

BIO390: Introduction to Bioinformatics

Lecture I: What is Bioinformatics?

Michael Baudis - 2018-09-18



University of
Zurich^{UZH}

BIO390: Course Schedule

- 2018-09-18: Michael Baudis - **What is Bioinformatics?** Introduction and Resources
- 2018-09-25: Christian von Mering - **Sequence Informatics**
- 2018-10-02: Mark Robinson - **Statistical Bioinformatics**
- 2018-10-09: Katja Bärenfaller - **Proteomics**
- 2018-10-16: Izaskun Mallona - **Regulatory Genomics and Epigenomics**
- 2018-10-23: Daniel Stekhoven - **Clinical Bioinformatics**
- 2018-10-30: Shinichi Sunagawa - **Metagenomics**
- 2018-11-06: Amedeo Caflisch - **Molecular Dynamics**
- 2018-11-13: Elif Ozkirimli Olmez - **Protein Interactions**
- 2018-11-20: Fabio Rinaldi - **Text Mining I**
- 2018-11-27: Fabio Rinaldi - **Text Mining II**
- 2018-12-04: Andreas Wagner - **Biological Networks**
- 2018-12-11: Michael Baudis - **Human Genome Data - Risks & Opportunities || Feedback**
- 2018-12-18: Michael Baudis - **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 OLAT
- written exam at end of course (== 14th course - December 19th)
- 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

Some Recommended Books

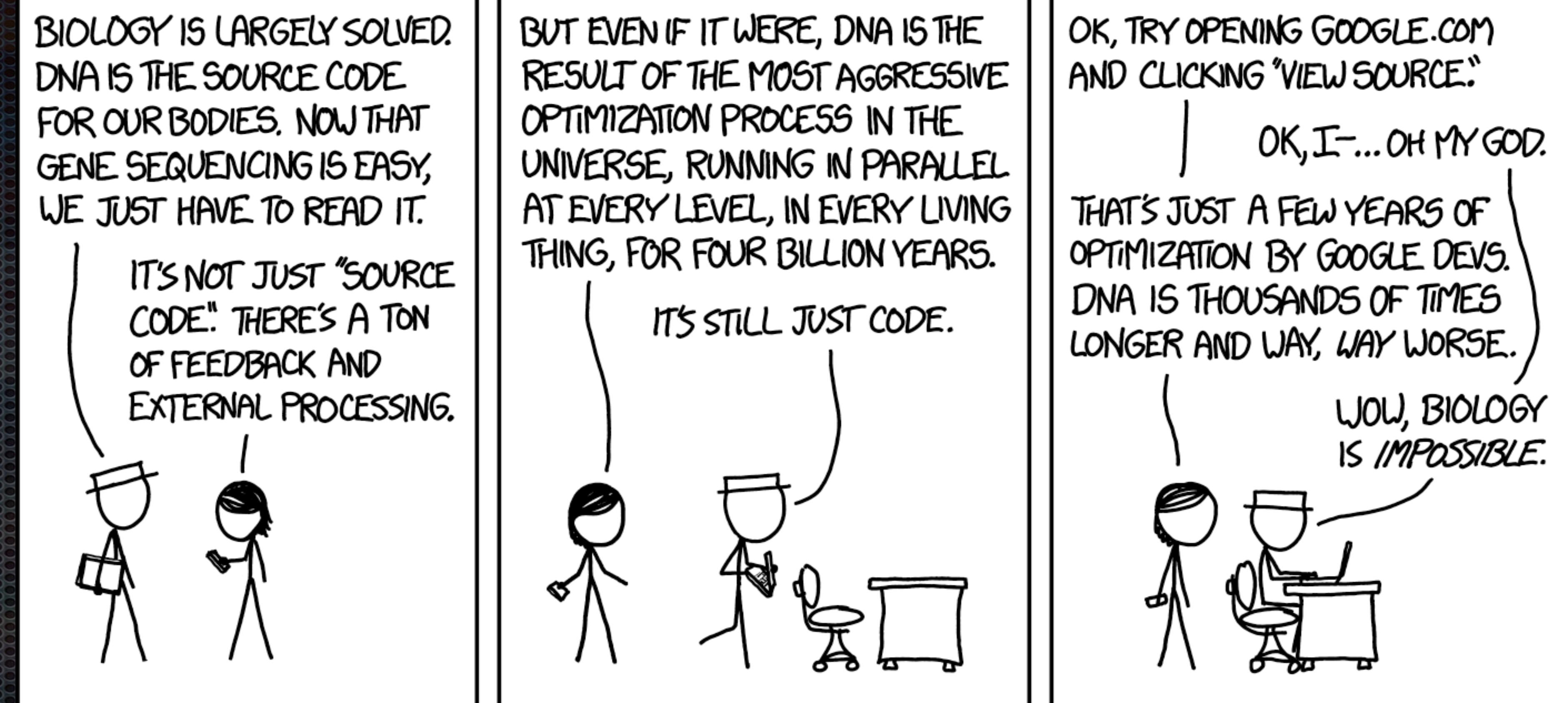
- Anna Tramontano: Introduction to Bioinformatics
- Robert Gentleman: R Programming for Bioinformatics
- John Maindonald & W. John Braun: Data Analysis and Graphics Using R
- Andy Hector: The New Statistics with R
- Neil C. Jones & Pavel A. Pevzner: Bioinformatics Algorithms
- Edward Tufte: The Visual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



Why Bioinformatics?

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

Biology is *impossibly* complex - But bioinformatics might help



So, What is Bioinformatics?

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

What is Bioinformatics?



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

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

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



What is Bioinformatics?

Bioinformatics **uses** informatics tools for analyses

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

What is Bioinformatics?

Bioinformatics **develops** informatics tools for analyses

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

What is Bioinformatics?

biological data

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

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

What is Bioinformatics?



Bioinformatics **analyzes**

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

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

What is Bioinformatics?



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

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

What is Bioinformatics?

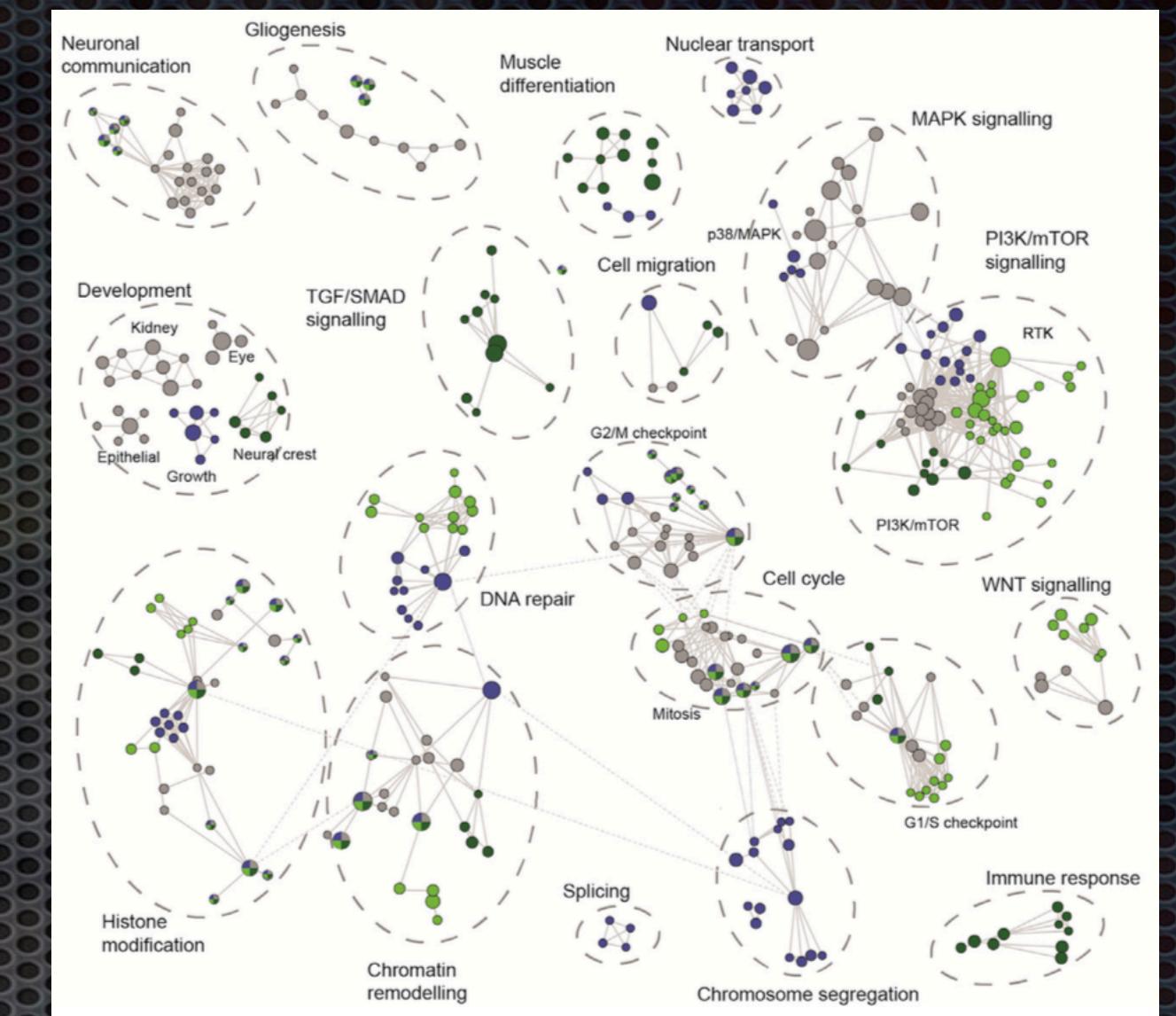
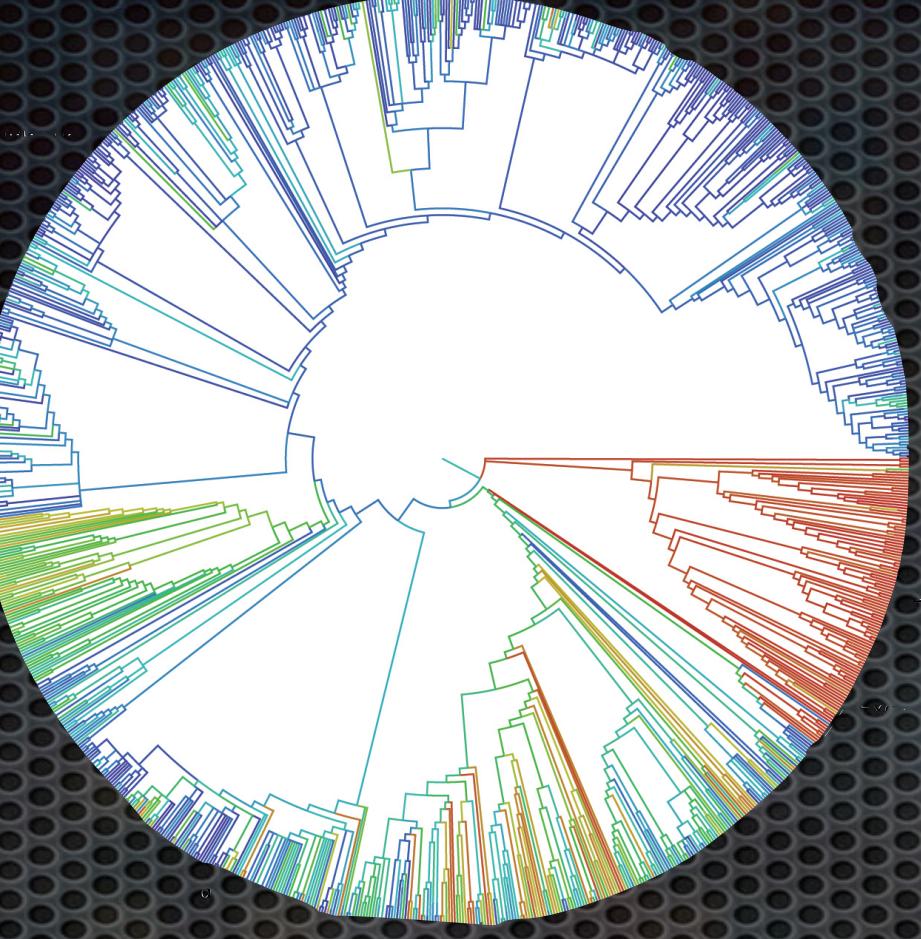


- Bioinformatics is "the science that uses the instruments of informatics to analyze biological data in order to **formulate hypotheses** about life." (Anna Tramontano)

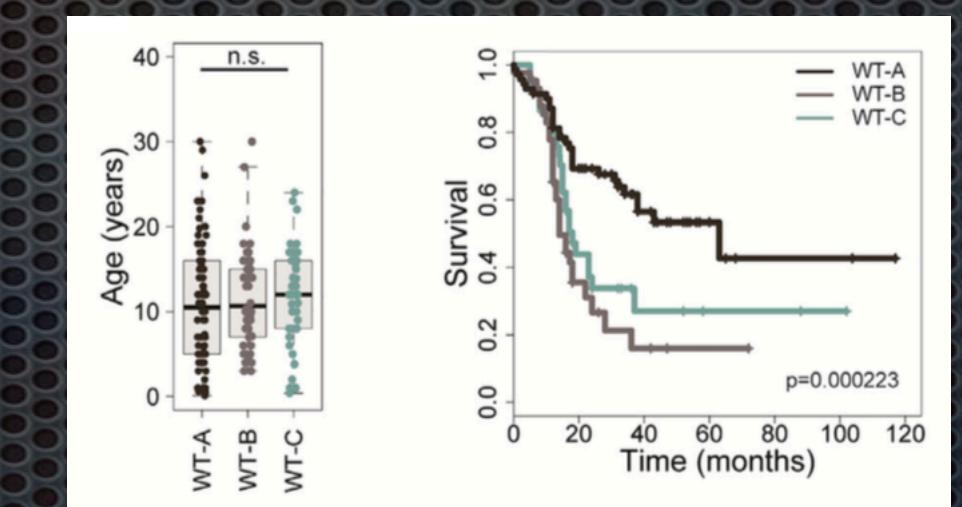
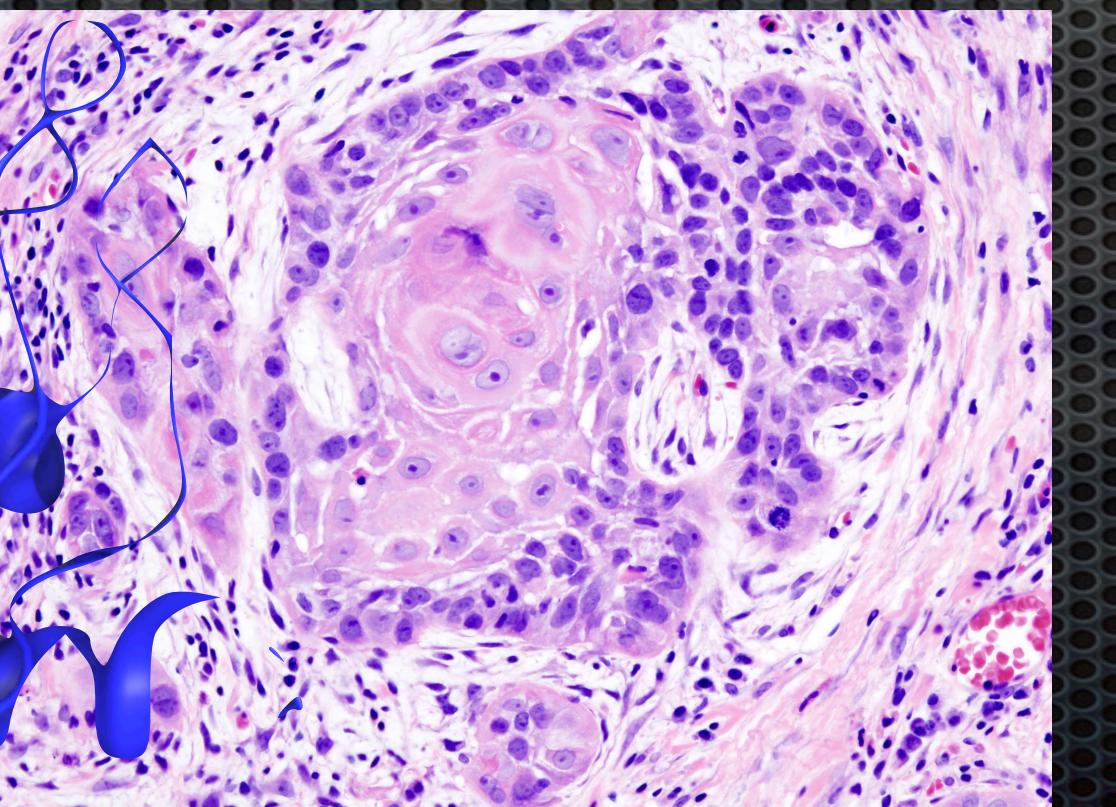
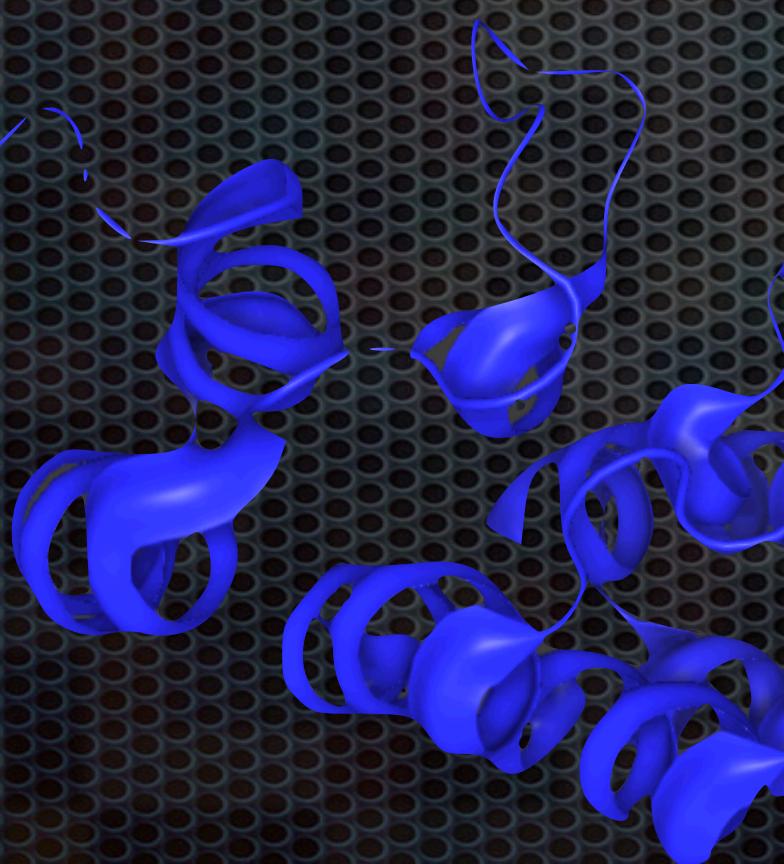
b : an interpretation of a practical situation or condition taken as the ground for action



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

Who is a Bioinformatician?

Bioinformatician

- strong biological knowledge
- provides hypothesis and/or dataset
- sufficient statistical and computational expertise to correctly use bioinformatics tools & develop workflows (scripting ...)
- expert user of informatics tools
- may get a Nobel

Bioinformatician

- sufficient biological background
- provides statistical, analysis methods
- sufficient biologic or medical background to understand problems presented and identify pitfalls and hidden biases arising from data generation methods
- developer of informatics tools
- may get rich

Who is a Bioinformatician?

Bioinformatician

- strong biological knowledge
- provides hypothesis and/or dataset
- sufficient statistical and computational expertise to correctly use bioinformatics tools & develop workflows (scripting ...)
- expert user of informatics tools
- may get a Nobel

Bioinformatician

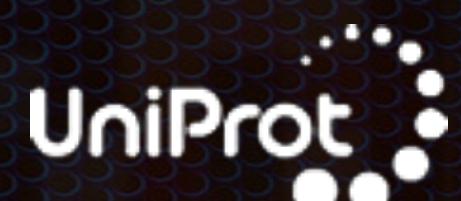
- sufficient biological background
- provides statistical, analysis methods
- sufficient biologic or medical background to understand problems presented and identify pitfalls and hidden biases arising from data generation methods
- developer of informatics tools
- may get rich

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...)

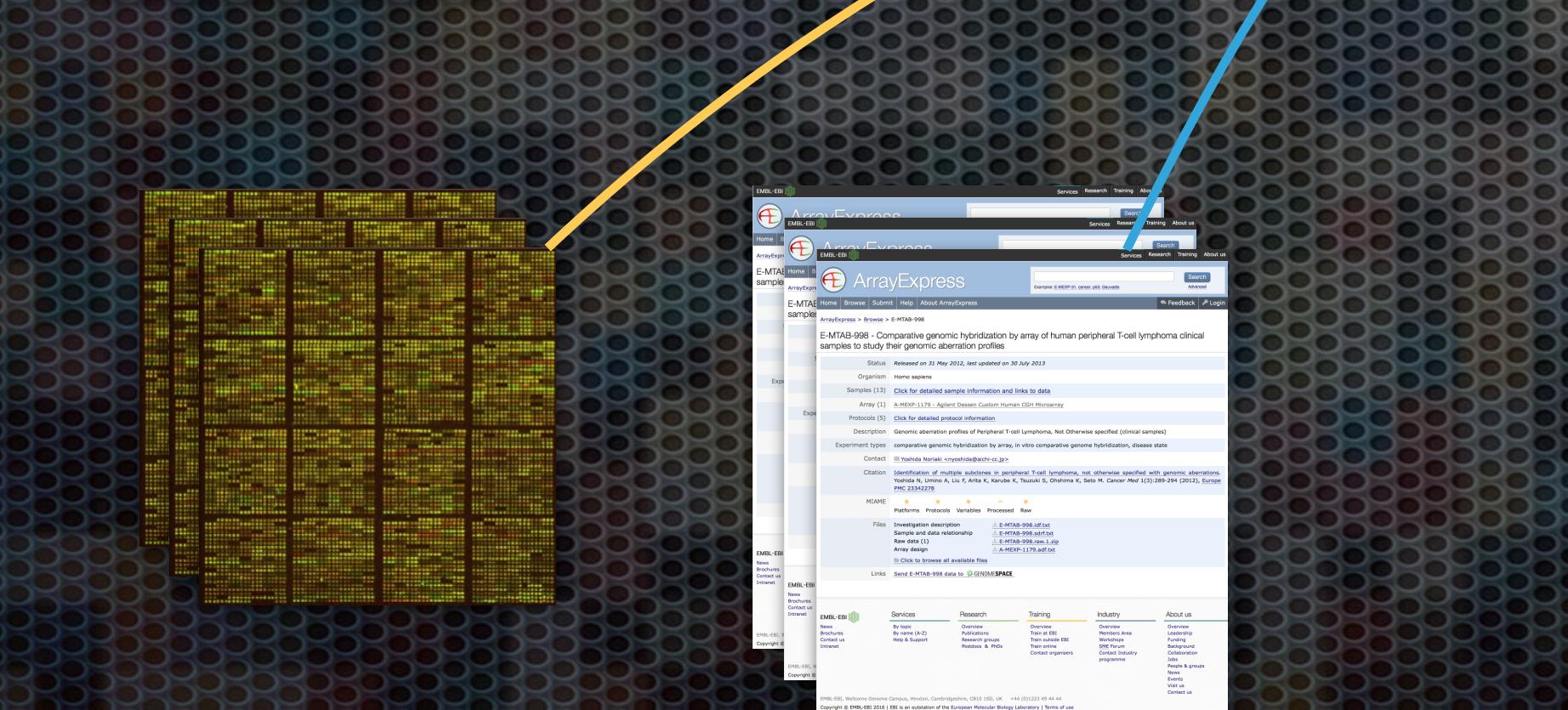
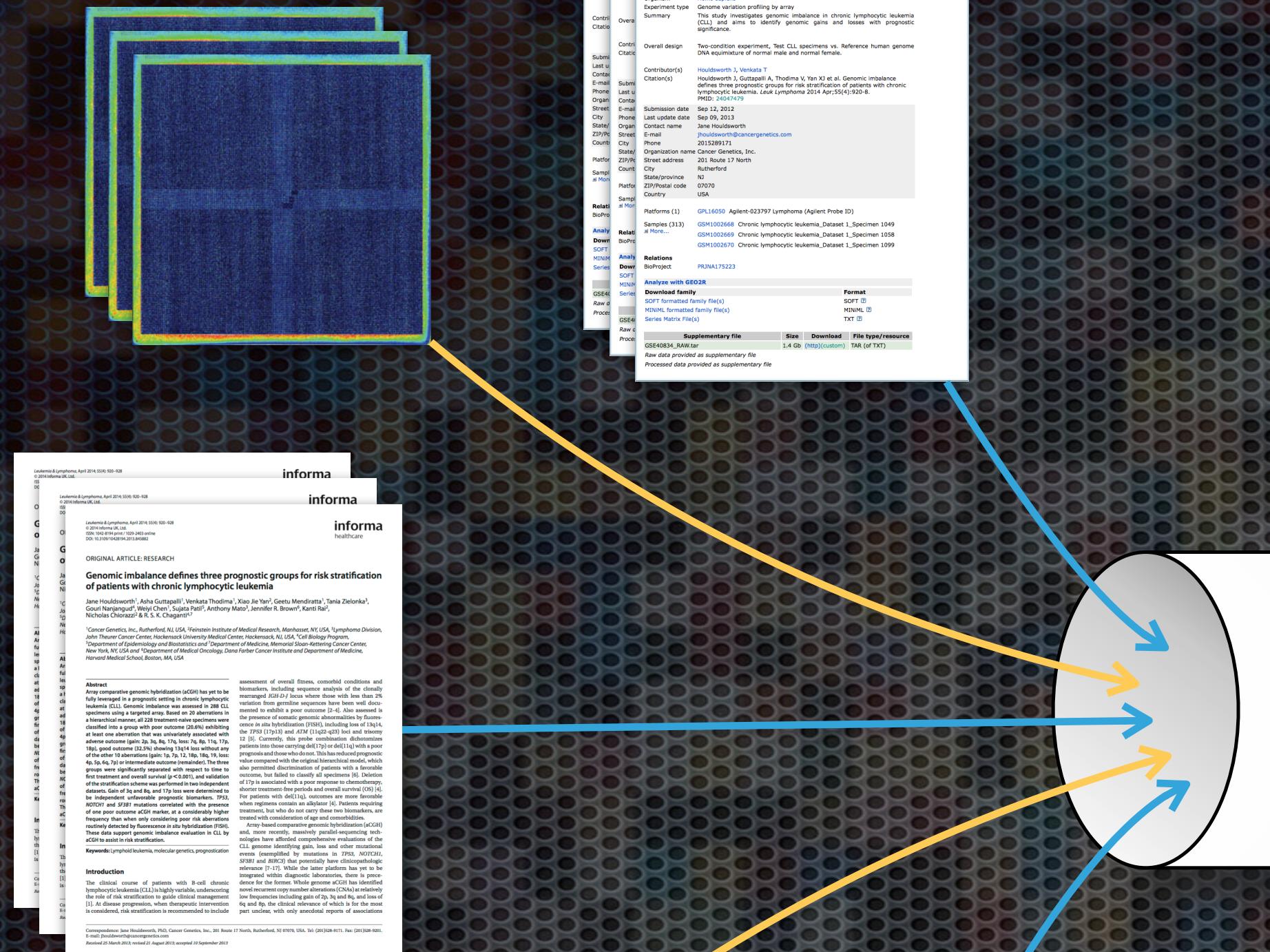


SRA



ARRAYMAP DATA PIPELINE

BIOCURATION BIOINFORMATICS

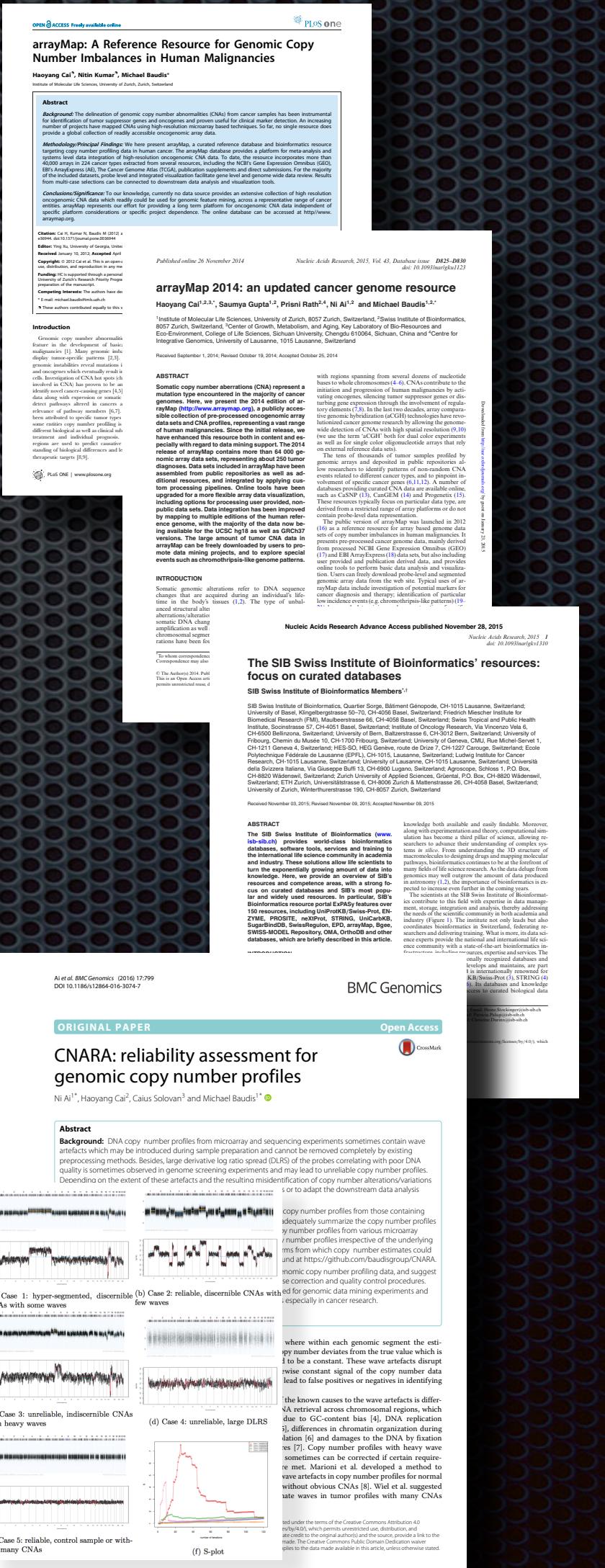
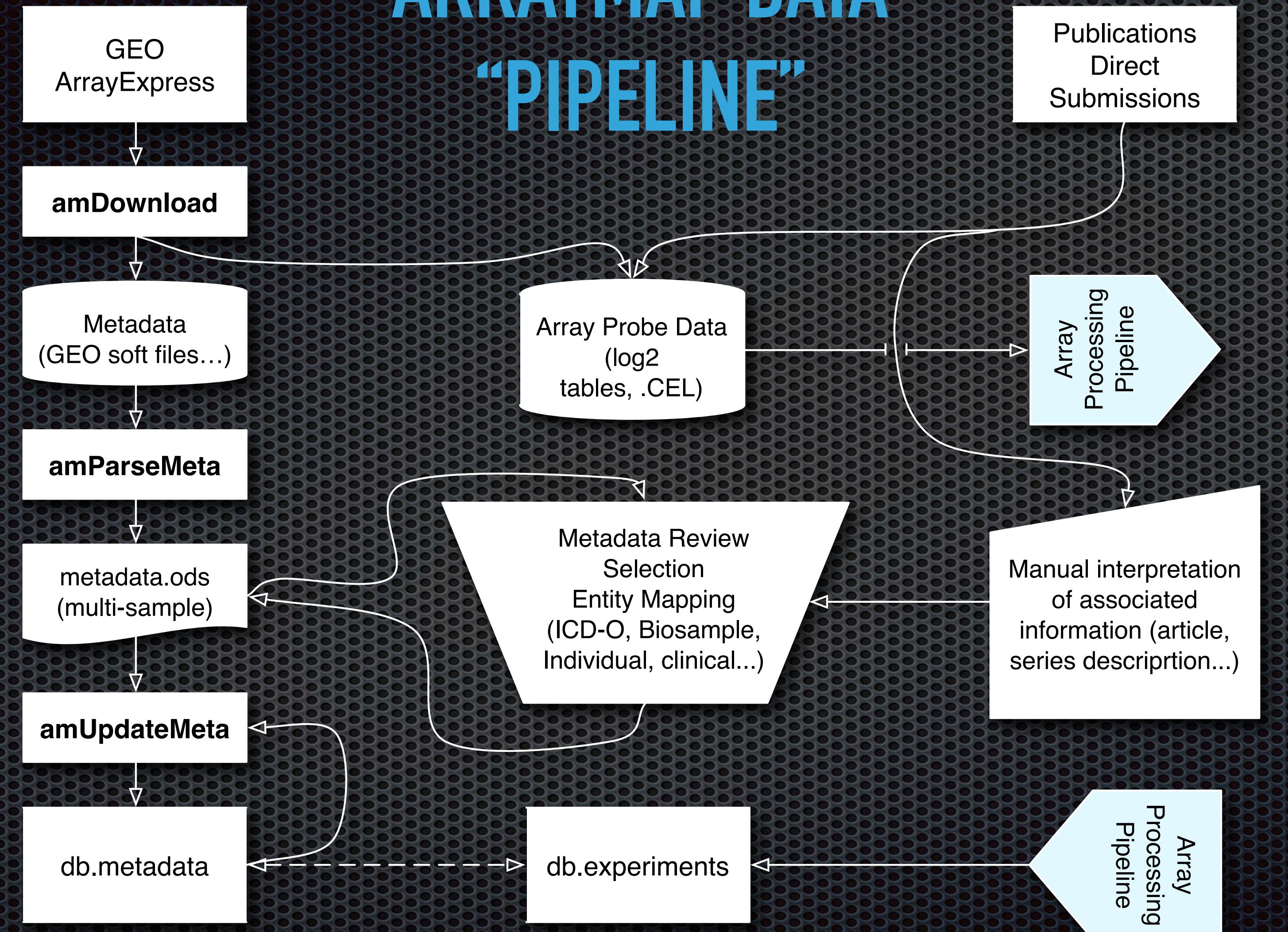


arrayMap



BIOINFORMATICS & CURATION

ARRAYMAP DATA “PIPELINE”



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

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

The image consists of three main parts:

- Top Left:** A screenshot of a Mac OS X file info dialog for a file named "GSM1904006.CEL". The file is 69.1 MB and was modified on 3 February 2016 at 17:46. The "General" section shows it is an FLC animation. The "Preview" section shows a small thumbnail with a large red "X" over it, labeled "not a movie...".
- Bottom Left:** A screenshot of a BED file example. It shows a list of genomic tracks for chromosome 7, each with start, end, strand, and itemRGB values. The first few lines are:

```
browser position chr7:127471196-127495720
browser hide all
track name="ItemRGBDemo" description="Item RGB"
chr7 127471196 127472363 Pos1 0 + 127472363
chr7 127472363 127473530 Pos2 0 + 127472363 127473530 255,0,0
chr7 127473530 127474697 Pos3 0 + 127473530 127474697 255,0,0
chr7 127474697 127475864 Pos4 0 + 127474697 127475864 255,0,0
chr7 127475864 127477031 Neg1 0 - 127475864 127477031 0,0,255
chr7 127477031 127478198 Neg2 0 - 127477031 127478198 0,0,255
chr7 127478198 127479365 Neg3 0 - 127478198 127479365 0,0,255
chr7 127479365 127480532 Pos5 0 + 127479365 127480532 255,0,0
chr7 127480532 127481699 Neg4 0 - 127480532 127481699 0,0,255
```
- Right Column:** A vertical list of various file formats, each preceded by a small blue square icon.
 - 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

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**

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": {"$oid": "558e5bfccad9a82d95838b62a"}, "CLINICALGROUP": "Carcinomas: hepatic ca."}, {"FOLLOWUP": "", "PMID": "11485905", "BIOSAMPLEID": "AM_BS_HCChypo-won-H18", "ICDMORPHOLOGYCODE": "8170/3", "ICDMORPHOLOGY": "Hepatocellular carcinoma, NOS", "CLINICALGROUP": "Carcinomas: hepatic ca.", "_id": {"$oid": "558e5c48ad9a82d95839185a"}, "NCIT:CODE": "C3099", "SEX": "male"}]
```

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

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

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

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



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

- □ 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](#)

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

BIO390: Introduction to Bioinformatics

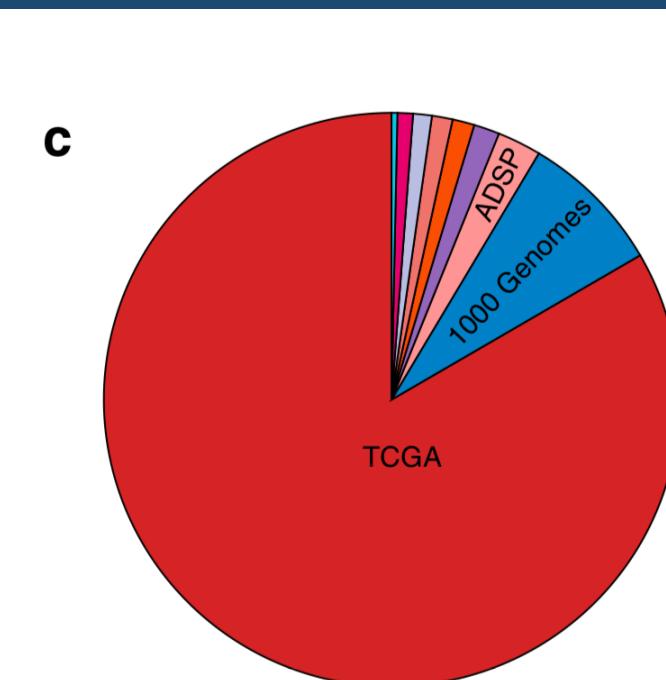
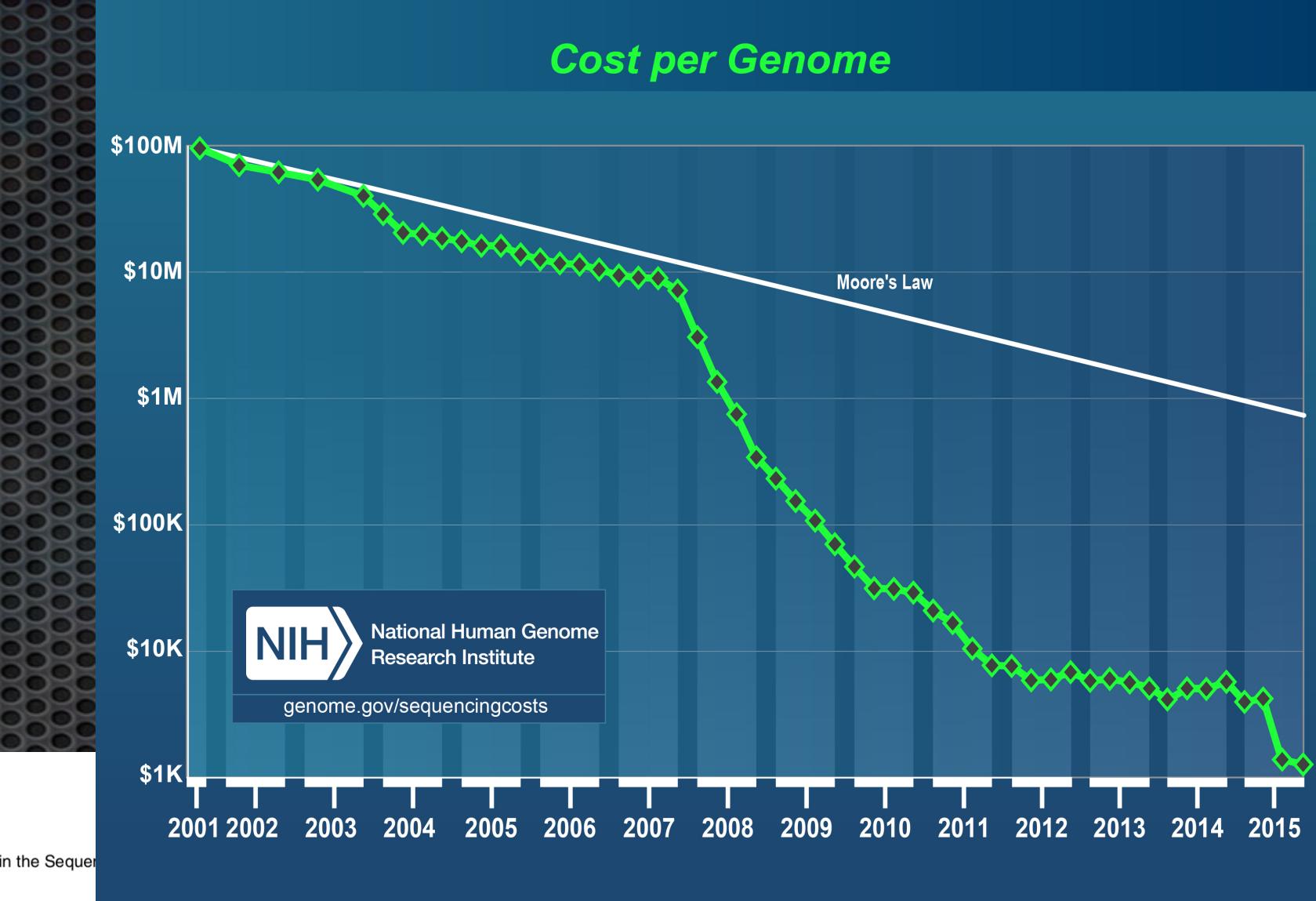
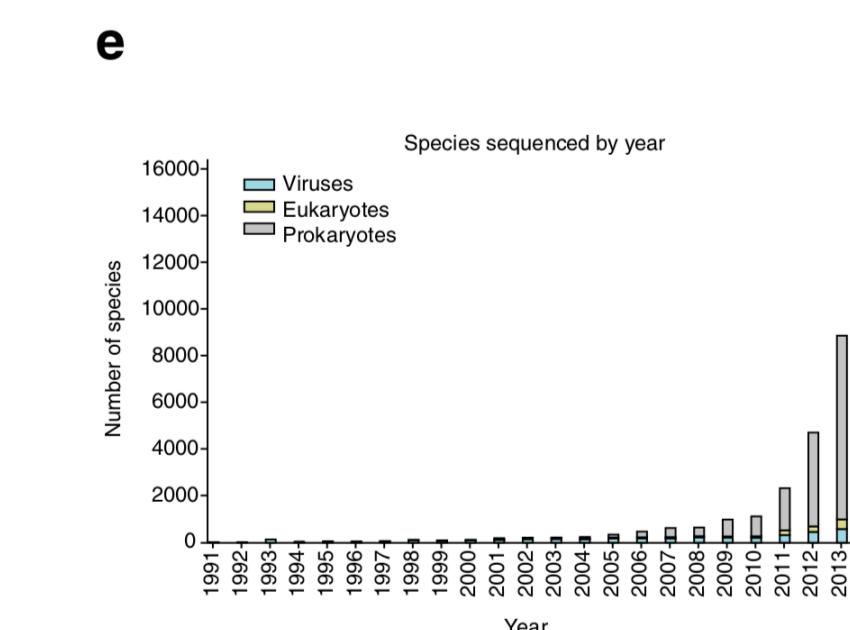
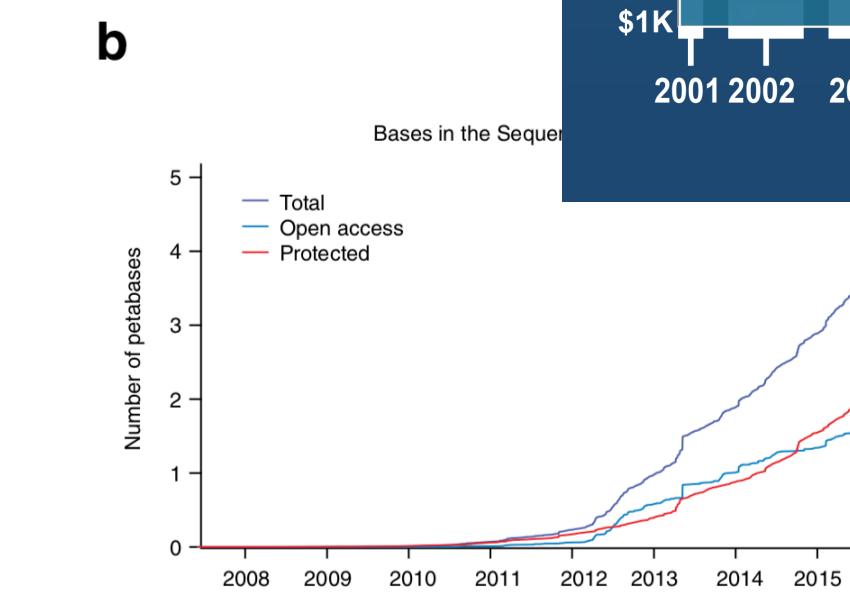
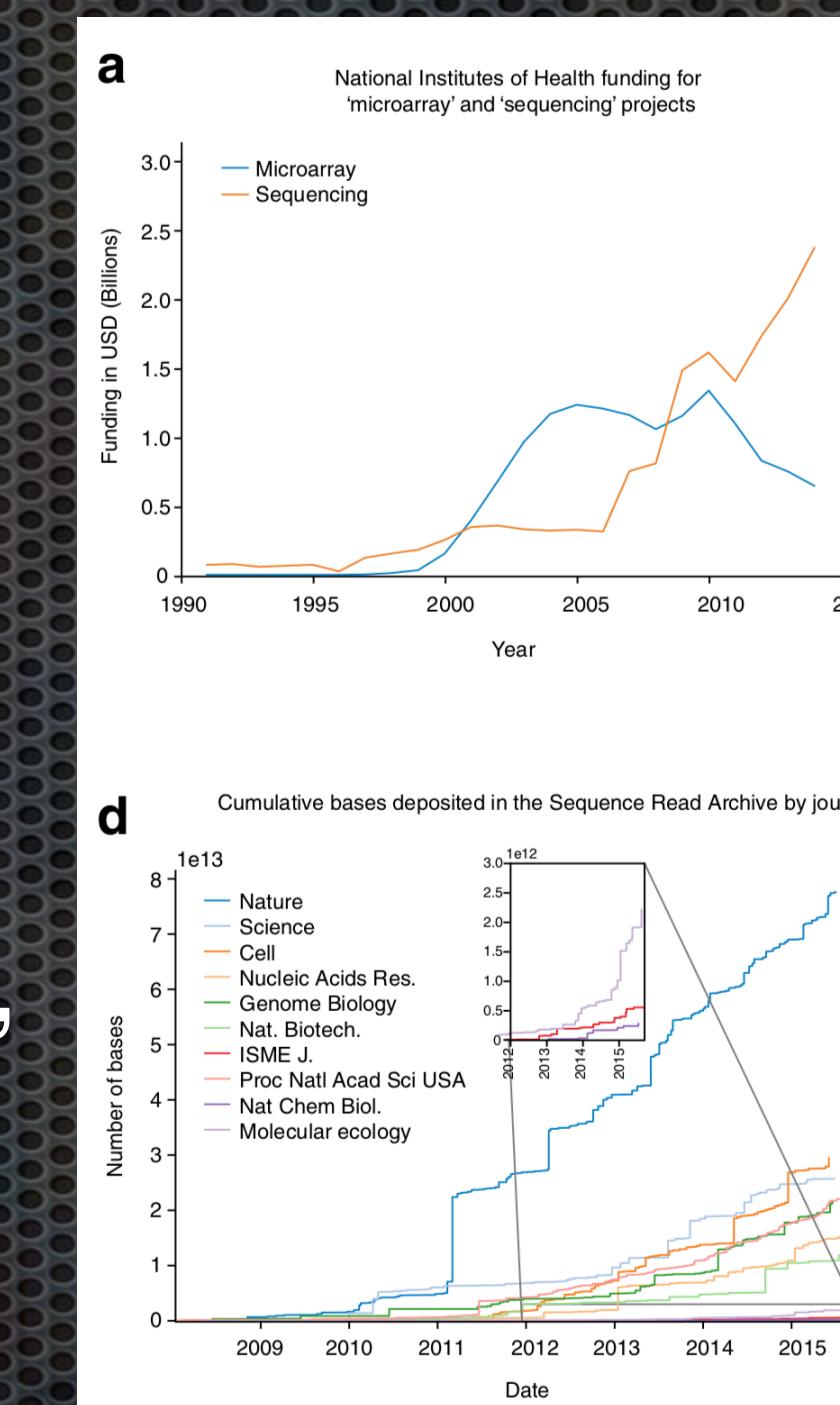
Lecture I: What is Bioinformatics? Examples from
Sequencing & Human Genome Variation

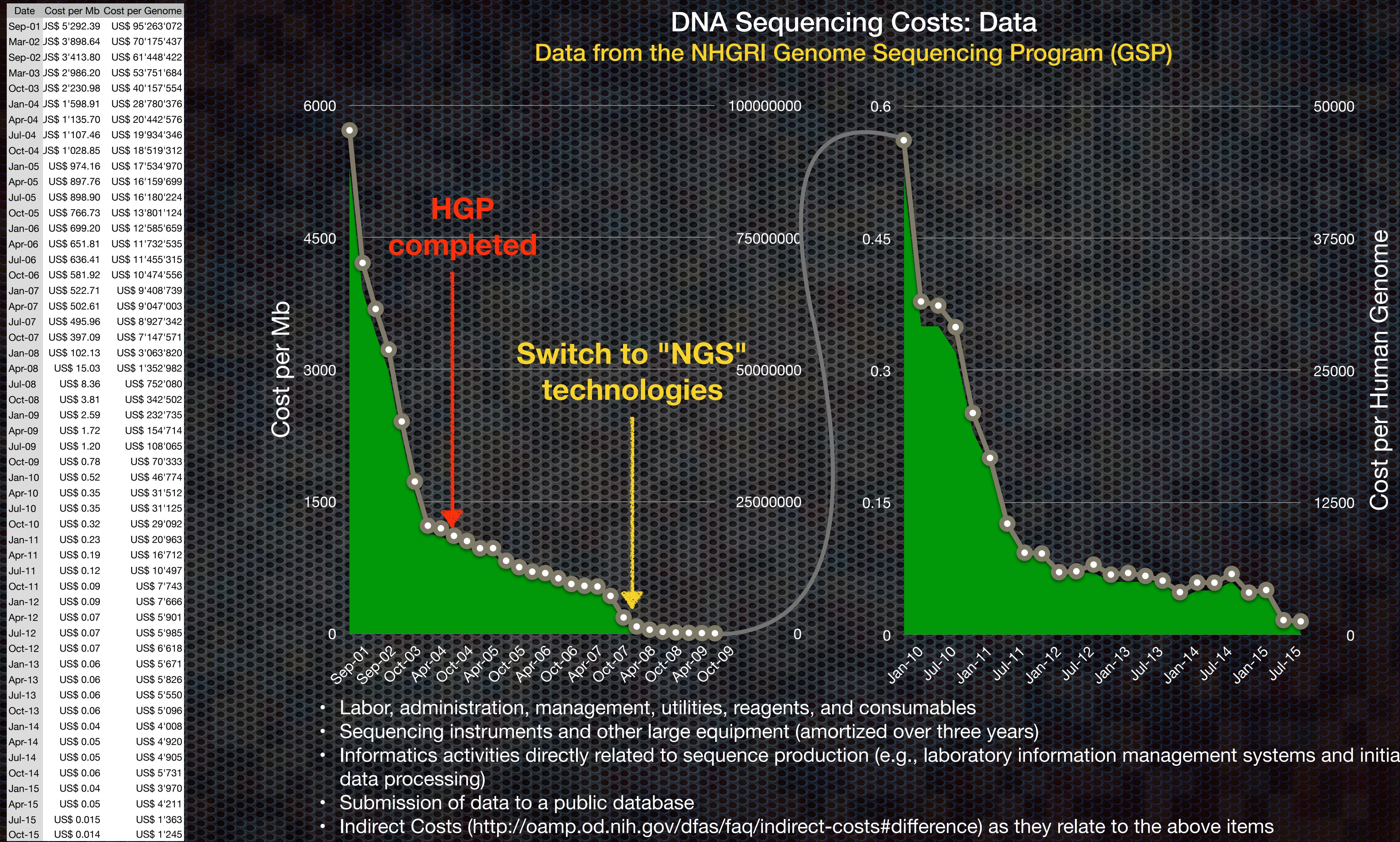


University of
Zurich^{UZH}

Sequencing everywhere, and everything

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

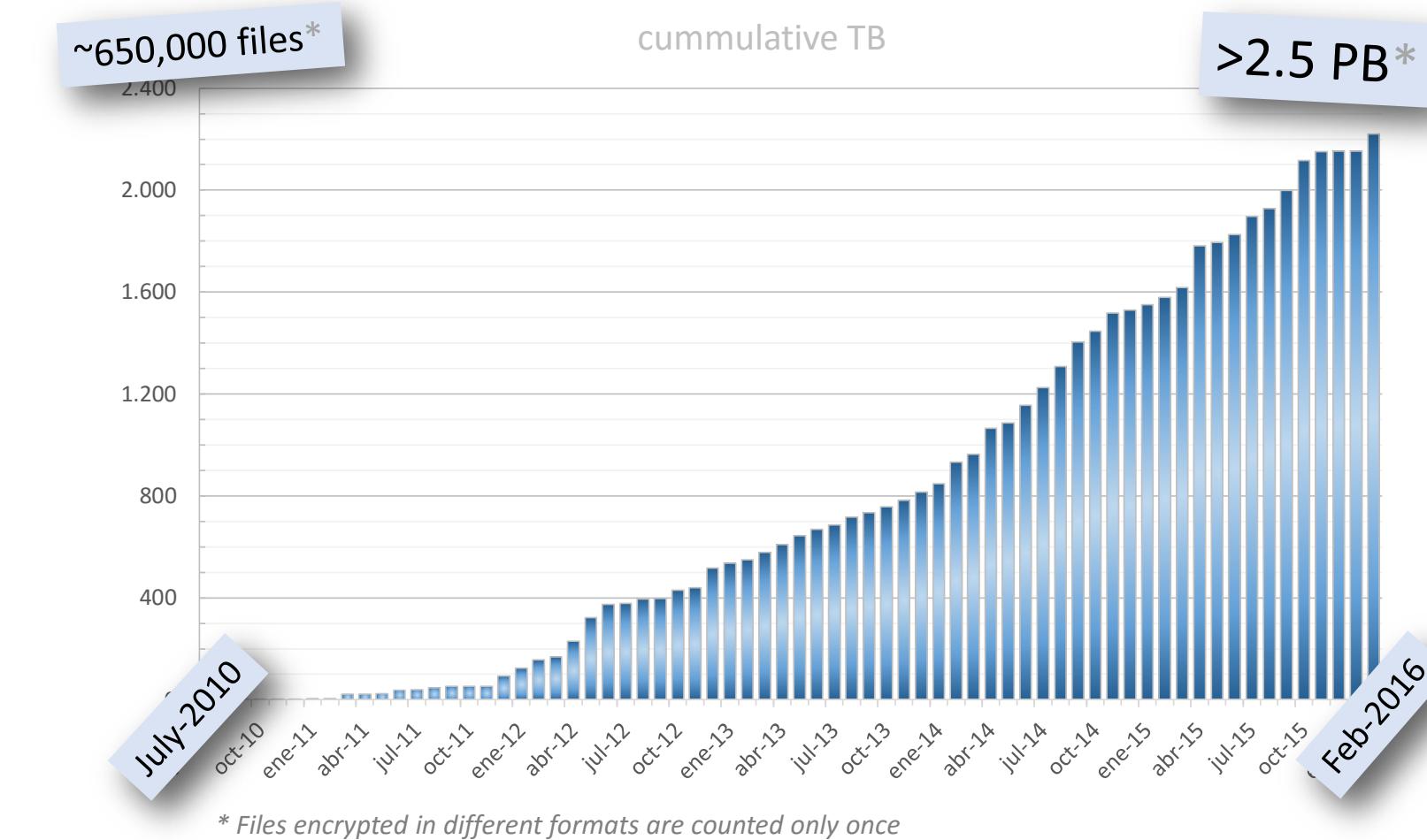




Bioinformatics - Databases & Data Driven Science

- "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)

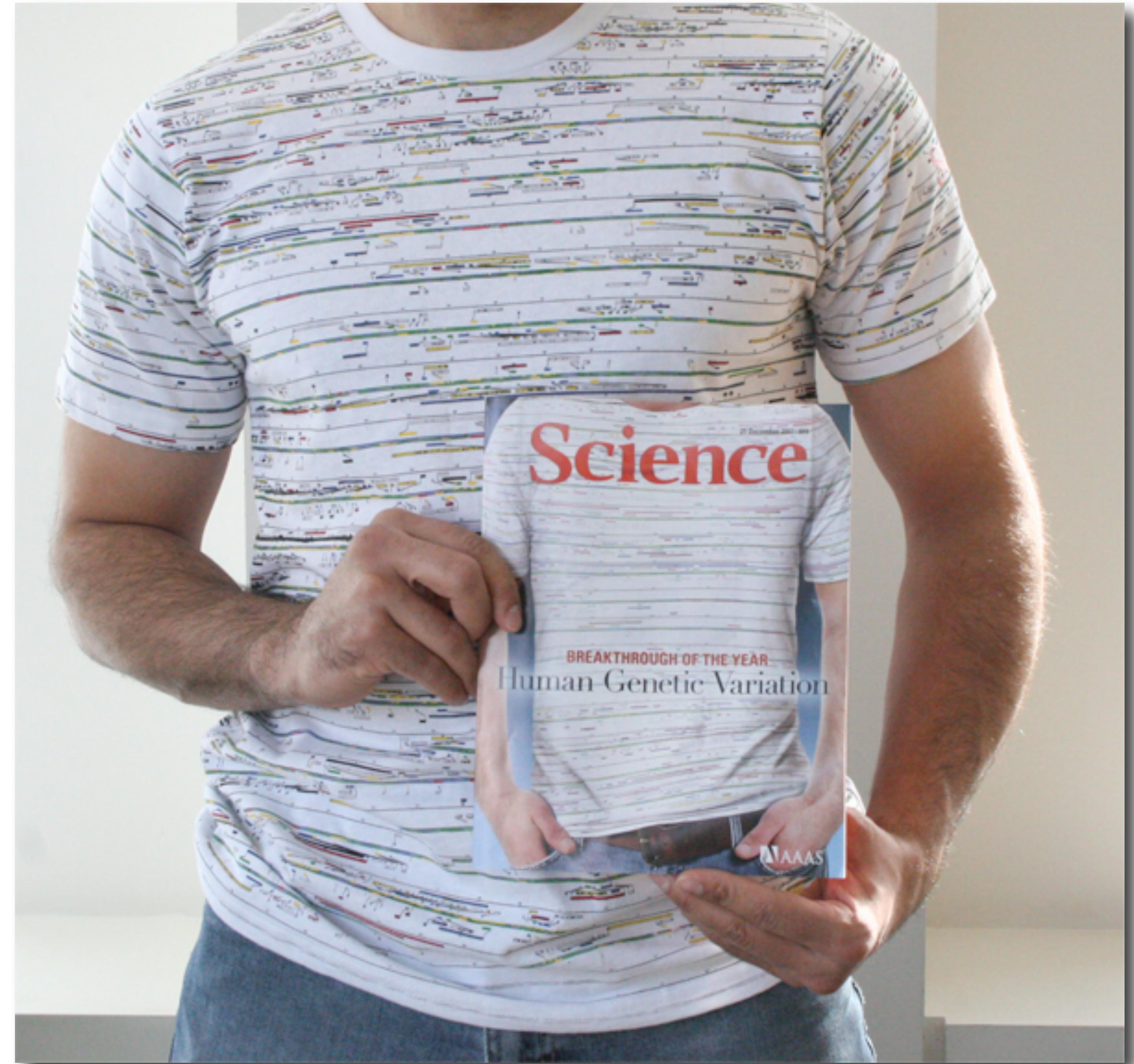
The EGA contains a growing amount of data



6

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

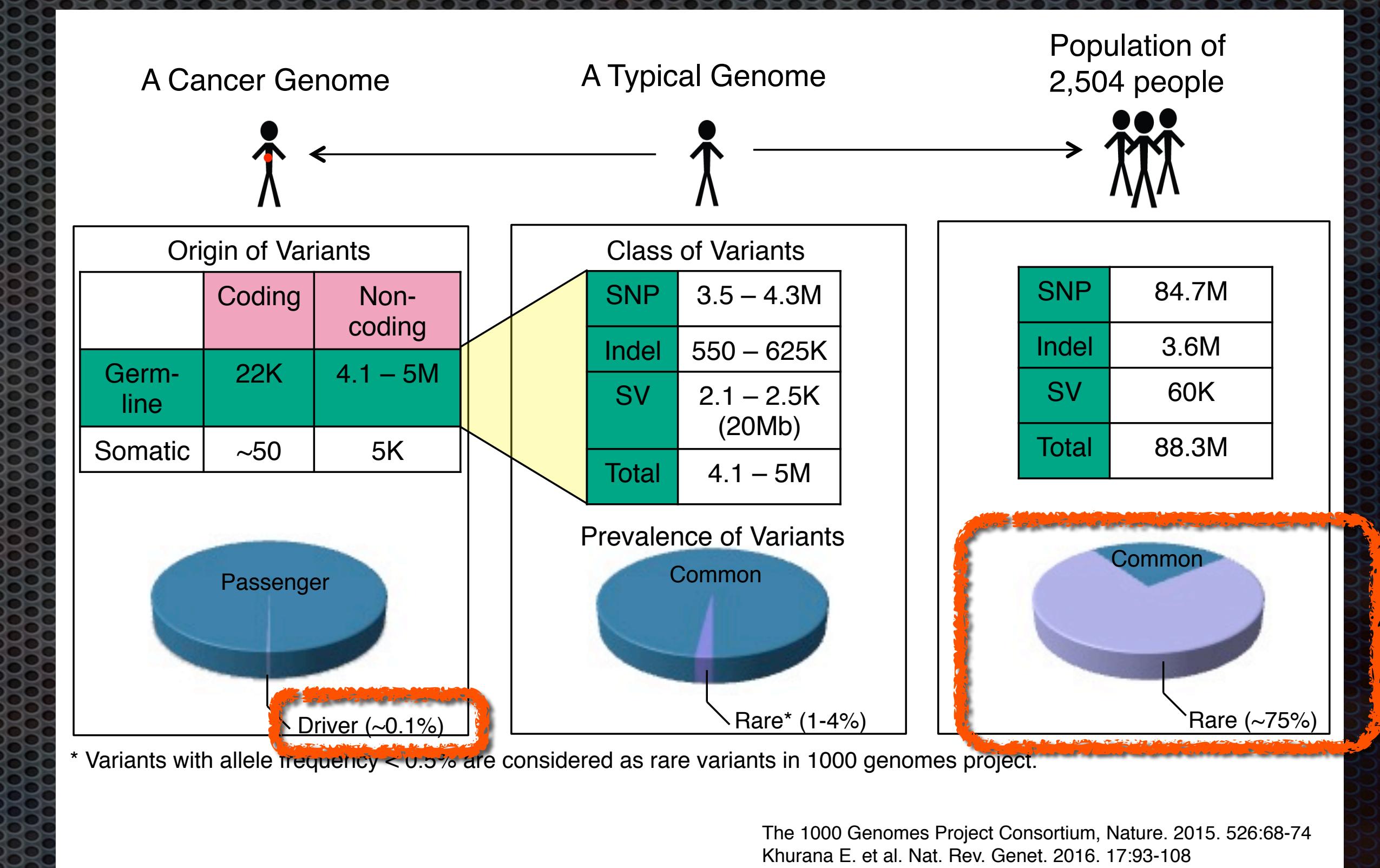


Conclusions from the analysis of variation in the human genome

- 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.

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

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



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}
 - $2 * 3 * 10^9 \text{ b} = 6,000,000,000 \text{ b}$
 - perfect genome (no overhead): ~715 MB
 - 1PB => ~1400000 genomes
- according to Swiss online store (Sep 2017) ~45'000CHF (100x10TB 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

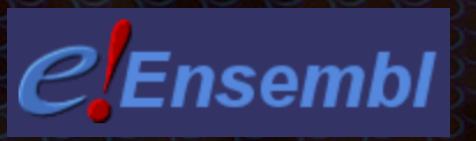
- aggregates information about genomic variation and its relationship to human health

- 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.



Beacon Network

Search Beacons

Search all beacons for allele

GRCh37 ▾ 10:118969015 C / CT Search

Response All None
 Found 16
 Not Found 27
 Not Applicable 22

Organization All None
 AMPLab, UC Berkeley
 BGI
 BioReference Laborato...
 Brazilian Initiative on ...
 BRCA Exchange
 Broad Institute
 Centre for Genomic R...
 Centro Nacional de A...
 Curoverse
 EMBL European Bio...
 Global Alliance for G...
 Google
 Institute for Systems ...
 Instituto Nacional de ...

Response	All	None
<input checked="" type="checkbox"/> Found	16	
<input type="checkbox"/> Not Found	27	
<input type="checkbox"/> Not Applicable	22	

BioReference BioReference Hosted by BioReference Laboratories Found

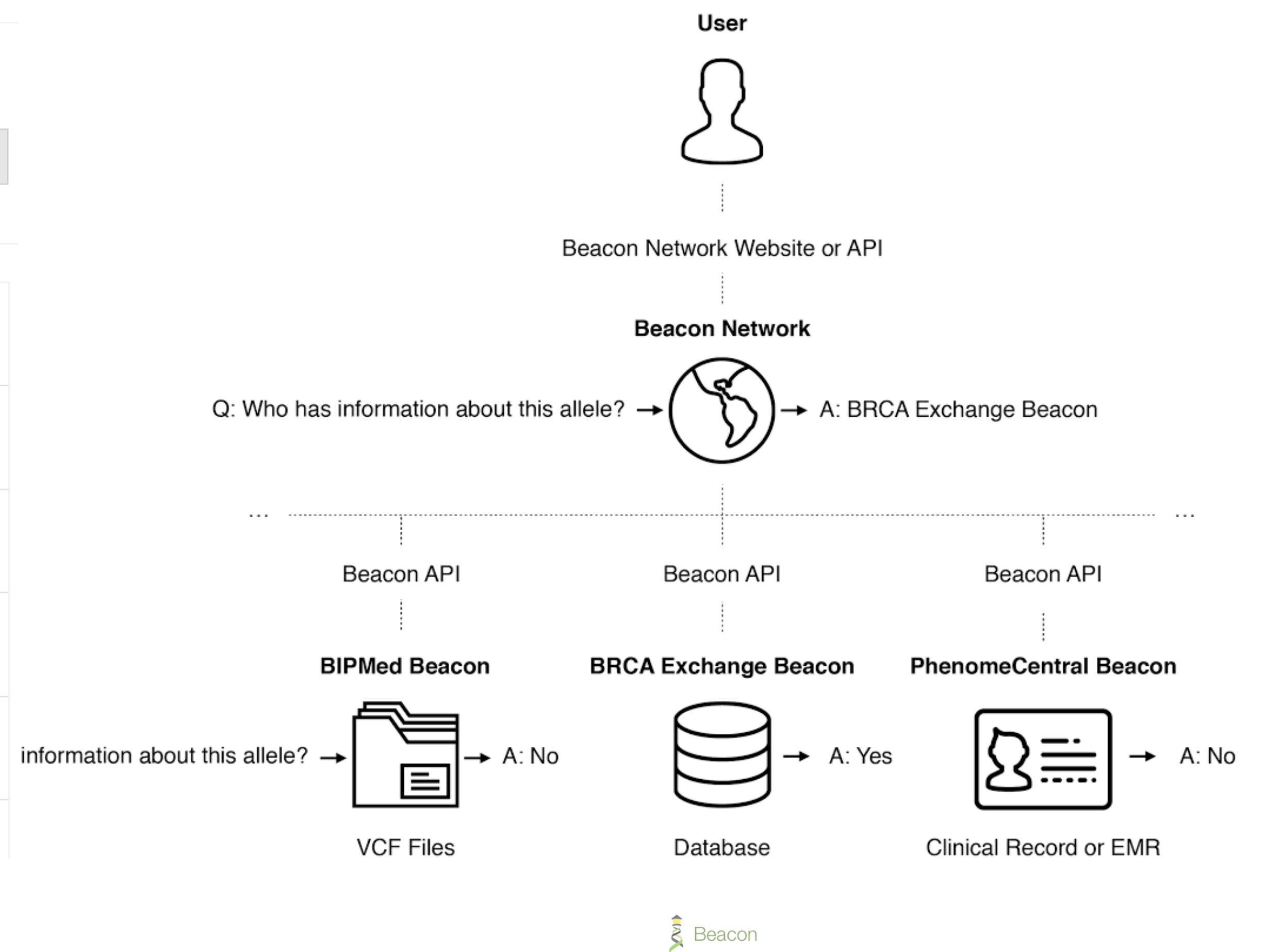
Catalogue of Somatic Mutations in Cancer Catalogue of Somatic Mutations in Cancer Hosted by Wellcome Trust Sanger Institute Found

Cell Lines Cell Lines Hosted by Wellcome Trust Sanger Institute Found

Conglomerate Conglomerate Hosted by Global Alliance for Genomics and Health Found

COSMIC COSMIC Hosted by Wellcome Trust Sanger Institute Found

dbGaP: Combined GRU Catalog and NHLBI Exome Seq... dbGaP: Combined GRU Catalog and NHLBI Exome Seq... Found



Date	Tag	Title
2018-01-24	v0.4.0	Beacon
2016-05-31	v0.3.0	Beacon

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

Beacon+

This forward looking Beacon interface implements additional, planned features beyond the current GA4GH specifications. [Info](#)

Query

Dataset: DIPG (CNV + selected SNV)
Reference name*: 17
Genome Assembly*: GRCh36 / hg18
Variant type*: SNV / indel
Position*: 7577121
Ref. Base(s)*: G
Alt. Base(s)*: A
Bio-ontology: pgx:icdom:9380_3

Beacon Query

Response

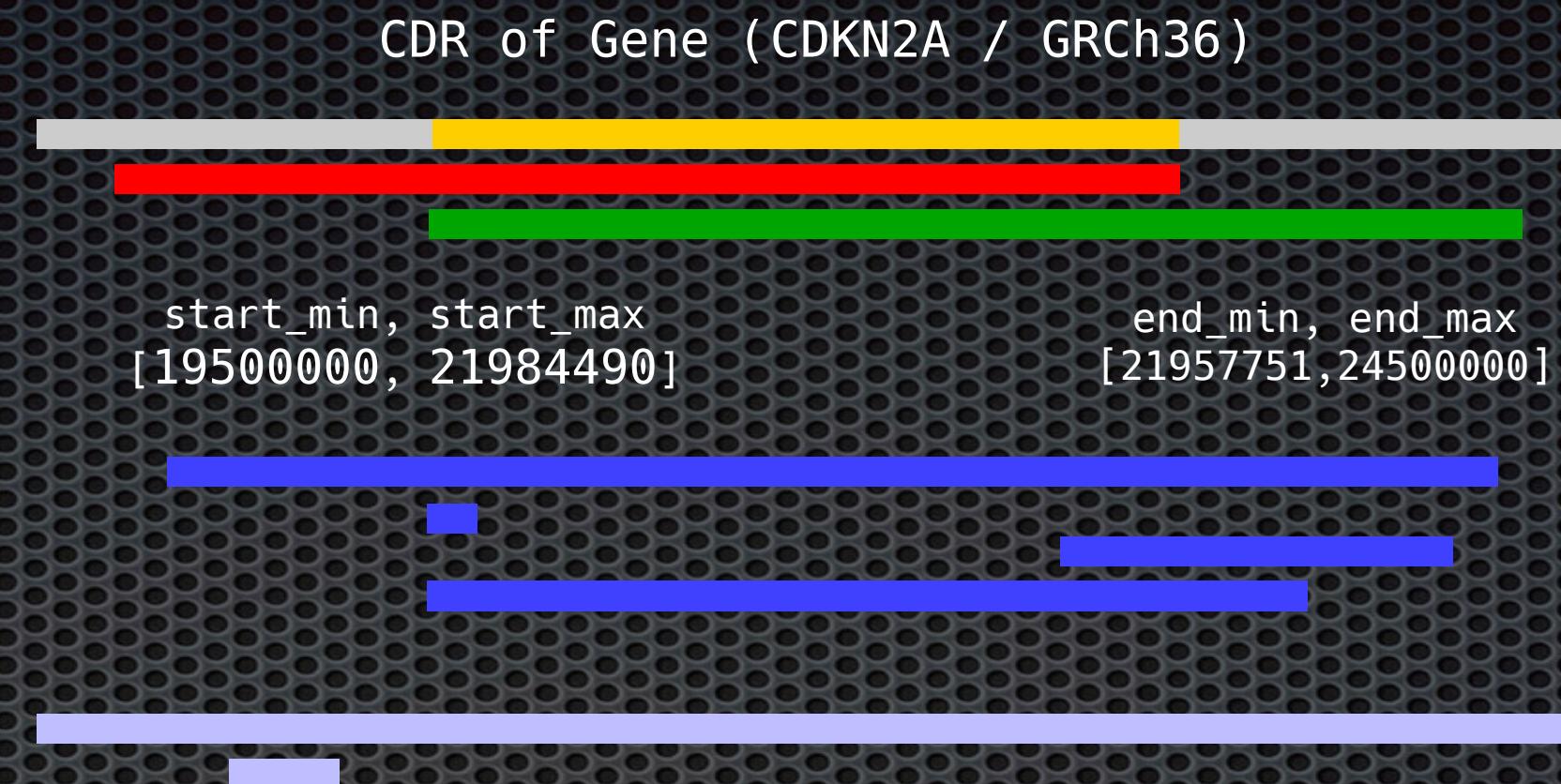
Dataset	Chro.	Assembly	Var. Type	Start Min	Start Max	End Min	End Max	Pos.	Ref.	Alt.	Bio Query	Call Count	Samples	f	Query
arraymap	9	GRCh36	DEL	19000000	21984490	21900000	25000000				pgx:icdom:8140_3	3781	403	0.0065	show JSON
dipg	17	GRCh36	SNV			7577121		G	A	pgx:icdom:9380_3	21	20	0.0187	show JSON	

arrayMap  University of Zurich UZH  This Beacon implementation is developed by the Computational Oncogenomics Group at the University of Zurich, with support from the SIB Technology group and ELIXIR.   

```

"allele_request" : {
  "$and" : [
    { "reference_name" : "9" },
    { "variant_type" : "DEL" },
    { "start" : { "$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&variants.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"
      ]
    }
  ]
}

```



Match using query ranges “at least one base in interval affected”

Example “focal” matches

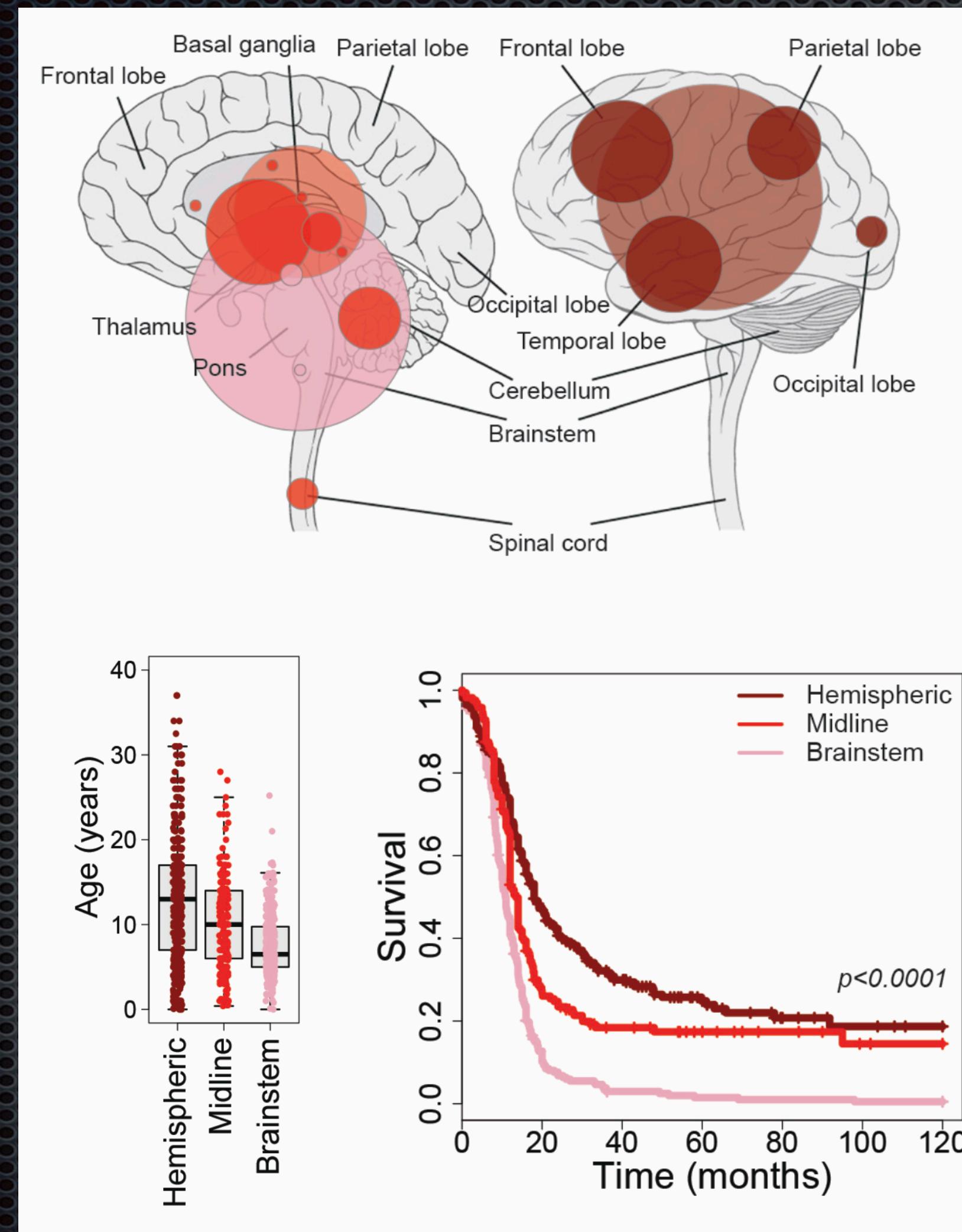
Mismatches

- 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



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)

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



```
{  
  "_id" : ObjectId("591eb73709d374442fcc6b2a"),  
  "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" } ] },  
    }  
  }
```

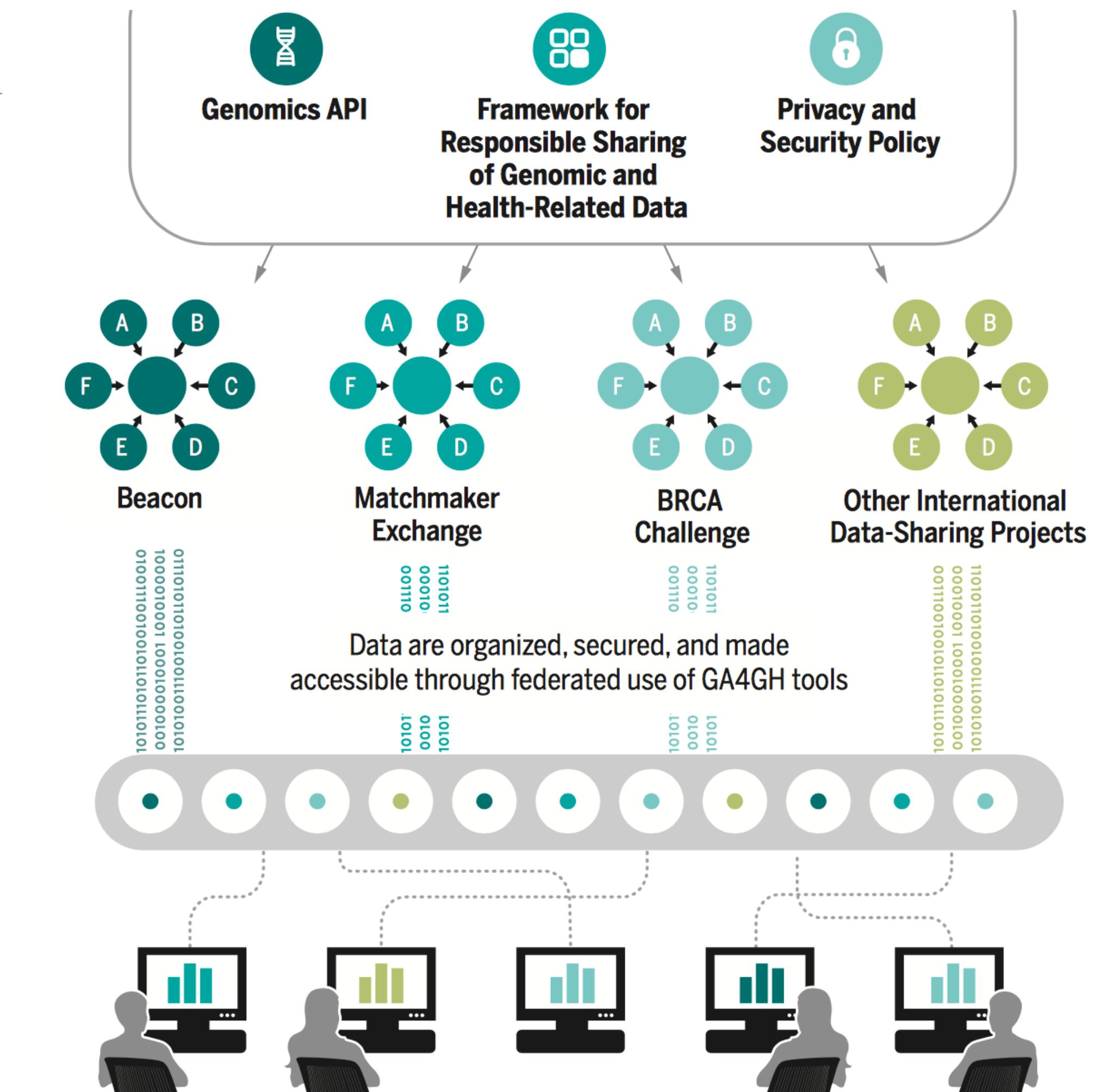
GA4GH API promotes sharing



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.



BIO390: Course Schedule

- 2018-09-18: Michael Baudis - **What is Bioinformatics? Introduction and Resources**
- 2018-09-25: Christian von Mering - **Sequence Informatics**
- 2018-10-02: Mark Robinson - **Statistical Bioinformatics**
- 2018-10-09: Katja Bärenfaller - **Proteomics**
- 2018-10-16: Izaskun Mallona - **Regulatory Genomics and Epigenomics**
- 2018-10-23: Daniel Stekhoven - **Clinical Bioinformatics**
- 2018-10-30: Shinichi Sunagawa - **Metagenomics**
- 2018-11-06: Amedeo Caflisch - **Molecular Dynamics**
- 2018-11-13: Elif Ozkirimli Olmez - **Protein Interactions**
- 2018-11-20: Fabio Rinaldi - **Text Mining I**
- 2018-11-27: Fabio Rinaldi - **Text Mining II**
- 2018-12-04: Andreas Wagner - **Biological Networks**
- 2018-12-11: Michael Baudis - **Human Genome Data - Risks & Opportunities || Feedback**
- 2018-12-18: Michael Baudis - **Exam (multiple choice)**



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



Global Alliance
for Genomics & Health



arraymap.org

progenetix.org

info.baudisgroup.org

sib.swiss/baudis-michael

imls.uzh.ch/en/research/baudis