BIO390: Introduction to Bioinformatics

Lecture I: What is Bioinformatics?

Michael Baudis - 2019-09-17

University of Zurich[UZH]

BIO390: Course Schedule

- 2019-09-17: Michael Baudis - What is Bioinformatics? Introduction and Resources
- 2019-09-24: Christian von Mering - Sequence Bioinformatics
- 2019-10-01: Shinichi Sunagawa (ETHZ) - Metagenomics
- 2019-10-08: Mark Robinson - Statistical Bioinformatics
- 2019-10-15: Izaskun Mallona - Regulatory Genomics and Epigenomics
- 2019-10-22: Abdullah Kahraman (USZ) - Molecular Interaction Networks
- 2019-10-29: Katja Baerenfaller (SIAF) - Proteomics
- 2019-11-05: Amedeo Caflisch - Molecular Dynamics
- 2019-11-12: Elif Ozkirimli - Protein Structure and Interactions
- 2019-11-19: Christophe Dessimoz (UniL) - Sequence evolution and phylogenetics
- 2019-11-26: Valérie Barbie (SIB) - Clinical Bioinformatics
- 2019-12-03: Andreas Wagner - Biological Networks
- 2019-12-10: Alex Handler Wagner (WUSTL) & Michael Baudis - Human Genome Variation Resources
- 2019-12-17: Exam (Multiple Choice)

Course Information BIO390

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

Prof. Dr. Michael Baudis
Department of Molecular Life Sciences (IMLS)
University of Zurich Campus Irchel, Y-13F-01
CH-8057 Zurich
email mbaudis@imls.uzh.ch
web www.imls.uzh.ch/baudis

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

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.

Final Program 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**

This list is not exhaustive; additional information about “need to know” topics will be provided during the individual lectures.

Literature and Resources

- Literature links and recommendations
- Resource links (browsers and online repositories)

Links

- BIO390 HS 2019 in the UZH OLAT system
- BIO390 HS 2019 in the UZH directory

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

Lectures

Upcoming

2019-09-17

Michael Baudis - Introducing Bioinformatics

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 2019-09-17: more ...

2019-09-24

Christian von Mering - Sequence Bioinformatics

2019-09-24: more ...

2019-10-01

Shinichi Sunagawa - Metagenomics

2019-10-01: more ...

2019-10-08

Mark Robinson - Statistical Bioinformatics

2019-10-08: more ...

2019-10-15

Izaskun Mallona - Regulatory Genomics and Epigenomics

2019-10-15: more ...

2019-10-22

Abdullah Kahraman - Molecular Interaction Networks

2019-10-22: more ...

2019-10-29

Katja Baerenfaller - Proteomics

2019-10-29: more ...

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

People

Abdullah Kahraman, PhD

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

2019-12-10: more ...

Alex H. Wagner, PhD

- Instructor in Medicine
- Washington University School of Medicine, St. Louis, MO, U.S.A.
- Co-director, Variant Interpretation for Cancer Consortium (VICC)
- GKS VR Lead, Global Alliance for Genomics and Health (GA4GH)

2019-12-10: more ...

Amedeo Caflisch

- Professor of Computational Structural Biology
- Department of Biochemistry
- University of Zürich

1996-07-01: more ...

Andreas Wagner

- Professor and Chairman, Dept. of Evolutionary Biologyand 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 ...

Christophe Dessimoz, PhD

- SNSF Professor, University of Lausanne
- Associate Professor, University College London
- Group leader, Swiss Institute for Bioinformatics

2019-12-10: more ...

Elif Ozkirimli Olmez, PhD

- Associate Professor of Chemical Engineering, Bogazici University, Istanbul, Turkey ↗
- Visiting Scientist in Department of Biochemistry, Computational Structural Biology, University of Zürich

2019-08-28: more ...

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

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

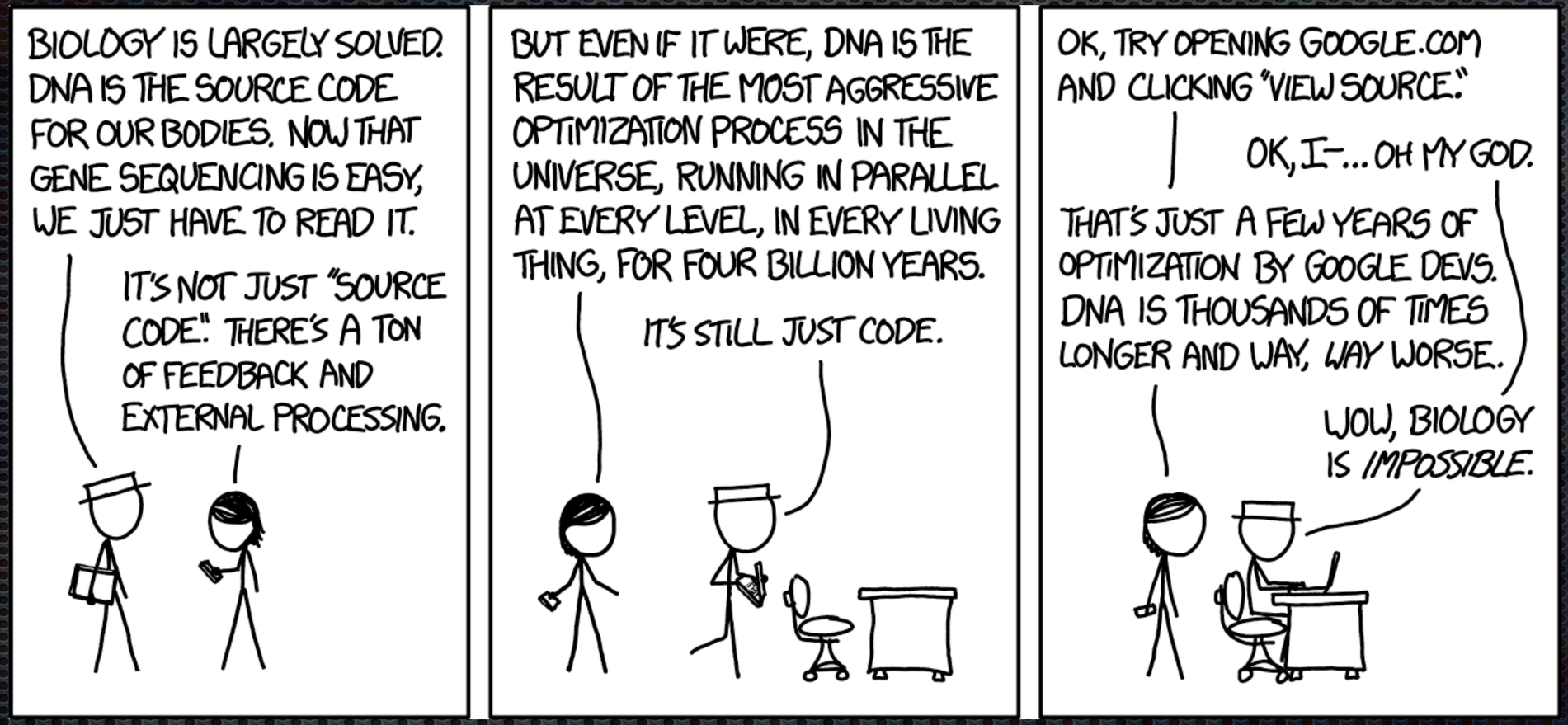
Some Recommended Books

- 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)

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



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

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?

science

noun | sci·ence | \ˈsī-ən(t)s\

Popularity: Top 1% of lookups

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

Merriam-Webster

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 ... *

adapted from "Catalyzing Inquiry at the Interface of Computing and Biology"; John C Wooley and Herbert S Lin. National Research Council (US) Committee on Frontiers at the Interface of Computing and Biology. Washington (DC): National Academies Press (US); 2005.

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?

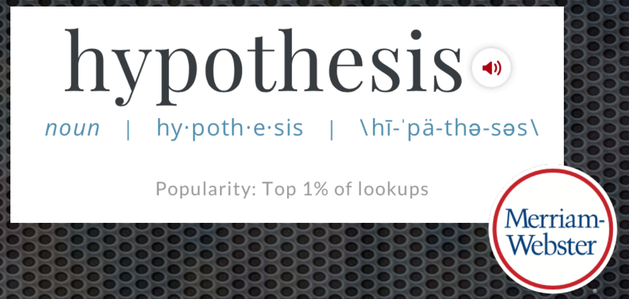
- Bioinformatics is "the science that uses the instruments of informatics to analyze biological data in order to **formulate hypotheses** about life." (Anna Tramontano)

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

What is Bioinformatics?

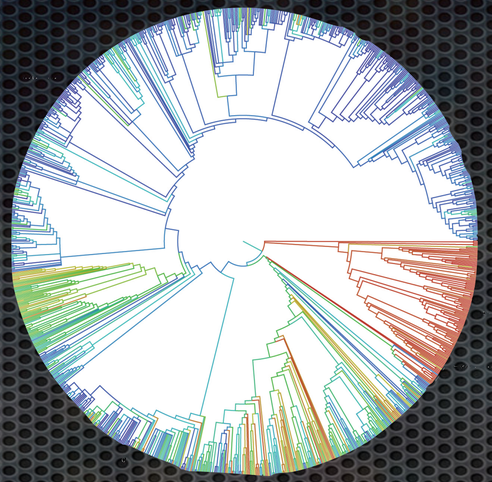
hypothesis driven science

data driven science

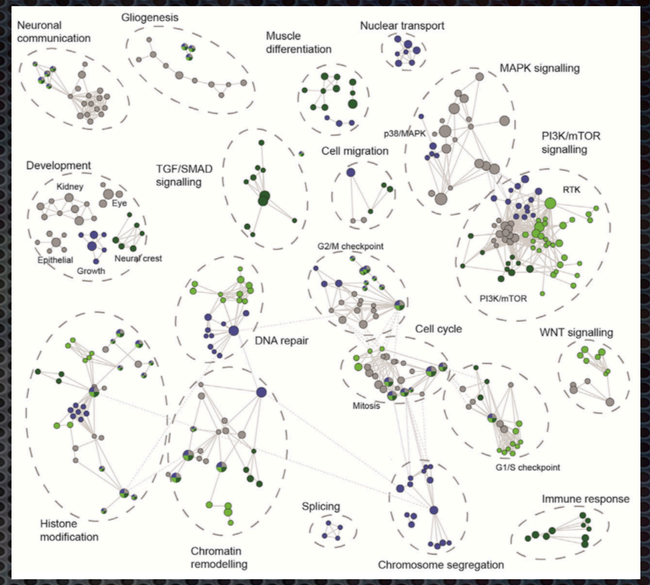


- Bioinformatics is "the science that uses the instruments of informatics to analyze biological data in order to **formulate hypotheses** about life." (Anna Tramontano)

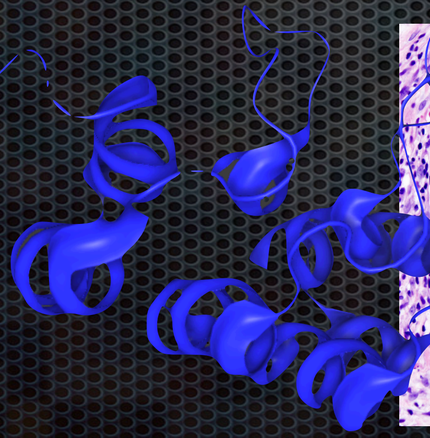
b : an interpretation of a practical situation or condition taken as the ground for action

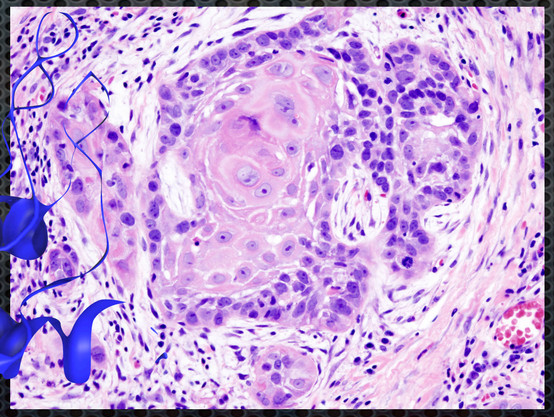


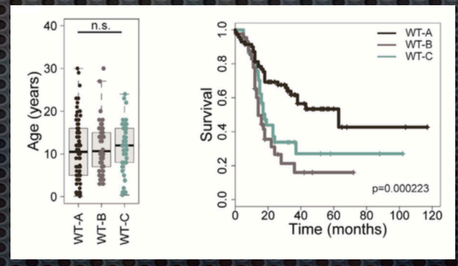
42



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





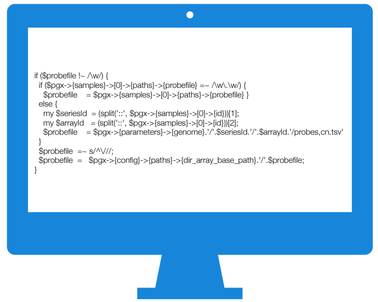


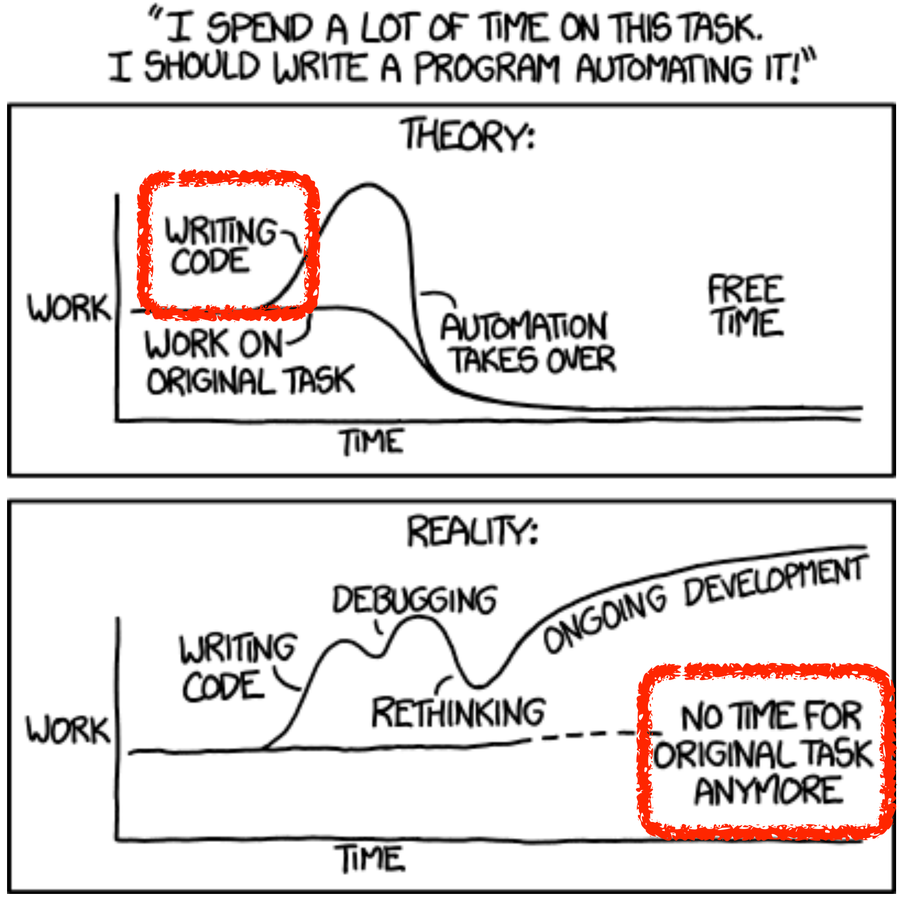
Sources: nextprot | opentreeoflife | wikipedia | MacKay et al., Cancer Cell (2017) | original photos

{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}





Randall Munroe - XKCD
https://xkcd.com/1319/

Who is a Bioinformatician?

Bioinformatician	Bio**informatician**
strong biological knowledge	sufficient biological background
provides hypothesis and/or dataset	provides statistical, analysis methods
sufficient statistical and computational expertise to correctly use bioinformatics tools & develop workflows (scripting ...)	sufficient biologic or medical background to understand problems presented and identify pitfalls and hidden biases arising from data generation methods
expert user of informatics tools	developer of informatics tools
may get a Nobel	may get rich

Who is a Bioinformatician?

Bioinformatician	Bio**informatician**
strong biological knowledge	sufficient biological background
provides hypothesis and/or dataset	provides statistical, analysis methods
sufficient statistical and computational expertise to correctly use bioinformatics tools & develop workflows (scripting …)	sufficient biologic or medical background to understand problems presented and identify pitfalls and hidden biases arising from data generation methods
expert user of informatics tools	developer of informatics tools
may get a Nobel	may get rich

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

ENA European Nucleotide Archive

STRING

arrayMap

KEGG Kyoto Encyclopedia of Genes and Genomes

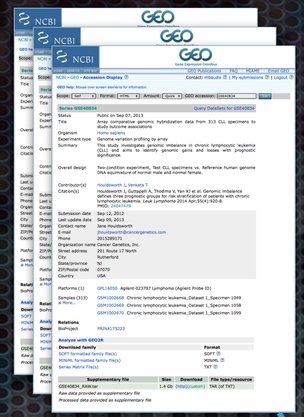
neXtprot

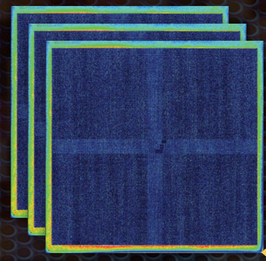
UniProt

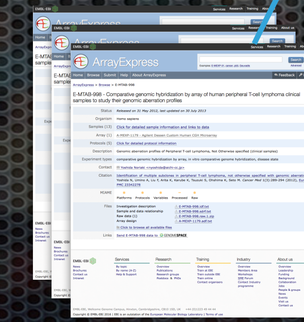
ARRAYMAP DATA PIPELINE

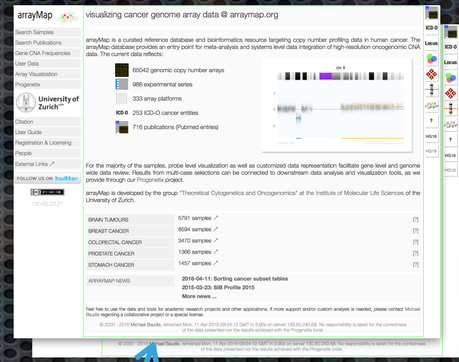
BIOCURATION

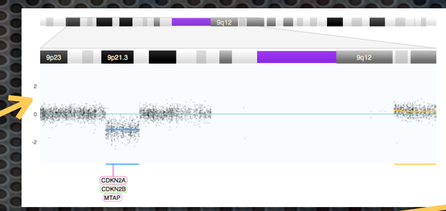
BIOINFORMATICS

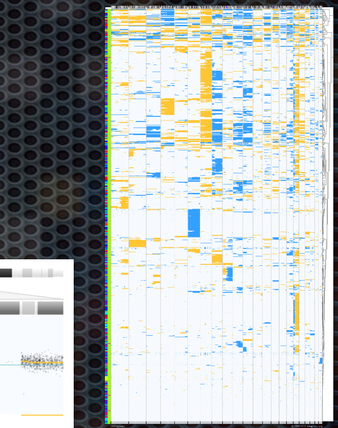


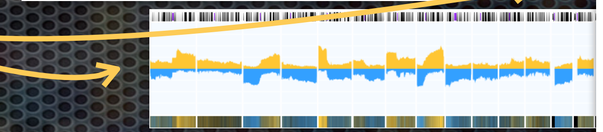


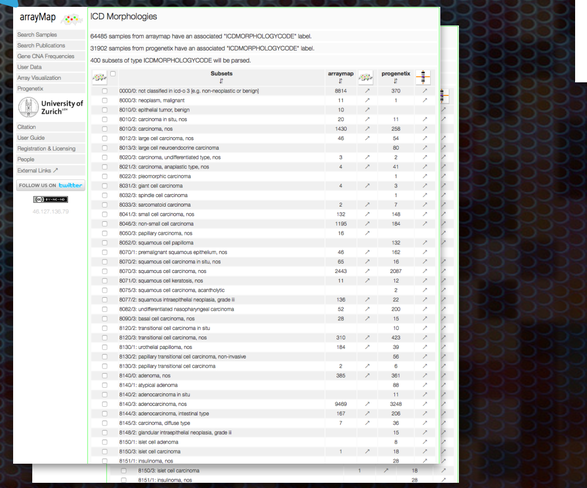












arrayMap

BIOINFORMATICS & CURATION
ARRAYMAP DATA
"PIPELINE"

GEO
ArrayExpress

amDownload

Metadata
(GEO soft files…)

amParseMeta

metadata.ods
(multi-sample)

amUpdateMeta

db.metadata

Array Probe Data
(log2
tables, .CEL)

Metadata Review
Selection
Entity Mapping
(ICD-O, Biosample,
Individual, clinical...)

db.experiments

Publications
Direct
Submissions

Array
Processing
Pipeline

Manual interpretation
of associated
information (article,
series descriprtion...)

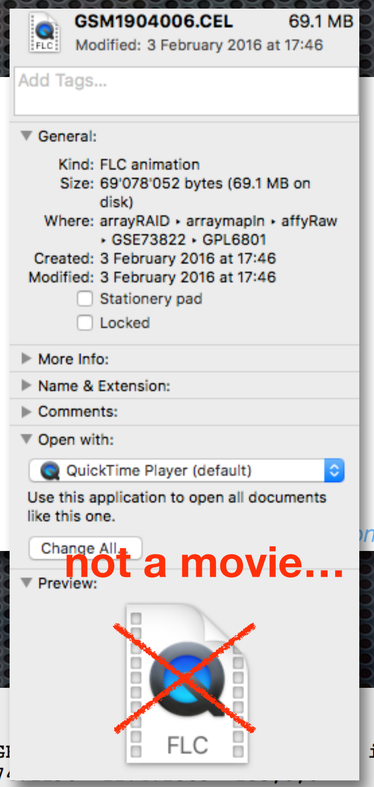
Array
Processing
Pipeline

arrayMap

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

me.ucsc.edu/FAQ/FAQformat.html

```
browser position chr7:127471196-127495720
browser hide all
track name="ItemRGBDemo" description="Item RGB                                    itemRgb="On"
chr7    127471196  127472363  Pos1  0  +  127                                    ,,,
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
```

BED file example

Bioinformatics: File Formats, Ontologies & **API**s

- 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**

http://progenetix.org/api/?db=progenetix&api_out=samples&api_doctype=json&icdm_m=817&randno=20

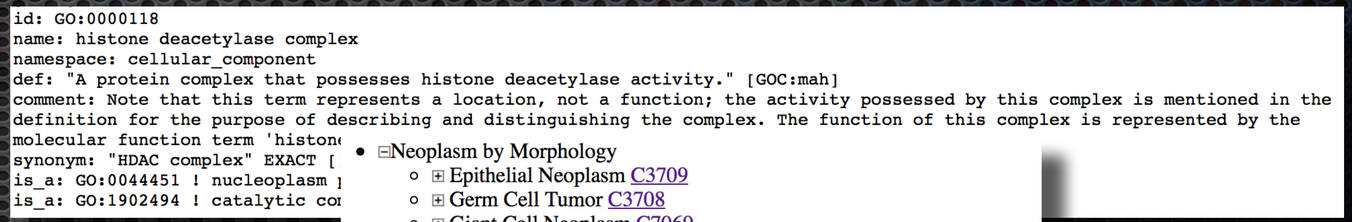

```
{“api_params":{"genome":"hg18", "db":"progenetix", "datatype":"sampledata", "count":20,
"call":"api_doctype=json&api_out=samples&db=progenetix&icdm_m=817&randno=20",
"scope":"samples"}, "data":[{"ICDMORPHOLOGY":"Liver cell adenoma", "NCIT:CODE":"C3758",
"ICDTOPOGRAPHYCODE":"C22", "_id":{"$oid":"558e5c2ead9a82d95838f76c"},
"CLINICALGROUP":"Carcinomas: hepatic ca.", "PMID":"15765123", "FOLLOWUP":"",
"BIOSAMPLEID":"AM_BS__HKCI-C2-DOR", "ICDMORPHOLOGYCODE":"8170/0",
"NCIT:TERM":"Hepatocellular Adenoma", "UID":"HKCI-C2-DOR",
"DIAGNOSISTEXT":"Hepatocellular carcinoma [cell line, doxorubicin resistant subclone]",
"DEATH":"", "ICDTOPOGRAPHY":"liver", "AGE":""}, {"FOLLOWUP":"", "PMID":"14578863",
"ICDMORPHOLOGYCODE":"8170/3", "BIOSAMPLEID":"AM_BS__PHCC-30",
"ICDMORPHOLOGY":"Hepatocellular carcinoma, NOS", "_id":
{"$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":"", "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":{"$oid":"558e5bfcad9a82d95838b62a"},
"CLINICALGROUP":"Carcinomas: hepatic ca."}, {"FOLLOWUP":"", "PMID":"11485905",
"BIOSAMPLEID":"AM_BS__HCChypo-won-H18", "ICDMORPHOLOGYCODE":"8170/3",
"ICDMORPHOLOGY":"Hepatocellular carcinoma, NOS", "CLINICALGROUP":"Carcinomas: hepatic
ca.”, "_id":{"$oid":"558e5c48ad9a82d95839185a"}, "NCIT:CODE":"C3099", "SEX":"male" . . .
```

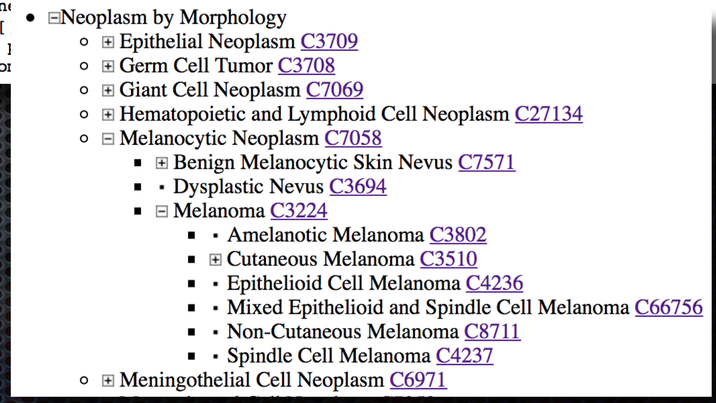
Bioinformatics: File Formats, **Ontologies** & APIs

- 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)





Ontologies in bioinformatics. R Stevens, C Wroe, P Lord and C Goble, Department of Computer Science, University of Manchester. Retrieved from http://homepages.cs.ncl.ac.uk/phillip.lord/download/publications/handbook.pdf

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

source: Colin Dewey (UWisc): "Introduction to Bioinformatics Fall 2016"

BIO390: Introduction to Bioinformatics

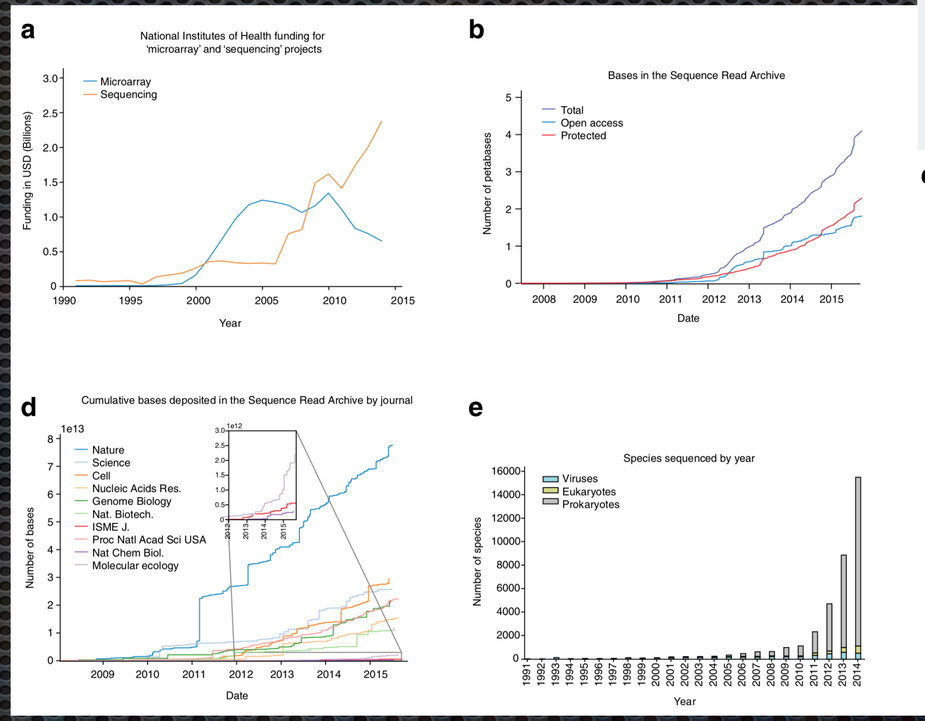
Lecture I: What is Bioinformatics? Examples from Sequencing & Human Genome Variation

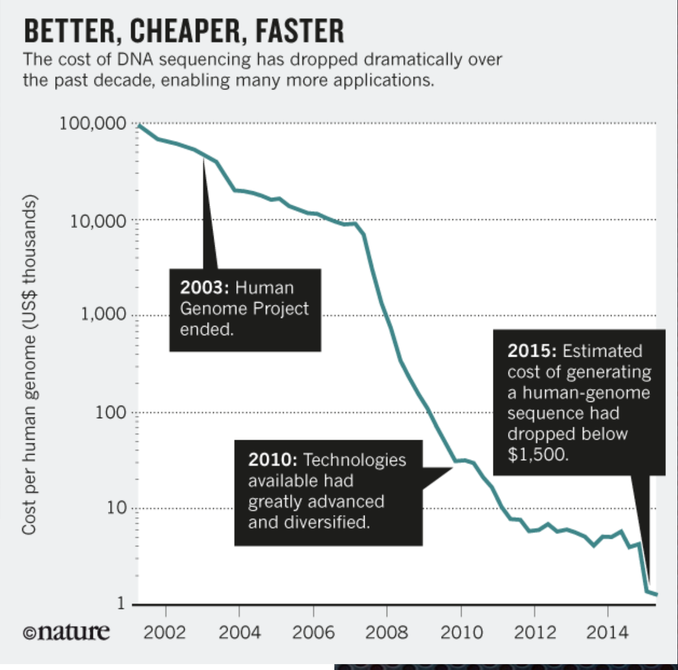
University of Zurich[UZH]

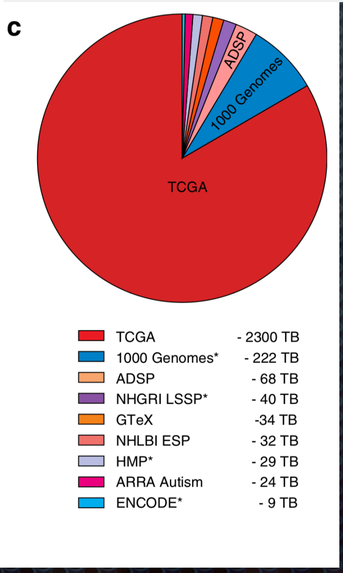
Sequencing everywhere, everything

- dramatic increase in genome sequencing, for human / disease related projects as well as basic biology (species identification, metagenome analyses, evolutionary modeling ...)

Muir et al. Genome Biology (2016) 17:53







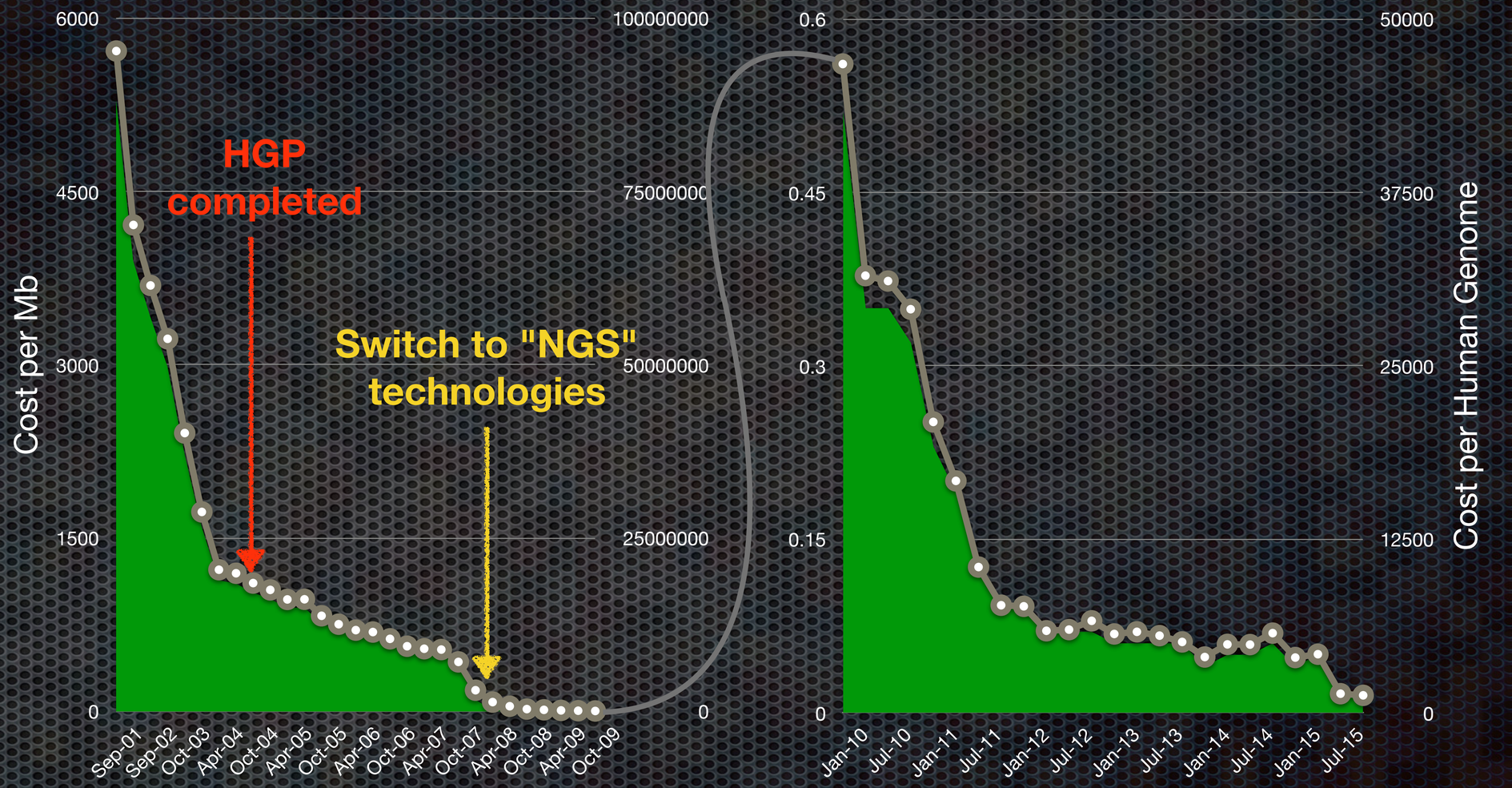
Nature, 2017-12-17

DNA Sequencing Costs: Data

Data from the NHGRI Genome Sequencing Program (GSP)

https://www.genome.gov/sequencingcostsdata/

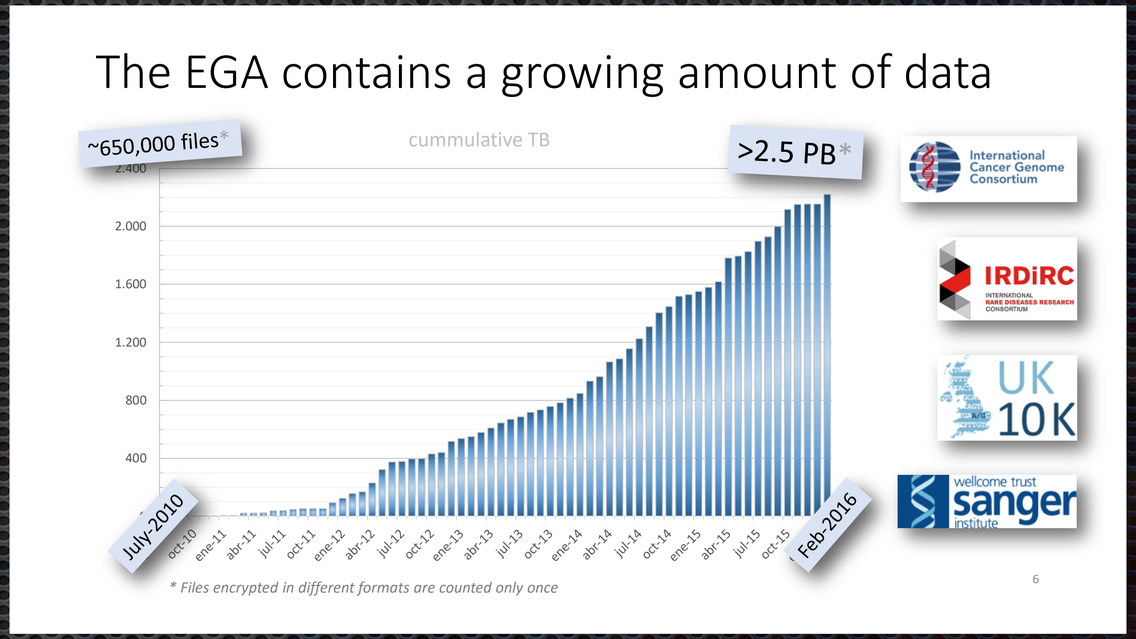
Date	Cost per Mb	Cost per Genome
Sep-01	JS$ 5'292.39	US$ 95'263'072
Mar-02	JS$ 3'898.64	US$ 70'175'437
Sep-02	JS$ 3'413.80	US$ 61'448'422
Mar-03	JS$ 2'986.20	US$ 53'751'684
Oct-03	JS$ 2'230.98	US$ 40'157'554
Jan-04	JS$ 1'598.91	US$ 28'780'376
Apr-04	JS$ 1'135.70	US$ 20'442'576
Jul-04	JS$ 1'107.46	US$ 19'934'346
Oct-04	JS$ 1'028.85	US$ 18'519'312
Jan-05	US$ 974.16	US$ 17'534'970
Apr-05	US$ 897.76	US$ 16'159'699
Jul-05	US$ 898.90	US$ 16'180'224
Oct-05	US$ 766.73	US$ 13'801'124
Jan-06	US$ 699.20	US$ 12'585'659
Apr-06	US$ 651.81	US$ 11'732'535
Jul-06	US$ 636.41	US$ 11'455'315
Oct-06	US$ 581.92	US$ 10'474'556
Jan-07	US$ 522.71	US$ 9'408'739
Apr-07	US$ 502.61	US$ 9'047'003
Jul-07	US$ 495.96	US$ 8'927'342
Oct-07	US$ 397.09	US$ 7'147'571
Jan-08	US$ 102.13	US$ 3'063'820
Apr-08	US$ 15.03	US$ 1'352'982
Jul-08	US$ 8.36	US$ 752'080
Oct-08	US$ 3.81	US$ 342'502
Jan-09	US$ 2.59	US$ 232'735
Apr-09	US$ 1.72	US$ 154'714
Jul-09	US$ 1.20	US$ 108'065
Oct-09	US$ 0.78	US$ 70'333
Jan-10	US$ 0.52	US$ 46'774
Apr-10	US$ 0.35	US$ 31'512
Jul-10	US$ 0.35	US$ 31'125
Oct-10	US$ 0.32	US$ 29'092
Jan-11	US$ 0.23	US$ 20'963
Apr-11	US$ 0.19	US$ 16'712
Jul-11	US$ 0.12	US$ 10'497
Oct-11	US$ 0.09	US$ 7'743
Jan-12	US$ 0.09	US$ 7'666
Apr-12	US$ 0.07	US$ 5'901
Jul-12	US$ 0.07	US$ 5'985
Oct-12	US$ 0.07	US$ 6'618
Jan-13	US$ 0.06	US$ 5'671
Apr-13	US$ 0.06	US$ 5'826
Jul-13	US$ 0.06	US$ 5'550
Oct-13	US$ 0.06	US$ 5'096
Jan-14	US$ 0.04	US$ 4'008
Apr-14	US$ 0.05	US$ 4'920
Jul-14	US$ 0.05	US$ 4'905
Oct-14	US$ 0.06	US$ 5'731
Jan-15	US$ 0.04	US$ 3'970
Apr-15	US$ 0.05	US$ 4'211
Jul-15	US$ 0.015	US$ 1'363
Oct-15	US$ 0.014	US$ 1'245



- Labor, administration, management, utilities, reagents, and consumables
- Sequencing instruments and other large equipment (amortized over three years)
- Informatics activities directly related to sequence production (e.g., laboratory information management systems and initial data processing)
- Submission of data to a public database
- Indirect Costs (http://oamp.od.nih.gov/dfas/faq/indirect-costs#difference) as they relate to the above items

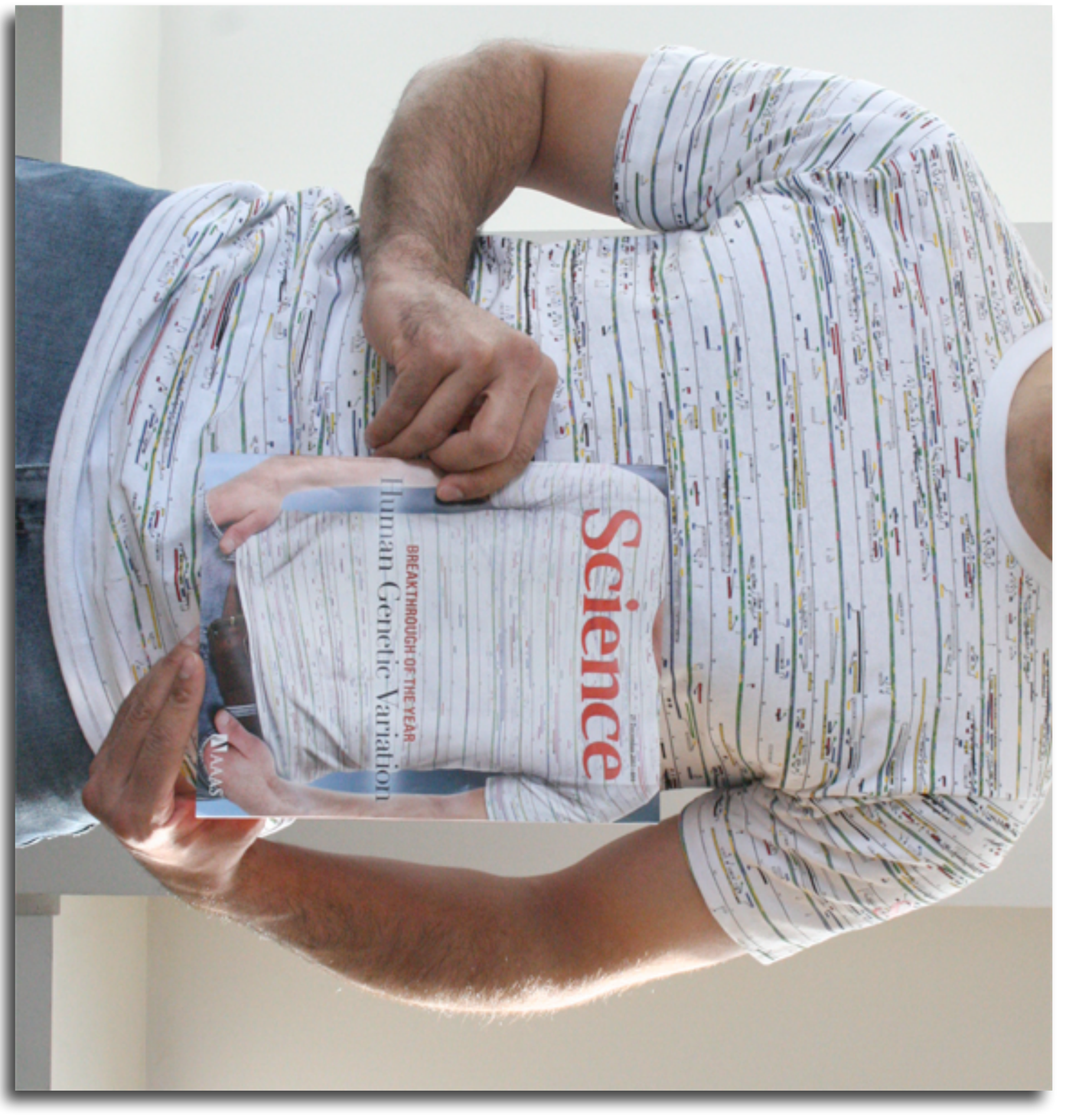
Bioinformatics - Databases & Data Driven Science

- "**post-genomic** era": shift in biology from being descriptive/qualitative to being analytical/quantitative ("**data driven science**")
- typical examples here are genomic screening assays to map genome variants, and statistically associate them with phenotypes/traits (e.g. Genome Wide Association Studies - GWAS)



population based and cancer research studies produce a rapidly increasing amount of genome sequence data - growth of the European Genome - Phenome Archive (EGA)

The trouble with human genome variation



Conclusions from the analysis of variation in the human genome

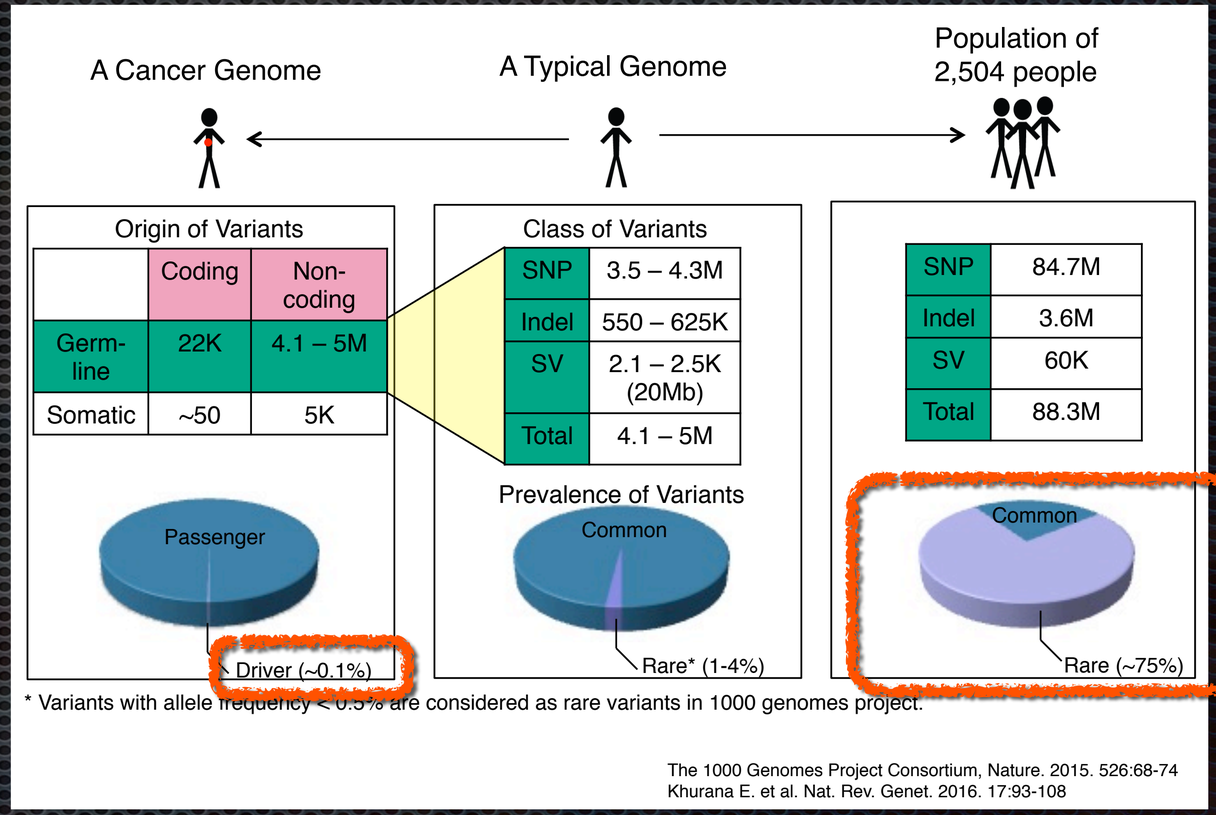
- 1. Humans are all very similar to each other
  - Two humans will show about 99.9% sequence identity with each other. In other words, only about 1 in 1’000 bp is different between two individuals.
  - Humans show about 98% sequence identity to chimps. So two humans are still much more similar to each 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.

Slide from Michael Hengartner’s Human Genome Lecture

Finding Somatic Mutations In Cancer:
Many Needles in a Large Haystack

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



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

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

What is a PB, for human genomes? It depends.

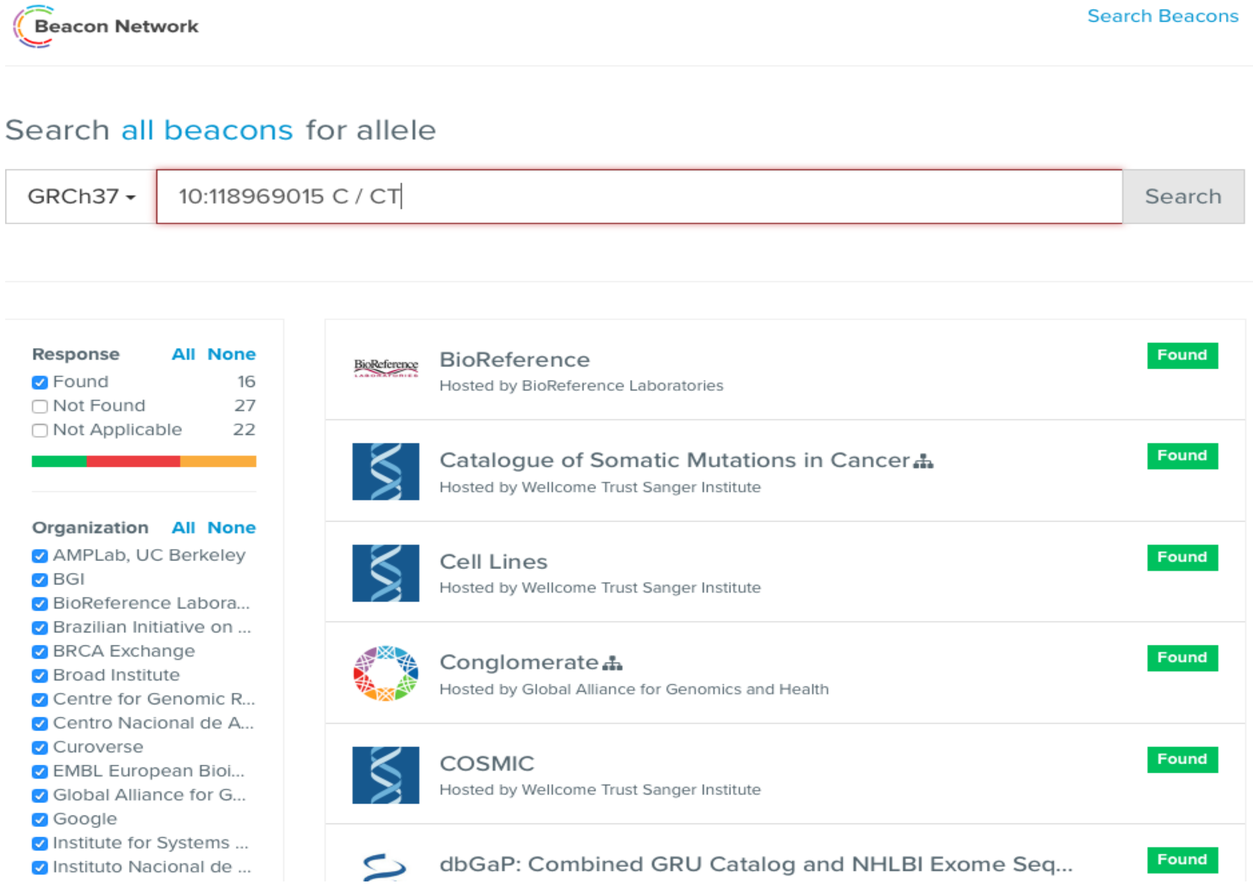
- 2 bits per base are sufficient to encode TCGA
  - using 00, 01, 10, 11
  - [TCGA]{3’000’000’000}
  - 2 * 3 * 10^9b = 6, 000, 000, 000 b
  - perfect genome (no overhead): ~715 MB
  - 1PB => ~1’400’000 genomes
- according to Swiss online store (Sep 2019) ~35'000CHF (65x16TB disks)
- this is less than a PhD position per year in Switzerland …
- (real costs are 2x that, + duplication, facilities, service … => ~500’000CHF)
- **However: A single 30x BAM file => 100GB**
- Still: 500’000CHF => 1PB => 10’000 genomes => 50CHF/genome (BAM format)

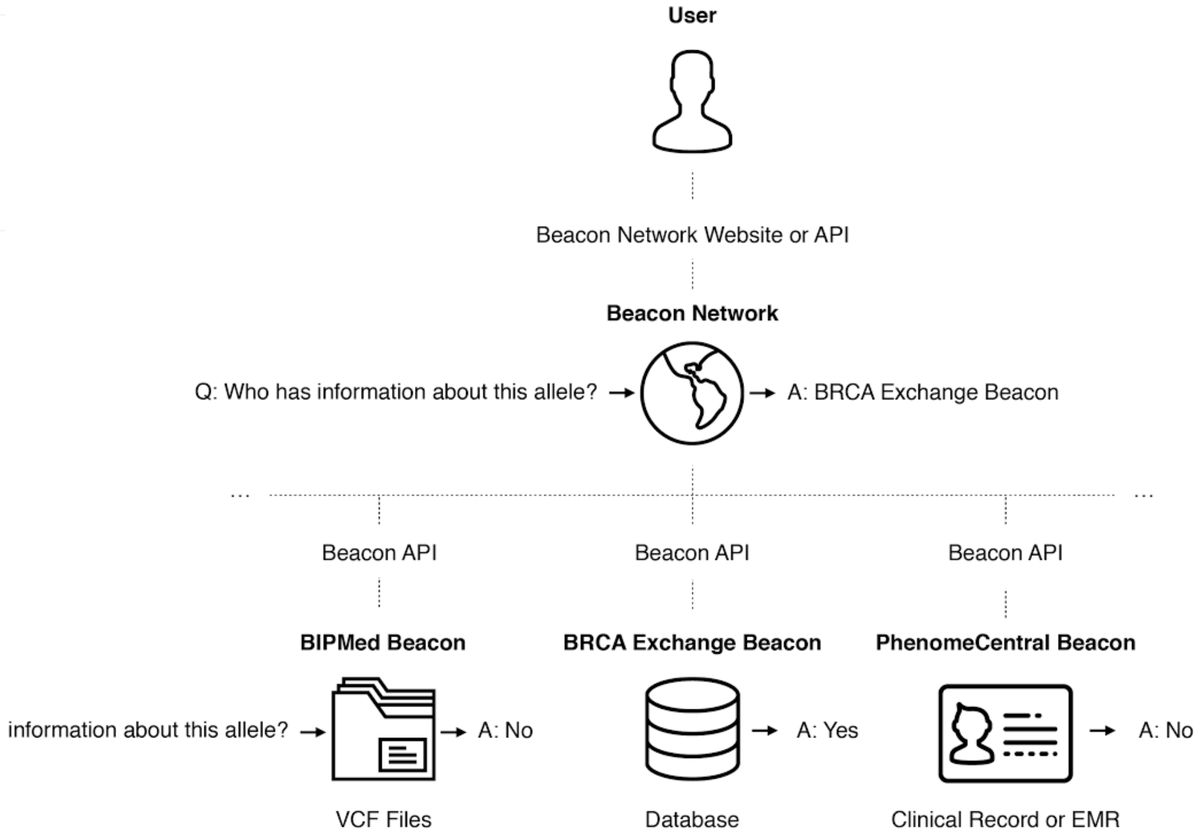
Reference Resources for Human Genome Variants

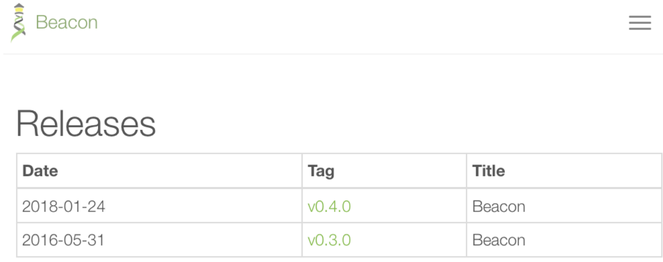
- NCBI:dbSNP
  - single nucleotide polymorphisms (SNPs) and multiple small-scale variations
  - including insertions/deletions, microsatellites, non-polymorphic variants
- NCBI:dbVAR
  - genomic structural variation
  - insertions, deletions, duplications, inversions, multinucleotide substitutions, mobile element insertions, translocations, complex chromosomal rearrangements
- NCBI:ClinVar
  - aggregates information about genomic variation and its relationship to human health
- EMBL-EBI:EVA
  - open-access database of all types of genetic variation data from all species
- Ensembl
  - portal for many things genomic...

Beacon Project

An open web service that tests the willingness of international sites to share genetic data.







Global Alliance for Genomics & Health

35+ Organizations

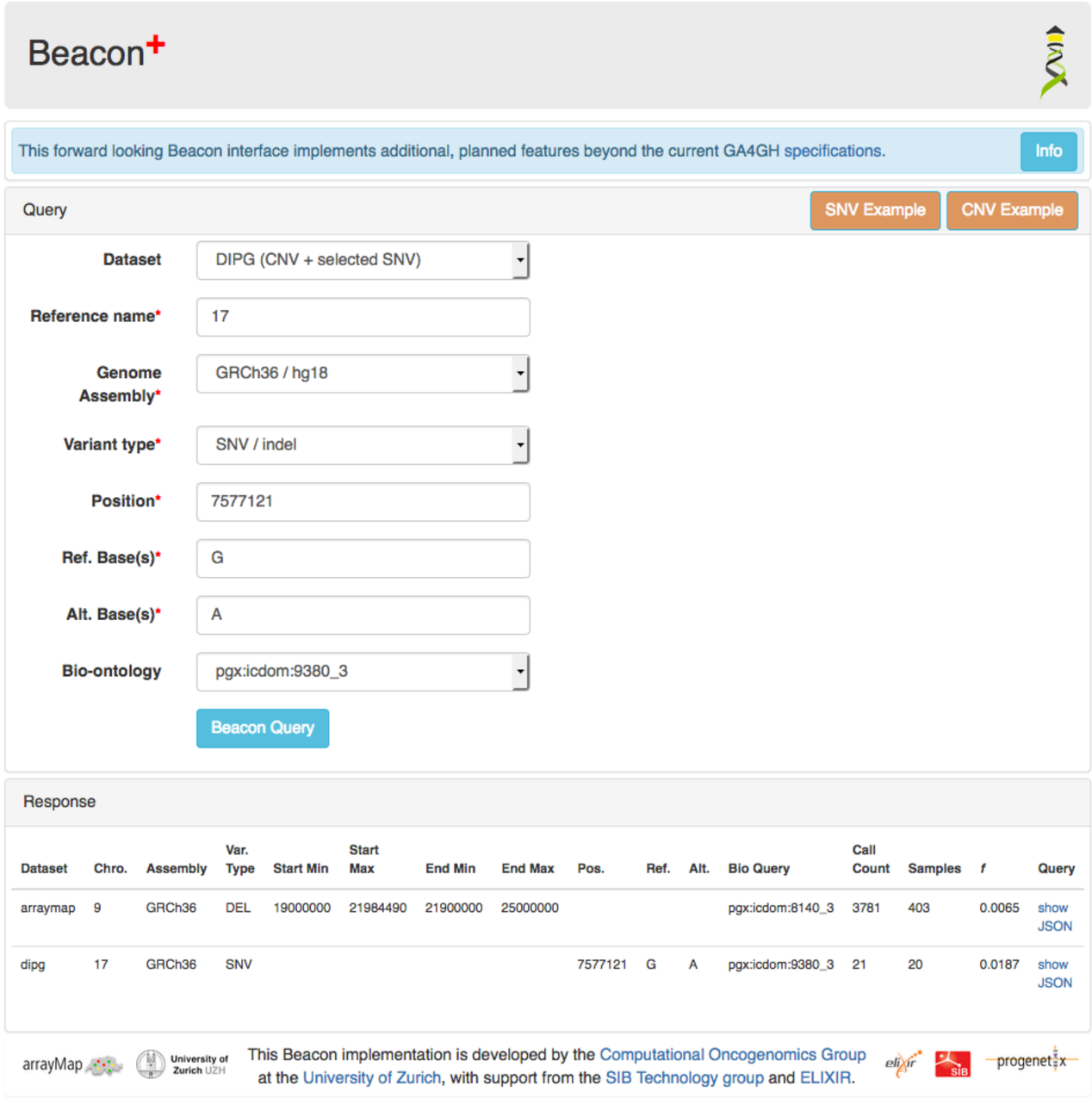
90+ Beacons

200+ Datasets

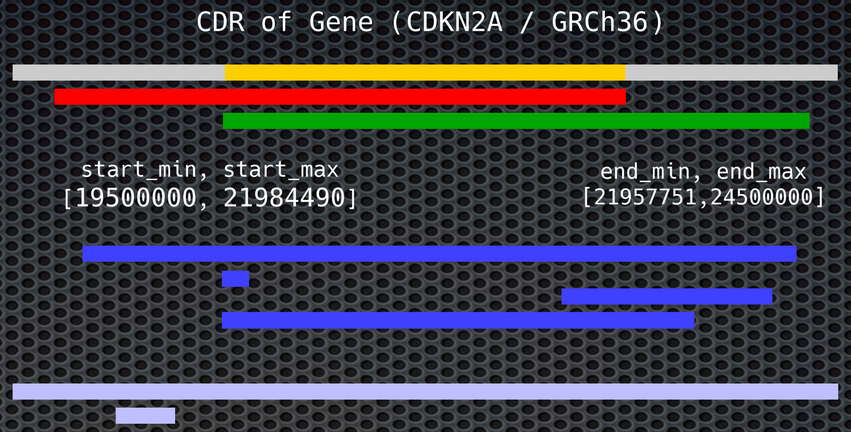
100K+ Individuals

Beacon+ Concept

- Implementation of cancer beacon prototype, backed by arrayMap and DIPG data set (MacKay *et al.*, Cancer Cell 2017, in print)
- structural variations (DUP, DEL) in addition to SNV
- diagnosis queries using ontology codes (NCIT, ICD-O)
- quantitative responses
- current version uses **GA4GH schema compatible** database



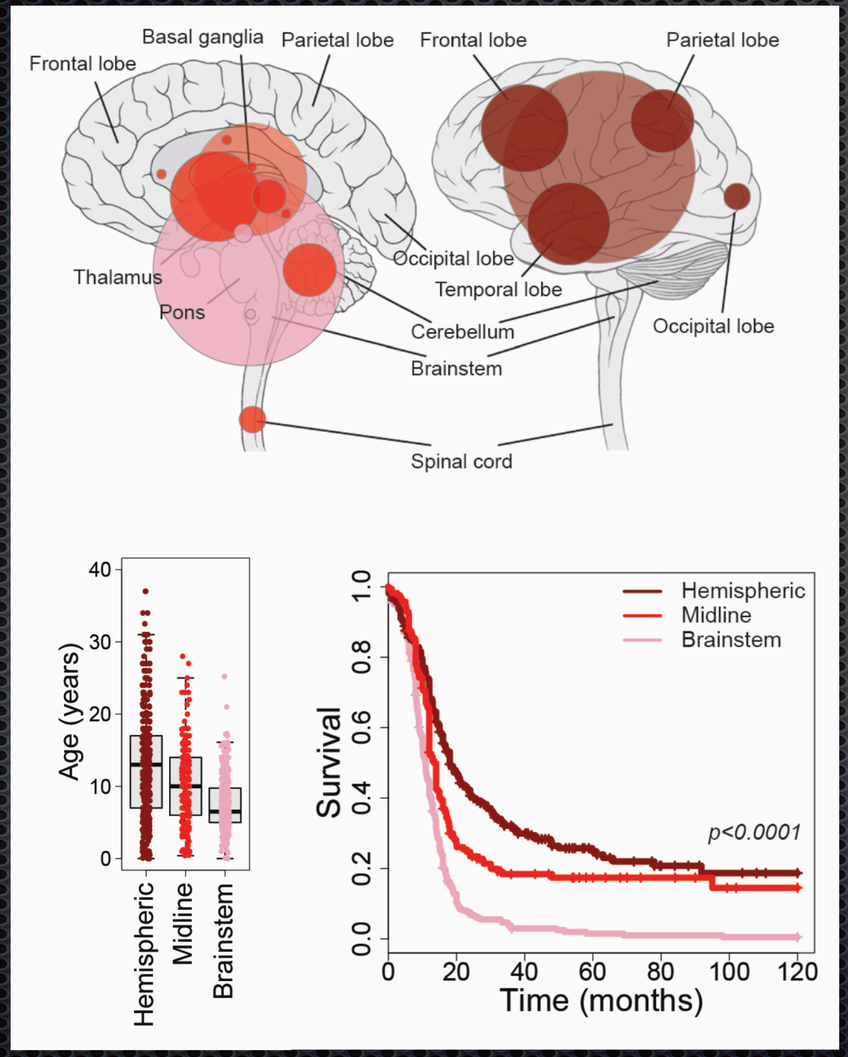
```
    { "reference_name" : "9" },
    { "variant_type" : "DEL" },
    { "start" : { "$gte" : 19500000 } },
    { "start" : { "$lte" : 21984490 } },
    { "end" : { "$gte" : 21957751 } },
    { "end" : { "$lte" : 24500000 } }
  ]
 },
 "api_version" : "0.4",
  "beacon_id" : "org.progenetix:progenetix-beacon",
  "exists" : true,
  "info" : {
   "query_string" :
"dataset_id=arraymap&variants.reference_name=chr9&assembly_id=GRCh36&va
riants.variant_type=DEL&variants.start_max=19000000&variants.start_min=
21984490&variants.end_min=21900000&variants.end_max=25000000&biosamples
.bio_characteristics.ontology_terms.term_id=pgx:icdom:9440_3",
   "version" : "Beacon+ implementation based on a development branch of
the beacon-team project: https://github.com/ga4gh/beacon-team/pull/94"
  },
  "url" : "http://progenetix.org/beacon/info/",
  "dataset_allele_responses" : [
   {
    "dataset_id" : "arraymap",
    "error" : null,
    "exists" : true,
    "external_url" : "http://arraymap.org",
    "sample_count" : 584,
    "call_count" : 3781,
    "variant_count" : 3244,
    "frequency" : 0.0094,
    "info" : {
     "description" : "The query was against database
\"arraymap_ga4gh\", variant collection \"variants_cnv_grch36\". 3781 /
59428 matched callsets for 3602919 variants. Out of 62105 biosamples in
the database, 2047 matched the biosample query; of those, 584 had the
variant.",
     "ontology_ids" : [
      "ncit:C3058",
      "pgx:icdom:9440_3",
      "pgx:icdot:C71.9",
      "pgx:icdot:C71.0"
     ]
```

- Beacon+ **range queries** allow the definition of a genome region of interest, containing a specified variant or potentially other position related feature
- “fuzzy” matching of region ends essential for inexact features
- quantitative reporting

Implementing real-world datasets for federated access using GA4GH schema specifications: pHGG

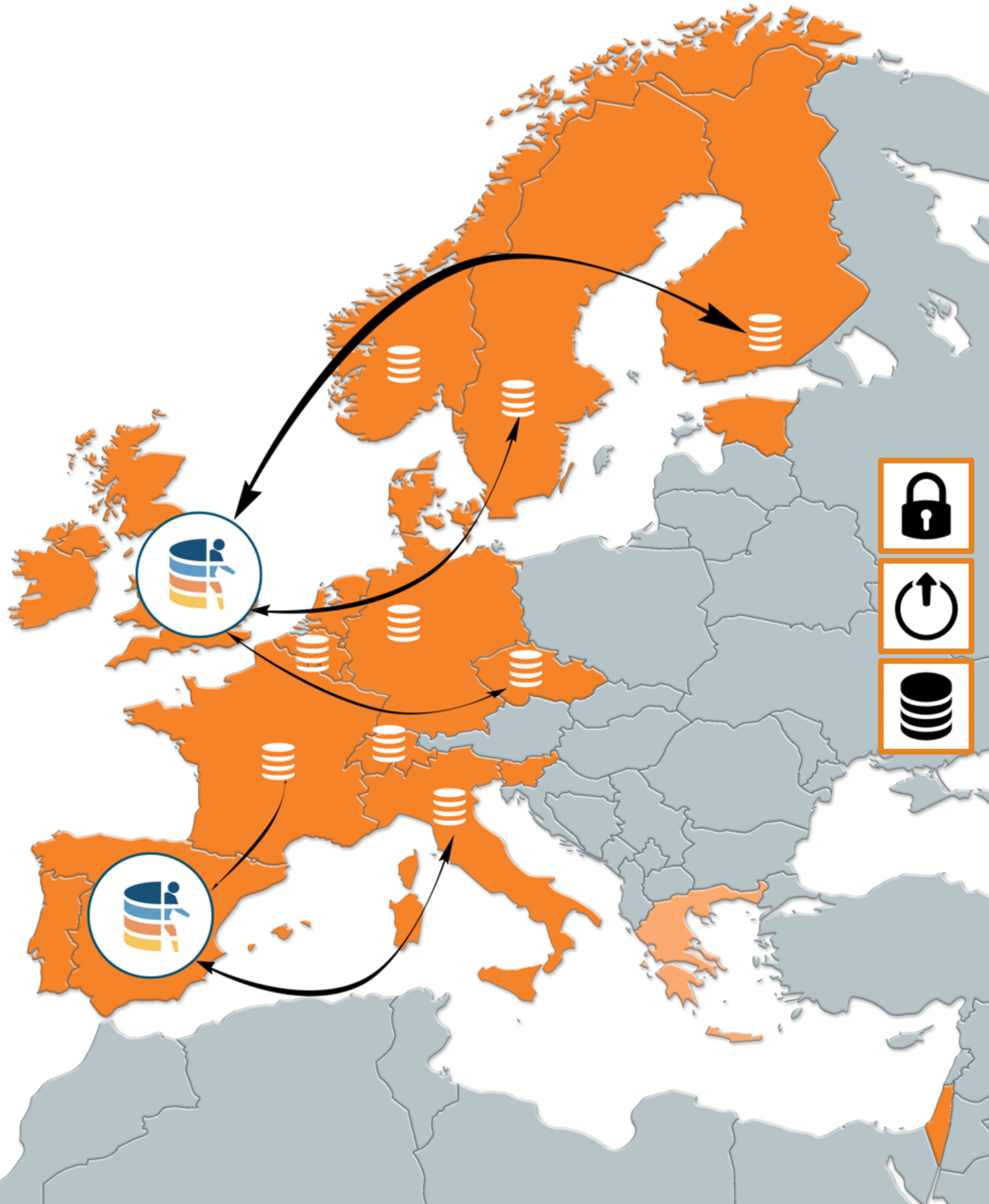
Mackay A, Jones C, Baudis M and many, many others: Integrated molecular meta-analysis of 1000 paediatric high grade and diffuse intrinsic pontine glioma (2017, Cancer Cell, in press)




```
"_id" : ObjectId("591eb73709d374421fc0b2a"),
"individual_id" : "DIPG_IND_0809",
"id" : "DIPG_BS_0809",
"name" : "pHGG_META_0809",
"description" : "glioma, paediatric, high grade",
"individual_age_at_collection" : {
   "age_class" : {
      "term" : "Adult onset",
      "term_id" : "HP:0003581"
   },
   "age" : "P17Y0M"
},
"bio_characteristics" : [
   {
      "ontology_terms" : [
         {
            "term_label" : "Glioma",
            "term_id" : "ncit:C3059"
         },
         {
            "term_label" : "Brain NOS",
            "term_id" : "pgx:icdot:C71.9"
         }
      ],
      "description" : "Juvenile high grade glioma",
   }
],
"external_identifiers" : [
   {
      "database" : "Pubmed",
      "identifier" : "25752754",
      "relation" : "reported_in"
   }
],
"attributes" : {
   "grade" : { "values" : [ { "string_value" : "4" } ] },
   "histone" : { "values" : [ { "string_value" : "wt" } ] },
}
```

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)



ELIXIR Human Genomics & Translational Data - technically

- Data
- Training
- Compute
- Interoperability
- Tools

elixir AAI

ELIXIR Beacon - GA4GH Driver Project

Data Discoverability

Federating lightweight discoverability of data, and datasets across ELIXIR

EUROPEAN GENOME-PHENOME ARCHIVE

Data Archival

Utilising the ELIXIR Deposition Databases to ensure secure, long-term, efficient archival of data

ELIXIR Federated Human Data Community - htsget/htsref

Federated Data Access

Coordinating a collection of interoperable EGA-like resources to ensure secure management of sensitive data across the ELIXIR Nodes

trait a Lygature platform

RD Connect

bio.tools

Data Analysis

Bringing 'analysis to data' via common workflow languages, workflows, containers, and tools

Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.

elixir

GA4GH API promotes sharing

GENOMICS

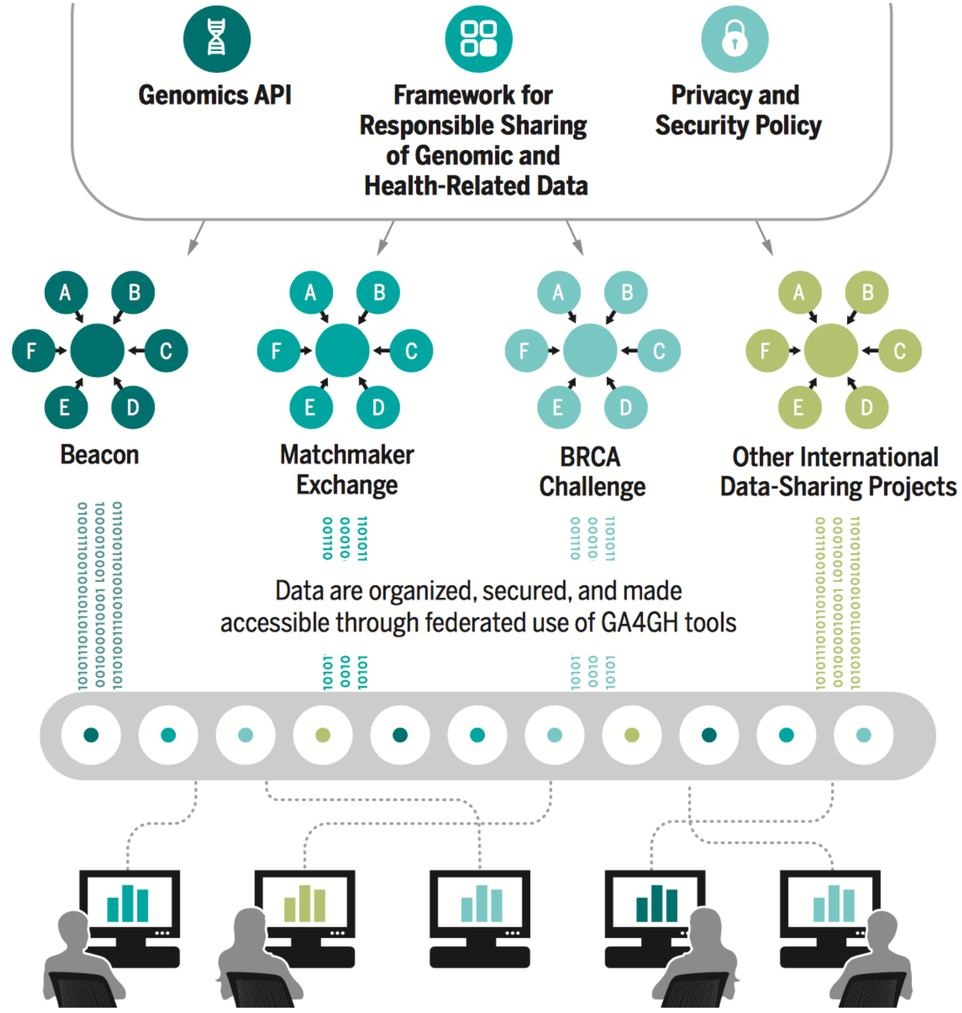
A federated ecosystem for sharing genomic, clinical data

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

The Global Alliance for Genomics and Health*

SCIENCE 10 JUNE 2016 • VOL 352 ISSUE 6291

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



Global Alliance for Genomics & Health

BIO390: Course Schedule

- 2019-09-17: Michael Baudis - What is Bioinformatics? Introduction and Resources
- **2019-09-24: Christian von Mering - Sequence Bioinformatics**
- 2019-10-01: Shinichi Sunagawa (ETHZ) - Metagenomics
- 2019-10-08: Mark Robinson - Statistical Bioinformatics
- 2019-10-15: Izaskun Mallona - Regulatory Genomics and Epigenomics
- 2019-10-22: Abdullah Kahraman (USZ) - Molecular Interaction Networks
- 2019-10-29: Katja Baerenfaller (SIAF) - Proteomics
- 2019-11-05: Amedeo Caflisch - Molecular Dynamics
- 2019-11-12: Elif Ozkirimli - Protein Structure and Interactions
- 2019-11-19: Christophe Dessimoz (UniL) - Sequence evolution and phylogenetics
- 2019-11-26: Valérie Barbie (SIB) - Clinical Bioinformatics
- 2019-12-03: Andreas Wagner - Biological Networks
- 2019-12-10: Alex Handler Wagner (WUSTL) & Michael Baudis - Human Genome Variation Resources
- 2019-12-17: Exam (Multiple Choice)

University of Zurich[UZH]

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

arraymap.org

progenetix.org

info.baudisgroup.org

sib.swiss/baudis-michael

imls.uzh.ch/en/research/baudis

Global Alliance
for Genomics & Health