

BIO390: Introduction to Bioinformatics

Lecture I: What is Bioinformatics?

Michael Baudis | 2022-09-20

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 20st)
- 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 website & OLAT for
additional course information**

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

 **CompbioZurich**  Search  compbiozurich.github.io ☆1 ⚡8

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

- Autumn semesters
- 1 x 2h / week
- Tue 08:00-09:45
- UZH Irchel campus, **Y03-G-85**
- OLAT [lecture recordings](#)
- Course language is English

Some [very approximate learning goals](#) may provide you with additional guidance - but please be aware that those may not be particularly adjusted to a given course edition.

Upcoming

 September 20, 2022

What is Bioinformatics? Introduction and Resources

BIO390 UZH HS22 - INTRODUCTION TO BIOINFORMATICS

 **CompbioZurich**  Search  compbiozurich.github.io ☆1 ⚡8

What is Bioinformatics? Introduction and Resources

BIO390 UZH HS22 - Introduction to Bioinformatics

Michael Baudis

The first day of the "Introduction to Bioinformatics" lecture series starts with a general introduction into the field and a description of the lecture topics, timeline and procedures.

Topics covered in the lecture are e.g.:

- a term definition for *bioinformatics*
- the relation of hypothesis driven and data driven science, with respect to bioinformatics
- categories of bioinformatics tools and data
- research areas and topics
- the varying emphasis on "bio" and "informatics"
- databases (primary vs. derived) and data curation
- data collection & curation
- file formats, ontologies & APIs across areas/topics (w/o details)
- "not-bioinformatics"

At the end of the lecture, a - very - brief introduction into the relevance of different aspects of bioinformatics for human genome variation is being given.

Q & A¹

WHAT IS MEANT WITH PROSE IN BIOINFORMATICS?

Prose refers to written text, e.g. the body of scientific publications or descriptions of experimental procedures, which may contain much

1. Thanks to Katja W. in HS21 for starting some questions!
2. One shouldn't write in MS Word, IMO. Use LaTeX, Markdown, OpenOffice ...

Links

- [\[2021 lecture slides\] \(PDF\)](#)

 September 20, 2022

Source: New York Times | SUSAN DYNARSKI NOV. 22, 2017



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)



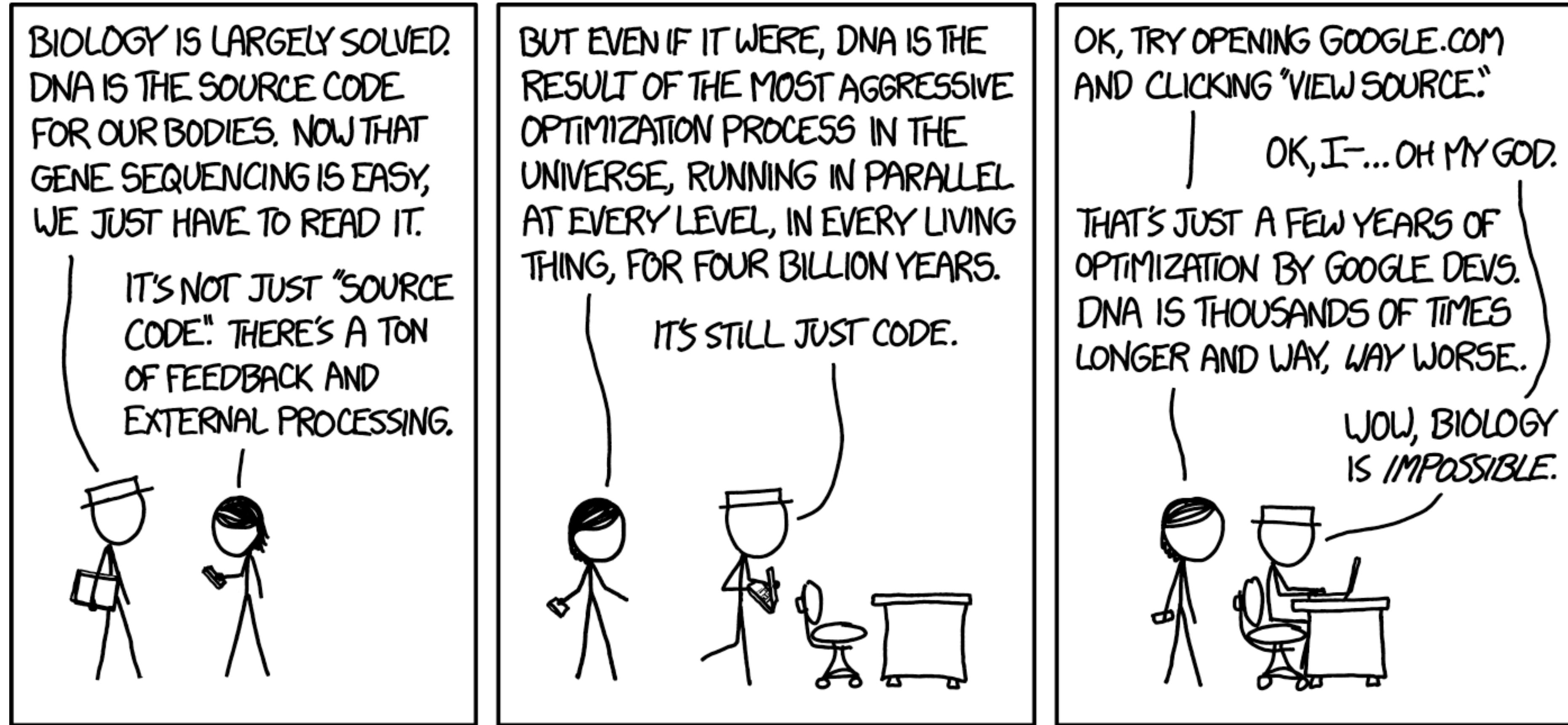
BIO390: Course Schedule

- 2022-09-20: Michael Baudis - What is Bioinformatics? Introduction and Resources
- 2022-09-27: Christian von Mering - Sequence Bioinformatics
- 2022-10-04: Mark Robinson - Statistical Bioinformatics
- 2022-10-11: Valentina Boeva (ETHZ) - Machine Learning for Biological Use Cases
- 2022-10-18: Izaskun Mallona - Regulatory Genomics and Epigenomics
- 2022-10-25: Shinichi Sunagawa (ETHZ) - Metagenomics
- 2022-11-01: Katja Baerenfaller (SIAF) - Proteomics
- 2022-11-08: Andreas Wagner - Biological Networks
- 2022-11-15: Patrick Ruch - Text mining & Search Tools
- 2022-11-22: Ahmad Aghaebrahimian (ZHAW) - Semantic Web
- 2022-11-29: Michael Baudis - Building a Genomics Resource
- 2022-12-06: Valérie Barbie (SIB) - Clinical Bioinformatics
- 2022-12-13: Michael Baudis - Genome Data & Privacy | Feedback
- 2022-12-20: Exam (Multiple Choice)

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

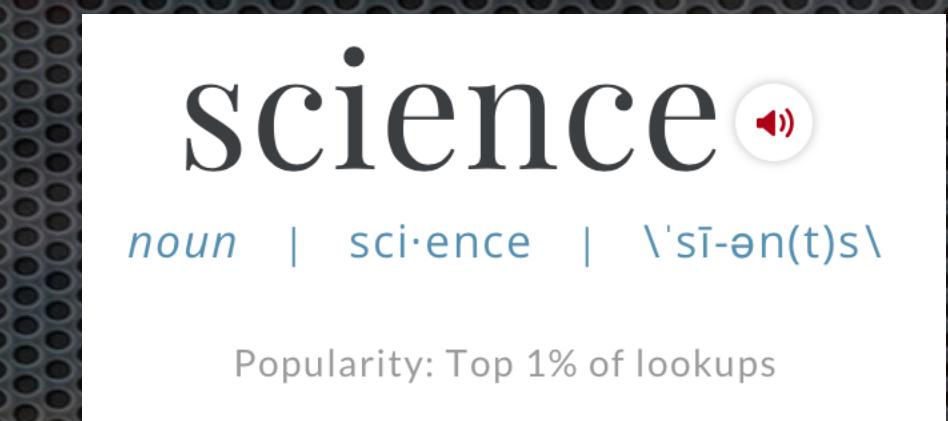
Biology is *impossibly* complex - But bioinformatics might help



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

Merriam-
Webster

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?

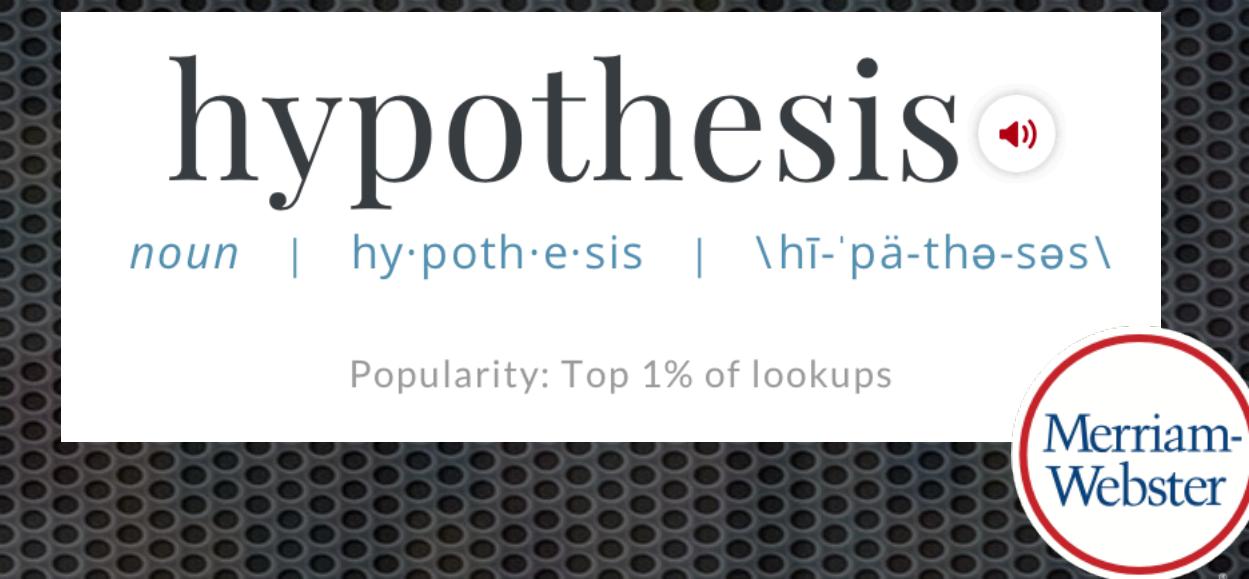


Bioinformatics **analyzes**

- 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 driven science

data driven science

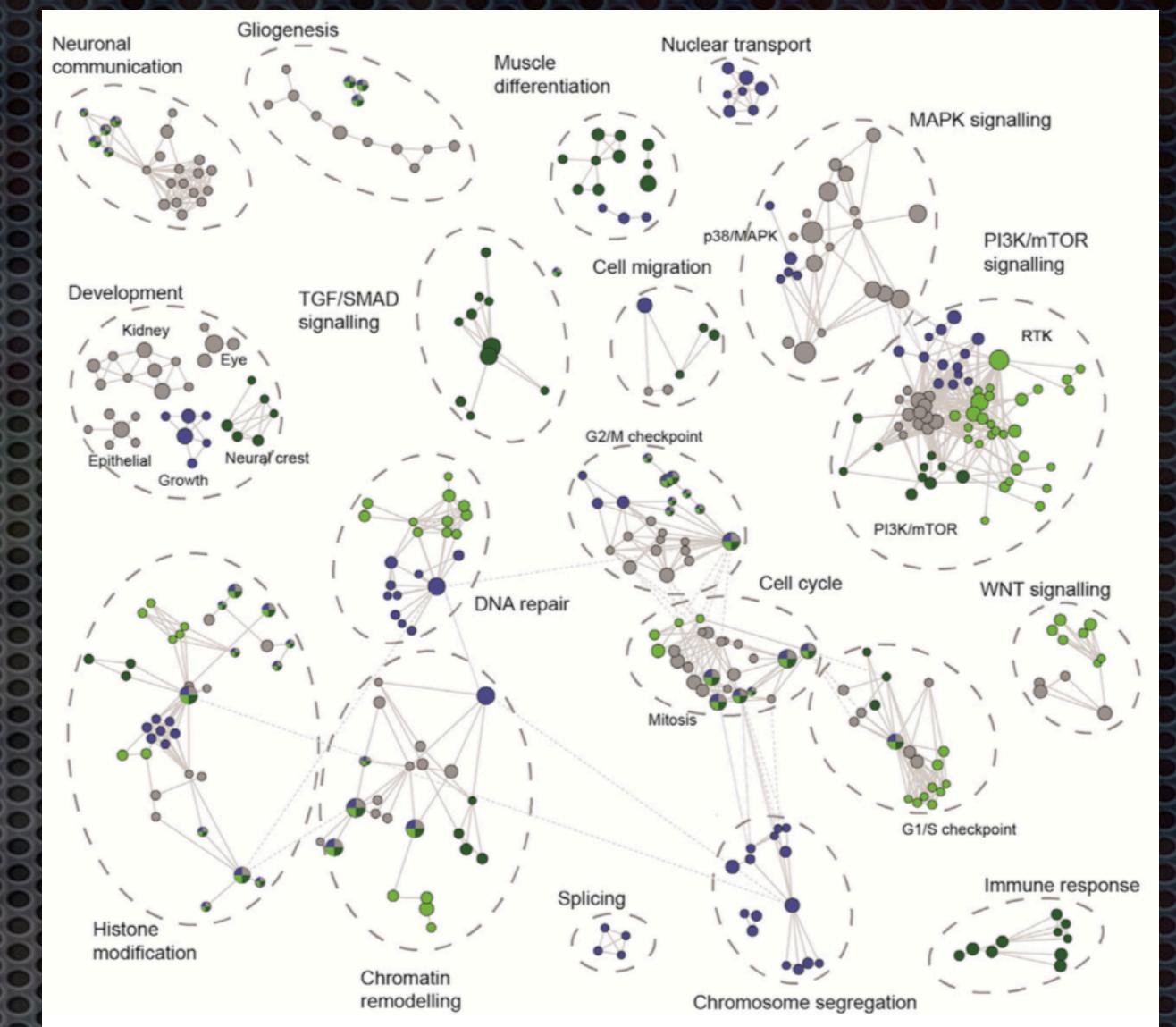
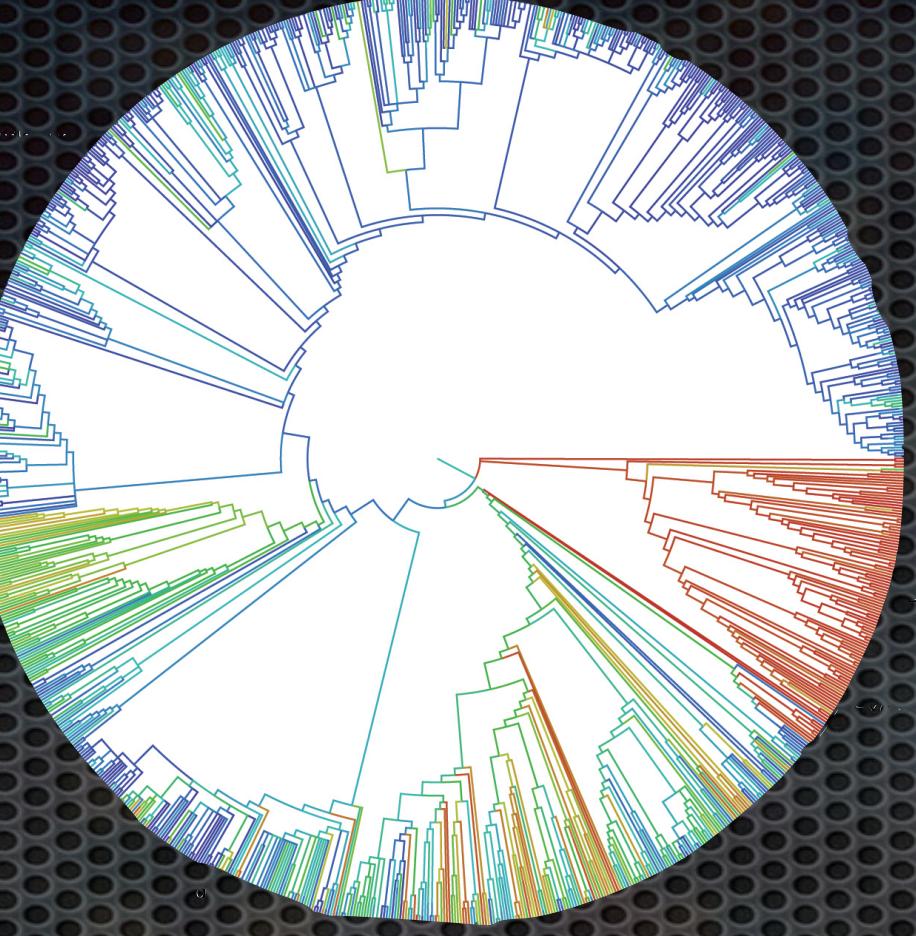


- Bioinformatics is "the science that uses the instruments of informatics to analyze biological data in order to **formulate hypotheses** about life." (Anna Tramontano)

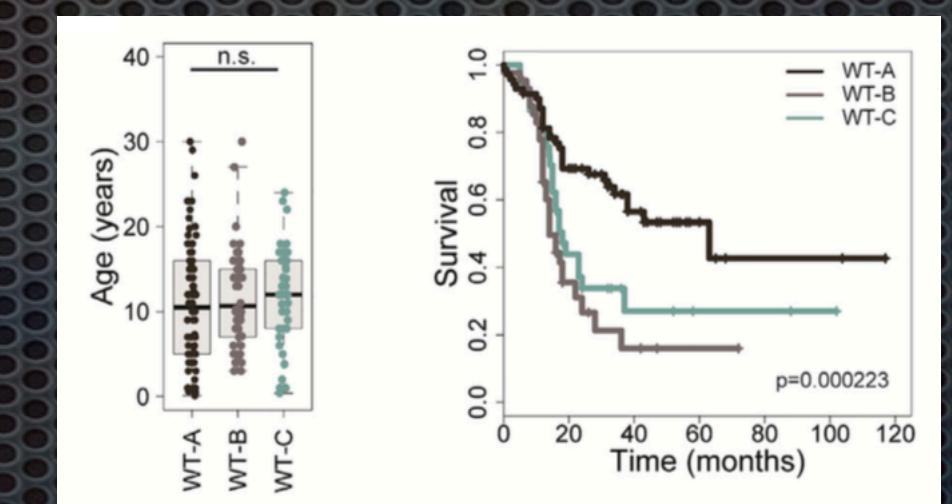
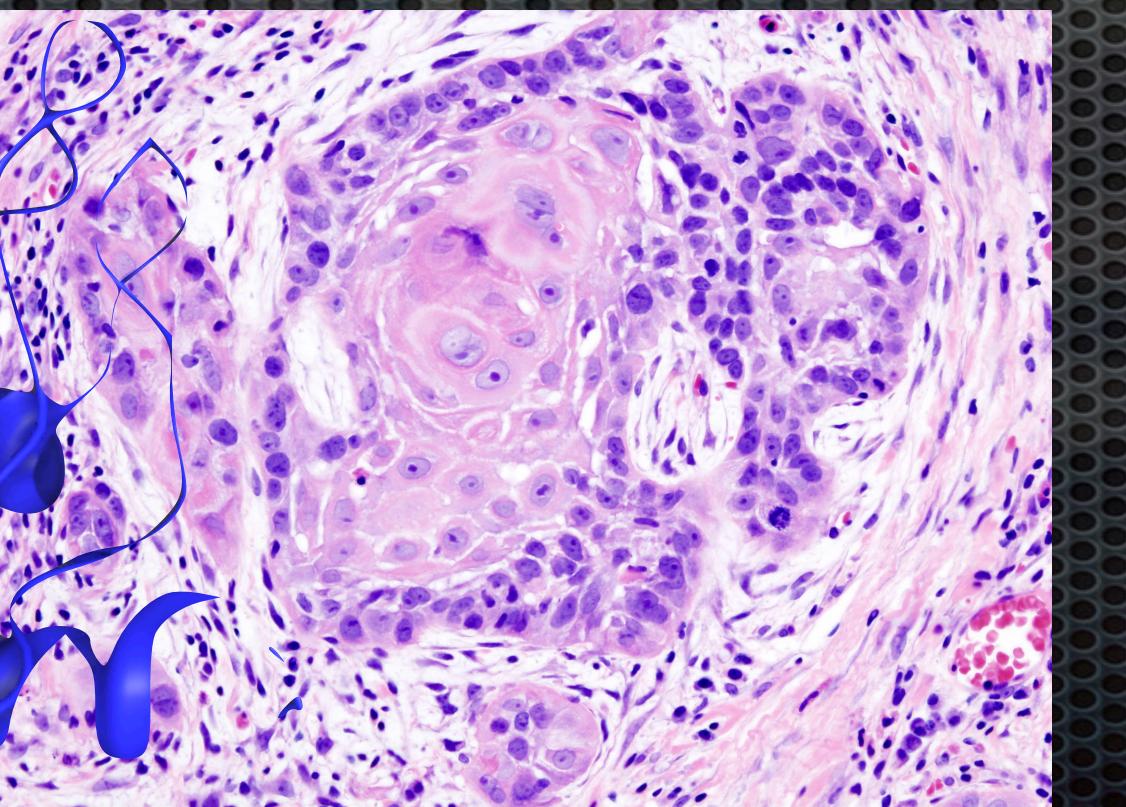
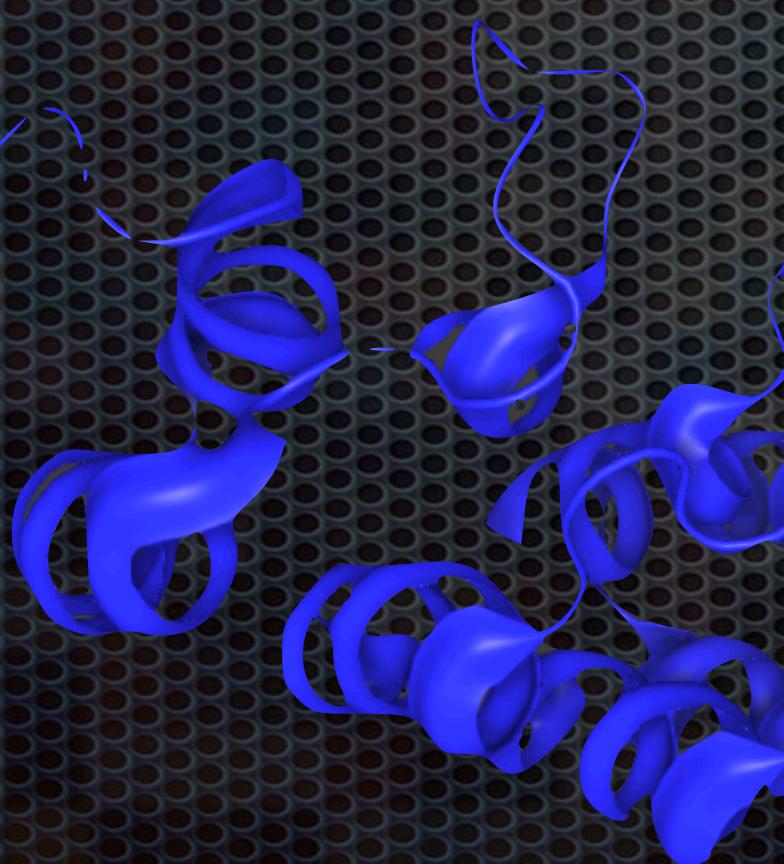
b : an interpretation of a practical situation or condition taken as the ground for action



42



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

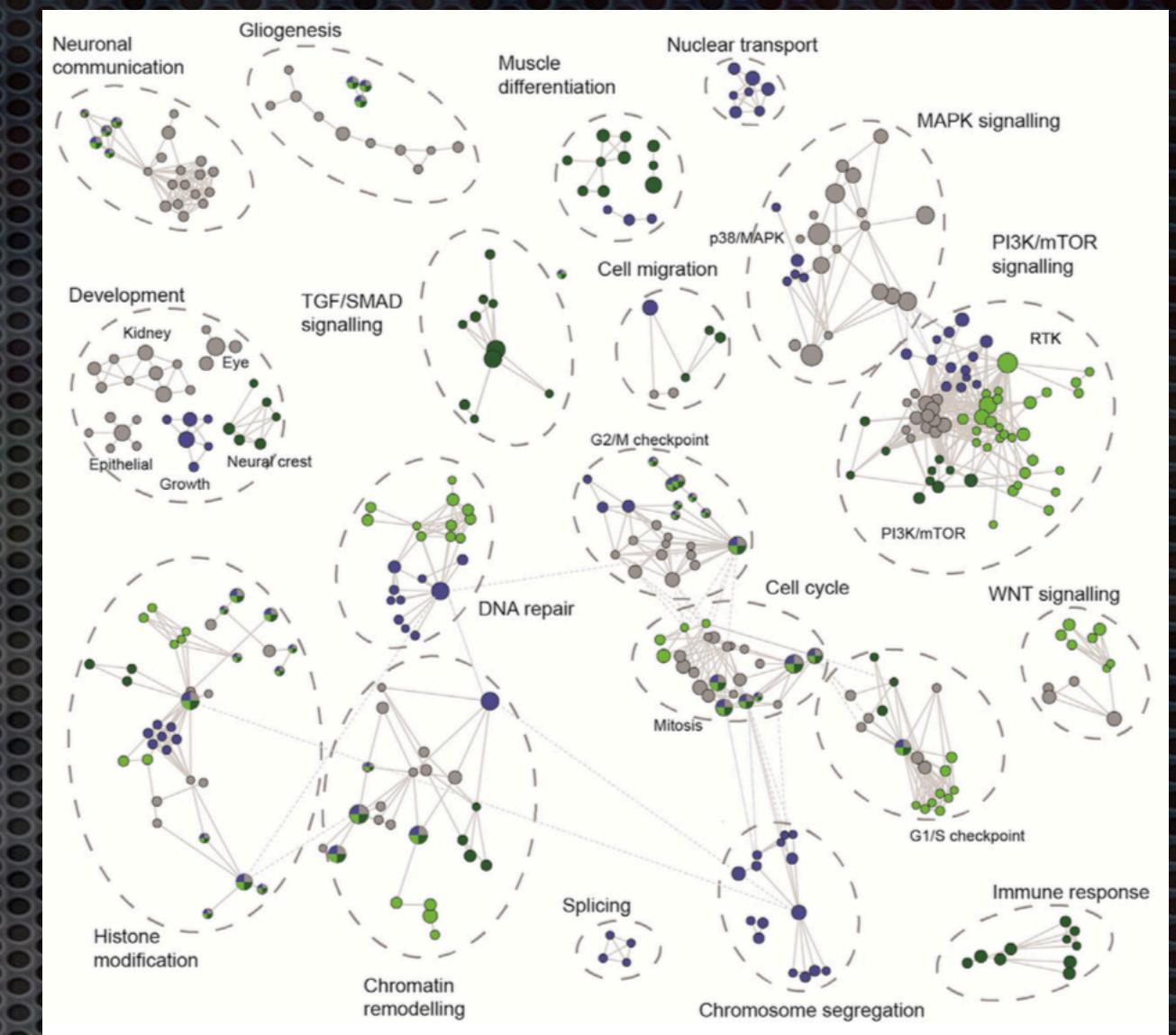
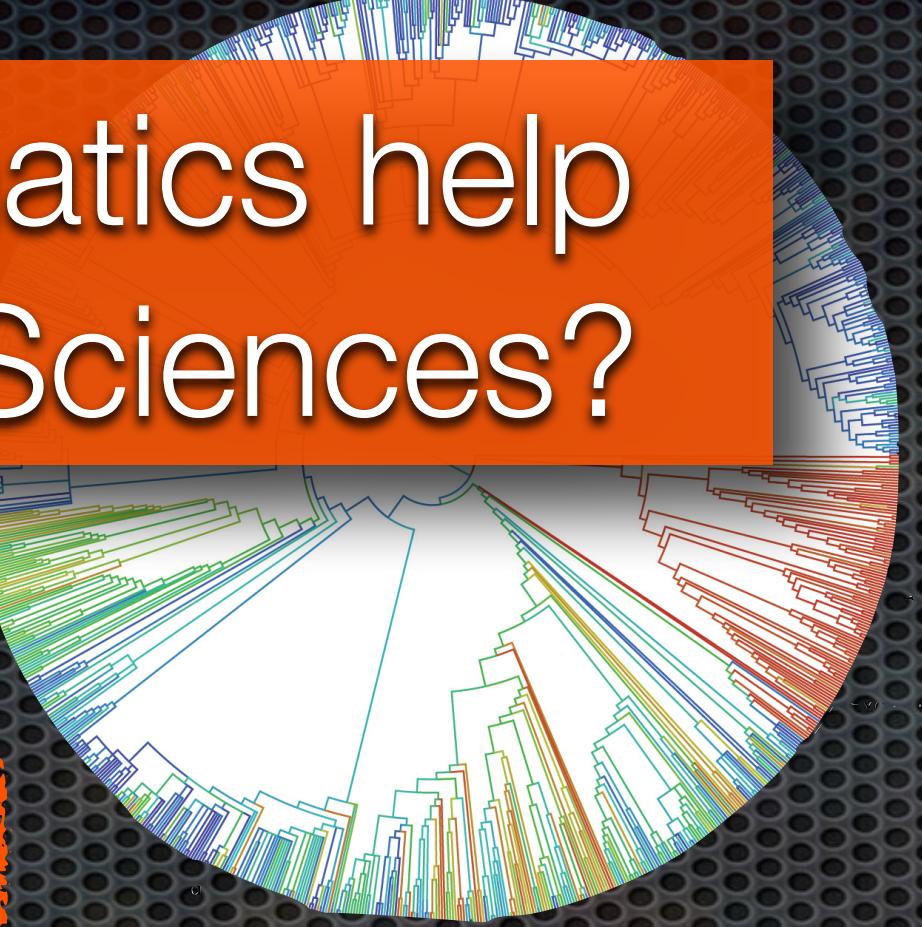


Sources: nextprot | opentreeoflife | wikipedia | MacKay et al., Cancer Cell (2017) | original photos

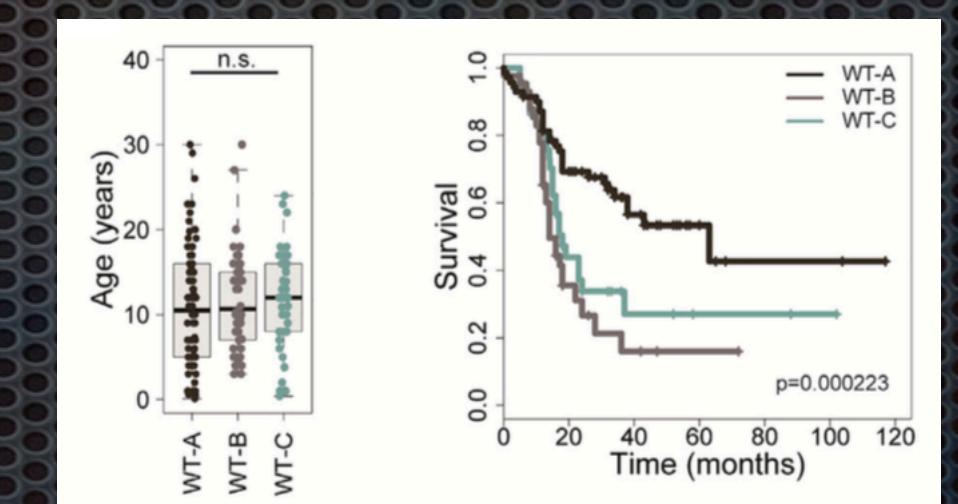
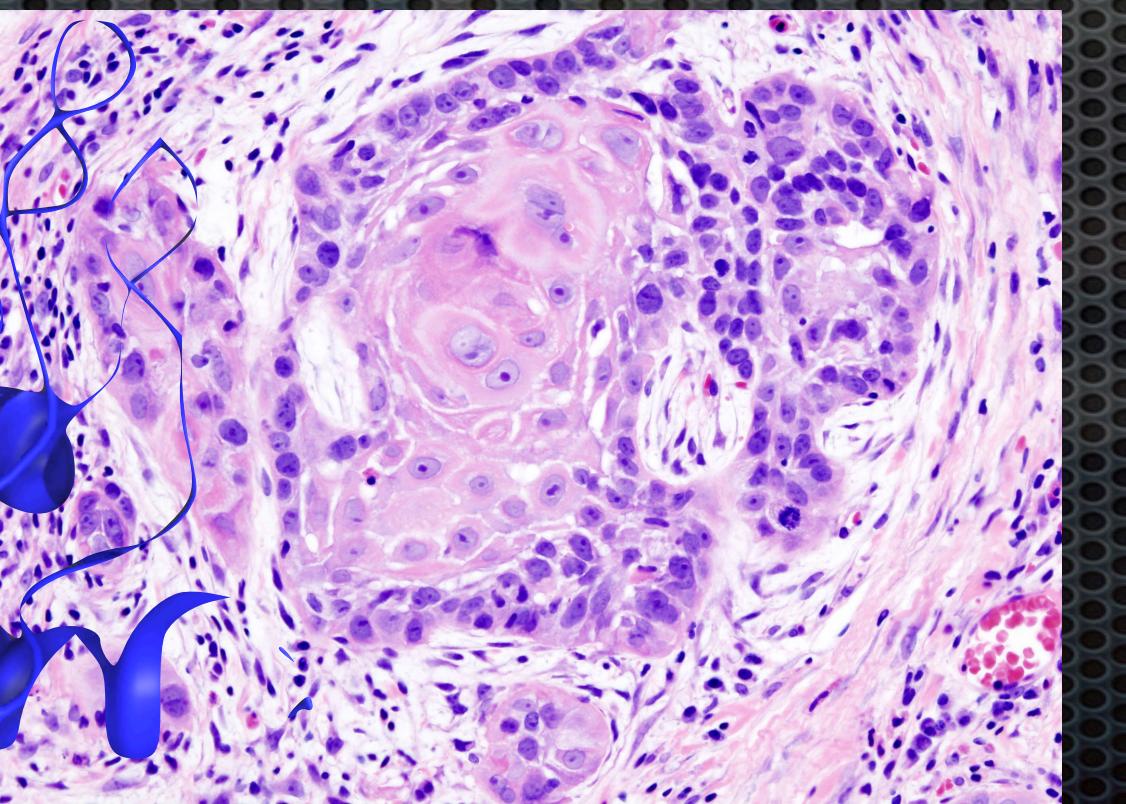
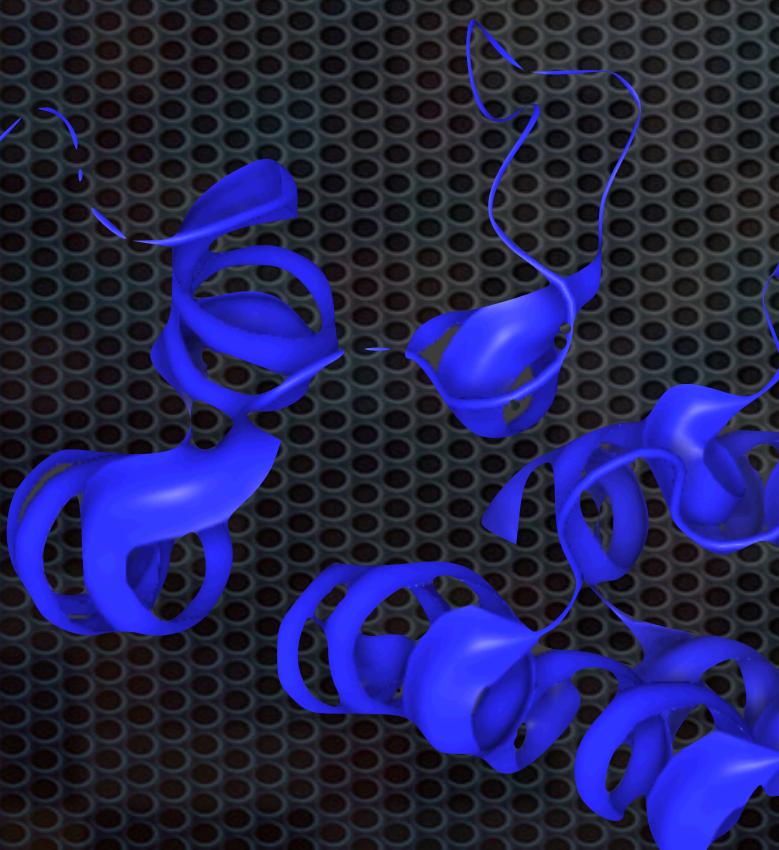


How can Bioinformatics help with the **42** of Life Sciences?

42



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

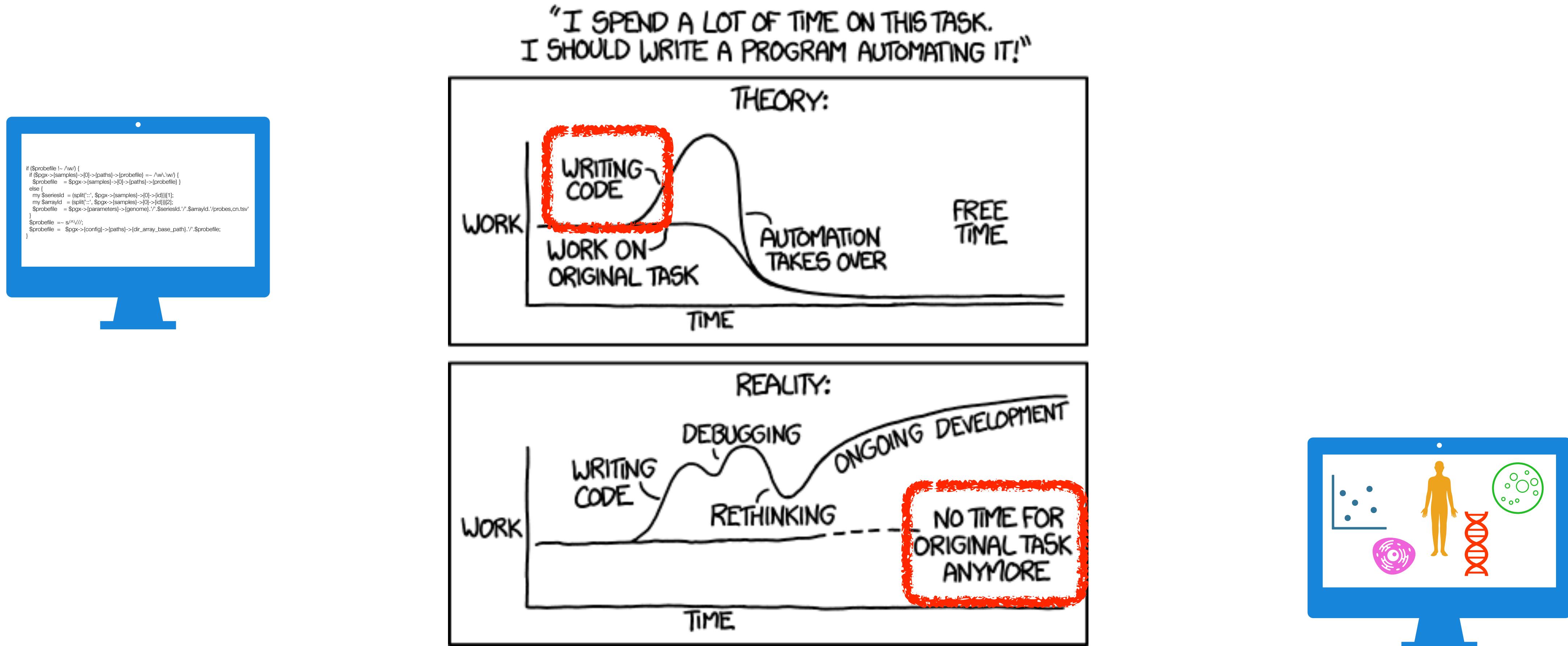


Sources: nextprot | opentreeoflife | wikipedia | MacKay et al., Cancer Cell (2017) | original photos

{bio_informatics_science}



{bio_informatics_science}



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

Bio**informatician**

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

Bio**informatician**

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

What do Bioinformaticians work on?

Hypothesis & Data Driven Approaches to Biological Topics

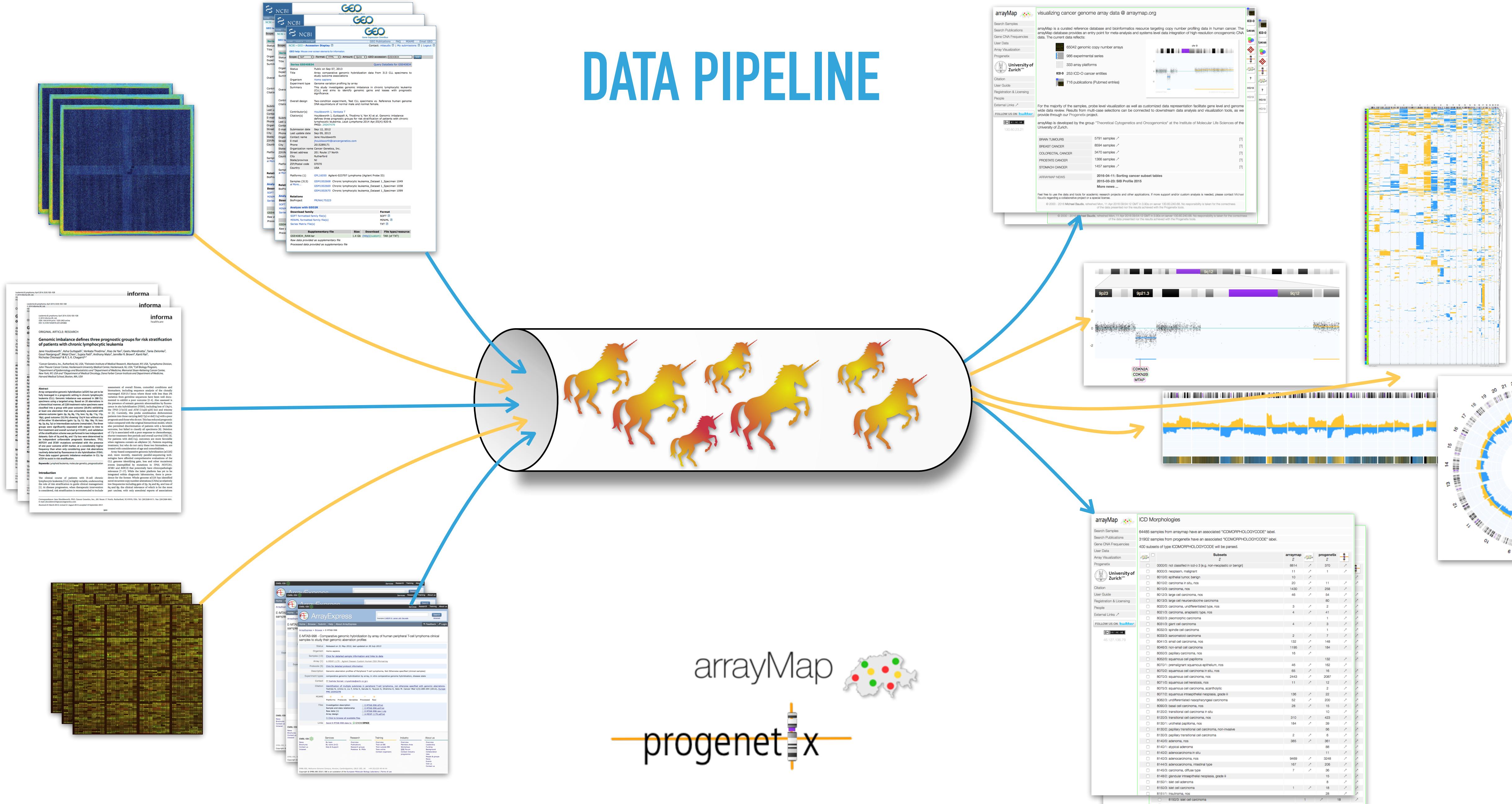
- protein **structure** definition
- DNA/RNA/protein **sequence** analysis
- **quantitative** analysis of "-omics" and cytometry data
- **functional** enrichment of target data (e.g. genes, sequence elements)
- **evolutionary** reconstruction and "tree of life" questions
- **image processing** for feature 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

EVER
CLEANING
THE DATA

DATA PIPELINE

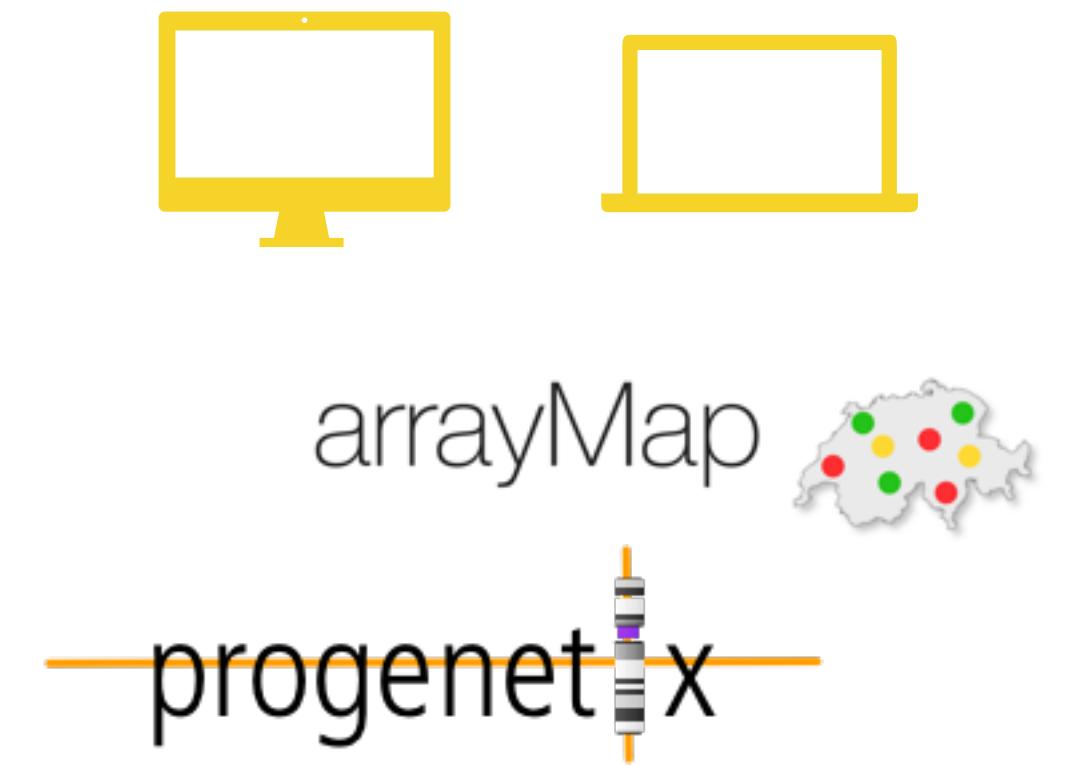


DATA PIPELINE

BIOCURATION BIOINFORMATICS

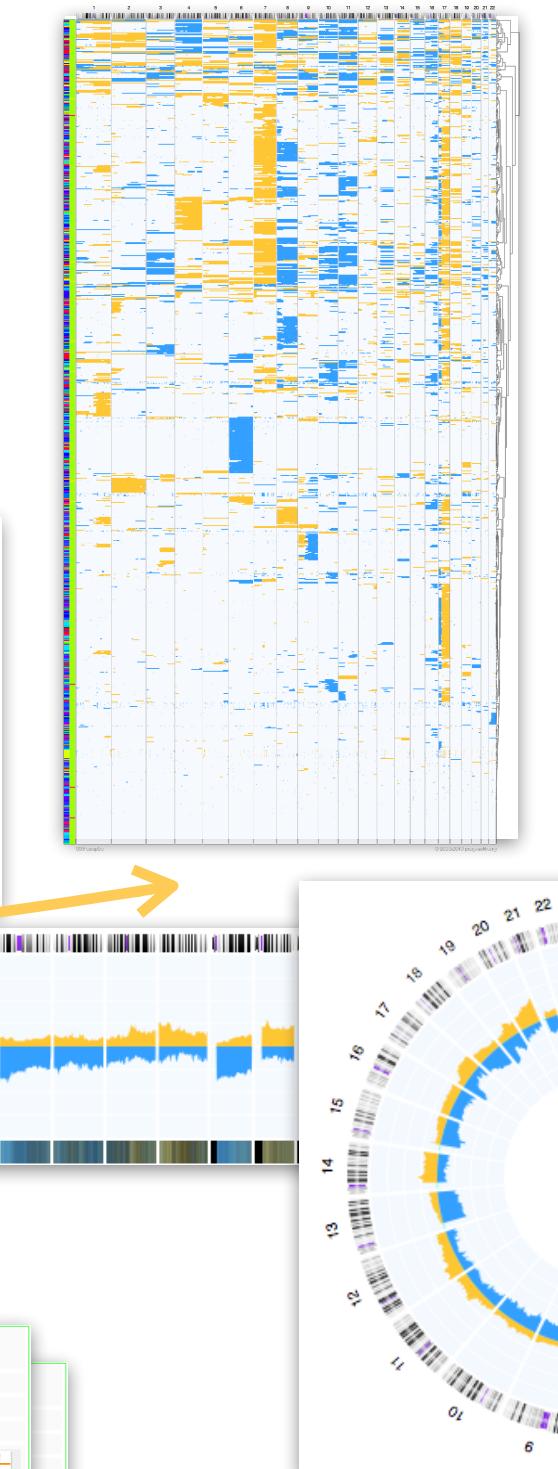
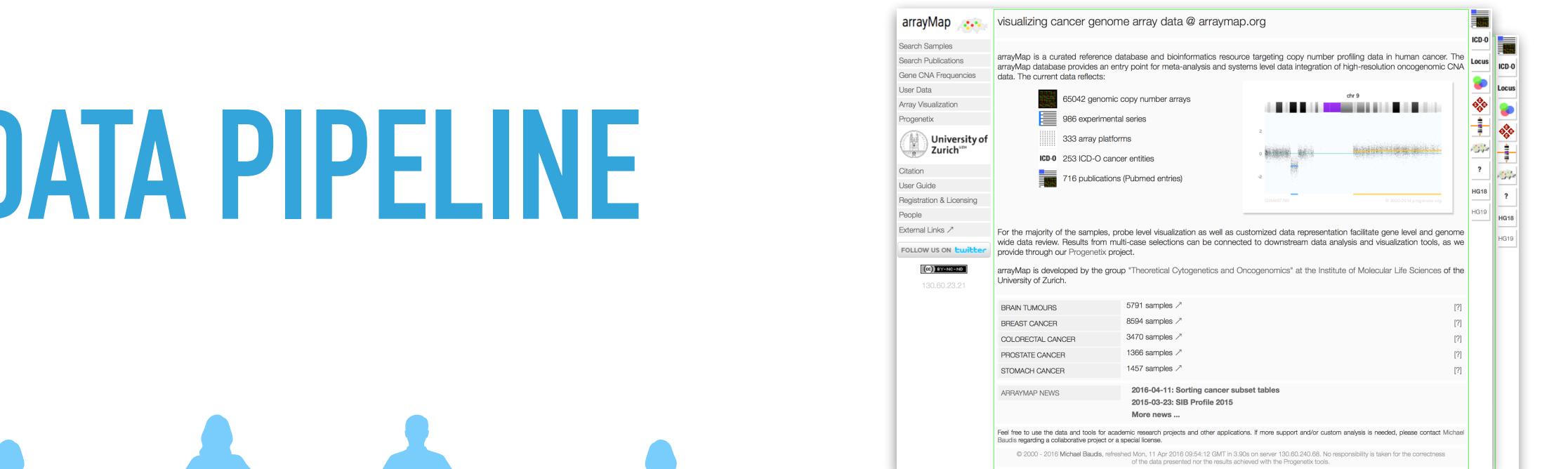


NCBI GEO Accession Display for Series GSE40034. The page shows detailed information about the study, including the title "Genomic imbalance analysis by array CGH in chronic lymphocytic leukemia", the author "Houldsworth J, Guttapalli A, Thodani V, Yan X, et al.", and the date "Public on Sep 07, 2013". It also displays a histogram of sample sizes and a scatter plot of CNA frequencies.



arrayMap

progenetix



BIOCURATION

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)

The image consists of three main parts. At the top right is a file information dialog box for a file named "GSM1904006.CEL" which is 69.1 MB in size and was modified on 3 February 2016 at 17:46. The dialog shows details like kind (FLC animation), size (69'078'052 bytes), and location (arrayRAID → arraymapIn → affyRaw → GSE73822 → GPL6801). It also includes sections for general settings, more info, and opening with applications (QuickTime Player is selected). A red arrow points from the text "not a movie..." to the movie camera icon in the preview section, which is crossed out with a large red X. Below the dialog is a screenshot of a BED file content. The file starts with "browser position chr7:127471196-127495720" and "browser hide all". It then lists genomic tracks for chromosome 7, each with a start and end position, strand (+/-), and itemRGB values. To the right of the file content is a vertical list of file formats, many of which are preceded by a small blue square icon.

File Formats List:

- 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

<http://genome.ucsc.edu/FAQ/FAQformat.html>

BED file example:

```
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 + 127472363 127473530 255,0,0
chr7 127473530 127474697 Pos3 0 + 127473530 127474697 255,0,0
chr7 127474697 127475864 Pos4 0 + 127474697 127475864 255,0,0
chr7 127475864 127477031 Neg1 0 - 127475864 127477031 0,0,255
chr7 127477031 127478198 Neg2 0 - 127477031 127478198 0,0,255
chr7 127478198 127479365 Neg3 0 - 127478198 127479365 0,0,255
chr7 127479365 127480532 Pos5 0 + 127479365 127480532 255,0,0
chr7 127480532 127481699 Neg4 0 - 127480532 127481699 0,0,255
```

not a movie...

File Formats: VCF

Genomic variant storage standard

- 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 Variant Call Format (VCF) Version 4.2 Specification

25 Jun 2020

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

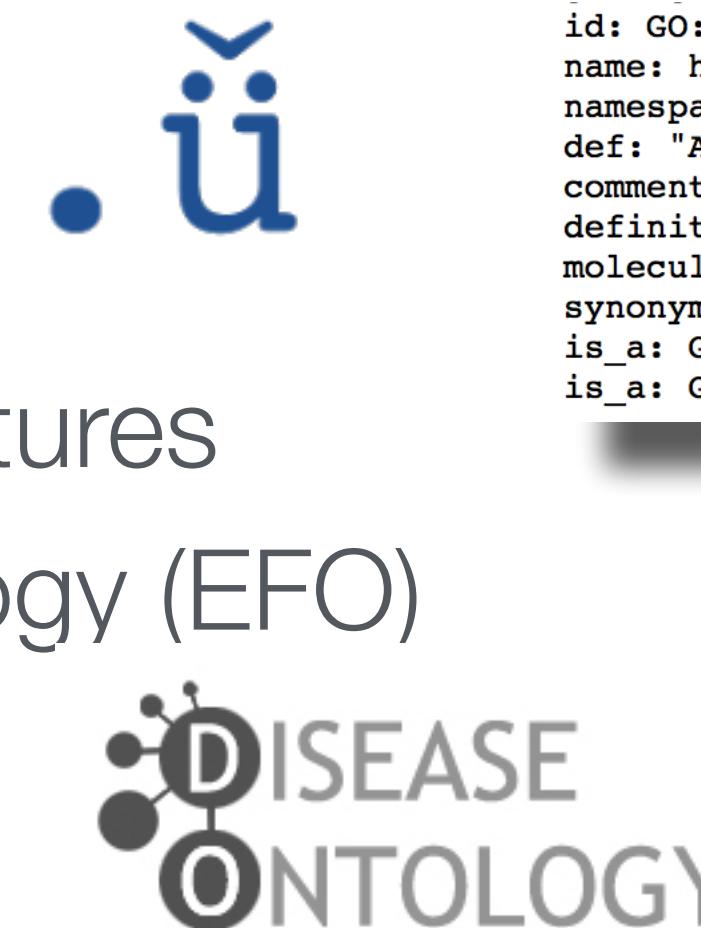
```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number 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)
- NCI 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 complex, not a single protein.
definition for the purpose of this ontology:
molecular function term 'histone deacetylase activity'
synonym: "HDAC complex" EXACT [C3709]
is_a: GO:0044451 ! nucleoplasm
is_a: GO:1902494 ! catalytic complex

complex is mentioned in the
lex is represented by the

- ☐ 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

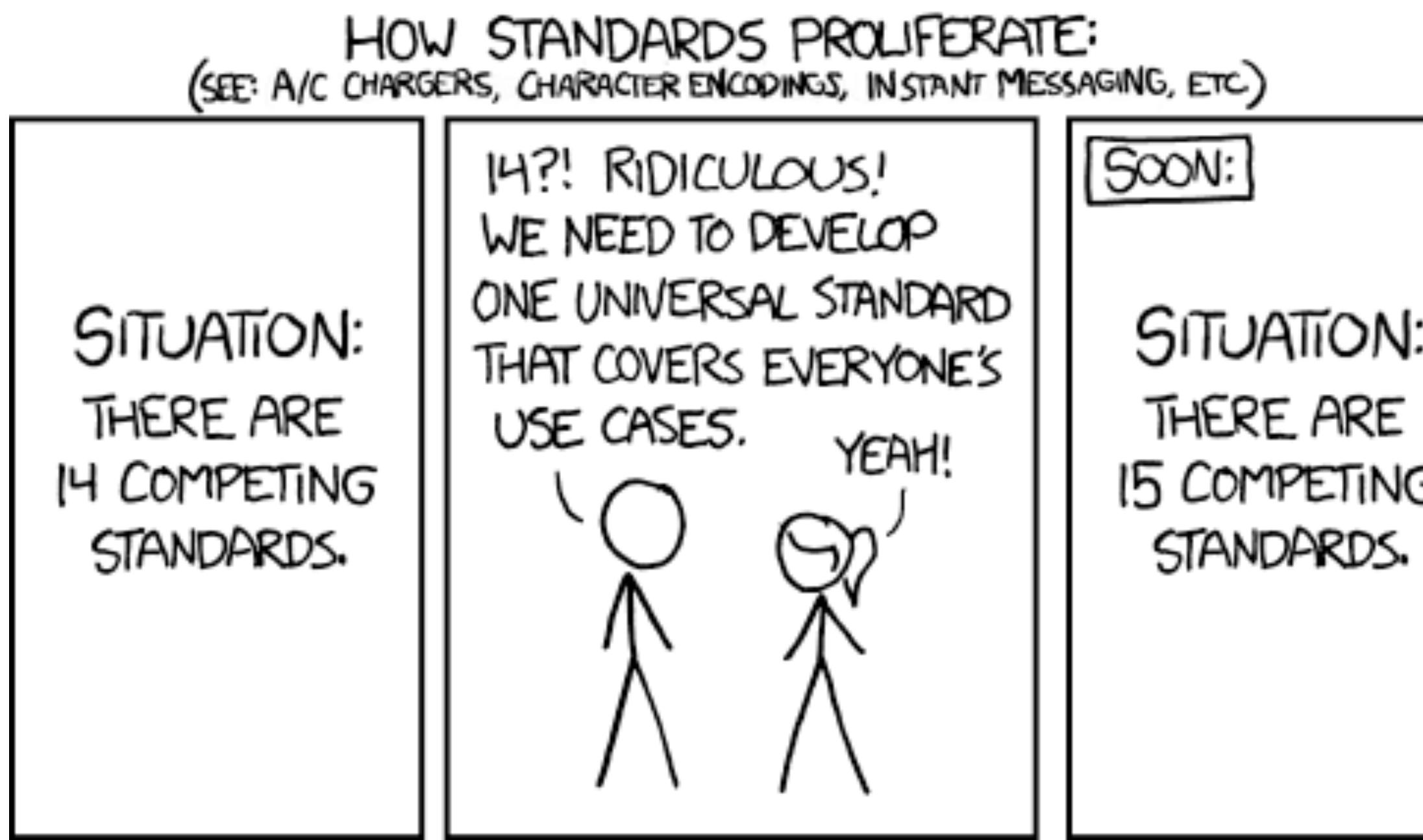
Data re-use depends on standardized, machine-readable metadata

- 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"
```

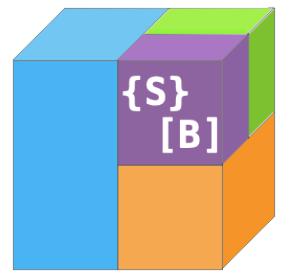
Standardized Data

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



xkcd

```
"label" : "no restriction",
"id" : "DUO:0000004"
},
"provenance" : {
  "material" : {
    "type" : {
      "id" : "EF0:0009656",
      "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 | ◦ Beacon API |
| Used by | ◦ Beacon ◦ Progenetix database schema (Beacon+ backend) |
| Contributors | ◦ Marc Fiume ◦ Michael Baudis ◦ Sabela de la Torre Pernas ◦ Jordi Rambla ◦ Beacon developers... |
| Source (v1.1.0) | ◦ 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 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 | ◦ |
| Used by | ◦ |
| Contributors | ◦ |
| Source (v1.0) | ◦ raw source [JSON] ◦ Github |

Attributes

Type: string
Pattern: ^\w[^:]+:\$
Description: A string that refers to a sender.
VR does not impose any contrain on data, the VR Specification RECOMMENDS that CURIEs are represented as **namespace:accession** or **name**. The VR specification also RECOMMENDS that the **reference** component is an absolute URI. URIs may be located in various contexts. Implementations MAY provide C 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 | ◦ Phenopackets |
| Used by | ◦ Phenopackets |
| Contributors | ◦ GA4GH Data Working Group ◦ Jules Jacobsen ◦ Peter Robinson ◦ Michael Baudis ◦ Melanie Courtot ◦ Isuru Liyanage |
| Source (v1.0.0) | ◦ 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-phenopackets/v1.0.0/Tissue.json [SRC] [HTML] |

Checksum sb-checksum ↗

| | |
|-------------------|---------------------------------|
| {S}[B] Status [i] | proposed |
| Provenance | ◦ GA4GH DRS (`develop` branch) |
| Used by | ◦ GA4GH DRS ◦ GA4GH TRS |
| Contributors | ◦ Susheel Varma |
| Source (v0.0.1) | ◦ 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

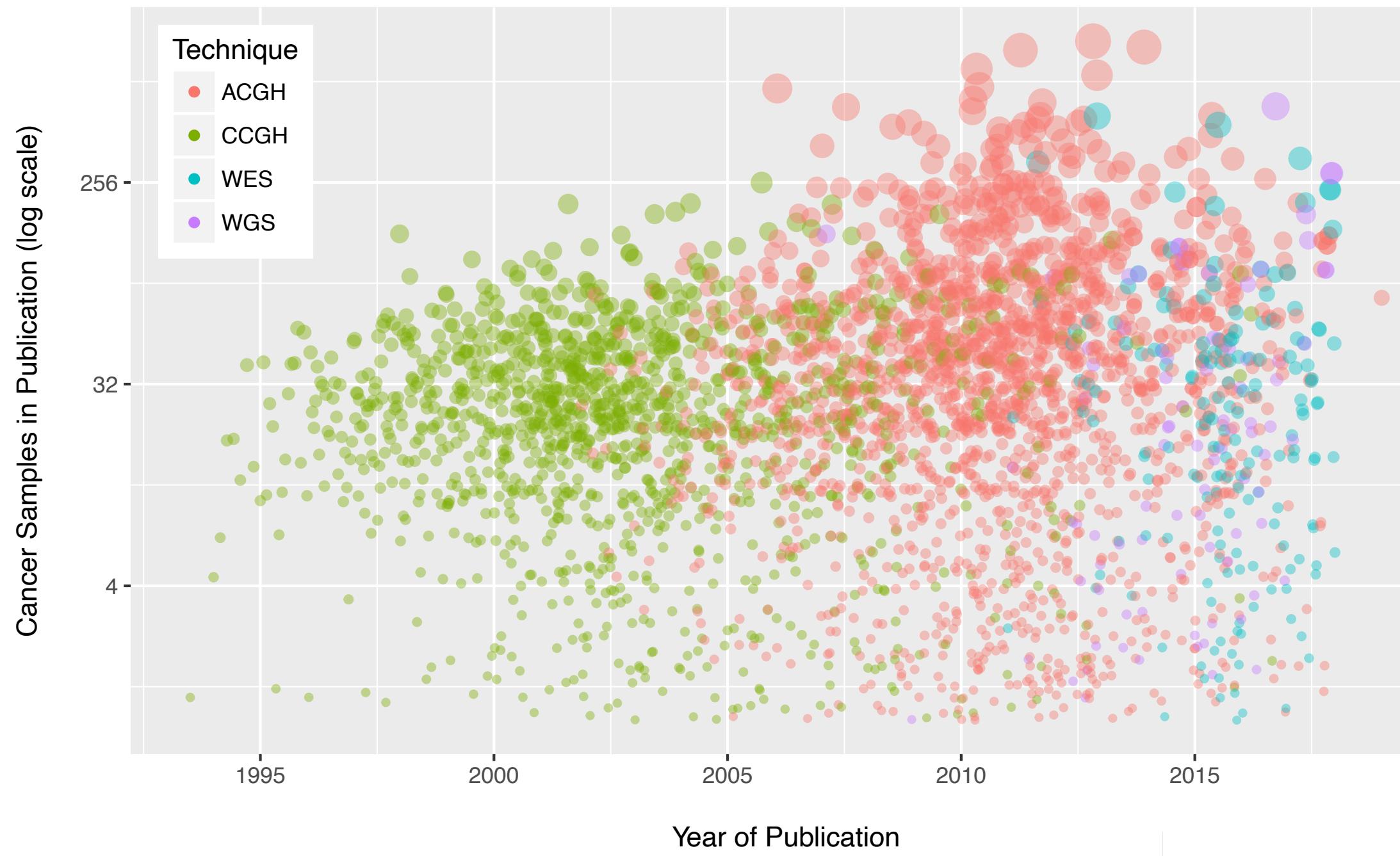
- databases can be accessed through Application Programming Interfaces
- *API : set of routines, protocols, and tools that specifies how software components interact, to exchange data and processing capabilities*
- web API example: implementing geographic maps, with parameters provided by the client (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  
  }  
}
```

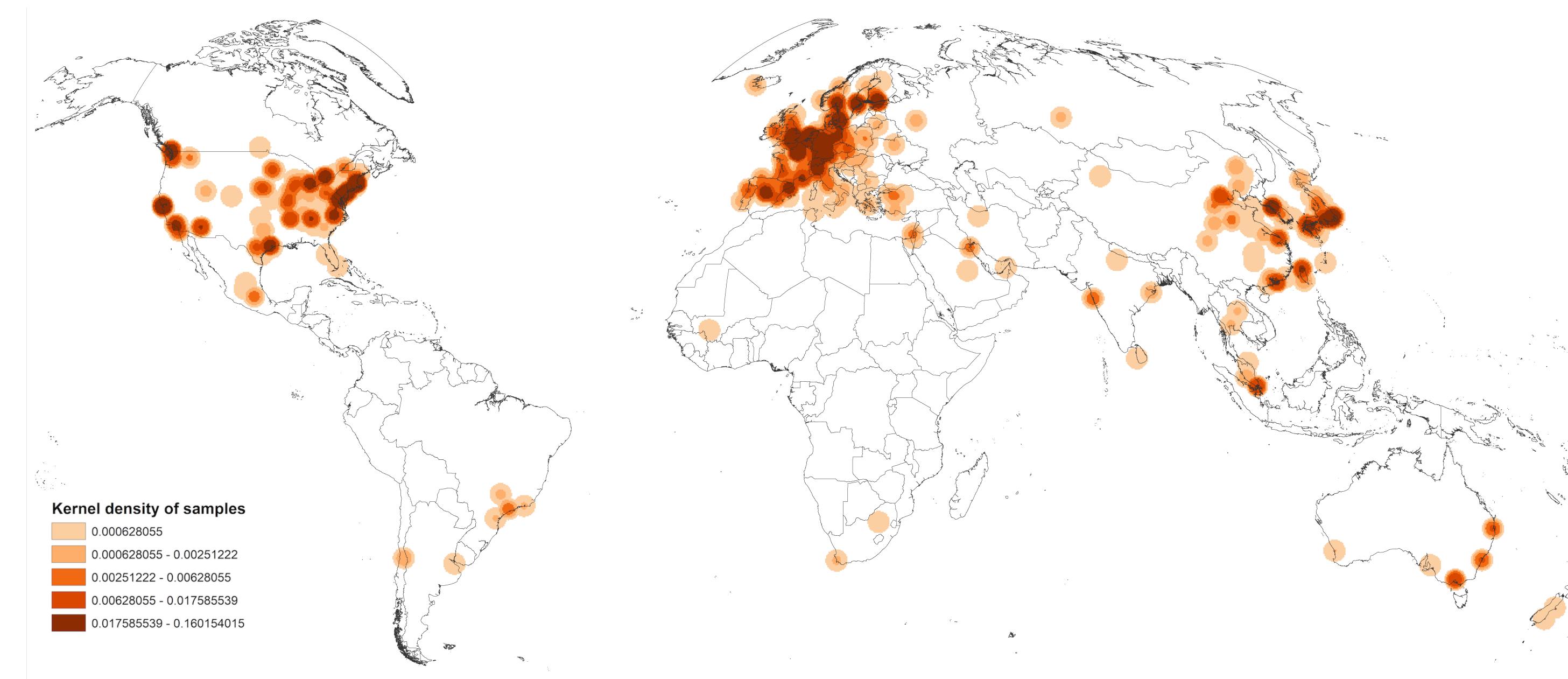
```
{  
  "meta": {  
    "apiVersion": "v2.0.0",  
    "beaconId": "org.progenetix.beacon",  
    "createDateTime": "2015-11-13 00:00:00",  
    "receivedRequestSummary": {  
      "apiVersion": "v2.0.0",  
      -----  
      "response": {  
        "resultSets": [  
          {  
            "exists": true,  
            "id": "progenetix",  
            "info": {  
              "counts": {  
                "callCount": 525,  
                "sampleCount": 515,  
                "variantCount": 247  
              }  
            },  
            "paginatedResultsCount": 247,  
            "results": [  
              {  
                "caseLevelData": [  
                  {  
                    "analysisId": "pgxcs-kftwbzza",  
                    "biosampleId": "pgxbs-kftviv0x",  
                    "id": "pgxvar-5c86619f09d374f2dc3bbfcda"  
                  },  
                  {  
                    "analysisId": "pgxcs-kftwbzza",  
                    "biosampleId": "pgxbs-kftviv0x",  
                    "id": "pgxvar-5c86619f09d374f2dc3bbfcdb"  
                  },  
                  {  
                    "analysisId": "pgxcs-kftwbzza",  
                    "biosampleId": "pgxbs-kftviv0x",  
                    "id": "pgxvar-5c86619f09d374f2dc3bbfcdc"  
                  },  
                  {  
                    "analysisId": "pgxcs-kftwbzza",  
                    "biosampleId": "pgxbs-kftviv0x",  
                    "id": "pgxvar-5c86619f09d374f2dc3bbfcdd"  
                  },  
                  {  
                    "analysisId": "pgxcs-kftwbzza",  
                    "biosampleId": "pgxbs-kftviv0x",  
                    "id": "pgxvar-5c86619f09d374f2dc3bbfcde"  
                  }  
                ]  
              }  
            ]  
          }  
        ]  
      }  
    }  
  }
```

Data Science: Meta-Studies of Metadata



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.



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.

Who is a Bioinformatician?

BIOLOGY IS LARGELY SOLVED.
DNA IS THE SOURCE CODE
FOR OUR BODIES. NOW THAT
GENE SEQUENCING IS EASY,
WE JUST HAVE TO READ IT.

|
IT'S NOT JUST "SOURCE
CODE." THERE'S A TON
OF FEEDBACK AND
EXTERNAL PROCESSING.



bio**in**formatician

BUT EVEN IF IT WERE, DNA IS THE
RESULT OF THE MOST AGGRESSIVE
OPTIMIZATION PROCESS IN THE
UNIVERSE, RUNNING IN PARALLEL
AT EVERY LEVEL, IN EVERY LIVING
THING, FOR FOUR BILLION YEARS.

|
IT'S STILL JUST CODE.



OK, TRY OPENING GOOGLE.COM
AND CLICKING "VIEW SOURCE."

|
OK, I-... OH MY GOD.

|
THAT'S JUST A FEW YEARS OF
OPTIMIZATION BY GOOGLE DEVs.
DNA IS THOUSANDS OF TIMES
LONGER AND WAY, WAY WORSE.

|
WOW, BIOLOGY
IS IMPOSSIBLE.



Randall Munroe: <https://xkcd.com/1605/>

bio**in**formatician

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?

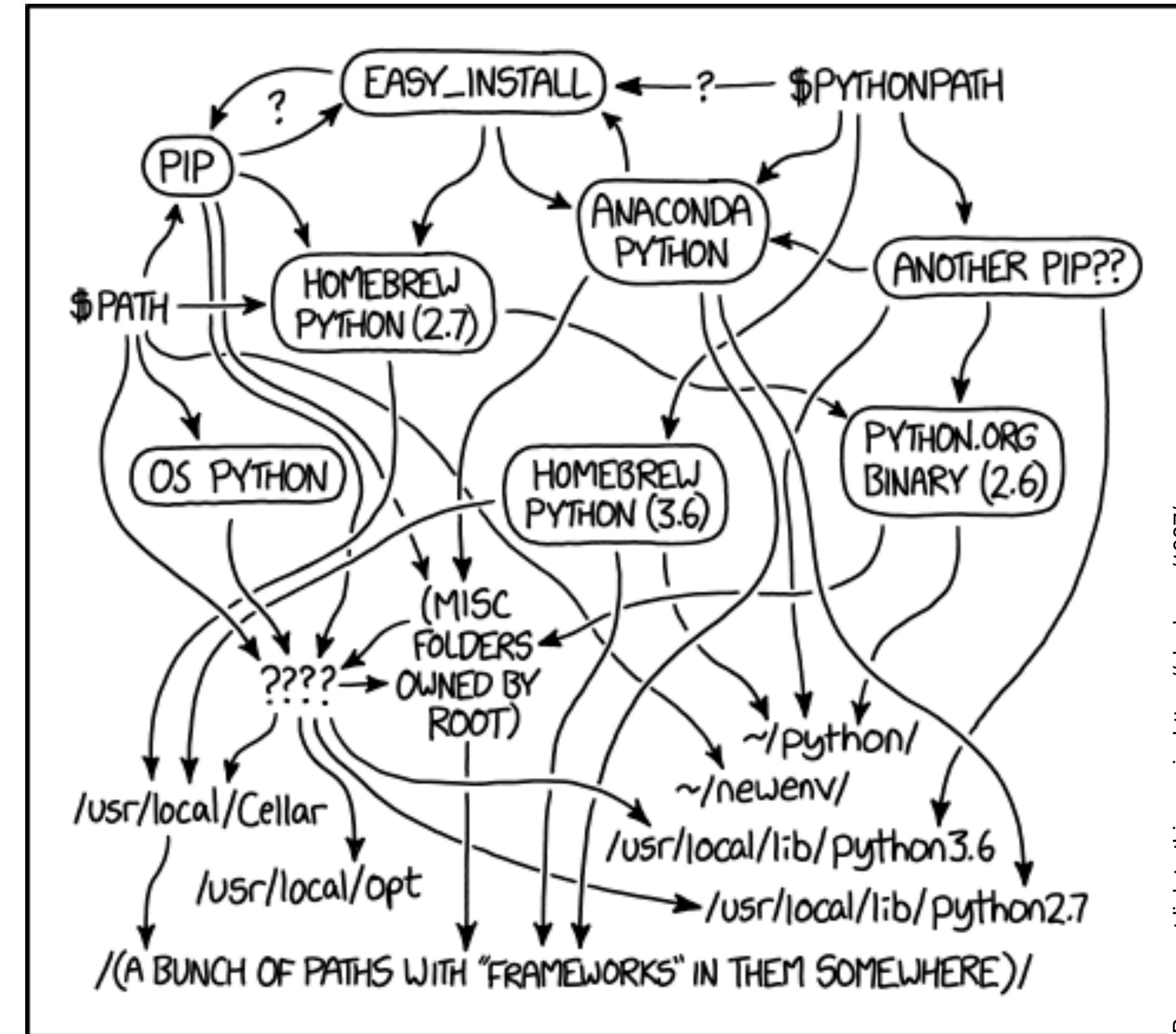
- **Bioinformatics**

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...



Permanent link to this comic: <https://xkcd.com/1987/>

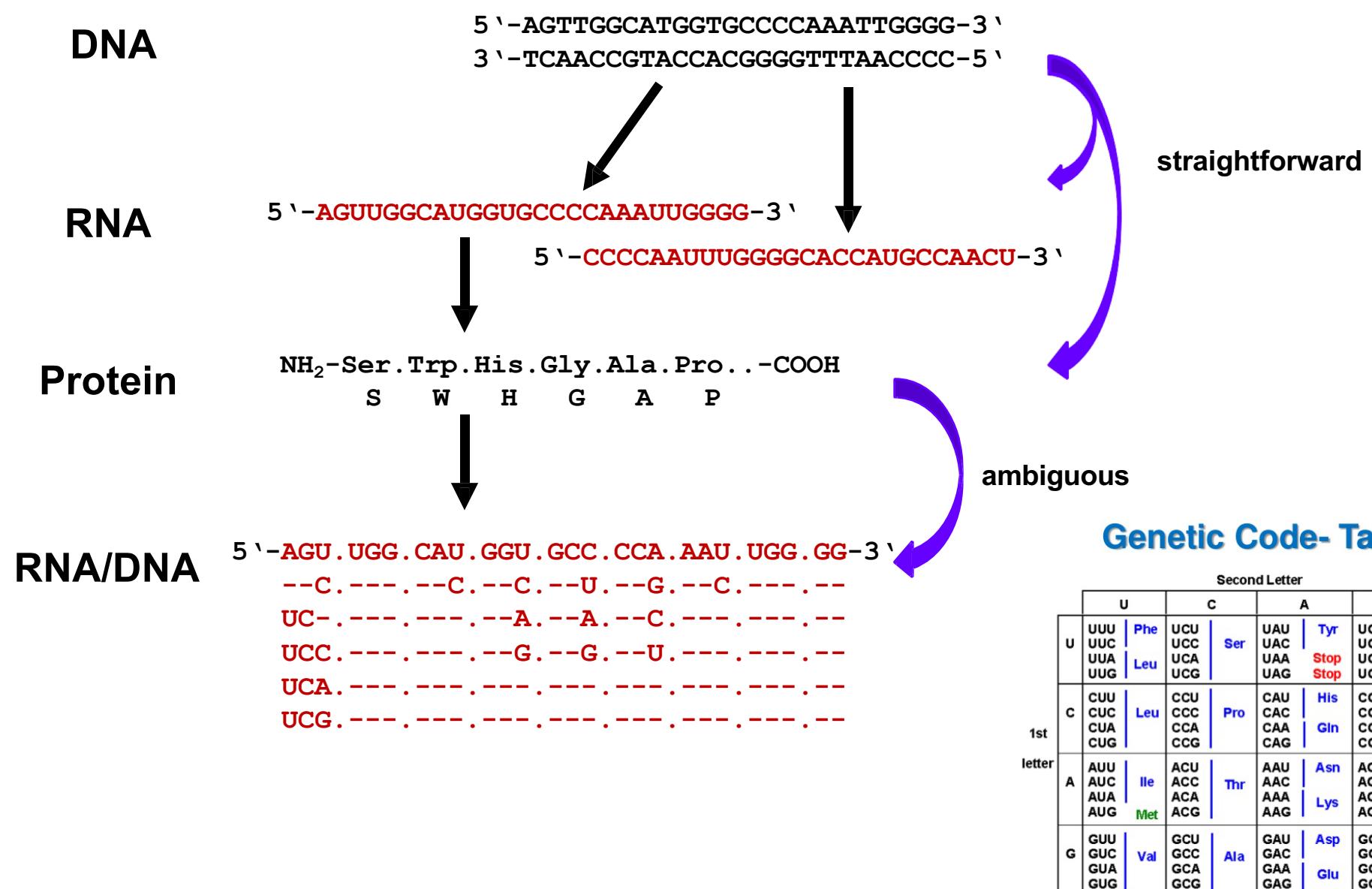
BIO390: Course Schedule

- 2022-09-20: Michael Baudis - What is Bioinformatics? Introduction and Resources
- **2022-09-27: Christian von Mering - Sequence Bioinformatics**
- 2022-10-04: Mark Robinson - Statistical Bioinformatics
- 2022-10-11: Valentina Boeva (ETHZ) - Machine Learning for Biological Use Cases
- 2022-10-18: Izaskun Mallona - Regulatory Genomics and Epigenomics
- 2022-10-25: Shinichi Sunagawa (ETHZ) - Metagenomics
- 2022-11-01: Katja Baerenfaller (SIAF) - Proteomics
- 2022-11-08: Andreas Wagner - Biological Networks
- 2022-11-15: Patrick Ruch - Text mining & Search Tools
- 2022-11-22: Ahmad Aghaebrahimian (ZHAW) - Semantic Web
- 2022-11-29: Michael Baudis - Building a Genomics Resource
- 2022-12-06: Valérie Barbie (SIB) - Clinical Bioinformatics
- 2022-12-13: Michael Baudis - Genome Data & Privacy | Feedback
- 2022-12-20: 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

AAGCTTACCAAAATTGAAGGGACGTTGACGTAGGGGGACGCTTTAG
GACGCTTTAGTTAGGCCACCGGTATTTAGC

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

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

AAGCTTACCAAAATTGAAGGGACGTTGACGTAGGGGGACGCTTTAG
AATCTAGCAATTATTGAAGGGACGTTGACGAAGGGGTTCGCTACCG

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

AAGCTTACCAAAATTGAAGGGACGTTGACGTAGGGGGACGCTTTAG
AATCTAGCAATTATTGAAGGGACGTTGACGAAGGGGTTCGCTACCG

Statistical Bioinformatics

Mark Robinson



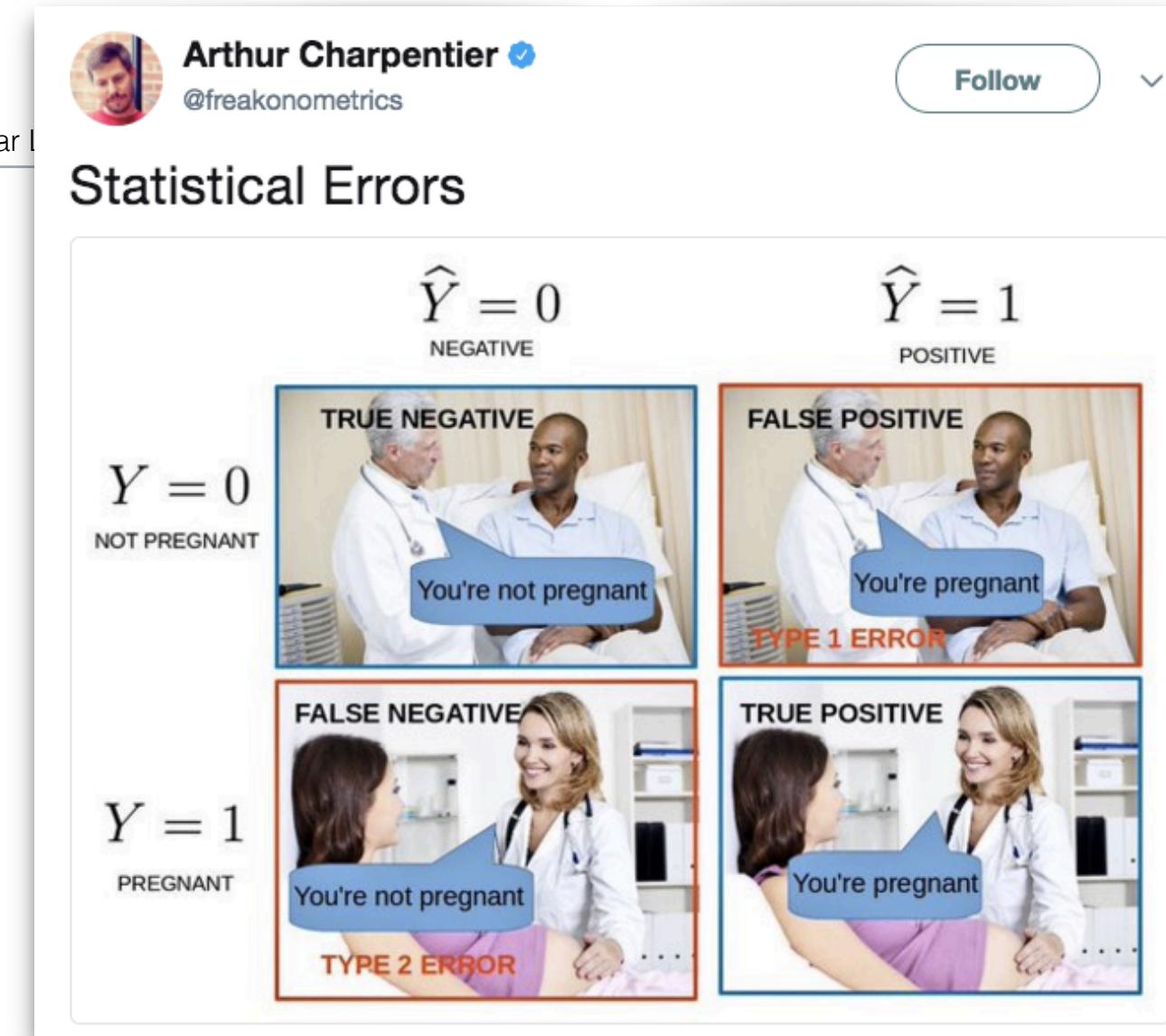
University of
Zurich ^{UZH}

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



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

40

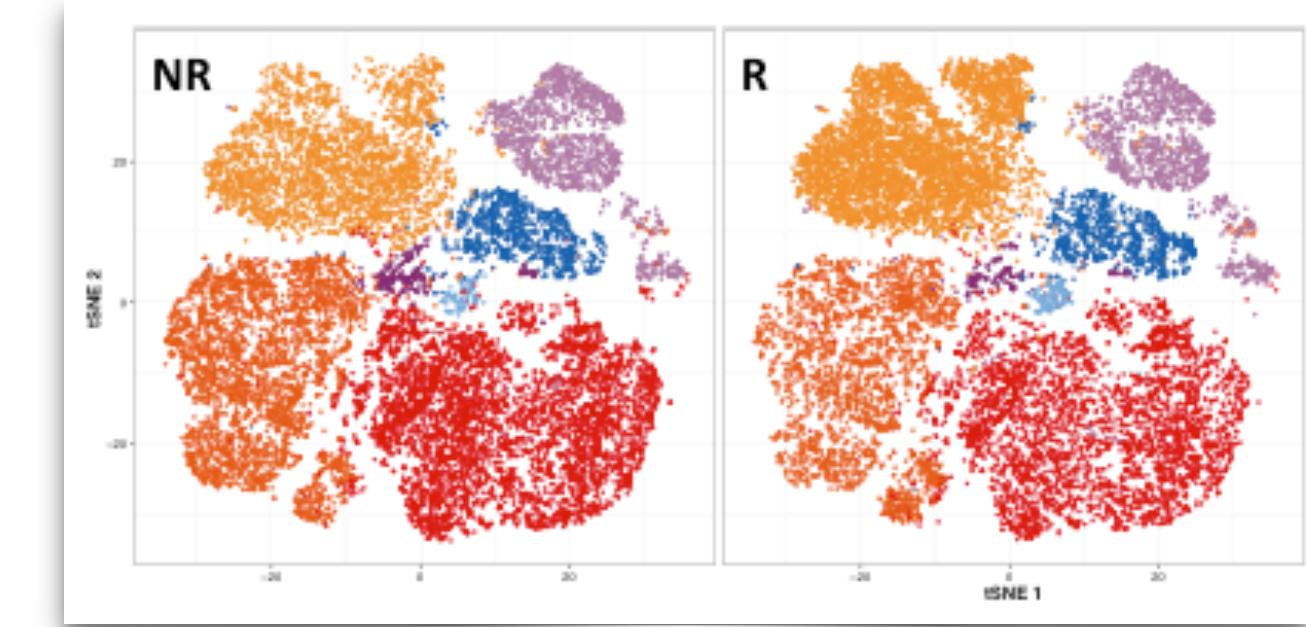


University of
Zurich ^{UZH}

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.

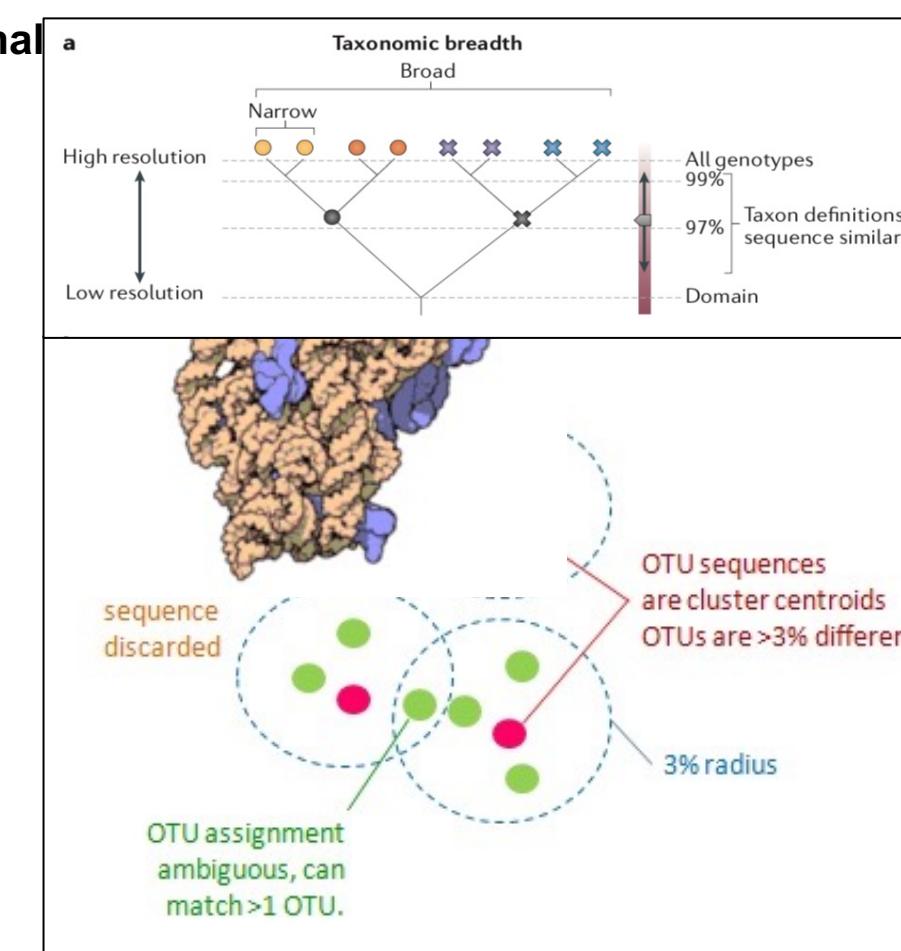
30

Metagenomics

Shinichi Sunagawa (ETHZ)

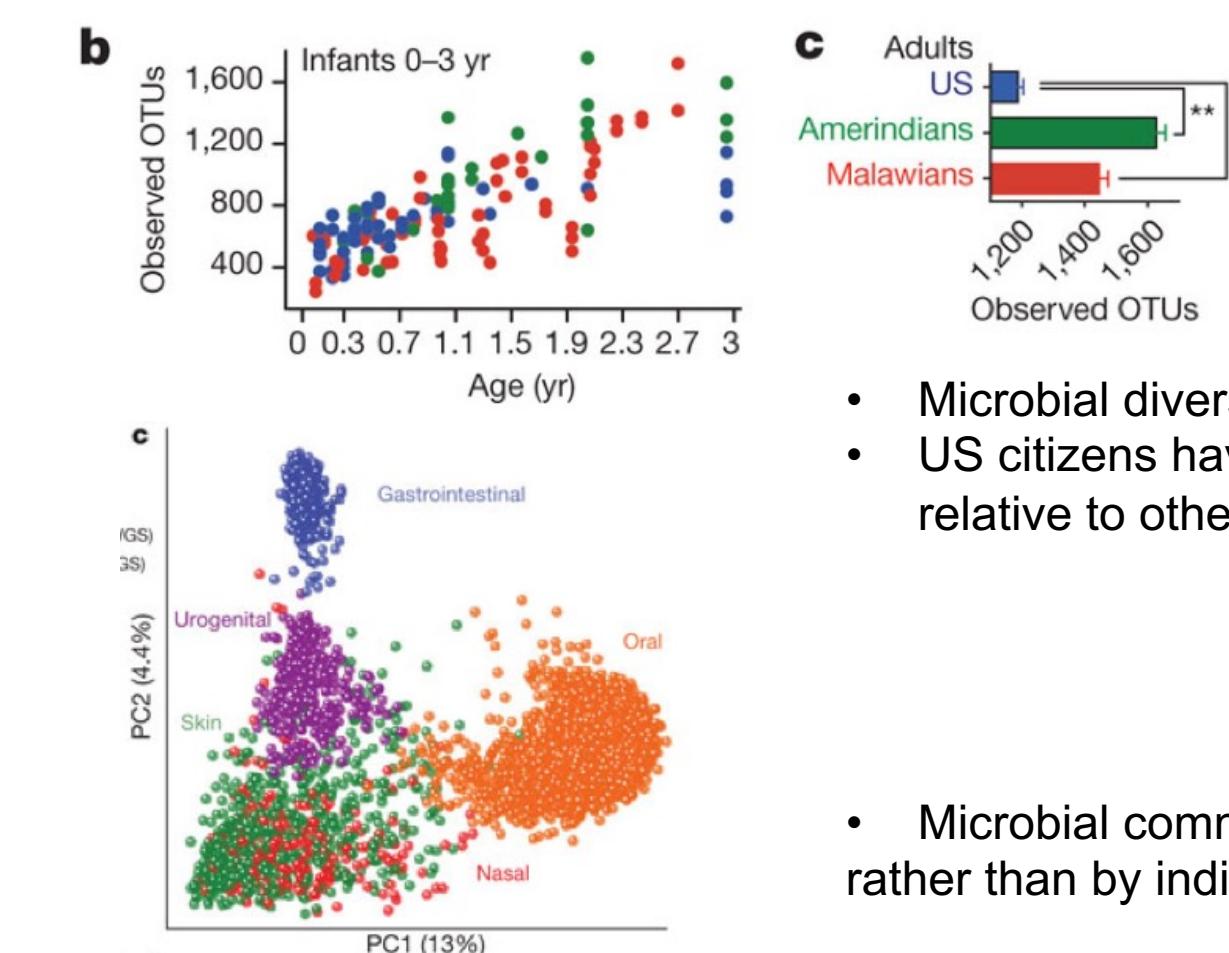
Review: 16S rRNA-based Operational Taxonomic Units (OTUs)

- 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)



Metagenomics Part I | 26-Oct-21 | 10

Applied examples I



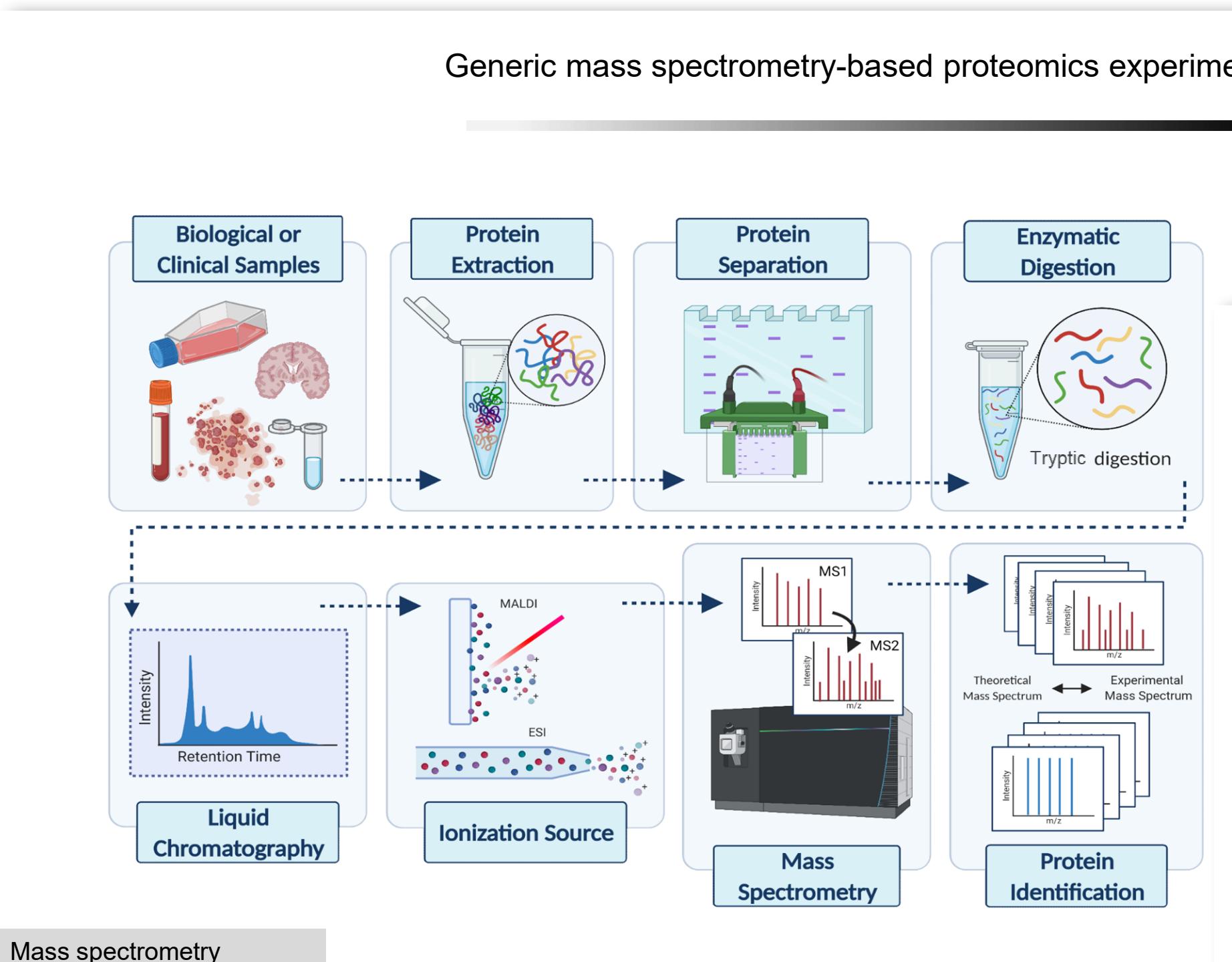
- Microbial diversity in human gut increases with age
- US citizens have harbor less diverse gut microbiota relative to other populations
- Microbial communities cluster by human body site rather than by individual

Top: Yatsunenko *et al.* *Nature* (2012); Bottom: C Huttenhower *et al.* *Nature* (2012)

Proteomics

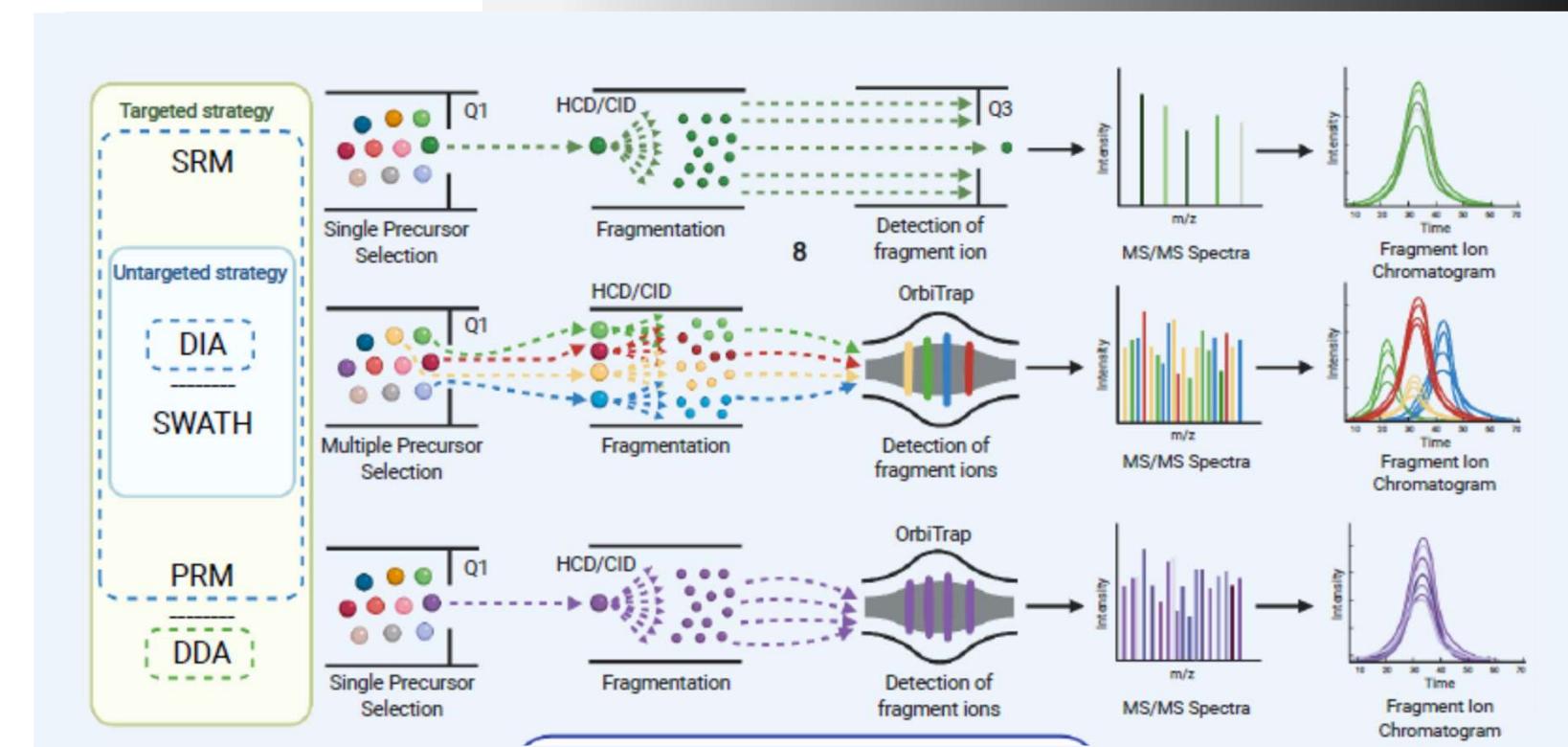
Katja Bärenfaller (SIAF)

Generic mass spectrometry-based proteomics experiment



Mass spectrometry

Hypothesis-driven, targeted bottom-up proteomics approaches



Radzikowska et al., EAACI Position Paper, in revision

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

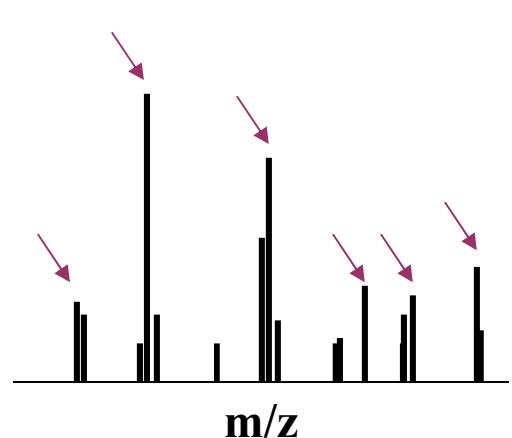
DIA/SWATH: Data Independent Acquisition/Sequential Windowed Acquisition of All Theoretical Mass Spectra

fragment ion signals from a precursor ion

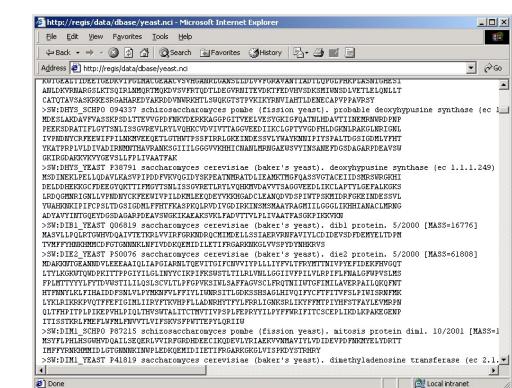
Peptide Mass Fingerprint

Identifying peptides using an MS spectrum:

List of peptide masses
from MS scan



Sequence database



Search
algorithm

Identified peptide/protein

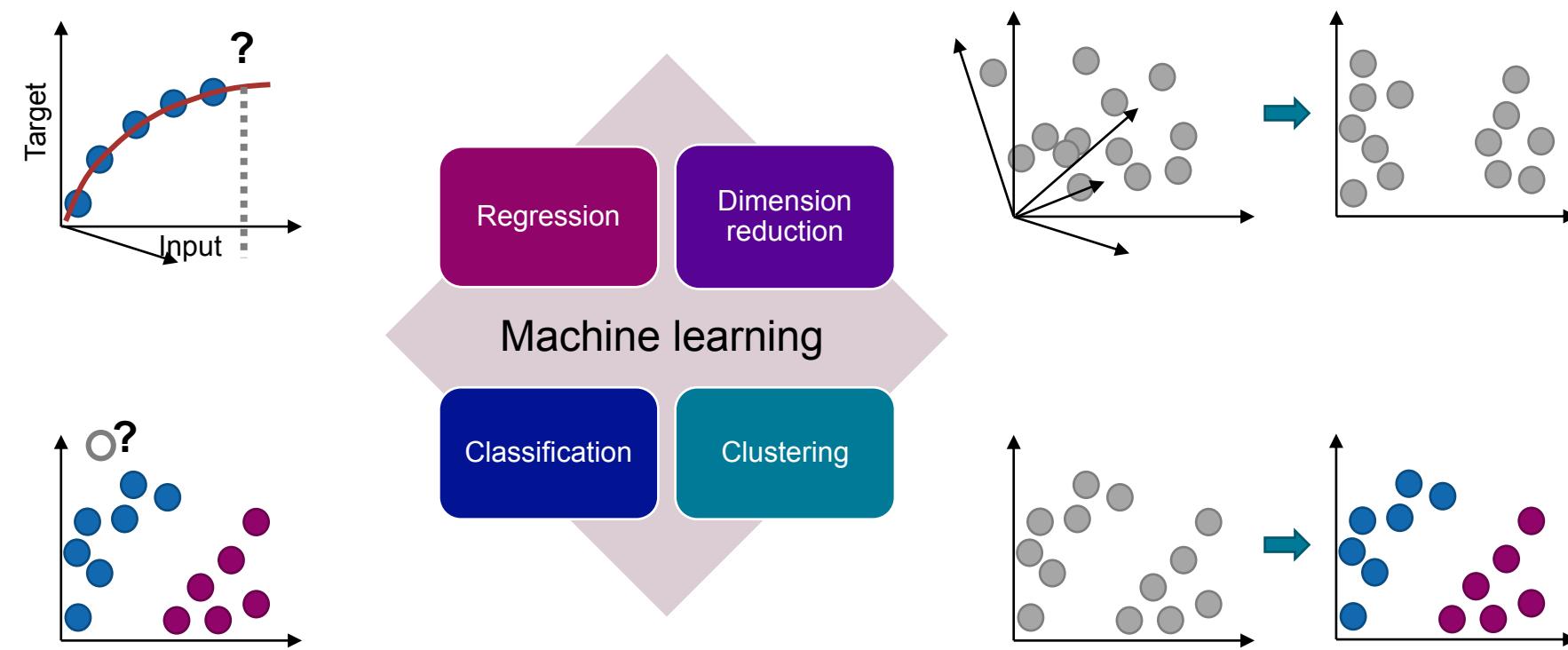
- Peptide spectrum assignment with Peptide Mass Fingerprinting is only advisable with samples of low complexity and small sequence databases, as the number of all possible peptides with a given mass over charge is huge in large sequence databases.

Mass spectrometry

Machine Learning for Biological Use Cases

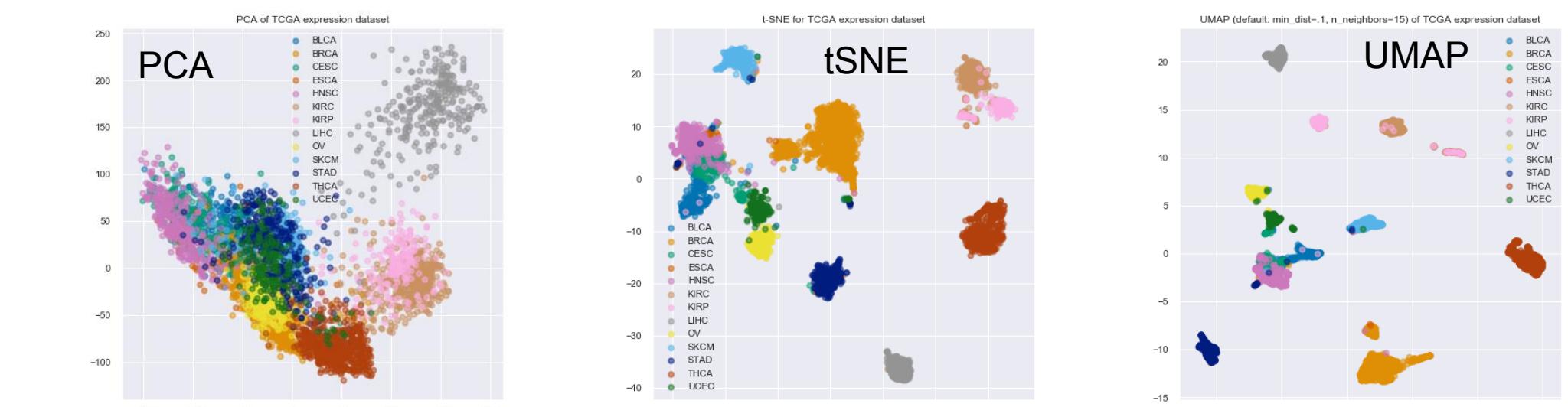
Valentina Boeva (ETHZ)

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)

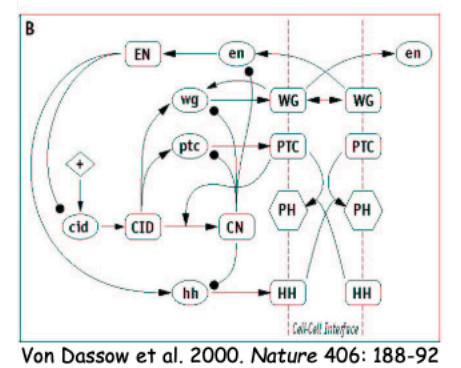
Biological Networks

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

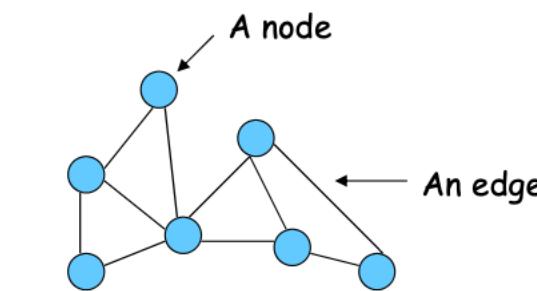
2. Genome-scale networks
(hundreds to thousands of gene products)

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



Mathematical characterization based on qualitative
understanding of network topology

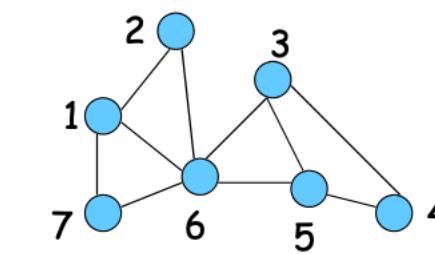
Graphs



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)

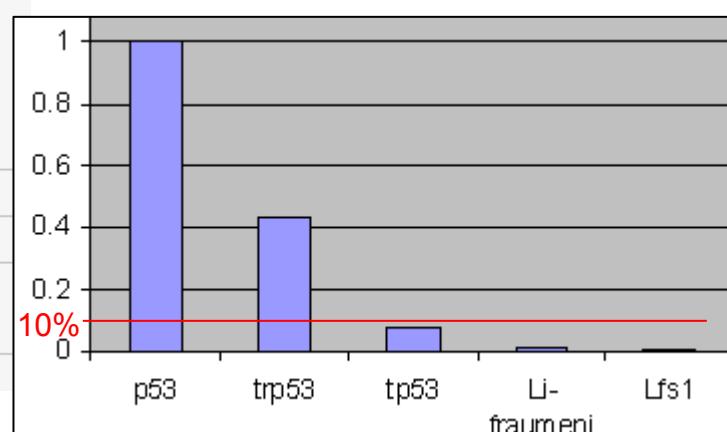
Features

- 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 !

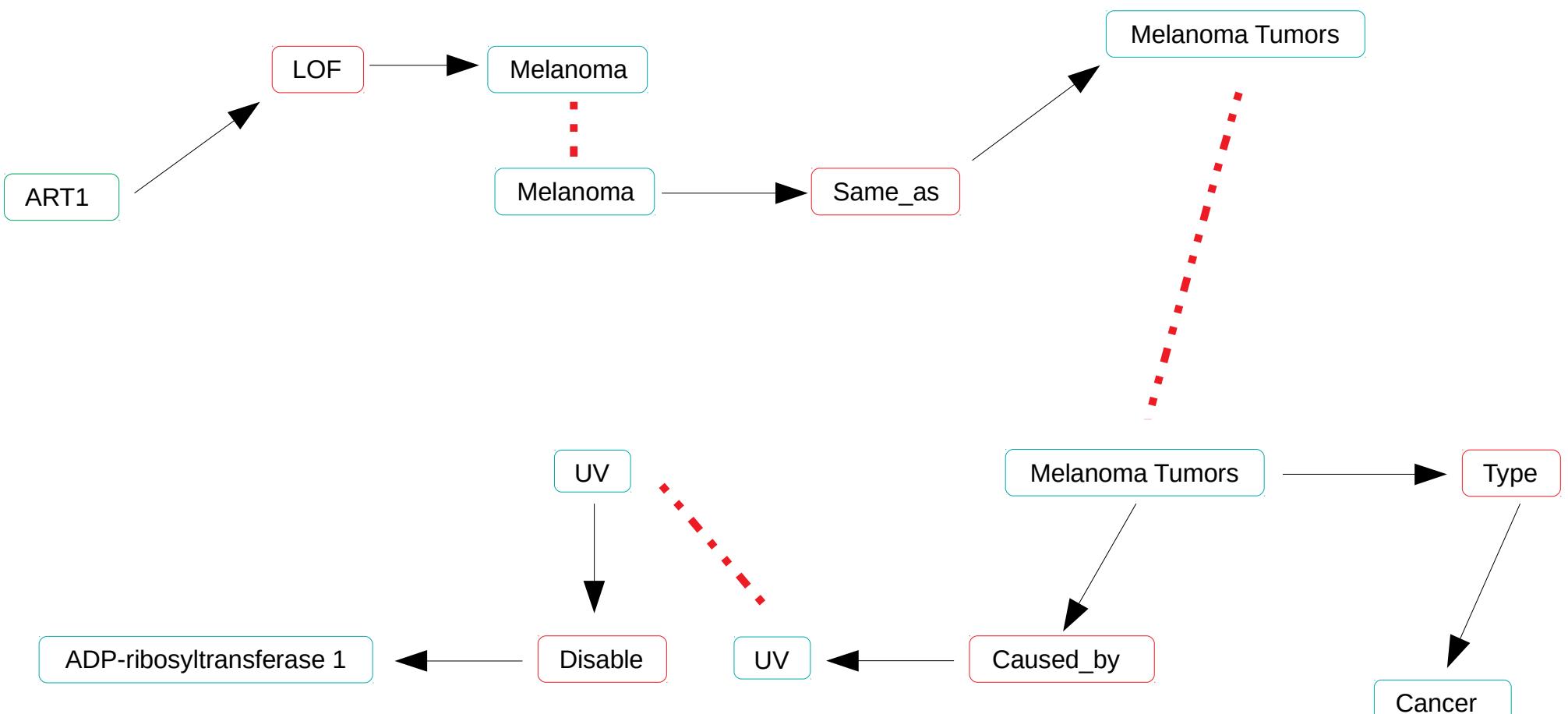
| | | |
|-------------------------------------|---------------------|--|
| □ Antigen NY-CO-13 | Protein | SwissProt:P04637 |
| □ Cellular tumor antigen p53 | Protein [preferred] | SwissProt:P04637 |
| □ FLJ92943 | Gene | EntrezGene:7157 |
| □ LFS1 | Gene | EntrezGene:7157 HGNC:11998 |
| □ Li-Fraumeni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

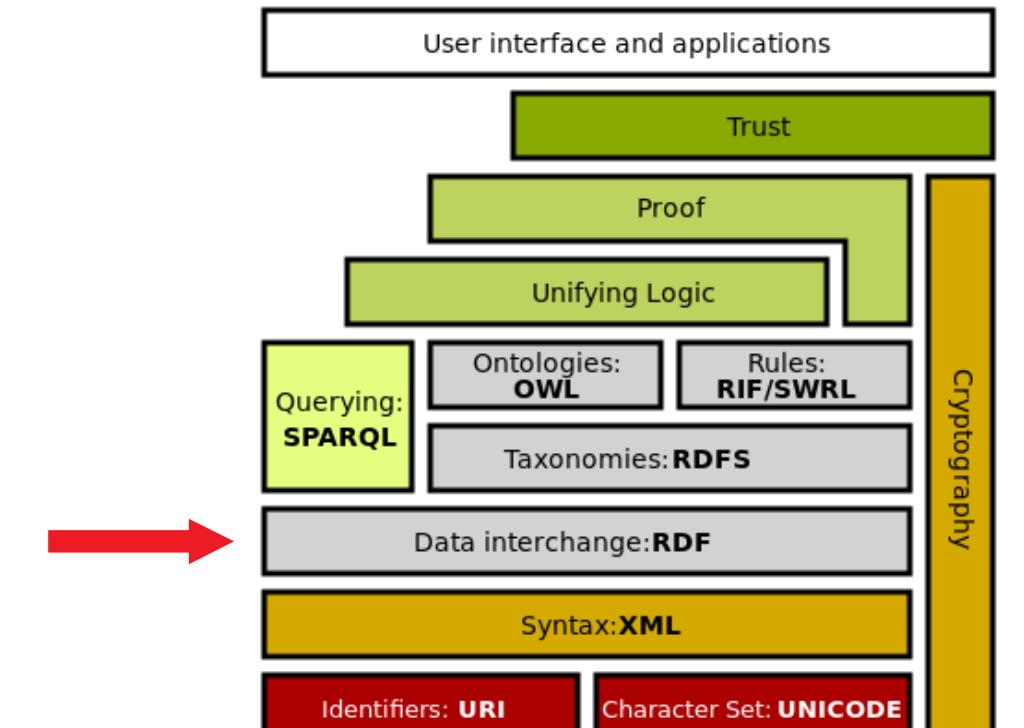
Predicates are labeled

Predicates are directed

RDF is a graph model

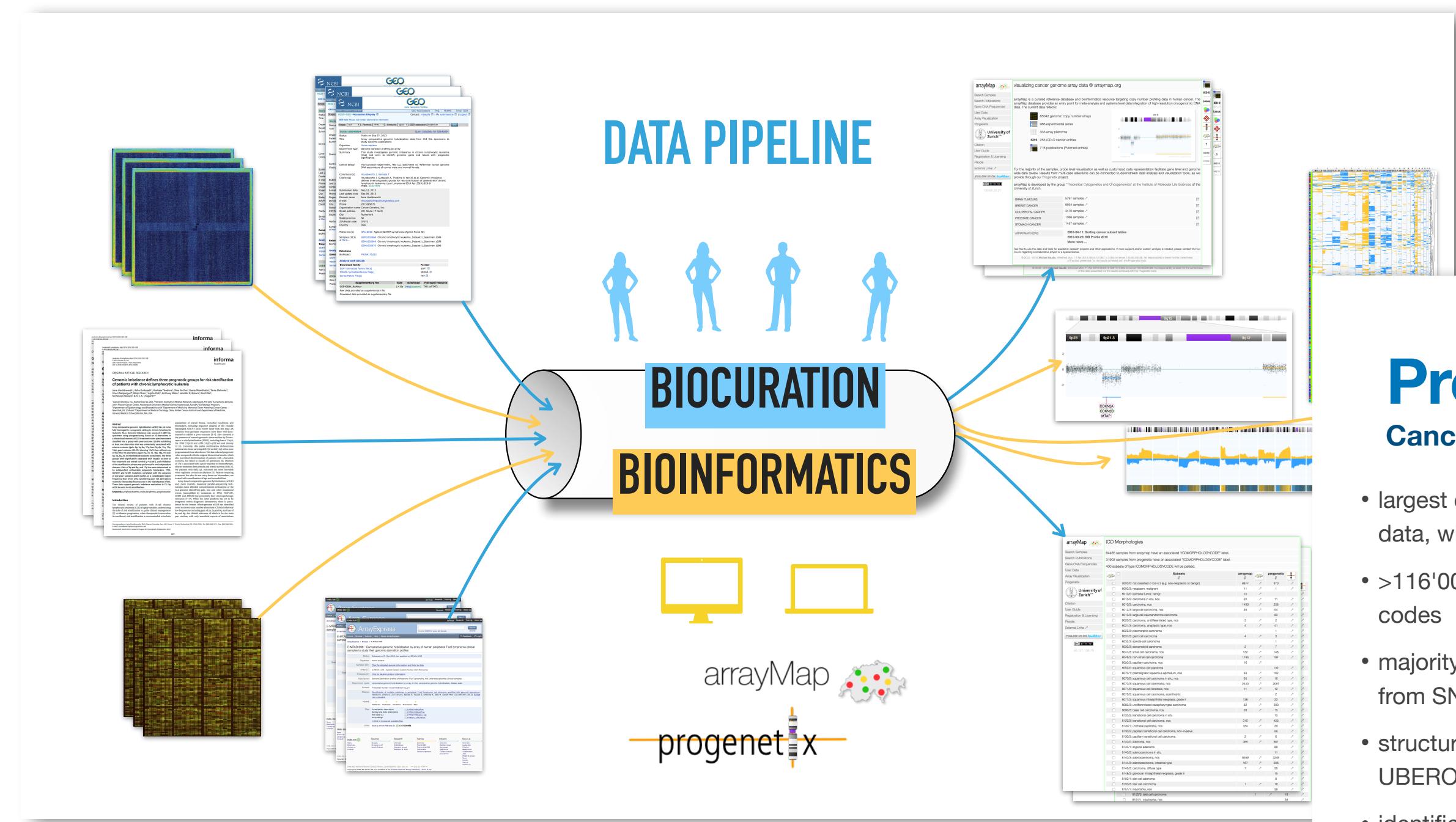
RDF serialization:

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



Building a Genomics Resource

Michael Baudis



progenetix

... using
archaic
tools



Let's
build a
database!

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

The screenshot shows the Progenetix homepage with the following sections:

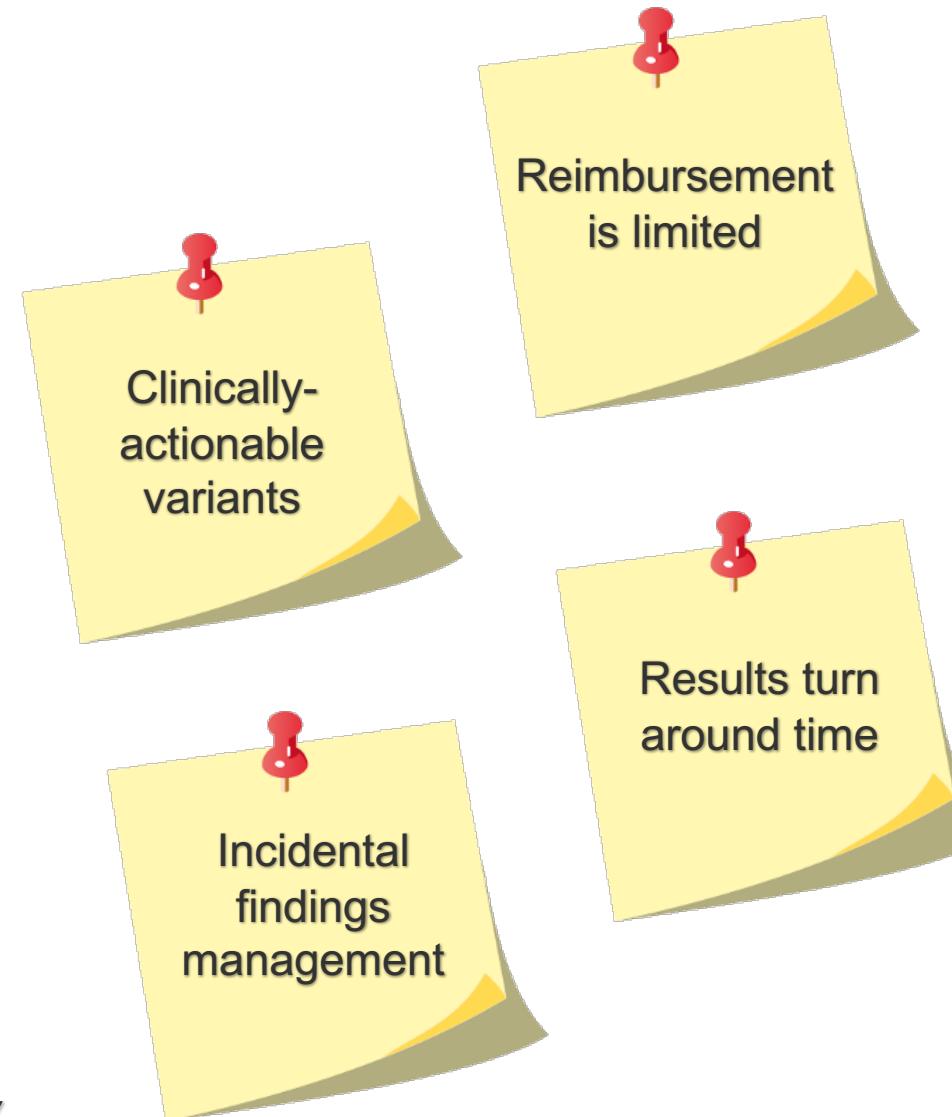
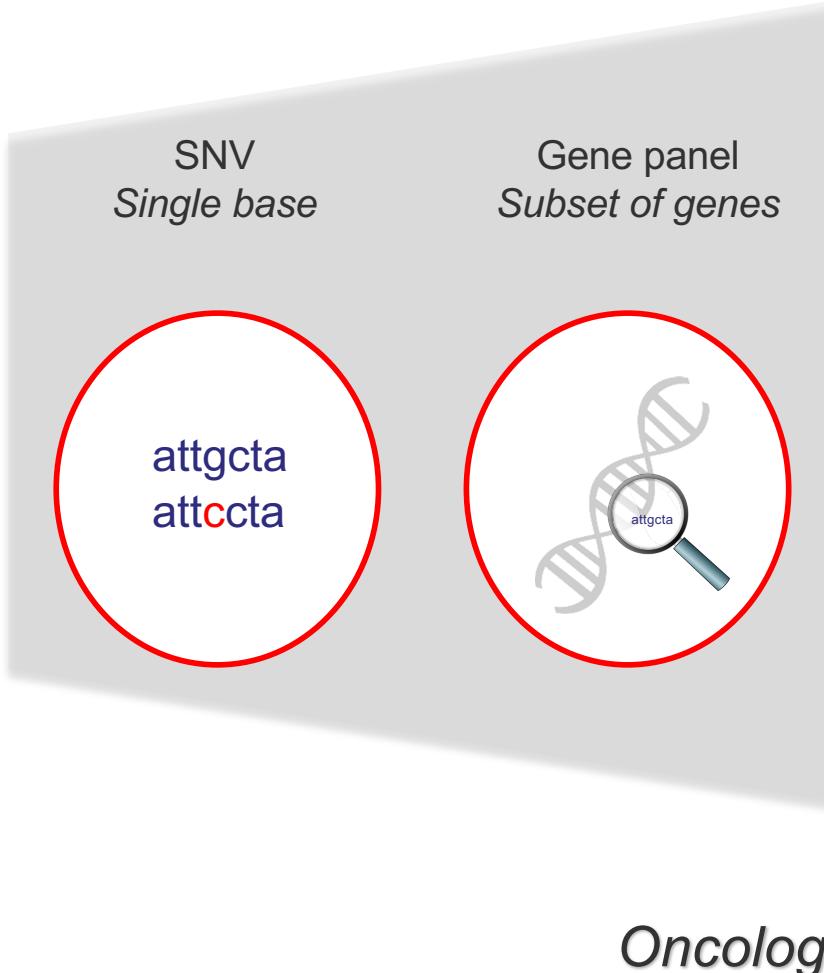
- progenetix**: Logo and navigation links.
- Cancer genome data @ progenetix.org**: Overview of mutation data in cancer, focusing on copy number abnormalities (CNV/CNA).
- Studies & Cohorts**: Links to arrayMap, TCGA Samples, DIPG Samples, Gao & Baudis, 2021, and Cancer Cell Lines.
- Publication DB**: Services like NCIt Mappings and UBERON Mappings.
- Upload & Plot**: Download Data.
- Beacon+**: Progenetix Info, About Progenetix, Use Cases, Documentation, and Baudisgroup @ UZH.
- Progenetix Use Cases**: Local CNV Frequencies, example for aggregated CNV data in 362 samples in Breast Cancer by AJCC v6 Stage.
- Cancer CNV Profiles**: Example for regional copy number gains and losses for all 22 autosomes.
- Cancer Genomics Publications**: Overview of publications, mentioning 4025 annotated references.

© 2000 - 2021 Progenetix Cancer Genomics Information Resource by the Computational Oncogenomics Group at the University of Zurich and the Swiss Institute of Bioinformatics SIB is licensed under CC BY 4.0. No responsibility is taken for the correctness of the data presented nor the results achieved with the Progenetix tools.

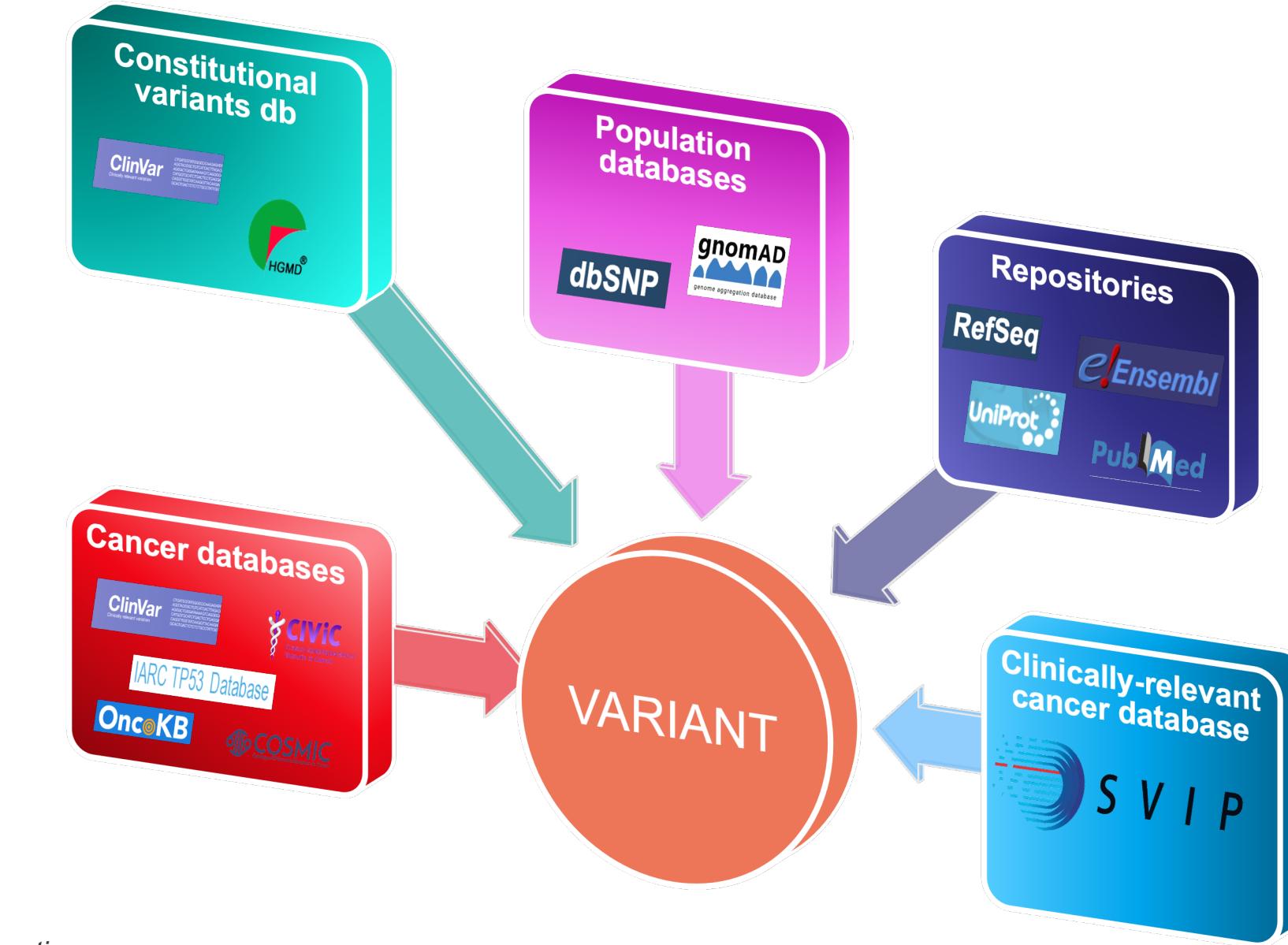
Clinical Bioinformatics

Valérie Barbié (Director SIB Clinical Bioinformatics)

Scale matters



Knowledge bases



Genomic Data & Privacy: Risks & Opportunities

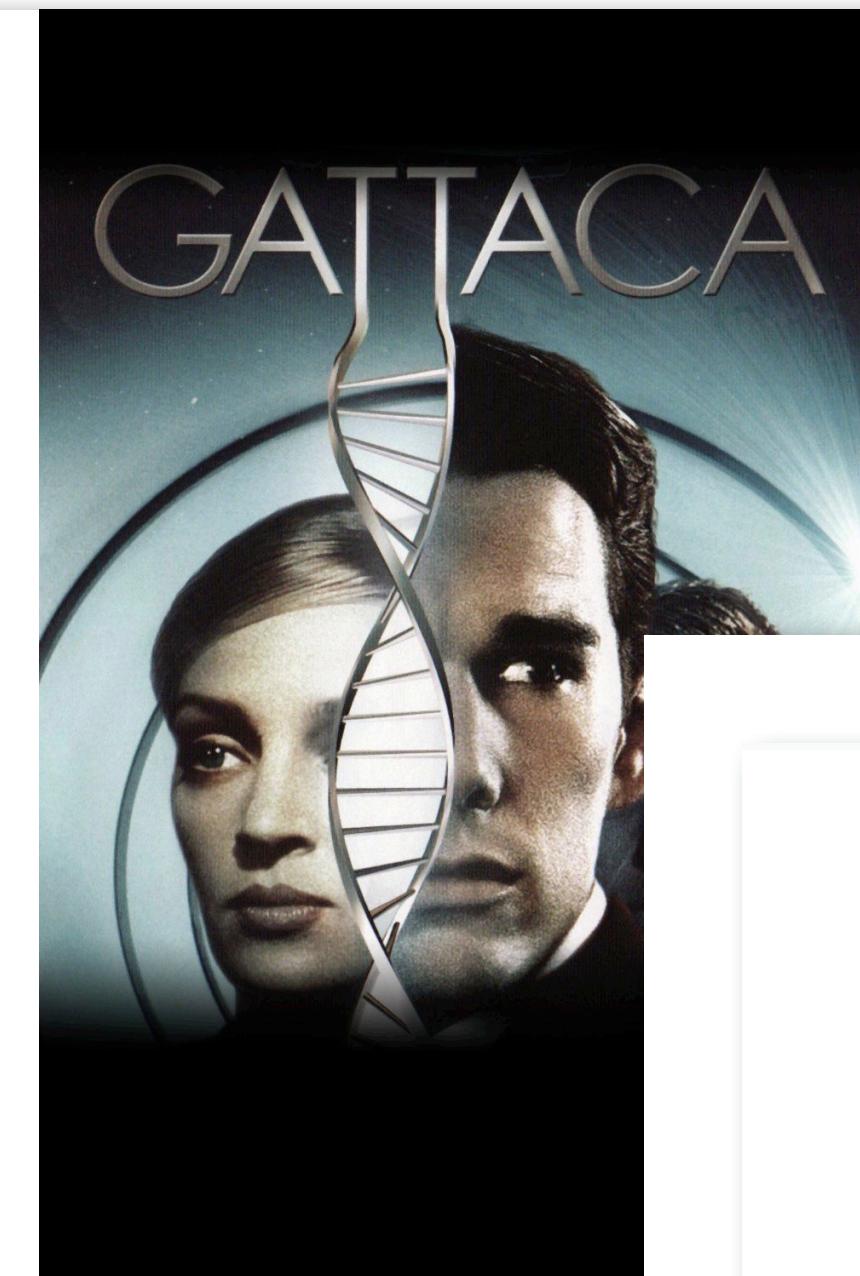
Michael Baudis

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

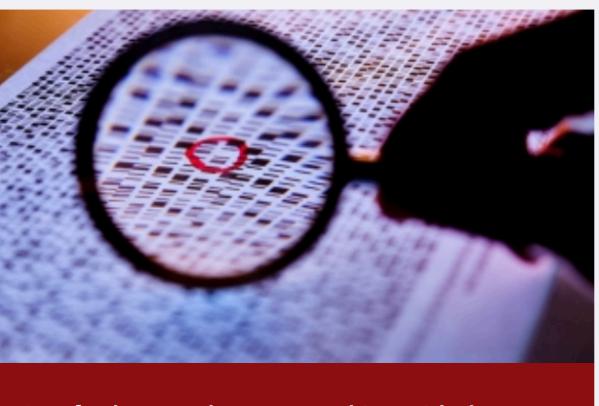
OCT 29
2015

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

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

Human Genetics, also bears importantly on the search as those from different people at a crime



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*

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 unreplicable 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>

BIO390: Introduction to Bioinformatics

Lecture I: What are Bioinformaticians doing? Example of Developing "Federated Human Data" concepts

The trouble with human genome variation



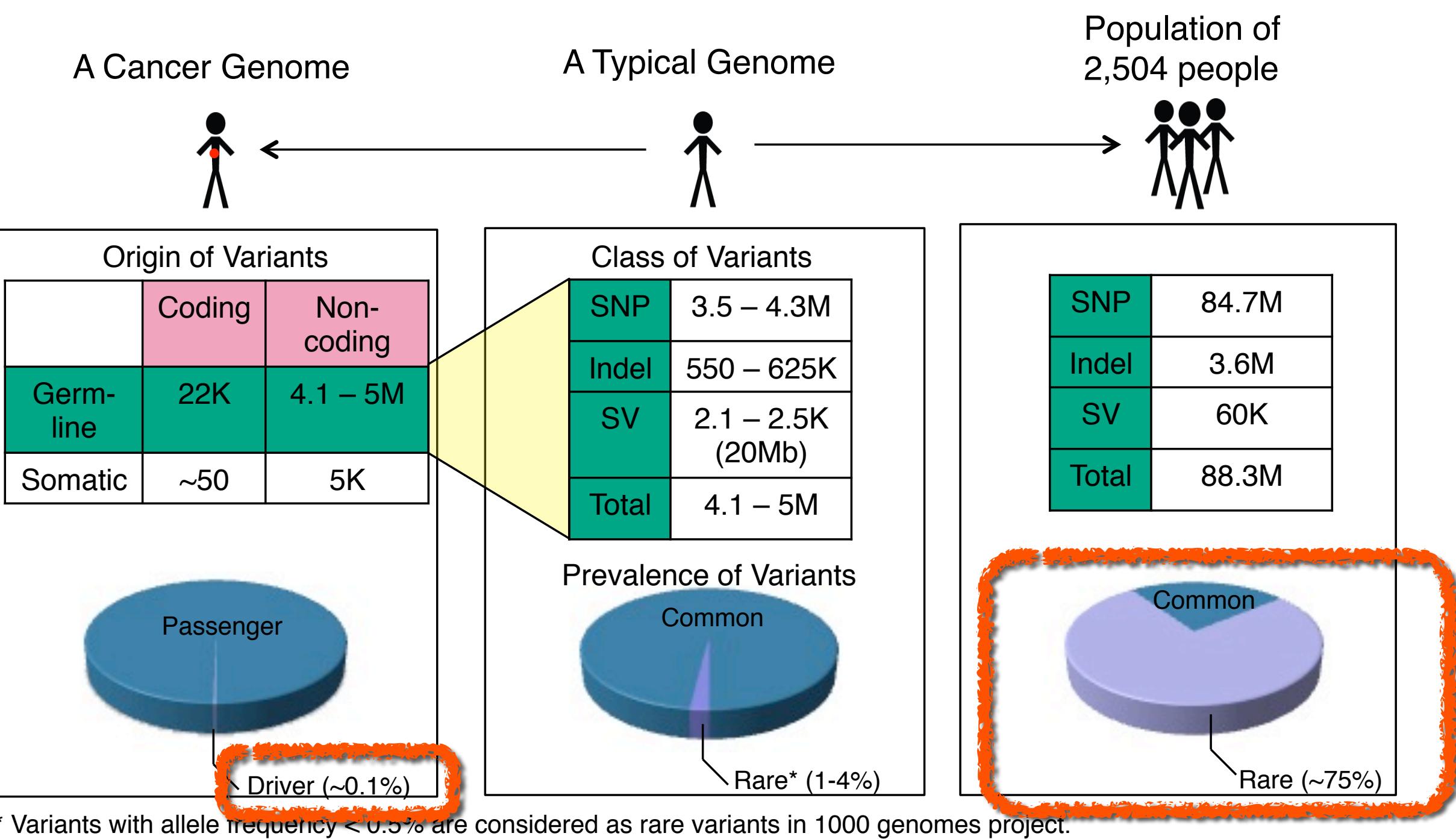
Conclusions from the analysis of variation in the human genome

- 1. Humans are all very similar to each other
 - Two humans will show about 99.9% sequence identity with each other. In other words, only about 1 in 1'000 bp is different between two individuals.
 - Humans show about 98% sequence identity to chimps. So two humans are still much more similar to each other than either is to the monkey.
- 2. Humans are very different from each other
 - Two typical humans will likely have over 1'000'000 independent sequence differences in their genomes.

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

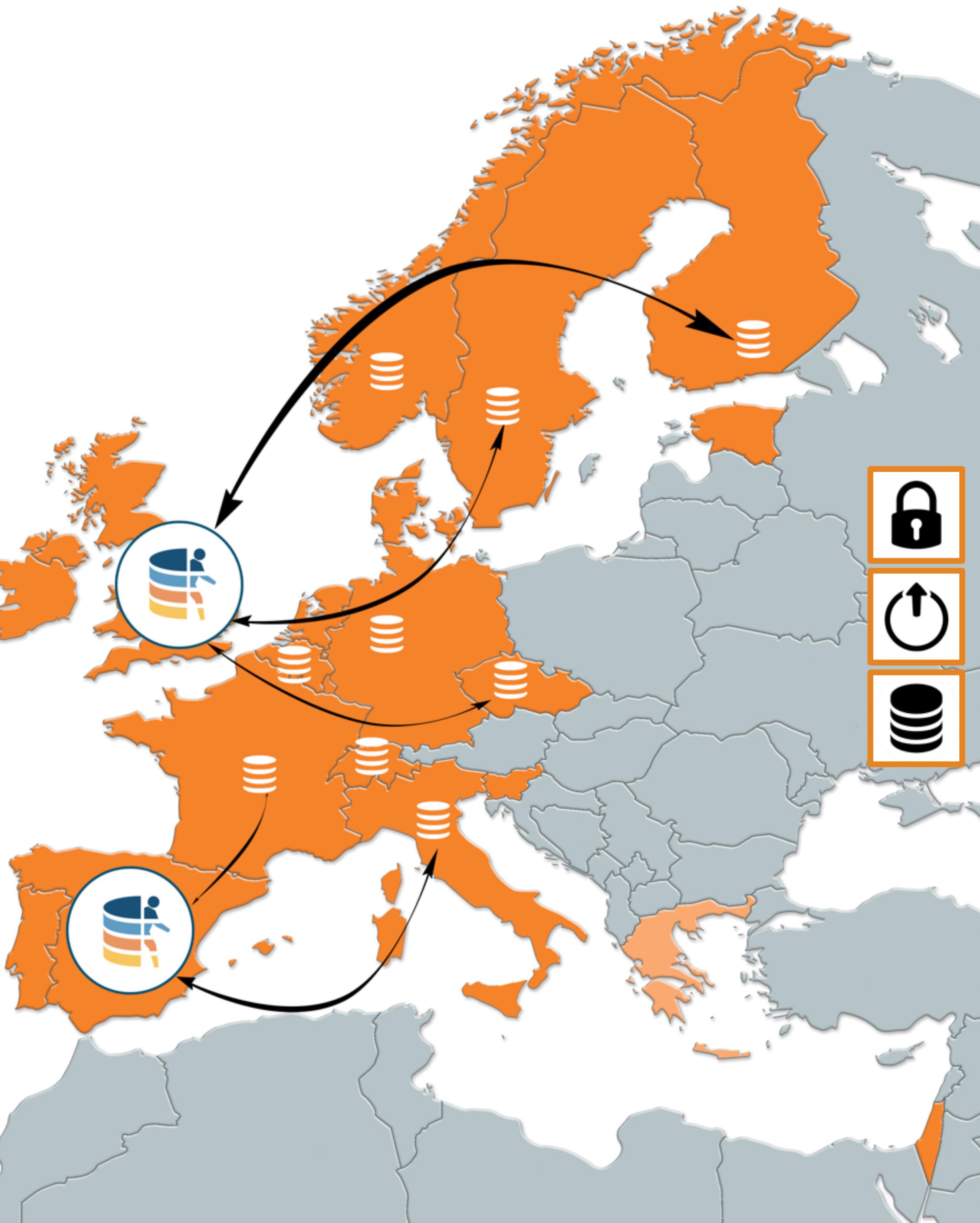


The 1000 Genomes Project Consortium, Nature. 2015. 526:68-74
Khurana E. et al. Nat. Rev. Genet. 2016. 17:93-108

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

Federation of human genome data

- Many national datasets from human research participants needs to be stored locally
- ELIXIR developing a federation with shared metadata (FAIR) and local data store (secure)
- Linking local EGA to national clouds – and international access (ELIXIR-AAI)



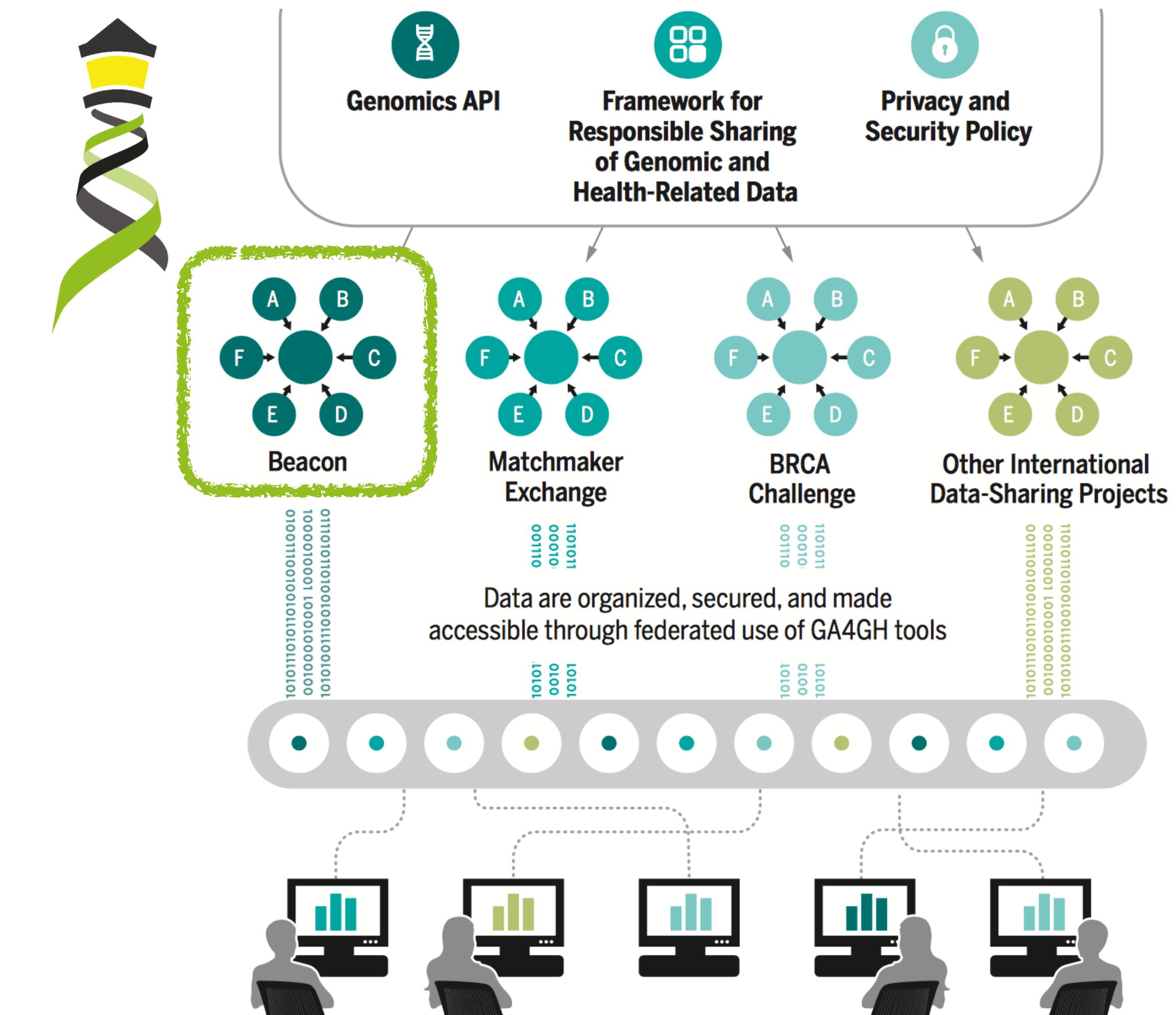


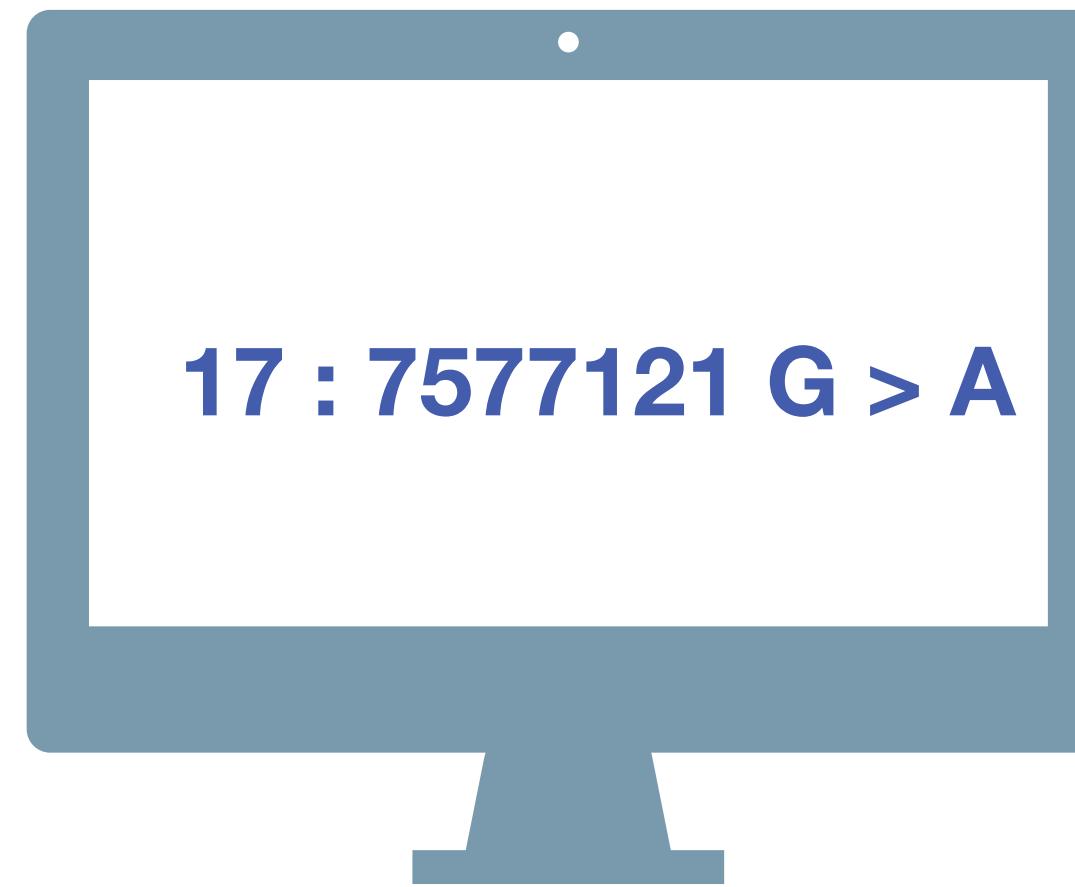
GENOMICS

A federated ecosystem for sharing genomic, clinical data

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

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





Beacon

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

YES | NO | \0

Beacon v2 Requests

POSTing Queries

- Beacon v2 supports a mix of dedicated endpoints with REST paths
- POST requests using JSON query documents
- final syntax for core parameters still in testing stages

```
{  
  "$schema": "beaconRequestBody.json",  
  "meta": {  
    "apiVersion": "2.0",  
    "requestedSchemas": [  
      {  
        "entityType": "individual",  
        "schema": "https://progenetix.org/services/schemas/Phenopacket/"  
      }  
    ],  
    "query": {  
      "requestParameters": {  
        "datasets": {  
          "datasetIds": ["progenetix"]  
        }  
      },  
      "filterLogic": "OR"  
    },  
    "pagination": {  
      "skip": 0,  
      "limit": 10  
    },  
    "filters": [  
      { "id": "NCIT:C4536" },  
      { "id": "NCIT:C95597" },  
      { "id": "NCIT:C7712" }  
    ]  
  }  
}
```



Progenetix & Beacon v1->2

Handover elements in Beacon responses

- Progenetix utilizes handovers to deliver data matched by the Beacon queries
- These handovers are interpreted by the front end to populate different parts of the UI, w/o the need of active selection
- Handovers are either standard Beacon v2 paths or dedicated custom functions

```
"results_handovers": [
  {
    "description": "create a CNV histogram from matched callsets",
    "handoverType": {"id": "pgx:handover:cnvhistogram", "label": "CNV Histogram"},
    "url": "https://progenetix.org/cgi-bin/PGX/cgi/samplePlots.cgi?method=cnvhistogram&accessid=aff0f73f-6dbf-45e5-91ba-04f19e3621bb"
  },
  {
    "description": "retrieve data of the biosamples matched by the query",
    "handoverType": {"id": "pgx:handover:biosamples", "label": "Biosamples"},
    "url": "https://progenetix.org/beacon/biosamples/?accessid=61b68a59-2160-41e4-a17d-0cf128841a57"
  },
  {
    "description": "retrieve variants matched by the query",
    "handoverType": {"id": "pgx:handover:variants", "label": "Found Variants (.json)" },
    "url": "https://progenetix.org/beacon/variants/?method=variants&accessid=5cced529-3acf-4156-b121-6ae7e5e63d0c"
  },
  {
    "description": "Download all variants of matched samples - potentially huge dataset...",
    "handoverType": {"id": "pgx:handover:callsetsvariants", "label": "All Sample Variants (.json)" },
    "url": "https://progenetix.org/beacon/variants/?method=callsetsvariants&accessid=61b68a59-2160-41e4-a17d-0cf128841a57"
  },
  {
    "description": "map variants matched by the query to the UCSC browser",
    "handoverType": {"id": "pgx:handover:bedfile2ucsc", "label": "Show Variants in UCSC" },
    "url": "http://genome.ucsc.edu/cgi-bin/hgTracks?org=human&db=hg38&position=chr9:21531306-22492891&hgt.customText=https://progenetix.org/tmp/5cced529-3acf-4156-b121-6ae7e5e63d0c.bed"
  }
]
```

ELIXIR Beacon Network



- developed under lead from ELIXIR Finland
- **authenticated access** w/ ELIXIR AAI
- **incremental extension**, starting with ELIXIR Beacon resources adhering to the **latest specification** (contrast to legacy networks)
- service details provided by individual Beacons, using **GA4GH service-info**
- **registration service**
 - integrator** throughout ELIXIR Human Data
 - starting point for "**beyond ELIXIR**" **feature rich** federated Beacon services

GRCh38 ▾ 17 : 7577121 G > A

[Example variant query](#) [Advanced Search](#)

baudisgroup at UZH and SIB
Progenetix Cancer Genomics Beacon+

Beacon+ provides a forward looking implementation of the Beacon API, with focus on structural variants and metadata based on the cancer and reference genome profiling data represented in the Progenetix oncogenomic data resource (<https://progenetix.org>).

[Visit Us](#) · [Beacon API](#) · [Contact Us](#)

National Bioinformatics Infrastructure Sweden
SweFreq Beacon

Beacon API Web Server based on the GA4GH Beacon API

[Visit Us](#) · [Beacon API](#) · [Contact Us](#)

LCSB at University of Luxembourg
ELIXIR.LU Beacon

ELIXIR.LU Beacon

[Visit Us](#) · [Beacon API](#) · [Contact Us](#)

Research Programme on Biomedical Informatics
DisGeNET Beacon

Variant-Disease associations collected from curated resources and the literature

[Visit Us](#) · [Beacon API](#) · [Contact Us](#)

European Genome-Phenome Archive (EGA)
EGA Beacon

This [Beacon](https://beacon-project.io/) is based on the GA4GH Beacon [v1.1.0](https://github.com/ga4gh/beacon/specification/blob/develop/beacon.yaml)

[Visit Us](#) · [Beacon API](#) · [Contact Us](#)

University of Tartu Institute of Genomics, Estonia
Beacon at the University of Tartu, Estonia

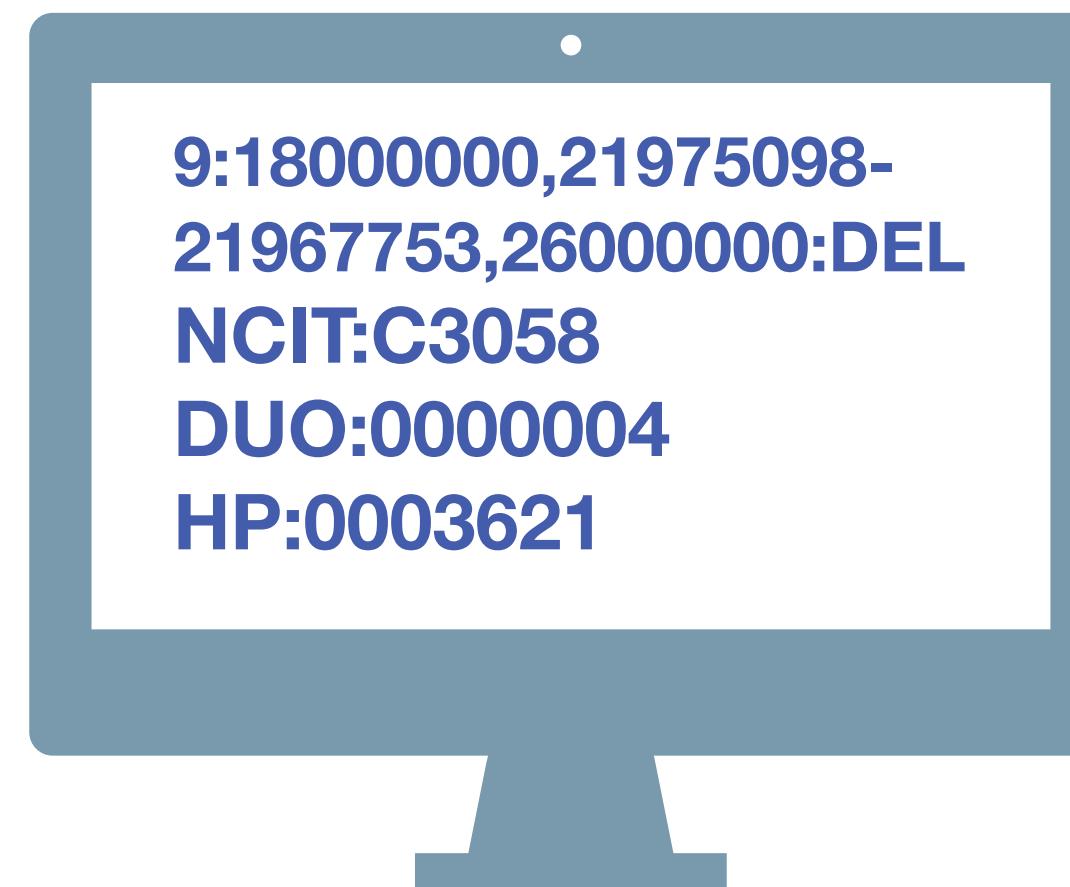
Beacon API Web Server based on the GA4GH Beacon API

[Visit Us](#) · [Beacon API](#) · [Contact Us](#)

CSC - IT Center for Science Production Beacon

Beacon API Web Server based on the GA4GH Beacon API

[Visit Us](#) · [Beacon API](#) · [Contact Us](#)

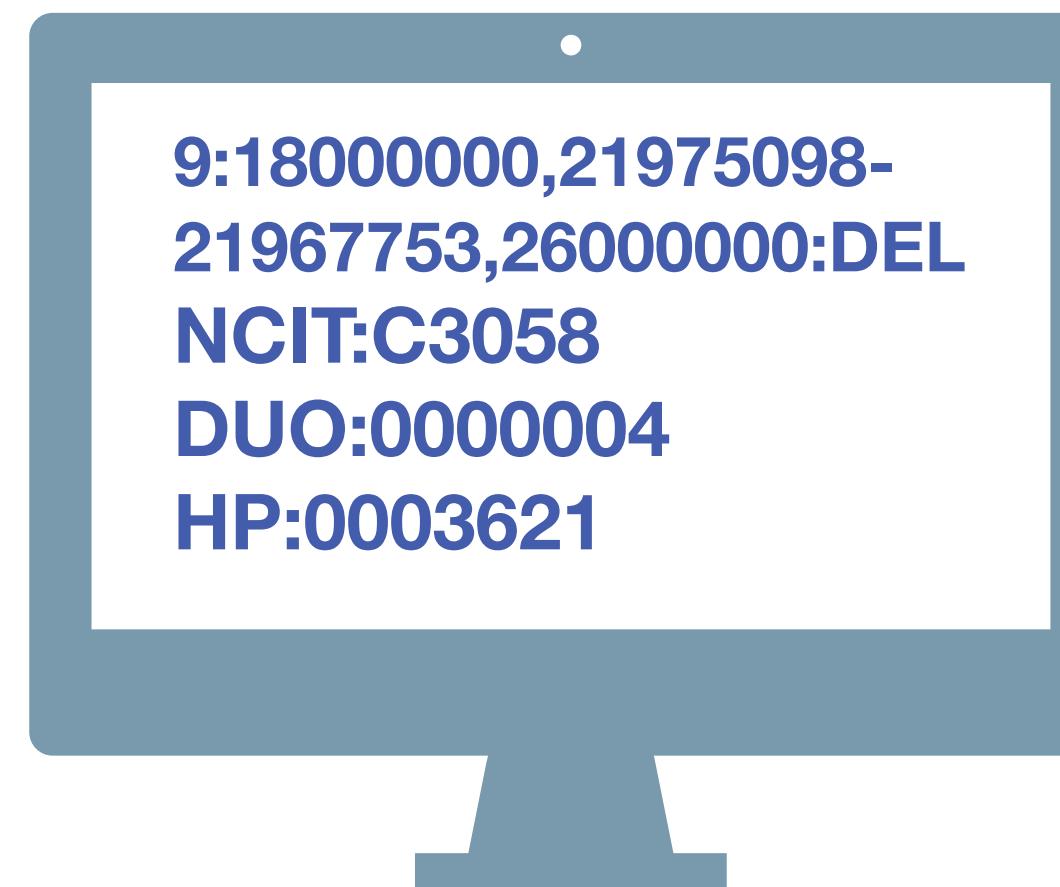


Have you seen deletions in this region on chromosome 9 in Glioblastomas from a juvenile patient, in a dataset with unrestricted access?

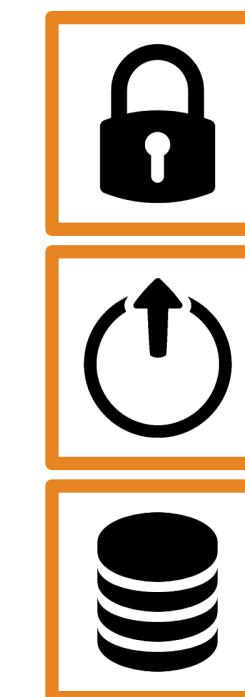


Beacon v2 API

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



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".





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

