

# **BIO390: Introduction to Bioinformatics**

## **Lecture I: What is Bioinformatics?**

**Michael Baudis | 2020-09-15**

# 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 15th)
- Organizer:

Prof. Dr. Michael Baudis

Department of Molecular Life Sciences (IMLS)

University of Zurich Campus Irchel, Y-13F-01

CH-8057 Zurich

email [michael@baud.is](mailto:michael@baud.is)

web [info.baudisgroup.org](http://info.baudisgroup.org)

**Please use website & OLAT for**

- **ZOOM link**
- **"on site" sign-up**
- **additional course information**

**<https://comppbiozurich.org/UZH-BIO390/>**



## UZH BIO390

Introduction to Bioinformatics

### News and Updates

Lectures

Teachers

### Examples, Guides & FAQ

#### Related Sites

CompbioZurich

UZH392 course

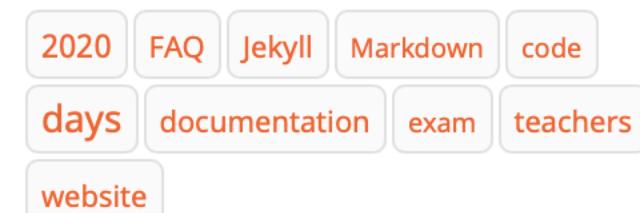
Baudisgroup at UZH

#### Github Projects

compbiozurich

progenetix

#### Tags



# UZH BIO390 - Introduction to Bioinformatics Lecture Series

This is a repository for materials related to the BIO390 *Introduction to Bioinformatics* lecture series at the University of Zürich.

### Time & Place

- 1 x 2h / week
- Tue 08:00-09:45
- UZH Irchel campus, NEW Y15-G-60
- OLAT [On Site Reservation System](#)
- ZOOM live stream (links posted in OLAT)

### Course Language

- English

### 2020 COVID19 Regulations - On Site & Remote

Due to the limits imposed on on-site access, an “[On Site Reservation System](#)” has been established. It will allow a (rotating) set of students to attend the lectures on Irchel. With the current numbers of registered students we expect to be able to host ca. 50% per lecture on site.

Additional to this, remote options will be provided:

- lecture slides and tasks will be posted by the lecturers and linked from the daily pages and/or deposited in OLAT
- podcasts will be made available (with delay)
- optional/possibly: live streaming (ZOOM); please see OLAT

However, while those remote options will provide an alternative we still encourage in person attendance while observing the [university's safety concept](#) and regulations.

### Programme and day-by-day Schedule

#### 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.

#### Learning Goals

The overall learning goals - especially the (limited) set necessary for passing the test - will be updated throughout the semester.

- Core [Learning Goals](#)

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



## UZH BIO390

Introduction to Bioinformatics

### News and Updates

BIO390 HS20 Programme  
BIO390 HS19 Final Programme

### Lectures

#### Teachers

#### Examples, Guides & FAQ

### Related Sites

CompbioZurich  
UZH392 course  
Baudisgroup at UZH

### Github Projects

compbiozurich  
progenetix

### Tags

2020 FAQ Jekyll Markdown code  
days documentation exam teachers  
website

## BIO390 HS20 Programme

In the following you will find the program for the 2020 Autumn lectures in the "Introduction to Bioinformatics" series at the University of Zurich.

For on-site access information or live streaming please see the relevant information in UZH's "OLAT" platform (registered participants only):

- OLAT On Site Reservation System
- ZOOM live stream (links posted in OLAT)

Please also see the [individual pages](#) lecture pages for more information.

### Individual Lectures

2020-09-15

#### What is Bioinformatics? Introduction and Resources

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.

@mbaudis 2020-09-15: [more ...](#)

2020-09-22

#### Biological Sequence Informatics

Christian von Mering

2020-09-22: [more ...](#)

2020-09-29

#### Statistical Bioinformatics

Mark Robinson

2020-09-29: [more ...](#)

2020-10-06

#### Text Mining

Patrick Ruch (University of Geneva)

2020-10-06: [more ...](#)

2020-10-13

#### Proteomics

Katja Baerenfaller

In proteomics one of the important bioinformatics tasks is to generate lists of reliably identified peptides and proteins in mass spectrometry-based experiments. For this, amino acid sequences are assigned to measured tandem mass spectra. The quality of the peptide spectrum assignments are scored and criteria are applied that allow to distinguish the good from the bad hits and to estimate the quality of the dataset.

2020-10-13: [more ...](#)

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





[date ↓] [date ↑] [Z → A]

## Pages tagged "teachers"

### Abdullah Kahraman, PhD

- Clinical Bioinformatics
- USZ, Institute for Pathology and Molecular Pathology

2019-12-10: [more ...](#)



### Andreas Wagner

- Professor and Chairman, Dept. of Evolutionary Biology and Environmental Studies
- University of Zurich

2006-01-01: [more ...](#)



### Christian von Mering

- Professor of Statistical Genomics
- Institute of Molecular Life Sciences
- IMLS Director of the Institute
- University of Zurich

2007-04-01: [more ...](#)



### Izaskun Mallona Gonzalez, PhD

- Postdoctoral research scientist, Department of Molecular Mechanisms of Disease, University of Zürich ↗

2019-08-28: [more ...](#)



### Katja Baerenfaller, PhD

- Group leader, Swiss Institute of Allergy and Asthma Research
- Swiss Institute of Bioinformatics SIB

2019-12-10: [more ...](#)



### Mark Robinson

- Associate Professor of Statistical Genomics
- Institute of Molecular Life Sciences
- University of Zurich

2011-09-01: [more ...](#)



### Michael Baudis

- Professor of Bioinformatics, Institute of Molecular Life Sciences, University of Zurich
- Swiss Institute of Bioinformatics SIB

2007-08-01: [more ...](#)



### Valerie Barbie, PhD

- Head of Clinical Bioinformatics, Swiss Institute of Bioinformatics SIB
- Swiss Variant Interpretation Platform SVIP

2019-12-10: [more ...](#)



## UZH BIO390

Introduction to Bioinformatics

### News and Updates

### Lectures

### Teachers

### Examples, Guides & FAQ

### Related Sites

CompbioZurich

UZH392 course

Baudisgroup at UZH

### Github Projects

compbiozurich

progenetix

### Tags

- 2020
- FAQ
- Jekyll
- Markdown
- code
- days
- documentation
- exam
- teachers
- website



## UZH BIO390

Introduction to Bioinformatics

[News and Updates](#)

[General Info](#)

[Lectures](#)

[Teachers](#)

[Examples, Guides & FAQ](#)

[Related Sites](#)

[CompbioZurich](#)

[UZH392 course](#)

[Baudisgroup at UZH](#)

[Github Projects](#)

[compbiozurich](#)

[progenetix](#)

[Tags](#)

[FAQ](#)

[Jekyll](#)

[Markdown](#)

[code](#)

[days](#)

[documentation](#)

[exam](#)

[teachers](#)

[website](#)

## UZH BIO390 - Learning Goals

This page indicates some of the learning goals, as emphasised by the different lecturers. Some points will have been discussed in different lectures; accordingly, exam questions may not refer to information of one specific presentation.

### Bioinformatics: Definition & Concepts

- definition of "Bioinformatics" (cf. Anna Tramontano)
- categories of informatics tools used in bioinformatics
- hypothesis versus data driven science
- areas of bioinformatics/bioinformaticians, in contrast to ("pure" modelling, statistics etc.)
- 3 main categories of biological data, and example resources
- definition of API
- common sequence related file formats
- hierarchies and relationships as 2 main principles of ontologies
- areas of "not-bioinformatics", and why

### Sequence Analysis

- substitution matrices
- BLAST

### Statistical Bioinformatics

- statistical evidence for a change in the means
- usage of gene expression profiling
- dimensionality reduction
- central limit theorem
- multiple testing correction
- parameters for hierarchical clustering

### Bioinformatics tools: Statistics & Graphics in R & BioConductor

- What is tidy data?
- ideas behind ggplot: components of a ggplot, arrangement of input data ... (no actual code writing needed)
- interpret common types of plots, e.g. barplot, boxplot, histogram
- effect of data transformation (e.g. log) on common types of plots

### Regulatory Genomics and Epigenomics

- secondary/tertiary human genome structure
- functional genome content
- transcription factors & genome interaction
- chemical genome modifications, their effectors and results
- ChIP-Seq



University of  
Zurich



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

# 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)



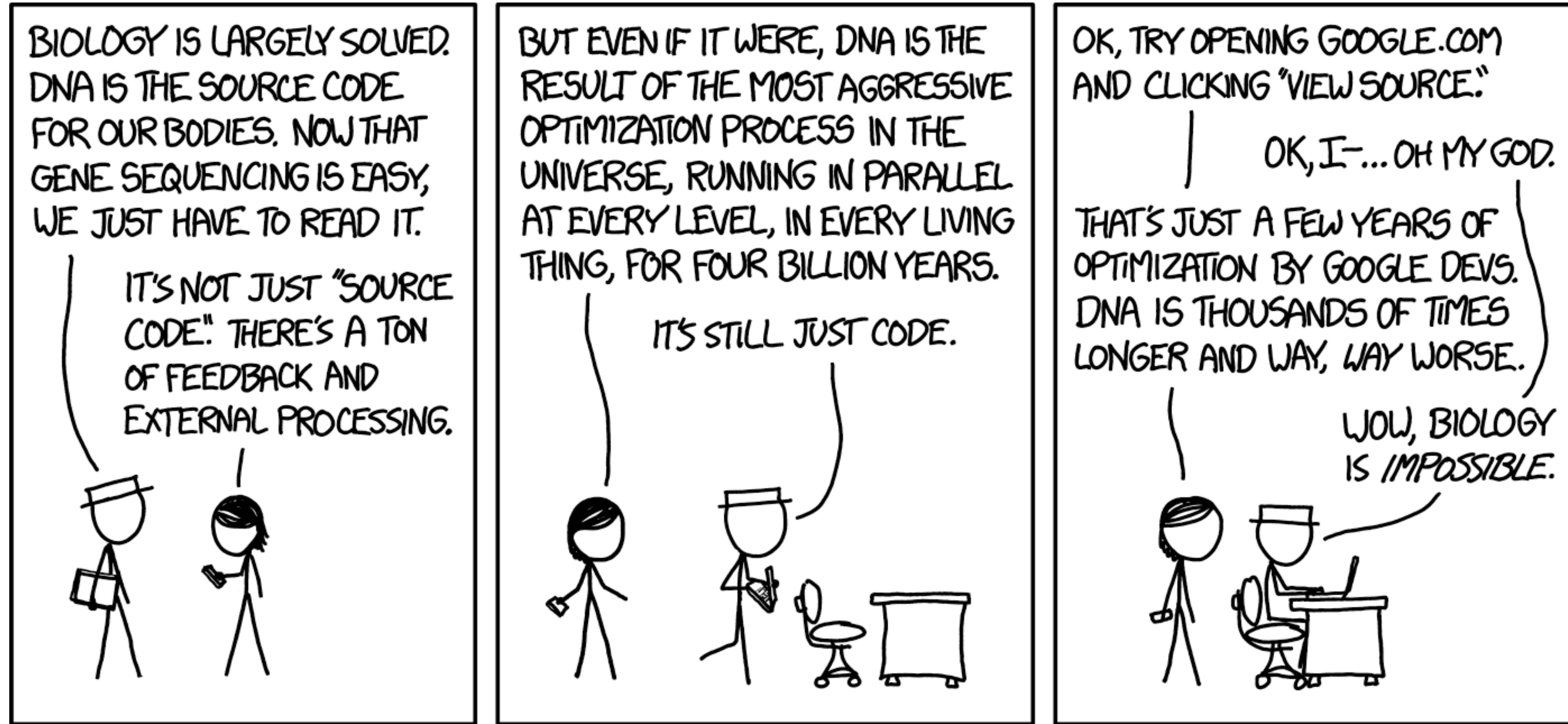
# BIO390: Course Schedule

- 2020-09-12: Michael Baudis - What is Bioinformatics? Introduction and Resources
- 2020-09-22: Christian von Mering - Sequence Bioinformatics
- 2020-09-29: Mark Robinson - Statistical Bioinformatics
- 2020-10-06: Patrick Ruch (UniGe) - Text Mining
- 2020-10-13: Katja Baerenfaller (SIAF) - Proteomics
- 2020-10-20: Shinichi Sunagawa (ETHZ) - Metagenomics
- 2020-10-27: Valentina Boeva (ETHZ) - Machine Learning for Biological Use Cases
- 2020-11-03: Izaskun Mallona - Regulatory Genomics and Epigenomics
- 2020-11-10: Andreas Wagner - Biological Networks
- 2020-11-17: Abdullah Kahraman (USZ) - Molecular Interaction Networks
- 2020-11-24: Valérie Barbie (SIB) - Clinical Bioinformatics
- 2020-12-01: Michael Baudis - Building a Genomics Resource
- 2020-12-08: Michael Baudis - Genome Data & Privacy
- 2020-12-15: 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

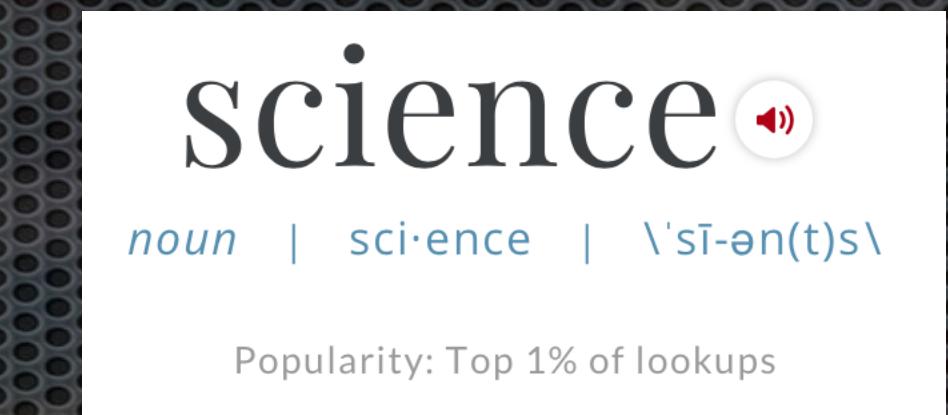
# Biology is *impossibly* complex - But bioinformatics might help



# 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



# 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?

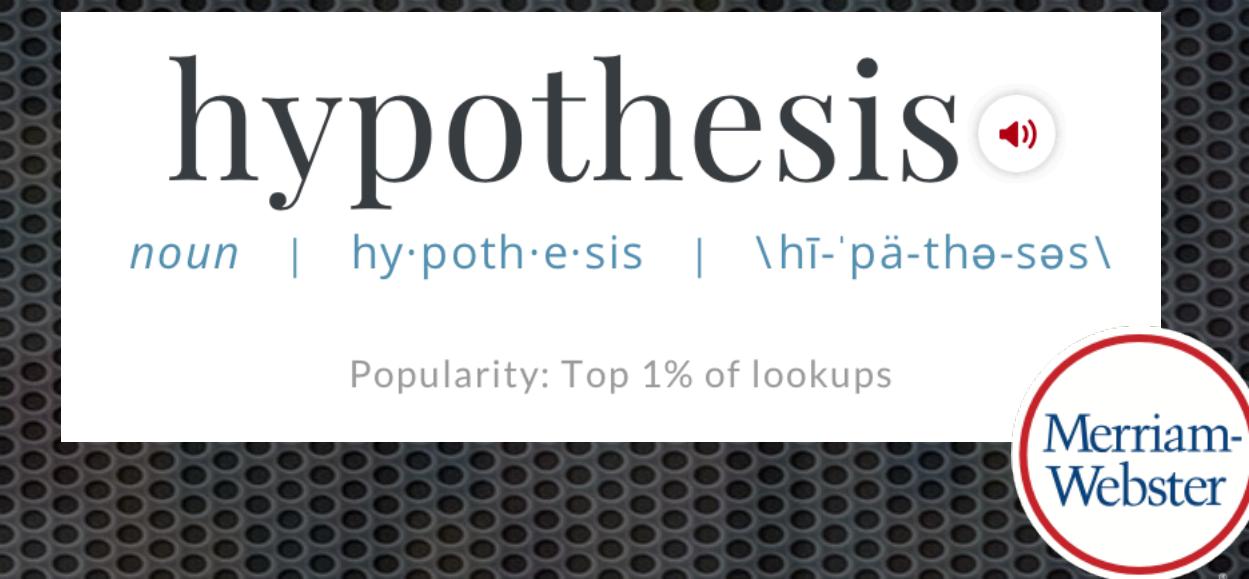


Bioinformatics **analyzes**

- 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

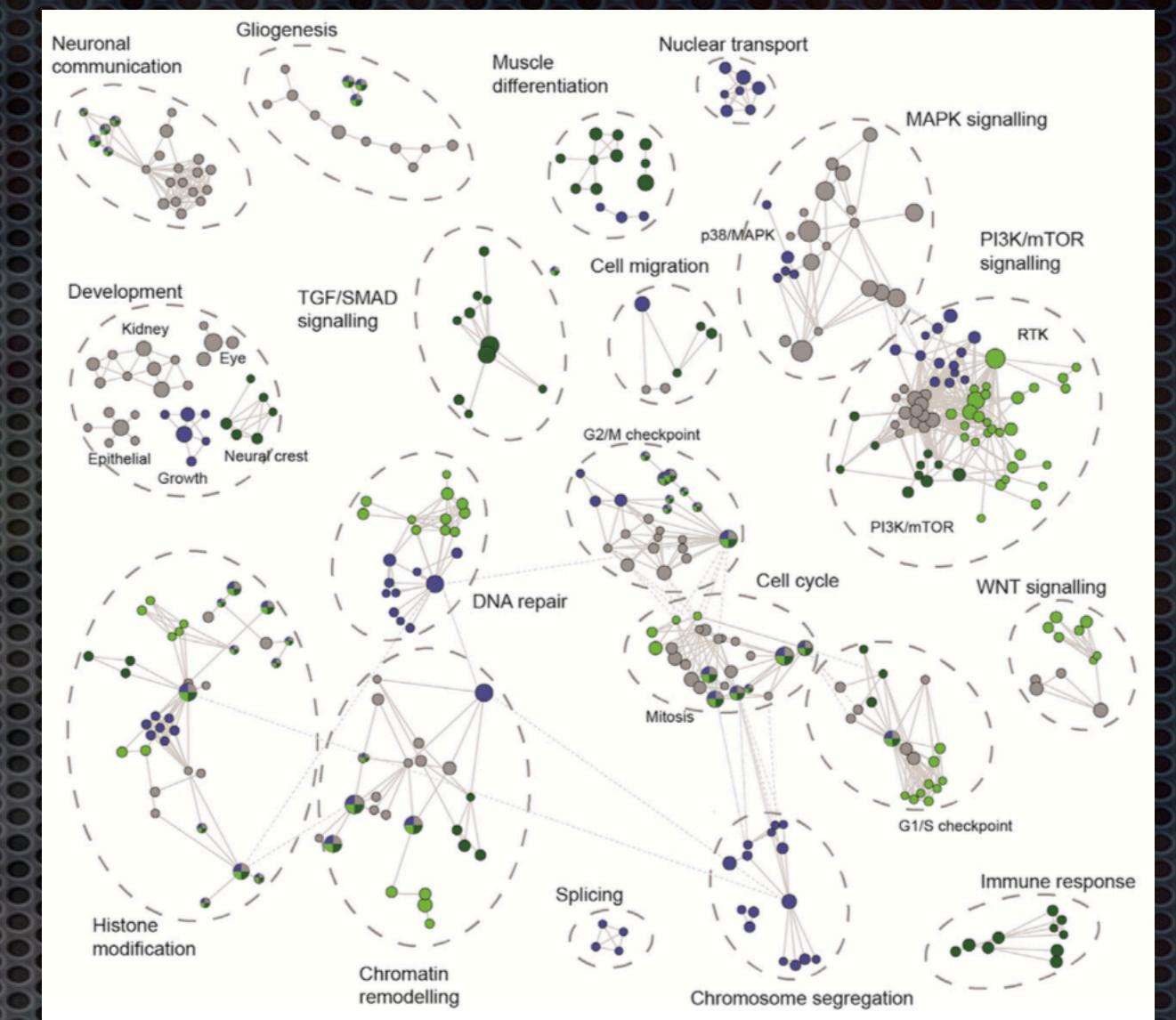
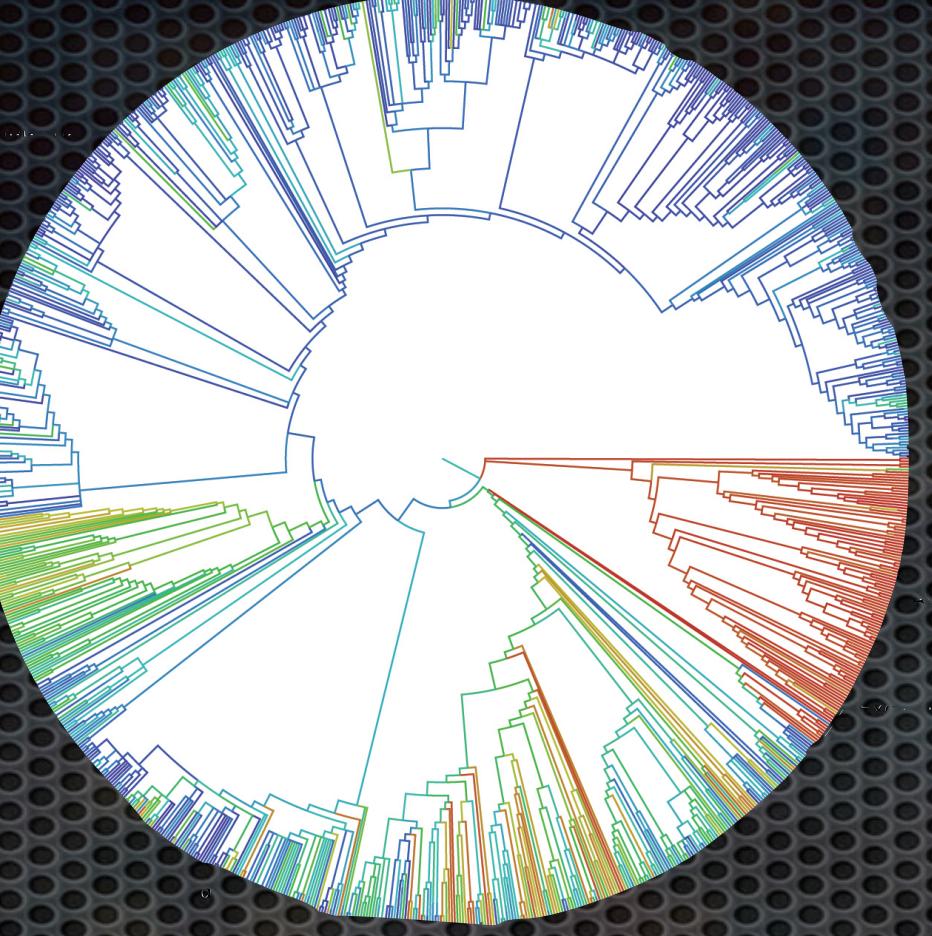


- Bioinformatics is "the science that uses the instruments of informatics to analyze biological data in order to **formulate hypotheses** about life." (Anna Tramontano)

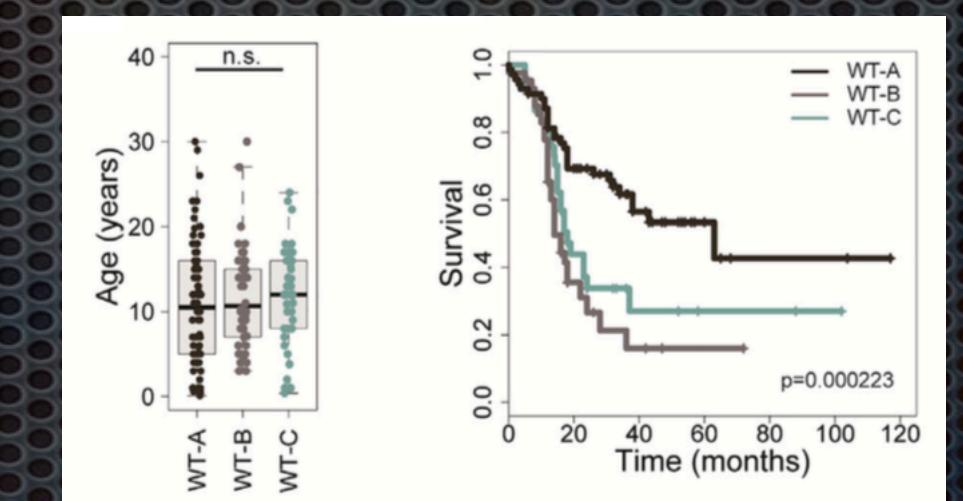
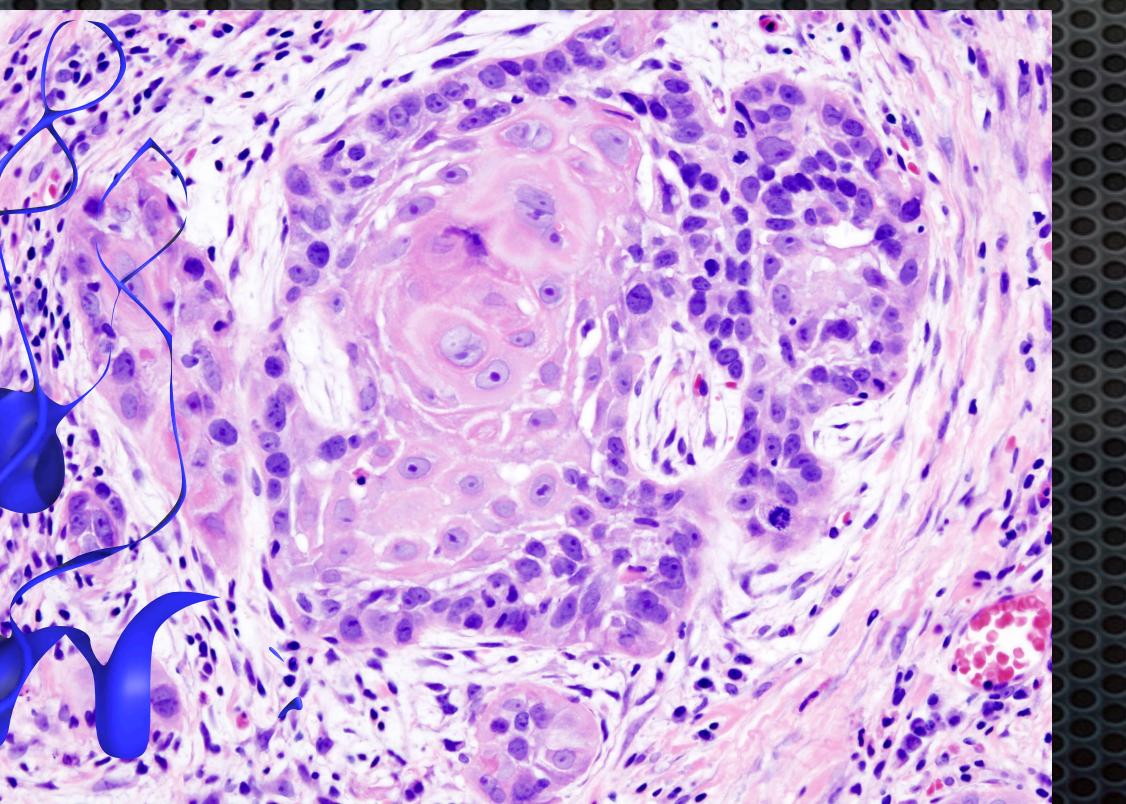
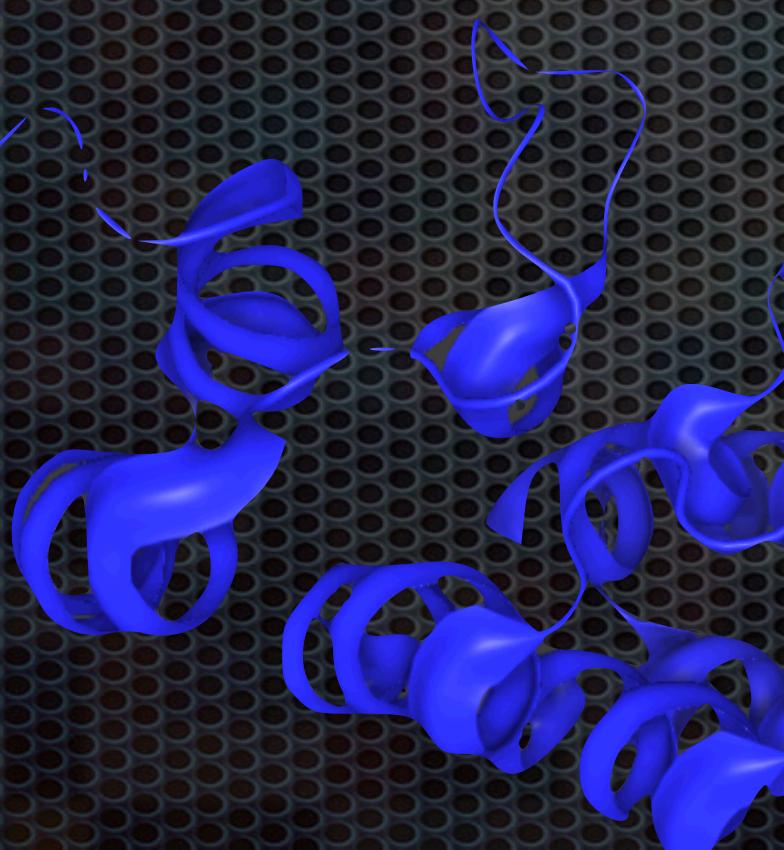
b : an interpretation of a practical situation or condition taken as the ground for action



# 42



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

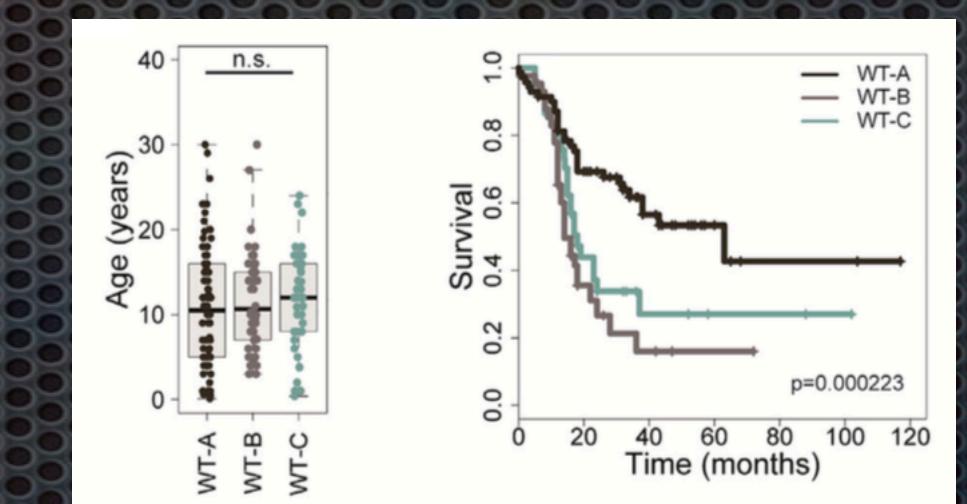
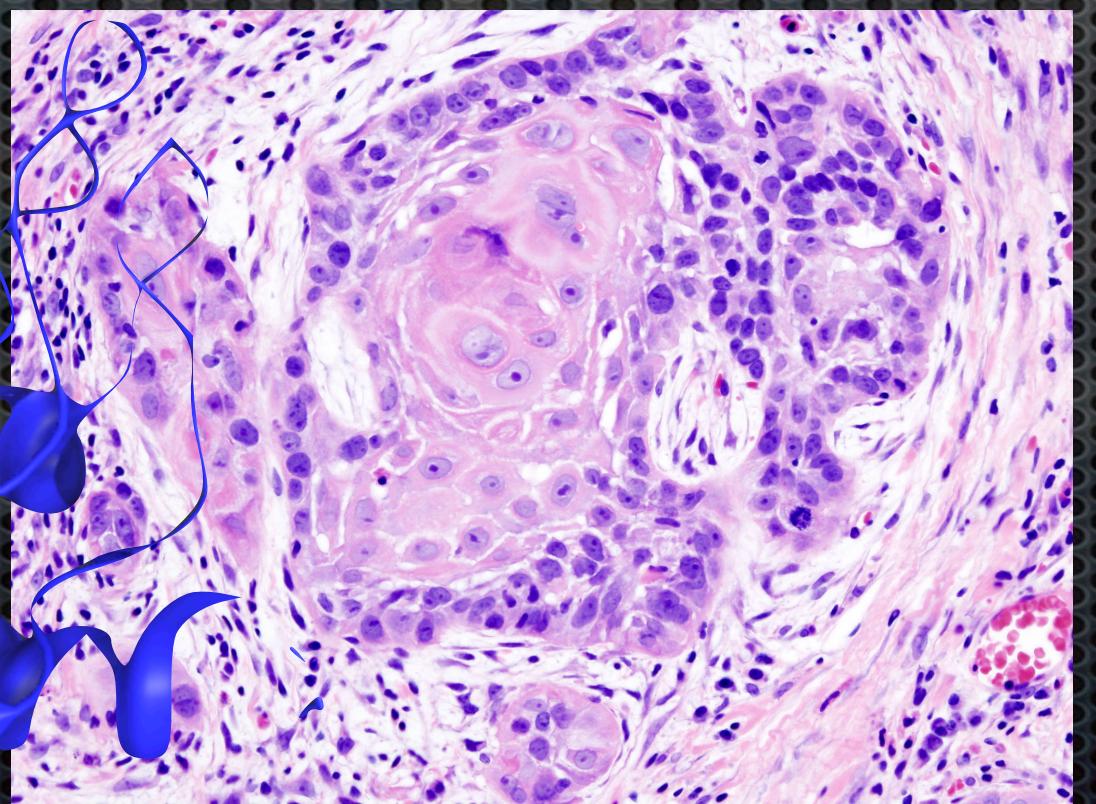
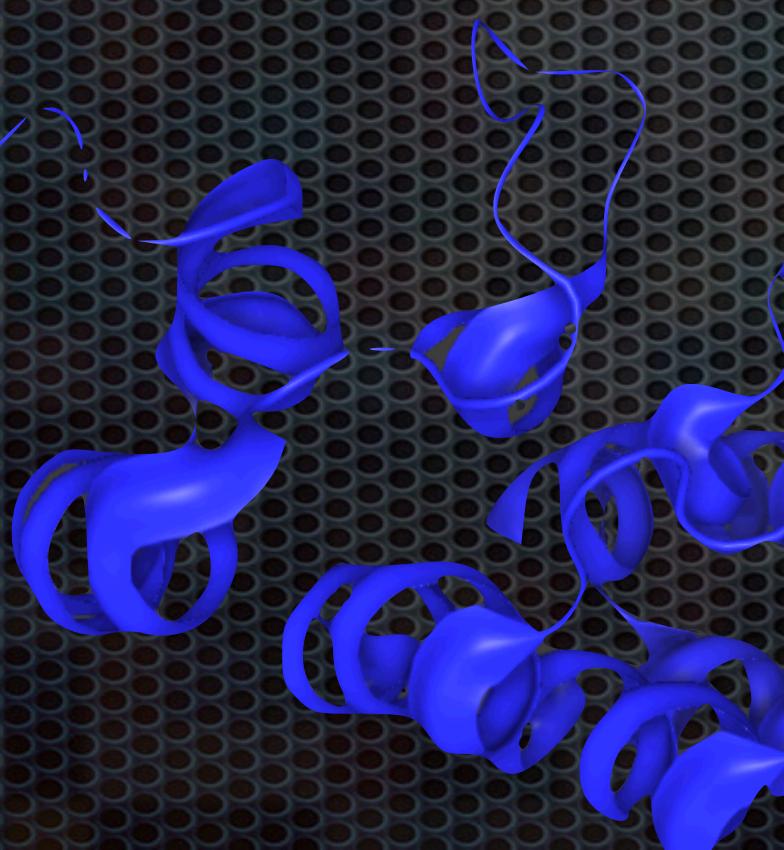


Sources: nextprot | opentreeoflife | wikipedia | MacKay et al., Cancer Cell (2017) | original photos

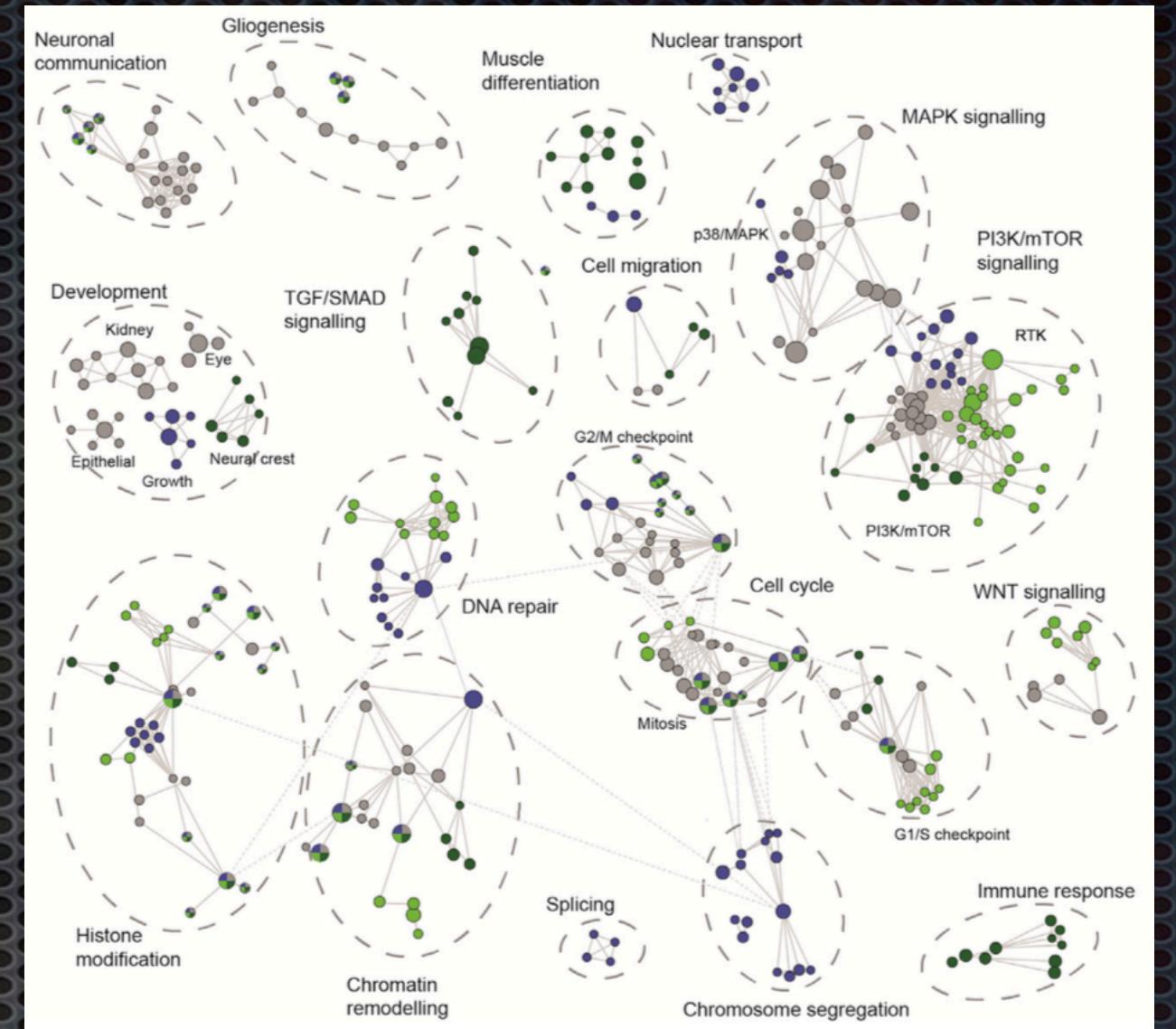
# Homework Assignment - Why could Bioinformatics be considered the **42** of Life Sciences?

**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



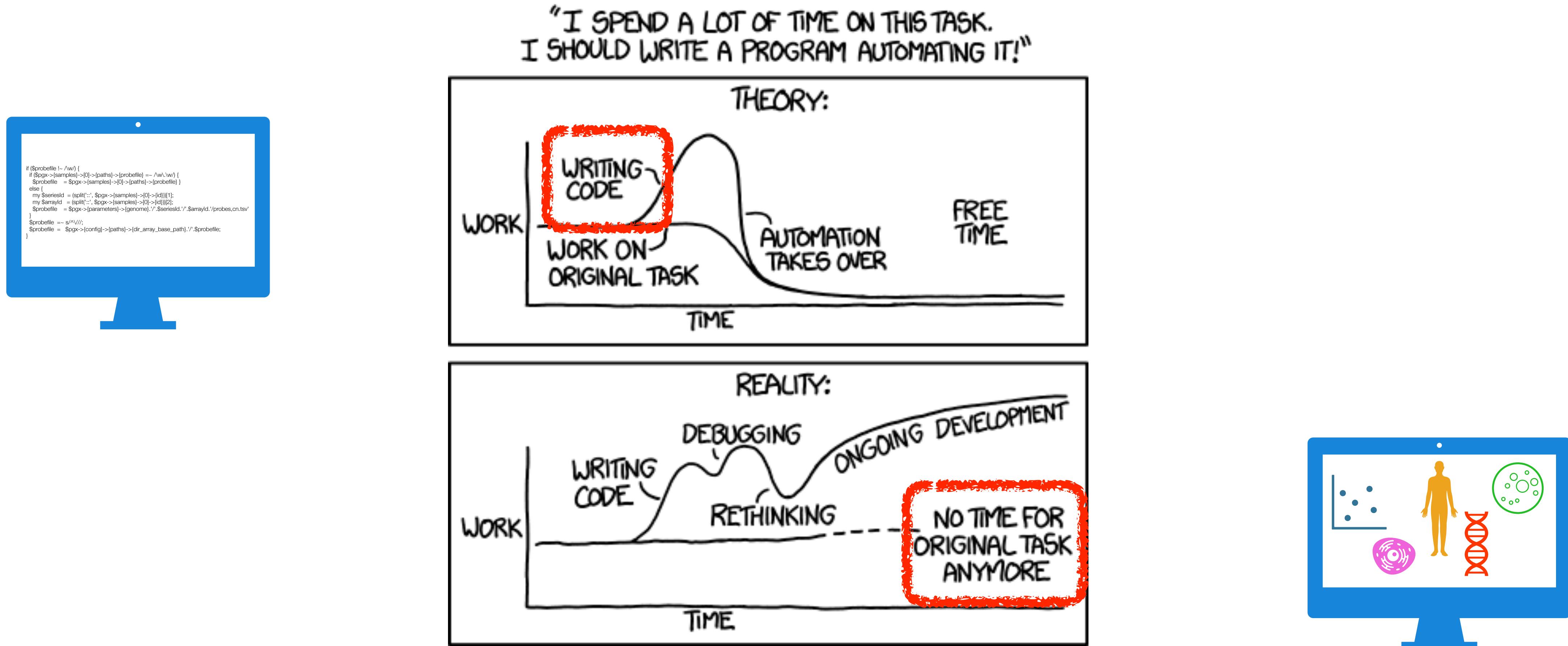


# {bio\_informatics\_science}

---



# {bio\_informatics\_science}



# Who is a Bioinformatician?

## 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 biologic or medical background to understand problems presented and identify pitfalls and hidden biases arising from data generation methods
- developer of informatics tools
- may get rich

# Who is a Bioinformatician?

## 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 biologic or medical background to understand problems presented and identify pitfalls and hidden biases arising from data generation methods
- developer of informatics tools
- may get rich

flux

# 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

# What do Bioinformaticians work on?

- 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
- ...

# Bioinformatics: Data **Categories** & **Databases**

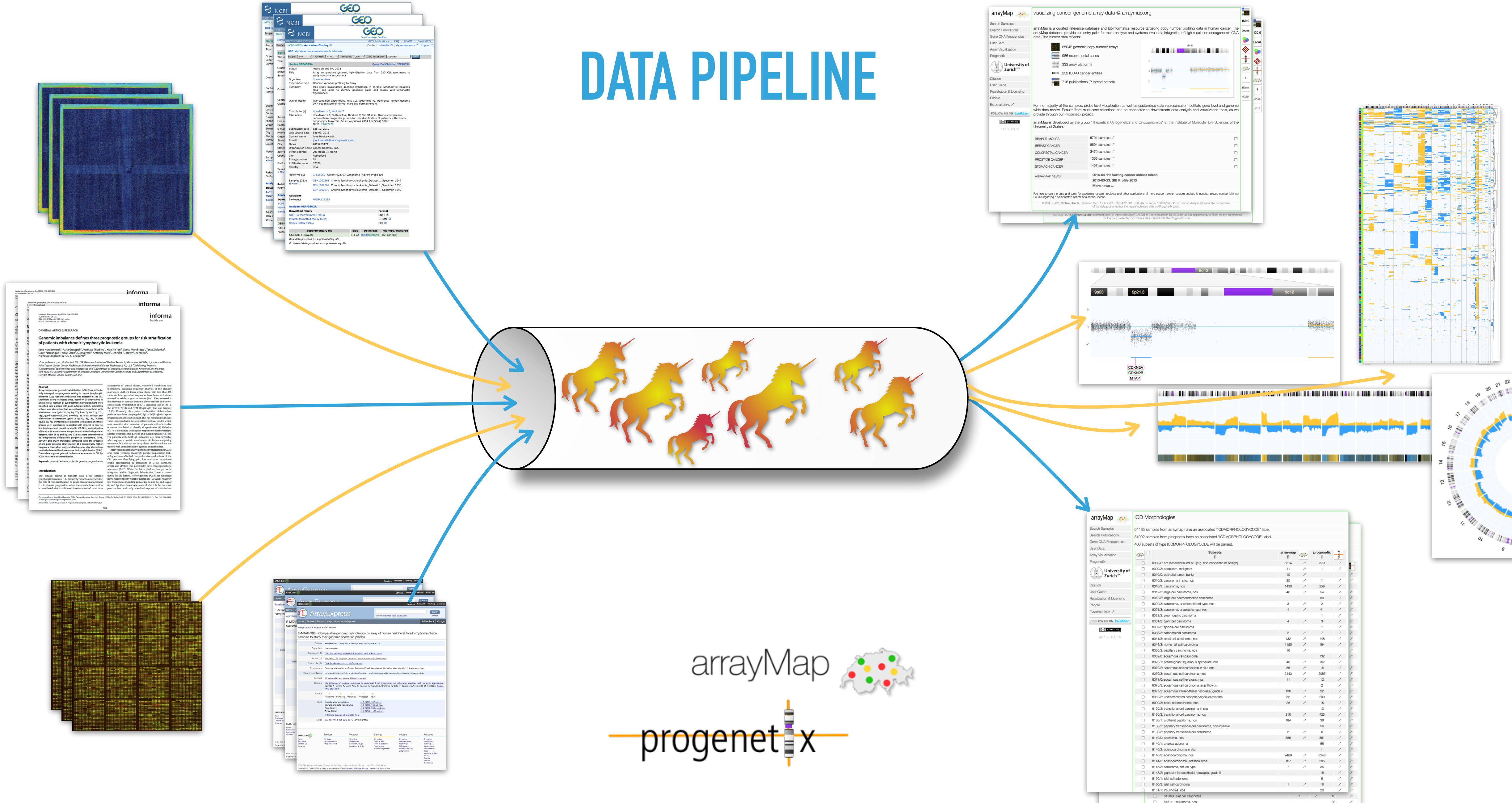
- biological data comes in **3 main categories**:
  - **sequence** data (nucleic acids, aminoacids)
  - **structural** data (DNA, RNA, proteins; intracellular organisation, tissues ...)
  - **functional** data (interactions in time and space)
- data storage & retrieval: importance of local and connected **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** - information resources providing agglomerated & **curated** data derived from primary sources (e.g. UniprotKB, nextProt, String, KEGG, arrayMap...)



SRA



# DATA PIPELINE



# DATA PIPELINE

## BIOCURATION BIOINFORMATICS



NCBI GEO Accession Display

Series GSE640034 Public on Sep 07, 2013

Organism: Human

Experiment type: Genomic variation profiling by array

Summary: This study investigates genomic variation in chronic lymphocytic leukemia (CLL) specimens with prognostic significance.

Overall design: Overall design experiment, Test vs. Specimens vs. Reference human genome

Contributor(s): Houldsworth J, Venkata T, Guttagji A, Thoduri V, Yan XI et al.

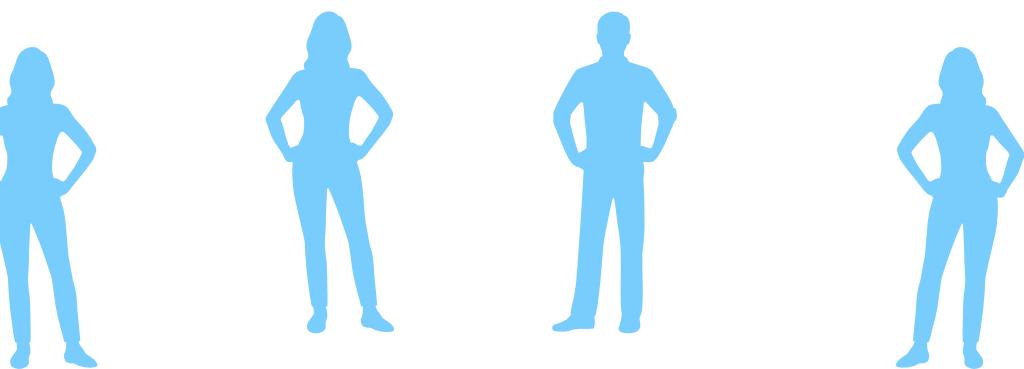
Phone: +41 61 267 32 32

Address: University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

Sample ID: GSE640034

Platform: Agilent G1317P Lymphoma (Agilent Probe ID)

Supplementary file: GSE64034.RAW.tar



arrayMap

985 experimental series

333 array platforms

253 ICD-O cancer entities

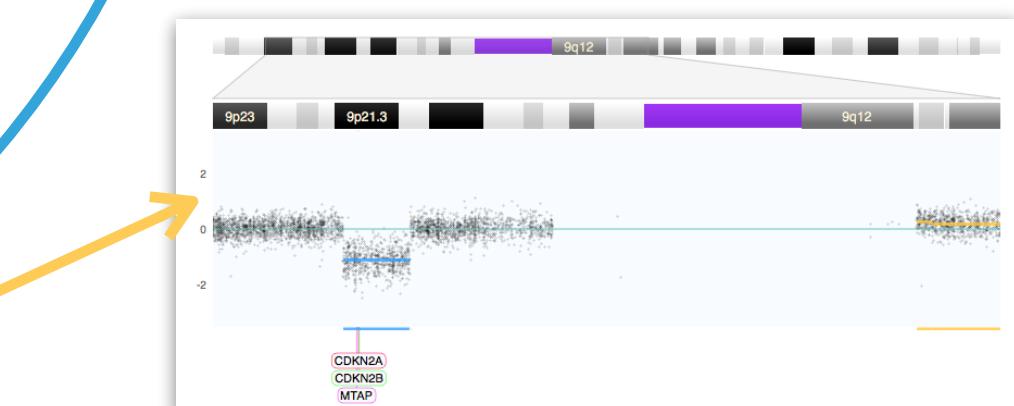
716 publications (PubMed entries)

For the majority of the samples, probe level visualization as well as customized data representation facilitate gene level and genome wide data review. Results from multi-case selections can be connected to downstream data analysis and visualization tools, as we provide through our Progenetix project.

arrayMap is developed by the group "Theoretical Cytogenetics and Oncogenomics" at the Institute of Molecular Life Sciences of the University of Zurich.

Platforms (1): GPR100, Agilent G1317P Lymphoma (Agilent Probe ID)

Supplementary file: GSE64034.RAW.tar



informa healthcare

ORIGINAL ARTICLE RESEARCH

Genomic imbalance defines three prognostic groups for risk stratification of patients with chronic lymphocytic leukemia

Jane Houldsworth<sup>1</sup>, Asha Guttapalli<sup>1</sup>, Venkata Thoduri<sup>1</sup>, Xiao Jie Yan<sup>1</sup>, Geeta Mendiratta<sup>1</sup>, Tamja Zelenka<sup>2</sup>, Gouri Nangisetty<sup>3</sup>, Wei Chen<sup>3</sup>, Supratik Pati<sup>3</sup>, Anthony Mato<sup>3</sup>, Jennifer R. Brown<sup>3</sup>, Kanti Rai<sup>4</sup>

<sup>1</sup>Cancer Genetics, Inc., Rutherford, NJ, USA; <sup>2</sup>Weinstein Institute of Medical Research, Manhattan, NY, USA; <sup>3</sup>Lymphoma Division, Department of Hematology and Oncology, Department of Medicine, Division of Hematology/Oncology, Department of Epidemiology and Biostatistics, Department of Medicine, Memorial Sloan-Kettering Cancer Center, New York, NY, USA; <sup>4</sup>Department of Hematology/Oncology, Dana Farber Cancer Institute and Department of Medicine, Harvard Medical School, Boston, MA, USA

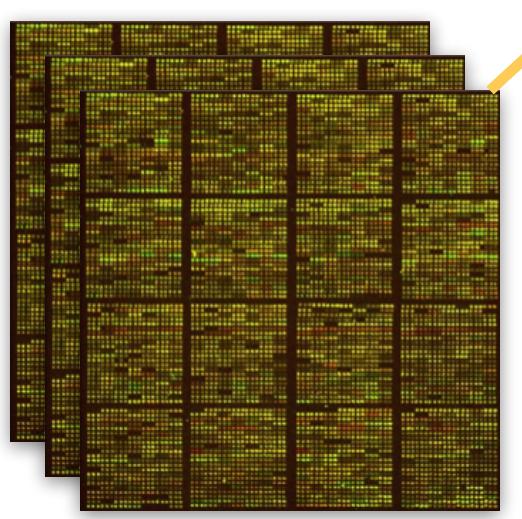
Abstract

Abstract: Several recent studies have demonstrated that genomic imbalance (GCI) has been fully leveraged in a prognostic setting in chronic lymphocytic leukemia (CLL). In this study, we used a targeted array-based genomic analysis to identify novel prognostic biomarkers in CLL. We analyzed 20 aberrations in 100 CLL specimens using a targeted array. Based on 20 aberrations in each specimen, we identified 100 specimens with a gain or loss of at least one of these 20 aberrations. These 100 specimens were then classified into a group with one or more gains (20R) exhibiting a gain of at least one of the 20 aberrations, a group with one or more losses (20L) exhibiting a loss of at least one of the 20 aberrations, and a group with no gains or losses (20N). The 20R group had a significantly higher first treatment and overall survival (P < 0.001), and validation of this finding was performed using an independent dataset. Gains of 9q and 17q loss of 13q were determined to be the most significant prognostic markers. We also found that patients with 9q gain and 17q loss had a significantly shorter median free period and overall survival (OS) (45 months vs. 70 months, P = 0.001). Patients requiring chemotherapy had a significantly shorter OS than those not treated with chemotherapy (45 months vs. 60 months, P = 0.001). NOVOTNY and SP18 mutations correlated with the presence of 9q gain and 17q loss. Our results demonstrate that GCI, when regarded as an allelic loss, is a prognostic marker in CLL. These data support genomic imbalance evaluation in CLL by using a targeted array-based approach.

Keywords: chronic leukemia, molecular genetics, prognostication

Introduction

The clinical course of patients with B-cell chronic lymphocytic leukemia (CLL) is highly variable, undergoing disease progression, remission, and relapse. When integrated into diagnostic laboratories, these are problematic. Comparative genomic hybridization (CGH) is a relatively new technique that allows for the detection of chromosomal gains and losses. CGH has been used to detect chromosomal imbalances in CLL specimens, but failed to classify all specimens [6]. Deletion of chromosomes 11q and 17q has been associated with shorter median free period and overall survival (OS) [4]. Chromosomal gains of 9q and 17q have been associated with prognosis in CLL [5]. Patients requiring chemotherapy have a significantly shorter OS than those not treated with chemotherapy (45 months vs. 60 months, P = 0.001). NOVOTNY and SP18 mutations correlated with the presence of 9q gain and 17q loss. Our results demonstrate that GCI, when regarded as an allelic loss, is a prognostic marker in CLL. These data support genomic imbalance evaluation in CLL by using a targeted array-based approach.



ArrayExpress

E-MTAB-998 Comparative genomic hybridization array of human peripheral T-cell lymphoma clinical samples to study their genomic aberration profiles

Organism: Homo sapiens

Sample: E-MTAB-998

Description: Genomic aberration profiles of Peripheral T-cell Lymphoma, not otherwise specified (clinical sample)

Experiment type: comparative genomic hybridization, array, in vitro, comparative genomic hybridization, disease state

Context: E-MTAB-998

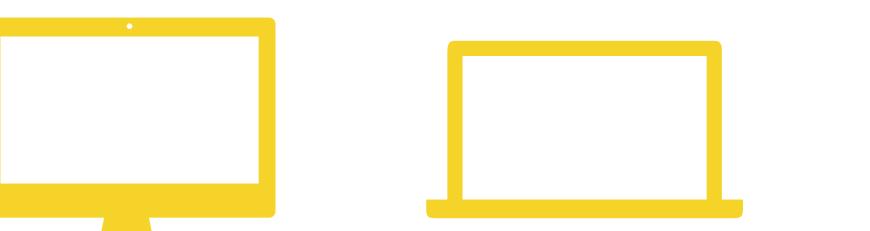
Platform: Agilent G1317P Human Oligo Microarray

Investigation description: By request of the investigator, this dataset is available under a license agreement. Please contact the investigator for further information.

Sample and data relationship: E-MTAB-998\_998.vcf.gz

Data analysis: E-MTAB-998.vcf.gz

Links: Send E-MTAB-998 data to GENOMEPAPE



arrayMap

progenetix

arrayMap

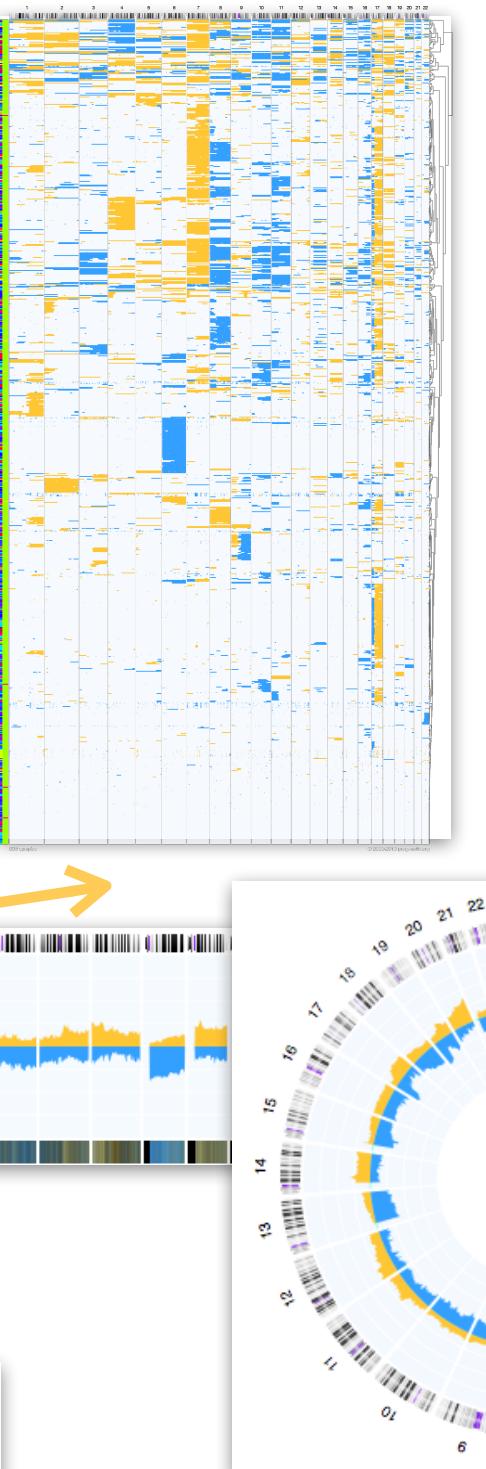
ICD Morphologies

64485 samples from arraymap have an associated "ICDMORPHOLOGYCODE" label.

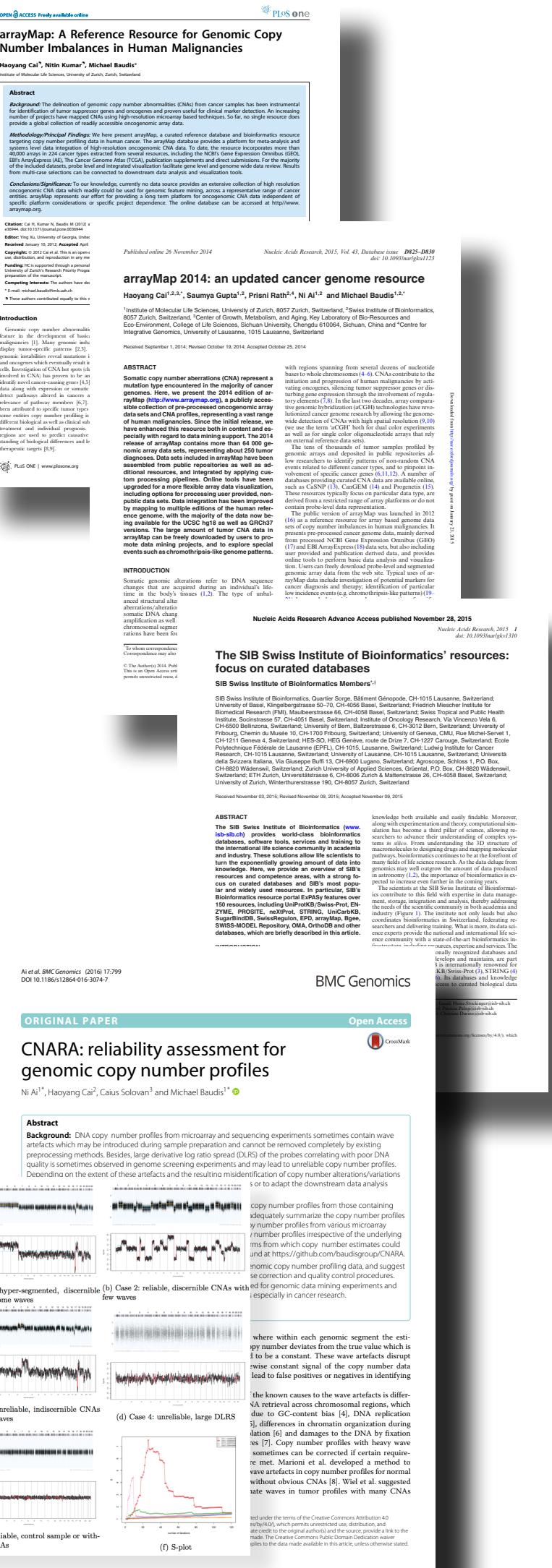
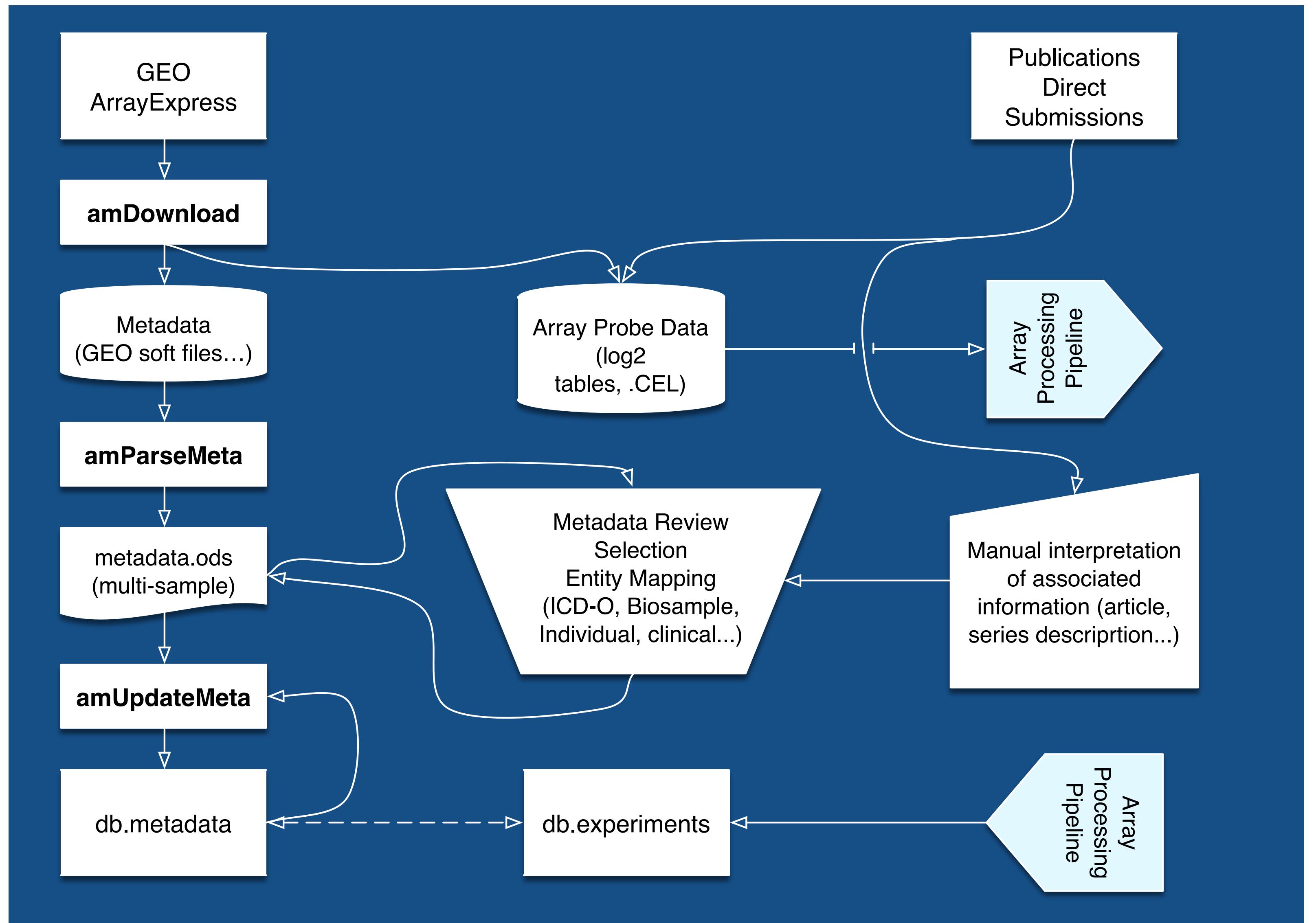
31922 samples from progenetix have an associated "ICDMORPHOLOGYCODE" label.

400 subsets of type ICDMORPHOLOGYCODE will be parsed.

| Subsets   | arrayMap      | progenetix |
|---|---------------|------------|
| 00000: not classified in icd-3 (e.g. non-neoplastic or benign)        | 8614 ↗ 370 ↗  |            |
| 00003: neoplasm, malignant  | 11 ↗ 1 ↗      |            |
| 00100: carcinoma, benign  | 10 ↗ 11 ↗     |            |
| 00102: carcinoma, in situ, nos  | 20 ↗ 258 ↗    |            |
| 00120: large cell carcinoma, nos                                      | 46 ↗ 54 ↗     |            |
| 00200: squamous cell carcinoma, nos                                   | 80 ↗ 60 ↗     |            |
| 00210: carcinoma, undifferentiated type, nos                          | 3 ↗ 2 ↗       |            |
| 00213: carcinoma, anaplastic type, nos                                | 4 ↗ 41 ↗      |            |
| 00220: giant cell carcinoma   | 1 ↗ 1 ↗       |            |
| 00303: giant cell carcinoma   | 4 ↗ 3 ↗       |            |
| 00333: adenocarcinoma, nos  | 1 ↗ 7 ↗       |            |
| 00413: small cell carcinoma, nos                                      | 132 ↗ 148 ↗   |            |
| 00500: basal cell carcinoma, nos                                      | 119 ↗ 184 ↗   |            |
| 00503: papillary carcinoma, nos                                       | 16 ↗ 16 ↗     |            |
| 00701: meningothelial squamous epithelium, nos                        | 46 ↗ 162 ↗    |            |
| 00702: squamous cell carcinoma, nos                                   | 65 ↗ 16 ↗     |            |
| 00703: squamous cell carcinoma, nos                                   | 2443 ↗ 2087 ↗ |            |
| 00704: squamous cell carcinoma, nos                                   | 11 ↗ 12 ↗     |            |
| 00754: squamous cell carcinoma, acantholytic                          | 132 ↗ 22 ↗    |            |
| 00800: differentiated/recapitulating carcinoma                        | 52 ↗ 200 ↗    |            |
| 00900: basal cell carcinoma, nos                                      | 28 ↗ 15 ↗     |            |
| 01200: transitional cell carcinoma, in situ                           | 10 ↗ 10 ↗     |            |
| 01203: transitional cell carcinoma, nos                               | 310 ↗ 423 ↗   |            |
| 01300: urothelial/papillary transitional cell carcinoma, non-invasive | 184 ↗ 39 ↗    |            |
| 01303: papillary transitional cell carcinoma                          | 2 ↗ 6 ↗       |            |
| 01400: adrenocortical carcinoma                                       | 385 ↗ 361 ↗   |            |
| 01402: adrenocortical carcinoma                                       | 88 ↗ 88 ↗     |            |
| 01403: adrenocarcinoma, in situ                                       | 11 ↗ 11 ↗     |            |
| 01403: adrenocarcinoma, nos   | 9469 ↗ 3248 ↗ |            |
| 01443: adrenocarcinoma, intestinal type                               | 167 ↗ 206 ↗   |            |
| 01453: carcinoma, diffuse type  | 7 ↗ 36 ↗      |            |
| 01500: sex cell adenoma   | 8 ↗ 15 ↗      |            |
| 01501: carcinoma, nos   | 1 ↗ 18 ↗      |            |
| 01502: sex cell carcinoma   | 1 ↗ 28 ↗      |            |
| 01511: insularoma, nos  | 29 ↗ 29 ↗     |            |



# Bioinformatics & Data Curation - arrayMap data “Pipeline”



# 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. 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 tracks for chromosomes 7 and 12, with columns for chromosome, start, end, 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 Demo"
chr7 127471196 127472363 Pos1 0 + 127472363
chr7 127472363 127473530 Pos2 0 + 127472363 127473530 255,0,0
chr7 127473530 127474697 Pos3 0 + 127473530 127474697 255,0,0
chr7 127474697 127475864 Pos4 0 + 127474697 127475864 255,0,0
chr7 127475864 127477031 Neg1 0 - 127475864 127477031 0,0,255
chr7 127477031 127478198 Neg2 0 - 127477031 127478198 0,0,255
chr7 127478198 127479365 Neg3 0 - 127478198 127479365 0,0,255
chr7 127479365 127480532 Pos5 0 + 127479365 127480532 255,0,0
chr7 127480532 127481699 Neg4 0 - 127480532 127481699 0,0,255
```

*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
```

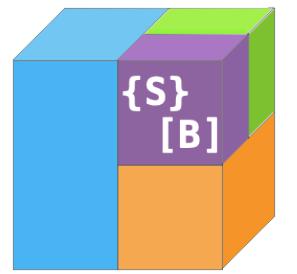
# Bioinformatics: File Formats, Ontologies & APIs

- 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

```
{"api_params": {"genome": "hg18", "db": "progenetix", "datatype": "sampledata", "count": 20, "call": "api_doctype=json&api_out=samples&db=progenetix&icdm_m=817&randno=20", "scope": "samples"}, "data": [{"ICDMORPHOLOGY": "Liver cell adenoma", "NCIT:CODE": "C3758", "ICDTOPOGRAPHYCODE": "C22", "_id": {"$oid": "558e5c2ead9a82d95838f76c"}, "CLINICALGROUP": "Carcinomas: hepatic ca.", "PMID": "15765123", "FOLLOWUP": "", "BIOSAMPLEID": "AM_BS_HKCI-C2-D0R", "ICDMORPHOLOGYCODE": "8170/0", "NCIT:TERM": "Hepatocellular Adenoma", "UID": "HKCI-C2-D0R", "DIAGNOSISTEXT": "Hepatocellular carcinoma [cell line, doxorubicin resistant subclone]", "DEATH": "", "ICDTOPOGRAPHY": "liver", "AGE": ""}, {"FOLLOWUP": "", "PMID": "14578863", "ICDMORPHOLOGYCODE": "8170/3", "BIOSAMPLEID": "AM_BS_PHCC-30", "ICDMORPHOLOGY": "Hepatocellular carcinoma, NOS", "_id": {"$oid": "558e5c36ad9a82d9583901bf"}, "CLINICALGROUP": "Carcinomas: hepatic ca.", "ICDTOPOGRAPHYCODE": "C22", "NCIT:CODE": "C3099", "ICDTOPOGRAPHY": "liver", "DEATH": "", "DIAGNOSISTEXT": "Hepatocellular carcinoma", "AGE": "", "NCIT:TERM": "Hepatocellular Carcinoma", "UID": "PHCC-30"}, {"DEATH": "", "DIAGNOSISTEXT": "Hepatocellular carcinoma [chronic Hepatitis B]", "ICDTOPOGRAPHY": "liver", "AGE": "P54Y2M", "NCIT:TERM": "Hepatocellular Carcinoma", "UID": "HCC-1997-14", "PMID": "8993981", "FOLLOWUP": "", "BIOSAMPLEID": "AM_BS_HCC-1997-14", "ICDMORPHOLOGYCODE": "8170/3", "ICDMORPHOLOGY": "Hepatocellular carcinoma, NOS", "NCIT:CODE": "C3099", "ICDTOPOGRAPHYCODE": "C22", "_id": {"$oid": "558e5bfccad9a82d95838b62a"}, "CLINICALGROUP": "Carcinomas: hepatic ca."}, {"FOLLOWUP": "", "PMID": "11485905", "BIOSAMPLEID": "AM_BS_HCChypo-won-H18", "ICDMORPHOLOGYCODE": "8170/3", "ICDMORPHOLOGY": "Hepatocellular carcinoma, NOS", "CLINICALGROUP": "Carcinomas: hepatic ca.", "_id": {"$oid": "558e5c48ad9a82d95839185a"}, "NCIT:CODE": "C3099", "SEX": "male"}]
```

[http://progenetix.org/api/?db=progenetix&api\\_out=samples&api\\_doctype=json&icdm\\_m=817&randno=20](http://progenetix.org/api/?db=progenetix&api_out=samples&api_doctype=json&icdm_m=817&randno=20)



## BeaconAlleleRequest beacon ↗

|                   |   |
|-------------------|---|
| {S}[B] Status [i] | implemented   |
| Provenance        | ◦ Beacon API  |
| Used by           | ◦ Beacon<br>◦ Progenetix database schema (Beacon+ backend)  |
| Contributors      | ◦ Marc Fiume<br>◦ Michael Baudis<br>◦ Sabela de la Torre Pernas<br>◦ Jordi Rambla<br>◦ Beacon developers... |
| Source (v1.1.0)   | ◦ raw source [JSON]<br>◦ 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       | <a href="https://schemablocks.org/schemas/beacon/v1.1.0/Chromosome">https://schemablocks.org/schemas/beacon/v1.1.0/Chromosome</a> [HTML] |
| referenceBases | string   |
| referenceName  | <a href="https://schemablocks.org/schemas/beacon/v1.1.0/Chromosome">https://schemablocks.org/schemas/beacon/v1.1.0/Chromosome</a> [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        | ◦ vr-spec                        |
| Used by           | ◦ vr-spec                        |
| Contributors      | ◦ Reece Hart<br>◦ Michael Baudis |

### Attributes

Type: object

Description: A CURIE is a Uniform Resource Identifier (URI) that identifies a single entity. It consists of a prefix followed by a namespace and a local identifier. The prefix is typically a well-known identifier for a specific domain, such as 'http://www.w3.org/2002/07/owl#' for the Web Ontology Language (OWL). The namespace is a URI that identifies the vocabulary or ontology being used. The local identifier is a unique identifier within that vocabulary.

VR does not impose any constraints on strings used as identifiers, the VR Specification RECOMMENDS that implementers use standard CURIEs.

String CURIEs are represented as [prefix:reference](#) (W3C Recommendation) or [namespace:accession](#) or [namespace:local\\_id](#) colloquially. The VR specification also RECOMMENDS that [prefix](#) be the [reference](#) component is an unconstrained string.

A CURIE is a URI. URIs may *locate* objects (i.e., specify where they are located) or identify resources (i.e., specify what they are). VR uses CURIEs primarily as a naming mechanism.

Implementations MAY provide CURIE resolution mechanisms. Using internal IDs in public messages is strongly discouraged.

### 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<br>◦ Jules Jacobsen<br>◦ Peter Robinson<br>◦ Michael Baudis<br>◦ Melanie Courtot<br>◦ Isuru Liyanage |

### 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 fraction from a gradient centrifugation.

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

FHIR mapping: Specimen.

### Properties

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

## Checksum sb-checksum ↗

|                   |                                |
|-------------------|--------------------------------|
| {S}[B] Status [i] | proposed                       |
| Provenance        | ◦ GA4GH DRS (`develop` branch) |
| Used by           | ◦ GA4GH DRS<br>◦ GA4GH TRS     |
| Contributors      | ◦ Susheel Varma                |

### 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

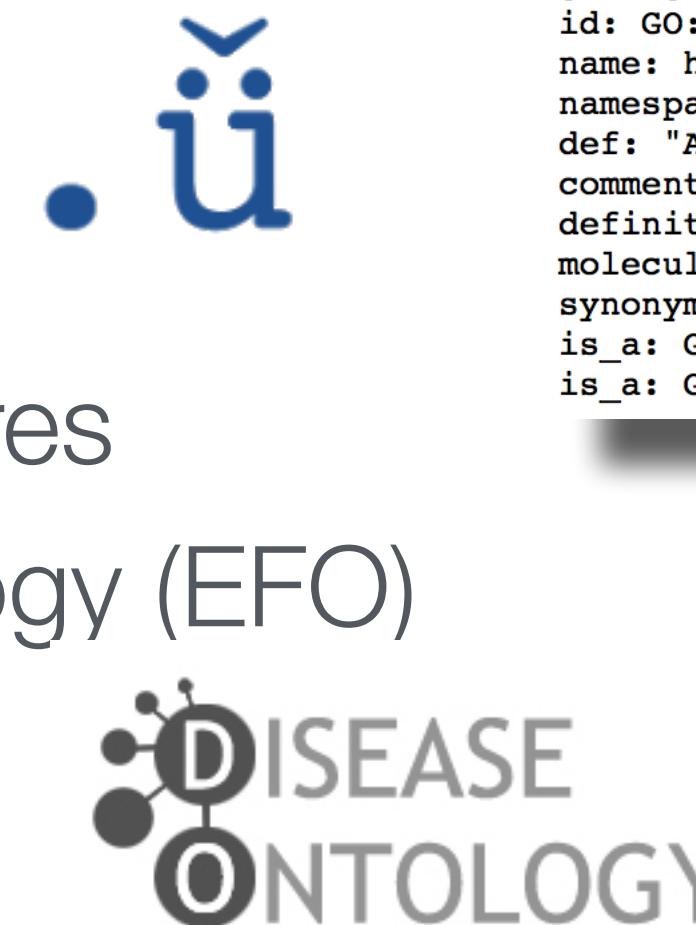
"sha-256"

# Bioinformatics: File Formats, Ontologies & APIs

- 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

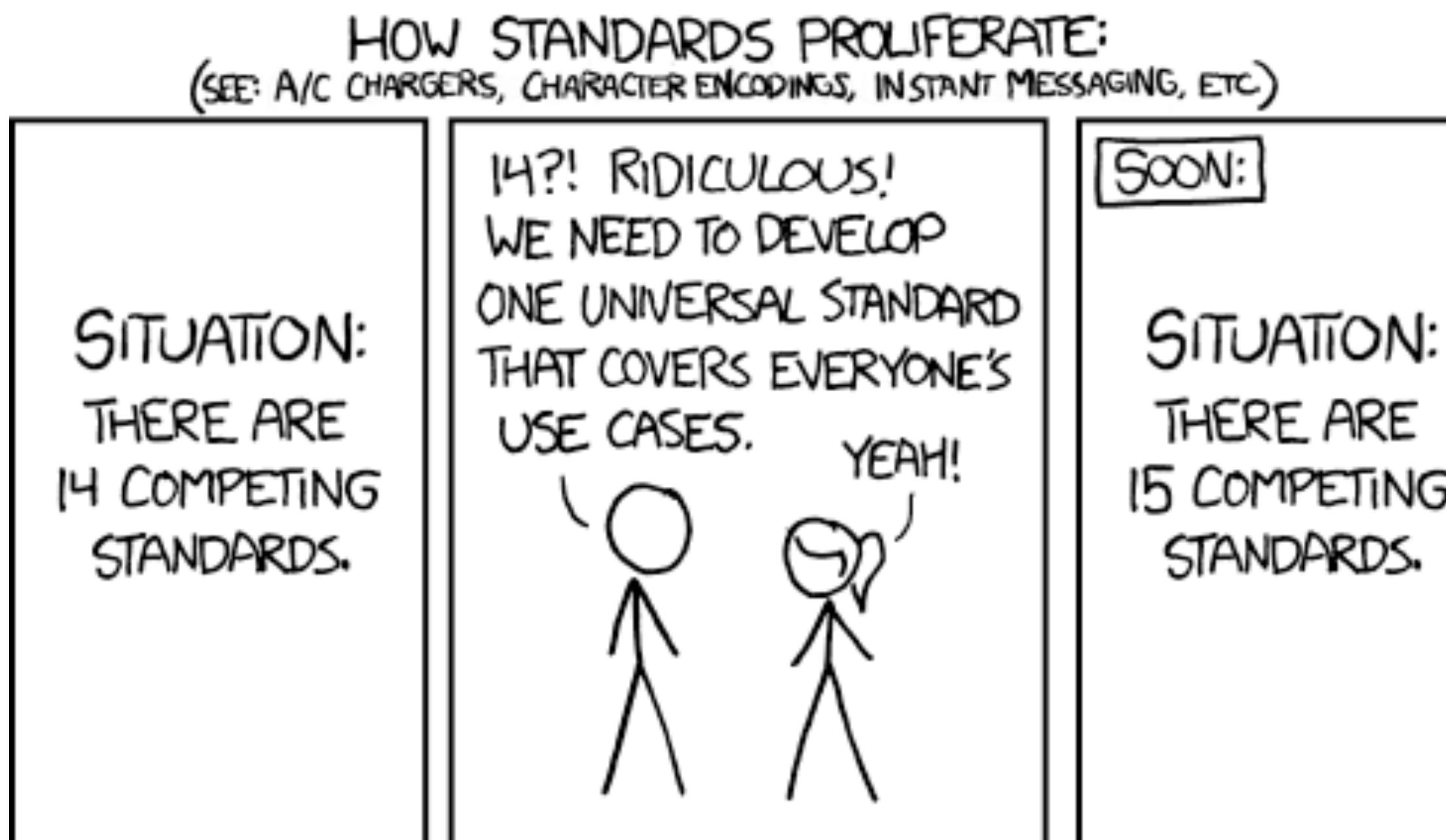
**Data re-use depends on standardized, machine-readable metadata**

- 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 ...)

```
"data_use_conditions" : {  
    "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  
            ]  
        },  
        "IS0-3166-alpha3" : "CHE"  
    },  
    {  
        "age": "P25Y3M2D"  
    }  
}
```

# Standardized Data

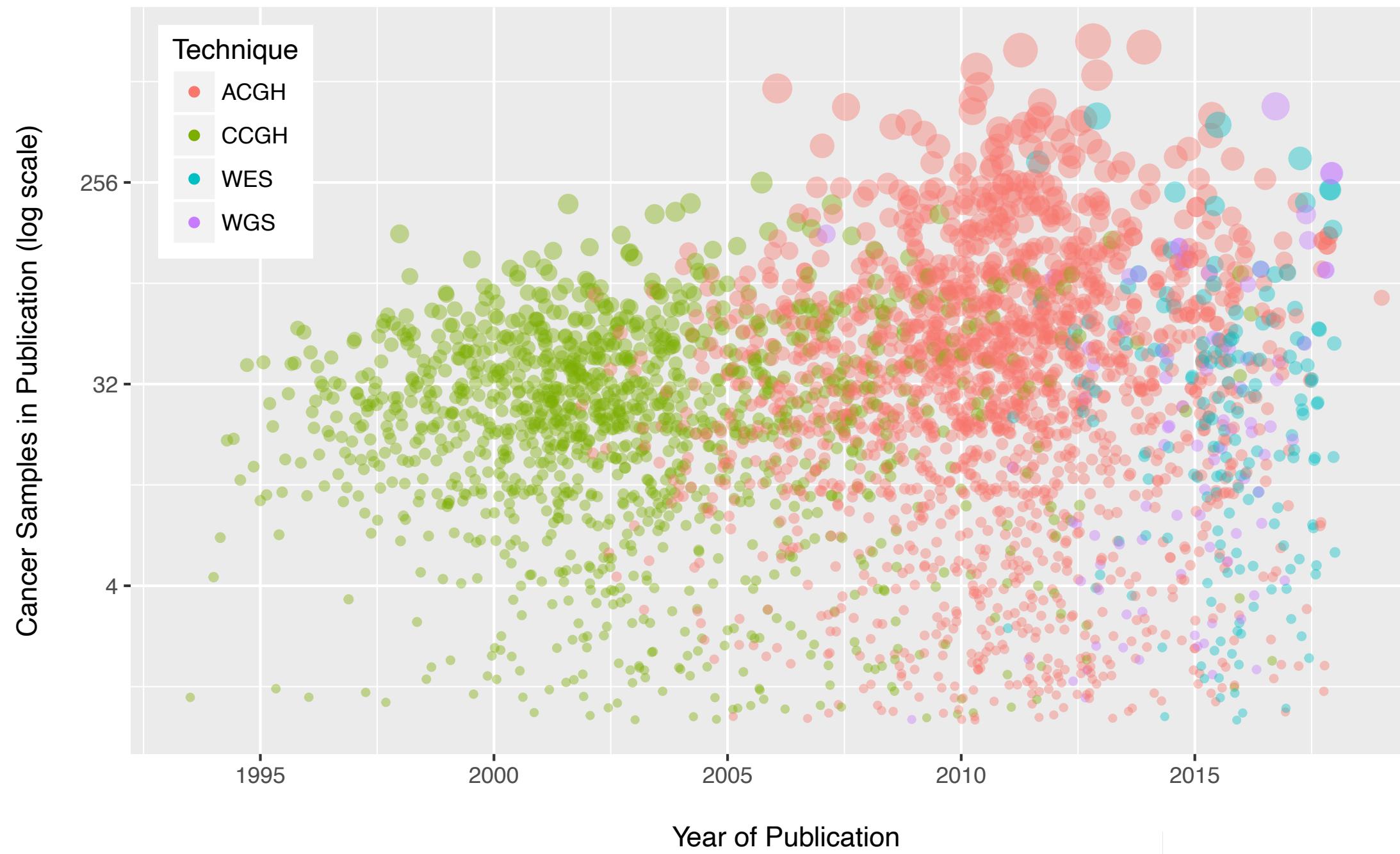
Data re-use depends on standardized, machine-readable metadata



xkcd

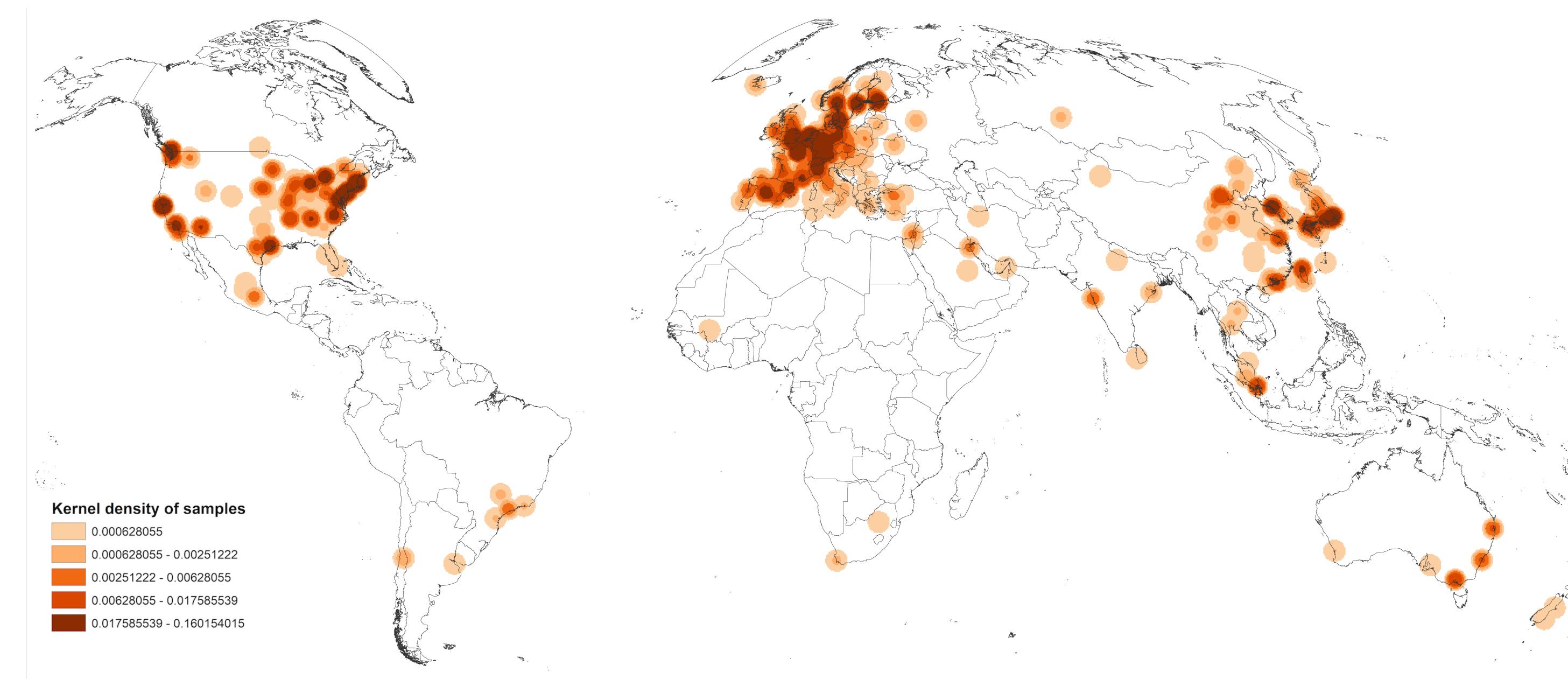
```
"data_use_conditions" : {  
    "label" : "no restriction",  
    "id" : "DUO:0000004"  
},  
  
"provenance" : {  
    "material" : {  
        "type" : {  
            "id" : "EFO:0009656",  
            "label" : "neoplastic sample"  
        }  
    },  
    "geo" : {  
        "label" : "Zurich, 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"  
    }  
}
```

# 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.

# 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**

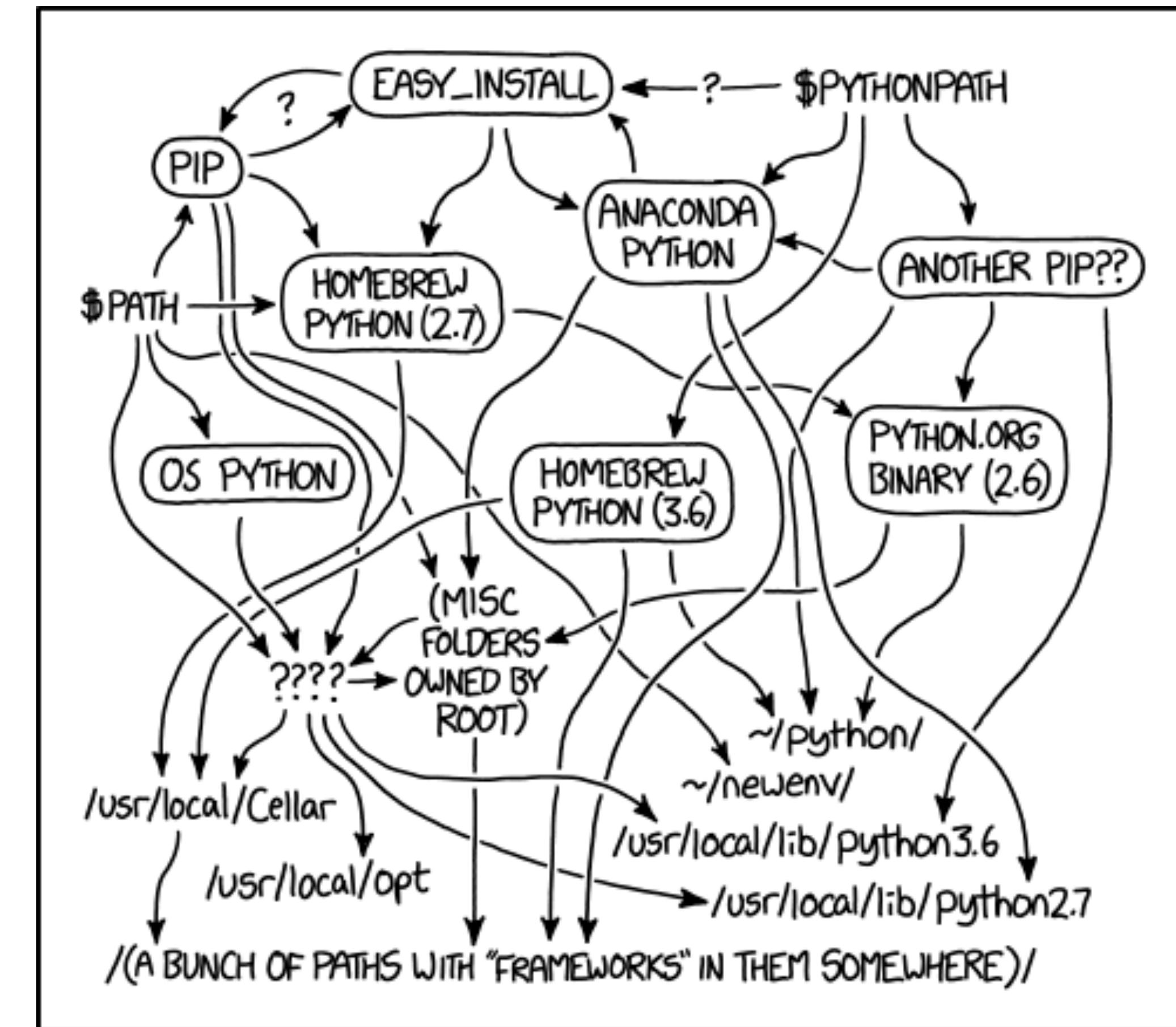
# Bioinformatics OR Computational / Systems Biology?

- 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



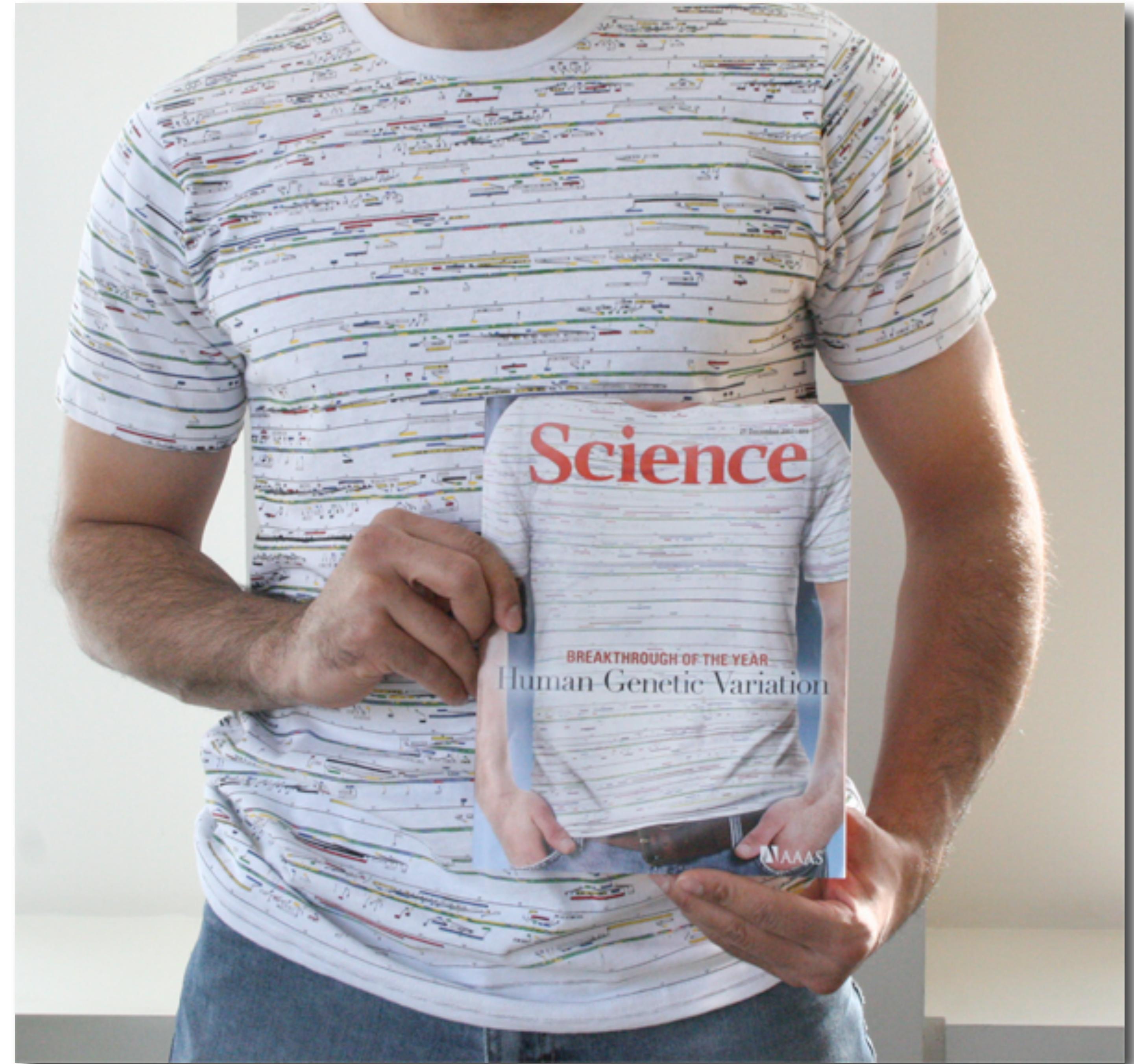
MY PYTHON ENVIRONMENT HAS BECOME SO DEGRADED  
THAT MY LAPTOP HAS BEEN DECLARED A SUPERFUND SITE.

# **BIO390: Introduction to Bioinformatics**

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

**Michael Baudis | 2020-09-15**

The trouble with human genome variation



# Conclusions from the analysis of variation in the human genome

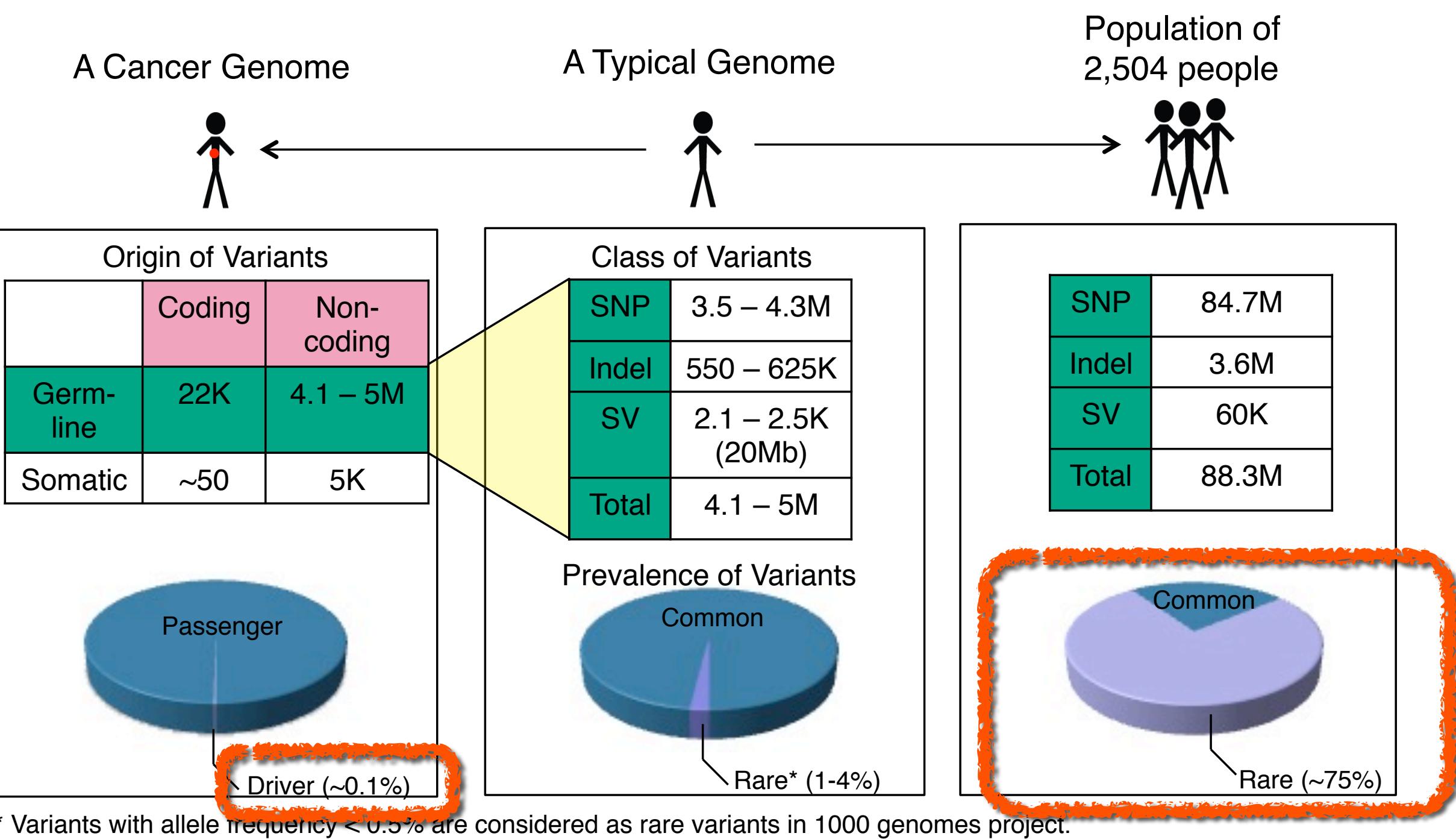
---

- 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

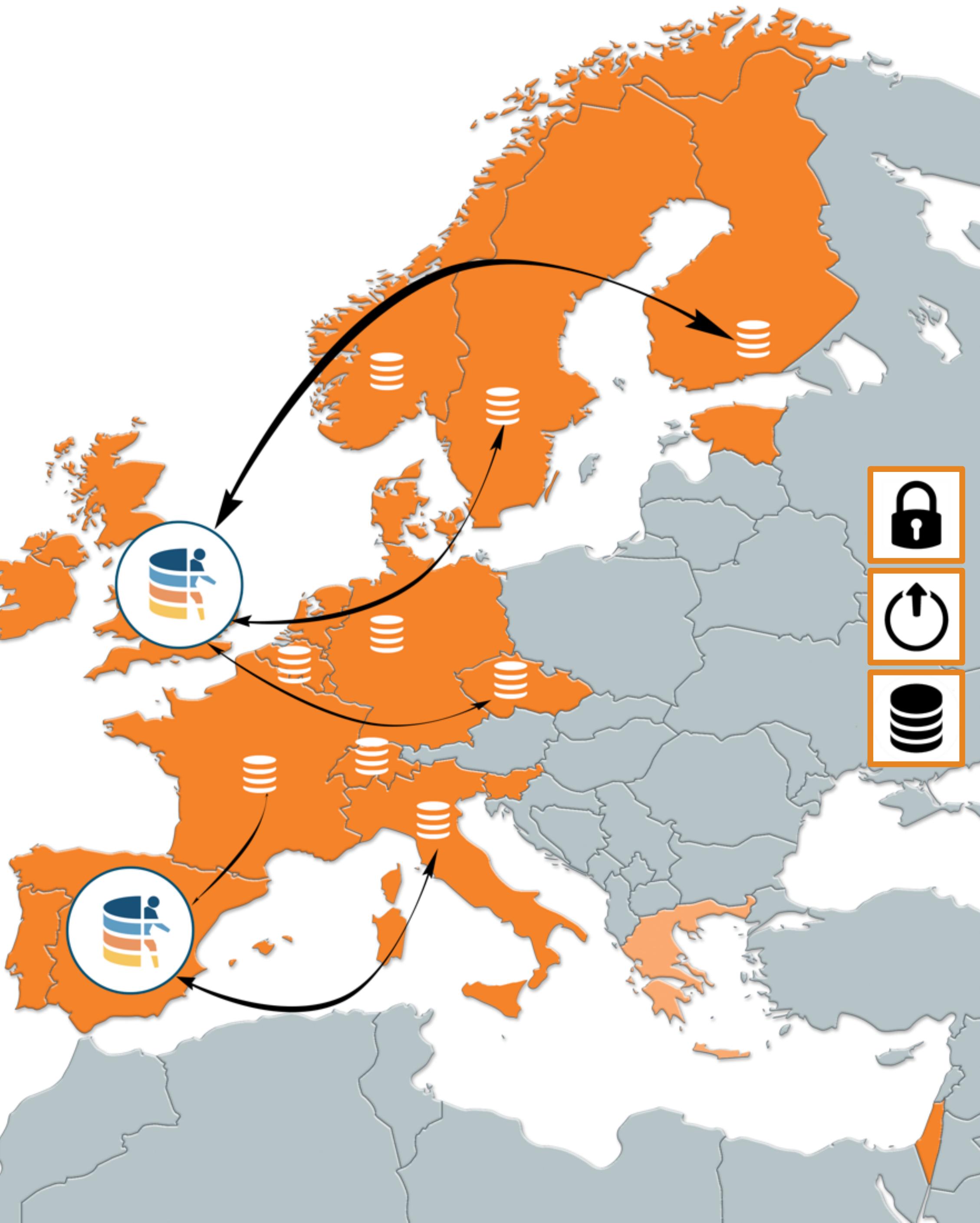


The 1000 Genomes Project Consortium, Nature. 2015. 526:68-74  
Khurana E. et al. Nat. Rev. Genet. 2016. 17:93-108

Graphic adapted from Mark Gerstein ([GersteinLab.org](http://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)



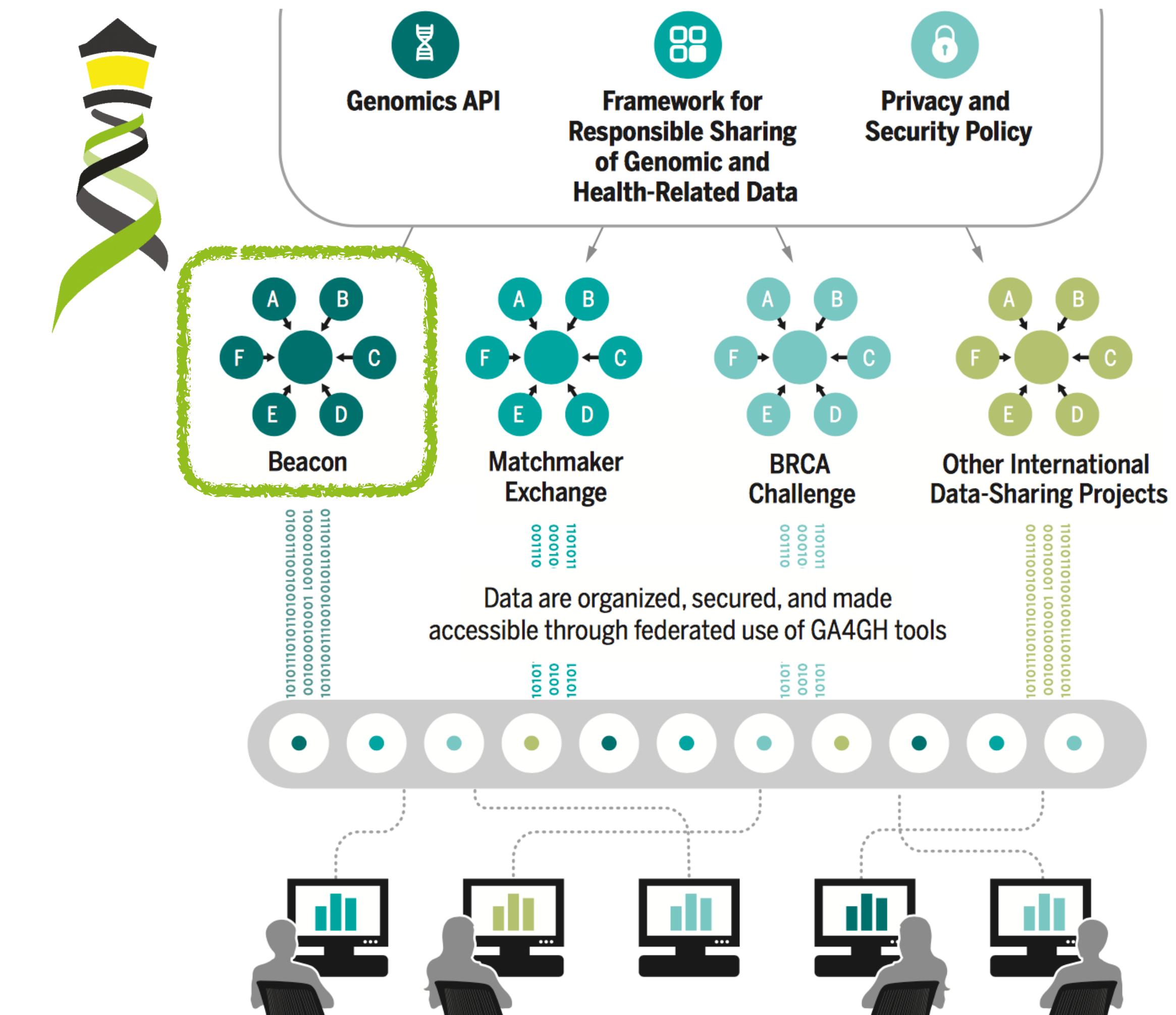


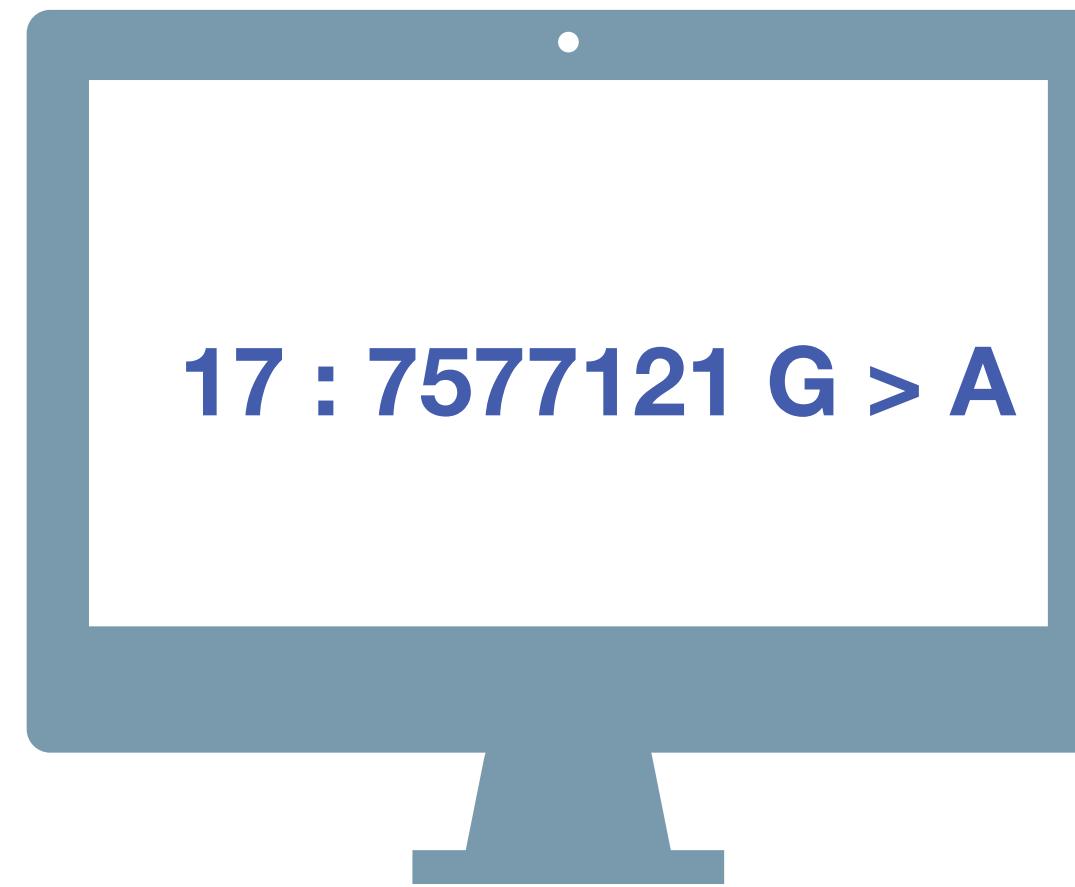
## GENOMICS

# *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**

# 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

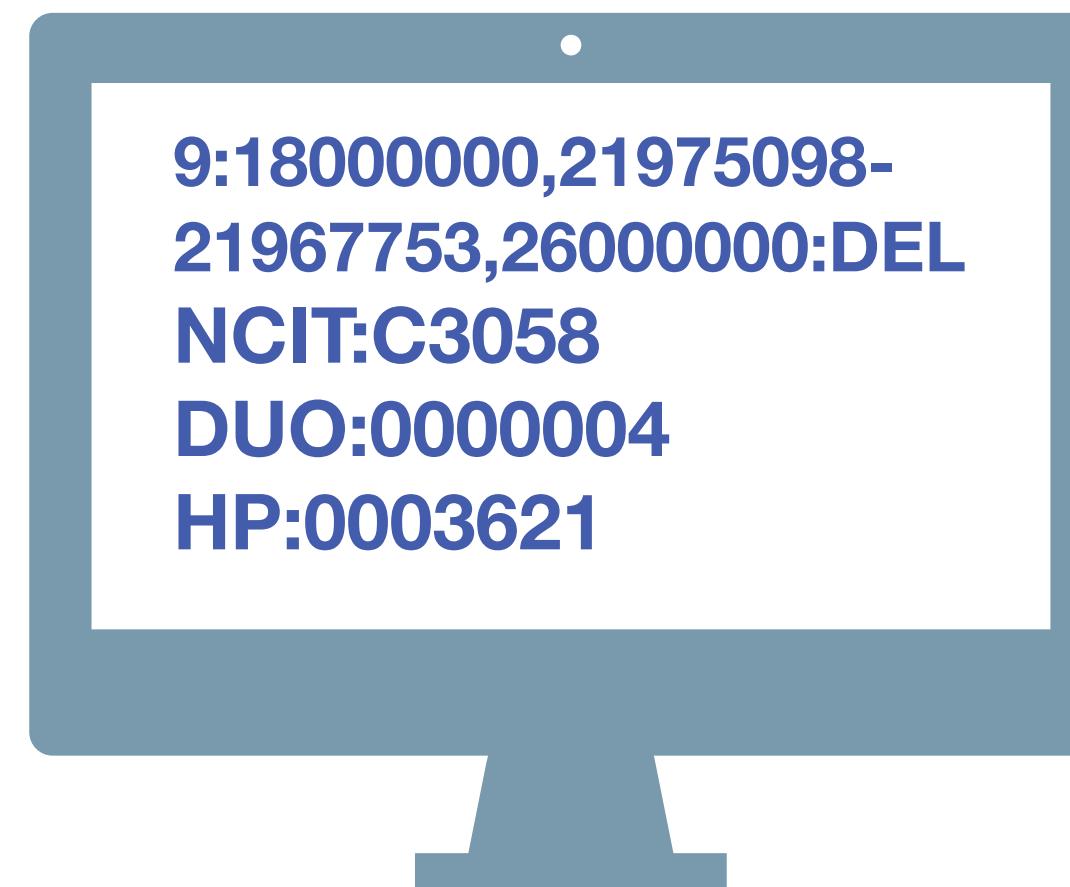
Beacon API Web Server based on the GA4GH Beacon API

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

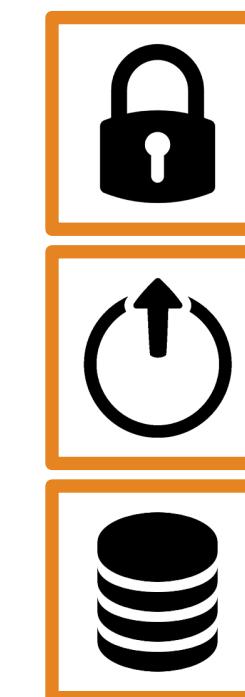
CSC - IT Center for Science Production Beacon

Beacon API Web Server based on the GA4GH Beacon API

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

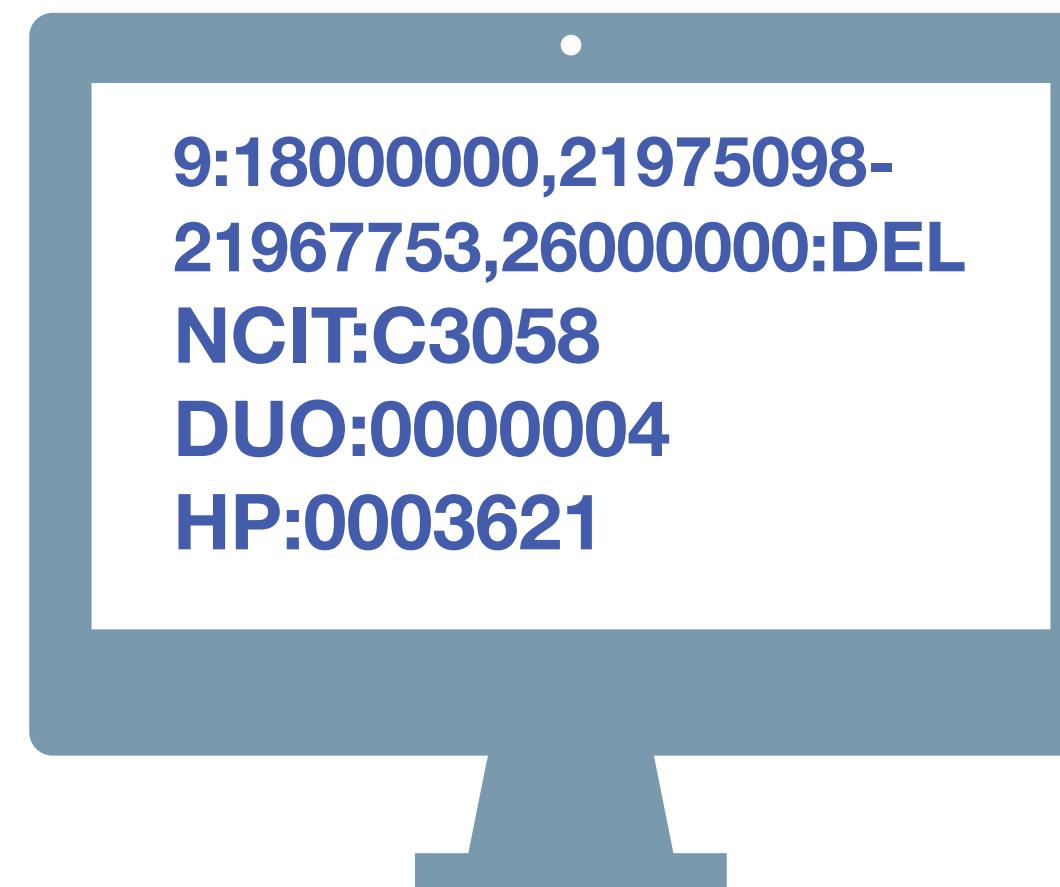


Have you seen deletions in this region on chromosome 9 in Glioblastomas from a juvenile patient, in a dataset with unrestricted access?

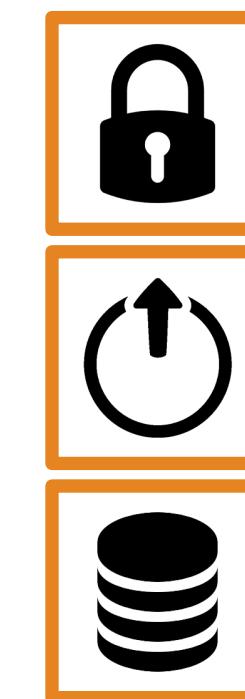


## Beacon v2 API

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

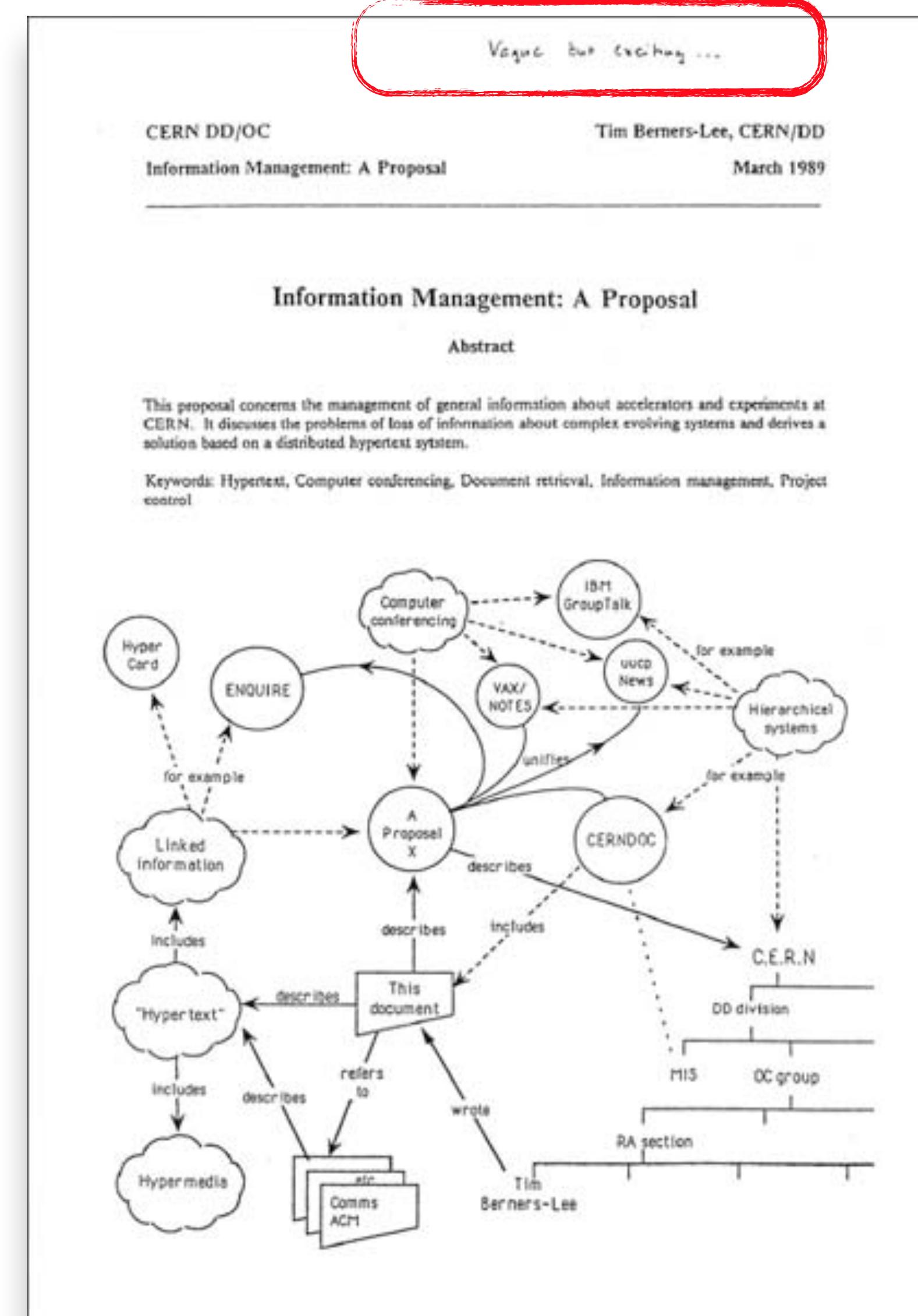


Have you seen deletions in this region on chromosome 9 in Glioblastomas from a juvenile patient, in a dataset with unrestricted access?



## Beacon v2 API

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



# 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

- 2020-09-12: Michael Baudis - What is Bioinformatics? Introduction and Resources
- **2020-09-22: Christian von Mering - Sequence Bioinformatics**
- 2020-09-29: Mark Robinson - Statistical Bioinformatics
- 2020-10-06: Patrick Ruch (UniGe) - Text Mining
- 2020-10-13: Katja Baerenfaller (SIAF) - Proteomics
- 2020-10-20: Shinichi Sunagawa (ETHZ) - Metagenomics
- 2020-10-27: Valentina Boeva (ETHZ) - Machine Learning for Biological Use Cases
- 2020-11-03: Izaskun Mallona - Regulatory Genomics and Epigenomics
- 2020-11-10: Andreas Wagner - Biological Networks
- 2020-11-17: Abdullah Kahraman (USZ) - Molecular Interaction Networks
- 2020-11-24: Valérie Barbie (SIB) - Clinical Bioinformatics
- 2020-12-01: Michael Baudis - Building a Genomics Resource
- 2020-12-08: Michael Baudis - Genome Data & Privacy
- 2020-12-15: Exam (Multiple Choice)



University of  
Zurich<sup>UZH</sup>



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

[arraymap.org](http://arraymap.org)  
[progenetix.org](http://progenetix.org)  
[info.baudisgroup.org](http://info.baudisgroup.org)  
[sib.swiss/baudis-michael](http://sib.swiss/baudis-michael)  
[imls.uzh.ch/en/research/baudis](http://imls.uzh.ch/en/research/baudis)  
[beacon-project.io](http://beacon-project.io)  
[schemablocks.org](http://schemablocks.org)



Global Alliance  
for Genomics & Health

