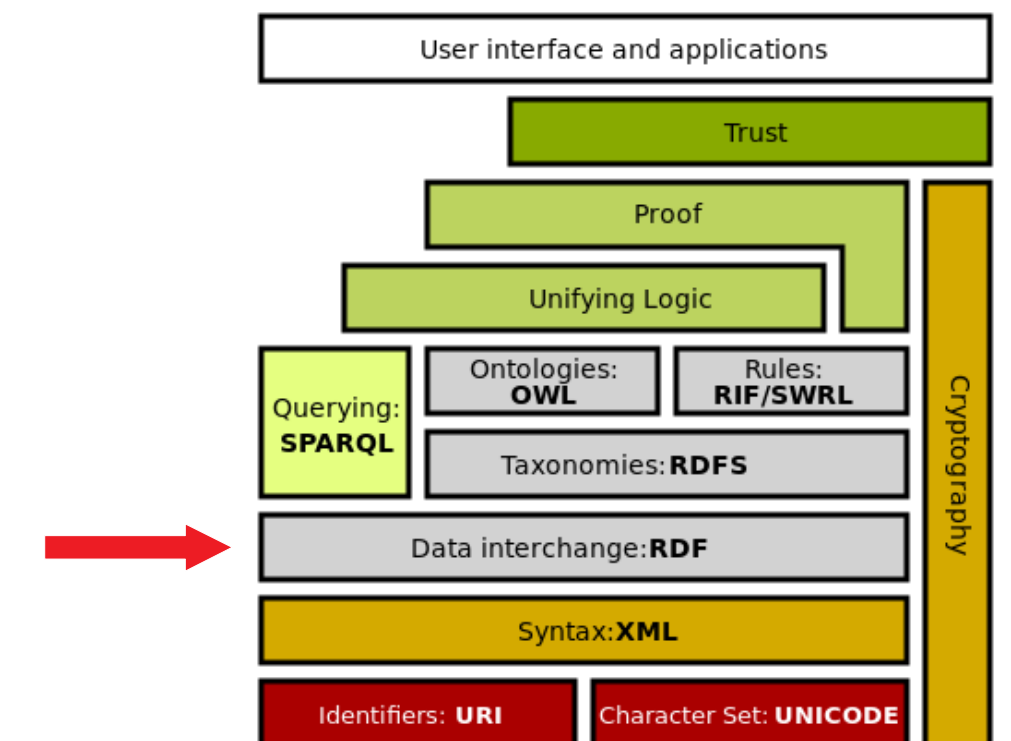
Triple structure: (subject, predicate, object)

- subject → a URI resource
- predicate → binary type URI
- object → a URI resource or literal

Predicates are labeled
 Predicates are directed
 RDF is a graph model

RDF serialization:

XML, N-triple, Turtle, TriG, JSON-LD



Building Genomics Resources

Qingyao Huang



Let's build a database!

... using archaic tools



Progenetix in 2021 Cancer Genomics Reference Resource

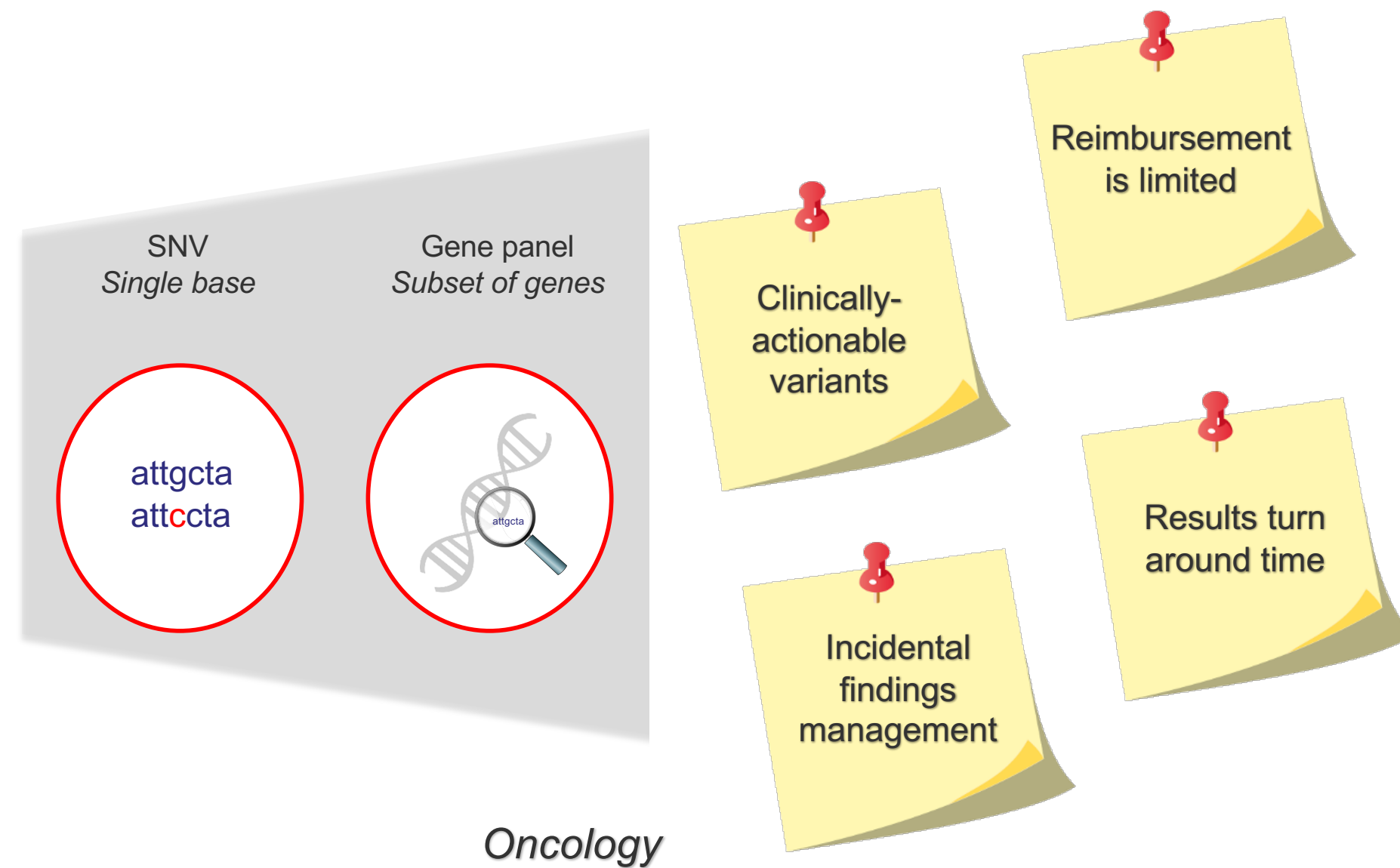
- largest open resource for curated cancer genome profiling data, with focus on copy number variations (CNV)
- >116'000 cancer CNV profiles, mapped to >800 NCIt codes
- majority of data from genomic arrays with ~50% overall from SNP platforms with original data re-processing
- structured diagnostic encodings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



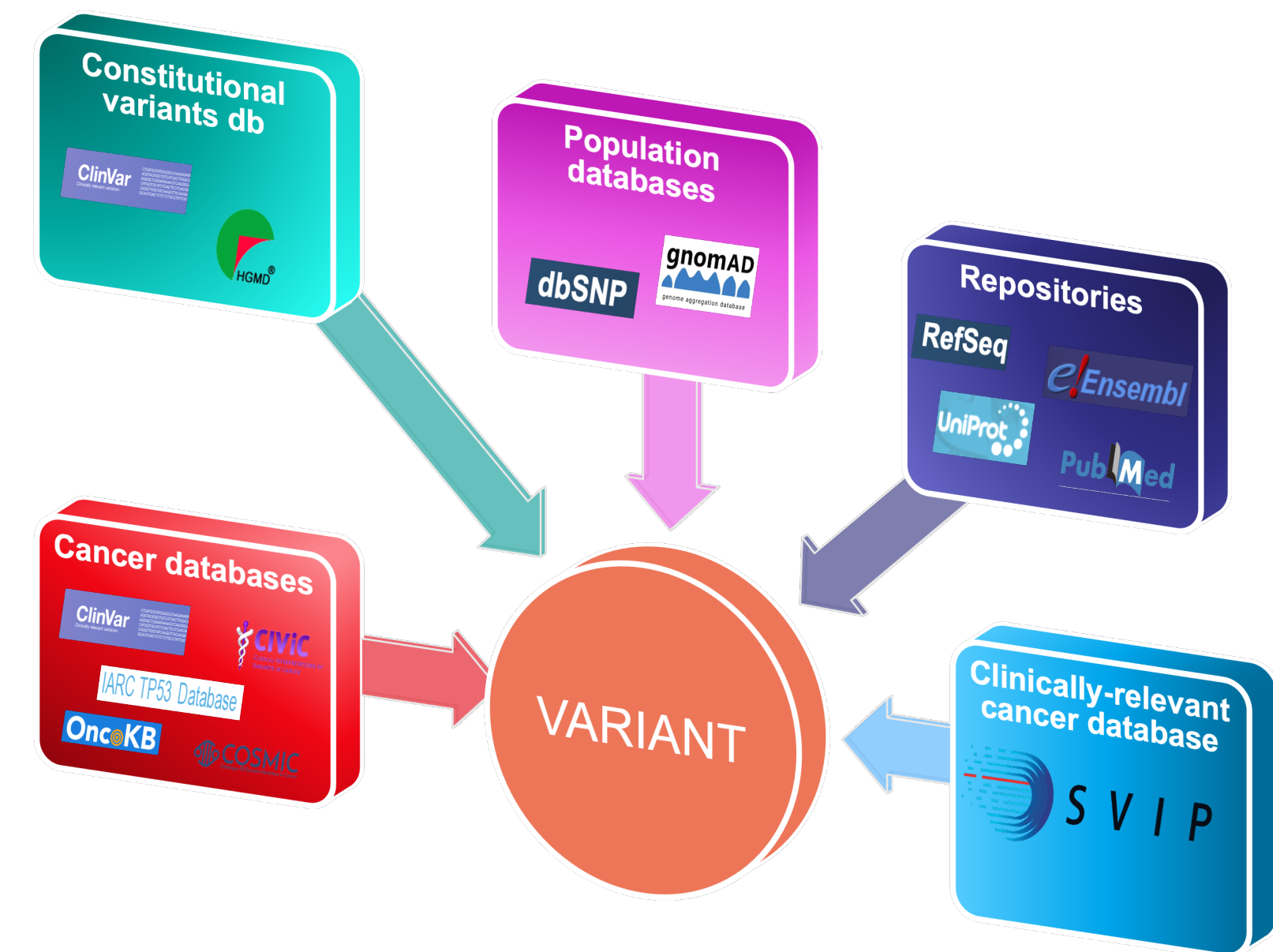
Clinical Bioinformatics

Valérie Barbié (Director SIB Clinical Bioinformatics)

Scale matters



Knowledge bases



Non exhaustive

Genomic Data & Privacy: Risks & Opportunities

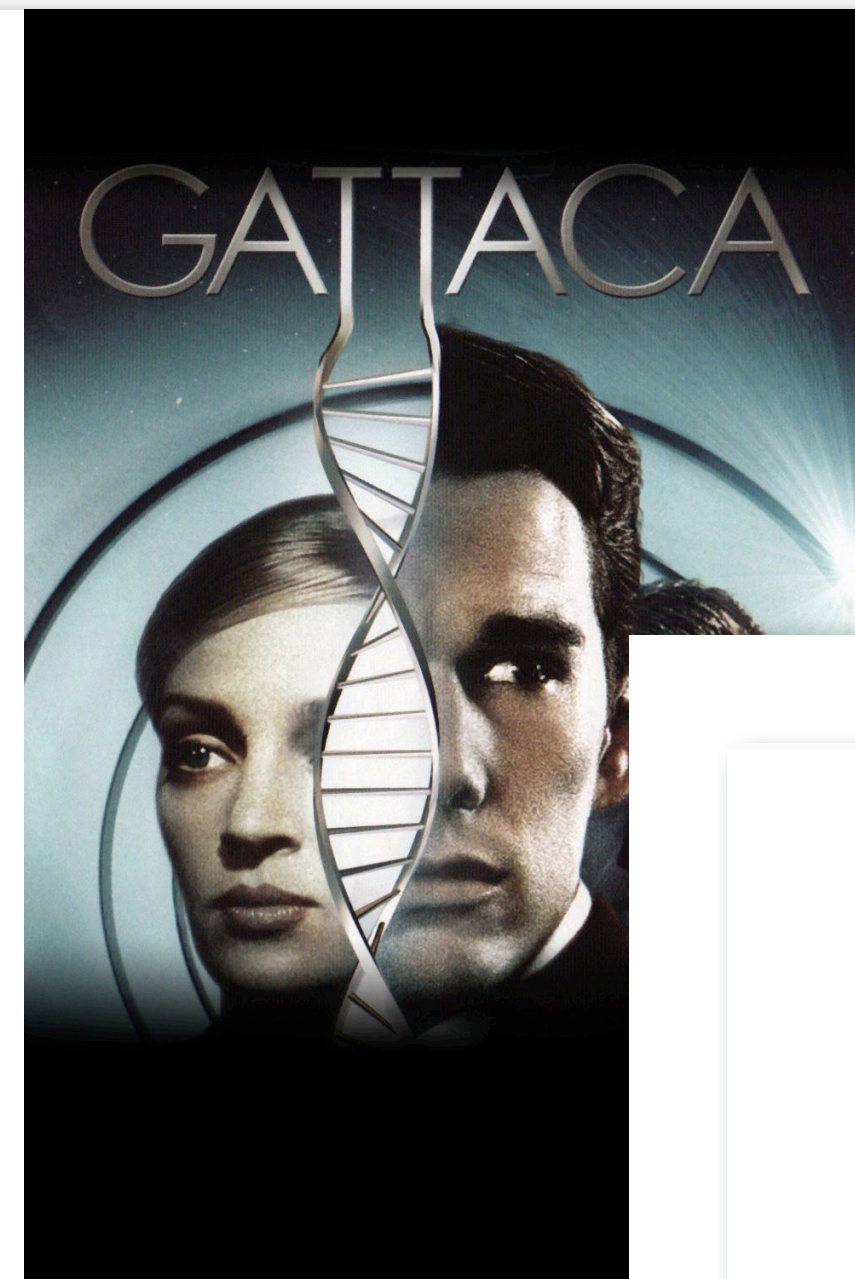
Michael Baudis

Gattaca (1997)

A genetically inferior man assumes the identity of a superior one in order to pursue his lifelong dream of space travel.

- genetic determinism
 - main character has been determined to be unsuitable for complex jobs based on genetic analysis
- genetic identification
 - the use of genetic sampling for personal identification is daily routine

With information from <https://www.imdb.com/title/tt0119177/>



Genome Beacons Compromise Security?

Querying for thousands of specific SNV occurrences in a genomic data pool can identify individuals in an anonymized genomic data collection

Stanford researchers identify potential security hole in genomic data-sharing network

Hackers with access to a person's genome might find out if that genome is in an international network of disease databases.

OCT 29 2015

Sharing genomic information among researchers is critical to the advance of biomedical research. Yet genomic data contains identifiable information and, in the wrong hands, poses a risk to individual privacy. If someone had access to your genome sequence — either directly from your saliva or other tissues, or from a popular genomic information service — they could check to see if you appear in a database of people with certain medical conditions, such as heart disease, lung cancer or autism.

Work by a pair of researchers at the Stanford University School of Medicine makes that genomic data more secure. [Suyash Shringarpure](#), PhD, a postdoctoral scholar in genetics, and [Carlos Bustamante](#), PhD, a professor of genetics, have



Stanford researchers are working with the Global Alliance for Genomics and Health to make genomic information in the Beacon Project more secure. *Science photo/Shutterstock*

genomic databases and how to prevent it. e for Genomics and Health on implementing

an Genetics, also bears importantly on the h as those from different people at a crime

Rapid re-identification of human samples

...

We developed a rapid, inexpensive, and portable strategy to re-identify human DNA using the MinION. Our strategy requires only ~60 min preparation and 5-30 minutes of MinION sequencing, works with low input DNA, and enables familial searches using Direct-to-Consumer genomic reference datasets. This method can be implemented in a variety of fields:



Forensics

Identification of abandoned material using DNA fingerprinting is a common practice. The main challenge currently being: time. Our method allows rapid sample preparation at the crime scene (see movie). We envision that the method can be adopted in the field for rapid checks, after a mass disaster, and can be adopted in border control to fight human trafficking.



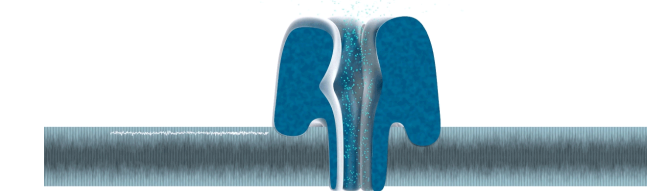
Clinic

Clinics process many samples, either for analysis or, for example, organ donations. These samples are DNA fingerprinted to prevent sample mix-up mistakes. Our method can be implemented in the clinic for rapid sanity-check of all incoming samples.



Cell line identification

Cross contamination of cell lines in science is a major problem. It results in unreproducible data, and clinical trials based on inaccurate findings. This problem costs billions of dollars per year. We envision labs can adopt our identification method to ensure the purity of the cell line, and detect contamination.



The MinION (Oxford Nanopore)
Source: Sophie Zaaijer
<https://medium.com/neodotlife/nanopore-6443c81d76d3>



University of
Zurich^{UZH}



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

progenetix.org
info.baudisgroup.org
sib.swiss/baudis-michael
imls.uzh.ch/en/research/baudis
beacon-project.io
schemablocks.org



Global Alliance
for Genomics & Health





University of Zurich ^{UZH}



Delayed responses since on research semester (ツ)

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

progenetix.org
info.baudisgroup.org
sib.swiss/audis-michael
imls.uzh.ch/en/research/audis
beacon-project.io
schemablocks.org



Global Alliance
for Genomics & Health

