

BIO390: Introduction to Bioinformatics

Lecture I: What is Bioinformatics?



Course Information BIO390

- Tuesdays at 08:00; 2x45min
- 13 presentations by different lecturers
- (unchecked) homework / preparation exercises w/ focus on test topics
- course language is English
- course slides may/should be made available through the website
- written exam at end of course (== 14th course - December 15)
- Organizer:

Prof. Dr. Michael Baudis

Department of Molecular Life Sciences (IMLS)

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

CH-8057 Zurich

email michael@baud.is

web info.baudisgroup.org

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

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

CompbioZurich

General Information

News & Events

Bioinformatics Groups

Lectures, Courses & Graduate Programs

Overview

UZH BIO390

UZH BIO392

Community Resources

Tools & Resources

Positions

BIO390 – Introduction to Bioinformatics

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.

 Requirements

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

 Course Info

- Autumn semesters
- Tuesdays at 08:00-09:45 (15min break)
- UZH Irchel campus, **Y03-G-85**
- [OLAT](#) - but not much there...
- No lecture recordings - we do **not record** the lectures since HS23 (regular attendance is expected) *but* there might be still 2022 [lecture recordings](#) available
- Course language is English

Some [learning goals](#) provide additional guidance - but please be aware that those may include details which not have been covered in the current semester and therefore won't be necessarily relevant for the exam.

Schedule

[Next: What is Bioinformatics? Introduction and Resources](#)BIO390 UZH HS25 - INTRODUCTION TO BIOINFORMATICS
08:00-09:45 @ UZH IRCHEL Y03-G-85 September 16, 2025**Michael Baudis**

The "What is Bioinformatics? Introduction and Resources" provides a general introduction into the field and a description of the lecture topics, timeline and procedures.

compbiozurich.org/teaching /UZH-BIO390/

[Upcoming: Statistical Bioinformatics](#)BIO390 UZH HS25 - INTRODUCTION TO BIOINFORMATICS
08:00-09:45 @ UZH IRCHEL Y03-G-85 September 23, 2025**Mark Robinson**

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

[→ Continue reading](#)[Upcoming: Biological Sequence Informatics](#)BIO390 UZH HS25 - INTRODUCTION TO BIOINFORMATICS
08:00-09:45 @ UZH IRCHEL Y03-G-85 September 30, 2025**Christian von Mering**

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

[Upcoming: Machine Learning for Biological Use Cases](#)BIO390 UZH HS25 - INTRODUCTION TO BIOINFORMATICS
08:00-09:45 @ UZH IRCHEL Y03-G-85 October 07, 2025**Valentina Boeva (ETHZ)**

Brief note: In this lecture V. Boeva will cover the standard machine learning methods used in the analysis of biological data: dimensionality reduction, clustering, classification and regression.

[→ Continue reading](#)

BIO390: Course Schedule

- 2026-09-15: Michael Baudis - What is Bioinformatics? Introduction and Resources
- 2025-09-22: Christian von Mering - Sequence Bioinformatics
- 2025-09-29: Mark Robinson - Statistical Bioinformatics
- 2026-10-06: Virginie Uhlmann - Image Data & Analysis
- 2026-10-13: Valérie Barbie (SIB) - Clinical Bioinformatics
- 2026-10-20: Valentina Boeva (ETHZ) - Machine Learning for Biological Use Cases
- 2026-10-27: Katja Baerenfaller (SIAF) - Proteomics
- 2026-11-03: Shinichi Sunagawa (ETHZ) - Metagenomics
- 2026-11-10: Patrick Ruch (HesGe) - Text mining & Search Tools
- 2026-11-17: Anuradha Mukherjee - Biological Networks & Evolution
- 2026-11-24: Michael Baudis - Building Biological Information Resources
- 2026-12-01: Patrizia Palagi (SIB) - FAIR Data Standards, Methods and Education in Bioinformatics
- 2026-12-08: Michael Baudis - Genome Data & Privacy | Feedback
- 2026-12-15: Exam (Multiple Choice, mostly)



Some Recommended Books

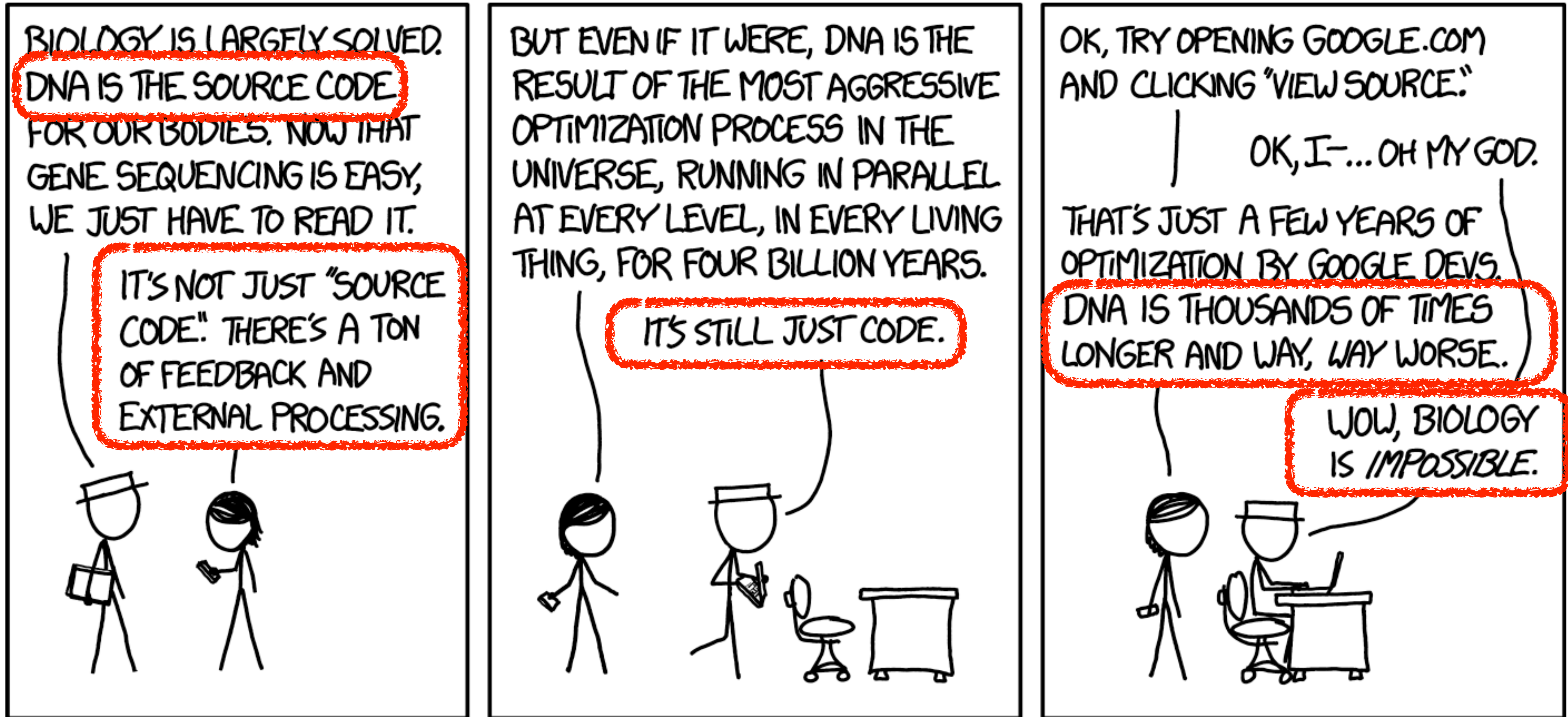
- Anna Tramontano: Introduction to Bioinformatics
- Susan Holmes and Wolfgang Huber: Statistics for Biology
- Robert Gentleman: R Programming for Bioinformatics
- John Maindonald & W. John Braun: Data Analysis and Graphics Using R
- Andy Hector: The New Statistics with R
- Neil C. Jones & Pavel A. Pevzner: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



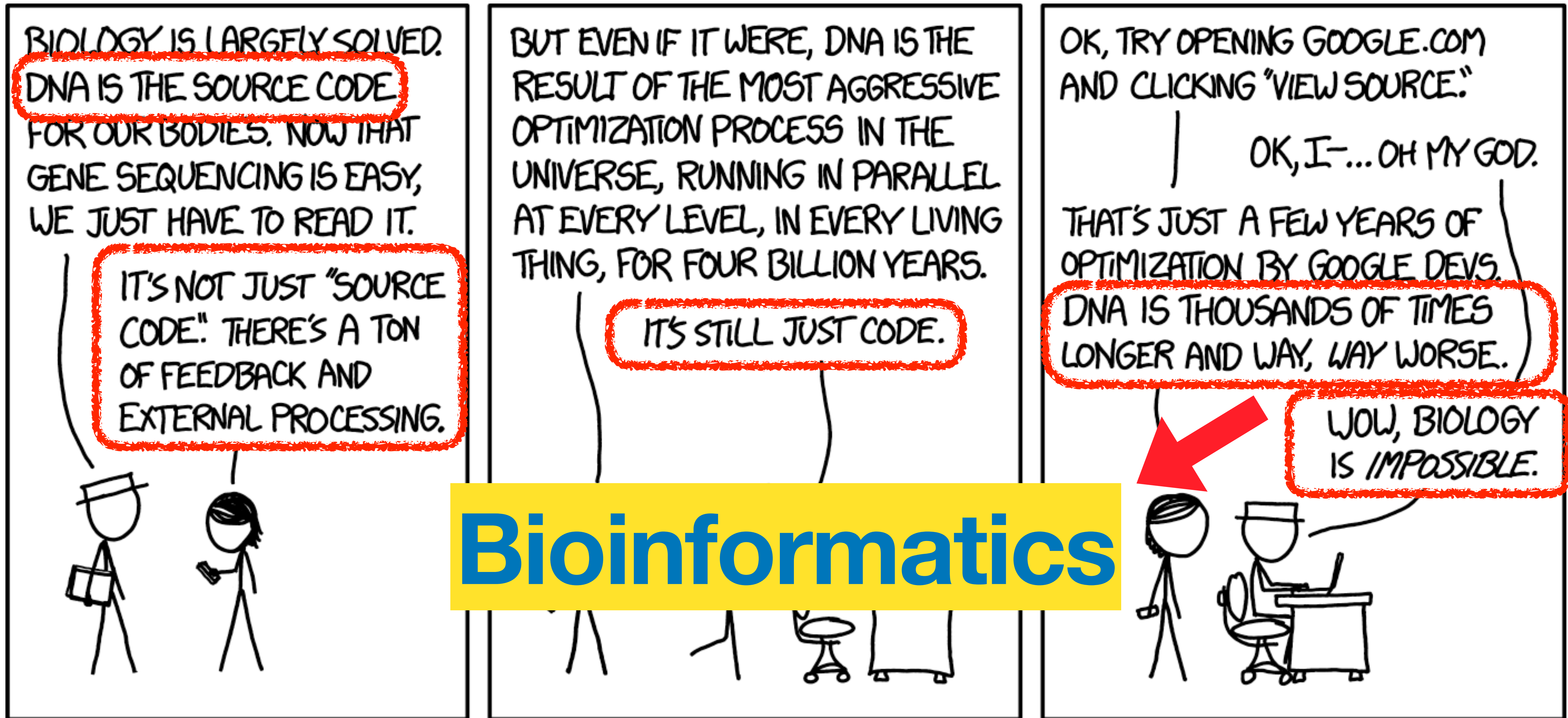
Why Bioinformatics?

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

Biology is *impossibly complex* - But bioinformatics might help



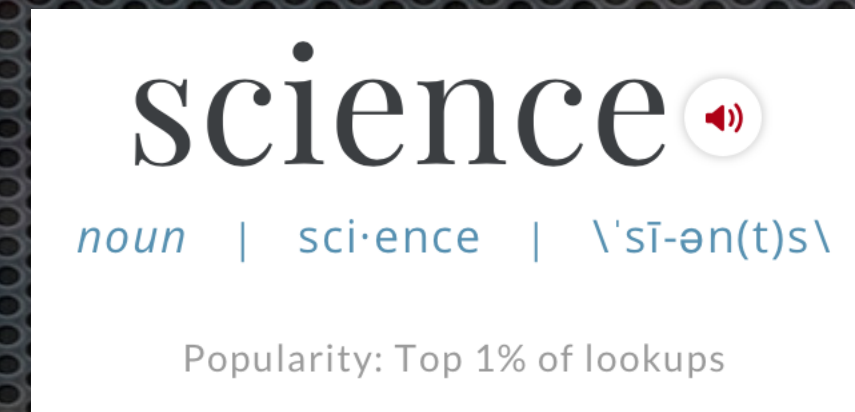
Biology is *impossibly complex* - But bioinformatics might help



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)

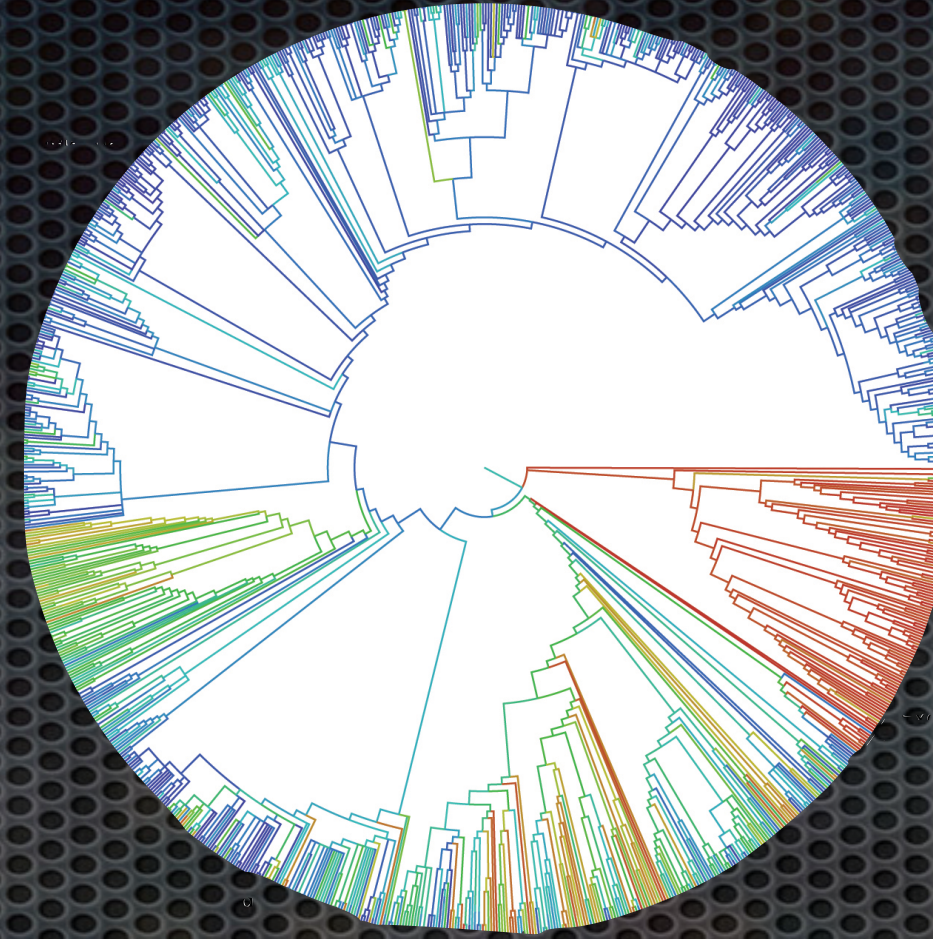
b : an interpretation of a practical situation or condition taken as the ground for action

What is Bioinformatics?

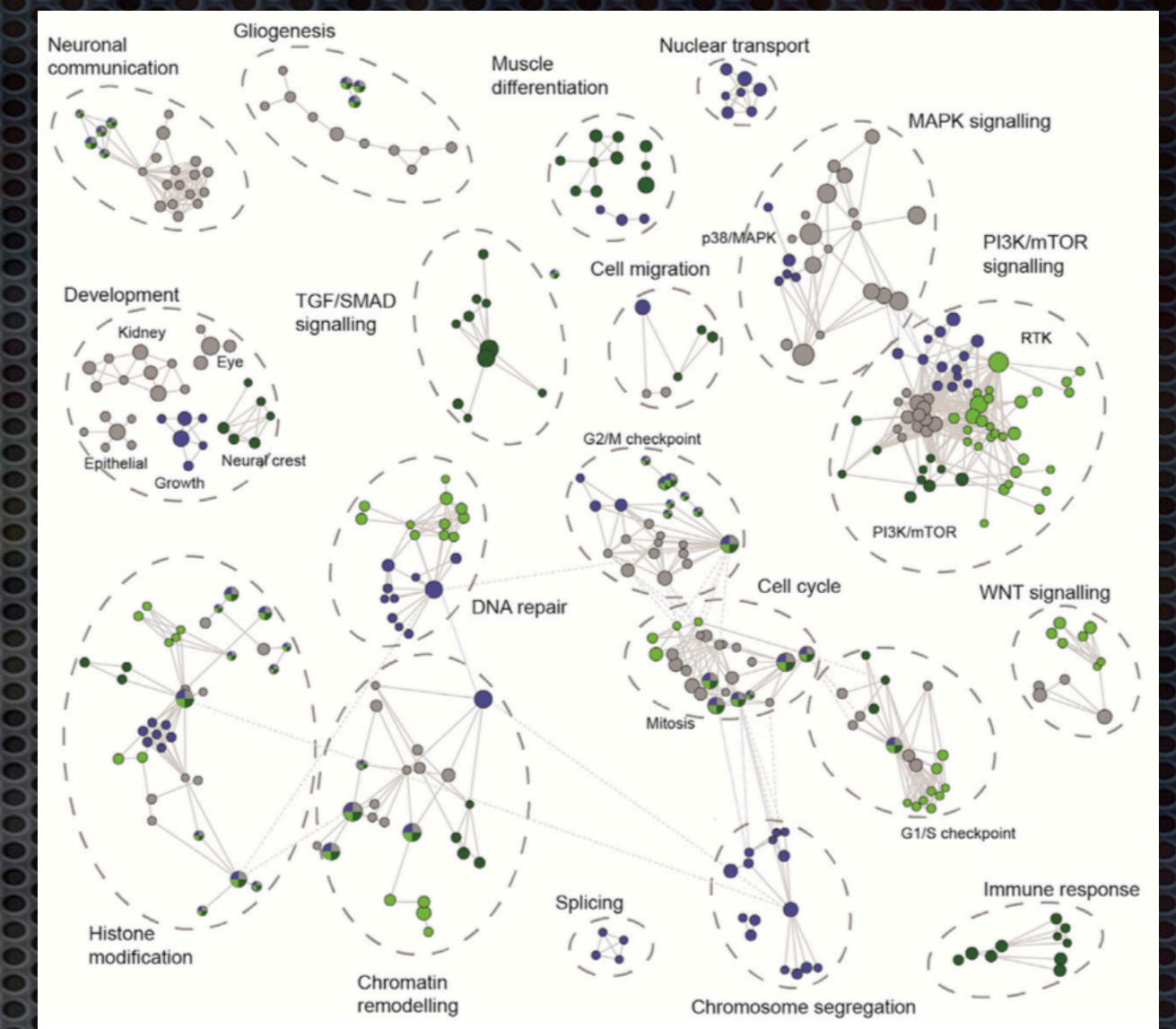


- ✦ Bioinformatics is "the science that uses the instruments of informatics to analyze biological data in order to **formulate hypotheses** about life." (Anna Tramontano)

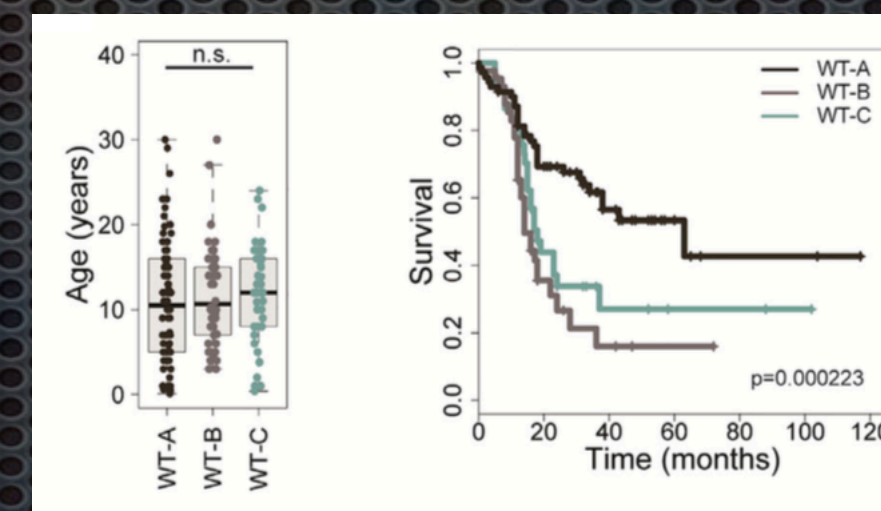
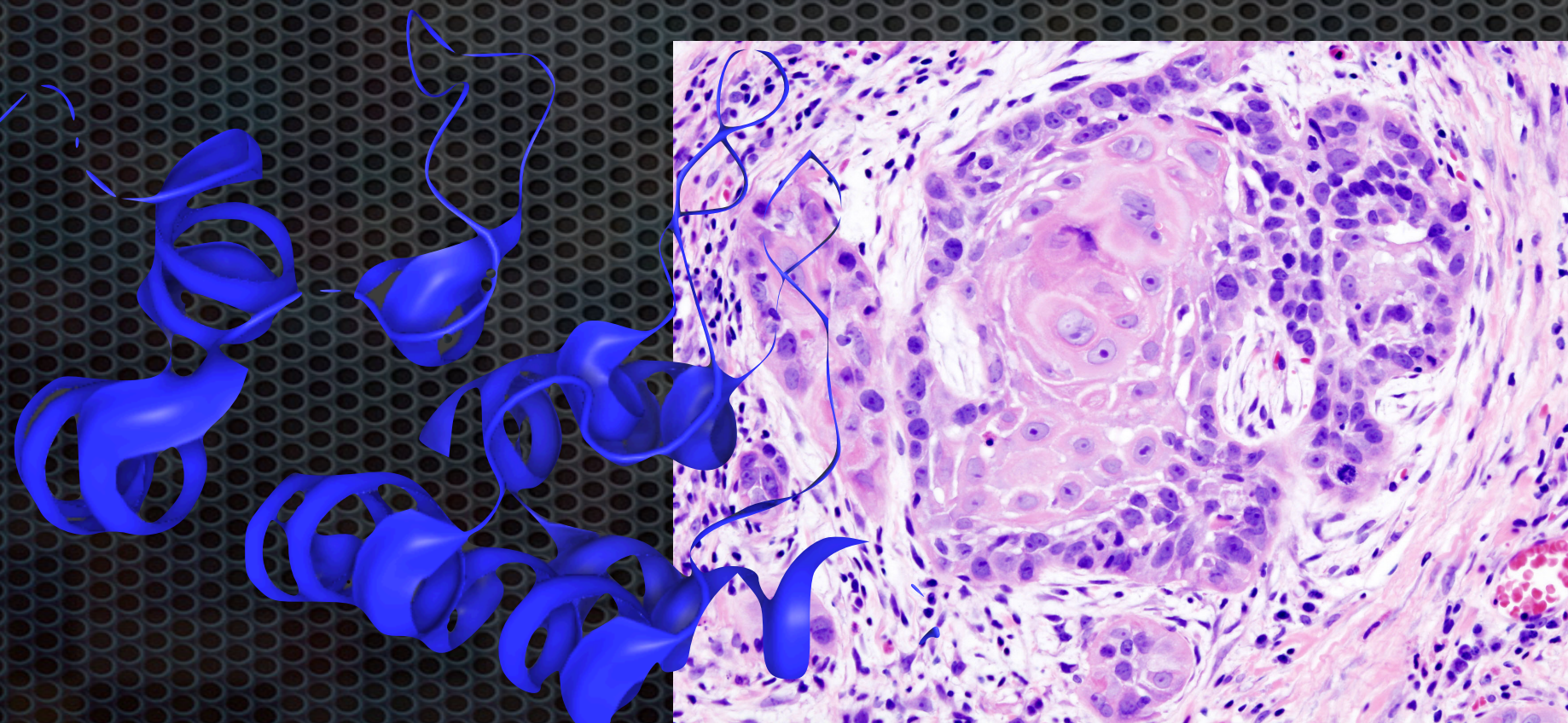
b : an interpretation of a practical situation or condition taken as the ground for action



42



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

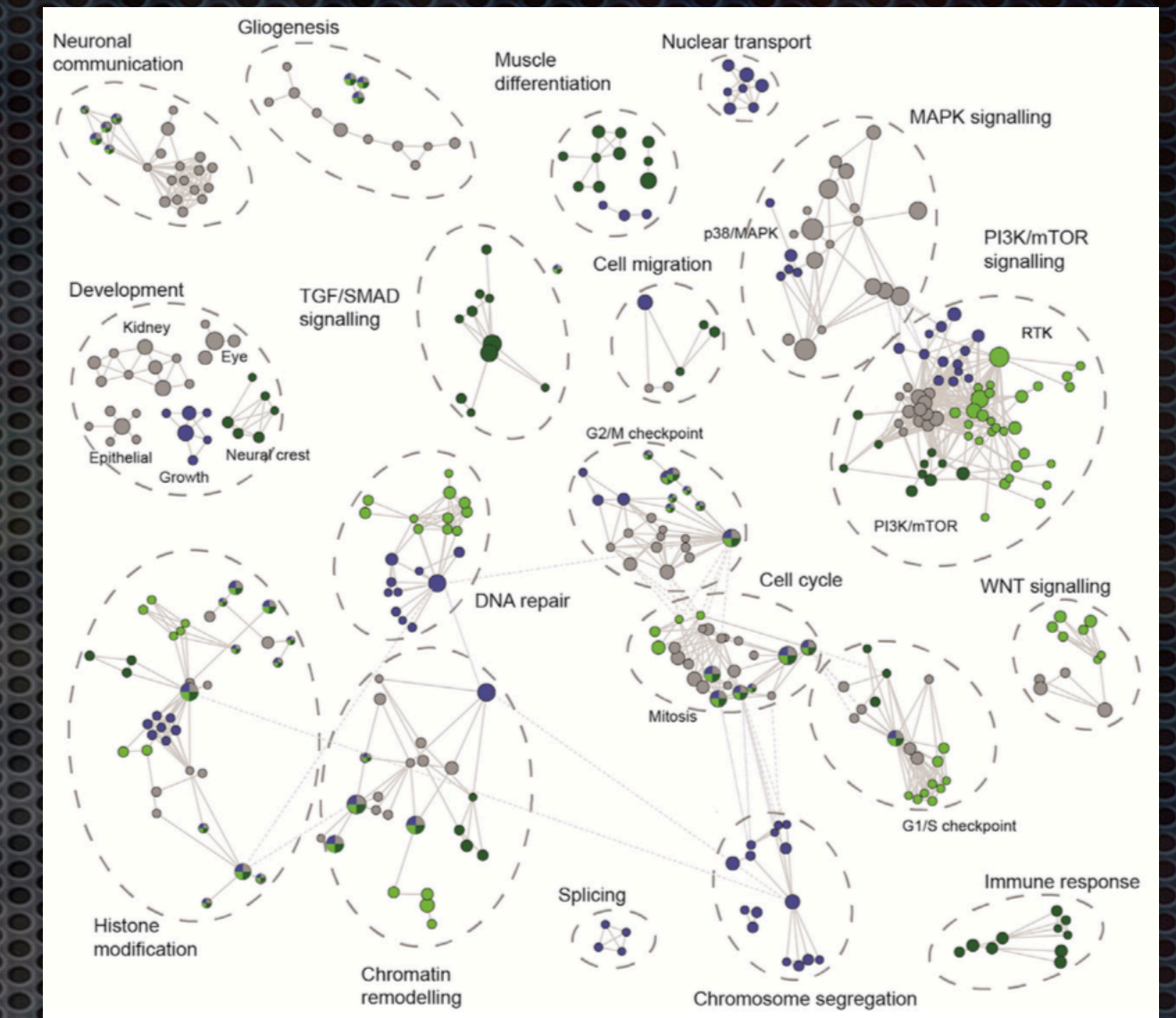
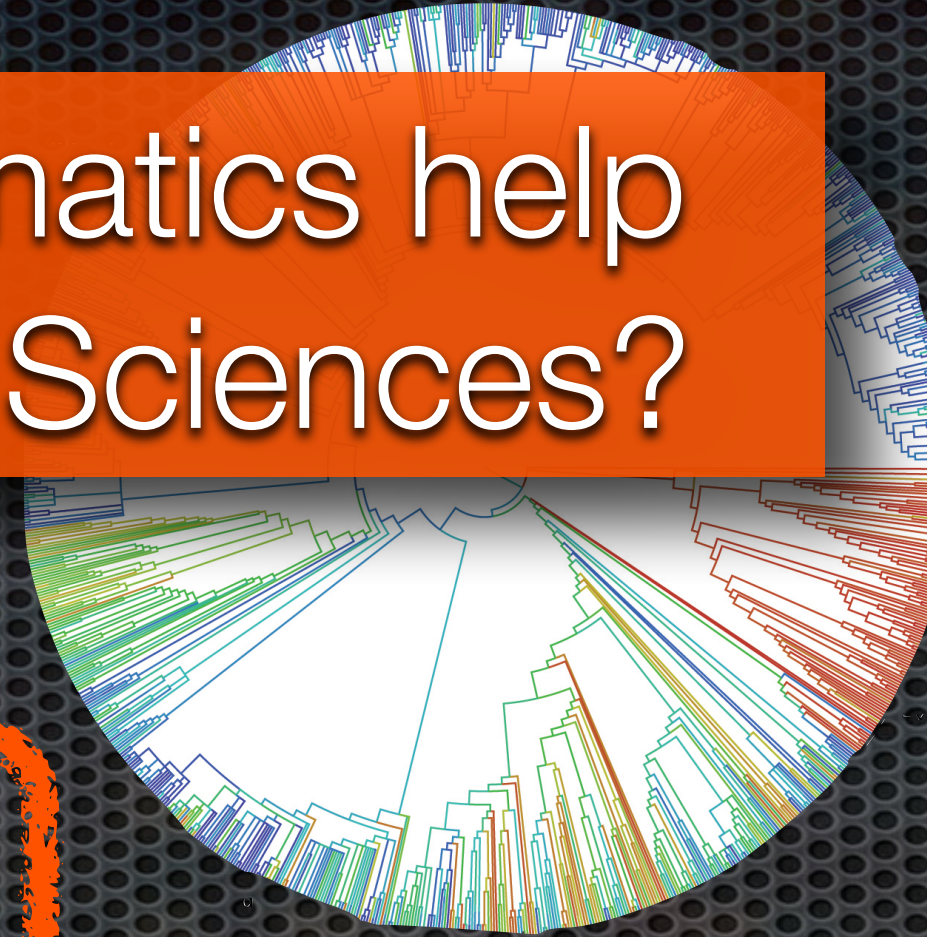


Sources: nextprot | opentreeoflife | wikipedia | MacKay et al., Cancer Cell (2017) | original photos

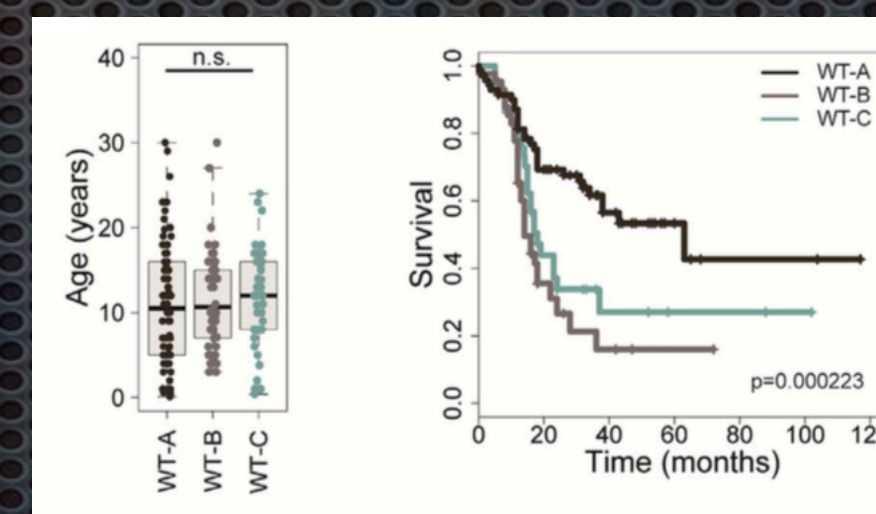
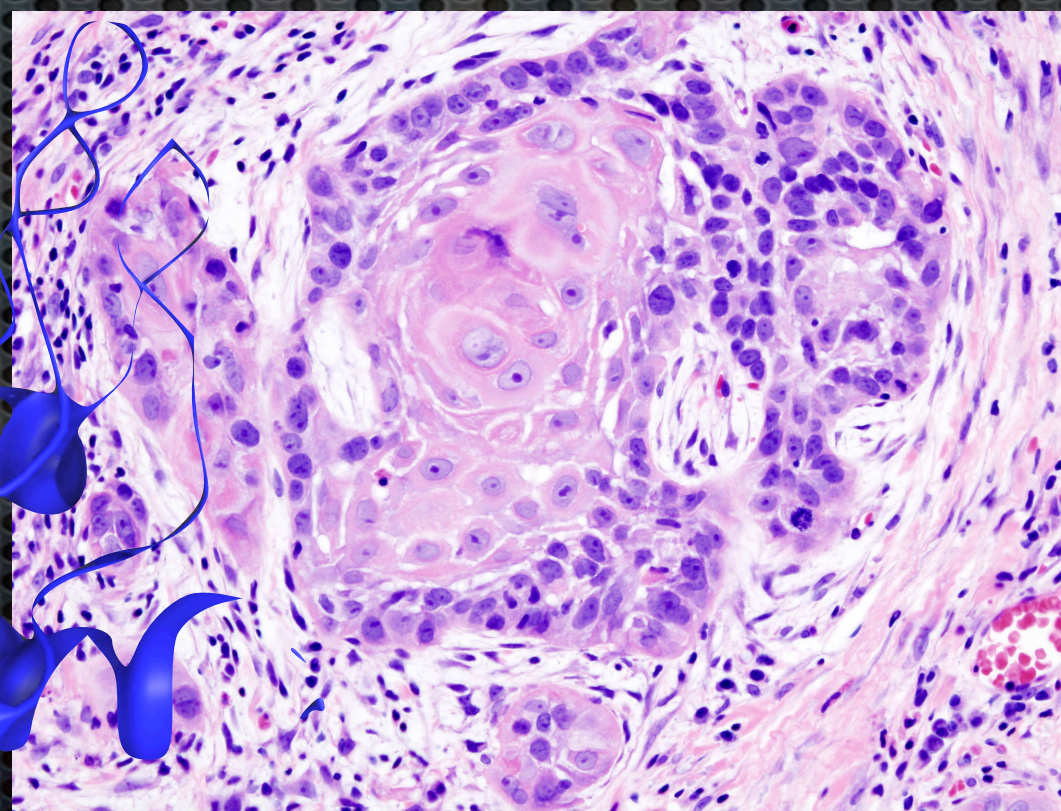
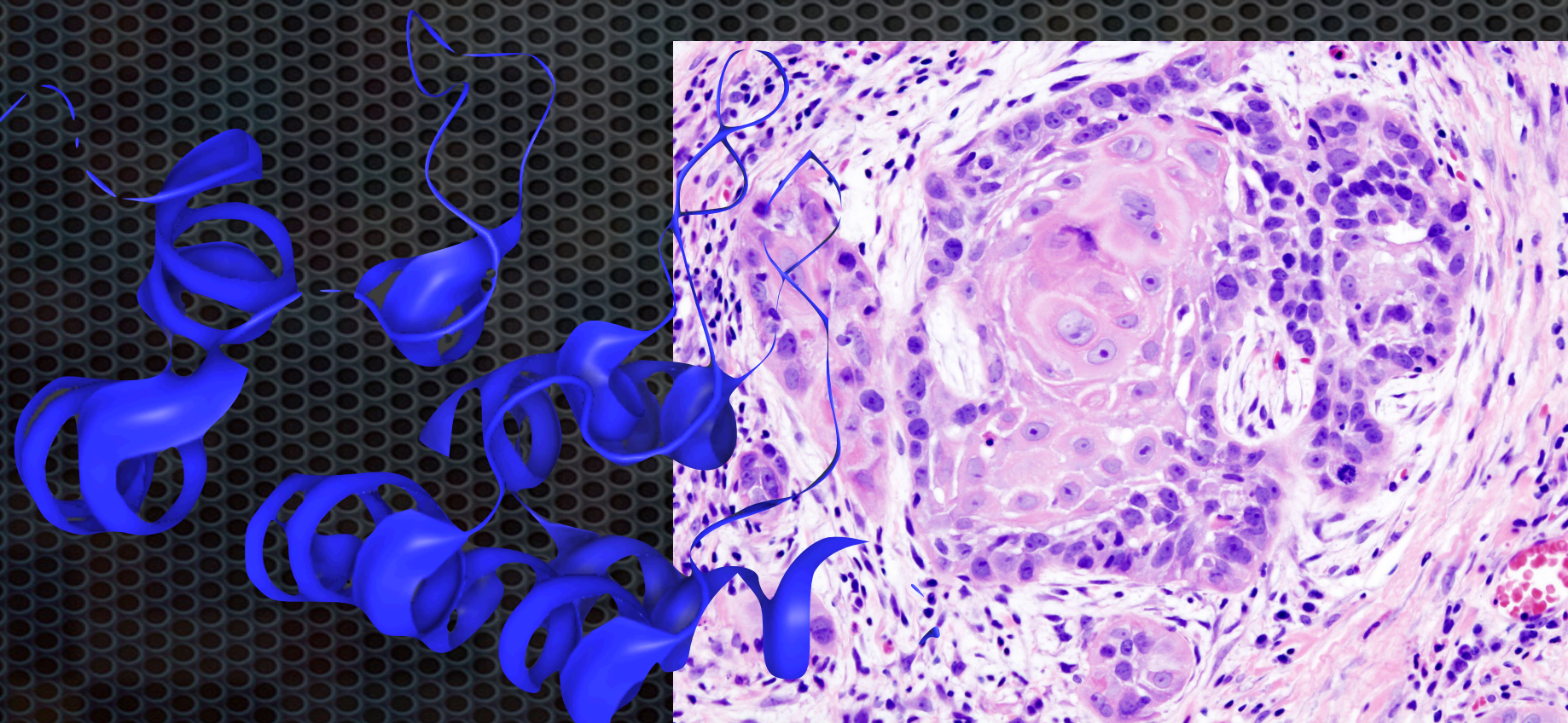


How can Bioinformatics help with the **42** of Life Sciences?

42



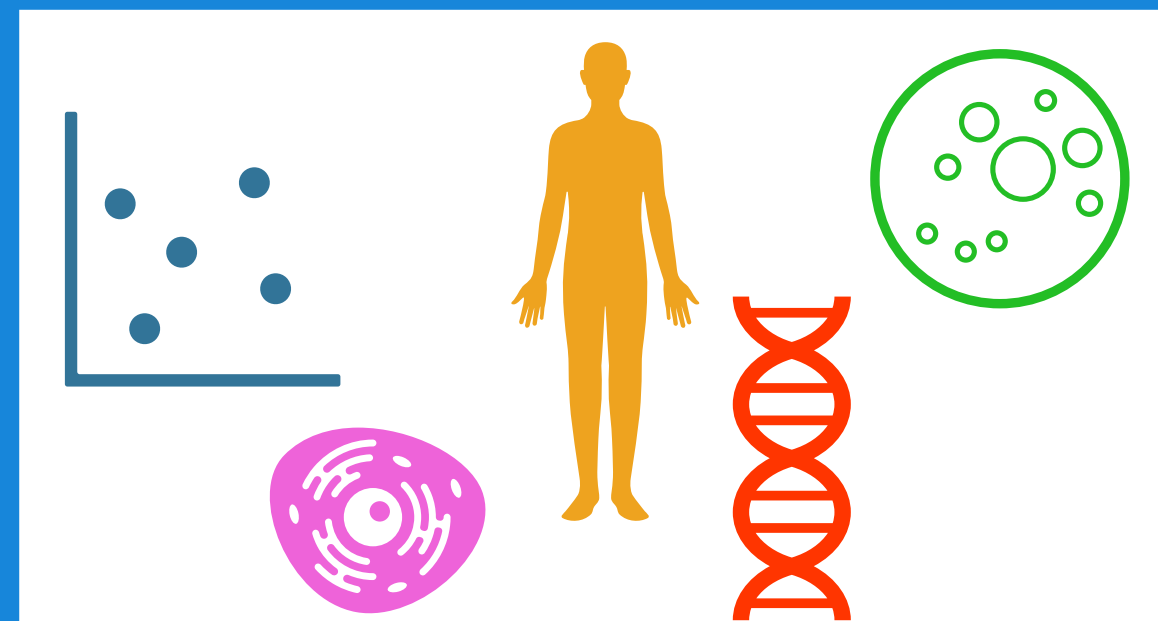
- ✦ Bioinformatics is "the science that uses the instruments of informatics to analyze biological data in order to formulate hypotheses about **life**." (Anna Tramontano)



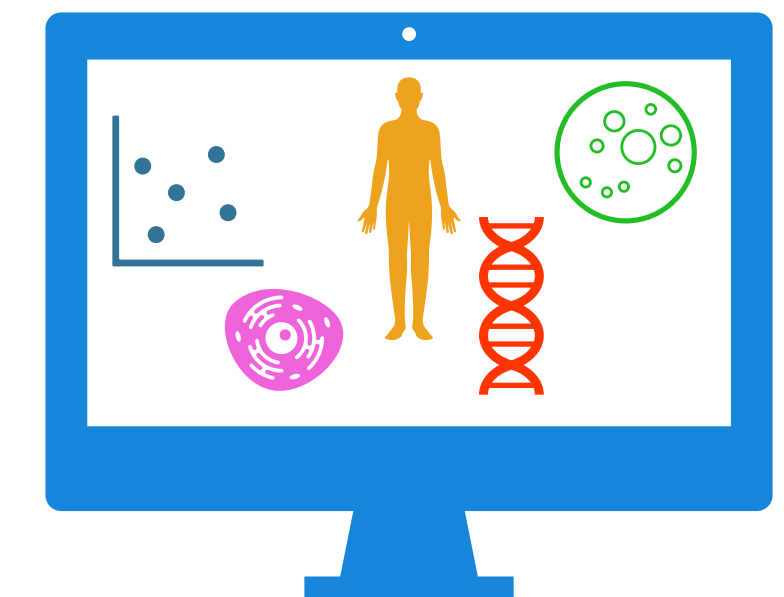
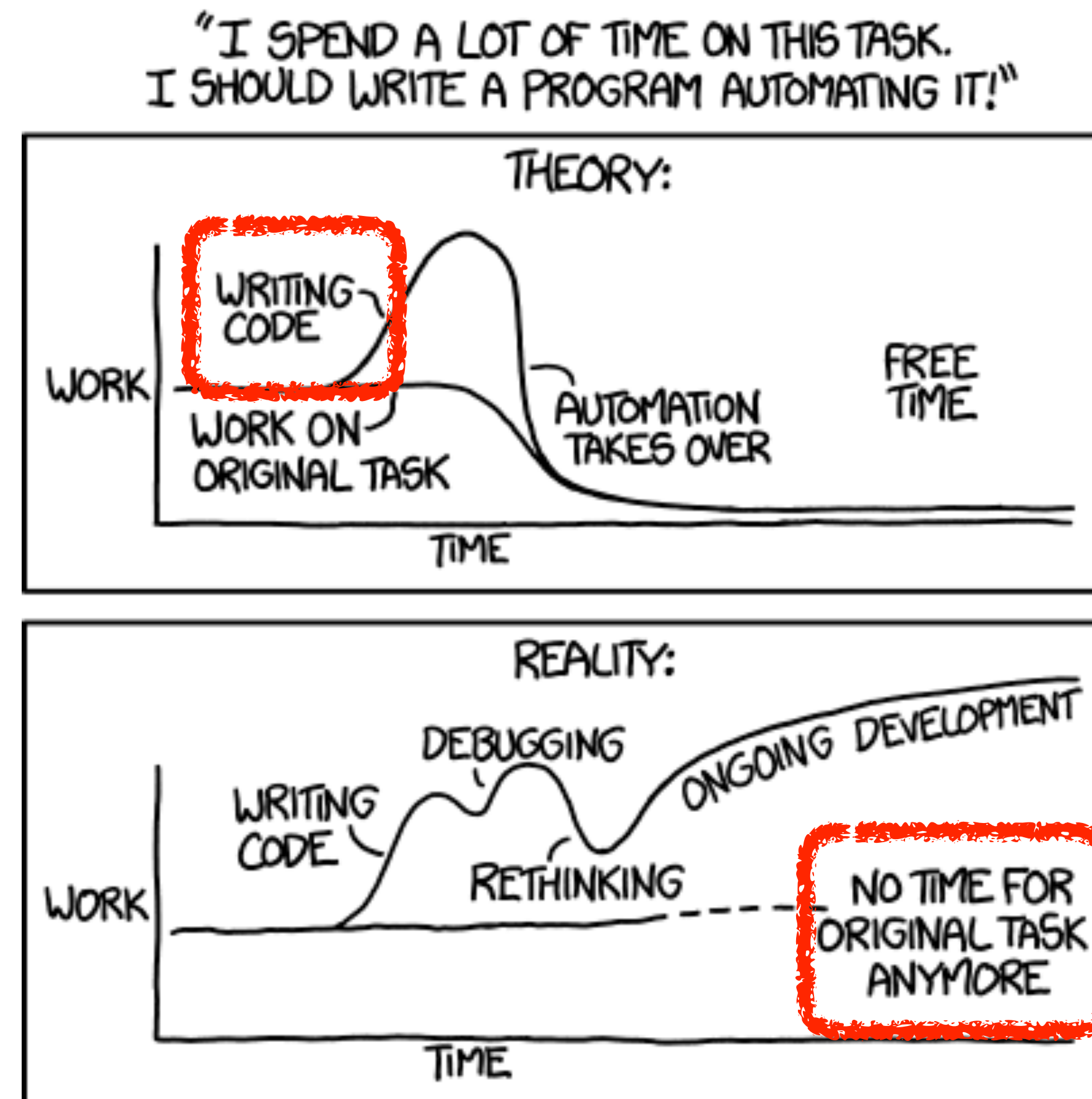
Sources: nextprot | opentreeoflife | wikipedia | MacKay et al., Cancer Cell (2017) | original photos

{bio_informatics_science}

```
if ($probeFile !~ /\w/) {  
  if ($pgx->{samples}->[0]->{paths}->{probeFile} =~ /\w\.\w/) {  
    $probeFile = $pgx->{samples}->[0]->{paths}->{probeFile} }  
  else {  
    my $seriesId = (split(':', $pgx->{samples}->[0]->{id}))[1];  
    my $arrayId = (split(':', $pgx->{samples}->[0]->{id}))[2];  
    $probeFile = $pgx->{parameters}->{genome}.'/'.'$seriesId.'/'.'$arrayId.'/'  
    probes.cn.tsv'  
  }  
  $probeFile =~ s/^\///;  
  $probeFile = $pgx->{config}->{paths}->{dir_array_base_path}.'/'.'$probeFile;  
}
```



{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

Bioinformatician

sufficient biological background

provides statistical, analysis methods

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

developer of informatics tools

may get rich

Bioinformatician

strong biological knowledge

provides hypothesis and / or dataset

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

expert **user** of informatics tools

may get a Nobel

Bioinformatician

sufficient biological background

provides statistical, analysis methods

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

developer of informatics tools

may get rich

flux

What do Bioinformaticians work on?

Hypothesis & Data Driven Approaches to Biological Topics

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

Data sets in tutorials



Data sets in the wild

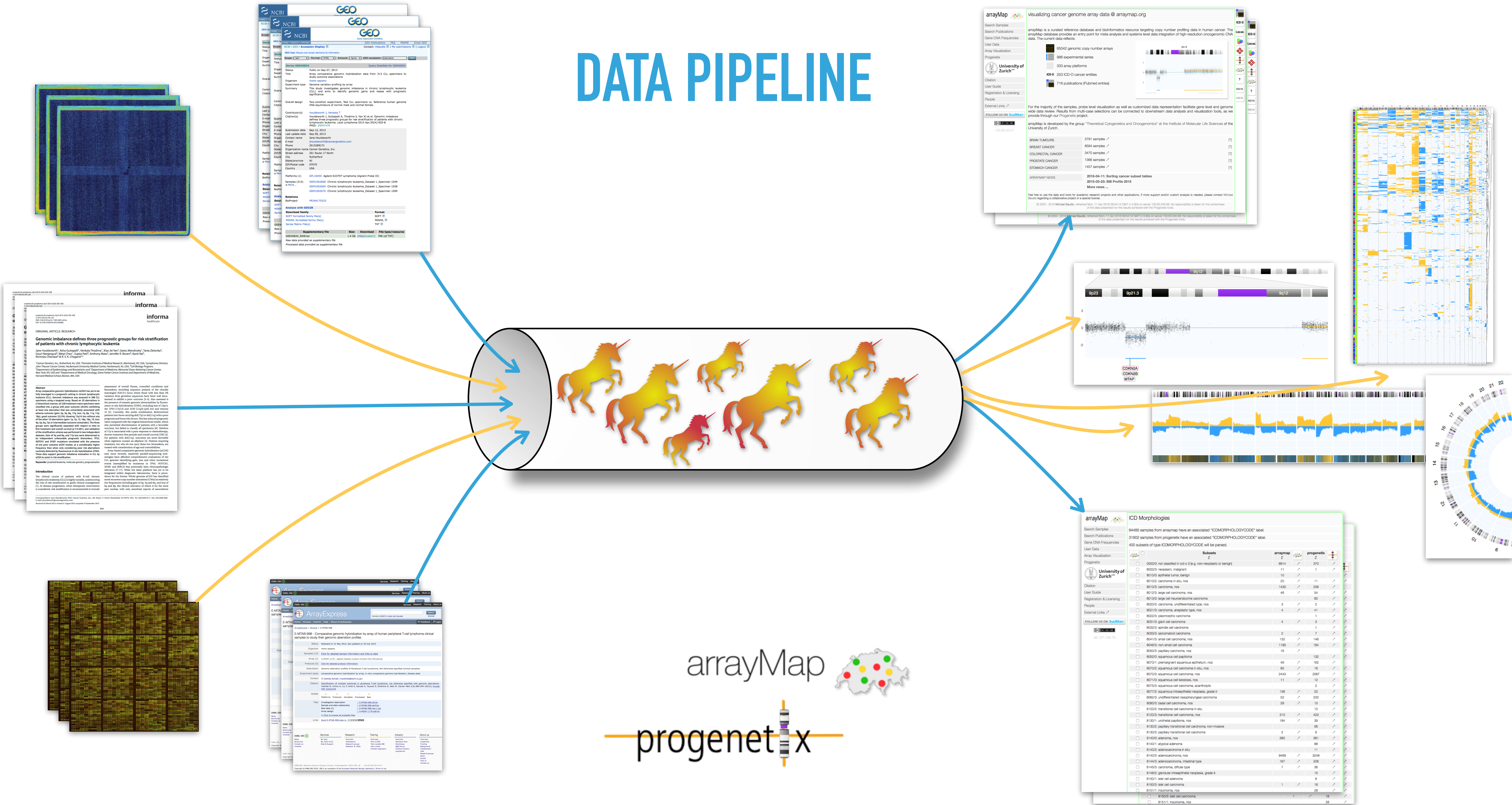




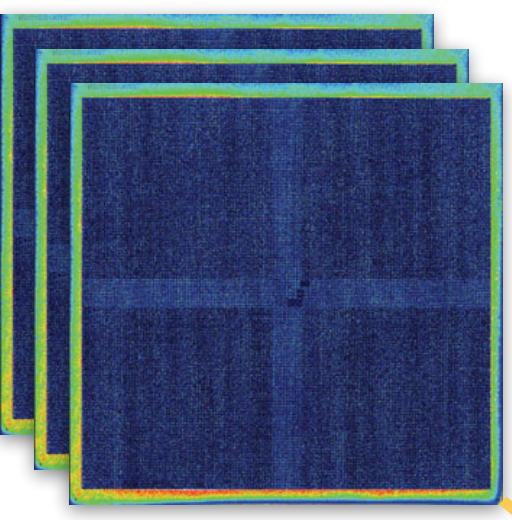
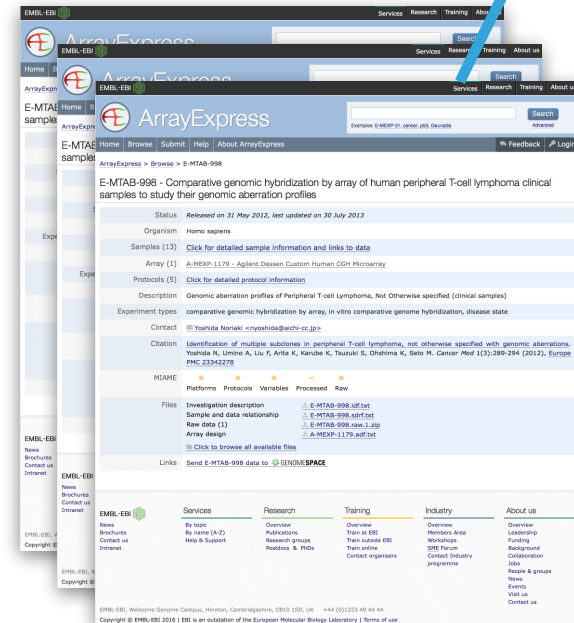
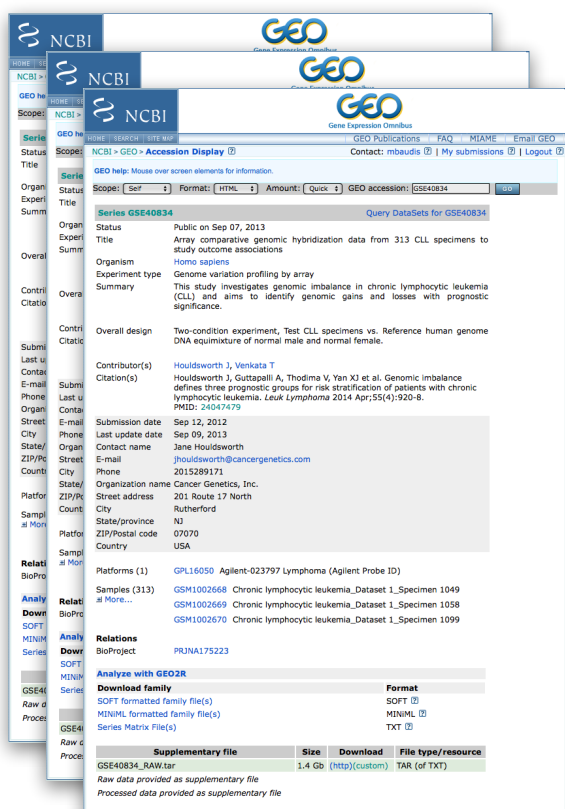
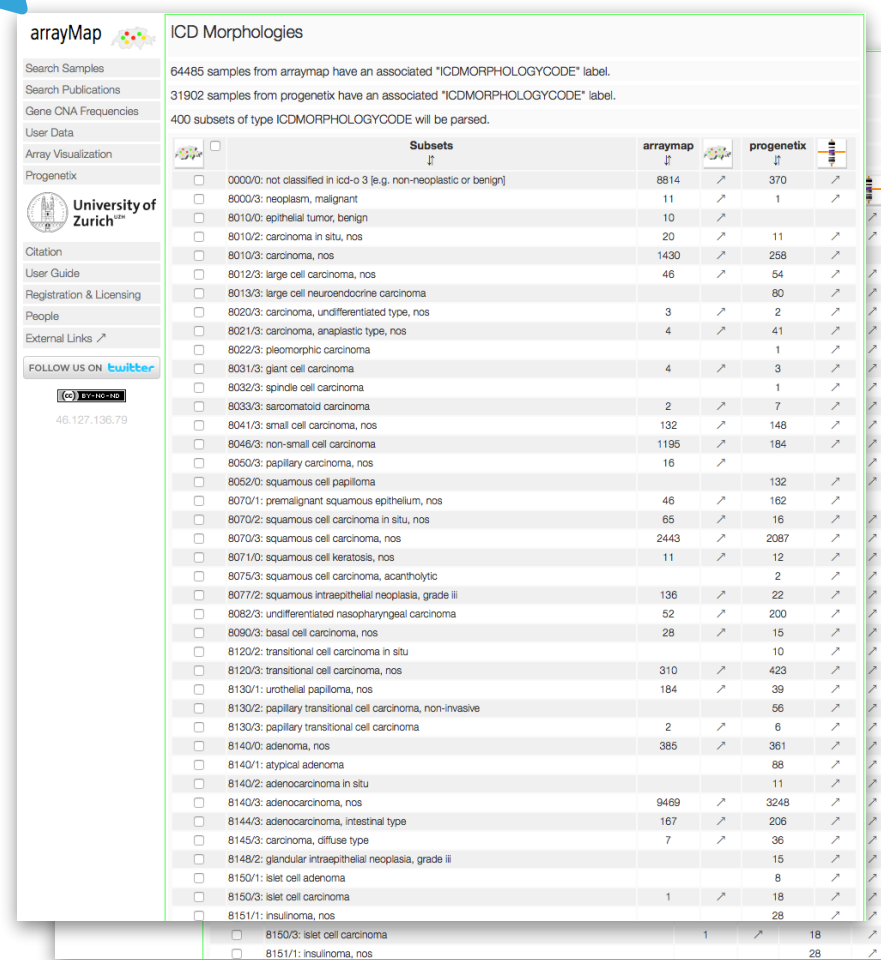
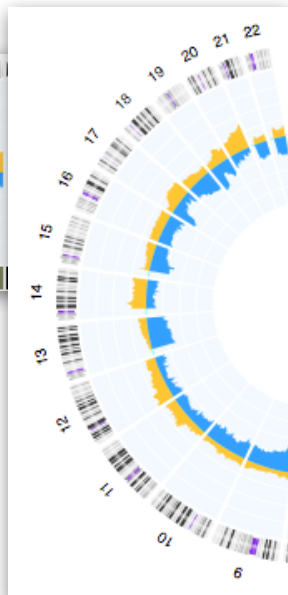
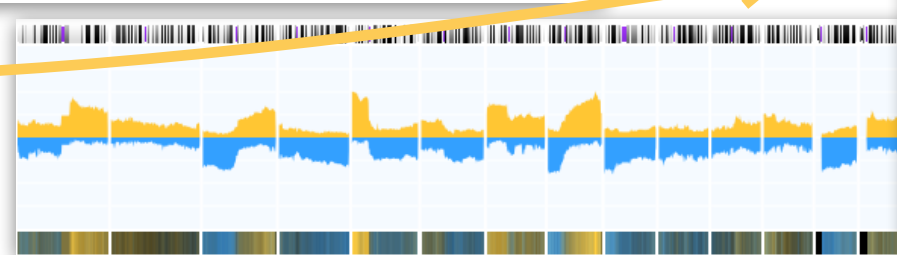
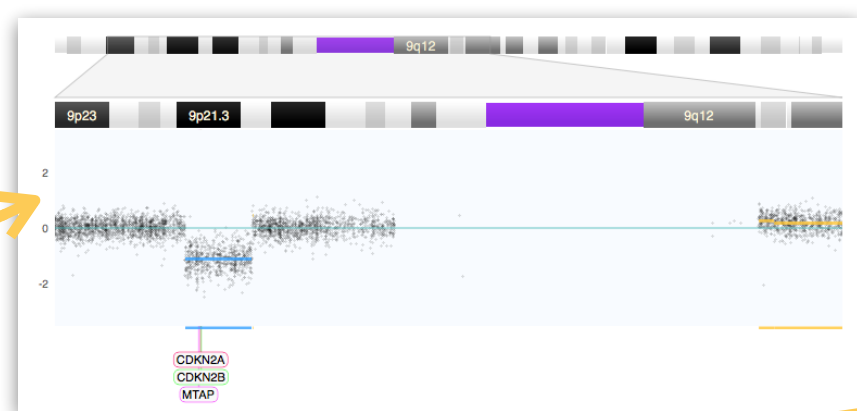
**FITTING
THE MODEL**

**CLEANING
THE DATA**

DATA PIPELINE



BIOINFORMATICS



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...)



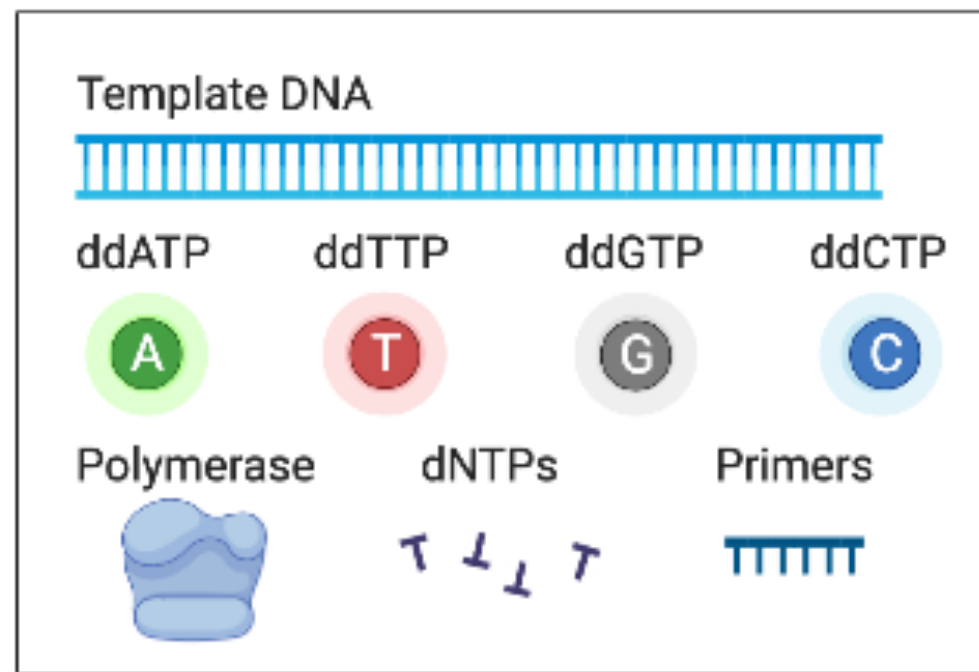
Bioinformatics: "High Dimensional" / -omics Data

- "-omics" data encompasses any data derived from DNA sequences
 - ➔ DNA, RNA (cDNA), proteins(fragment)s
 - ➔ germline or somatic
 - ➔ microorganisms (specific detection or screening - metagenomics)
- qualitative and quantitative measurements
 - ➔ **sequence variations**
 - ➔ **methylation status**
 - ➔ **DNA/RNA abundance, relative and absolute**

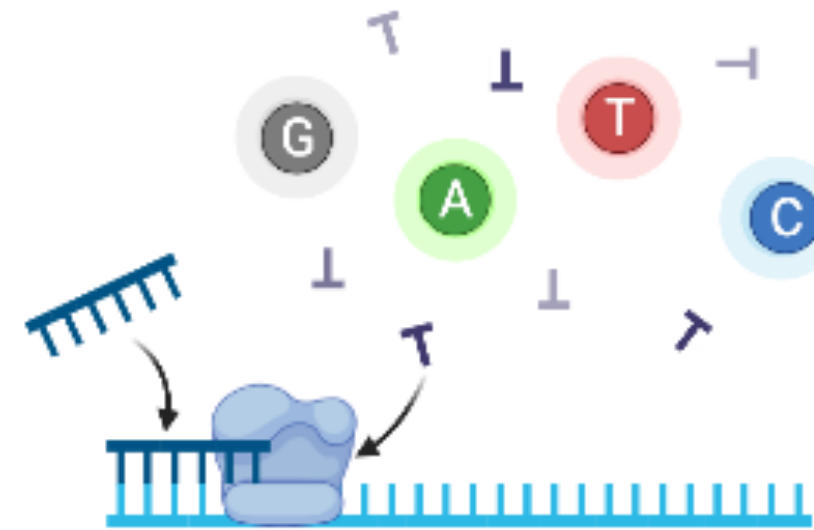
Genome Analysis

Sanger Sequencing

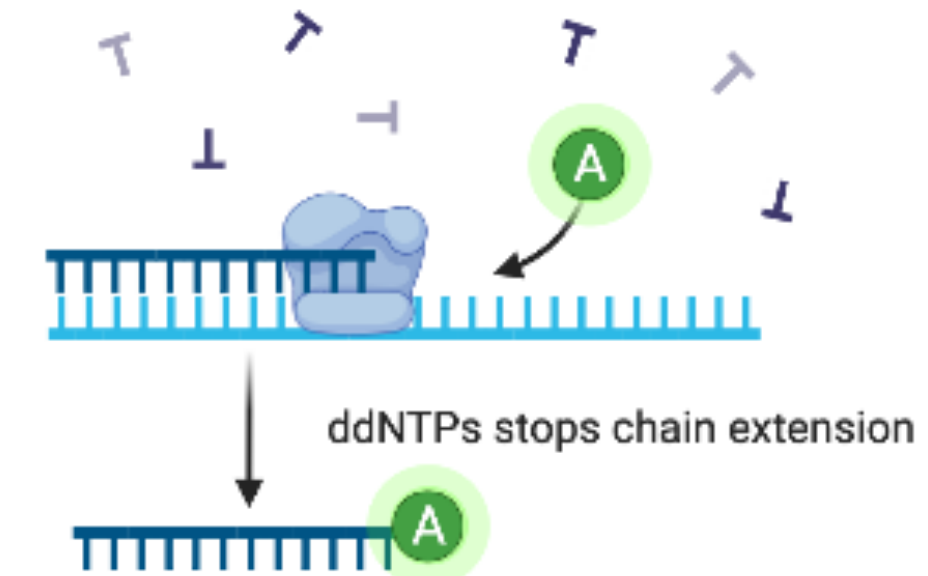
Reagents



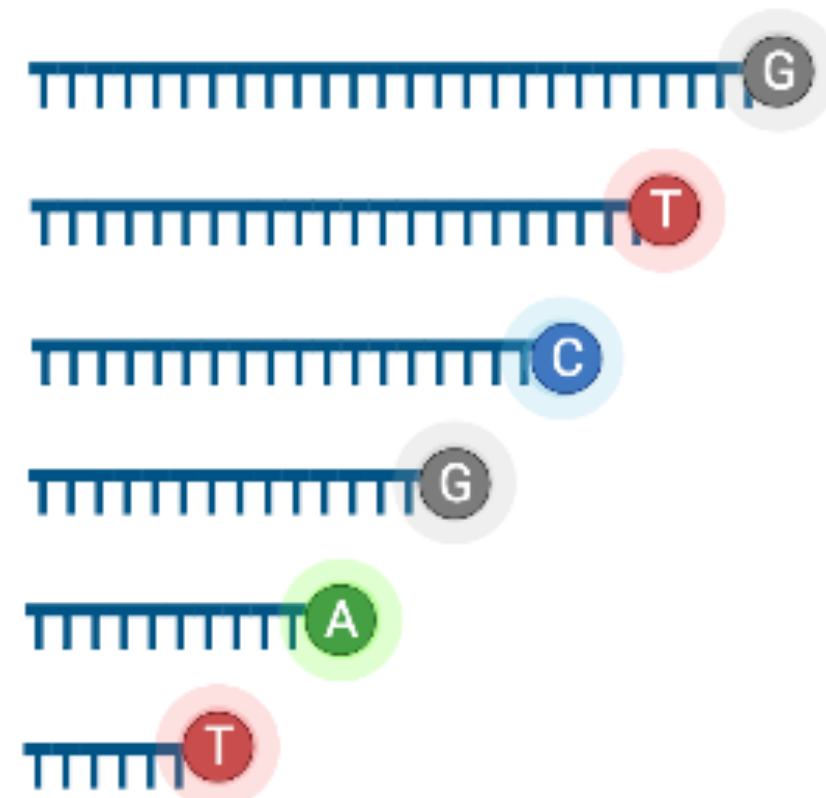
① Primer annealing and chain extension



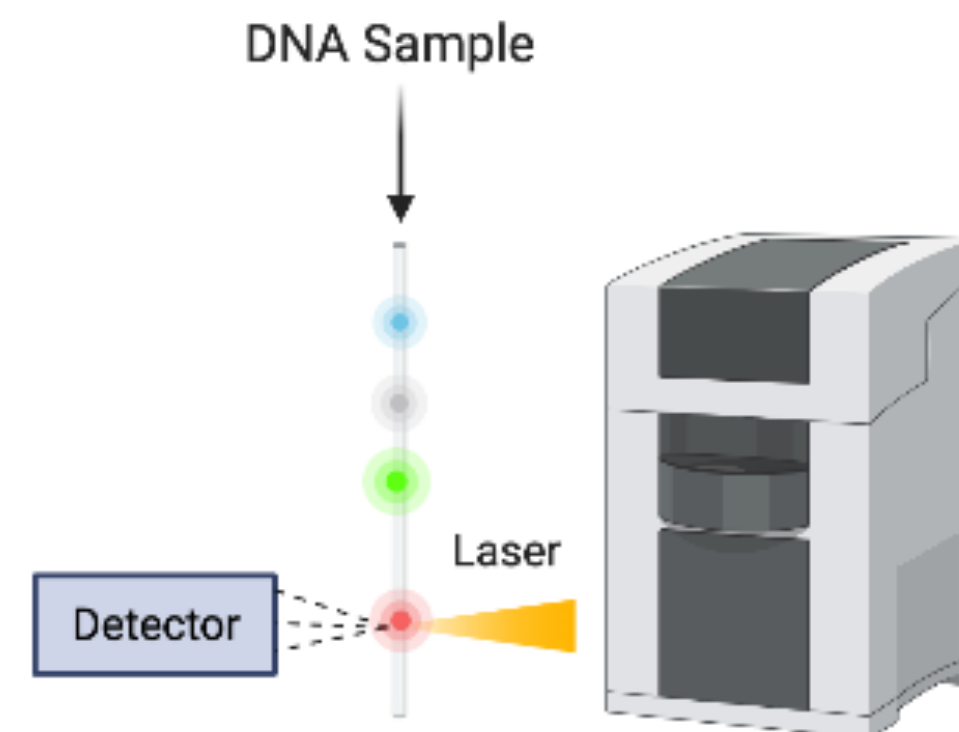
② ddNTP binding and chain termination



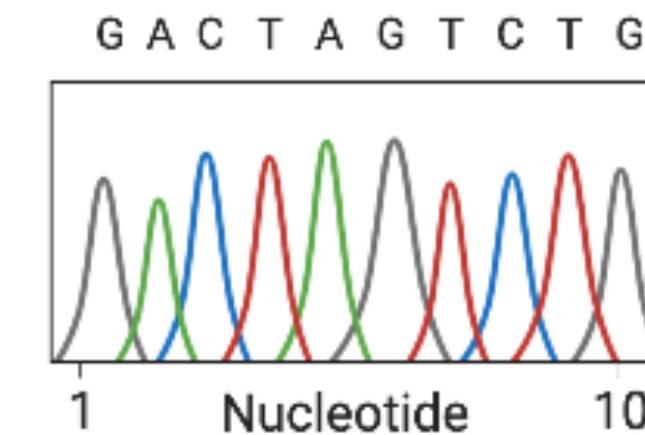
③ Fluorescently labelled DNA sample



④ Capillary gel electrophoresis and fluorescence detection



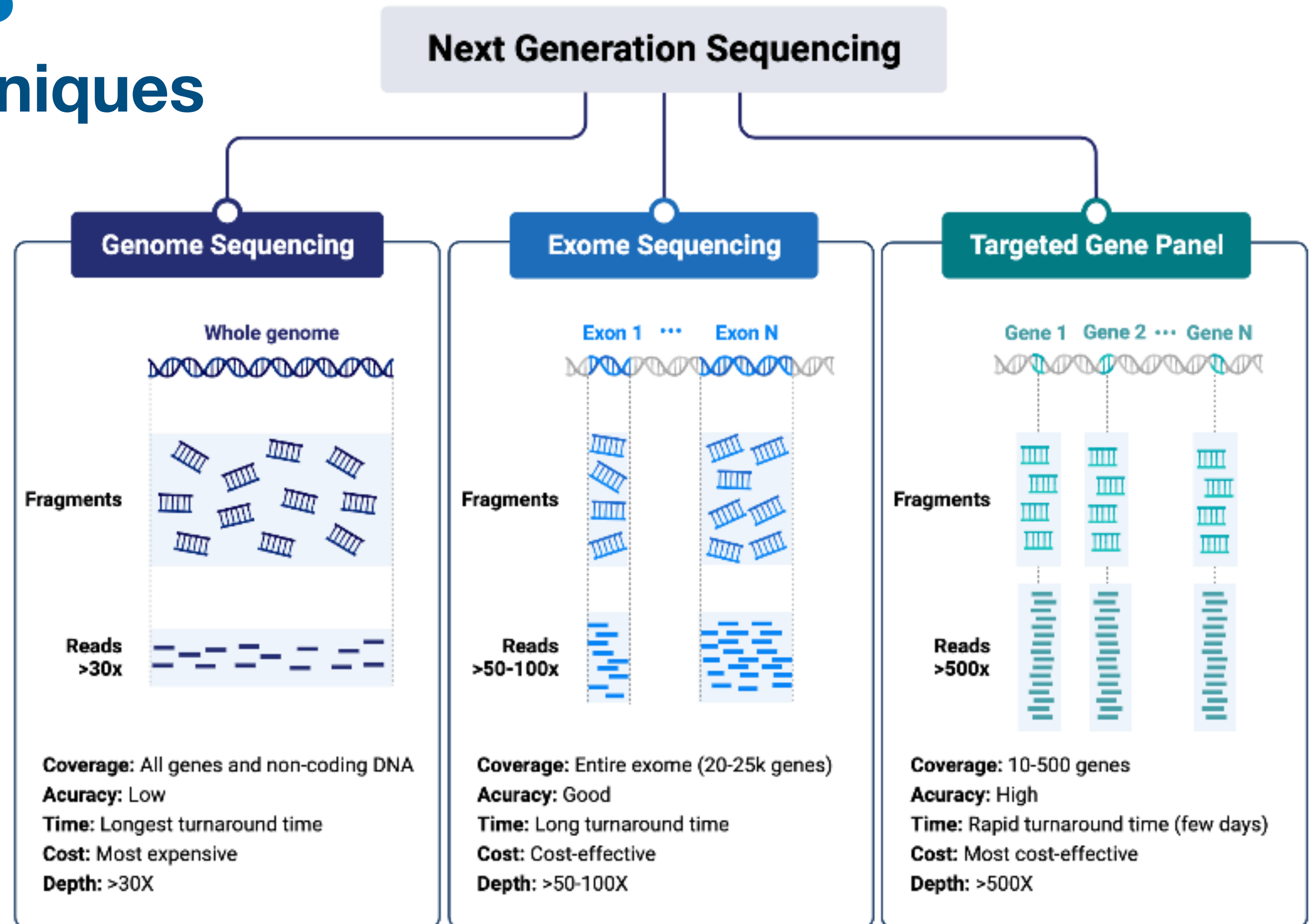
⑤ Sequence analysis and reconstruction



Genome Analysis

Comparison of NGS Techniques

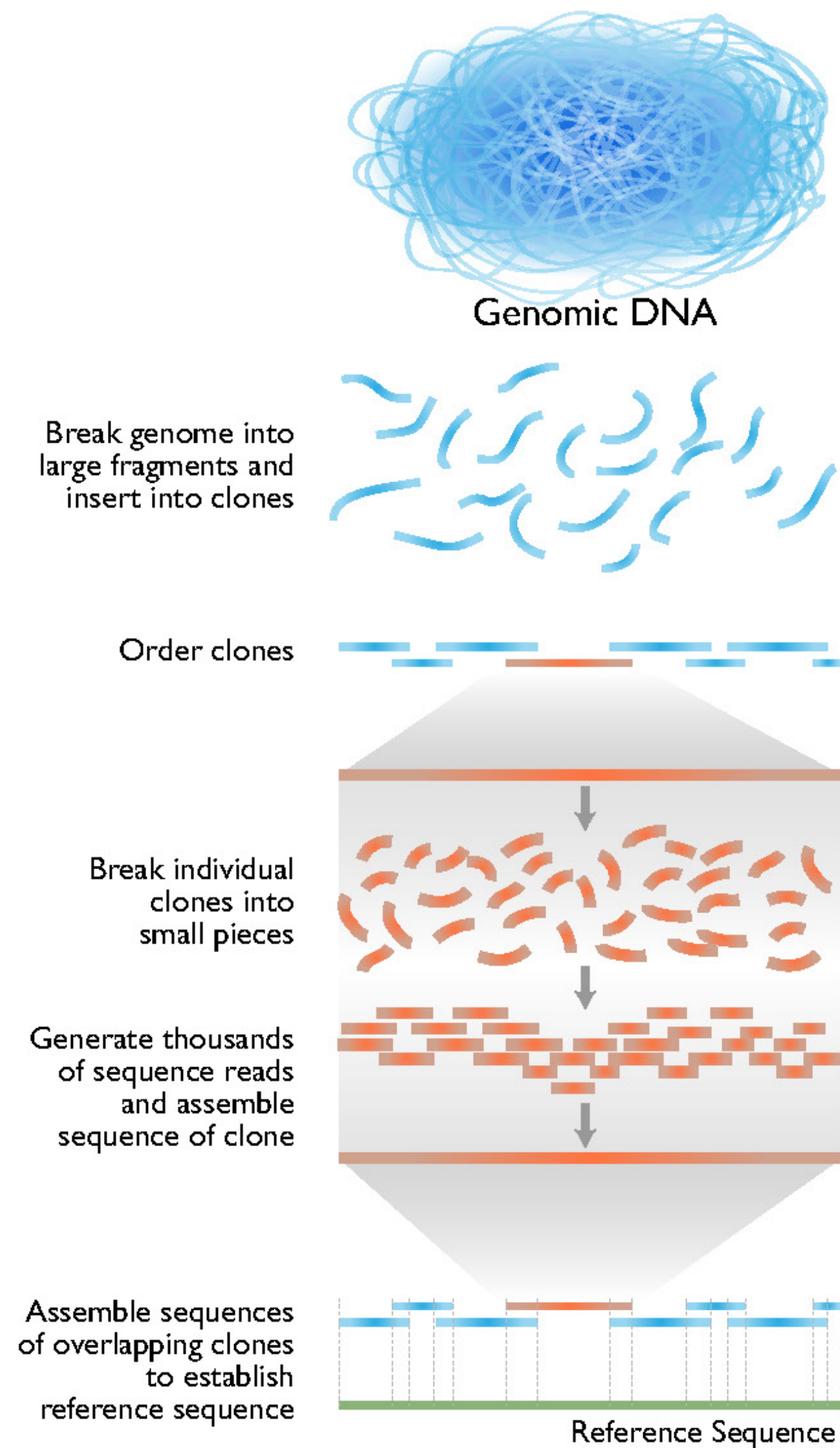
- current NGS technologies present various compromises between coverage, precision, speed & accuracy
- the **bottleneck** in WGS is not in read production costs but in **interpretation** and storage
- widespread use of incomplete genome profiling technologies avoids problems of hard to interpret or not "actionable" results but limits the expansion of our "knowledge horizon"



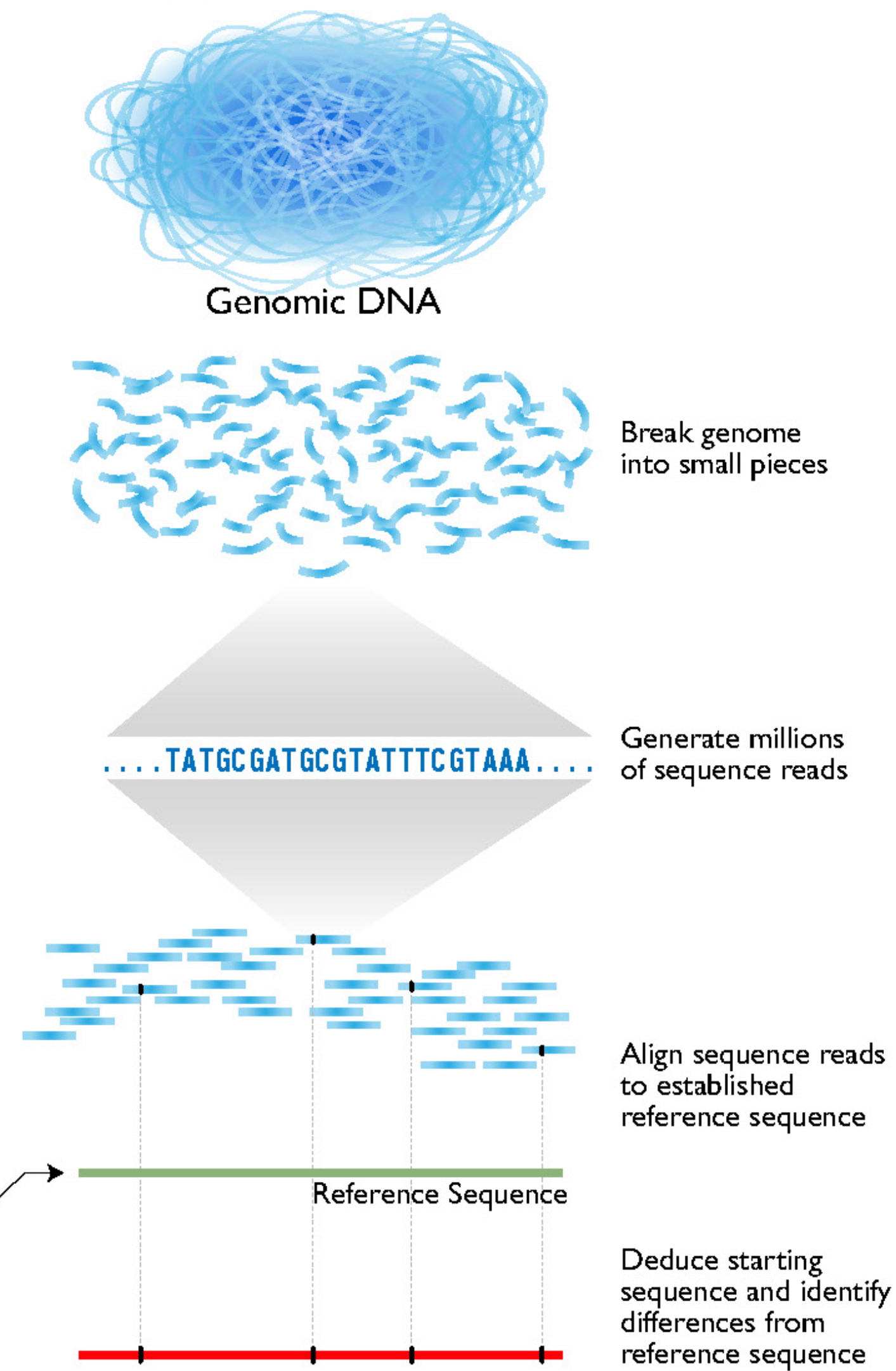
Template adapted from: Dr. Roshini Abraham
Clinical Immunologist at Nationwide Children's Hospital

Human Genome Sequencing

Generating a Reference Genome Sequence (e.g., Human Genome Project)



Generating a Person's Genome Sequence (e.g., Circa ~2016)



Alignment

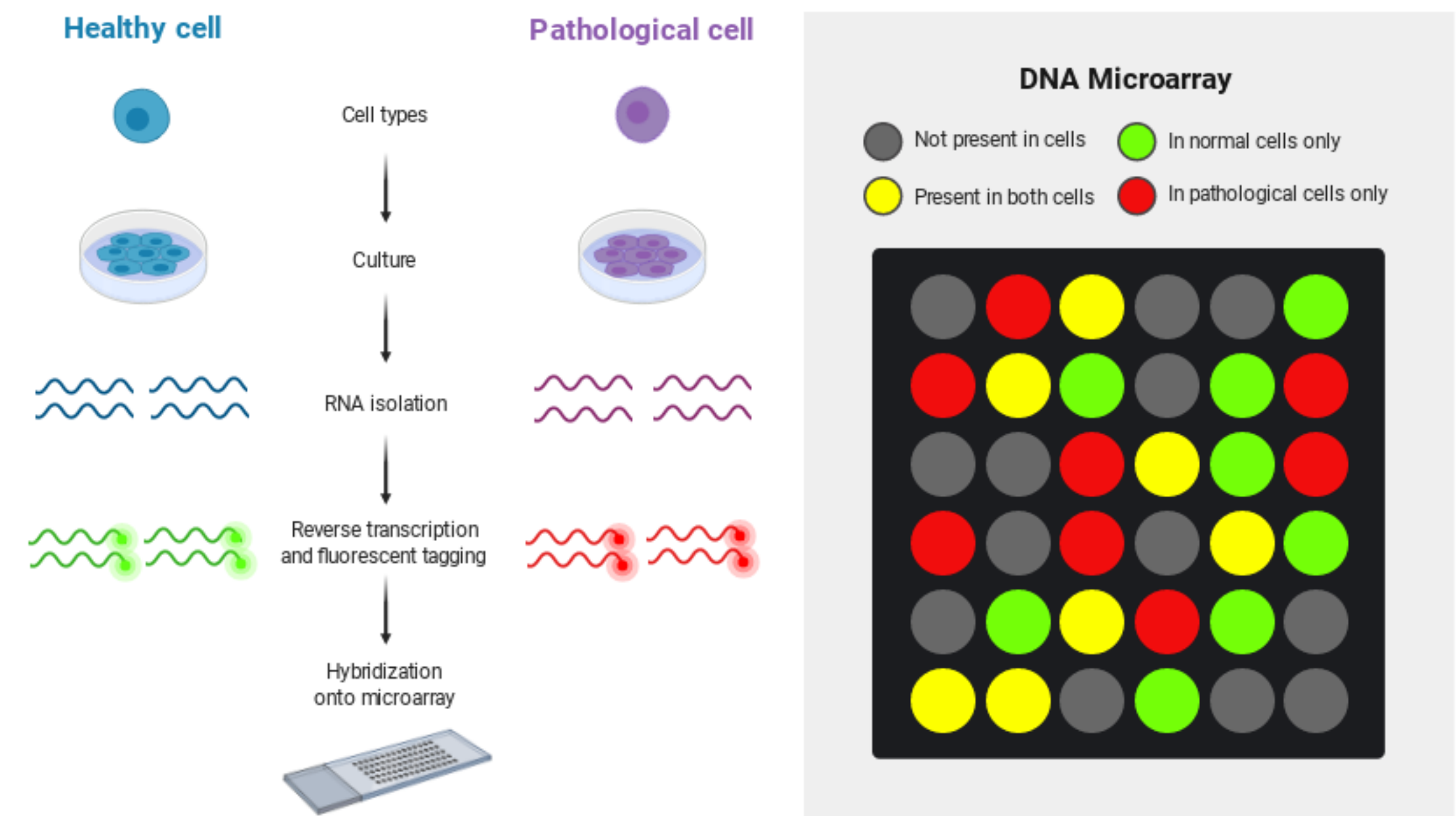
Variant Calling

Array Comparative Genomic Hybridization

Molecular Hybridization Technology for Genomic Imbalance Screening

- Molecular-cytogenetic technique to identify genomic copy number variations (CNV/ CNA) for given sequences (multi-kb to Mb)
- based on *in situ* suppression **hybridization** of labeled **genomic** tumor and reference DNA against spotted or synthesized DNA clones or oligonucleotides
- analysis of relative fluorescence ratio allows **semi-quantitative copy number** read-out
- **direct** attribution of involved target genes through known sequence content

- Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, Dohner H, Cremer T, Lichter P. Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer*. 1997; 4 (20):399-407.
- Albertson DG, Pinkel D. Genomic microarrays in human genetic disease and cancer. *Hum Mol Genet*. 2003; 12 Spec No 2 R145-52.
- Bignell GR, Huang J, Greshock J, Watt S, Butler A, West S, Grigorova M, Jones KW, Wei W, Stratton MR, Futreal PA, Weber B, Shaperro MH, Wooster R. High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Res*. 2004; 14 (2):287-295.



Reprinted from "DNA Microarray", August 2019, retrieved from <https://app.biorender.com/biorender-templates/figures/all/t-5e41b61b0dd2690088b72481-dna-microarray> © 2022 by BioRender.

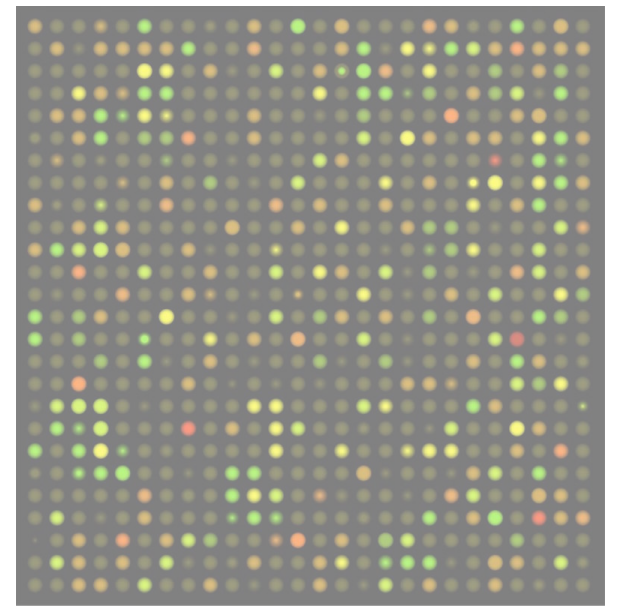
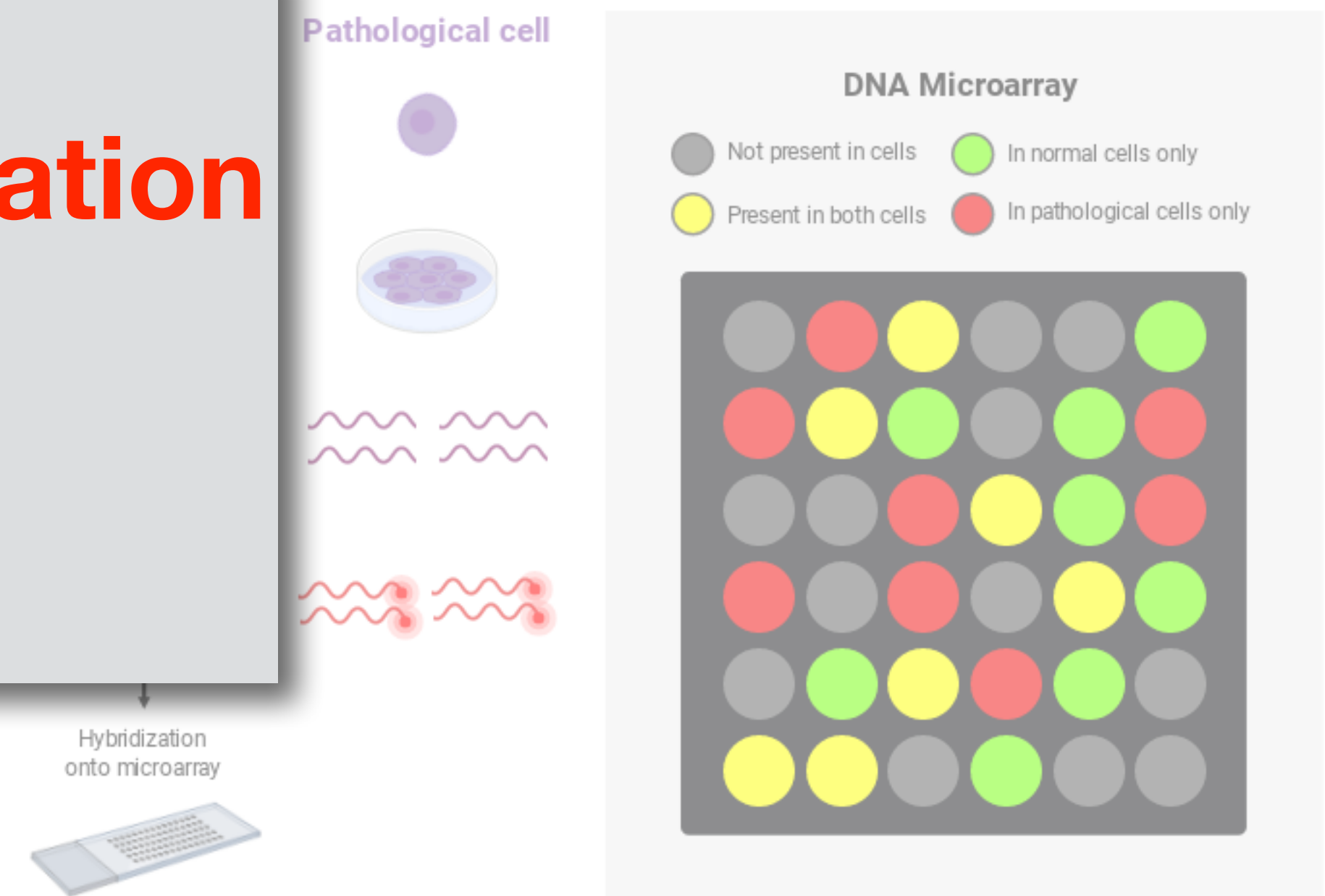
Array Comparative Genomic Hybridization

Molecular Hybridization Technology for Genomic Imbalance Screening

- Molecular-cytogenetic technique to identify genomic copy number variations (CNV/ CNA) for given sequences
- based on *in situ* suppression of **hybridization** of labeled test and reference DNA against synthesized DNA clones on a chip
- analysis of relative fluorescence intensities for **semi-quantitative copy number** read out
- **direct** attribution of involved target genes through known sequence content

- Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, Dohner H, Cremer T, Lichter P. Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer*. 1997; 4 (20):399-407.
- Albertson DG, Pinkel D. Genomic microarrays in human genetic disease and cancer. *Hum Mol Genet*. 2003; 12 Spec No 2 R145-52.
- Bignell GR, Huang J, Greshock J, Watt S, Butler A, West S, Grigorova M, Jones KW, Wei W, Stratton MR, Futreal PA, Weber B, Shaperro MH, Wooster R. High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Res*. 2004; 14 (2):287-295.

**Mapping
Background Calibration
Segmentation
CNV Calling
(Liftover)**



Reprinted from "DNA Microarray", August 2019, retrieved from <https://app.biorender.com/biorender-templates/figures/all/t-5e41b61b0dd2690088b72481-dna-microarray> © 2022 by BioRender.

Genome Lifter

Moving between genome editions

- different **reference genome editions** lead to shifting positions of defined elements such as genes
- local regions are frequently stable between genome editions
- shifts from change in regional lengths are defined in "chain files"
- chain files serve as guides for positional **remapping** using ***lifter*** methods

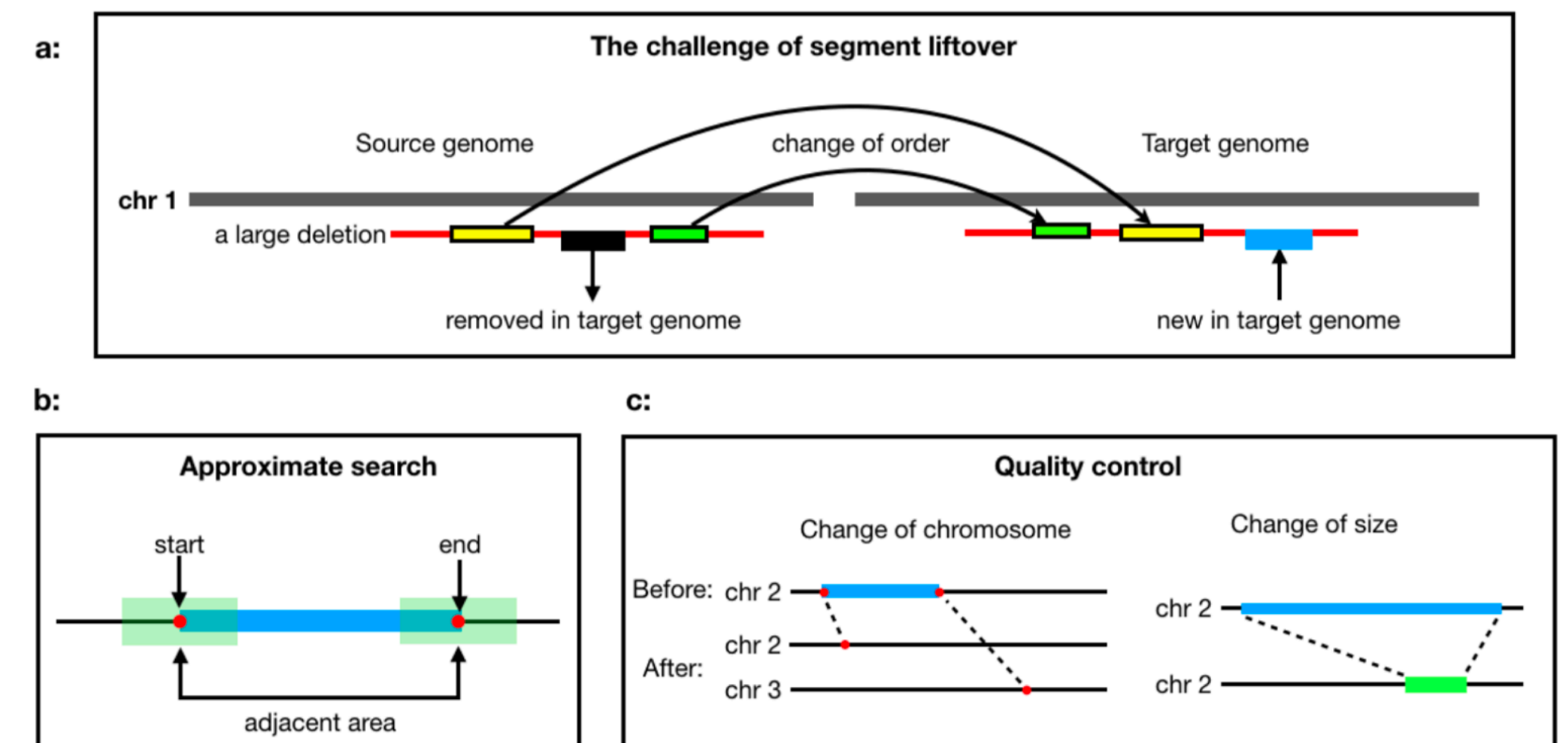
SOFTWARE TOOL ARTICLE

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

Bo Gao ^{1,2}, Qingyao Huang^{1,2}, Michael Baudis ^{1,2}

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

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

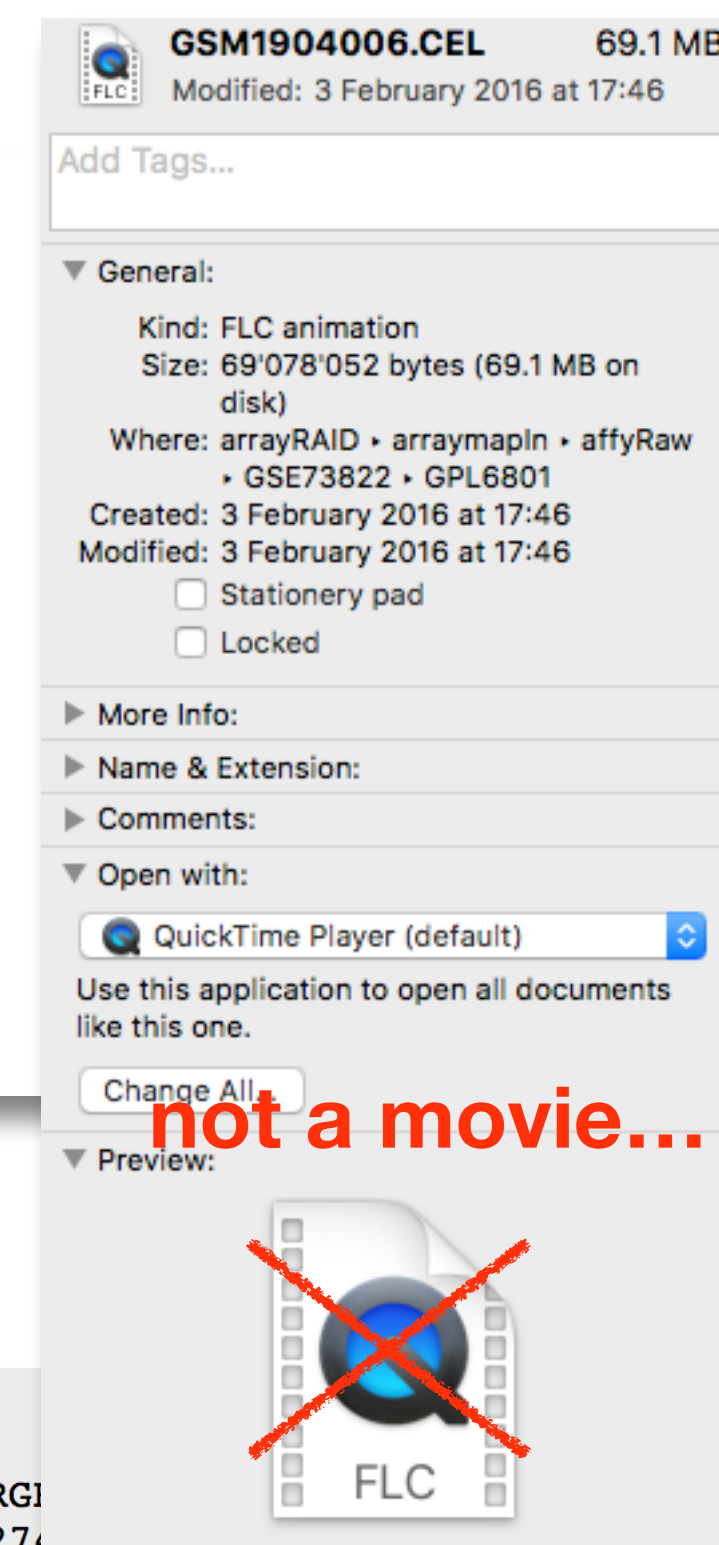


Bioinformatics: File Formats, Ontologies & APIs

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

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

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



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

BED file example

File Formats: VCF

Genomic variant storage standard

- The 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

The Variant Call Format Specification
VCFv4.5 and BCFv2.2

9 Oct 2024

The master version of this document can be found at <https://github.com/samtools/hts-specs>.
This printing is version 84d0d2b from that repository, last modified on the date shown above.

```
##fileformat=VCFv4.5
##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 NAO0001 NAO0002
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
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
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
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
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
```


The VCF file format

Standard for genomic variant representation

Example

VCF header		<pre>##fileformat=VCFv4.0 ##fileDate=20100707 ##source=VCFtools ##reference=NCBI36 ##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele"> ##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership"> ##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype"> ##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)"> ##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)"> ##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth"> ##ALT=<ID=DEL,Description="Deletion"> ##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant"> ##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant"></pre>										Mandatory header lines
		<pre>#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2</pre>										Optional header lines (meta-data about the annotations in the VCF body)
Body		1	1	.	ACG	A,AT	.	PASS	.	GT:DP	1/2:13	0/0:29
		1	2	rs1	C	T,CT	.	PASS	H2;AA=T	GT:GQ	0 1:100	2/2:70
		1	5	.	A	G	.	PASS	.	GT:GQ	1 0:77	1/1:95
		1	100	.	T		.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20

Deletion (points to in ALT)

SNP (points to A,AT in ALT)

Large SV (points to SVTYPE=DEL;END=300 in INFO)

Insertion (points to T,CT in ALT)

Other event (points to H2;AA=T in INFO)

Reference alleles (GT=0) (points to 0/0:29 in SAMPLE2)

Alternate alleles (GT>0 is an index to the ALT column) (points to 1/1:12:3 in SAMPLE2)

Phased data (G and C above are on the same chromosome) (points to 0|1:100 in SAMPLE1)

	CS1	CS2	CS3	CS4	CS5	...
V1		C	C			
V2		C		C	C	
V3	C					
V4			C	C		
V5					C	
...						

VariantSet (points to the entire grid)

Variant (points to a single row, e.g., V3)

CallSet (points to a single column, e.g., CS2)

Call (points to a single cell, e.g., V5, CS5)

Variant
Call
Format

- stores the results of a single or multiple interpretations of genome sequencing datasets, in comparison to a reference genome
- standard format for file-based storage of human genome variants

Bioinformatics: File Formats, Ontologies & APIs

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

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

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



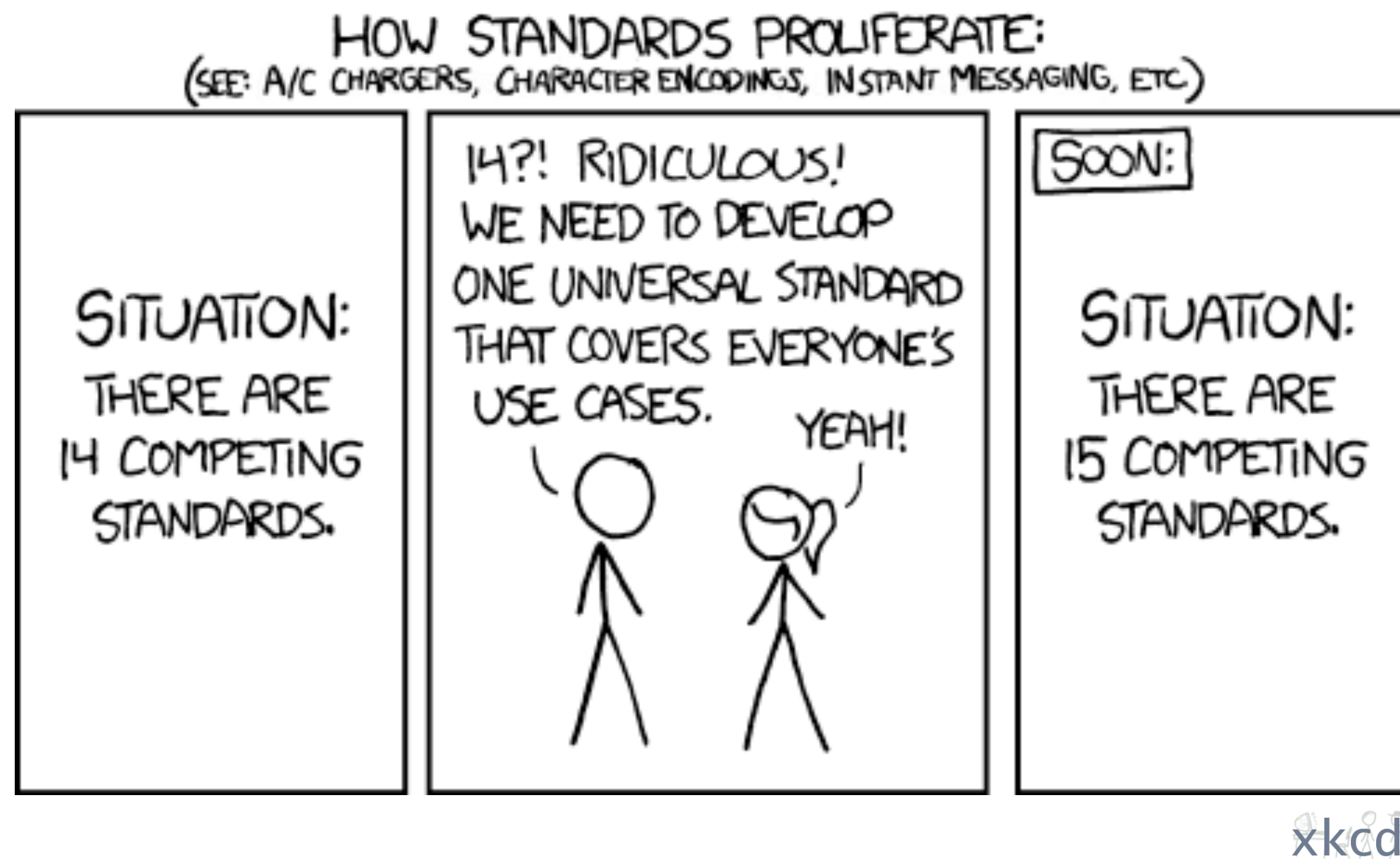
```
id: GO:0000118
name: histone deacetylase complex
namespace: cellular_component
def: "A protein complex that possesses histone deacetylase activity." [GOC:mah]
comment: Note that this term represents a location, not a function; the activities possessed by this complex is mentioned in the
definition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

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



Standardized Data

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



```
"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"
}
```


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 **accessions**, frequently in form of **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 ...)

```
"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"
}
```


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 coding access

- other international

- ▶ ISO (ISO 8601 time & period, ISO 3166 country codes ...)
- ▶ IETF (GeoJSON ...)
- ▶ W3C (CURIE ...)

CURIE

Compact uniform resource identifiers, or CURIEs, are an important formalism for referencing biomedical entities. CURIEs are URIs in which the general part of the URI has been replaced with a compact, registered prefix.

<https://cthoyt.com/2021/09/14/curies.html>

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


Genomic Copy Number Variation Nomenclatures in Computational (File/Schema) Formats



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

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



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

labelSeg: segment annotation for tumor copy number alteration profiles

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

Dates and Times

There is only the ISO way!

- Y = year
- M = month
- D = day
- H = hour
- M = minute
- S = second
- .S = decimal fraction of a second

The specification of a time point is given through the concatenation of

- a date in YYYY-MM-DD
- the designator "T" indicating a following time description
- the time of day in HH:MM:SS.SSS form, where "SSS" represents a decimal fraction of a second
- a time zone offset in relation to UTC

Examples

- year (YYYY)
 - 2015
 - Time points with year granularity are both common for obfuscated personal data as well as technical metadata (e.g. year of publication of an analysis).
- date (e.g. date of birth) in YYYY-MM-DD
 - 2015-02-10
 - This represents the standard way of representing a specific day, e.g. a date of birth.
- time stamp in milliseconds in YYYY-MM-DDTHH:MM:SS.SSS
 - 2015-02-10T00:03:42.123Z
 - Timepoints with millisecond granularity are typical use cases for timing computer generated entries, e.g. the time of a record's update ("updateTime").



<https://xkcd.com/1179/>

Field	Flags	Type	Limits	Description
id		string	0..1	The 'logical' identifier of the Entity in the system of record, e.g. a UUID. This 'id' is unique within a given system, but may or may not be globally unique outside the system. It is used within a system to reference an object from another.
name		string	0..1	A primary name for the entity.
description		string	0..1	A free-text description of the Entity.
aliases		string	0..m	Alternative name(s) for the Entity.
extensions		Extension	0..m	A list of extensions to the Entity, that allow for capture of information not directly supported by elements defined in the model.
digest		string	0..1	A sha512t24u digest created using the VRS Computed Identifier algorithm.
expressions		Expression	0..m	
type		string	1..1	MUST be "CopyNumberCount"
location		iriReference Location	1..1	The location of the subject of the copy count.

```
{
  "$schema": "beaconRequestBody.json",
  "meta": {
    "apiVersion": "2.0",
    "requestedSchemas": [
      {
        "entityType": "genomicVariation",
        "schema": "https://raw.githubusercontent.com/ga4gh-beacon/beacon-requests/main/schemas/genomicVariation.json"
      }
    ]
  },
  "query": {
    "requestParameters": {
      "g_variant": {
        "referenceName": "NC_000017.11",
        "start": [7577120],
        "referenceBases": "G",
        "alternateBases": "A"
      }
    },
    "requestedGranularity": "record",
    "pagination": {
      "skip": 0,
      "limit": 5
    }
  }
}
```


Bioinformatics: File Formats, Ontologies & APIs

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

Bioinformatics: File Formats, Ontologies & APIs

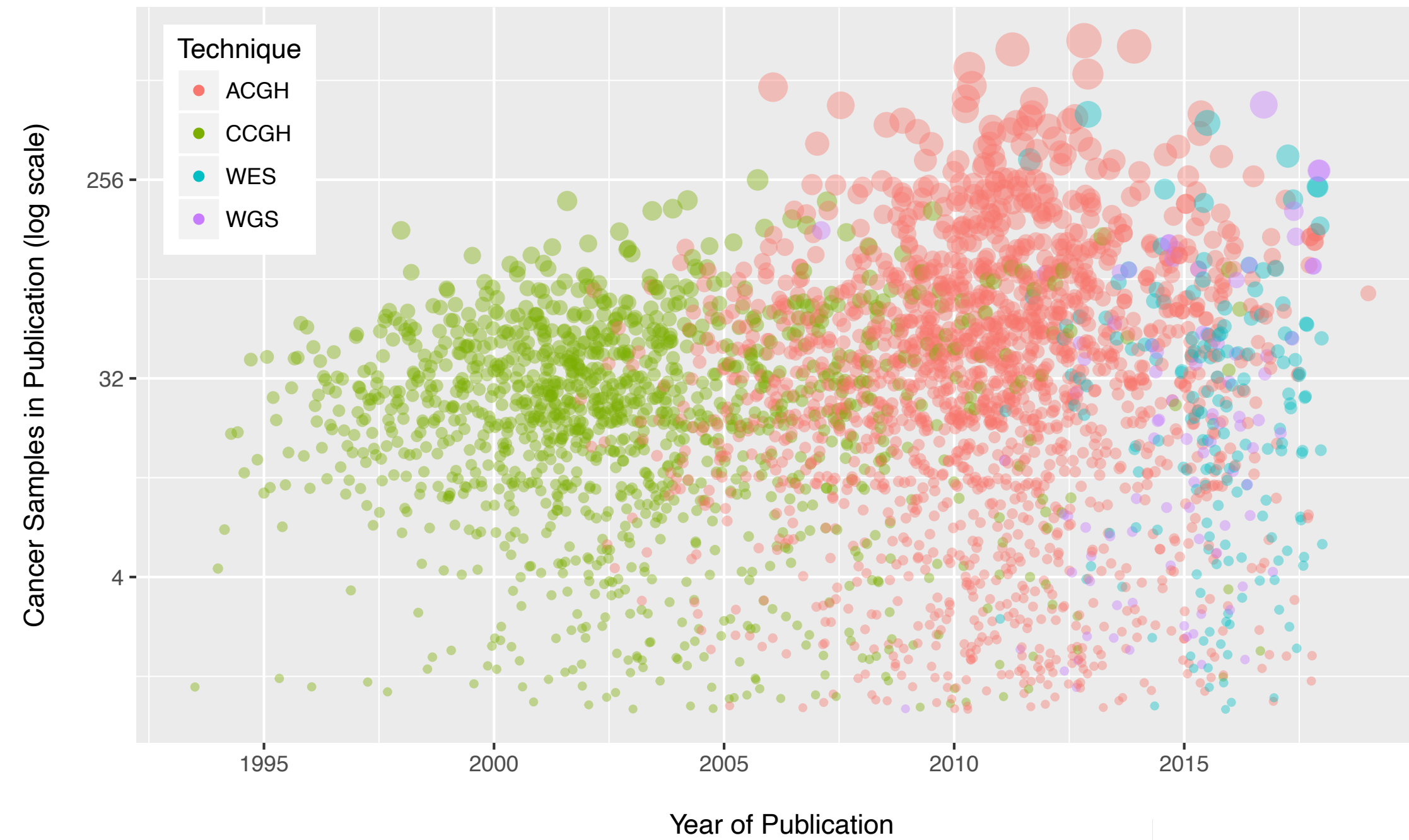
```
{
  "$schema": "https://raw.githubusercontent.com/ga4gh-beacon/
beacon-v2/main/framework/json/requests/beaconRequestBody.json",
  "meta": {
    "apiVersion": "2.0",
    "requestedSchemas": [
      {
        "entityType": "genomicVariation",
        "schema": "https://raw.githubusercontent.com/
ga4gh-beacon/beacon-v2/main/models/json/beacon-v2-default-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

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

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

Metadata

Publication Landscape of Cancer CNV Profiling

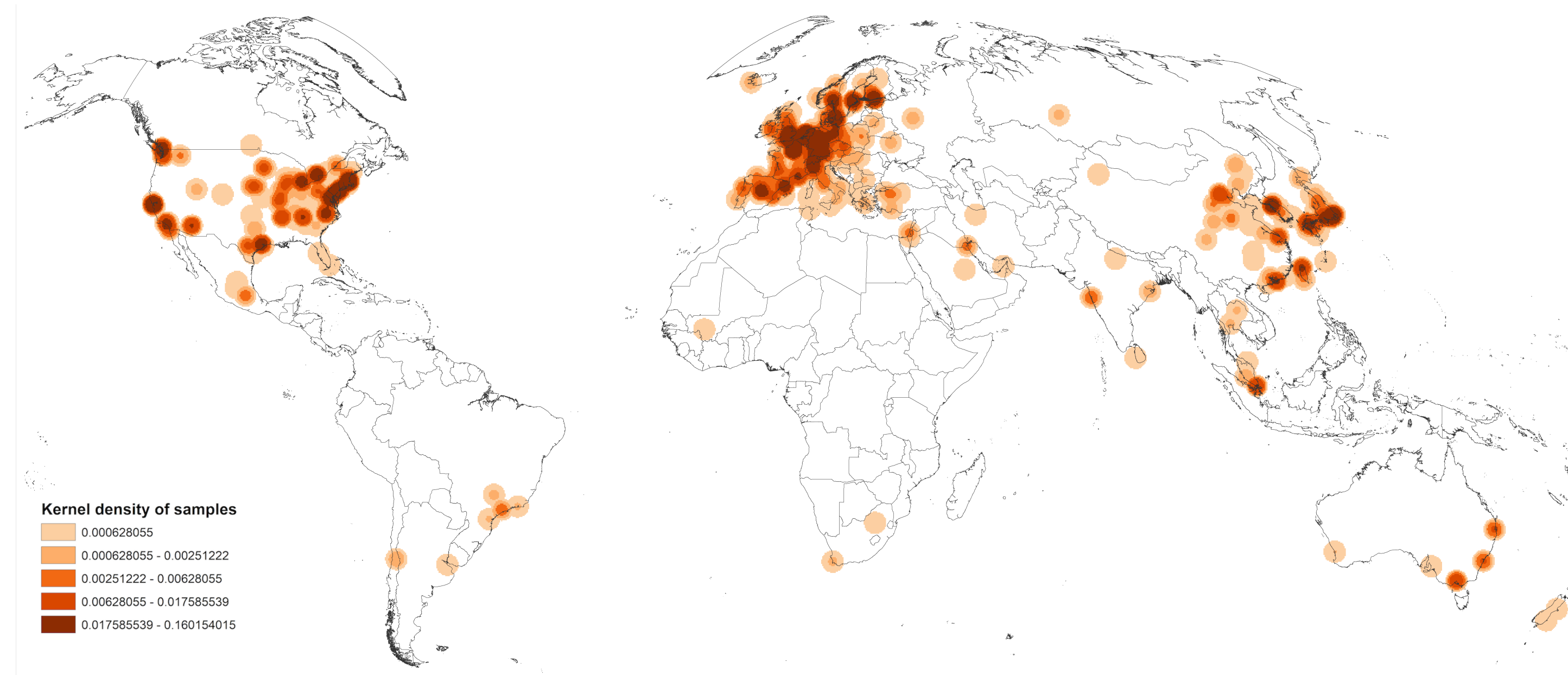


Publication statistics for cancer genome screening studies. The graphic shows our assessment of publications reporting whole-genome screening of cancer samples, using molecular detection methods (chromosomal CGH, genomic array technologies, whole exome and genome sequencing).

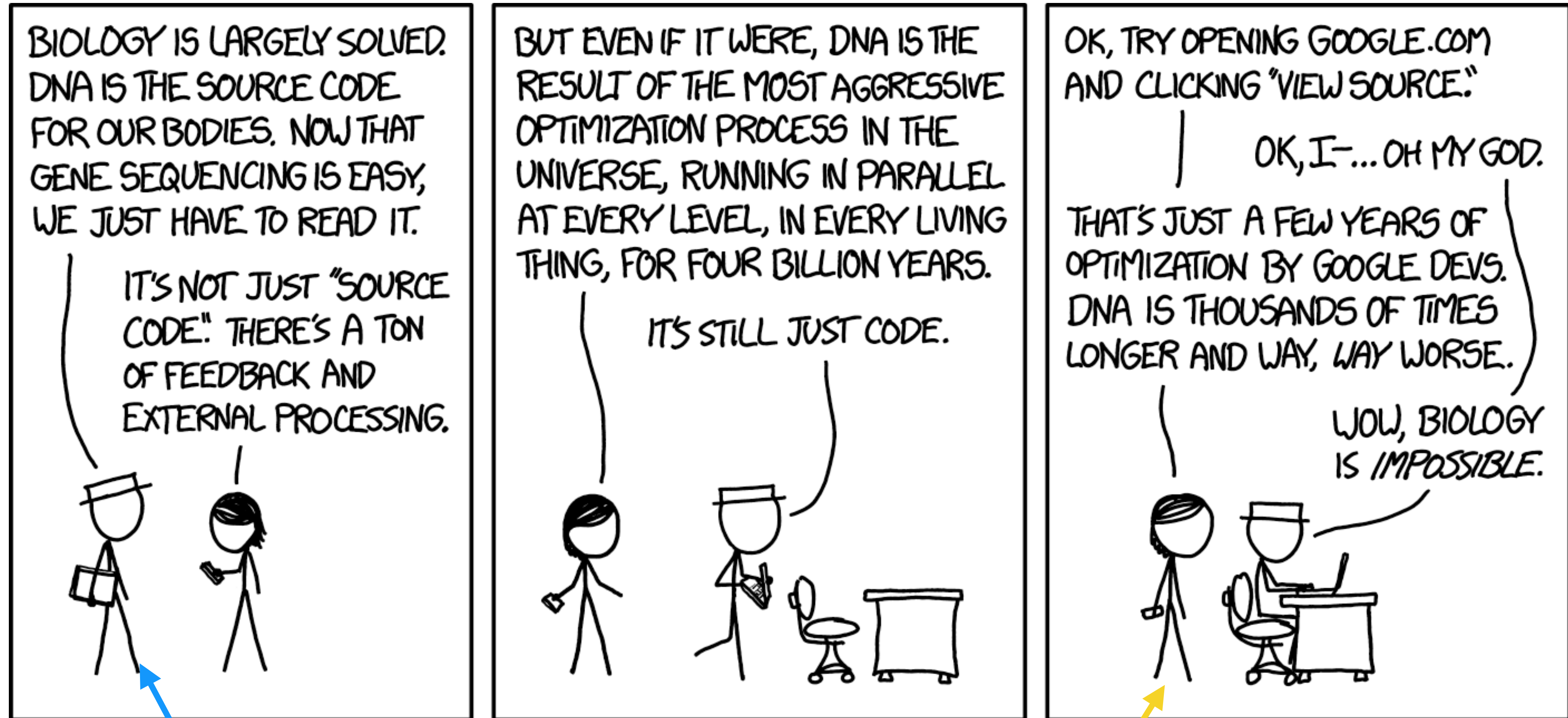
For the years 1993-2018, we found 3'229 publications reporting 174'530 individual samples in single series from 1 to more than 1000 samples. Y-axis and size of the dots correspond to the sample number; the color codes indicate the technology used.

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

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



Who is a Bioinformatician?



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

informatician

bioinformatician

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

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

Bioinformatics OR Computational / Systems Biology?

- **Bioinformatics**

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

- **Computational Biology**

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

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

Data Science

- **Computational Biology**

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

How about “AI”?

- **Bioinformatics**

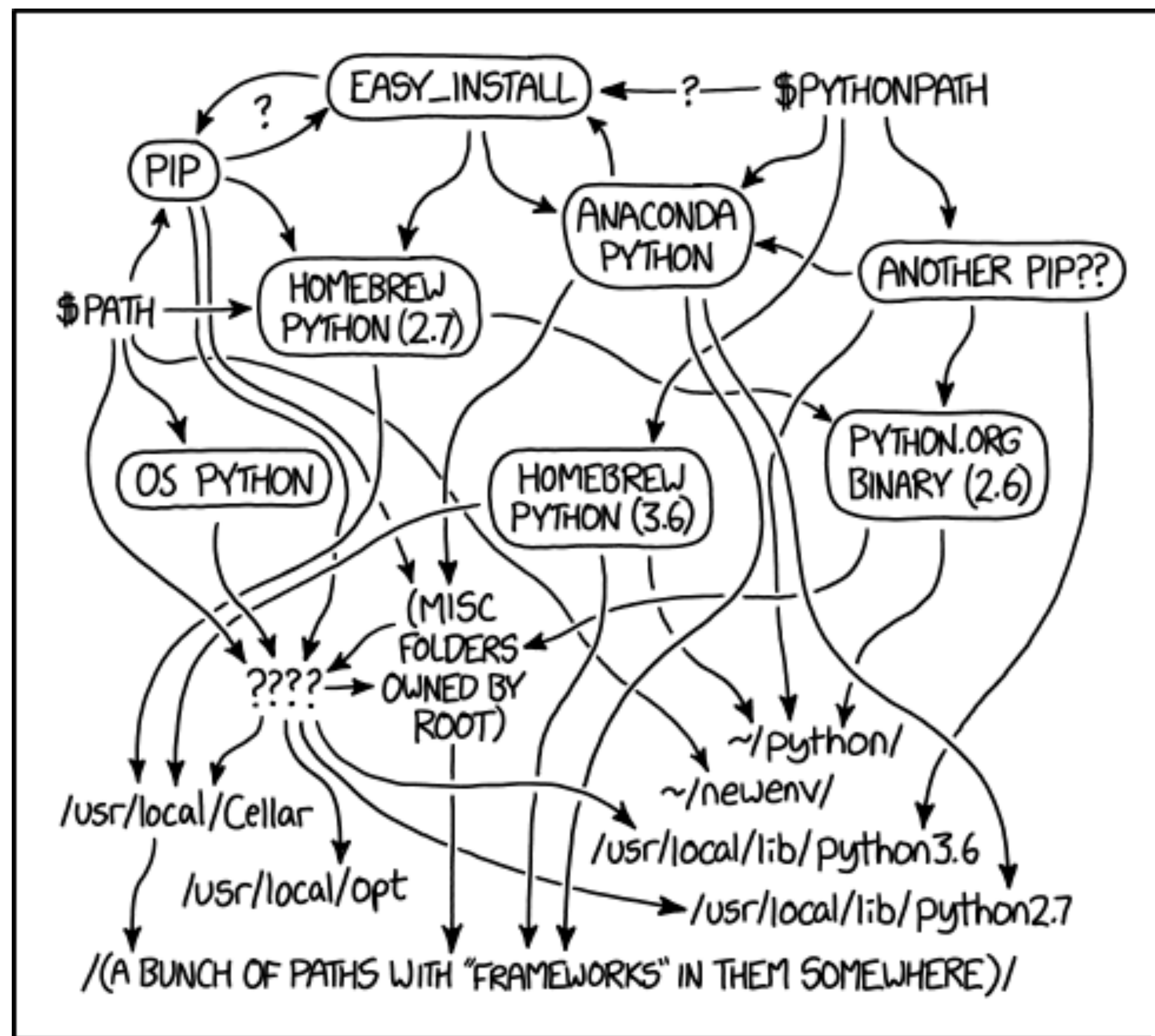
Application of machine learning methods/techniques (ML/AI) is part of bioinformatics workflows (text mining, pattern detection ...)

Training of domain specific models with large proportion of “bio” knowledge/specific data curation becomes relevant

- **Not bioinformatics**

Foundational model development (?)

But in reality that is what bioinformaticians do...



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



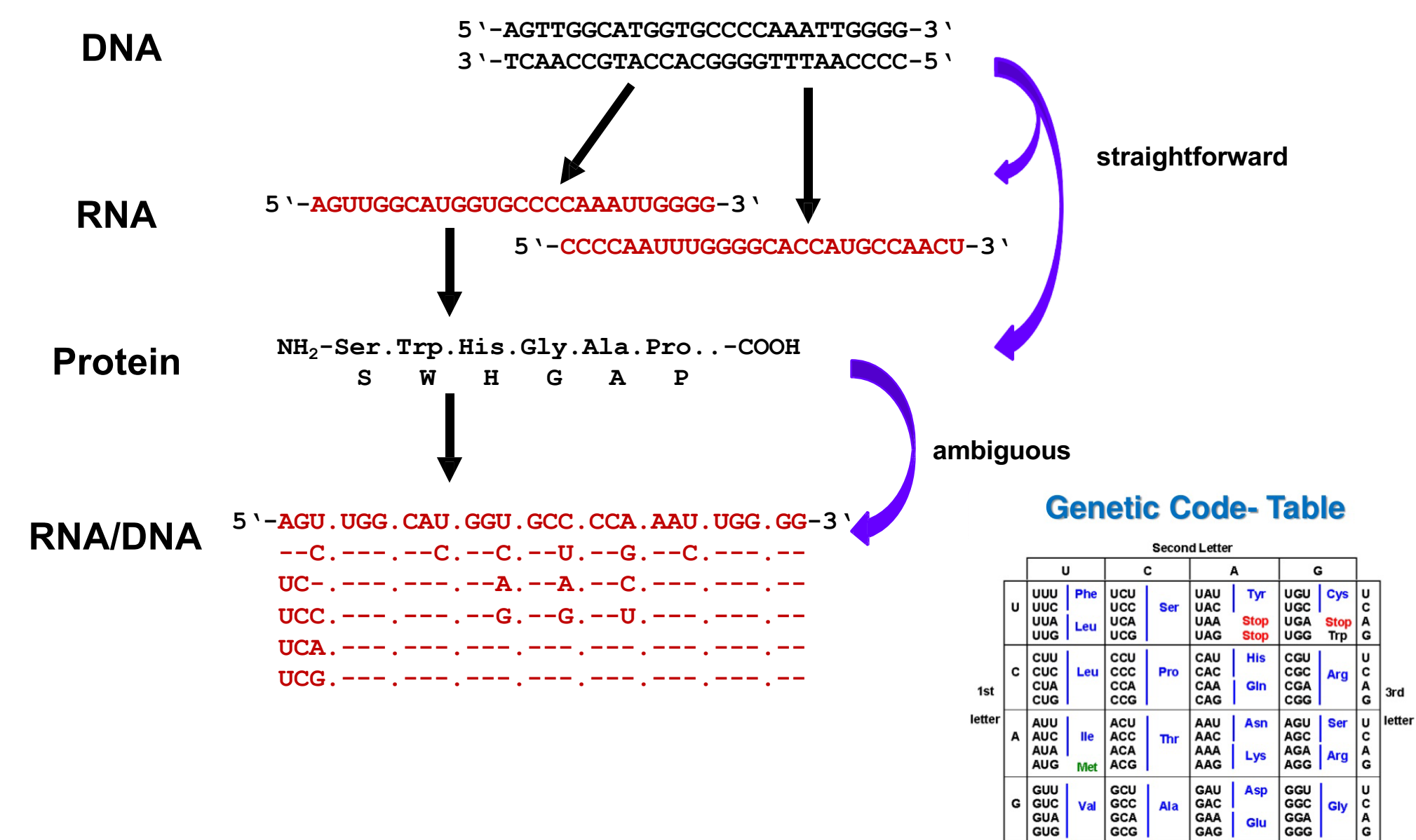
BIO390: Course Schedule

- 2026-09-15: Michael Baudis - What is Bioinformatics? Introduction and Resources
- 2025-09-22: Christian von Mering - Sequence Bioinformatics
- 2025-09-29: Mark Robinson - Statistical Bioinformatics
- 2026-10-06: Virginie Uhlmann - Image Data & Analysis
- 2026-10-13: Valérie Barbie (SIB) - Clinical Bioinformatics
- 2026-10-20: Valentina Boeva (ETHZ) - Machine Learning for Biological Use Cases
- 2026-10-27: Katja Baerenfaller (SIAF) - Proteomics
- 2026-11-03: Shinichi Sunagawa (ETHZ) - Metagenomics
- 2026-11-10: Patrick Ruch (HesGe) - Text mining & Search Tools
- 2026-11-17: Anuradha Mukherjee - Biological Networks & Evolution
- 2026-11-24: Michael Baudis - Building Biological Information Resources
- 2026-12-01: Patrizia Palagi (SIB) - FAIR Data Standards, Methods and Education in Bioinformatics
- 2026-12-08: Michael Baudis - Genome Data & Privacy | Feedback
- 2026-12-15: Exam (Multiple Choice, mostly)

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

AAGCTTACCAAAATTGAAGGGACGTTGACGTAGGGG**GACGCTTTAG**
GACGCTTTAGTTTAGCCACCGGTATTTAGC

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

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

AAGCTTACCAAAATTGAAGGGACGTTGACGTAGGGG**GACGCTTTAG**
AATCTAGCAATTATTGAAGGGACGTTGACGAAGGGGTT**CGCTACCG**

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

AAGCTTACCAAAATTGAAGGGACGTTGACGTAGGGGACGCTTTAG
AATCTAGCAATTATTGAAGGGACGTTGACGAAGGGGTT**CGCTACCG**

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

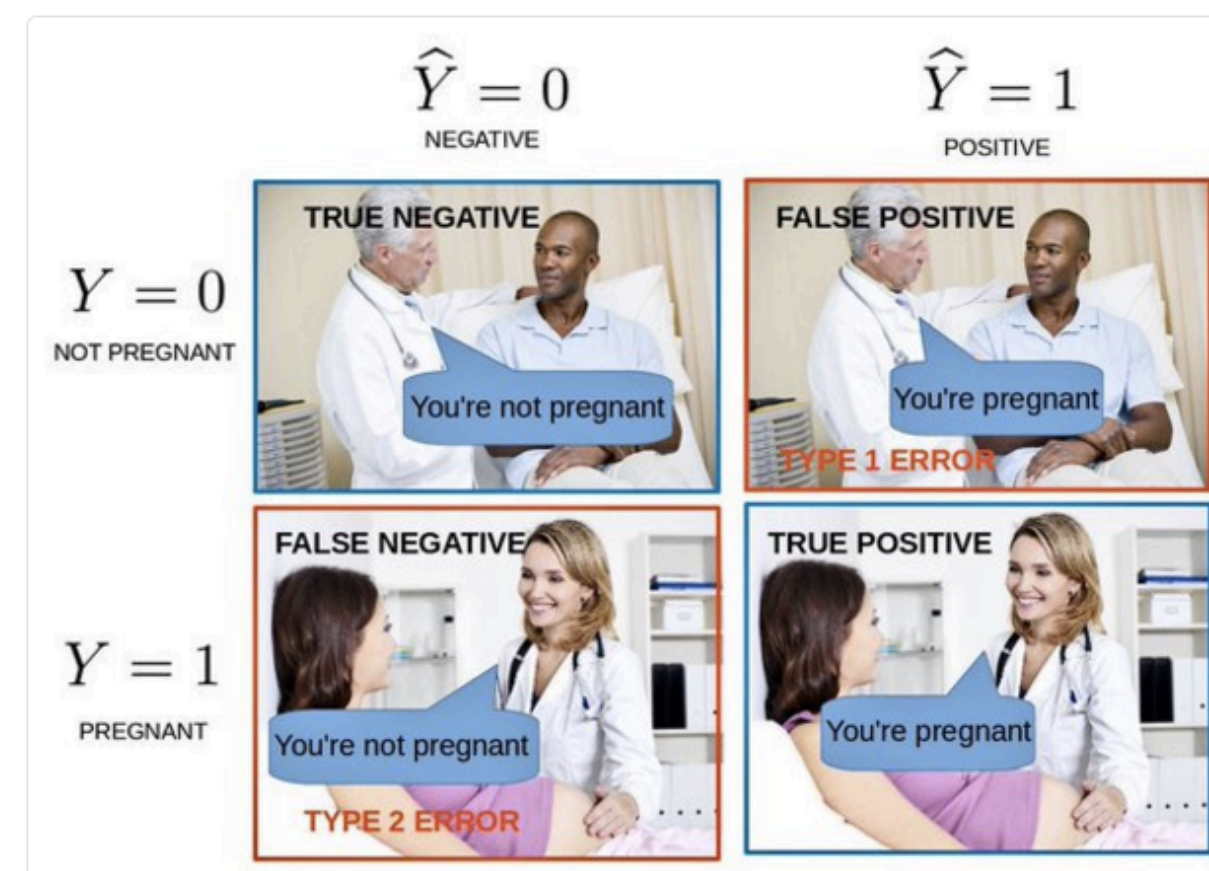
<https://twitter.com/freakonometrics/status/779060142239260672>



Arthur Charpentier
@freakonometrics

Follow

Statistical Errors



40



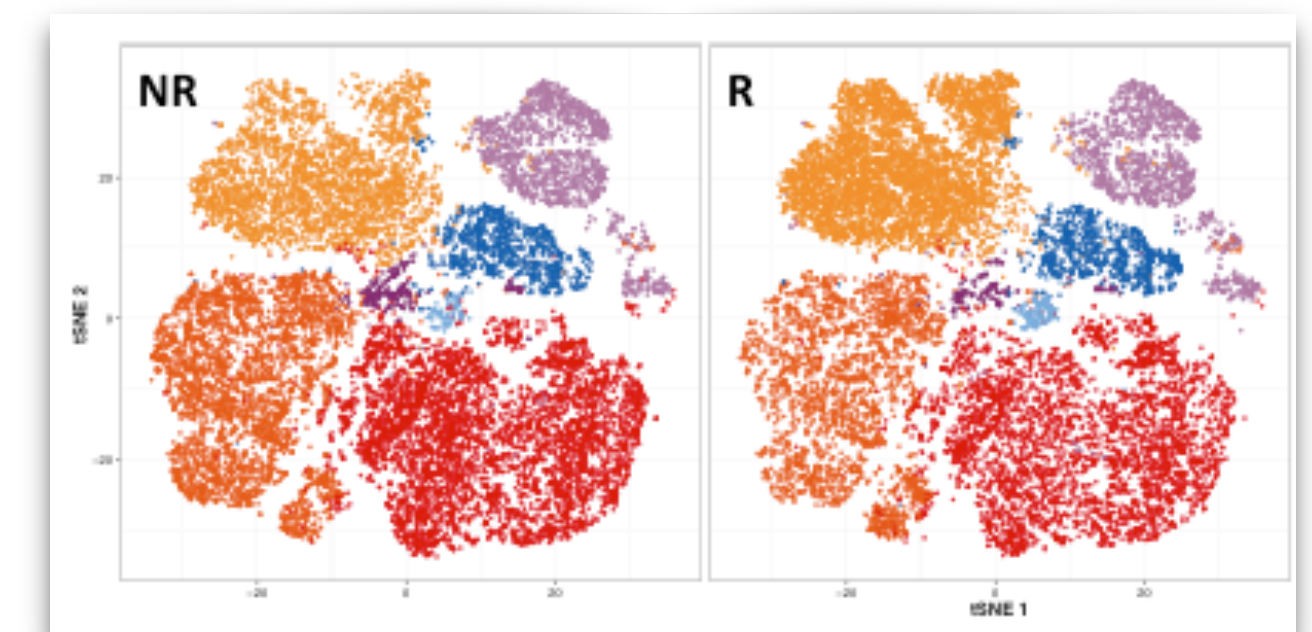
University of
Zurich^{UZH}

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



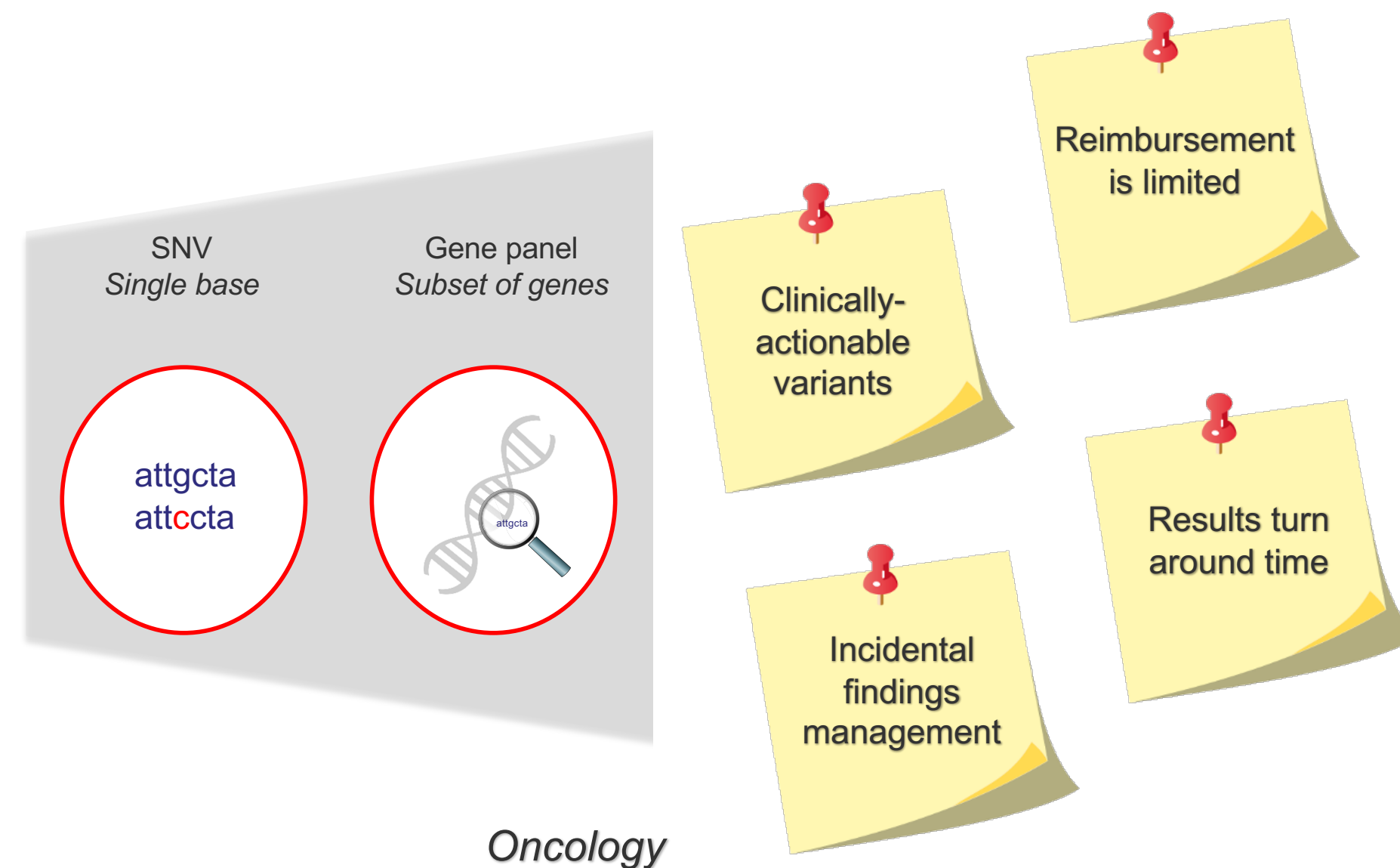
Under the hood: Generalized linear mixed model to
assess the change in relative abundance of
subpopulations.

30

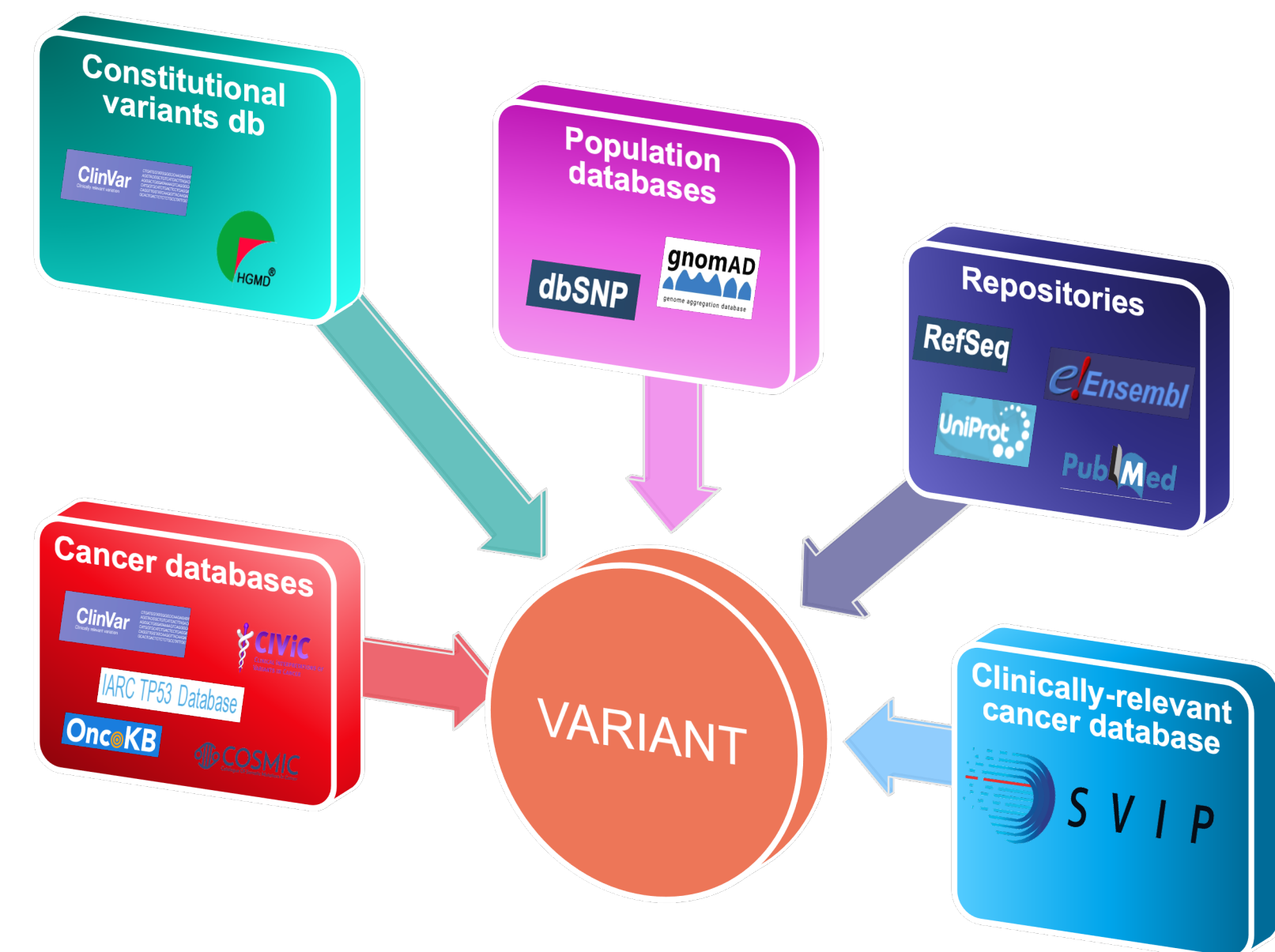
Clinical Bioinformatics

Valérie Barbié (Director SIB Clinical Bioinformatics)

Scale matters



Knowledge bases



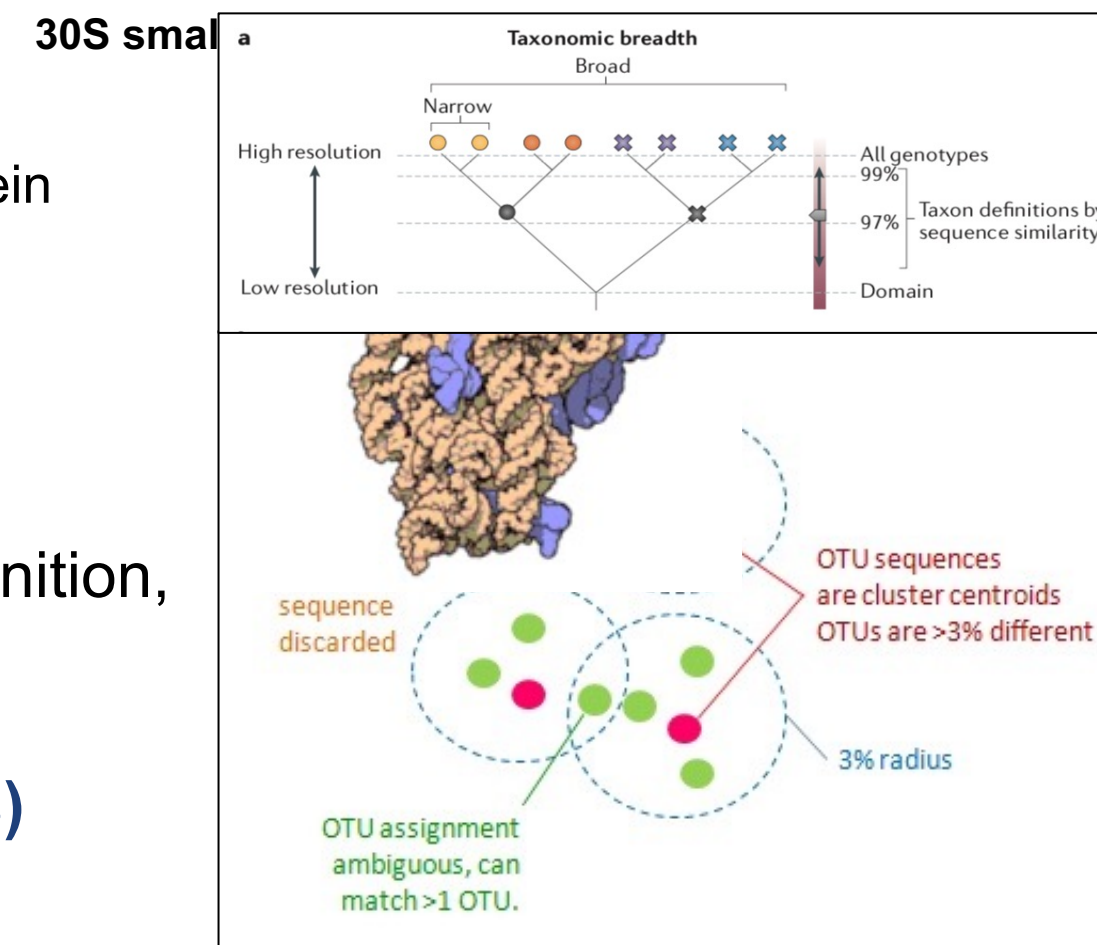
Non exhaustive

Metagenomics

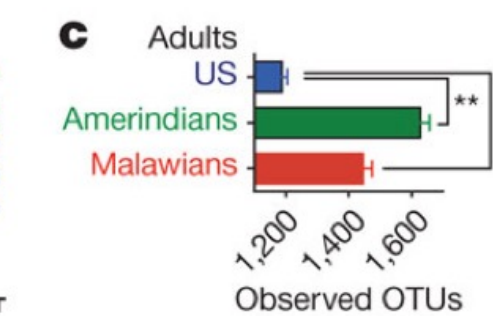
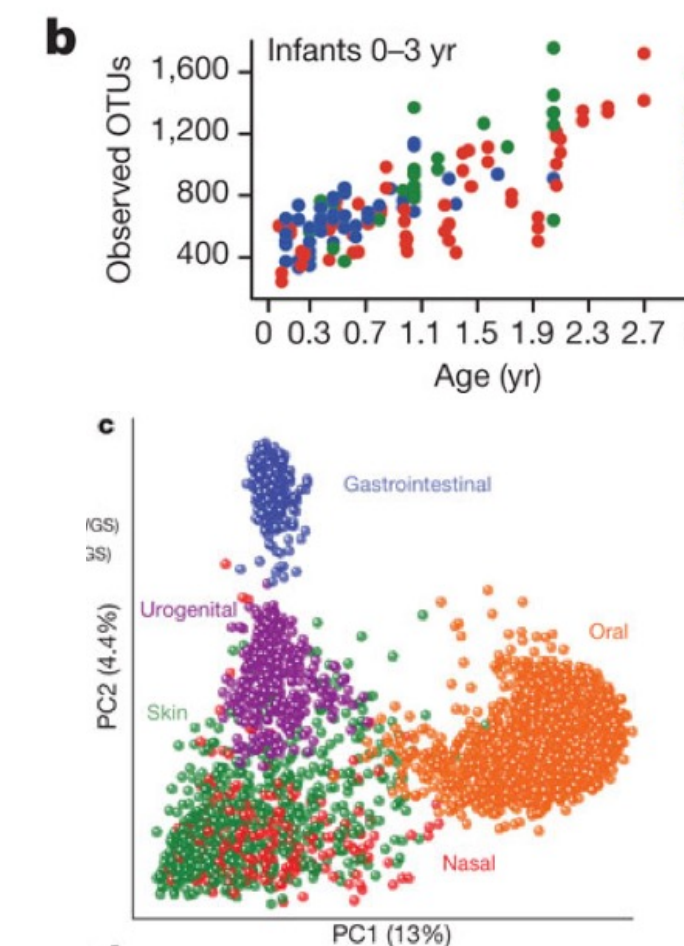
Shinichi Sunagawa (ETHZ)

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

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



Applied examples I

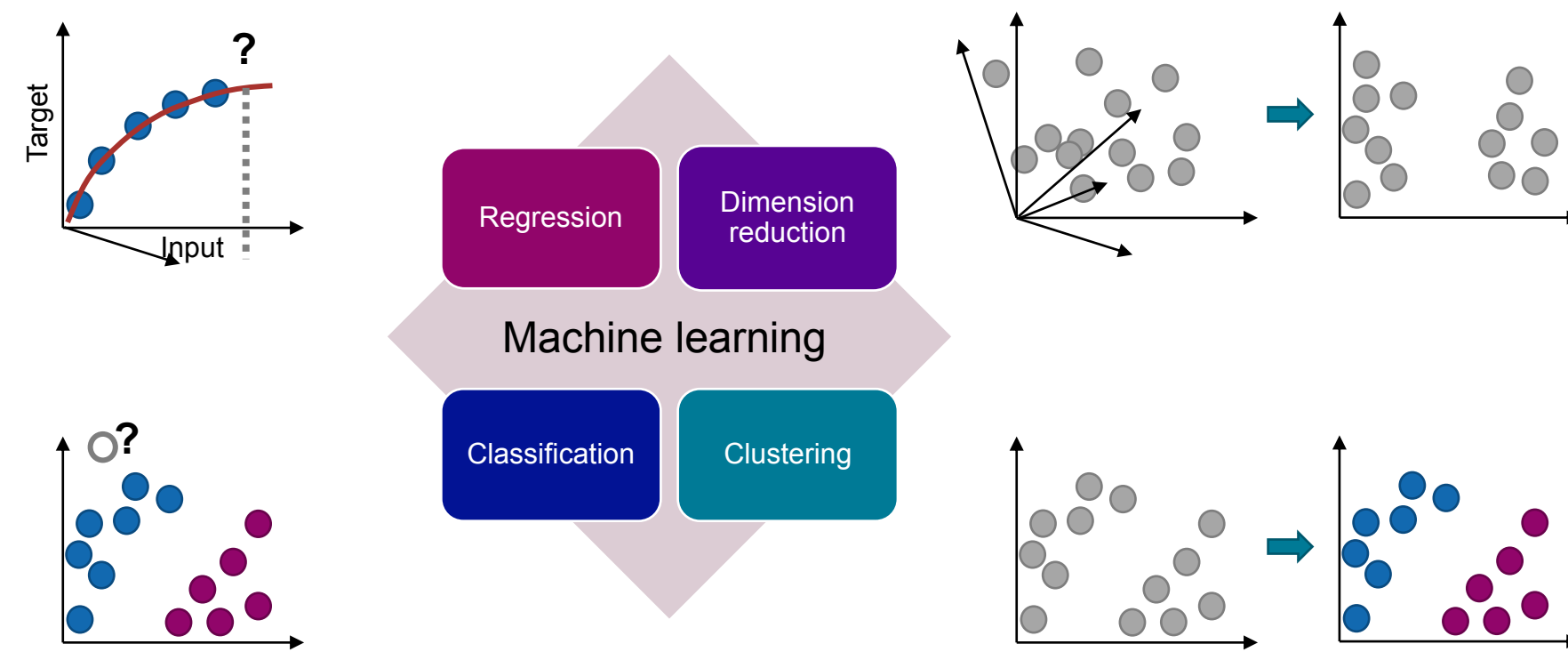


- Microbial diversity in human gut increases with age
- US citizens have harbor less diverse gut microbiota relative to other populations
- Microbial communities cluster by human body site rather than by individual

Machine Learning for Biological Use Cases

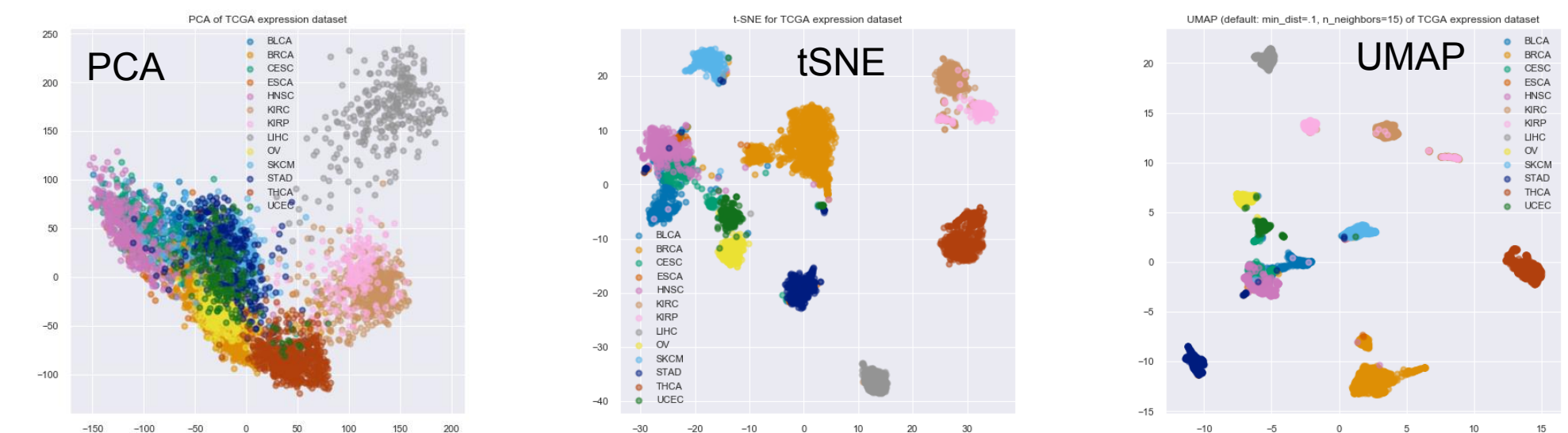
Valentina Boeva (ETHZ)

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

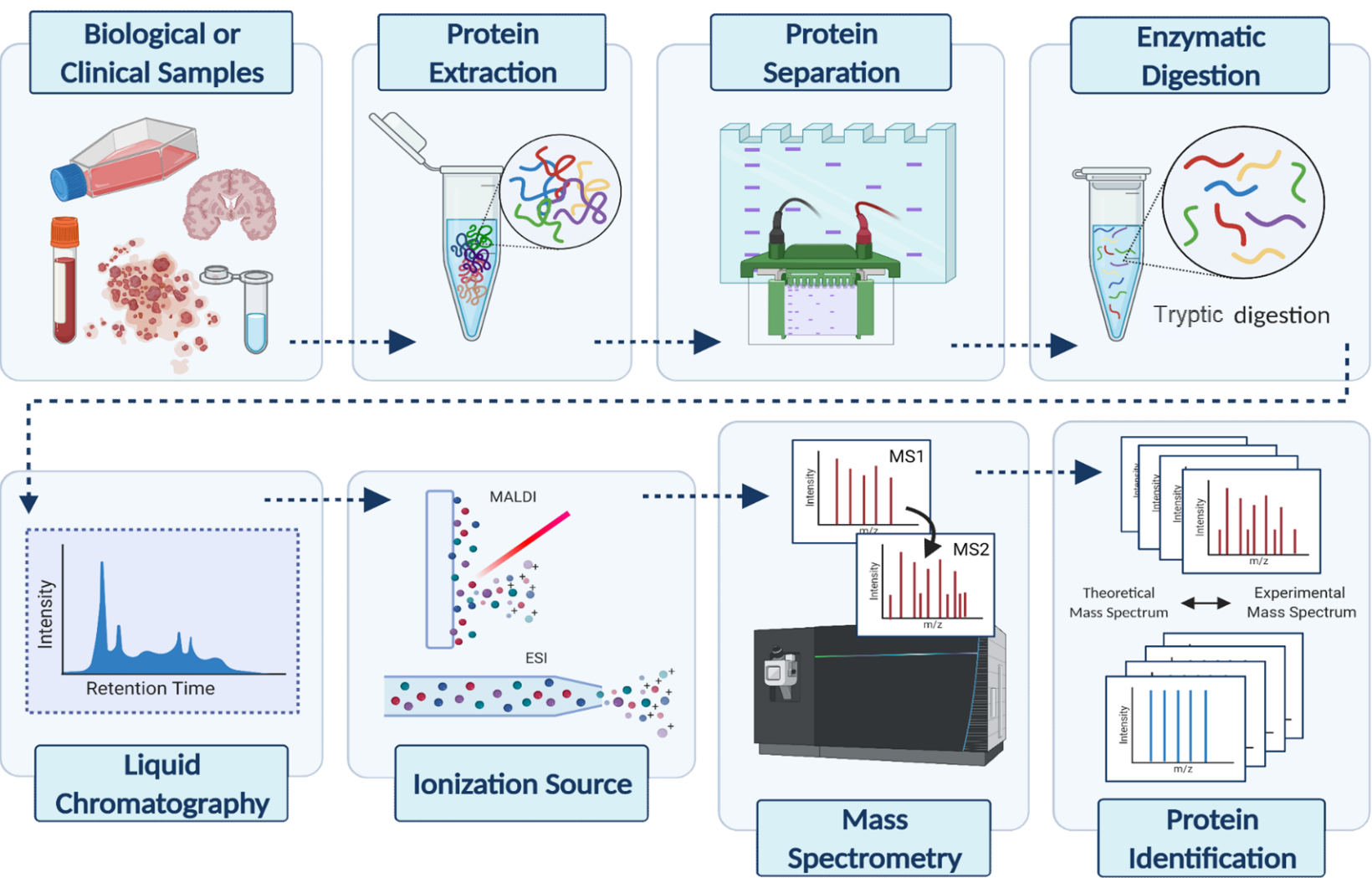


First introduced by [McInnes, L., Healy, J., ArXiv e-prints 1802.03426, 2018](#)

Proteomics

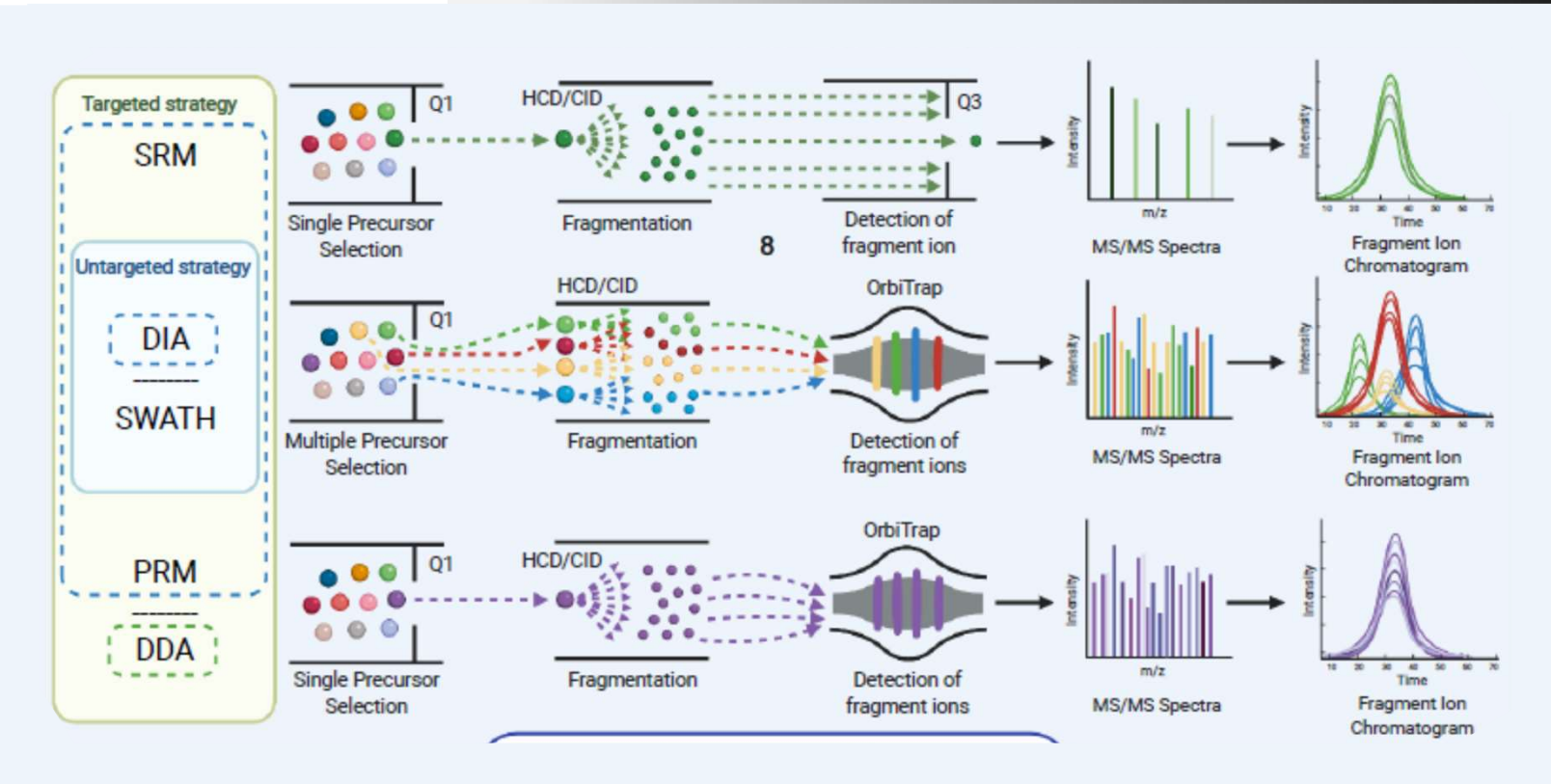
Katja Bärenfaller (SIAF)

Generic mass spectrometry-based proteomics experiment



Mass spectrometry

Hypothesis-driven, targeted bottom-up proteomics approaches



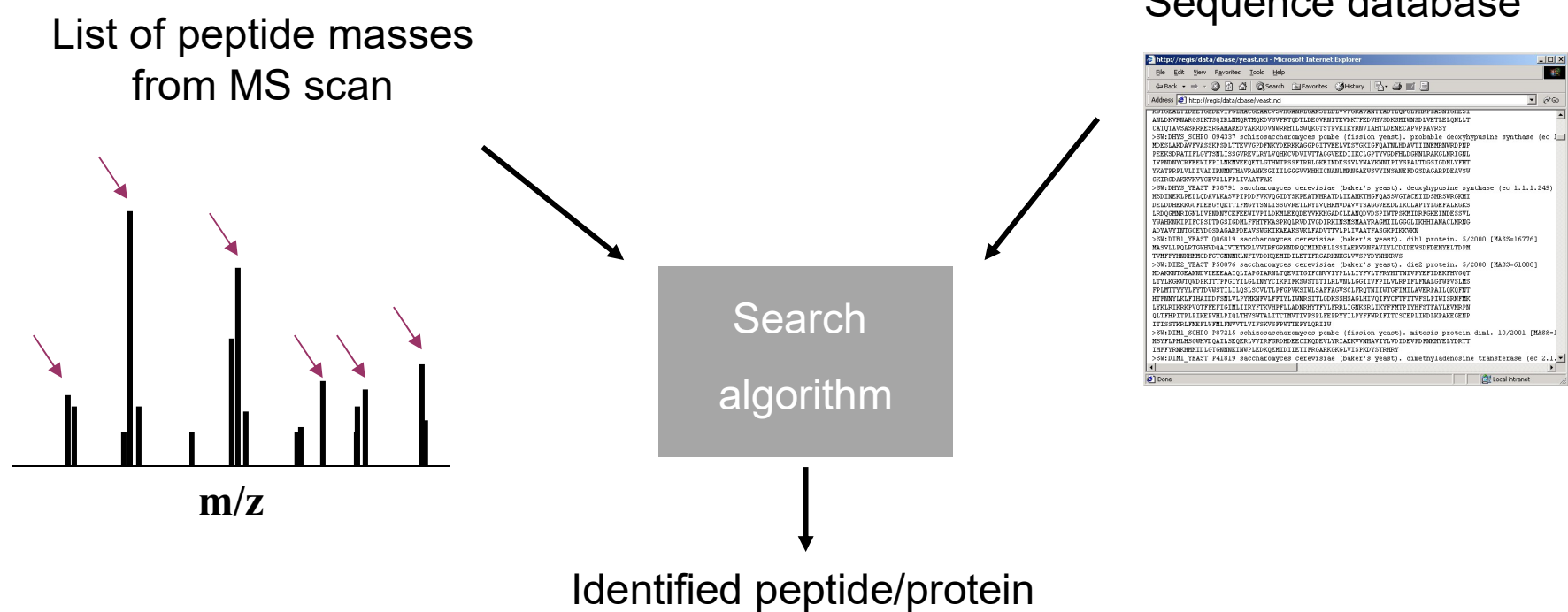
Radzikowska et al., EAACI Position Paper, in revision

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

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

Peptide Mass Fingerprint fragment ion signals from a precursor ion

Identifying peptides using an MS spectrum:



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

Mass spectrometry

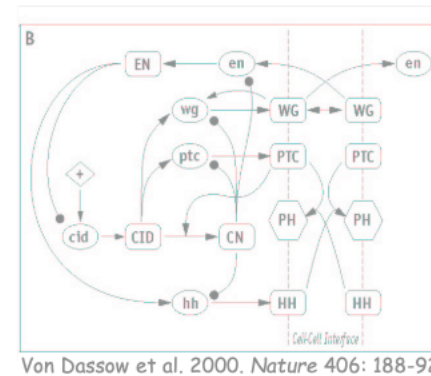
Biological Networks & Evolution

Anuradha Mukherjee

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
...



Von Dassow et al. 2000, Nature 406: 188-92

Mathematical characterization based on detailed, quantitative biochemical information

Cell-biological networks

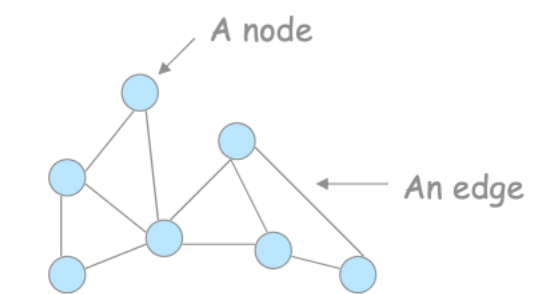
2. Genome-scale networks
(hundreds to thousands of gene products)

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



Mathematical characterization based on qualitative understanding of network topology

Graphs

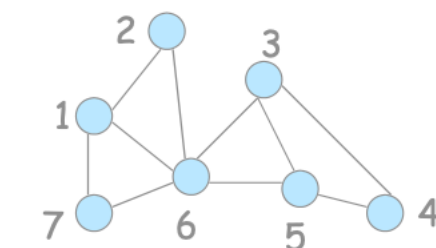


A graph $G=(V,E)$ comprises
a set V of nodes (vertices)
a set E of edges

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

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

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)

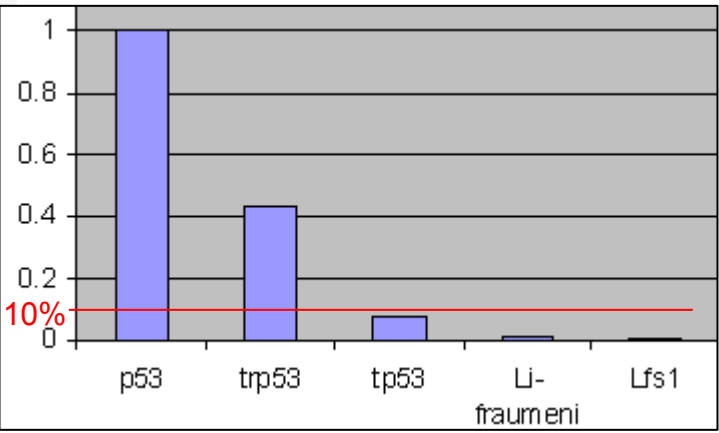
Features

- 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



Building Biological Information Resources

Michael Baudis



CNV Profiles

- ... by NCIT
- ... by ICD-O Morphology
- ... by ICD-O Site
- ... by TNM & Grade

Search Samples

arrayMap

- TCGA Data
- cBioPortal Studies

Publication DB

- Progenetix Use

NCIT - ICD-O Mappings

- UBERON Mappings

Upload & Plot

OpenAPI Paths and Examples

Cancer Cell Lines

Beacon+

Documentation

Baudisgroup @ UZH

Cancer genome data @ progenetix.org

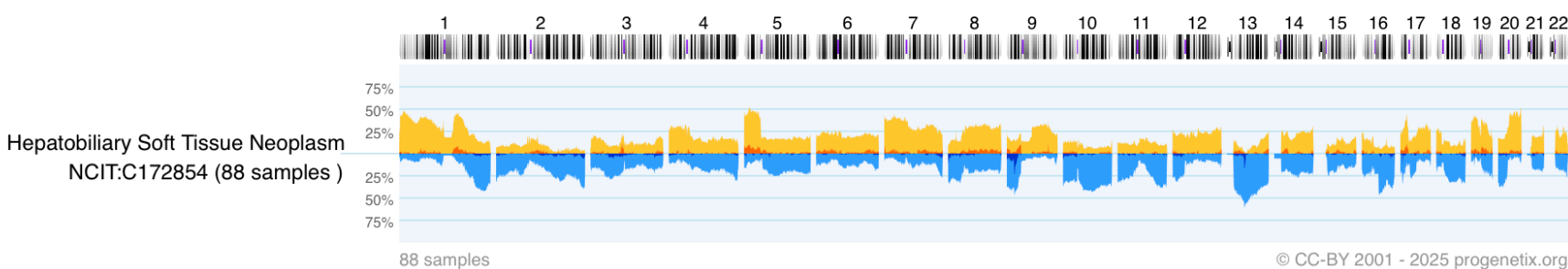
The Progenetix database provides an overview of mutation data in cancer, with a focus on copy number abnormalities (CNV / CNA), for all types of human malignancies. The data is based on *individual sample data* of currently **240600** samples from **1126** different cancer types (NCIT neoplasm classification)

Local CNV Frequencies

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

Cancer CNV Profiles

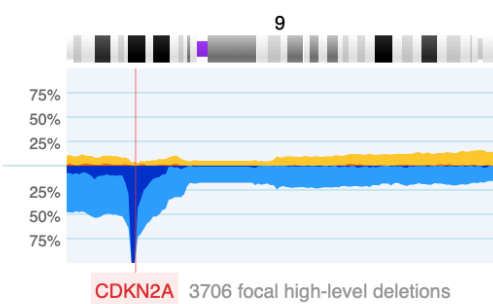
Frequency profiles of regional genomic gains and losses for all categories (diagnostic entity, publication, cohort ...) can be accessed through the respective Cancer Types pages (e.g. [NCIT Neoplasia Codes](#)) and compared through the [Compare CNV Profiles](#) option. Below is an example of aggregated CNV data in 88 samples in HePATObiliary Soft Tissue Neoplasm with the frequency of regional **copy number gains (high level)** and **losses (high level)** displayed for the 22 autosomes.



[Download SVG](#) | [Go to NCIT:C172854](#) | [Download CNV Frequencies](#)

Cancer Genomics Publications

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



© CC-BY 2001 - 2025 progenetix.org



Search Download Help My Data

Welcome to STRING

Protein-Protein Interaction Networks
Functional Enrichment Analysis

ORGANISMS | PROTEINS | INTERACTIONS
12535 | 59.3 mio | >20 bln

SEARCH

© STRING CONSORTIUM 2023



SIB - Swiss Institute of Bioinformatics



CPR - Novo Nordisk Foundation Center Protein Research



EMBL - European Molecular Biology Laboratory

ABOUT

Content
References
People
Statistics

INFO

Scores
Use scenarios
FAQs
Cookies/Privacy

ACCESS

Versions
APIs
Licensing
Usage

CREDITS

Funding
Datasources
Partners
Software



Global Alliance

for Genomics & Health

Collaborate. Innovate. Accelerate.

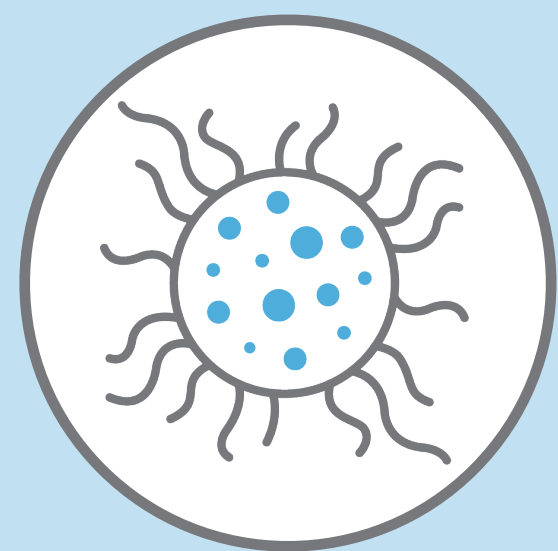
GENOMICS

A federated ecosystem for sharing genomic, clinical data

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



Global Genomic Data Sharing Can...



Demonstrate
patterns in health
& disease



Increase statistical
significance of
analyses



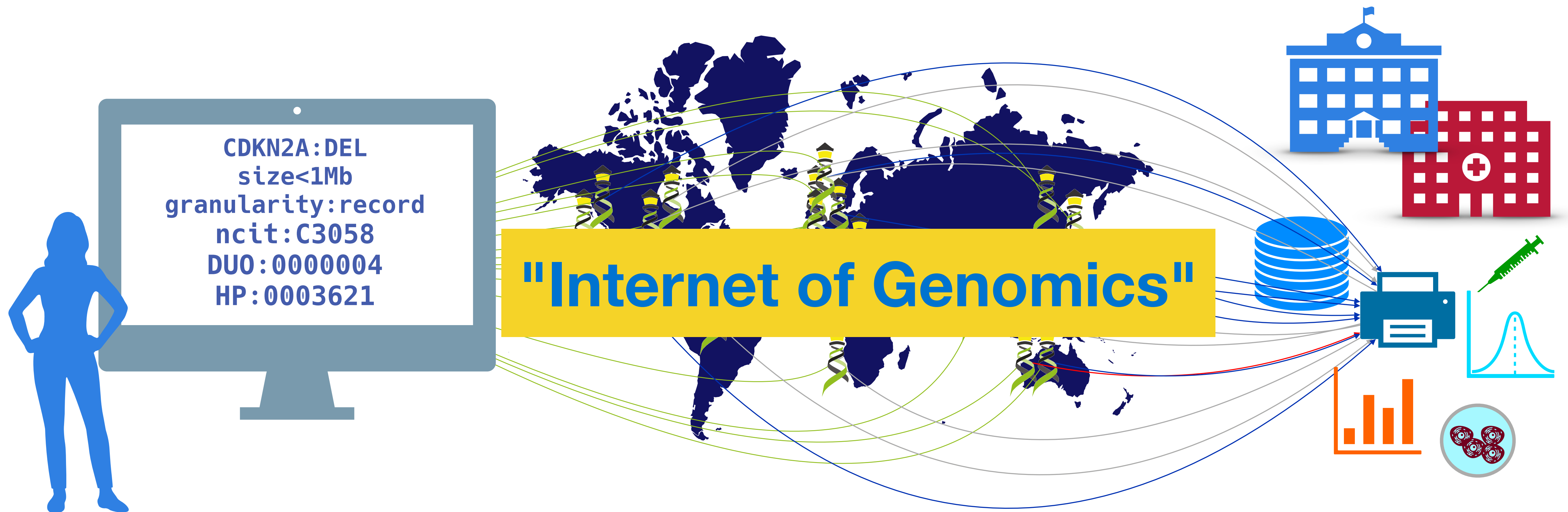
Lead to
“stronger” variant
interpretations



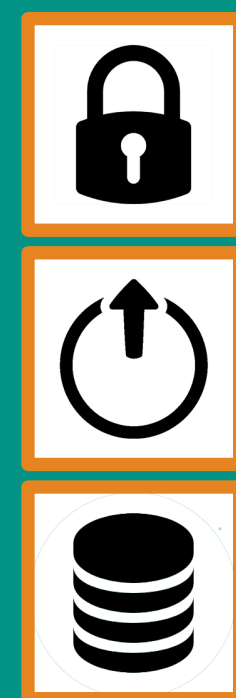
Increase
accurate
diagnosis



Advance
precision
medicine



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



Beacon API

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

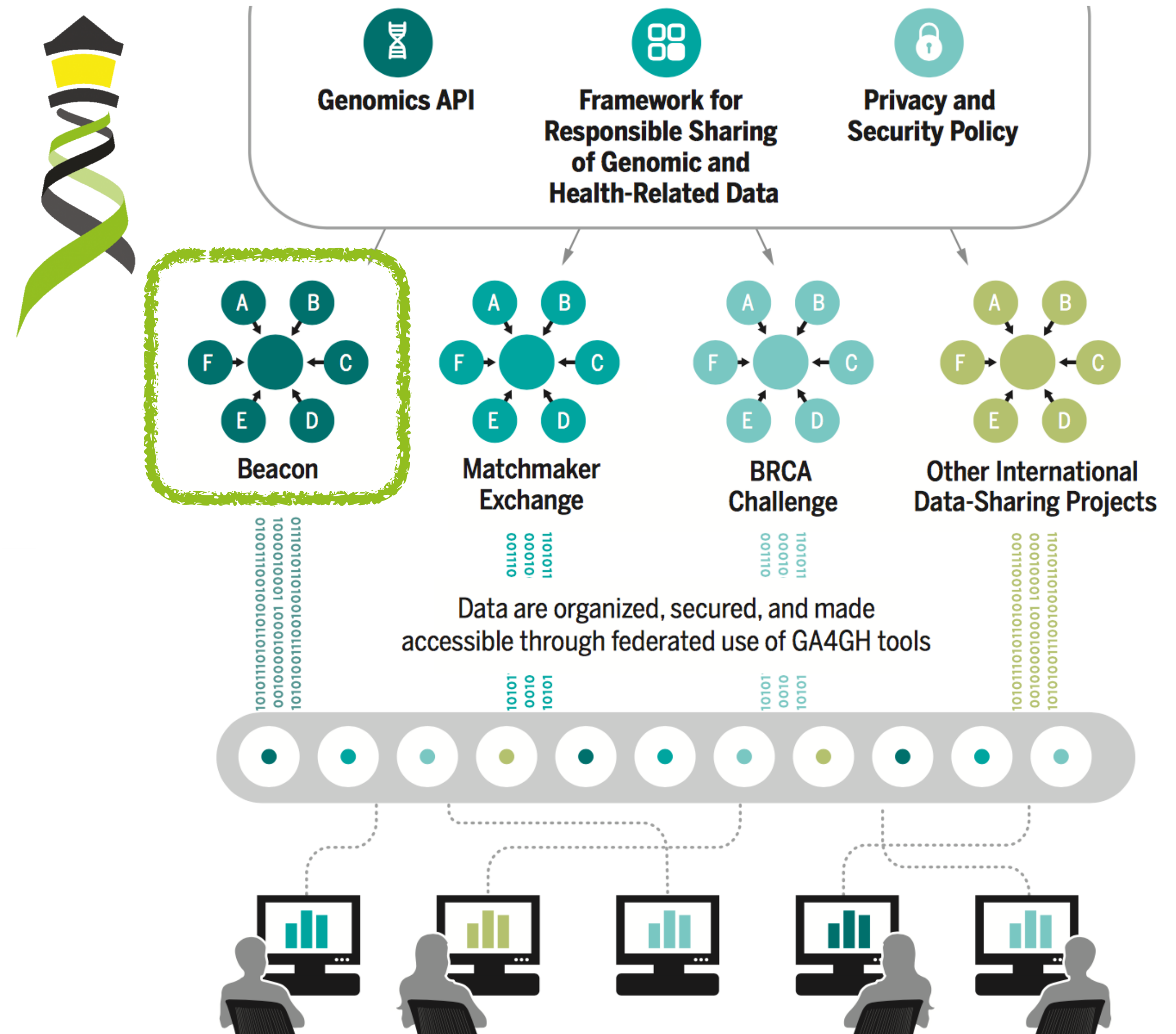


GENOMICS

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.



DNASTACK



Global Alliance
for Genomics & Health

Genomic Data & Privacy: Risks & Opportunities

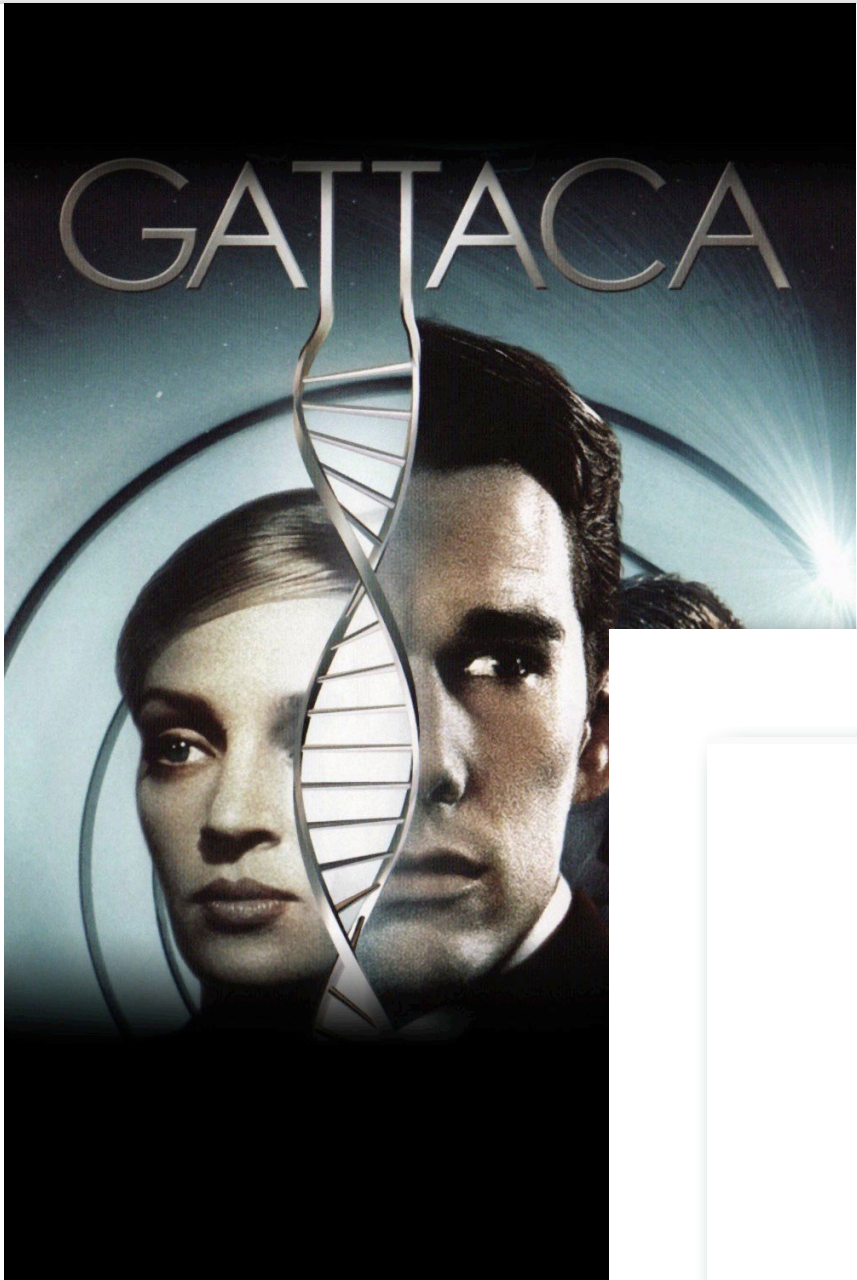
Michael Baudis

Gattaca (1997)

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

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

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



Genome Beacons Compromise Security?

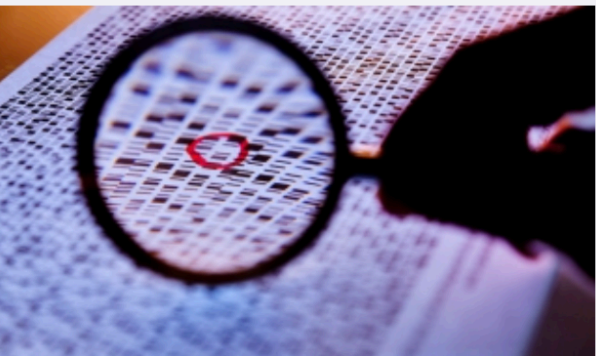
Querying for thousands of specific SNV occurrences in a genomic data pool can identify individuals in an anonymized genomic data collection

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

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

OCT 29 2015 Sharing genomic information among researchers is critical to the advance of biomedical research. Yet genomic data contains identifiable information and, in the wrong hands, poses a risk to individual privacy. If someone had access to your genome sequence — either directly from your saliva or other tissues, or from a popular genomic information service — they could check to see if you appear in a database of people with certain medical conditions, such as heart disease, lung cancer or autism.

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



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

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

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

Rapid re-identification of human samples

...

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



Forensics

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



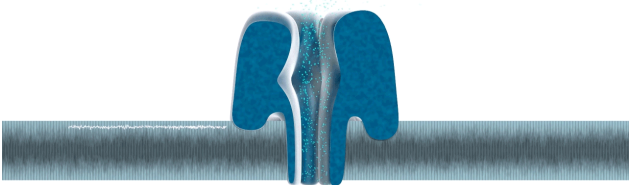
Clinic

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



Cell line identification

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



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



University of
Zurich ^{UZH}



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

progenetix.org
baudisgroup.org
sib.swiss/baudis-michael
mls.uzh.ch/en/research/baudis
genomebeacons.org



Global Alliance
for Genomics & Health

