BIO390: Introduction to Bioinformatics

Lecture I: What is Bioinformatics?

Michael Baudis - 2017-09-19

University of Zurich[UZH]

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

BIO390: Course Schedule

- 2017-09-19: Michael Baudis - What is Bioinformatics? Introduction Resources, Examples
- 2017-09-26: Christian von Mering - Sequence Informatics
- 2017-10-03: Mark Robinson - Statistical Bioinformatics
- 2017-10-10: Helen Lindsay - Bioinformatics tools: Statistics & Graphics in R & BioConductor
- 2017-10-17: Vinko Tosevski: Bioinformatics in flow- and mass cytometry
- 2017-10-24: Shinichi Sunagawa - Metagenomics
- 2017-10-31: Amedeo Caflisch - Molecular Dynamics
- 2017-11-07: Matthias Gstaiger - Proteomics
- 2017-11-14: Katja Bärenfaller - RNA structure prediction
- 2017-11-21: Daniel Stekhoven - Clinical Bioinformatics & Personalized Medicine
- 2017-11-28: Fabio Rinaldi - Text Mining
- 2017-12-05: Andreas Wagner - Biological Networks
- 2017-12-12: Tuncay Baubec - Bioinformatic approaches in gene regulation and epigenetics
- 2017-12-19: Exam

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

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?

analyze

verb | an·a·lyze | \ˈa-nə-ˌlīz\

Popularity: Top 1% of lookups

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?

hypothesis

noun | hy·poth·e·sis | \hī-ˈpä-thə-səs\

Popularity: Top 1% of lookups

Merriam-Webster

- 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

hypothesis

noun | hy·poth·e·sis | \hī-ˈpä-thə-səs\

Popularity: Top 1% of lookups

Merriam-Webster

- 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	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

flux

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

nextprot

STRING

UniProt

arrayMap

KEGG

ARRAYMAP

DATA PIPELINE

BIOINFORMATICS & CURATION
ARRAYMAP DATA
“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)

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

BED file example

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":{"$oid":"558e5bfcad9a82d95838b62a"},
"CLINICALGROUP":"Carcinomas: hepatic ca."}, {"FOLLOWUP":"", "PMID":"11485905",
"BIOSAMPLEID":"AM_BS__HCChypo-won-H18", "ICDMORPHOLOGYCODE":"8170/3",
"ICDMORPHOLOGY":"Hepatocellular carcinoma, NOS", "CLINICALGROUP":"Carcinomas: hepatic
ca.”, "_id":{"$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)

```
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 'histon
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

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: Course Schedule

- 2017-09-19: Michael Baudis - What is Bioinformatics? Introduction Resources, Examples
- 2017-09-26: Christian von Mering - Sequence Informatics
- 2017-10-03: Mark Robinson - Statistical Bioinformatics
- 2017-10-10: Helen Lindsay - Bioinformatics tools: Statistics & Graphics in R & BioConductor
- 2017-10-17: Vinko Tosevski: Bioinformatics in flow- and mass cytometry
- 2017-10-24: Shinichi Sunagawa - Metagenomics
- 2017-10-31: Amedeo Caflisch - Molecular Dynamics
- 2017-11-07: Matthias Gstaiger - Proteomics
- 2017-11-14: Katja Bärenfaller - RNA structure prediction
- 2017-11-21: Daniel Stekhoven - Clinical Bioinformatics & Personalized Medicine
- 2017-11-28: Fabio Rinaldi - Text Mining
- 2017-12-05: Andreas Wagner - Biological Networks
- 2017-12-12: Tuncay Baubec - Bioinformatic approaches in gene regulation and epigenetics
- 2017-12-19: Exam

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 ...)

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

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

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 => ~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)

The Variant Call Format (VCF)

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

- 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

Global Alliance for Genomics & Health

The VCF file format: Standard for variant calling

Example

VCF header

```
##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">
```

Body

```
#CHROM POS ID     REF  ALT    QUAL FILTER  INFO                 FORMAT     SAMPLE1   SAMPLE2
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    <DEL>   .   PASS    SVTYPE=DEL;END=300   GT:GQ:DP   1/1:12:3  0/0:20
```

Mandatory header lines

Optional header lines (meta-data about the annotations in the VCF body)

Reference alleles (GT=0)

Alternate alleles (GT>0 is an index to the ALT column)

Phased data (G and C above are on the same chromosome)

Deletion

SNP

Large SV

Insertion

Other event

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.

> 50 Beacons throughout the world

Global Alliance for Genomics & Health

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

SNV Example | CNV Example

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.

elixir | SIB | progenetix

```
  "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&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("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" } ] },
   }
}
```

BIO390: Course Schedule

- 2017-09-19: Michael Baudis - What is Bioinformatics? Introduction Resources, Examples
- **2017-09-26: Christian von Mering - Sequence Informatics**
- 2017-10-03: Mark Robinson - Statistical Bioinformatics
- 2017-10-10: Helen Lindsay - Bioinformatics tools: Statistics & Graphics in R & BioConductor
- 2017-10-17: Vinko Tosevski: Bioinformatics in flow- and mass cytometry
- 2017-10-24: Shinichi Sunagawa - Metagenomics
- 2017-10-31: Amedeo Caflisch - Molecular Dynamics
- 2017-11-07: Matthias Gstaiger - Proteomics
- 2017-11-14: Katja Bärenfaller - RNA structure prediction
- 2017-11-21: Daniel Stekhoven - Clinical Bioinformatics & Personalized Medicine
- 2017-11-28: Fabio Rinaldi - Text Mining
- 2017-12-05: Andreas Wagner - Biological Networks
- 2017-12-12: Tuncay Baubec - Bioinformatic approaches in gene regulation and epigenetics
- **2017-12-19: Exam**