UNIVERSITÄT BERN

De novo Genome & Transcriptome Assembly Course

Lecture 1 – Genome Properties and Fields of Application

Rémy Bruggmann – Interfaculty Bioinformatics Unit (IBU)

22.09.21

SIB Swiss Institute of Bioinformatics

Program of the Course

- 9.15 – 10.00 Lecture 1 – Introduction
- coffe break
- 11.30 – 12.30 Access reads and perform basic QC
- Lunch break
- 1.30 – 2.30 Lecture 2 – Data Generation (NGS, 3GS)

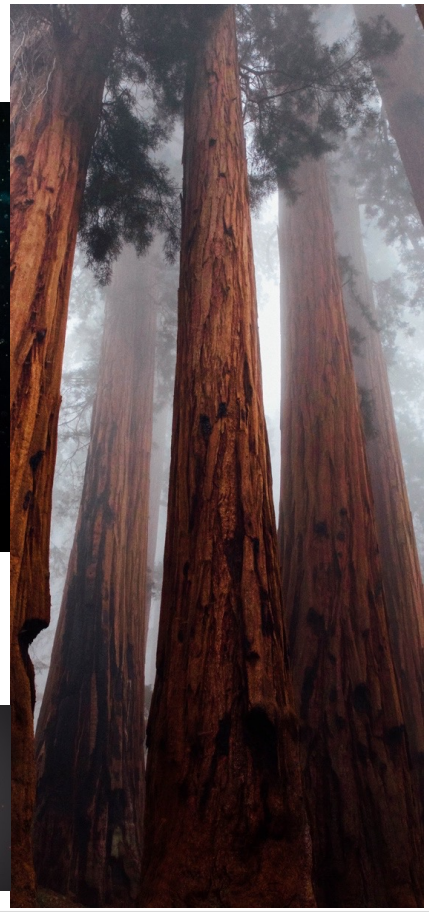
Genome Properties

- Which genomes to sequence?
- Why sequence and assemble genomes at all?

Different | Organisms – Genomes – Assemblies







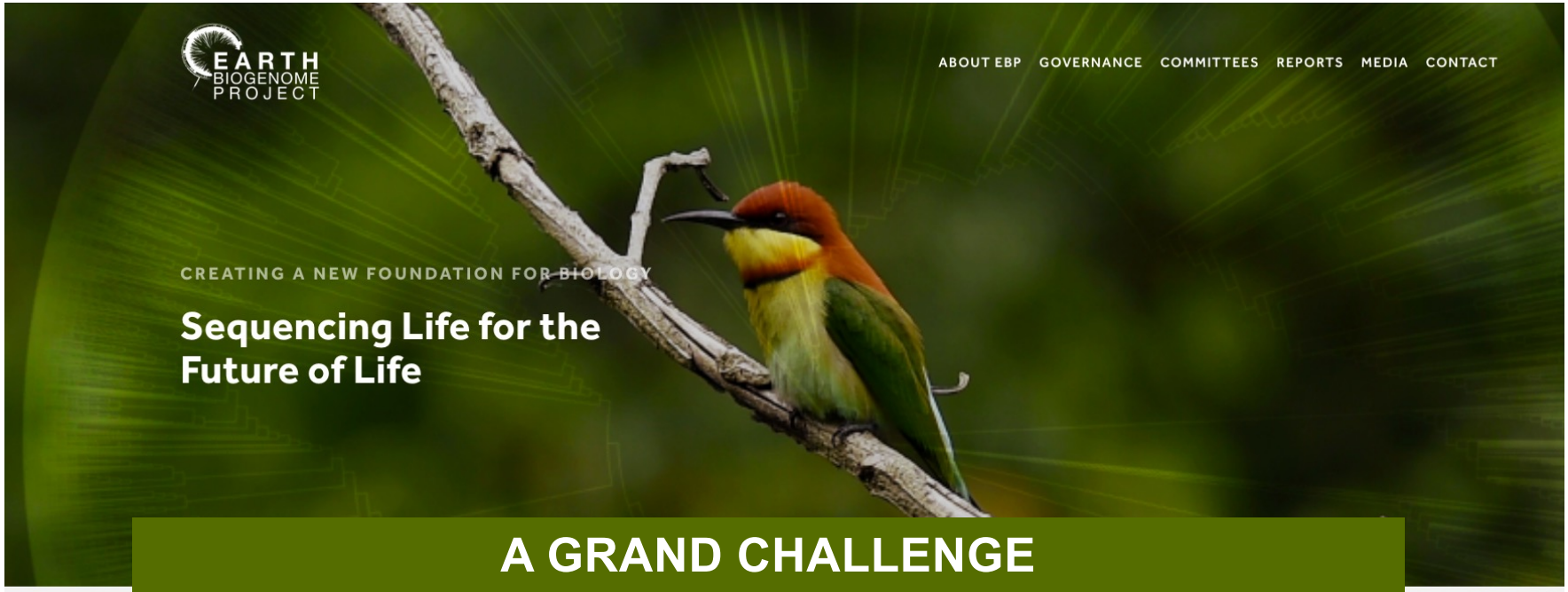


video from Claudia Eyer

fotos from unsplash

Initiatives

- Earth Biogenome Project (EBP)
- Darwin Tree of Life
- Vertebrate Genome Project (VGP)
- European Reference Genome Atlas (ERGA)



A GRAND CHALLENGE

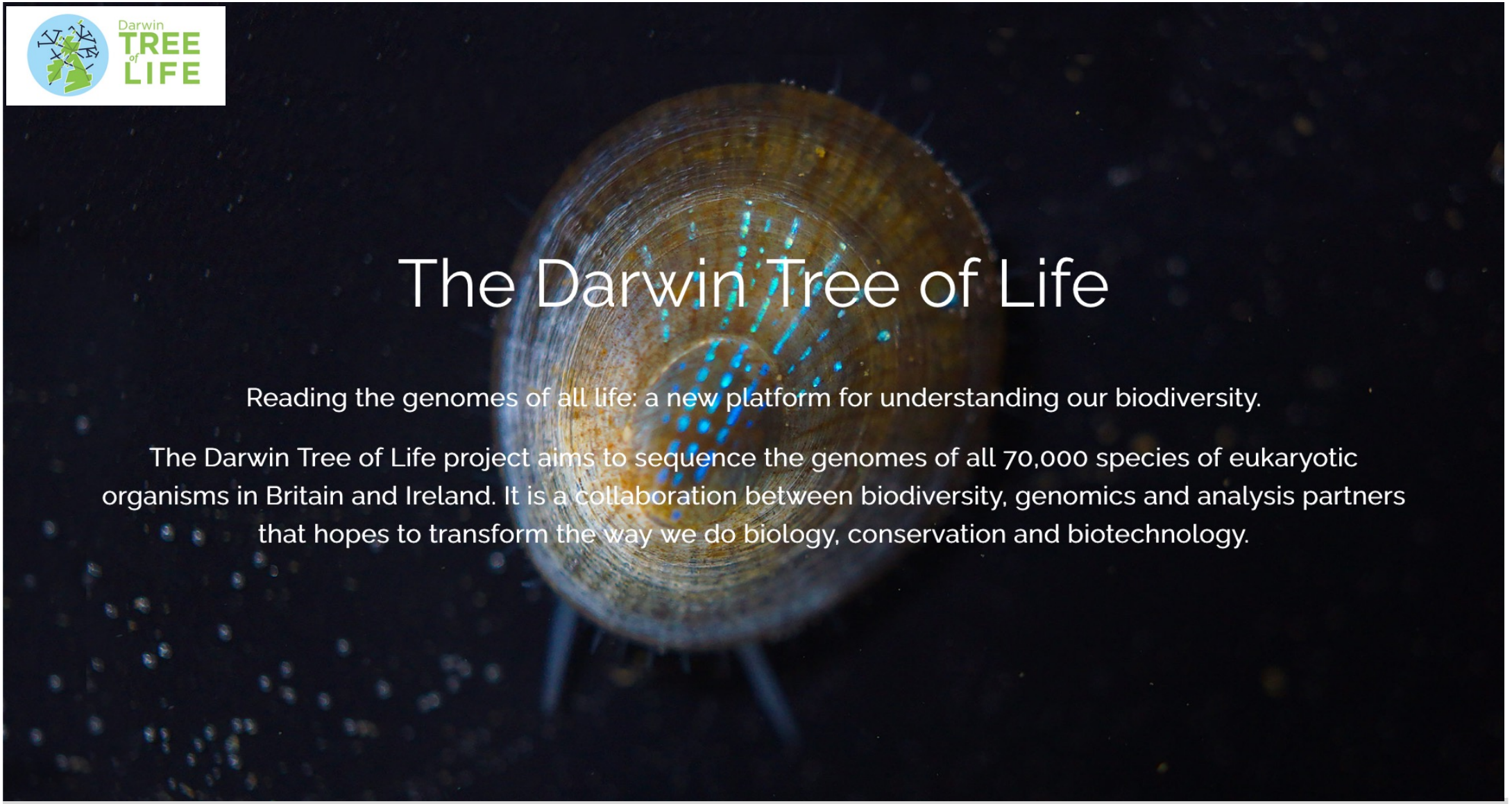
The Earth BioGenome Project (EBP), a moonshot for biology, aims to sequence, catalog and characterize the genomes of all of Earth's eukaryotic biodiversity over a period of ten years.

A GRAND CHALLENGE

The Earth BioGenome Project (EBP), a *moonshot* for biology, aims to sequence, catalog and characterize the genomes of all of Earth's eukaryotic biodiversity over a period of ten years.

A GRAND VISION

Create a new foundation for biology to drive solutions for preserving biodiversity and sustaining human societies.


Darwin
TREE of LIFE
The Darwin Tree of Life
Reading the genomes of all life: a new platform for understanding our biodiversity.
The Darwin Tree of Life project aims to sequence the genomes of all 70,000 species of eukaryotic organisms in Britain and Ireland. It is a collaboration between biodiversity, genomics and analysis partners that hopes to transform the way we do biology, conservation and biotechnology.

EUROPEAN REFERENCE GENOME ATLAS

UNIVERSITÄT BERN

OUR MISSION

European Reference Genome Atlas (ERGA)

UNIVERSITÄT BERN

- Aim of ERGA is to generate *de novo* genome assemblies that are
  - reference-quality
  - complete
  - error-free
- for species across the whole of European biodiversity
  - includes at least 200,000 species of higher organisms

Richard Durbin

*"We are **entering a new era of genomics**, in which reference sequences for all species will change the way we interact with the whole living world. The ERGA will coordinate European activities in this area and ensure that Europe plays a leading role in this fundamental initiative for 21st century science."*

Gene W. Myers

"An European-wide initiative like the ERGA is absolutely necessary given the scale of the proposed atlas. I belief this atlas **will revolutionize of conservation science, agricultural and animal science, and even contribute significantly to medicine**."

Genome Sequence – a Prerequisite for…

UNIVERSITÄT BERN

- precise genome editing
- conservation biology
- biodiversity
- evolutionary
- cancer genomics
- non-model
- population genetics
- personalized health
- basic research
- comparative genomics
- breeding (agriculture)
- …

How can we get to high quality genomes

The speed of the new DNA sequencing techniques has created a need for computer programs to handle the data produced. This paper describes simple programs designed specifically for use by people with little or no computer experience. The programs are for use on small computers and provide facilities for storage, editing and analysis of both DNA and amino acid sequences...

Beginning of Sequencing and Bioinformatics

Volume 4 Number 11 November 1977 **Nucleic Acids Research**

Sequence data handling by computer

R.Staden

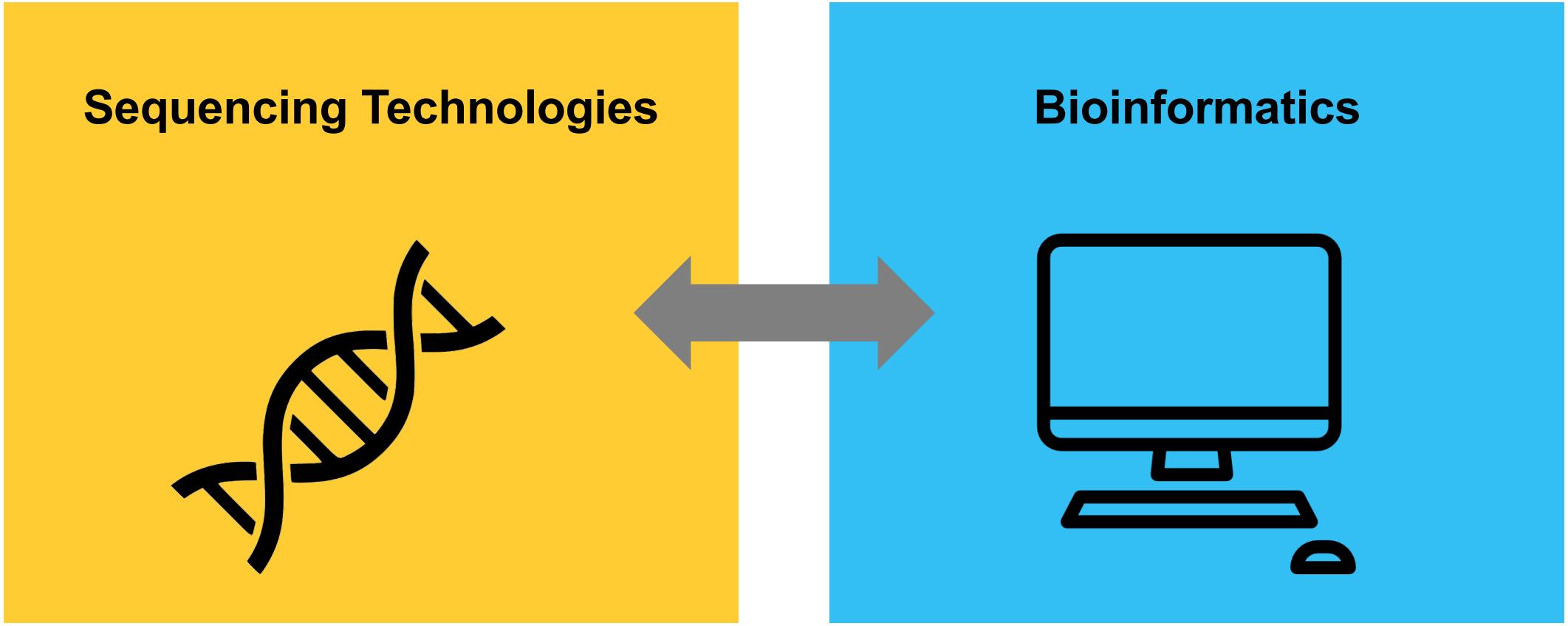
MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

Received 10 October 1977

ABSTRACT

The speed of the new DNA sequencing techniques has created a need for computer programs to handle the data produced. This paper describes simple programs designed specifically for use by people with little or no computer experience. The programs are for use on small computers and provide facilities for storage, editing and analysis of both DNA and amino acid sequences. A magnetic tape containing these programs is available on request.

Genome Assembly – a Combination of Sequencing and Bioinformatics

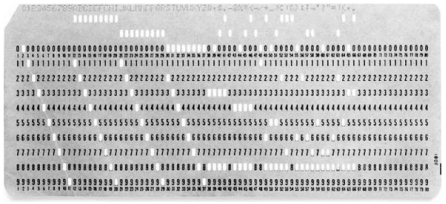


Sequencing | Bioinformatics | IT

Sequencing | Bioinformatics | IT

UNIVERSITÄT BERN

The Beginning of Bioinformatics



IBM 711 Punch card reader (1952)

- Punch cards in storage at a U.S. Federal records center in 1959
- All the data visible here could fit on a **4 GB** USB stick (one DVD)

First Bioinformatician

UNIVERSITÄT BERN

- Founder of bioinformatics and expert in programming Punchcards

Margaret Dayhoff (1925-83)

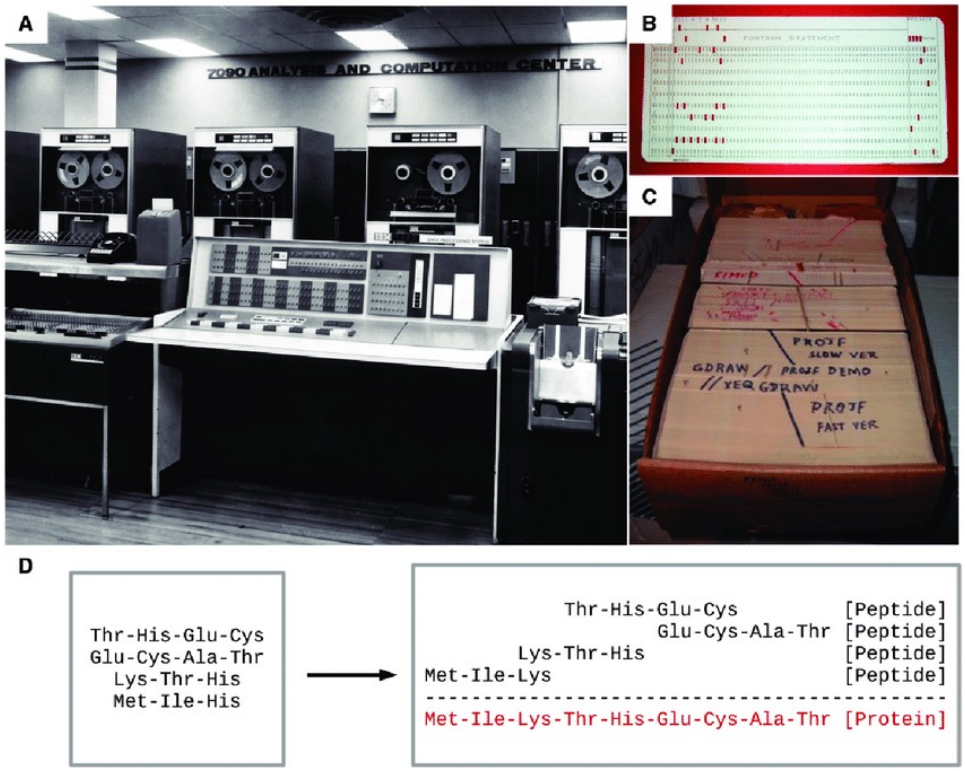
~ 1980

Foundation of Bioinformatics: Margaret Dayhoff

UNIVERSITÄT BERN

- Founder of bioinformatics and expert in programming Punchcards
- 1961 she developed computational tools (COMPROTEIN) to aid primary structure determination

  → **overlaps of amino/peptide acid sequences**
  → **first sequence alignment/assembly algorithm**



Foundation of Bioinformatics: Margaret Dayhoff

UNIVERSITÄT BERN

- Founder of bioinformatics and expert in programming Punchcards
- 1961 she developed computational tools (COMPROTEIN) to aid primary structure determination

  → **overlaps of amino/peptide acid sequences**
  → **first sequence alignment/assembly algorithm**

- introduced single-letter code for proteins
  → **data compression (punch cards)**

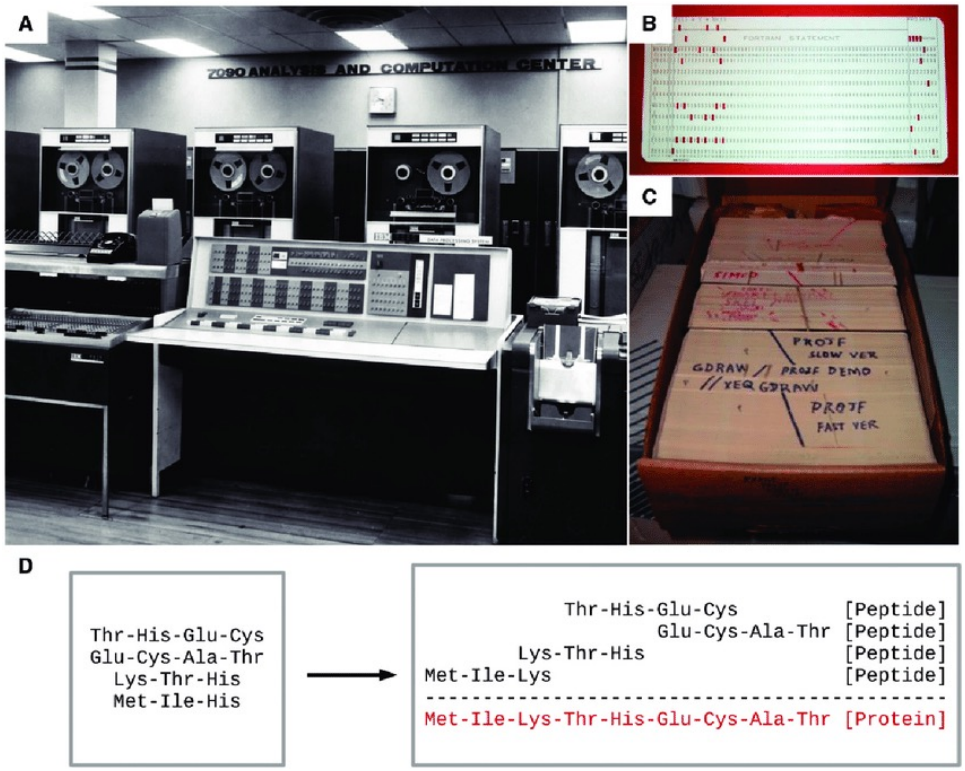
1-letter code	3-letter code	Amino acid	Possible codons
A	Ala	Alanine	GCA, GCC, GCG, GCT
B	Asx	Asparagine or Aspartic acid	AAC, AAT, GAC, GAT
C	Cys	Cysteine	TGC, TGT
D	Asp	Aspartic acid	GAC, GAT
E	Glu	Glutamic acid	GAA, GAG
F	Phe	Phenylalanine	TTC, TTT
G	Gly	Glycine	GGA, GGC, GGG, GGT
H	His	Histidine	CAC, CAT
I	Ile	Isoleucine	ATA, ATC, ATT
K	Lys	Lysine	AAA, AAG
L	Leu	Leucine	CTA, CTC, CTG, CTT, TTA, TTG
M	Met	Methionine	ATG
N	Asn	Asparagine	AAC, AAT
P	Pro	Proline	CCA, CCC, CCG, CCT
Q	Gln	Glutamine	CAA, CAG
R	Arg	Arginine	AGA, AGG, CGA, CGC, CGG, CGT
S	Ser	Serine	AGC, AGT, TCA, TCC, TCG, TCT
T	Thr	Threonine	ACA, ACC, ACG, ACT
V	Val	Valine	GTA, GTC, GTG, GTT
W	Trp	Tryptophan	TGG
X	X	Stop codon	TAA, TAG, TGA
Y	Tyr	Tyrosine	TAC, TAT
Z	Glx	Glutamine or Glutamic acid	CAA, CAG, GAA, GAG

Foundation of Bioinformatics: Margaret Dayhoff

UNIVERSITÄT BERN

- Founder of bioinformatics and expert in programming Punchcards
- 1961 she developed computational tools (COMPROTEIN) to aid primary structure determination

  → **overlaps of amino/peptide acid sequences**
  → **first sequence alignment/assembly algorithm**
- introduced single-letter code for proteins
  → **data compression (punch cards)**
- Fortran code was run on an IBM 7090 (70 Kflop/s); *iPhone XS: ~5 Teraflop/s → 71 million times faster!*



Foundation of Bioinformatics: Margaret Dayhoff

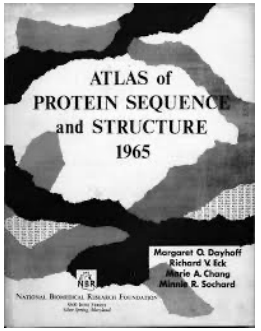
UNIVERSITÄT BERN

- 1965 publication of the
  "Atlas of Protein Sequence and Structure"

  → 1984 release of Protein Information Resource "PIR";

  → 2002, PIR, SwissProt and trEMBL
  merged into a single database UniProt
  (UNIversal PROTein resource).

- 1966: started to develop the PAM-Model (Point/Percent Accepted Mutation Matrix); substitution matrix



UniProt

EMBL-EBI PIR SIB

Foundation of Bioinformatics: Margaret Dayhoff

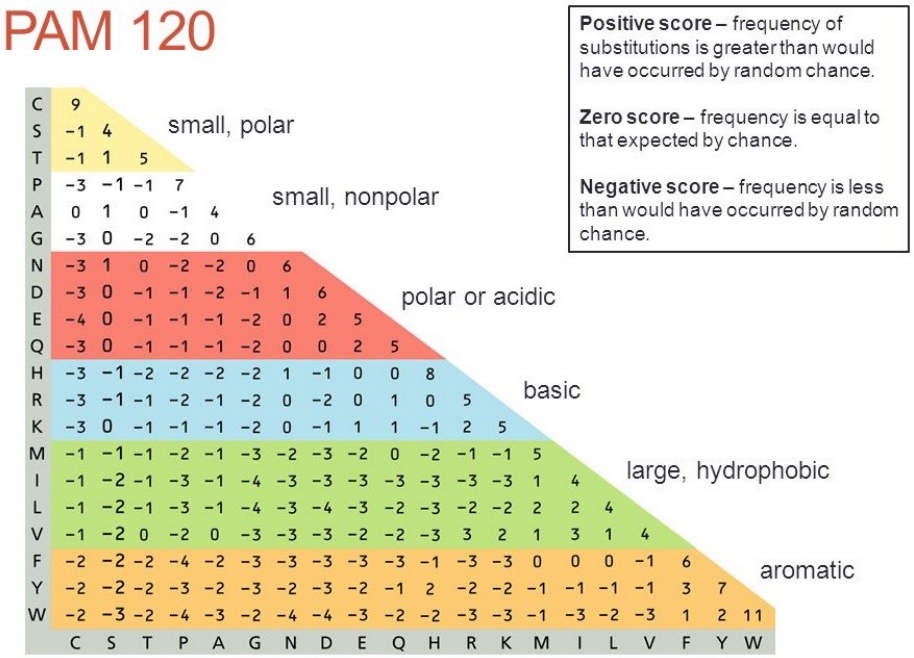
UNIVERSITÄT BERN

- 1965 publication of the **"Atlas of Protein Sequence and Structure"**

  → 1984 release of Protein Information Resource "PIR";

  → 2002, PIR, SwissProt and trEMBL merged into a single database UniProt (UNIversal PROTein resource).

- 1966: started to develop the PAM-Model (Point/Percent Accepted Mutation Matrix); substitution matrix



Foundation of Bioinformatics: Margaret Dayhoff

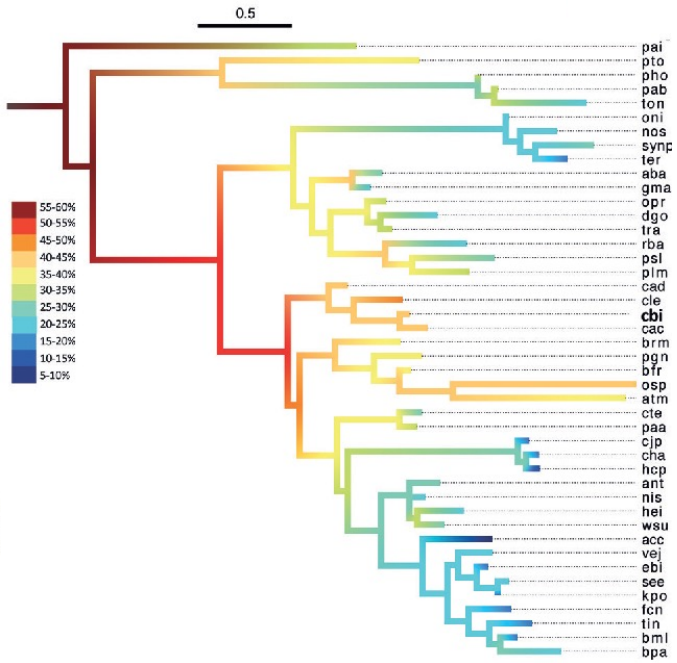
UNIVERSITÄT BERN

- 1965 publication of the **"Atlas of Protein Sequence and Structure"**

  → 1984 release of Protein Information Resource "PIR";

  → 2002, PIR, SwissProt and trEMBL merged into a single database UniProt (UNIversal PROTein resource).

- 1966: First programmatic inference of a phylogeny based on molecular/protein data.



ferredoxin

Foundation of Bioinformatics

UNIVERSITÄT BERN

What makes a bioinformatician?

- Development and sharing of computer programs/tools
- Resource Development
- Own discoveries based on biological data

Margaret Dayhoff (1925-83)

~ 1980

David Lipman (director of NCBI) stated in early 2000 that Dayhoff is the "mother and father of bioinformatics"

Sequencing Technologies

UNIVERSITÄT BERN

Sequencing Technologies

Bioinformatics

The Beginning: Sequencing

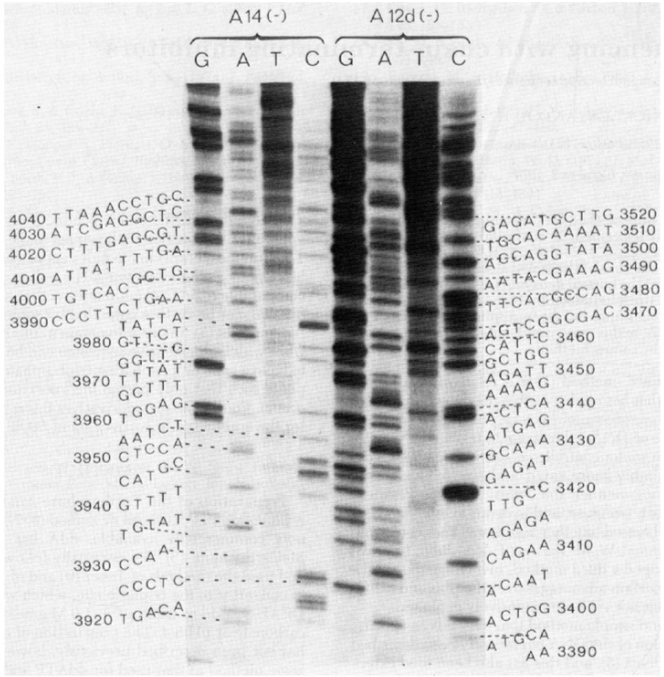
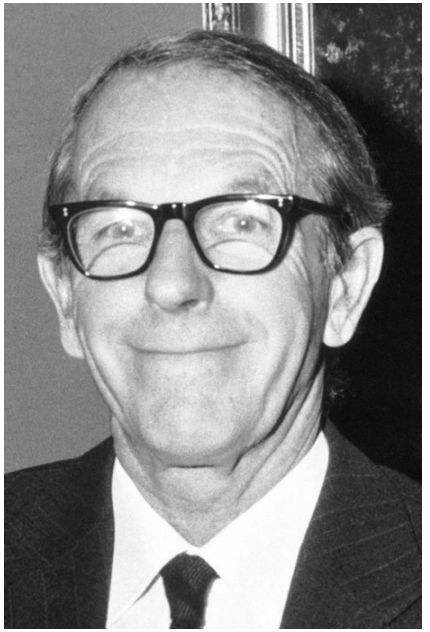


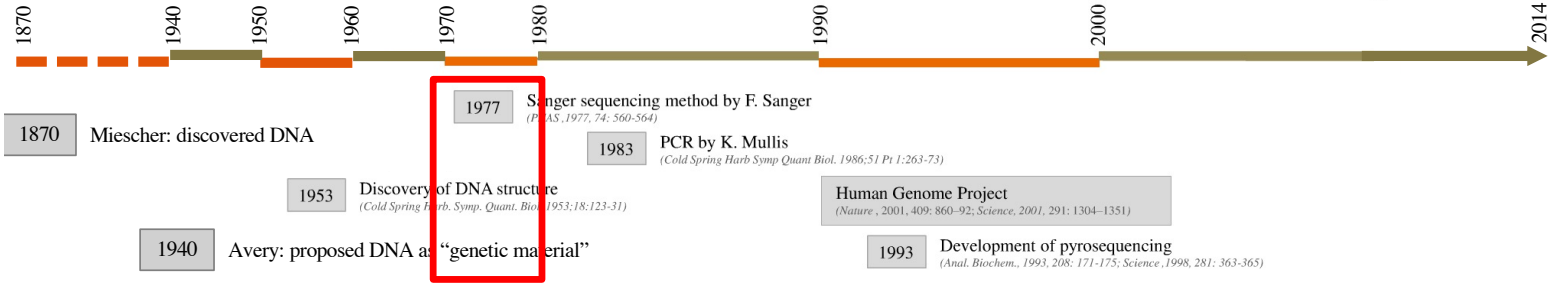
FIG. 1. Autoradiograph of the acrylamide gel from the sequence determination using restriction fragments A12d and A14 as primers on the complementary strand of ϕX174 DNA. The inhibitors used were (left to right) ddGTP, ddATP, ddTTP, and araCTP. Electrophoresis was on a 12% acrylamide gel at 40 mA for 14 hr. The top 10 cm of the gel is not shown. The DNA sequence is written from left to right and upwards beside the corresponding bands on the radioautograph. The numbering is as given in ref. 2.



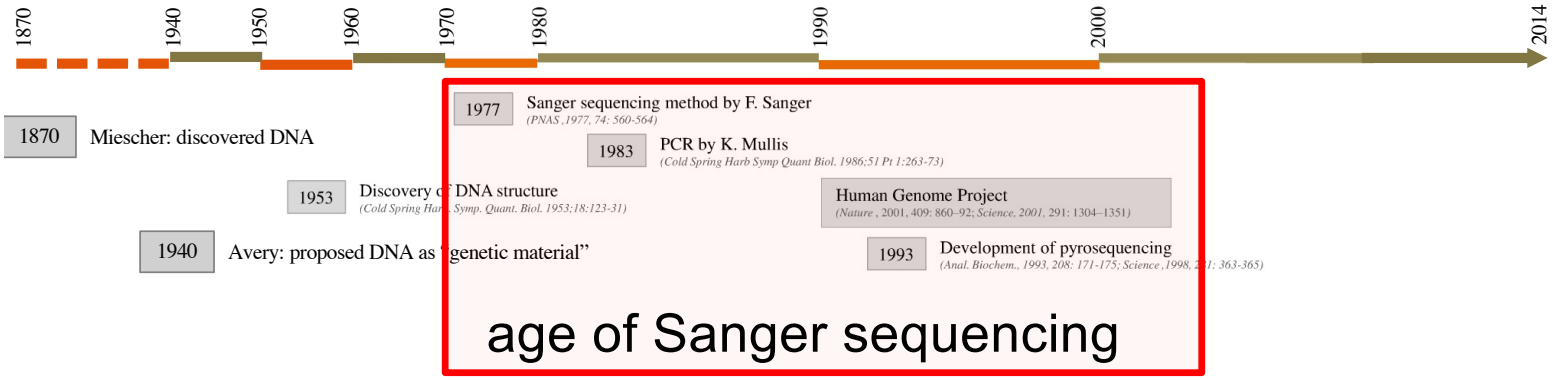
Nobelpreis in Chemie 1980

(zuvor schon Nobelpreis in Chemie 1958)

History of sequencing technologies (and DNA)



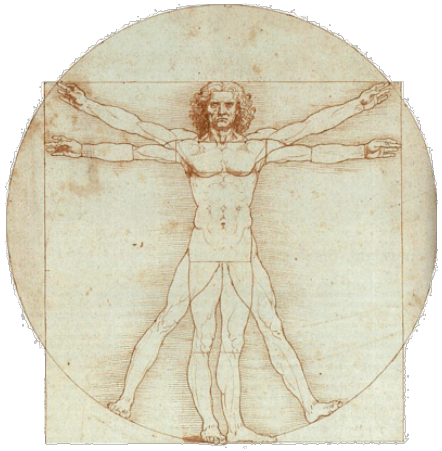
History of sequencing technologies (and DNA)



Human Genome Project (HGP)

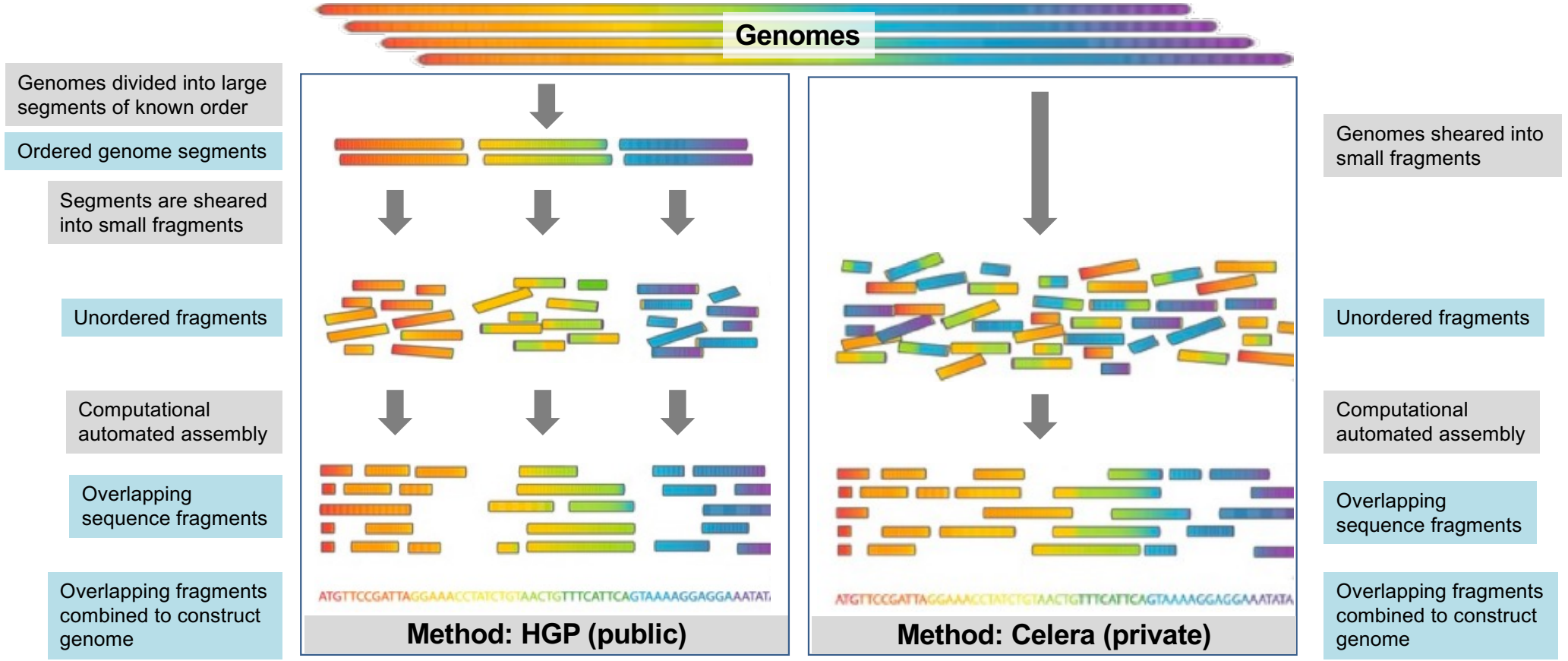
UNIVERSITÄT BERN

- 1985: Vision formulated by Robert Sinsheimer (NIH is uninterested in the first place)
- 1990: Start of the HGP
- 2001: Genome sequence published 4 years earlier than planned
- 2003: Official end of the HGP



Methodology used in the Human Genome Project

Shotgun Sequencing: Human Genome



First Publication of the Human Genome

Nature, 15. Februar 2001

Science, 16. Februar 2001

Sequencing costs of a human genome (Sanger)

UNIVERSITÄT BERN

- Throughput of one single instrument per 24 hours
  - 1.6 million bases (2304 sequences)
- Size of human genome (3.4×10^9)
  - 10x coverage needed,
    i.e. 3.4×10^{10} bp (34 billion bases)

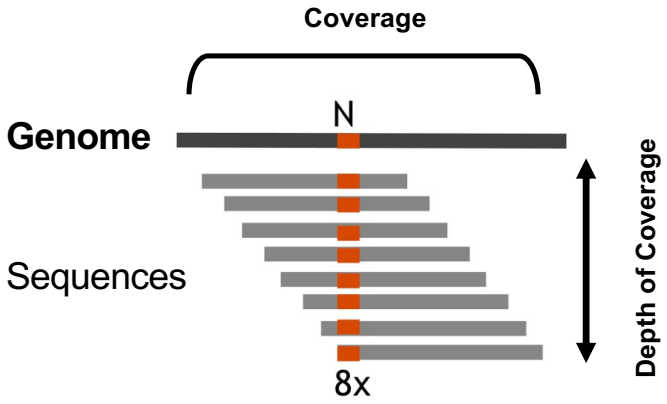
ABI 3730xl DNA Analyzer
"workhorse in HGP"

Throughput | Depth of Coverage | Coverage

UNIVERSITÄT BERN

$$\varnothing \text{ Depth of coverage} = \frac{\text{Througput [bp]}}{\text{Genome size [bp]}}$$

$$10x = \frac{34\ \text{Gbp}}{3.4\ \text{Gbp}}$$



Sequencing costs of a human genome (Sanger)

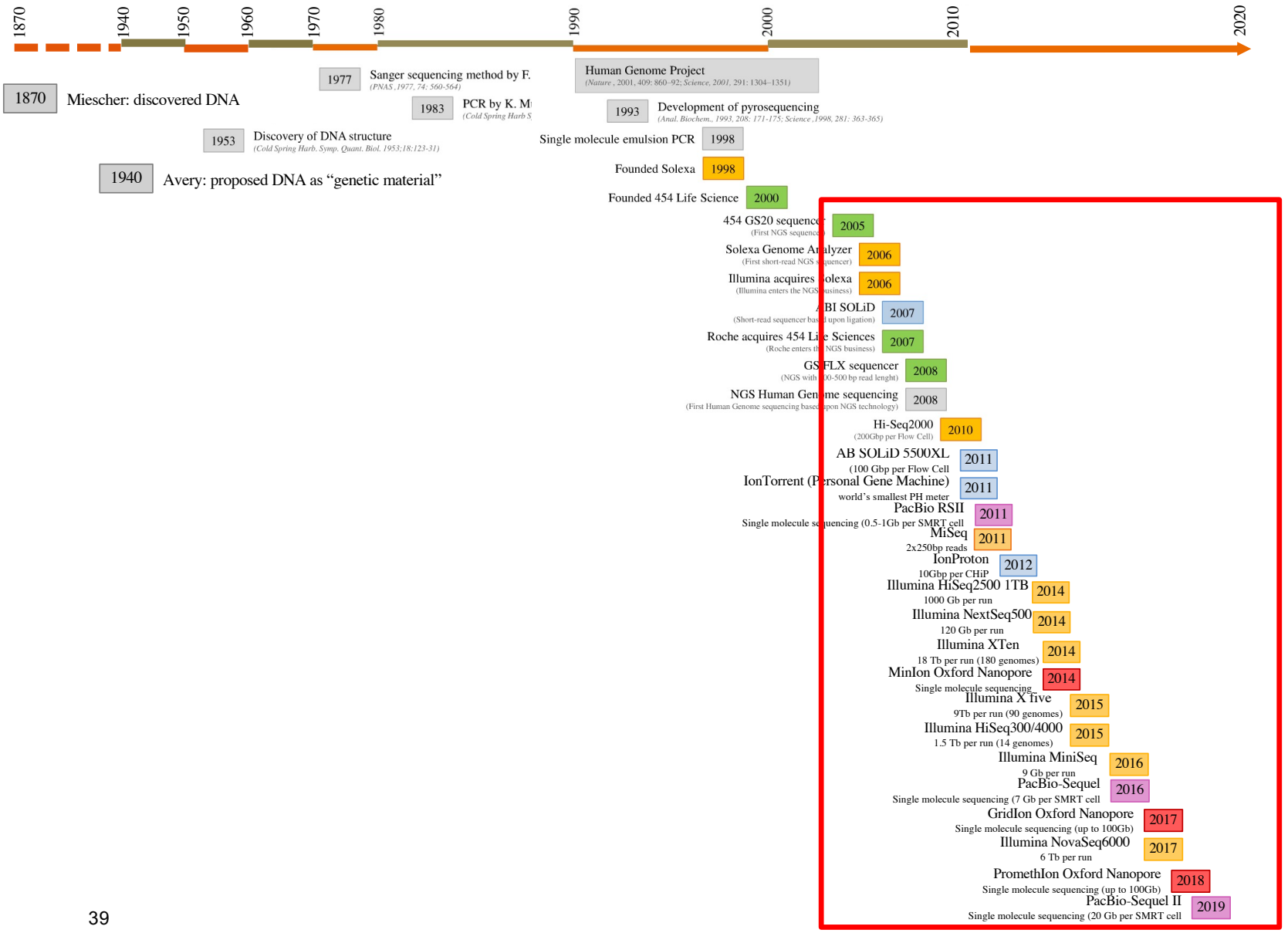
UNIVERSITÄT BERN

- Throughput of one single instrument per 24 hours
  - 1.6 million bases (2304 sequences)
- Size of human genome ($3.4x10^9$)
  - 10x coverage needed,
    i.e. $3.4x10^{10}$ bp (34 billion bases)
- Duration and costs to sequence the human at 10x
  - ~ 49 Millionen einzelne Sequenzreaktionen
  - costs per reaction: CHF 2-4
  - total costs : **CHF 100 - 200 Millionen**
  - duration about **58 years with a single instrument!**

Sequencing center end of 90's

Sequencing | Bioinformatics | IT

New sequencing methods required



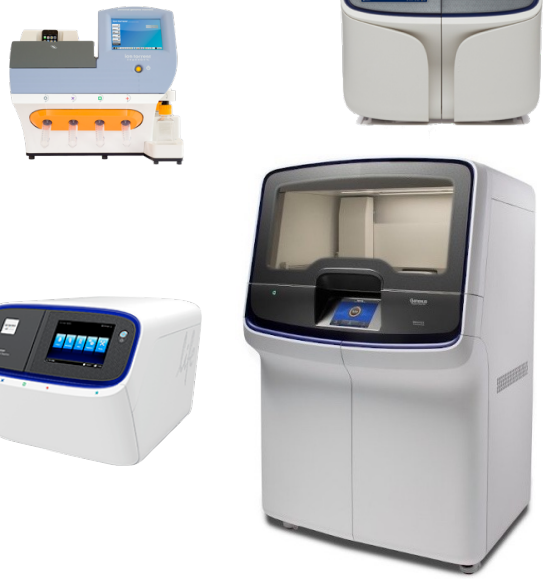
… to be continued…

Next Generation Sequencing (NGS) Instruments

UNIVERSITÄT BERN

Thermo Fisher (LifeTechnologies)

- IonTorrent (PGM)
- IonProton
- Ion GeneStudio S5
- Ion Torrent Genexus



Illumina

- MiSeq
- NextSeq500
- NextSeq550
- HiSeq2500
- HiSeq X ten
- NextSeq550
- HiSeq3000
- HiSeq4000
- HiSeq X five
- MiniSeq
- NextSeq1000/2000
- NovaSeq6000

images not to scale

3rd Generation Sequencing Instruments

UNIVERSITÄT BERN

Pacific Biosciences (PacBio)

PacBio RS II

PacBio Sequel I & II

Oxford Nanopore Technology (ONT)

MinION

MinION Mk1C

GridION

PromethION

images not to scale

Throughput and read length of NGS instruments

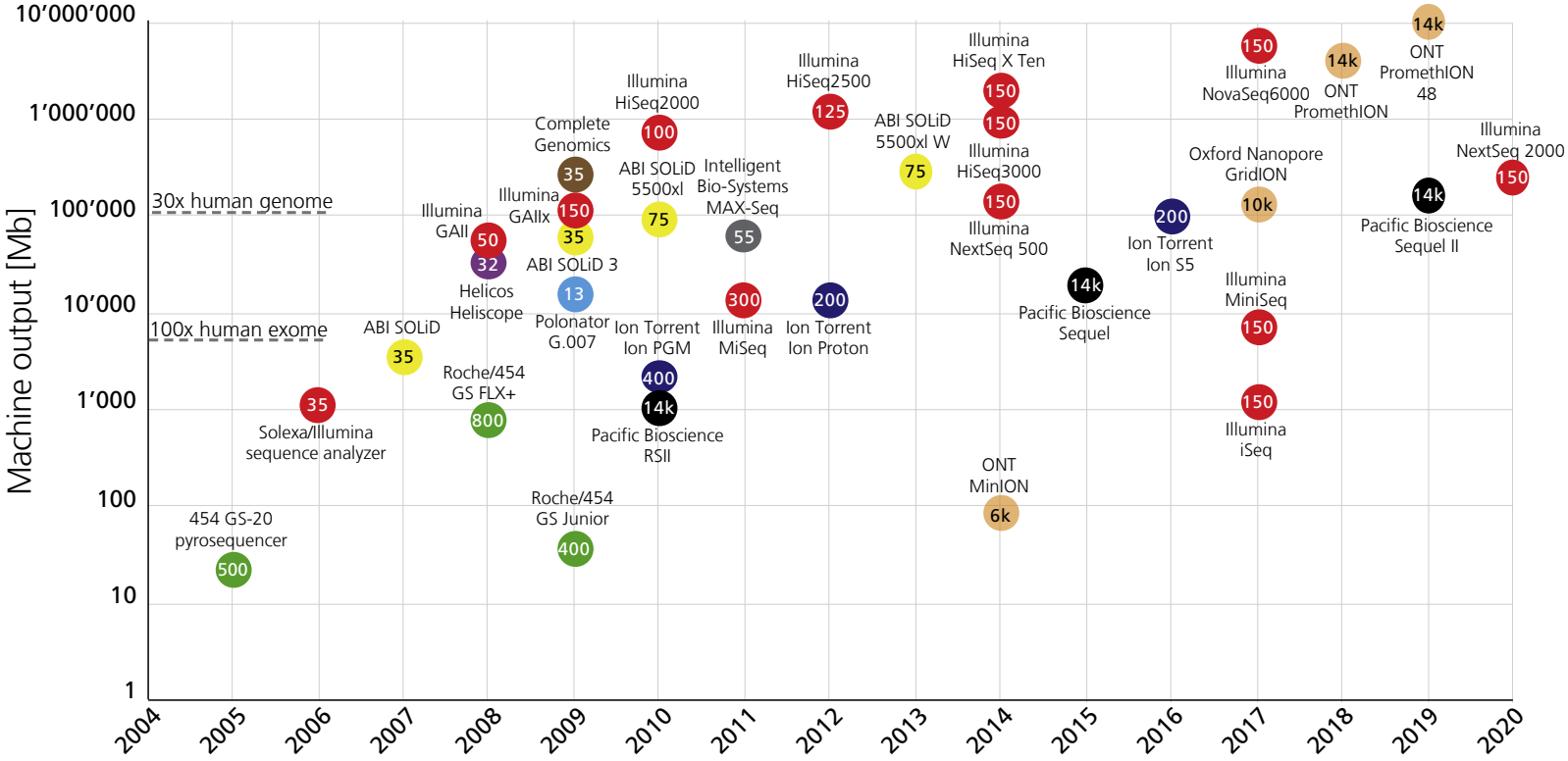
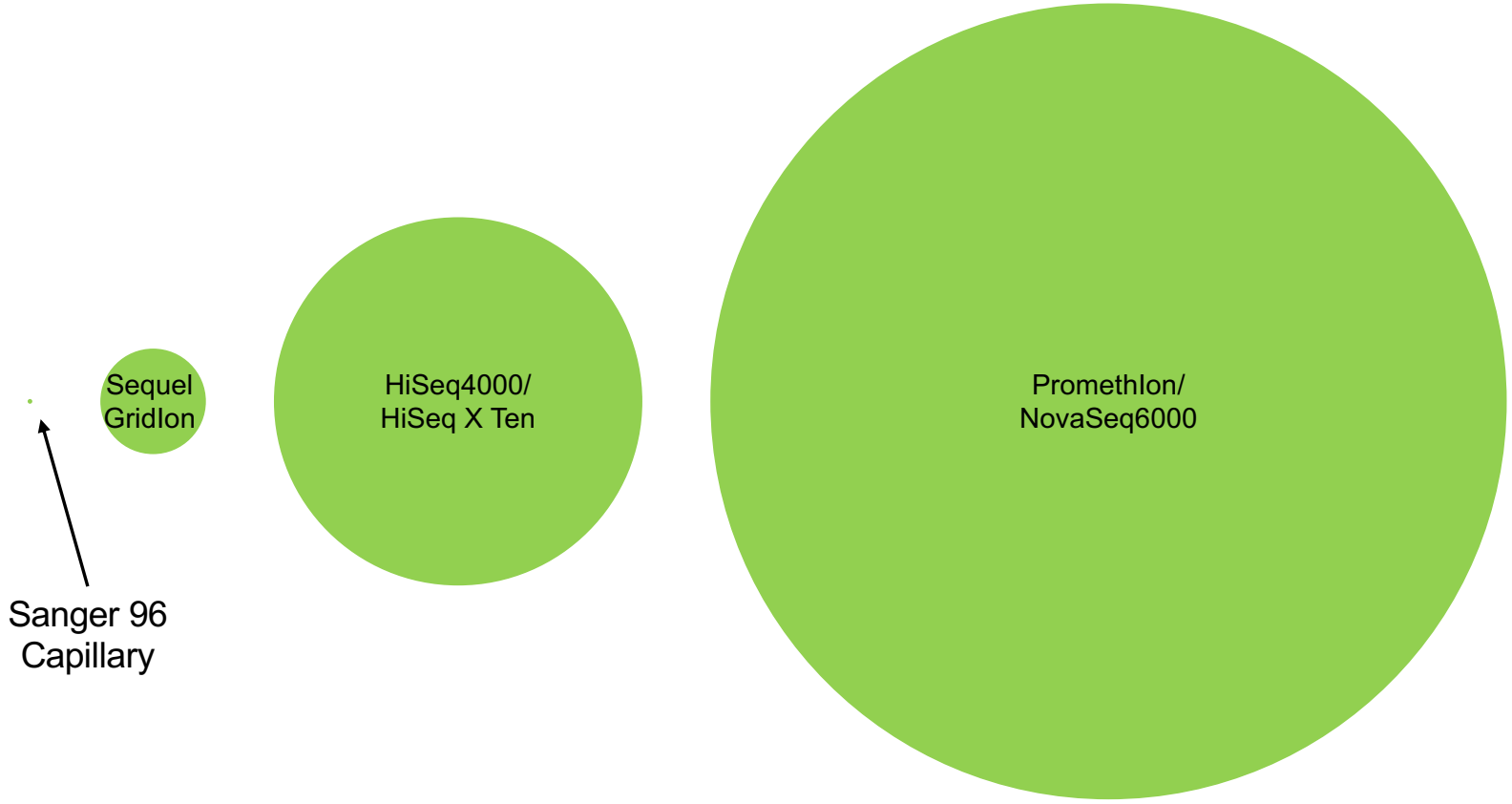
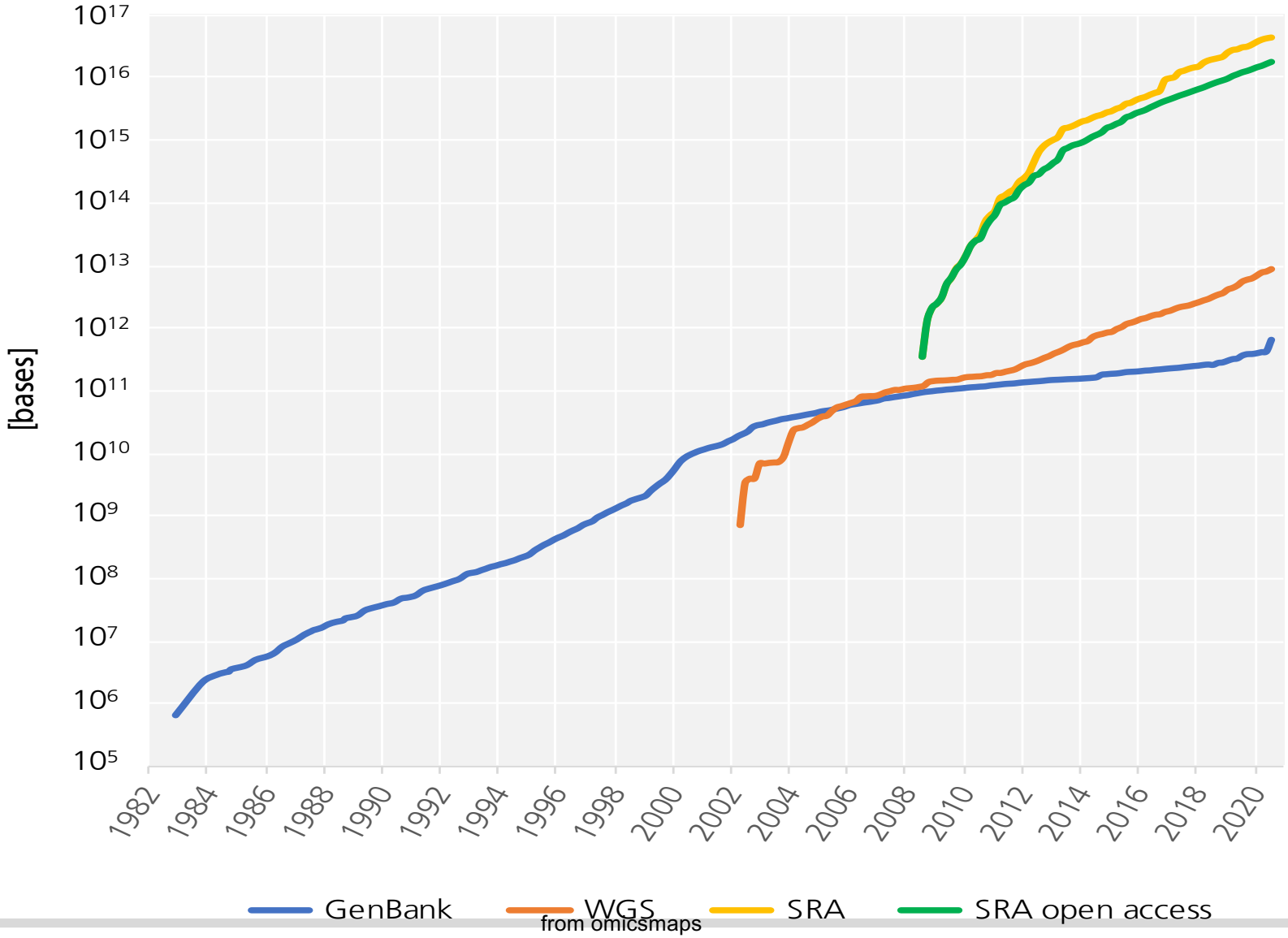


Figure complemented and adapted from Reuter et al. (2015)

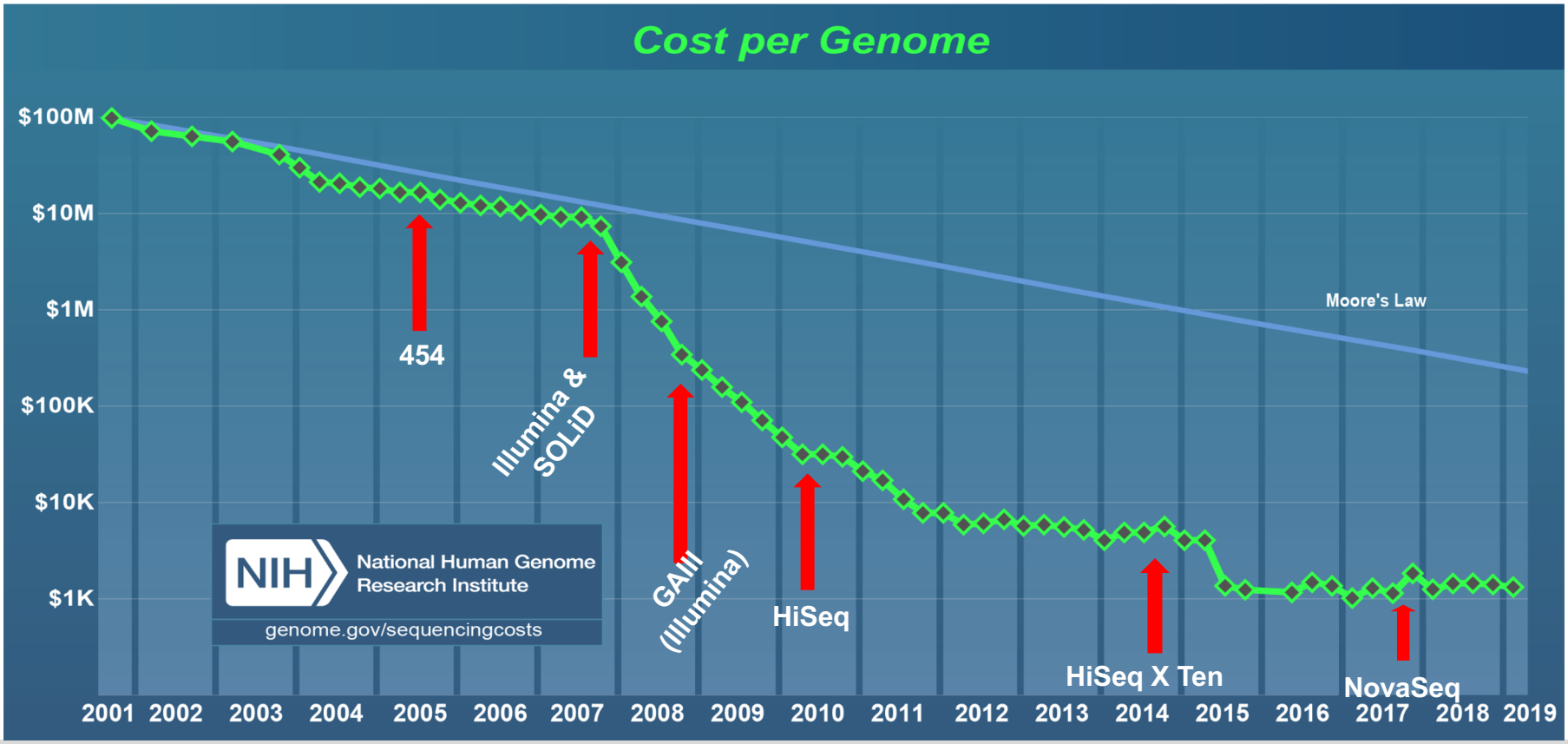
Throughput Comparison of Sequencing Instruments



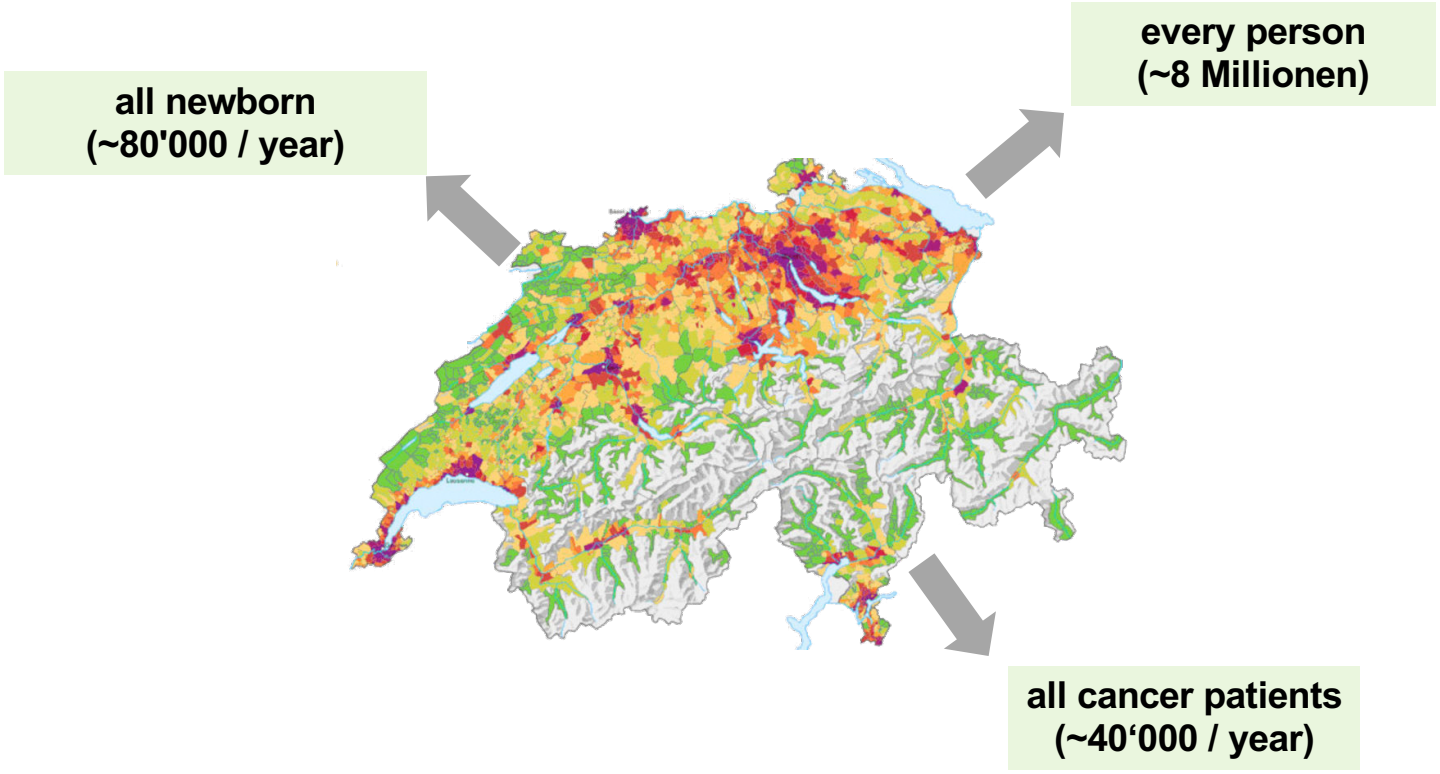


from omicsmaps

Cost per Human Genome over Time

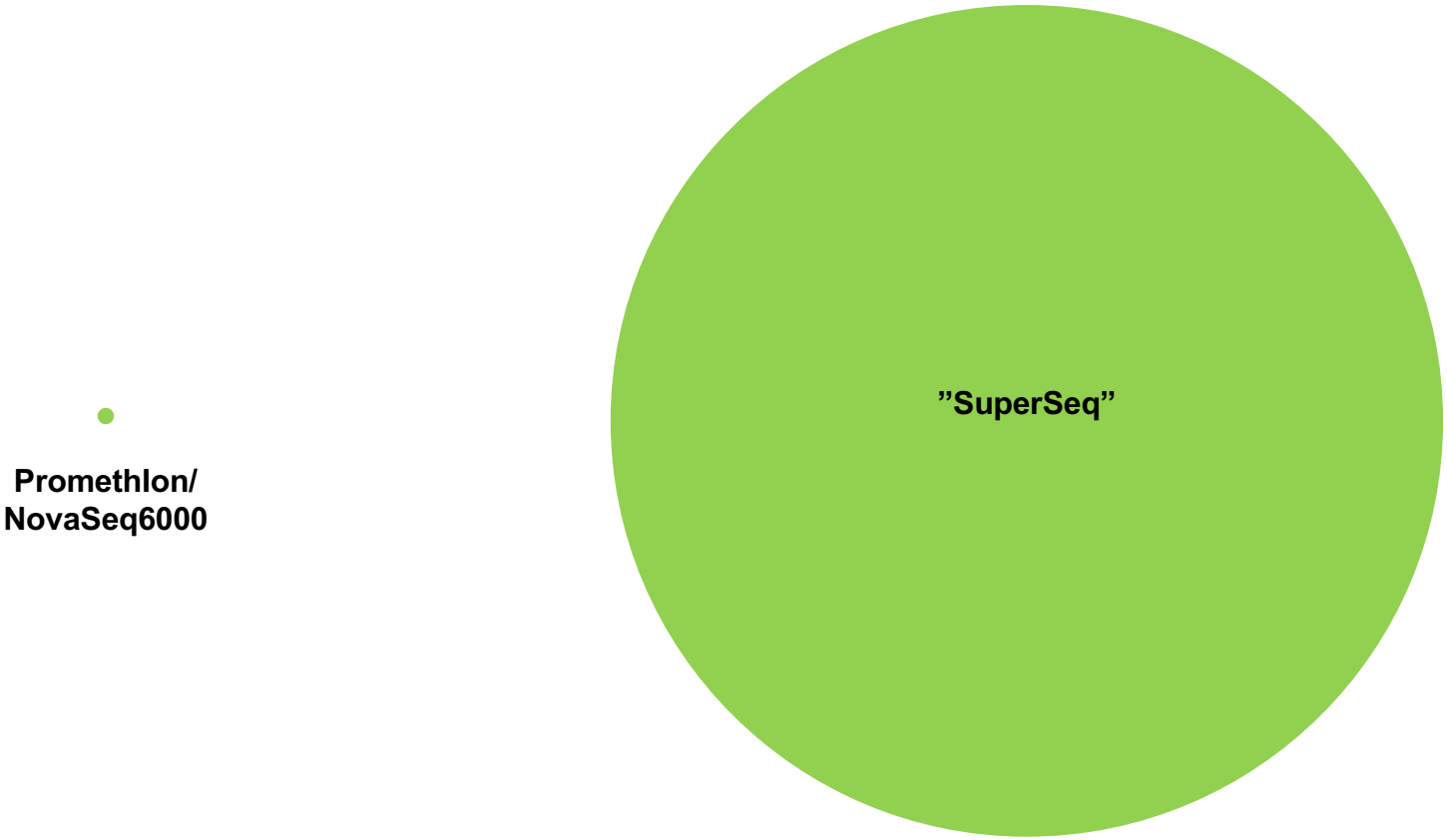


Future – Sequencing of all Genomes



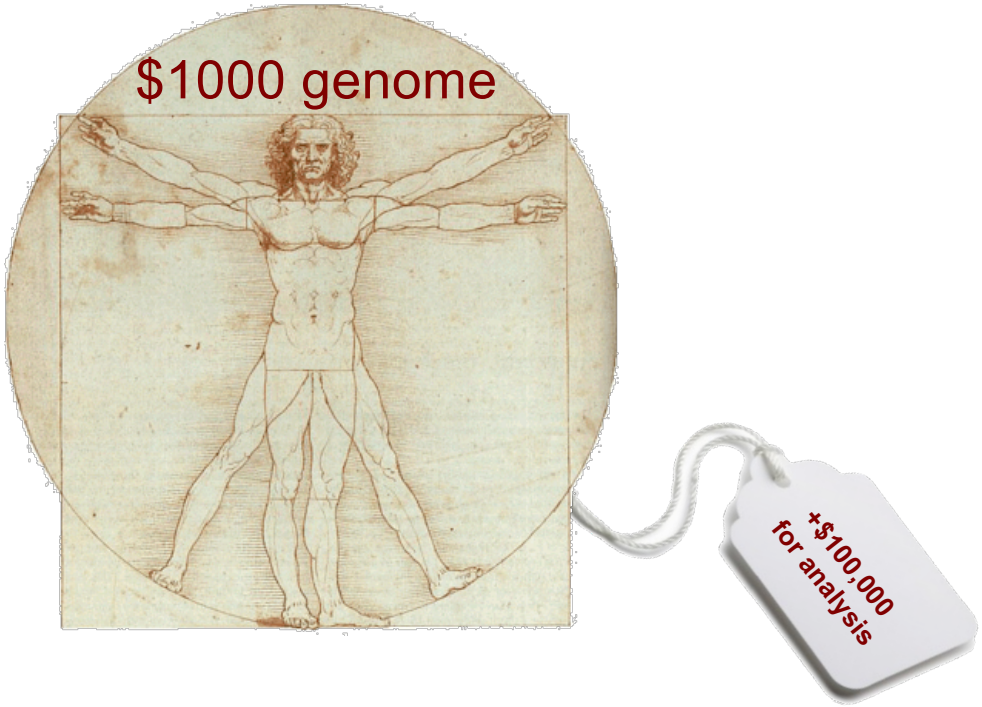
Sequencing | Bioinformatics | IT

Future – Super Sequencer



Sequencing | Bioinformatics | IT

True Cost of a Human Genome



Mardis *Genome Medicine* 2010, 2:84
http://genomemedicine.com/content/2/11/84

Genome Medicine

MUSINGS

The $1,000 genome, the $100,000 analysis?

Elaine R Mardis*

Having recently attended the Personal Genomes meeting at Cold Spring Harbor Laboratories (I was an organizer this year), I was struck by the number of talks that required for it to occur. I therefore offer the following as food for thought.

One source of difficulty in using resequencing

The $1,000 genome will come with a $100,000 analysis price tag?

The Human Genome

UNIVERSITÄT BERN

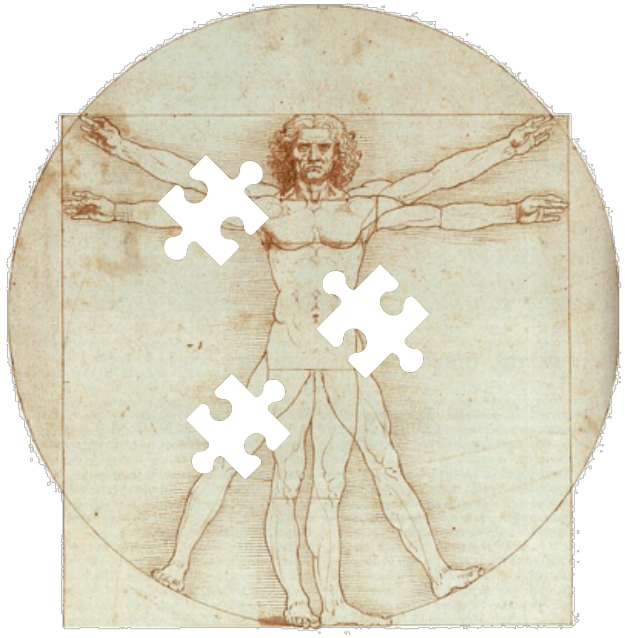
Psst, the human genome was never completely sequenced. Some scientists say it should be

By SHARON BEGLEY @sxbegle / JUNE 20, 2017

"As a matter of truth in advertising, **the 'finished' sequence isn't finished**," said Eric Lander, who led the lab at the Whitehead Institute that deciphered more of the genome for the government-funded Human Genome Project than any other. "I always say 'finished' is a term of art."

"**It's very fair to say the human genome was never fully sequenced**," Craig Venter, another genomics luminary, told STAT.

"**The human genome has not been completely sequenced and neither has any other mammalian genome** as far as I'm aware," said Harvard Medical School bioengineer George Church, who made key early advances in sequencing technology.



still missing sequences of the human genome
4-9%
(Georg Church, Karen Miga)

Telomere-to-telomere Initiative

Article

Telomere-to-telomere assembly of a complete human X chromosome

https://doi.org/10.1038/s41586-020-2547-7
Received: 30 July 2019
Accepted: 29 Ma
Published online:
Open access
Check for upd

Karen H. Miga[1,24], Sergey Koren[2,24], Arang Rhie[2], Mitchell R. Vollger[3], Ariel Gershman[4], Andrey Bzikadze[5], Shelise Brooks[6], Edmund Howe[7], David Porubsky[3], Glennis A. Logsdon[3], Valerie A. Schneider[8], Tamara Potapova[7], Jonathan Wood[9], William Chow[9], Joel Armstrong[1],

Article

The structure, function and evolution of a complete human chromosome 8

https://doi.org/10.1038/s41586-021-03420-7
Received: 4 September 2020
Accepted: 4 March 2021
Published online: 07 April 2021
Open access
Check for updates

Glennis A. Logsdon[1], Mitchell R. Vollger[1], PingHsun Hsieh[1], Yafei Mao[1], Mikhail A. Liskovykh[2], Sergey Koren[3], Sergey Nurk[3], Ludovica Mercuri[4], Philip C. Dishuck[1], Arang Rhie[3], Leonardo G. de Lima[5], Tatiana Dvorkina[6], David Porubsky[1], William T. Harvey[1], Alla Mikheenko[6], Andrey V. Bzikadze[7], Milinn Kremitzki[8], Tina A. Graves-Lindsay[8], Chirag Jain[3], Kendra Hoekzema[1], Shwetha C. Murali[1,9], Katherine M. Munson[1], Carl Baker[1], Melanie Sorensen[1], Alexandra M. Lewis[1], Urvashi Surti[10], Jennifer L. Gerton[5], Vladimir Larionov[2], Mario Ventura[4], Karen H. Miga[11], Adam M. Phillippy[3] & Evan E. Eichler[1,9]

The complete assembly of each human chromosome is essential for understanding human biology and evolution[1,2]. Here we use complementary long-read sequencing technologies to complete the linear assembly of human chromosome 8. Our assembly resolves the sequence of five previously long-standing gaps, including a 2.08-Mb centromeric α-satellite array, a 644-kb copy number polymorphism in the β-defensin gene cluster that is important for disease risk, and an 863-kb variable number tandem

b UNIVERSITÄT

The complete sequence of a human genome

Sergey Nurk[1,*], Sergey Koren[1,*], Arang Rhie[1,*], Mikko Rautiainen[1,*], Andrey V. Bzikadze[2], Alla Mikheenko[3], Mitchell R. Vollger[4], Nicolas Altemose[5], Lev Uralsky[6,7], Ariel Gershman[8], Sergey Aganezov[9], Savannah J. Hoyt[10], Mark Diekhans[11], Glennis A. Logsdon[4], Michael Alonge[9], Stylianos E. Antonarakis[12], Matthew Borchers[13], Gerard G. Bouffard[14], Shelise Y. Brooks[14], Gina V. Caldas[15], Haoyu Cheng[16,17], Chen-Shan Chin[18], William Chow[19], Leonardo G. de Lima[13], Philip C. Dishuck[4], Richard Durbin[21], Tatiana Dvorkina[3], Ian T. Fiddes[22], Giulio Formenti[23,24], Robert S. Fulton[25], Arkarachai Fungtammasan[18], Erik Garrison[11,26], Patrick G.S. Grady[10], Tina A. Graves-Lindsay[27], Ira M. Hall[28], Nancy F. Hansen[29], Gabrielle A. Hartley[10], Marina Haukness[11], Kerstin Howe[19], Michael W. Hunkapiller[30], Chirag Jain[1,31], Miten Jain[11], Erich D. Jarvis[23,24], Peter Kerpedjiev[32], Melanie Kirsche[9], Mikhail Kolmogorov[33], Jonas Korlach[30], Milinn Kremitzki[27], Heng Li[16,17], Valerie V. Maduro[34], Tobias Marschall[35], Ann M. McCartney[1], Jennifer McDaniel[36], Danny E. Miller[4,37], James C. Mullikin[14,29], Eugene W. Myers[38], Nathan D. Olson[36], Benedict Paten[11], Paul Peluso[30], Pavel A. Pevzner[33], David Porubsky[4], Tamara Potapova[13], Evgeny I. Rogaev[6,7,39,40], Jeffrey A. Rosenfeld[41], Steven L. Salzberg[9,42], Valerie A. Schneider[43], Fritz J. Sedlazeck[44], Kishwar Shafin[11], Colin J. Shew[20], Alaina Shumate[42], Yumi Sims[19], Arian F. A. Smit[45], Daniela C. Soto[20], Ivan Sović[30,46], Jessica M. Storer[45], Aaron Streets[5,47], Beth A. Sullivan[48], Françoise Thibaud-Nissen[43], James Torrance[19], Justin Wagner[36], Brian P. Walenz[1], Aaron Wenger[30], Jonathan M. D. Wood[19], Chunlin Xiao[43], Stephanie M. Yan[49], Alice C. Young[14], Samantha Zarate[9], Urvashi Surti[50], Rajiv C. McCoy[49], Megan Y. Dennis[20], Ivan A. Alexandrov[3,7,51], Jennifer L. Gerton[13], Rachel J. O'Neill[10], Winston Timp[8,42], Justin M. Zook[36], Michael C. Schatz[9,49], Evan E. Eichler[4,24,†], Karen H. Miga[11,†], Adam M. Phillippy[1,†]

[1-51] Affiliations are listed at the end
[*] Equal contribution
[†] Corresponding authors: Evan E. Eichler (eee@gs.washington.edu); Karen H. Miga (khmiga@ucsc.edu); Adam M. Phillippy (adam.phillippy@nih.gov)

Abstract
In 2001, Celera Genomics and the International Human Genome Sequencing Consortium published their initial drafts of the human genome, which revolutionized the field of genomics. While these drafts and the updates that followed effectively covered the euchromatic fraction of the genome, the heterochromatin and many other complex regions were left unfinished or erroneous. Addressing this remaining 8% of the genome, the Telomere-to-Telomere (T2T) Consortium has finished the first truly complete 3.055 billion base pair (bp) sequence of a human genome, representing the largest improvement to the human reference genome since its initial release. The new T2T-CHM13 reference includes gapless assemblies for all 22 autosomes plus Chromosome X, corrects numerous errors, and introduces nearly 200 million bp of novel sequence containing 2,226 paralogous gene copies, 115 of which are predicted to be protein coding. The newly completed regions include all centromeric satellite arrays and the short arms of all five acrocentric chromosomes, unlocking these complex regions of the genome to variational and functional studies for the first time.

Nurk S, Koren S, Rhie A, Rautiainen M, et al. The complete sequence of a human genome. bioRxiv, 2021.

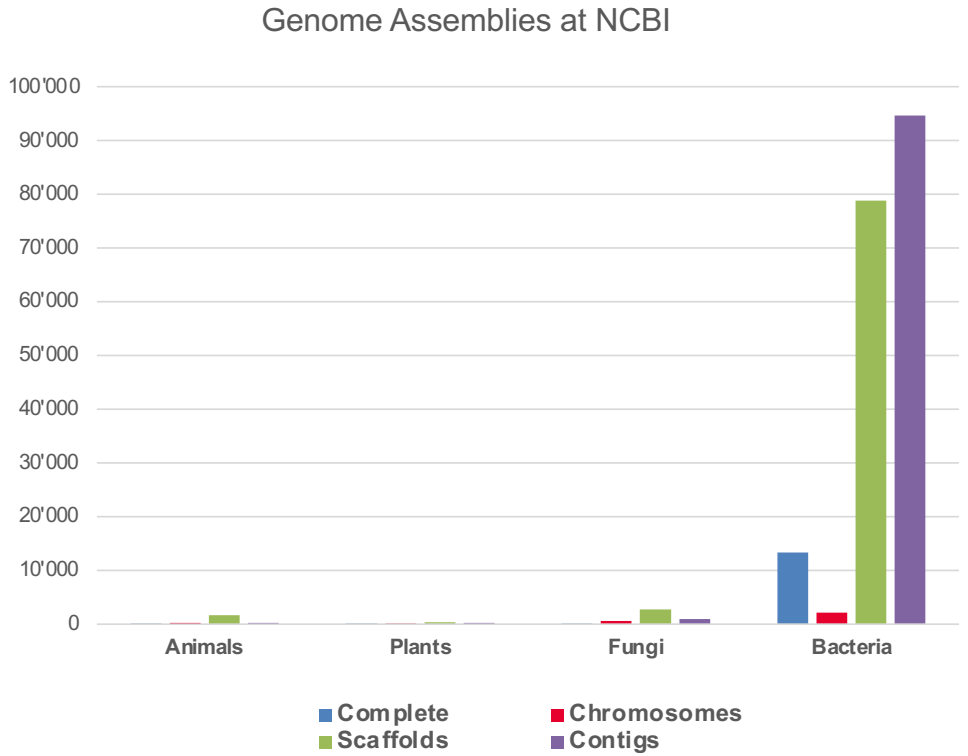
Genomes in Public Databases (NCBI)

Genomes in Public Databases (NCBI)

UNIVERSITÄT BERN

Assembly Level	Total	Complete		Chromosomes		Scaffolds		Contigs	
Animals	3'853	3	(0.1%)	695	(18.0%)	2'773	(72.0%)	382	(9.9%)
Plants	1'488	3	(0.2%)	416	(28.0%)	665	(44.7%)	404	(27.2%)
Fungi	6'886	86	(1.2%)	826	(12.0%)	4'390	(63.8%)	1'584	(23.0%)
Protists	943	18	(1.9%)	119	(12.6%)	593	(62.9%)	213	(22.6%)
Eukaryotes	13'170	110	(0.8%)	2'056	(15.6%)	8'421	(63.9%)	2'583	(19.6%)
Bacteria	276'806	19'884	(7.2%)	3'270	(1.2%)	102'252	(36.9%)	151'400	(54.7%)
Archaea	5'615	401	(7.1%)	25	(0.4%)	2'163	(38.5%)	3'026	(54.7%)
∑ Prokaryotes	282'421	20'285	(7.2%)	3'295	(1.2%)	104'415	(37.0%)	154'426	(54.7%)
Overall	295'591	20'395	(6.9%)	5'351	(1.8%)	112'836	(38.2%)	157'009	(53.1%)

*data accessed October 2020 at NCBI genomes
https://www.ncbi.nlm.nih.gov/genome/browse#!/overview/



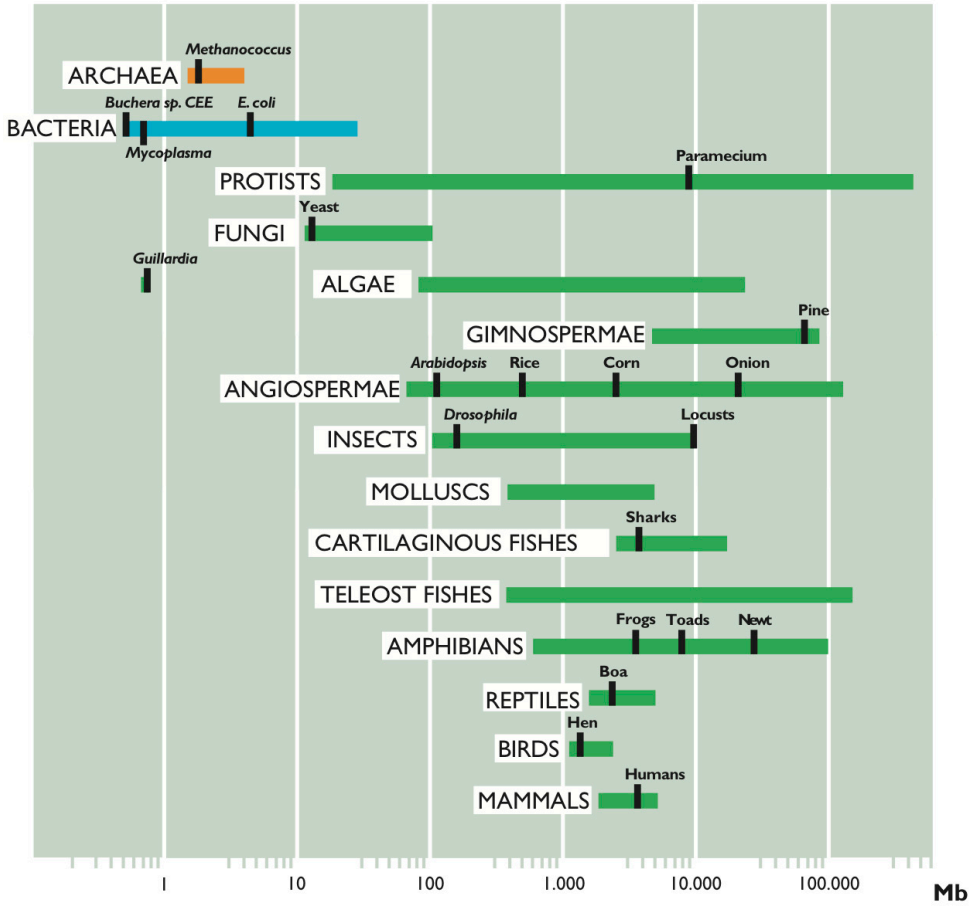
Why are only very few genomes complete?

UNIVERSITÄT BERN

- Genome size
- Heterozygosity levels
- Repeat-content
- GC-content
- Secondary structure
- Ploidy level

- Genome sizes range from 100 kbp to 150 Gbp
- The larger the genome, the more data is needed to assemble it (>50x usually)
- Compute resource needs grow with increased amount of data (running time and memory)
- Larger genomes do not necessarily have to be harder to assemble,
  although empirically this is often the case

Genome Sizes & Gene Density



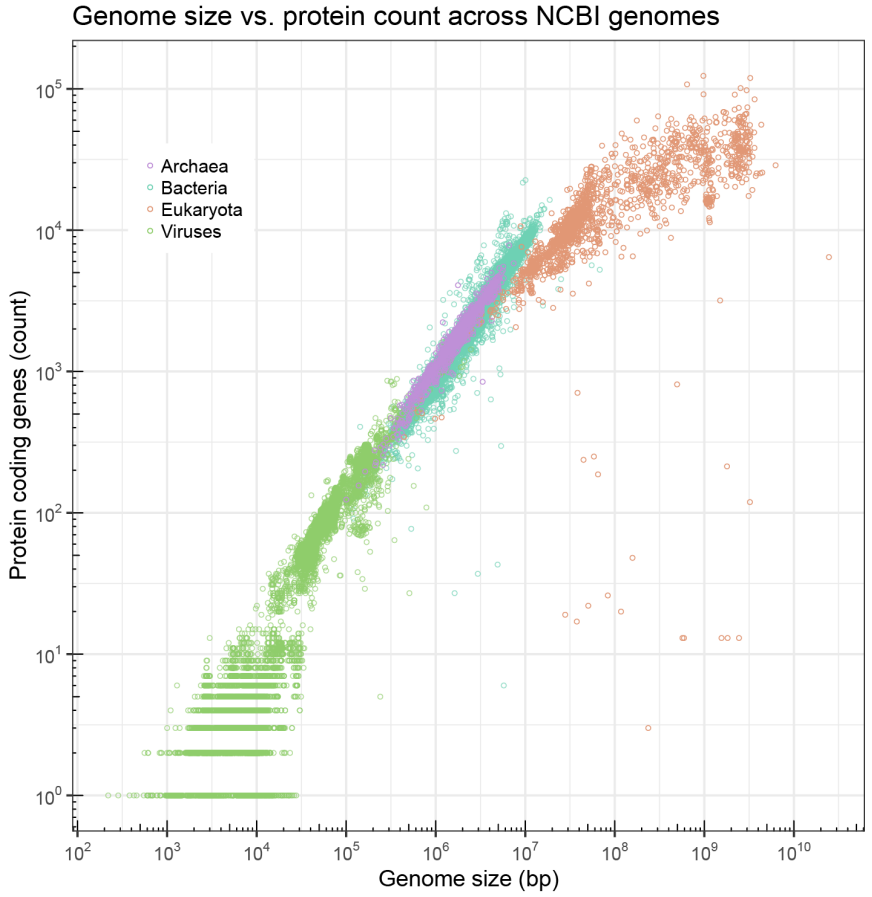
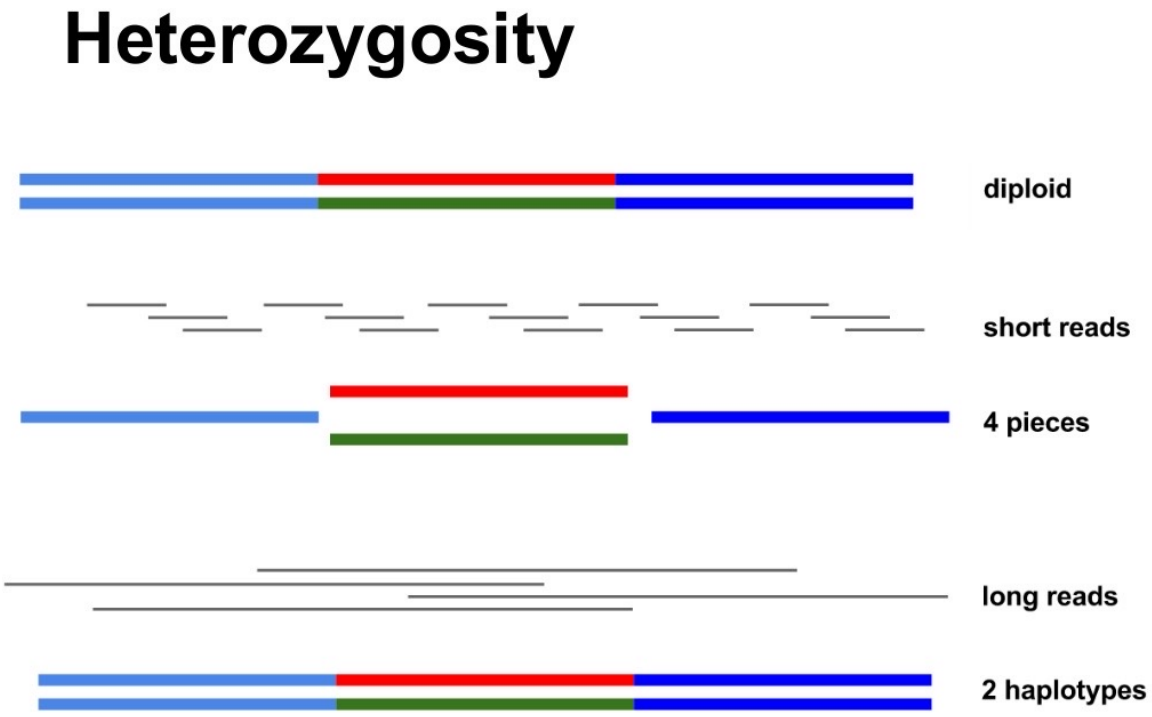


figure from Latorre, A., & Silva, F J. (2014). Mètode Science Studies Journal, 4.

from Wikipedia

Heterozygosity

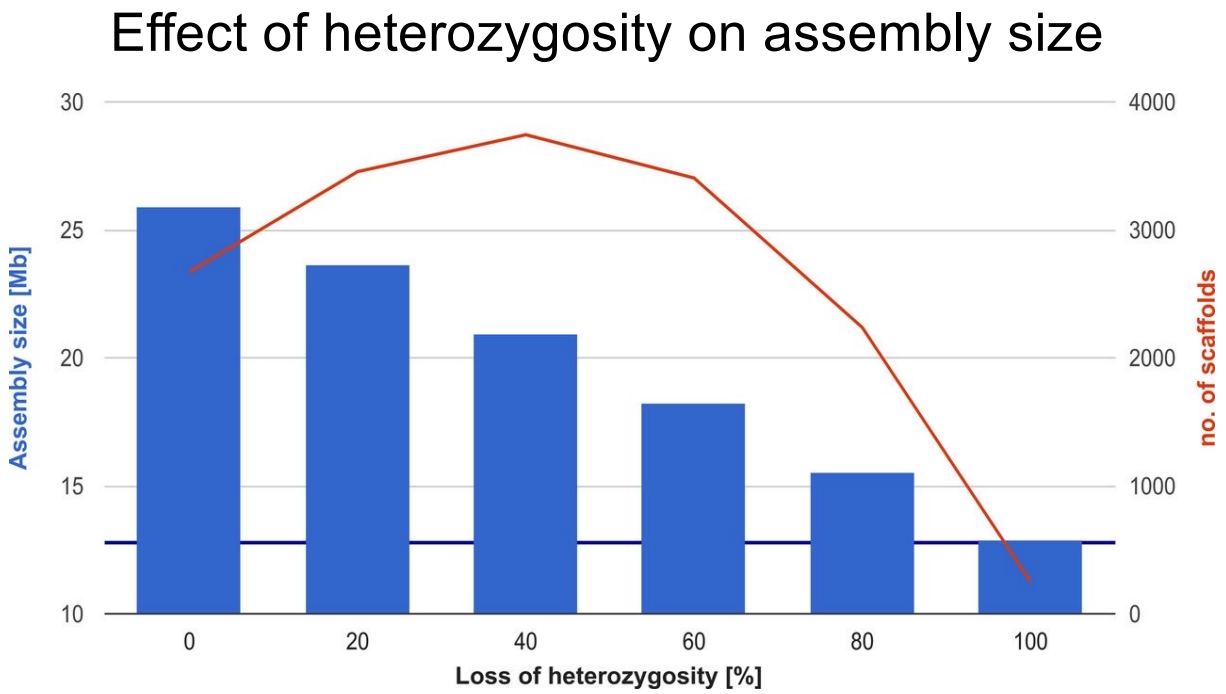


(Slide Torsten Seeman, Victorian Life Sciences Computation Initiative)

Heterozygosity

UNIVERSITÄT BERN

- Highly heterozygous regions tend to be assembled separately
- Homologous regions existing in multiple copies in the assembly
- Downstream problems in determining orthology for gene-based analyses, comparative genomics etc.



(Pryszcz and Gabaldon (2016) Nucl. Acids. res.)

De novo Genome Project Workflow

De novo Genome Project Workflow

1. Extract DNA (and RNA)
2. Choose best sequence technology for the project
3. Sequencing
4. Quality assessment and other pre-assembly investigations
5. Assembly
6. Assembly validation
7. Assembly comparisons
8. Repeat masking?
9. Annotation

DNA Extraction

- Extract much more DNA than you think you need
- Also remember to extract RNA for the annotation
- Single individual and haploid tissue if possible
- In particular for Illumina mate-pairs data and PacBio, a lot of high molecular weight DNA is critical!
- Extracting DNA for *de novo* assembly is very different from extractions intended for PCR
- Do several extractions if possible, and run them on a gel to get an idea of how fragmented the DNA is
- Try to remove contaminants from the extractions

Causes of DNA Degradation

UNIVERSITÄT BERN

- **Mechanical damage** during tissue homogenization.
- **Wrong pH and ionic strength** of extraction buffer.
- Incomplete removal / contamination with **nucleases**.
- **Phenol**: too old, or inappropriately buffered (**pH 7.8 – 8.0**); incomplete removal.
- Wrong pH of **DNA solvent** (acidic water).
- *Recommended: 1:10 TE for short-term storage, or 1xTE for long-term storage*.
- **Vigorous pipetting** (wide-bore pipet tips).
- **Vortexing** of DNA in high concentrations.
- Too many **freeze-thaw** cycles
- Debatable: sequence-dependent

What are the Main Contaminants?

Polysaccharides
Lypopolysaccharides
Growth media residuals

Chitin
Protein
Secondary metabolites
Pigments
Growth media residuals

Chitin
Fats
Proteins
Pigments

Polyphenols
Polysaccharides
Secondary metabolites
Pigments

What do Absorption Ratios Tell us?

UNIVERSITÄT BERN

Pure DNA 260/280: 1.8 – 2.0

< 1.8:

Too little DNA compared to other components of the solution; presence of organic contaminants: proteins and phenol; glycogen - **absorb at 280 nm**.

> 2.0:

High share of RNA.

Pure DNA 260/230: 2.0 – 2.2

<2.0:

Salt contamination, humic acids, peptides, aromatic compounds, polyphenols, urea, guanidine, thiocyanates (latter three are common kit components) – **absorb at 230 nm**.

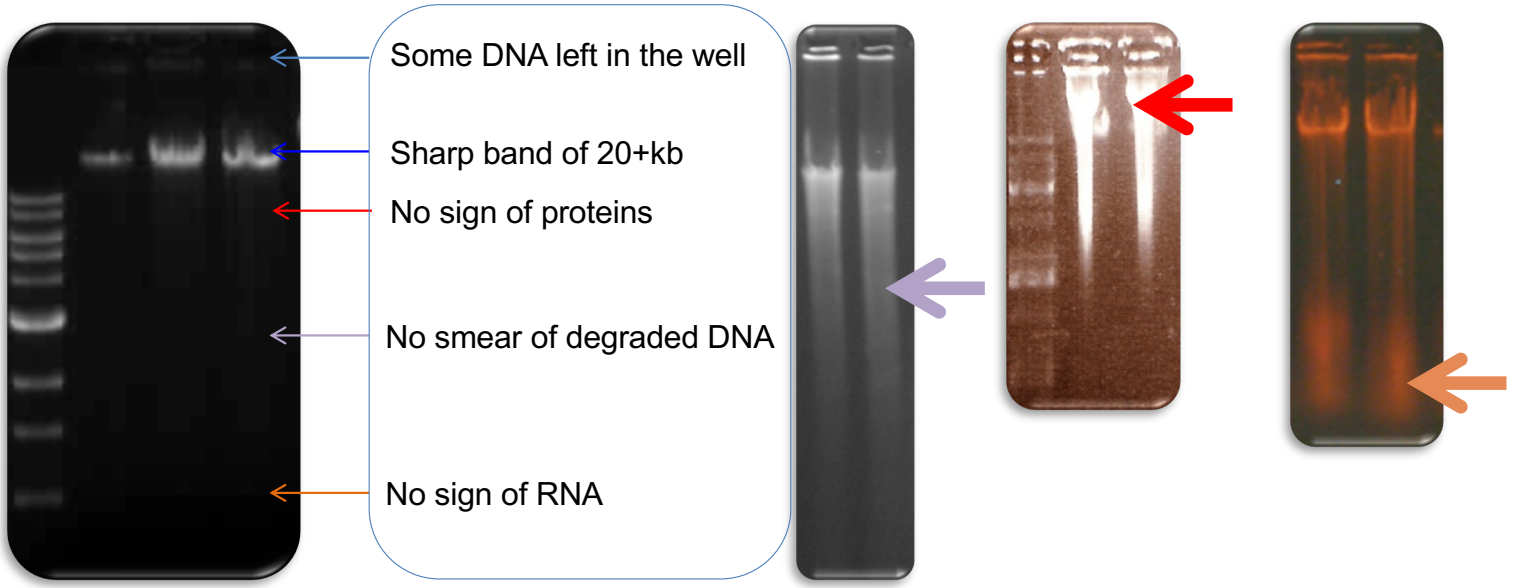
>2.2:

High share of RNA, very high share of phenol, **high turbidity,** dirty instrument, wrong blank.

Photometrically active contaminants

phenol
polyphenols
EDTA
thiocyanate
protein
RNA
nucleotides
(> 5 bp)

DNA Requirements

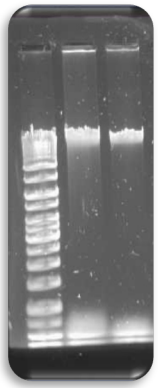


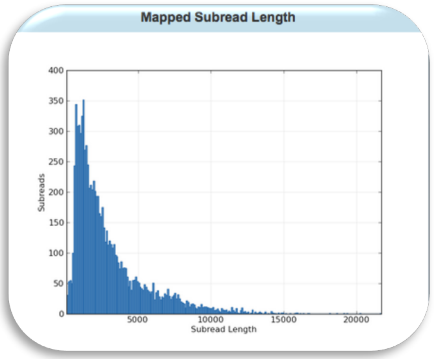
NanoDrop:

260/280 = 1.8 – 2.0
260/230 = 2.0 – 2.2

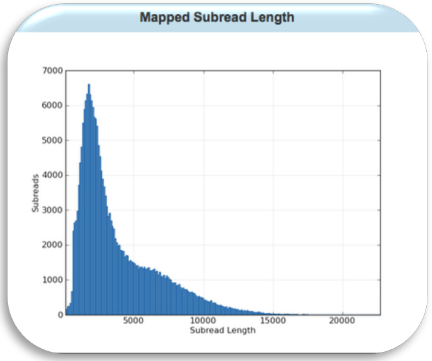
Qubit or Picogreen:

10 kb insert libraries: 3-5 ug
20 kb insert libraries: 10-20 ug



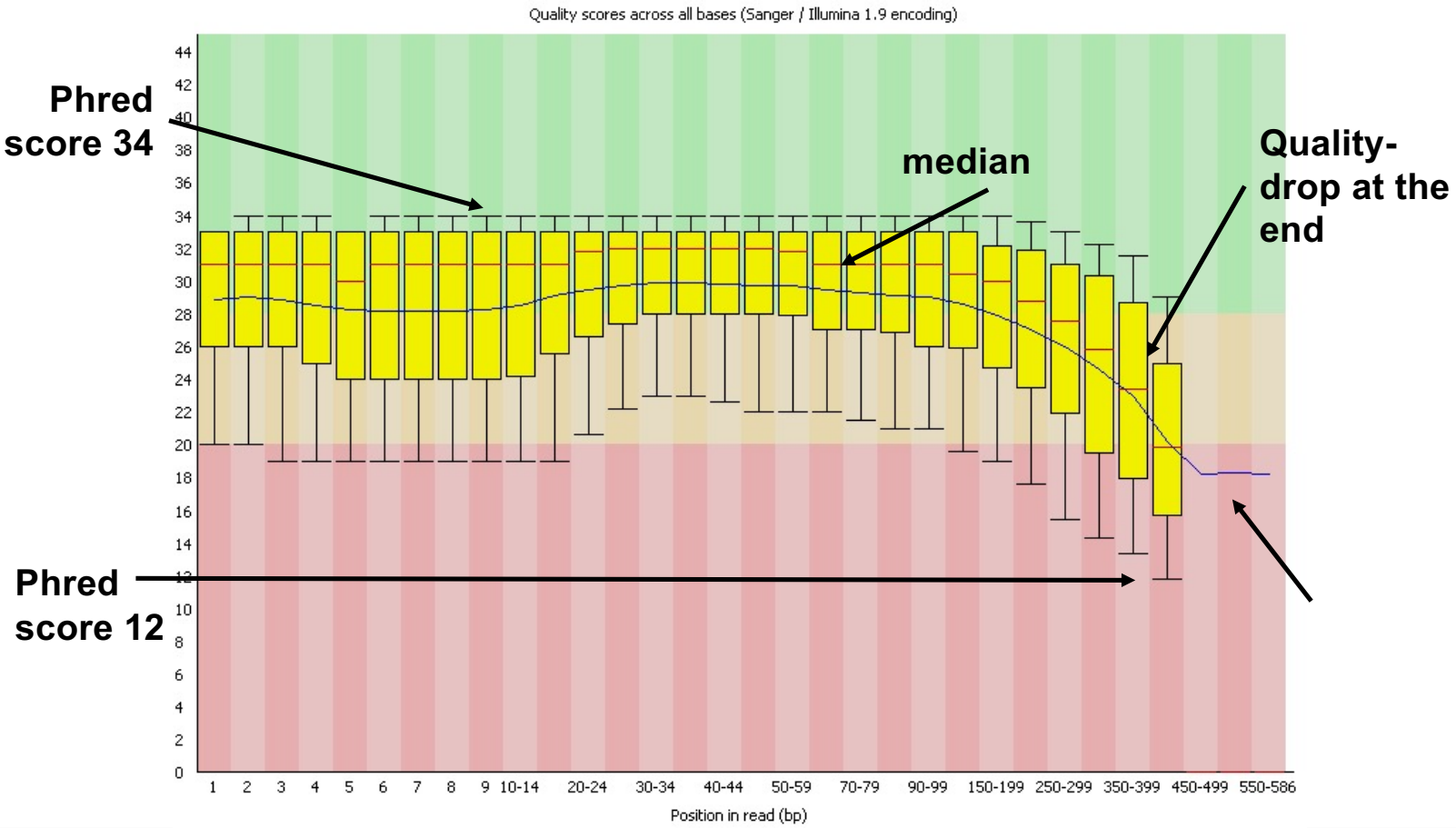


Polished Contigs	223	Max Contig Length	36,298
N50 Contig Length	2,932	Sum of Contig Lengths	480,087



Polished Contigs	9	Max Contig Length	1,508,929
N50 Contig Length	1,353,702	Sum of Contig Lengths	7,813,244

Quality of Reads

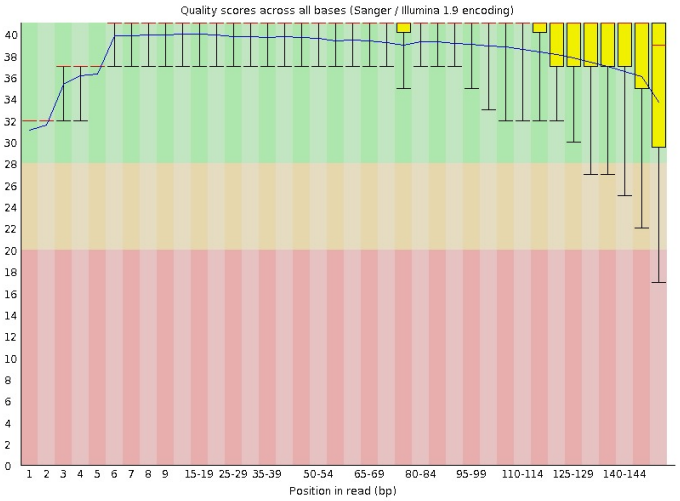


Quality Assessment of Paired Reads

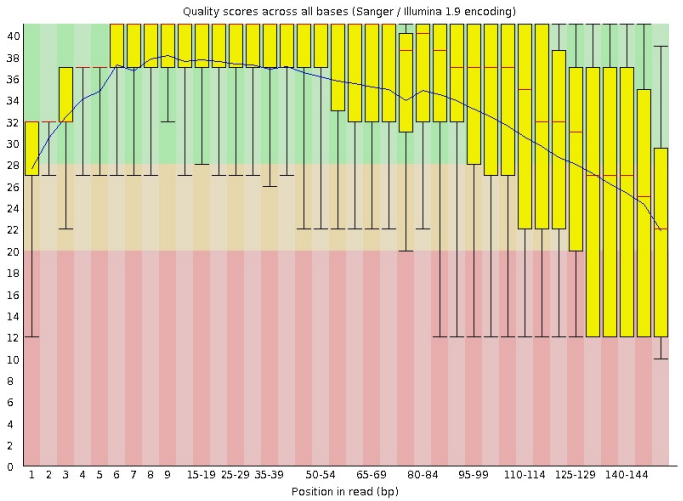
UNIVERSITÄT BERN

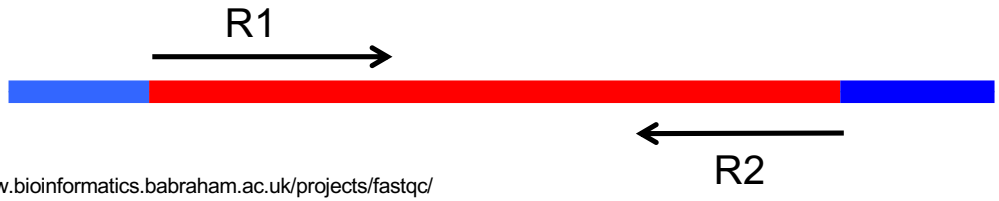
Detect biases in the sequencing & library preparation

R1



R2





Simon Andrews: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

FastA and FastQ Data Formats

UNIVERSITÄT BERN

FastA format

```
>61DFRAAXX100204:1:100:10494:3070/1 Description
AAACAACAGGGCACATTGTCACTCTTGTATTTGAAAAACACTTTCCGGCCAT
```

FastQ format

paired end information

```
@A00574:424:H7VVWDRXY:1:2101:29315:5008 1:N:0:GAATACCTAT+ACTCTATTGT
GCTTTGACATTGCTTTTGAGATAAAGTGCCTCGGTGTTCACCAAGGTGCCATCAAGATCAAAAATAATGGTAGTCG
+
FFFFFFFFFFFFFF:FFFFFFFFFFFFFFF:FFFFFFFFFF:FFFFFFFF:FFFFFFFFFFF:FF:FFFFF
```

paired end information

```
@A00574:424:H7VVWDRXY:1:2101:29315:5008 2:N:0:GAATACCTAT+ACTCTATTGT
GCCTGAGGCTGGGTATGCAGCCTTGCATGAAGGCCAATTAAAGGATTCCTGGCAAGAAGCAATGGCCGCGAGCCTT
+
FFFFFFFFF:FFFFFFFFFFFFFFFFFFFF:FFFFFFFF,F,FFF:FFFFFFFFFFFFFFFF:F,FFFFFFF
```

Phred Scores

Sanger definition and
Illumina version >= 1.3

$Q_{sanger} = -10 \log_{10}(p)$
$Q_{Illumina1.3+} = -10 \log_{10}(p)$

Illumina version < 1.3

$Q_{illumina<1.3} = -10 \log_{10}\left(\frac{p}{1-p}\right)$

$$p = 10^{\frac{-Q}{10}}$$

Phred	Error prob. p
3	50.12%
10	10.00%
15	3.16%
20	1.00%
25	0.32%
30	0.10%
35	0.03%
40	0.01%

The Phred quality score Q is an integer mapping of the probability p of the corresponding base call being incorrect

Excursus: ASCII, Computers and Chars

Computers can only handle 'ON' (1) or 'OFF' (0)

Letters are encoded by numerical values

The ASCII standard encodes letters by 7 bits → 128 chars possible

ASCII Code Chart

lower 4 bits by hexadecimal number

higher 4 bits by hexadecimal number	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL

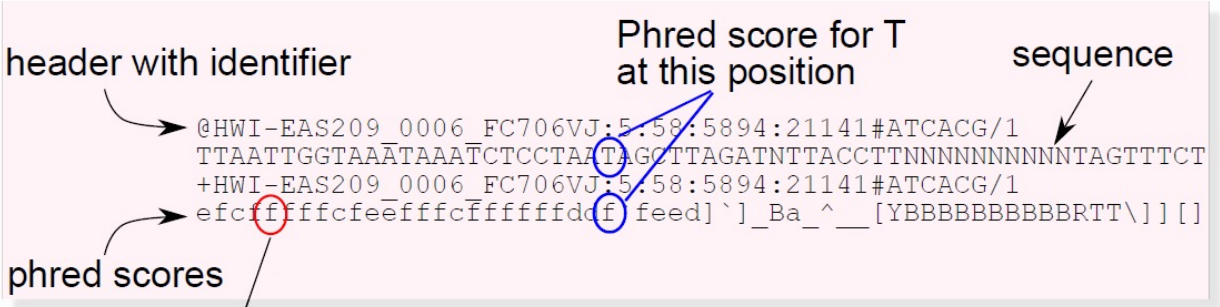
A
hex = 41
bin = 0100 0001
dec = 16*4 + 1*1 = 65

e
hex = 65
bin = 0110 1001
dec = 16*6 + 1*5 = 101

ascii = American Standard Code for Information Interchange

FASTQ Format: Encoding of Phred Scores

The Fastq format is an extension of the Fasta format
1. line : sequence identifier
2. line : sequence
3. line : sequence identifier repeated (optional)
4. line : quality scores encoded by ASCII with an offset



Phred score encoded by ASCII with offset

f → ascii(f) = 102
Q = 102 - 64 = 38
→ p_{error} = 0.016%

Sanger: Q-range 0-93, offset of 33

Illumina 1.3+: Q-range of 0-62, offset 64
0-2 special indicators eg. for read end

Illumina old: Q-range of -5-62, offset 59

Illumina (casava 1.8+): Q-range 0-93, offset 33

Data Freeze | Versioning

UNIVERSITÄT BERN

- genome assemblies are almost never finished
- genome annotations are never complete

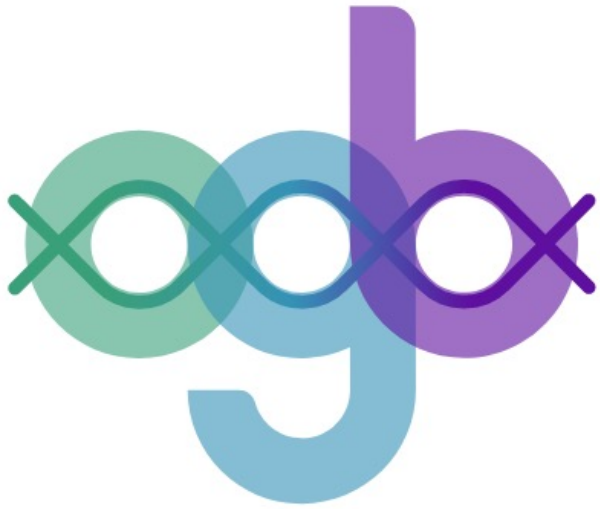
→ data freeze is important

→ assemblies and annotation are versioned

→ a particular version is static and does not change

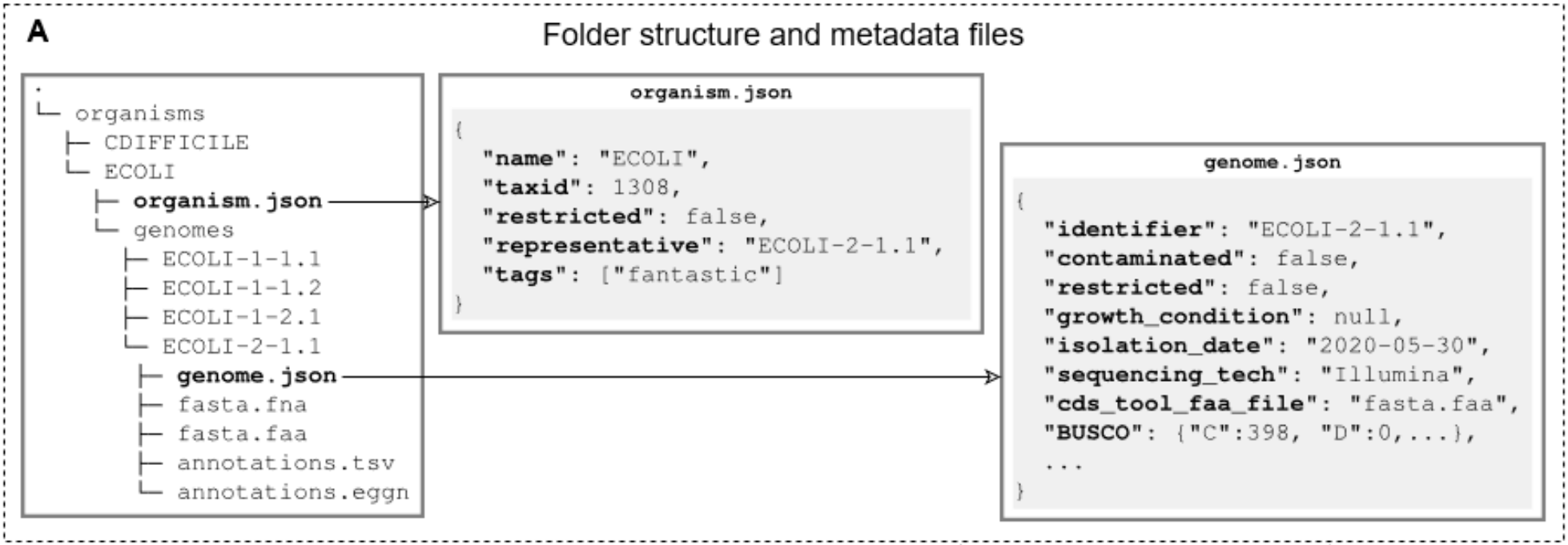
OpenGenomeBrowser

- We developed a genome browser to organize and analyze thousands of bacterial genomes
- Concept of versioning is crucial
- possible to work with different versions



OpenGenomeBrowser

Versioning



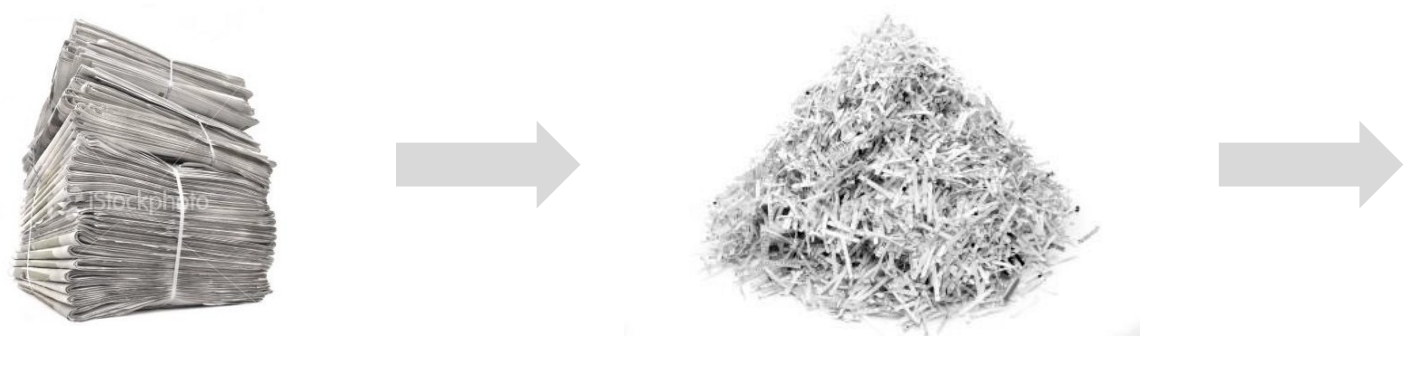
https://opengenomebrowser.bioinformatics.unibe.ch/

Now it is your turn – Your Project in this Course

UNIVERSITÄT BERN

- Perform all steps of a de novo genome assembly by yourself
- Start are subsampled read data which are from a Nature Communications publication
  - Nat Commun. 2020 Feb 20;11(1):989. doi: 10.1038/s41467-020-14779-y.
- Each student gets their own data set (PacBio reads)
- Each student will write an own report.
  - report has two parts: (i) assembly & (ii) annotation
    (annotation will be performed in the next course by Christian Parisod (6 weeks)
- At the end of this course (last day), each student presents the own work (15 minutes + discussion)

De Novo Genome Assembly



"From scratch"