Introduction

Isheng Jason Tsai

Introduction to NGS Data and Analysis
Lecture 1 ; v2020

Welcome!

This is the first lecture for Introduction to NGS Data and Analysis and “NGS analytics” lecture in TIGP B2 course.

This course is called

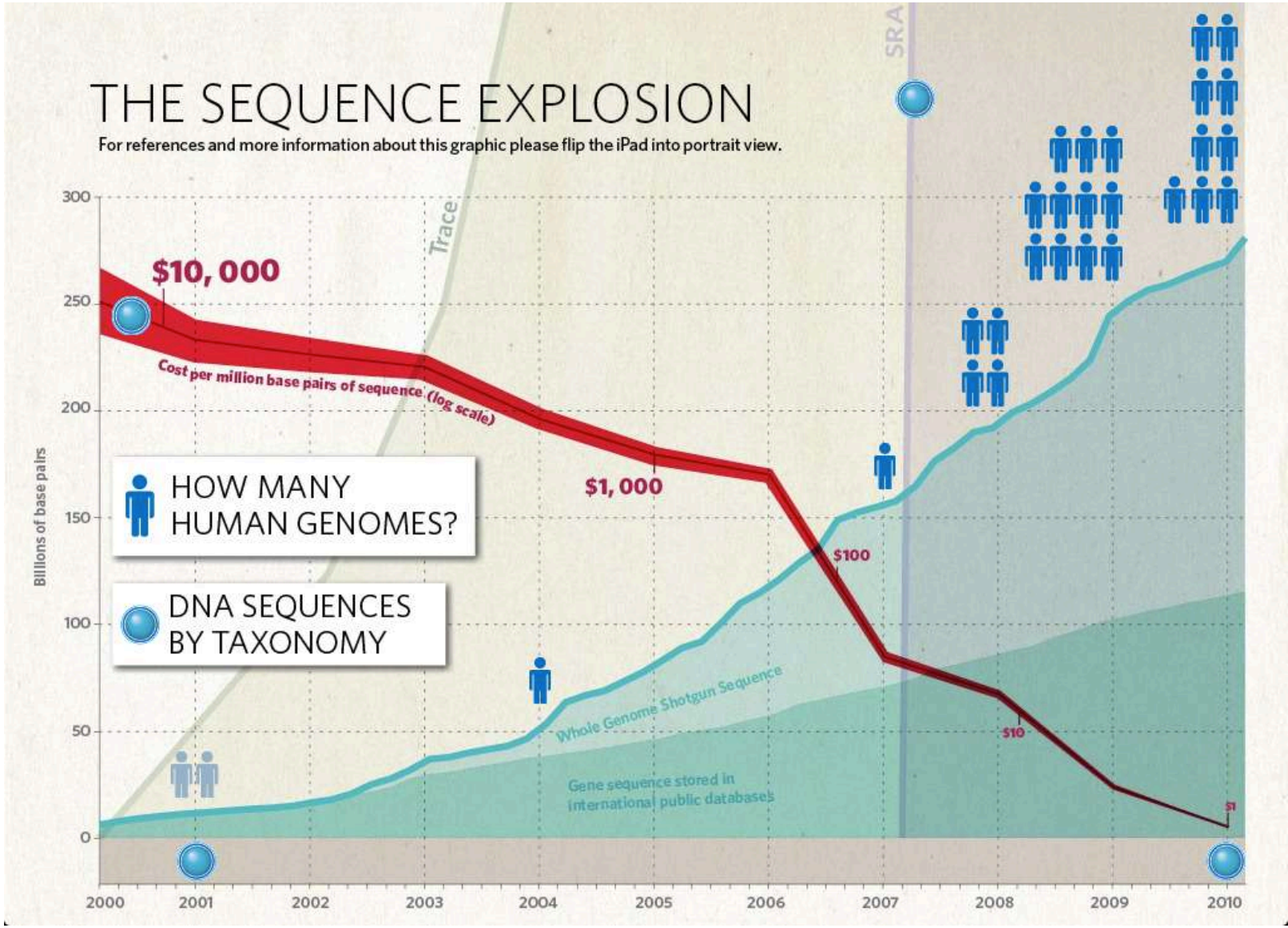
“Introduction to Next-Generation Sequencing (NGS) Data and Analysis”

Actually

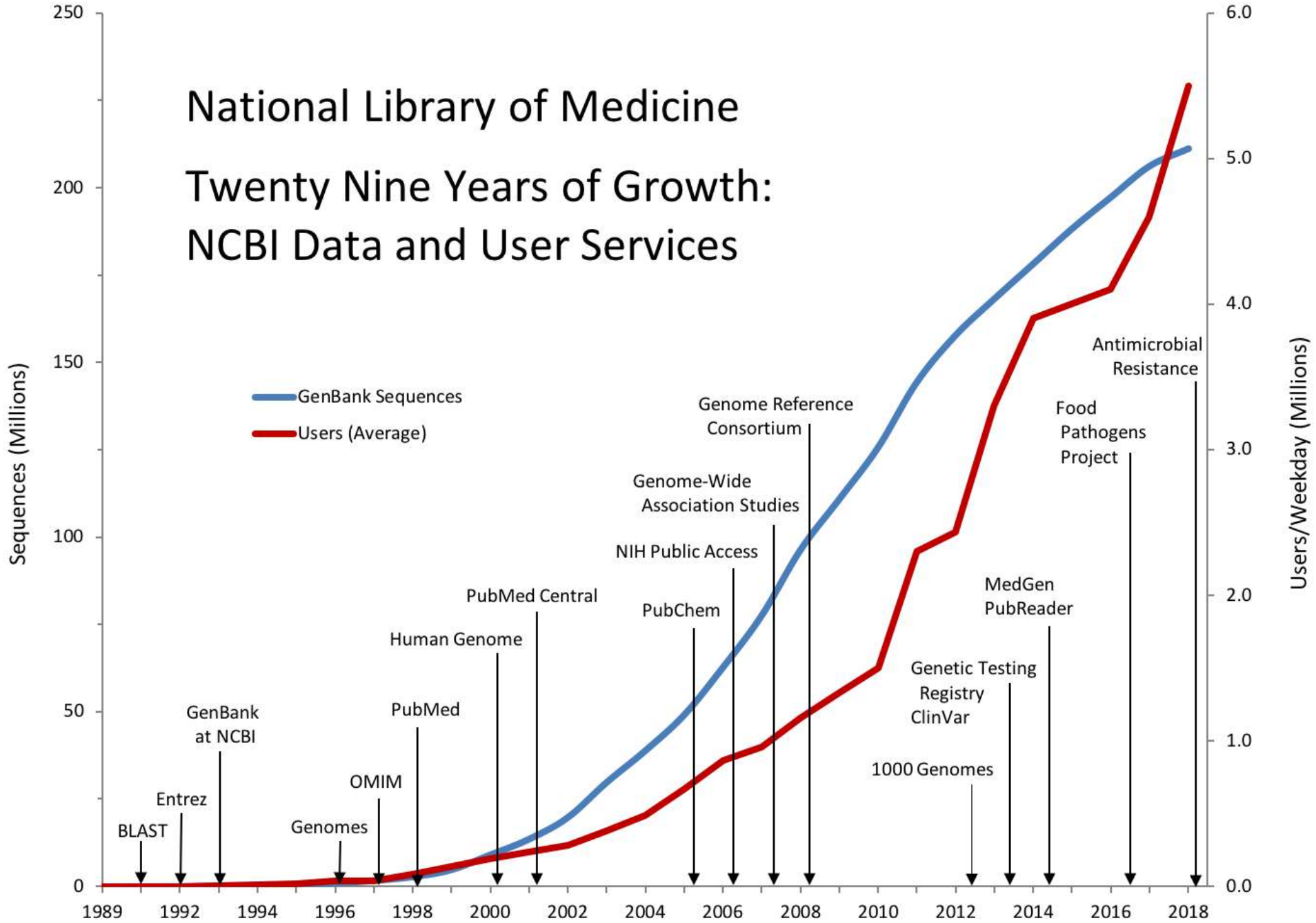
- I tried to change the course name to “**Introduction to Genomics sequencing: Data and Analysis**” but was too late
- Next Generation Sequencing is really “now” sequencing
- It won’t be so easy to tell you everything about NGS
  (it’s a bit like saying what can we do with PCR?)

What is NGS?

= Next generation sequencing,
= deep sequencing
= High Throughput Sequencing,
= Massively parallel sequencing
= 次世代定序
= 高速高量定序



http://www.nature.com/news/2010/100331/full/464670a.html



https://www.nlm.nih.gov/about/2020CJ.html

NGS = sequencing made cheaper, faster and higher throughput

Lecture outline

- 1. Course setup
- 2. History of bioinformatics / computational biology
- 3. History of sequencing and dawn of NGS
- 4. Advances in sequencing
- 5. Case studies
- 6. My journey

Course setup

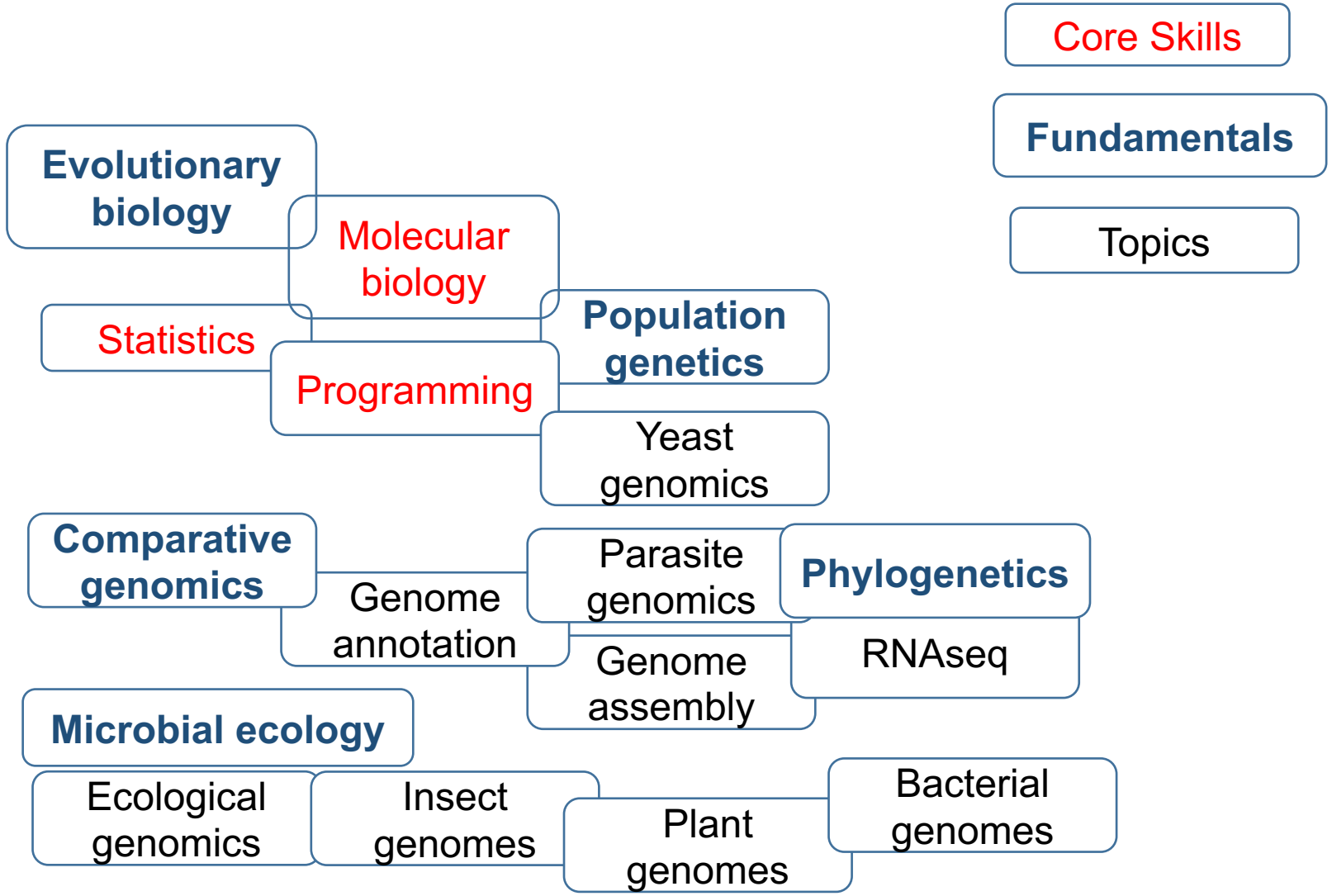
My background

Undergraduate:
Biochemistry and Genetics

2005-08 ; MSc & PhD:
Bioinformatics & Population genetics

2009-14 ; Postdoc:
Genomics & parasitology

2015 - ; Academia Sinica:
Microbial diversity & Bioinformatics



The primary goal of this course is for students to get a solid* background on the genomics discipline

* Bits I have experiences with and am comfortable in teaching

* To equip students with the necessary understanding to independently research further about genomic methodologies

This course is not designed for

- faint-hearted
- students who want to learn teach you the core skills such as programming languages
  (although some limited exercises will be run)
- human genomics;
  but more ecological context in general

For more human genomics oriented course:

國立台灣大學 National Taiwan University

次世代定序、生物資訊學與基因體醫學

陳倩瑜/許書睿/陳沛隆老師合授

陳沛隆 老師，歡迎進入 Ceiba 系統

課程資訊
教師資訊
公佈欄
課程內容
作業區
登出

語言 Language:
中文(Chinese)

更新：2018-11-02

訪客：914

課程資

友善列印

基本資訊				
課程名稱	次世代定序、生物資訊學與基因體醫學			
開課學期	108-2			
授課對象	基因體暨蛋白體醫學研究所	基因體暨蛋白體醫學研究所	基因體暨蛋白體醫學研究所	分子醫學學分學程
課號	Genom 7009	455 M0090		
班次				
上課時間	星期二 2,3,4			
上課地點	基醫509			
備註欄	本課程中文授課,使用英文教科書。不限定特定先修課程，但需具有主動以及跨領域學習之態度。與陳倩瑜共授	本課程中文授課,使用英文教科書。不限定特定先修課程，但需具有主動以及跨領域學習之態度。與陳倩瑜共授	本課程中文授課,使用英文教科書。不限定特定先修課程，但需具有主動以及跨領域學習之態度。與陳倩瑜共授	本課程中文授課,使用英文教科書。不限定特定先修課程，但需具有主動以及跨領域學習之態度。與陳倩瑜共授
課程網址				

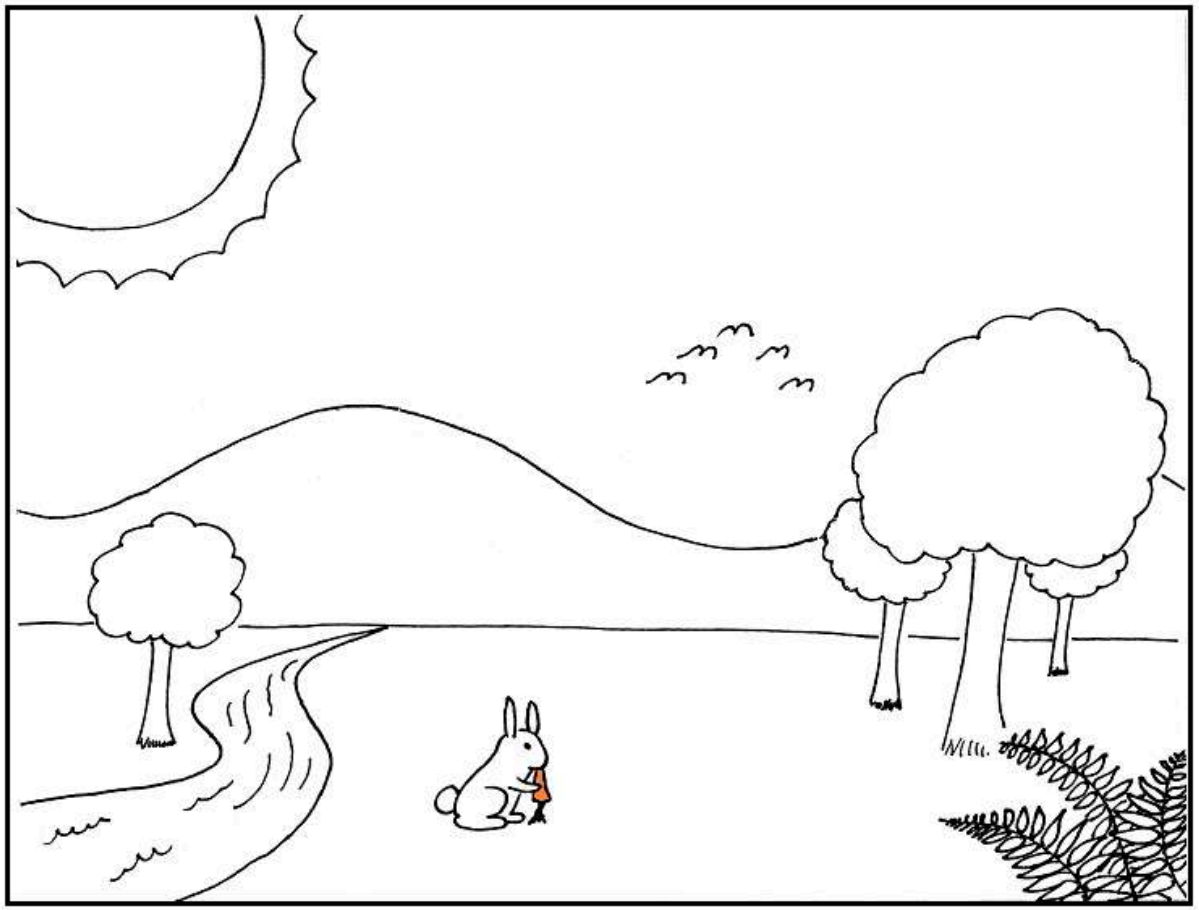
Layout of this course

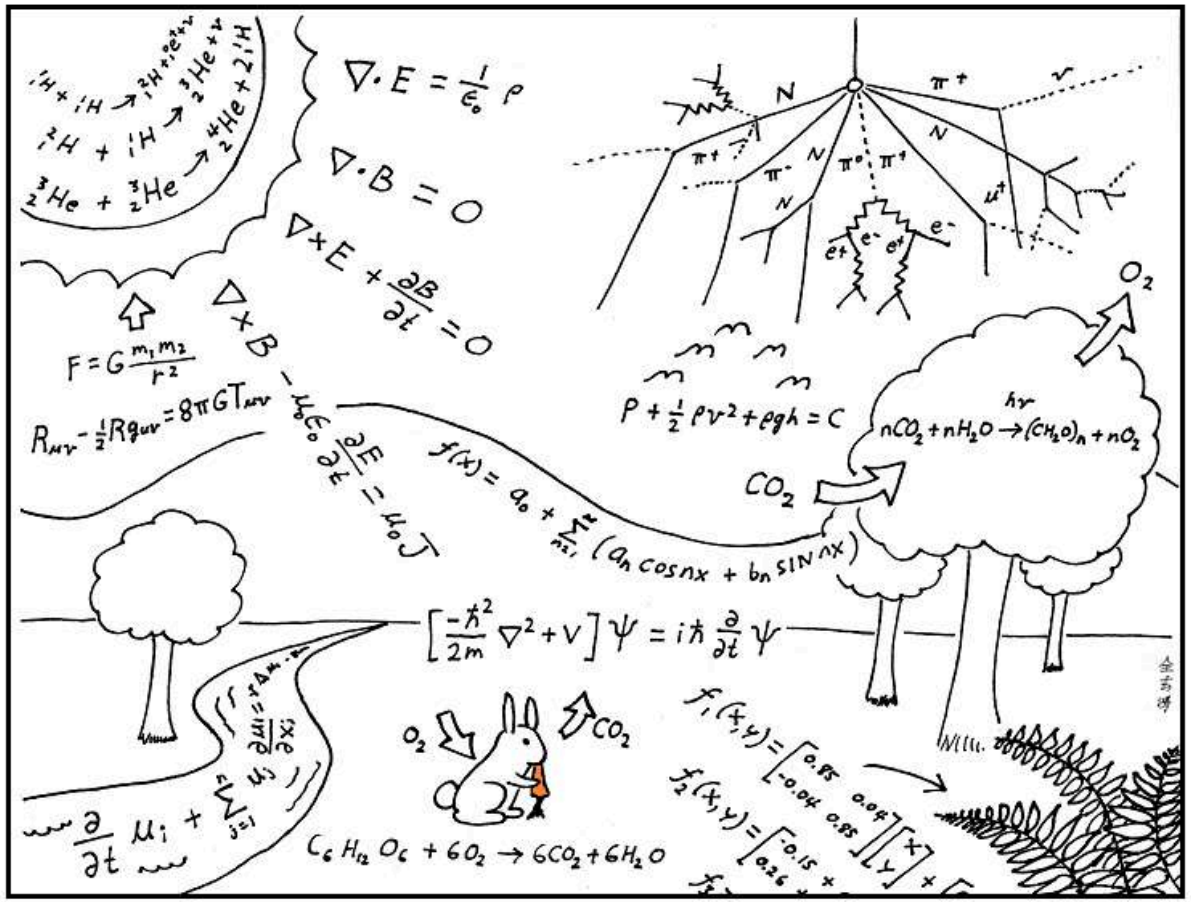
Week	Date	Topic
Week 1	3/4	Introductory lecture (Jason)
Week 2	3/11	Linux and R; basic usage (Jason)
Week 3	3/18	*Practical I: Statistics in R (Jason)
Week 4	3/25	Mapping and Case studies (Jason)
Week 5	4/1	Genome Assembly (Jason)
Week 6	4/8	Comparative Genomics (Jason)
Week 7	4/15	Transcriptomes (Jason)
Week 8	4/22	Alignment to phylogenies (Professor Jia-Ming Chang)
Week 9	4/29	Amplicon / Metagenomic (Jason)
Week 10	5/6	Population Genomics (Dr. John Wang)
Week 11	5/13	*Study week (no class; Protocol assignment due)
Week 12	5/20	*Midterm exam (Students)
Week 13	5/27	*Practical II: Nanopore sequencing (Dr. Huei-Mien Ke and Dr. Tom Lin)
Week 14	6/3	*Practical III: Mapping and Phylogeny in R (Jason)
Week 15	6/10	*Practical IV: RNAseq analysis in R (Jason)
Week 16	6/17	Experiences in NGS library preparation and construction (Dr. Meiyeh Lu)
Week 17	6/24	*Study week (no class; R assignment due)
Week 18	7/1	*Final presentation (Students)

Enrolled students only

Always start with a question

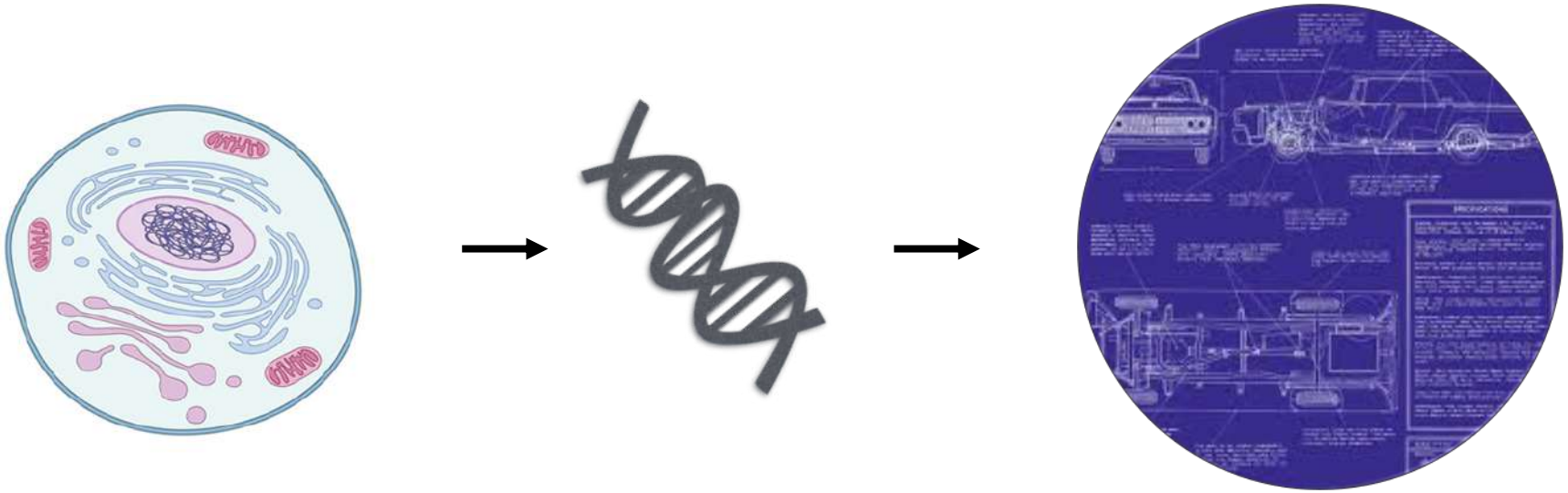
This is how scientists see the world





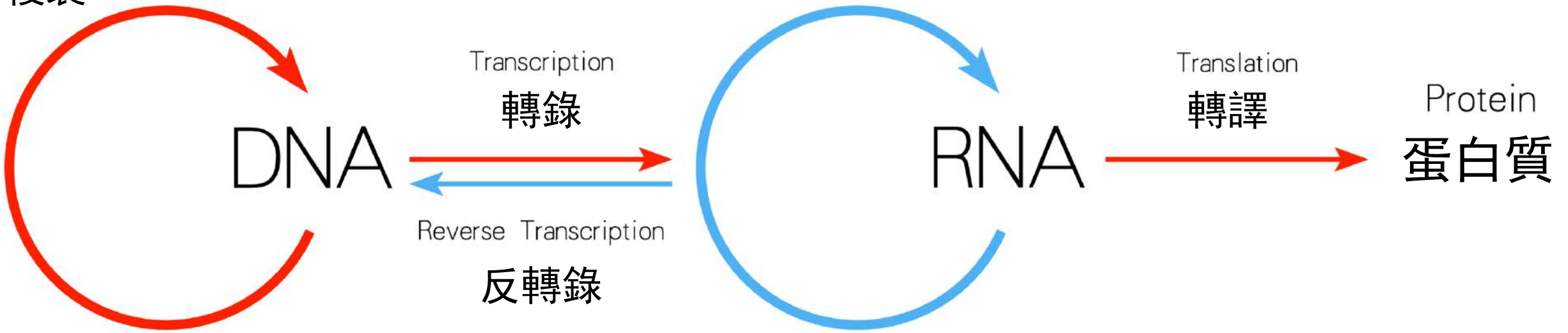
How? Who? Where? What?

Genome



Genome = Parts list of a single genome

複製



基因 (gene) : 一個有功能的DNA 片段
coding DNA: 可以轉譯成蛋白質的 DNA　　#noncoding
基因體 (genome): 物種一個細胞核內所有的DNA
定序 (sequencing) : 解析出DNA 序列 [ATCGTGACGTGACGTAC…]

Nowadays we usually call people who analyse sequence data, or lots of biological data -> bioinformaticians.

But what is bioinformatics?
Or Computational biology?

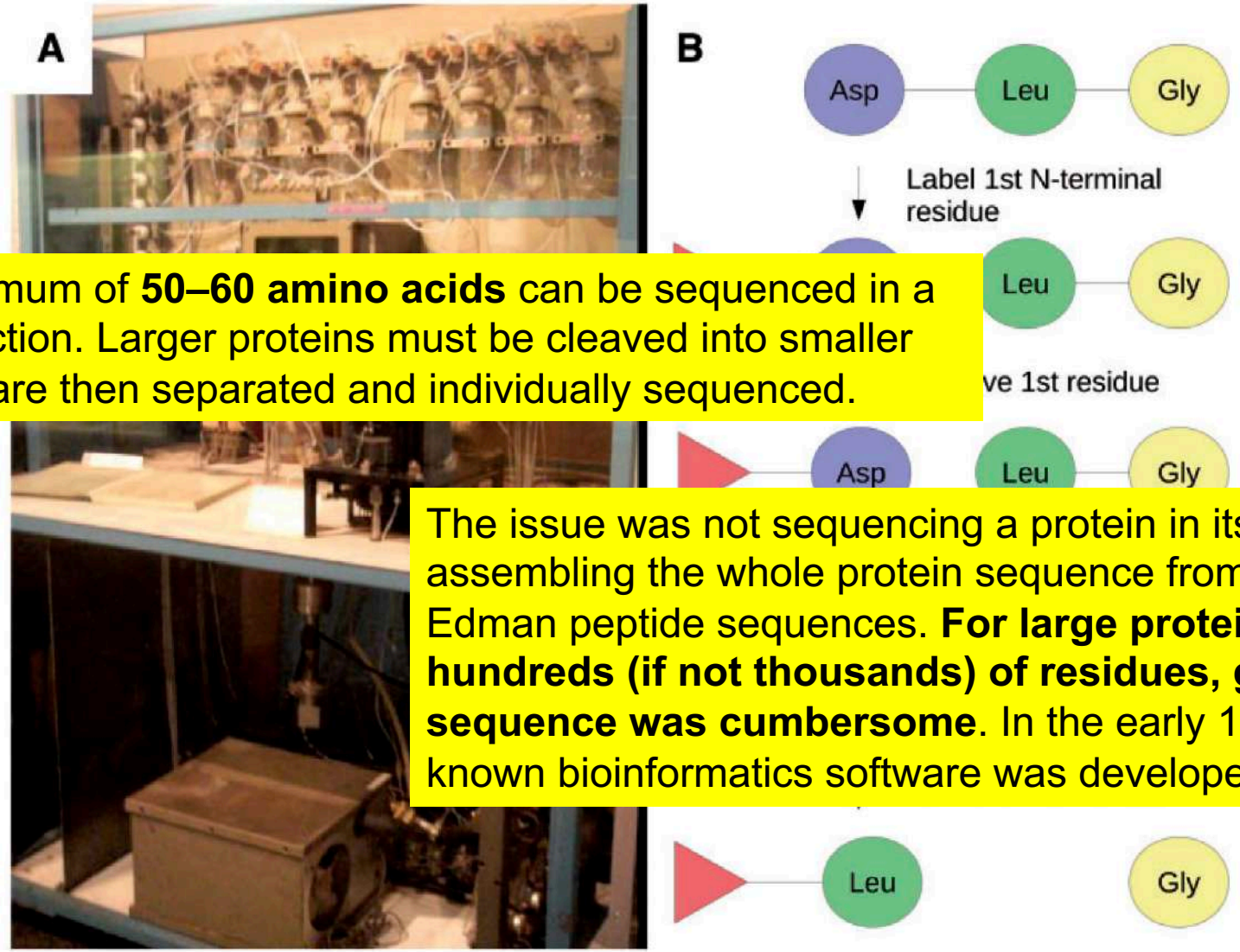
- the very beginnings of bioinformatics occurred **more than 50 years ago**, when desktop computers were still a hypothesis and DNA could not yet be sequenced."

- The foundations of bioinformatics were laid in **the early 1960s** the application of computational methods to protein sequence analysis (notably, *de novo* sequence assembly, biological sequence databases and substitution models).

- Later on, DNA analysis also emerged due to parallel advances in (i) molecular biology methods, which allowed easier manipulation of DNA, as well as its sequencing, and (ii) computer science, which saw the rise of increasingly miniaturized and more powerful computers, as well as novel software better suited to handle bioinformatics tasks. **In the 1990s through the 2000s, major improvements in sequencing technology, along with reduced costs, gave rise to an exponential increase of data**.

- The arrival of 'Big Data' has laid out **new challenges in terms of data mining and management**, calling for more expertise from computer science into the field.

A brief history of bioinformatics
Jeff Gauthier, Antony T Vincent, Steve J Charette, Nicolas Derome
Briefings in Bioinformatics (2018) https://doi.org/10.1093/bib/bby063



A theoretical maximum of **50–60 amino acids** can be sequenced in a single Edman reaction. Larger proteins must be cleaved into smaller fragments, which are then separated and individually sequenced.

The issue was not sequencing a protein in itself but rather assembling the whole protein sequence from hundreds of small Edman peptide sequences. **For large proteins made of several hundreds (if not thousands) of residues, getting back the final sequence was cumbersome**. In the early 1960s, one of the first known bioinformatics software was developed to solve this problem.

Figure 1. Automated Edman peptide sequencing. (**A**) One of the first automated peptide sequencers, designed by William J. Dreyer. (**B**) Edman sequencing: the first N-terminal amino acid of a peptide chain is labeled with phenylisothiocyanate (PITC, red triangle), and then cleaved by lowering the pH. By repeating this process, one can determine a peptide sequence, one N-terminal amino acid at a time.

A brief history of bioinformatics
Jeff Gauthier, Antony T Vincent, Steve J Charette, Nicolas Derome
Briefings in Bioinformatics (2018) https://doi.org/10.1093/bib/bby063

Dayhoff: the first bioinformatician

Margaret Dayhoff (1925-1983)

- Designed one letter amino acid code
- Trained in quantum chemistry and mathematics, she became interested in proteins and molecular evolution around 1960.
- to explore mathematical approaches for analysing amino-acid sequence data
- Her initial project was writing a series of FORTRAN programs to determine the amino-acid sequences of protein molecules.

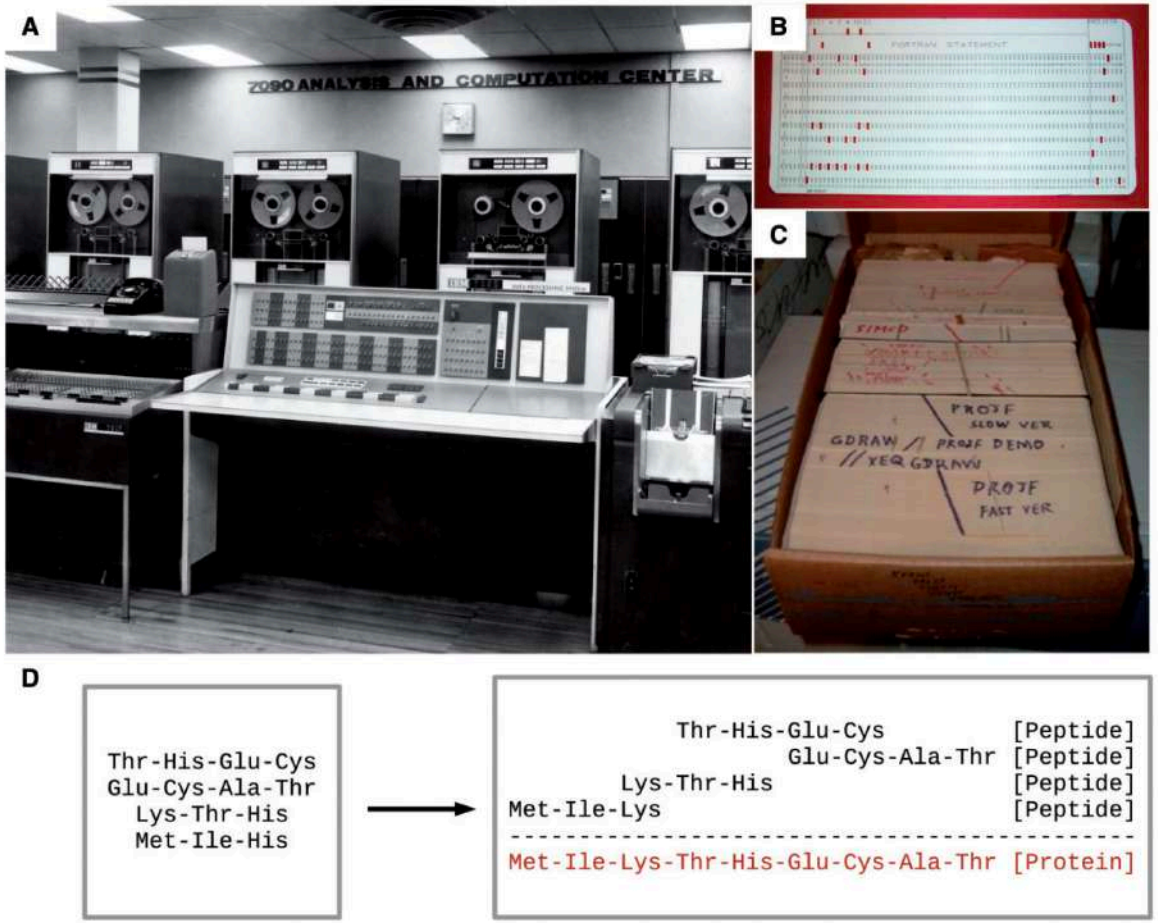


Figure 2. COMPROTEIN, the first bioinformatics software. (A) An IBM 7090 mainframe, for which COMPROTEIN was made to run. (B) A punch card containing one line of FORTRAN code (the language COMPROTEIN was written with). (C) An entire program's source code in punch cards. (D) A simplified overview of COMPROTEIN's input (i.e. Edman peptide sequences) and output (a consensus protein sequence).

A brief history of bioinformatics
Jeff Gauthier, Antony T Vincent, Steve J Charette, Nicolas Derome
Briefings in Bioinformatics (2018) https://doi.org/10.1093/bib/bby063

Hagen (2000) Nat Rev Genetics doi: 10.1038/35042090

Ancestral sequences and Molecular clock (Emile Zuckerkandl and Linus Pauling)

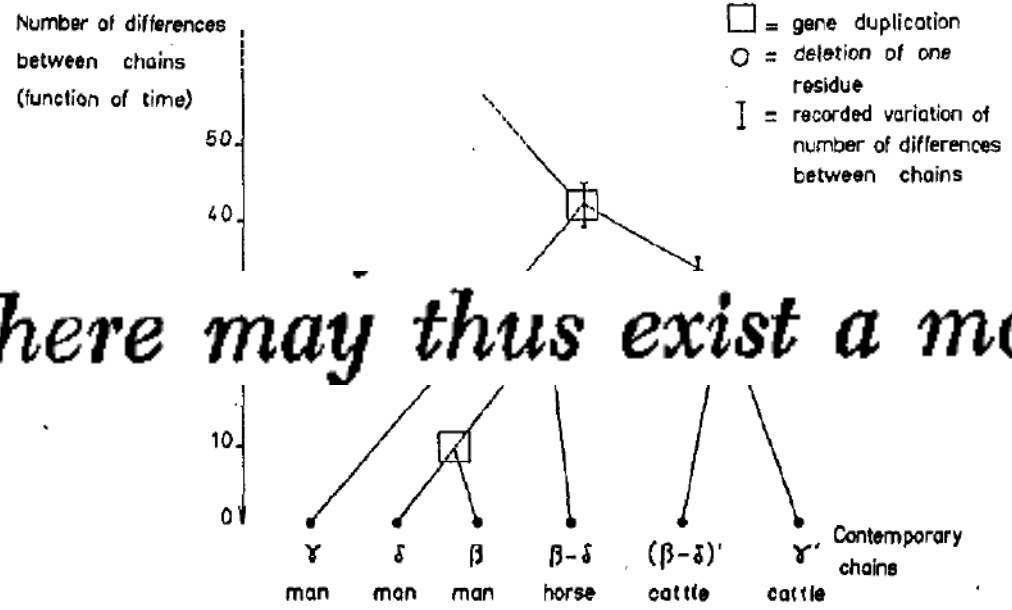
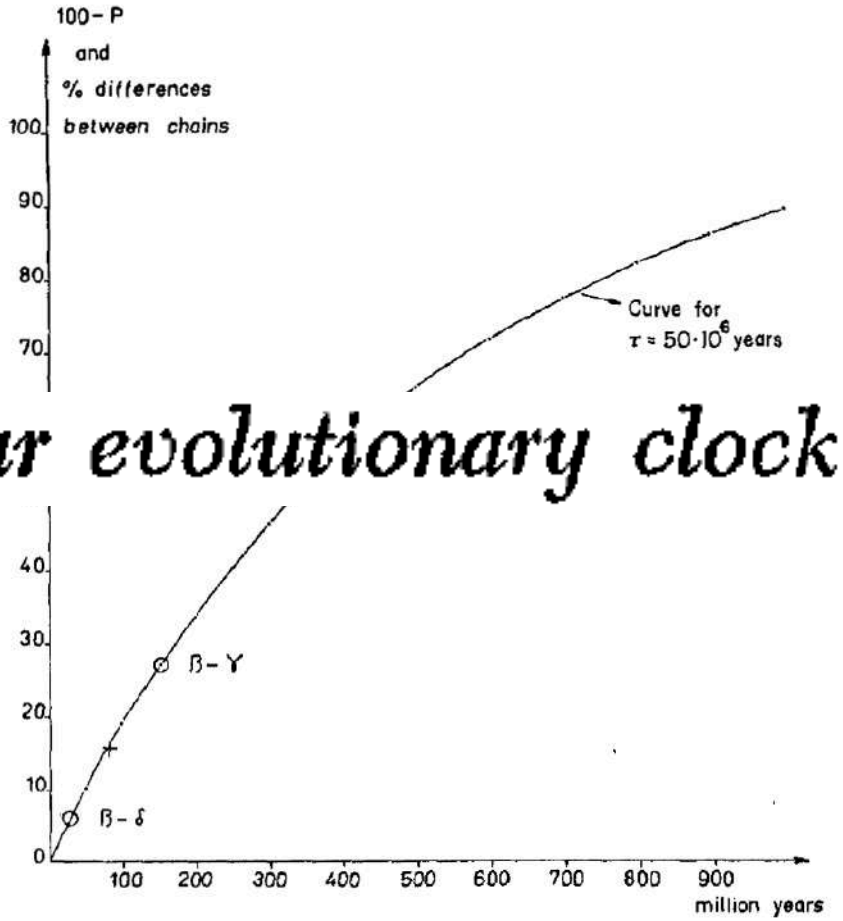


FIG. 4. Probable evolutionary relationship of some mammalian hemoglobin chains.

There may thus exist a molecular evolutionary clock.[1]

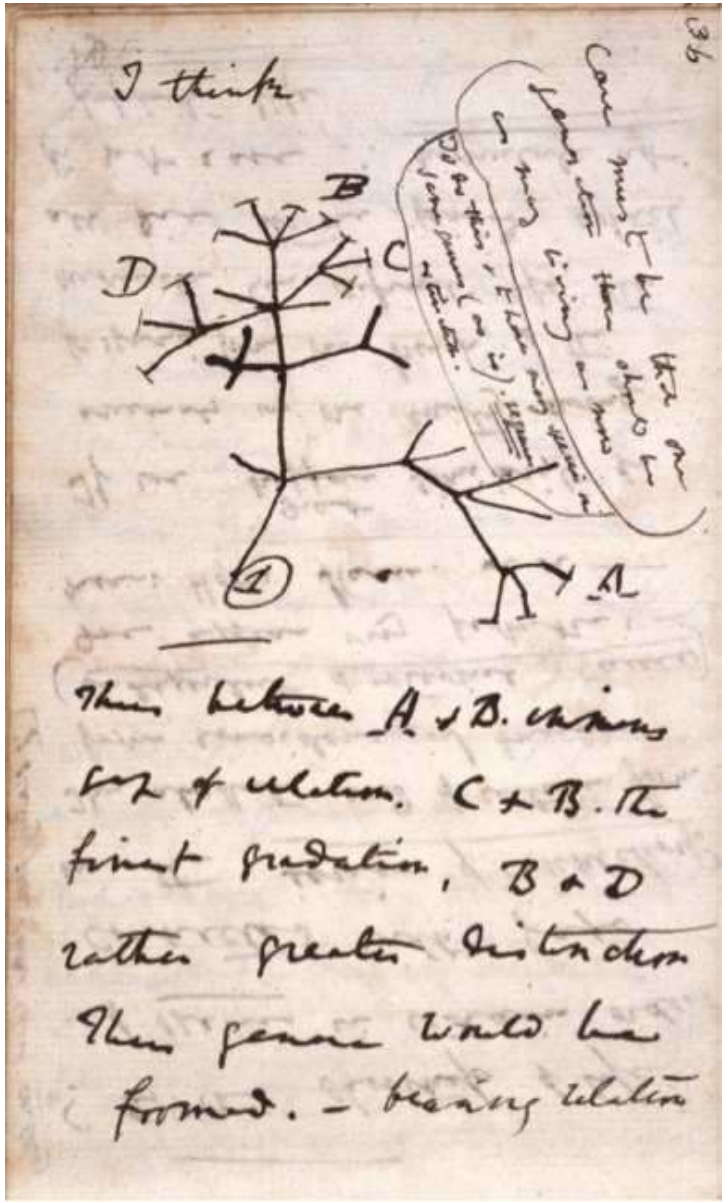


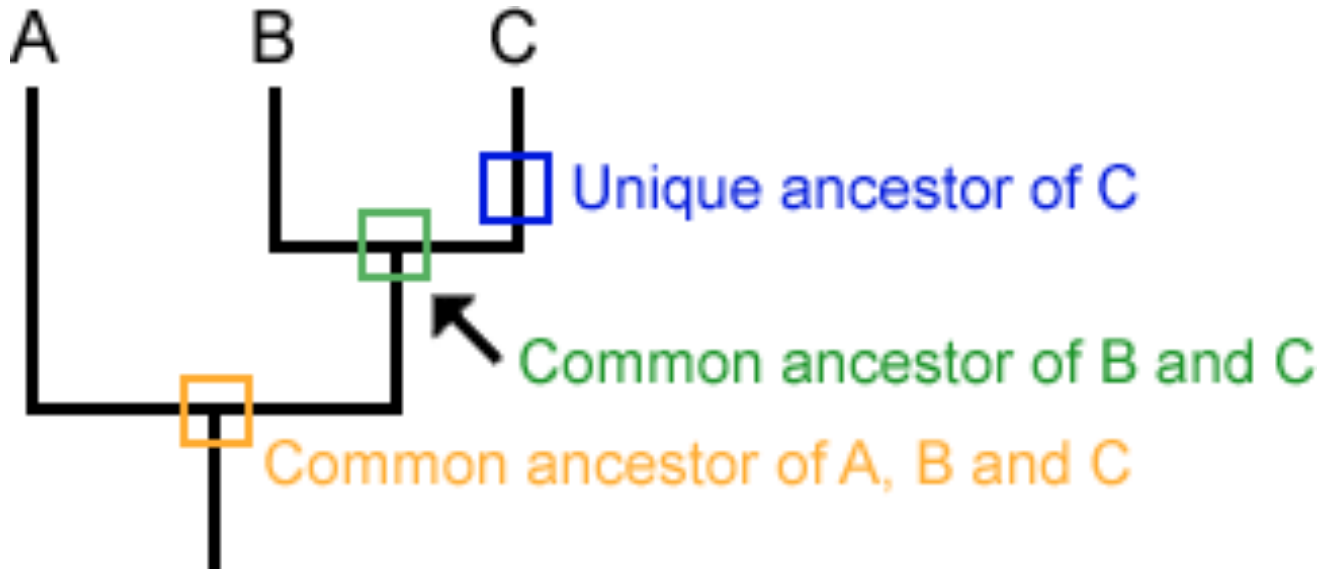
"Zuckerkandl and Pauling hypothesized that orthologous proteins evolved through divergence from a common ancestor. Consequently, by comparing the sequence of hemoglobin in currently extant organisms, it became possible to predict the 'ancestral sequences' of hemoglobin and, in the process, its evolutionary history up to its current forms"

Evolutionary divergence and convergence in proteins
Zuckerkandl, E. and Pauling, L (1965)

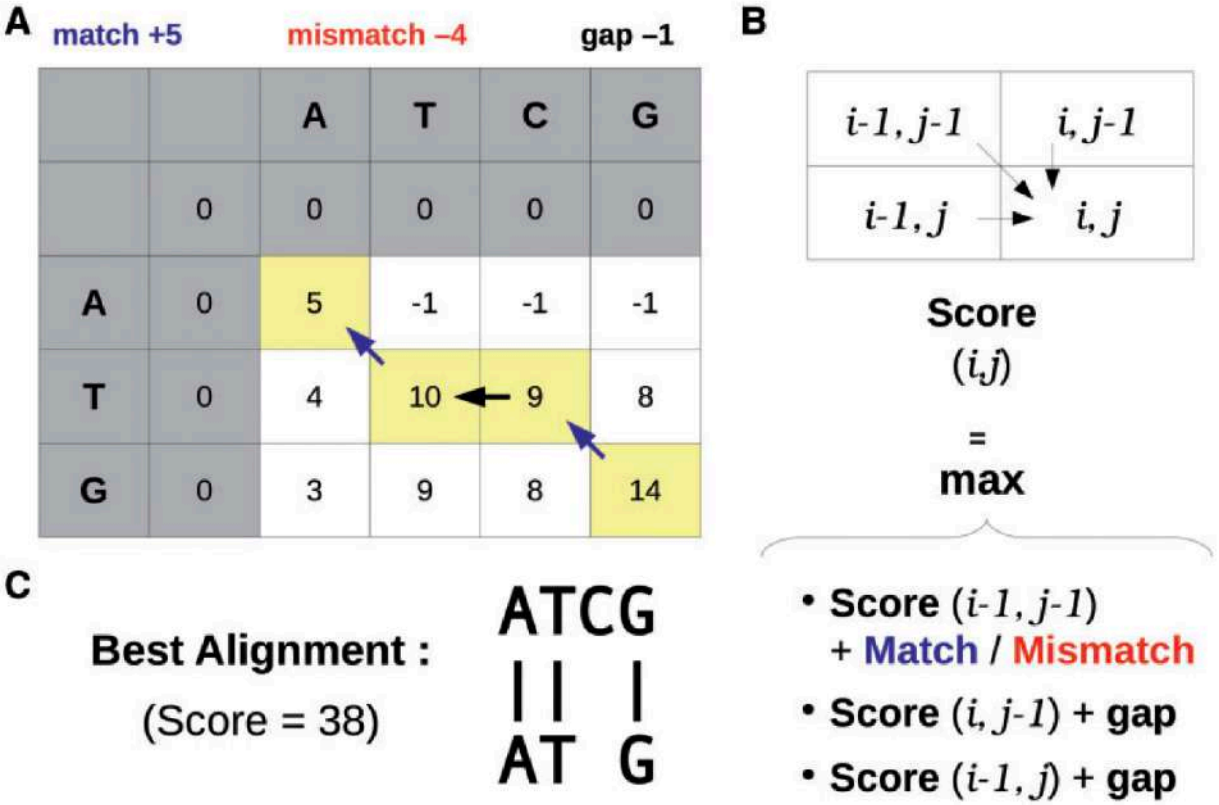
A brief history of bioinformatics
Jeff Gauthier, Antony T Vincent, Steve J Charette, Nicolas Derome
Briefings in Bioinformatics (2018) https://doi.org/10.1093/bib/bby063

Relationships between sequences recapitulate evolutionary relationships





A mathematical framework for sequence alignments



		A	T	C	G
	0	0	0	0	0
A	0	5	-1	-1	-1
T	0	4	10	9	8
G	0	3	9	8	14

Table 1. An excerpt of the PAM1 amino acid substitution matrix

10^4 P[a]		Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	... ...	Val V
Ala	A	9867	2	9	10	3	8	...	18
Arg	R	1	9913	1	0	1	10	...	1
Asn	N	4	1	9822	36	0	4	...	1
Asp	D	6	0	42	9859	0	6	...	1
Cys	C	1	1	0	0	9973	0	...	2
Gln	Q	3	9	4	5	0	9876	...	1
...	...	...	...	...	...	...	...	...	...
Val	V	13	2	1	1	3	2	...	9901

[a]Each numeric value represents the probability that an amino acid from the *i*-th column be substituted by an amino acid in the *j*-th row (multiplied by 10 000).

A brief history of bioinformatics
Jeff Gauthier, Antony T Vincent, Steve J Charette, Nicolas Derome
Briefings in Bioinformatics (2018) https://doi.org/10.1093/bib/bby063

1970-2000s – Paradigm shifts and parallel advances in biology and computer science

- Protein sequencing to DNA sequencing (faster / cheaper)
- Use DNA sequences to infer phylogenetic trees
- Sequence of marker genes and genomes
- Beyond sequences (structural bioinformatics)

- Faster computers
- GPUs
- Free software movement
- New Programming languages (Perl created by Larry Wall in 1987)

- Internet
- Online databases (NCBIs)

A brief history of bioinformatics
Jeff Gauthier, Antony T Vincent, Steve J Charette, Nicolas Derome
Briefings in Bioinformatics (2018) https://doi.org/10.1093/bib/bby063

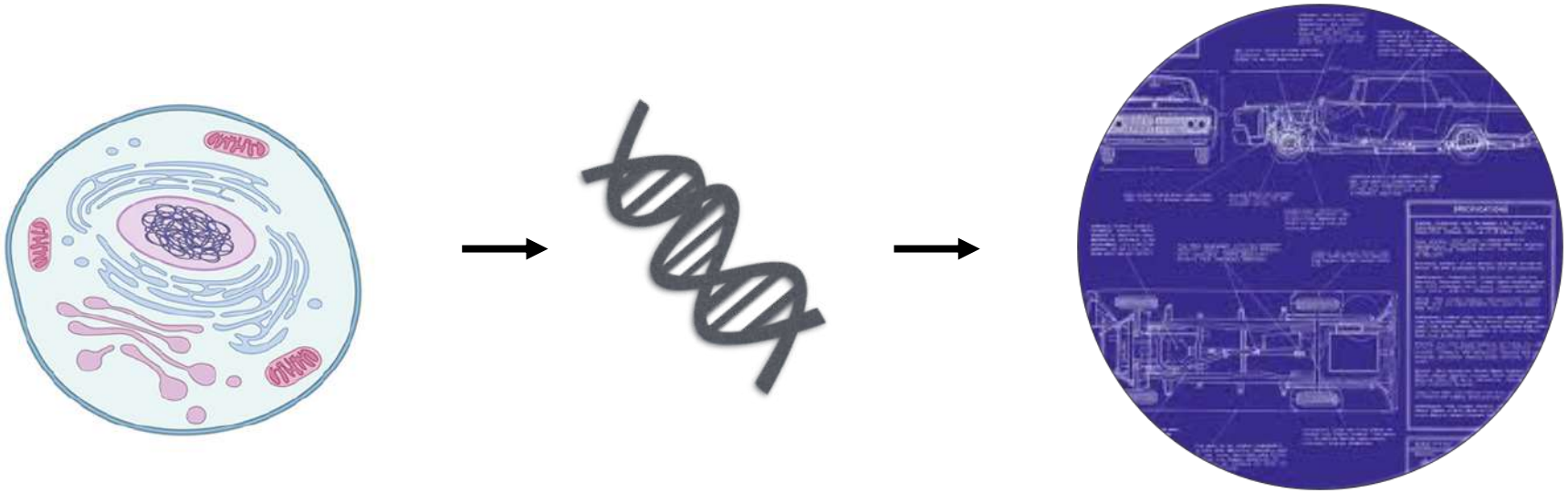
Summary (I)

- In order to analyse biological (sequence) data, you need to know:
  - how data were generated (experiments)
  - Organise lots of data (informatics / coding) *
  - Analyse (statistics / algorithm development) *
  - Interpretation (statistics / evolution / genetics) *

* It is difficult to generalize but hopefully this makes things easier
* Traditionally people who don't fall into the experimental category go here

A brief history of bioinformatics
Jeff Gauthier, Antony T Vincent, Steve J Charette, Nicolas Derome
Briefings in Bioinformatics (2018) https://doi.org/10.1093/bib/bby063

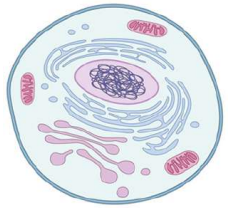
Genome



Genome = Parts list of a single genome

A genome project

Wet lab work (Lecture)



Extract DNA (or RNA)

Shear DNA (or RNA)

Sequence

Bioinformatics

Data QC

Variants

ATCG
ATGG
ATCG

Annotation (Lecture)

DNA or RNA Reads
50-500 bp

Mapping (Lecture 4)
RNAseq (Lecture 8)

Reads
50-500 bp

Assembly (Lecture)

Contigs
1kb – 100 kbp

Scaffolding

Scaffolds
Hopefully Mbp

Four situations you are most likely to encounter

Genome reference is available (for example, humans):

- Re-sequence (DNA, RNA)
- **Map** (align) sequence to the genome

Genome reference is NOT available

- **Assemble** the reads to get the genome

Counting:

- For a given region (gene) we want to know how much.→ gene expression or metagenomics

- **Statistics**

Many perceptions of NGS / genomics

What my parents think I do

What less friendly colleagues think we do

What more friendly colleagues think we do

What my friends think I do

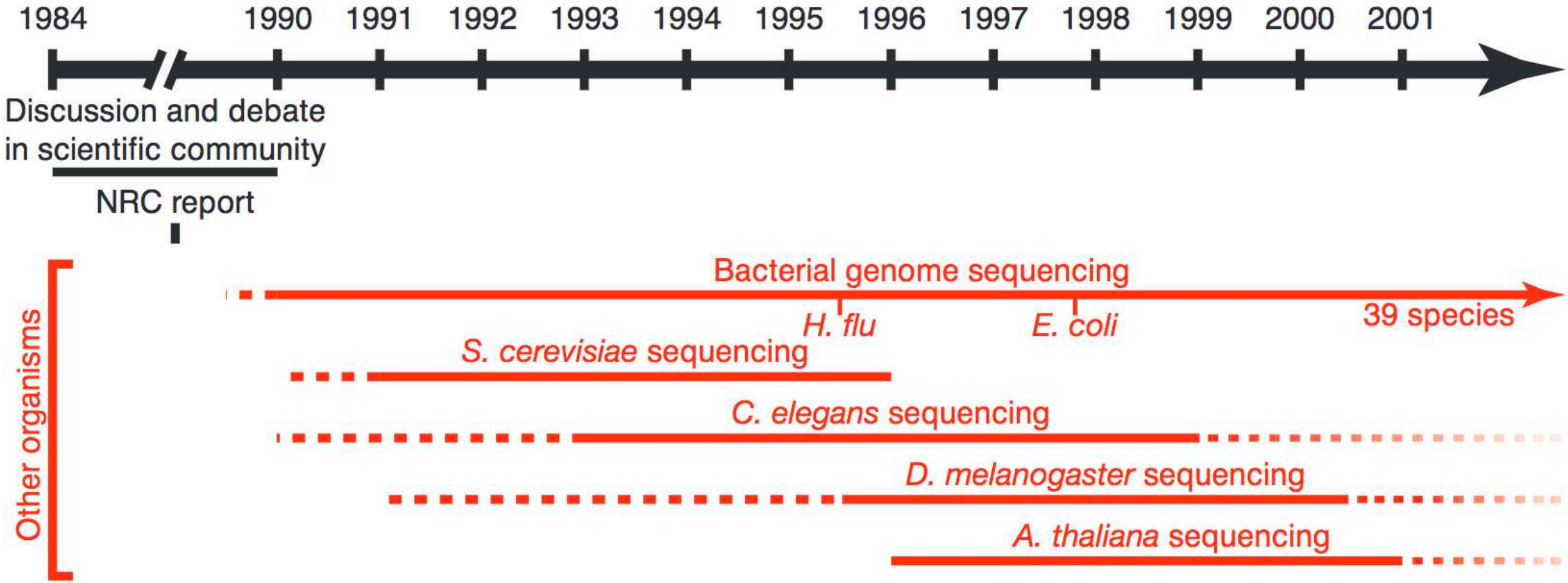
What we think we do

What we actually do

Why sequence a genome?

- Phylogenetic position
- Differences between species (comparative genomics)
- Variations between individuals (population genetics)
- Help to understand biology
- Of economic, agricultural, medical, ecology values

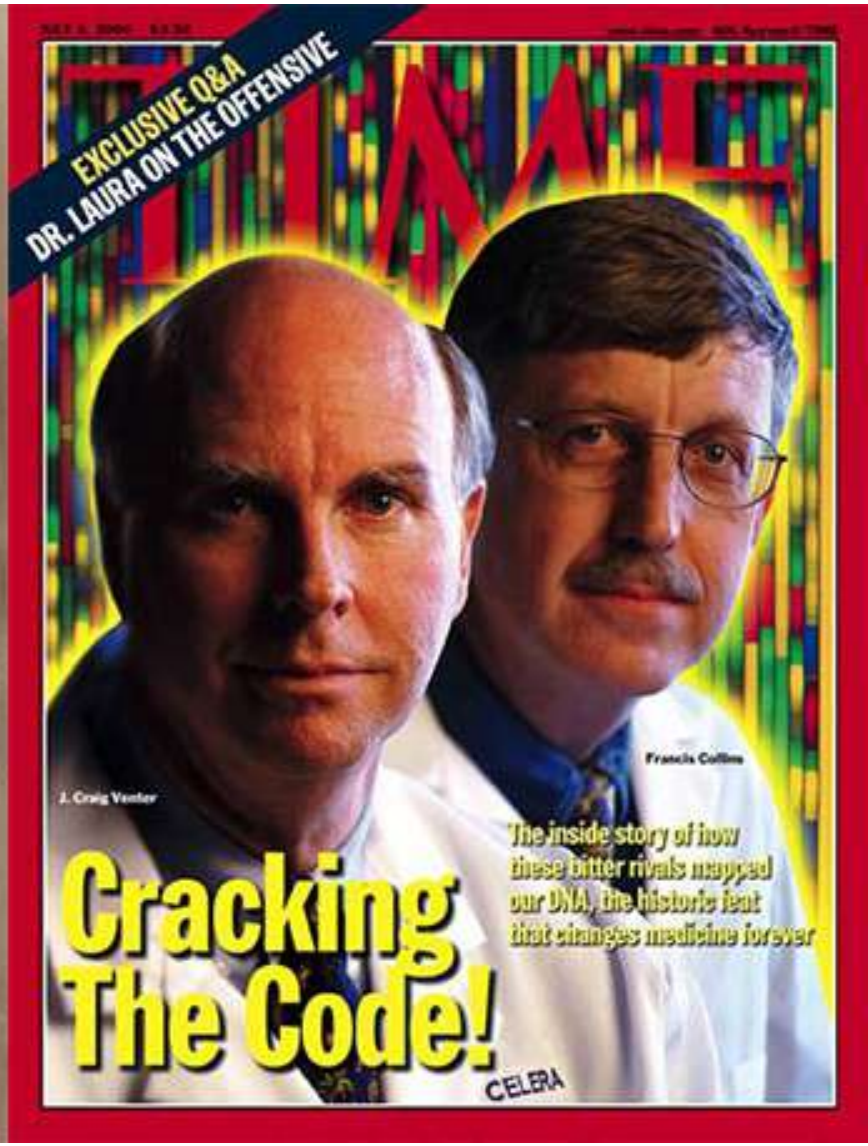
- **Help to understand biology**
- ~~Some lab just had the money~~ ; don’t do it



Nature **409**, 860-921(15 February 2001)
doi:10.1038/35057062


nature
the human genome


Science
THE HUMAN GENOME
AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE

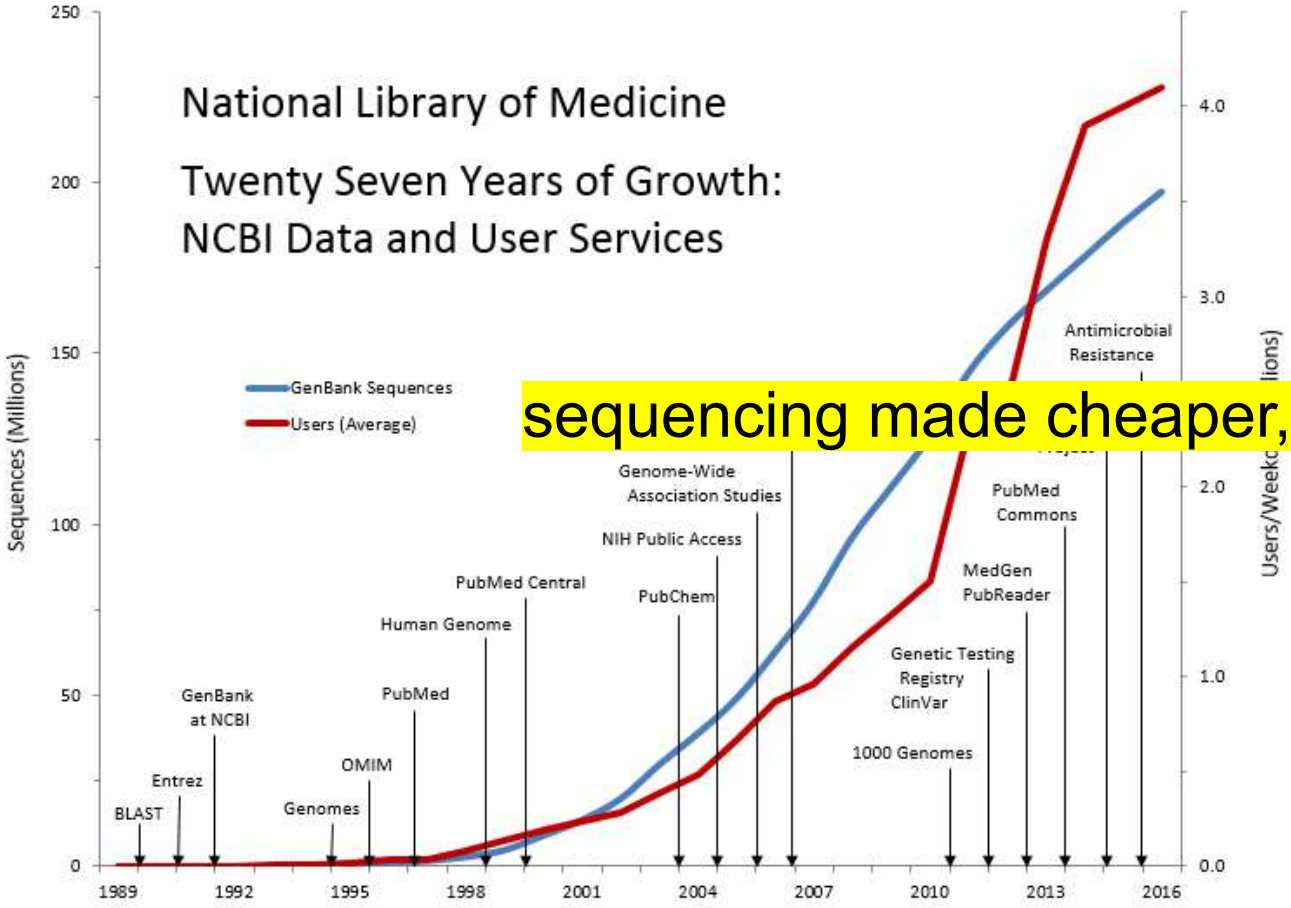

EXCLUSIVE Q&A
DR. LAURA ON THE OFFENSIVE
J. Craig Venter
Francis Collins
The inside story of how these bitter rivals mapped our DNA, the historic feat that changes medicine forever
Cracking The Code!
CELERA

Calculating the economic impact of the Human Genome Project

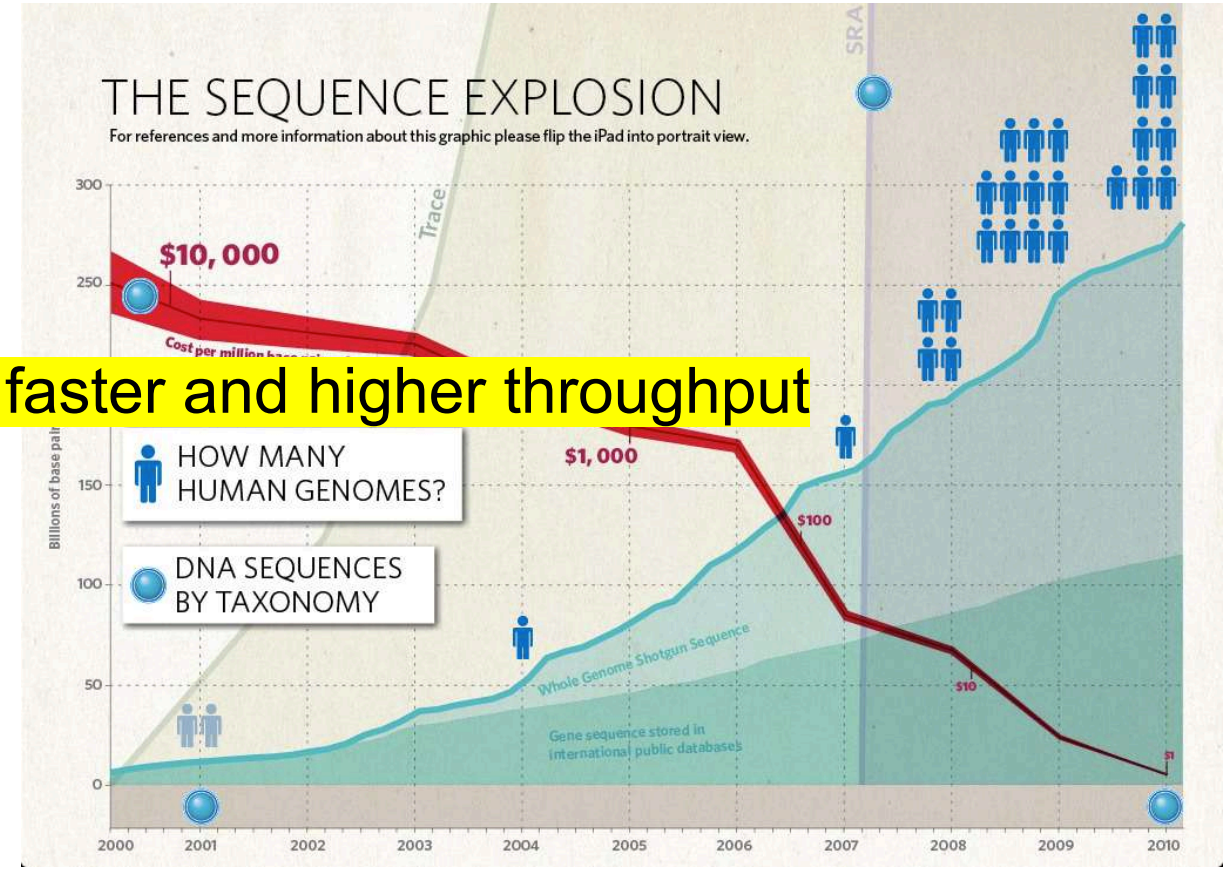
Public funding of scientific R&D has a significant positive impact on the wider economy, but quantifying the exact impact of research can be difficult to assess. A new report by research firm Battelle Technology Partnership Practice estimates that **between 1988 and 2010, federal investment in genomic research generated an economic impact of $796 billion**, which is impressive considering that Human Genome Project (HGP) spending **between 1990-2003 amounted to $3.8 billion**. This figure equates to a return on investment (ROI) of 141:1 (that is, every $1 invested by the U.S. government generated $141 in economic activity). The report was commissioned by Life Technologies Foundation.

https://www.genome.gov/27544383/calculating-the-economic-impact-of-the-human-genome-project/

2000-2010s – Second generation sequencing and associated challenges



sequencing made cheaper, faster and higher throughput



https://www.nlm.nih.gov/about/https://www.nlm.nih.gov/about/2018CJ.html

http://www.nature.com/news/2010/100331/full/464670a.html

A brief history of bioinformatics
Jeff Gauthier, Antony T Vincent, Steve J Charette, Nicolas Derome
Briefings in Bioinformatics (2018) https://doi.org/10.1093/bib/bby063

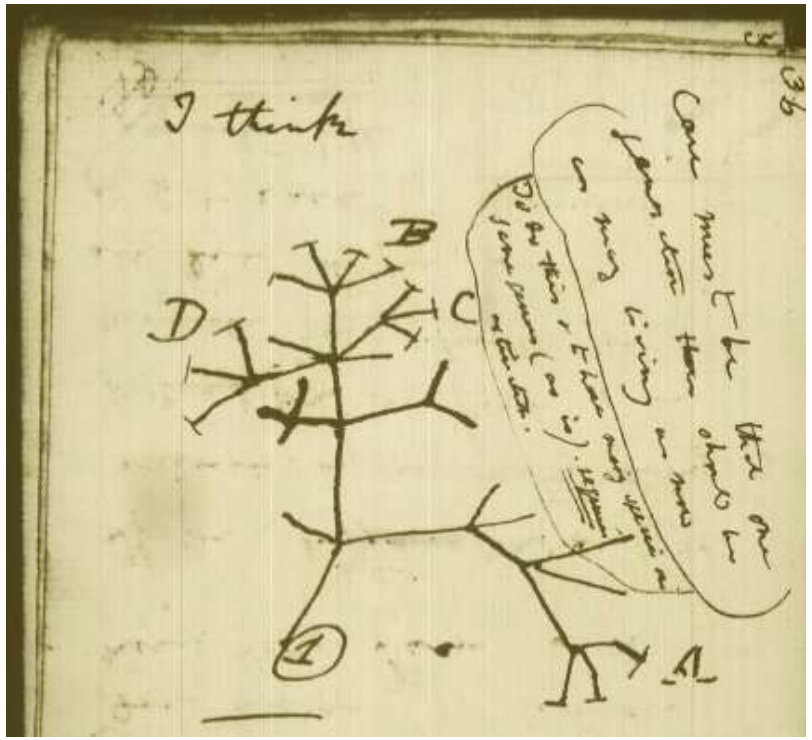
Project setup

- Assemble
  - *De novo s*equencing a species (Comparative genomics)
- Map
  - Sequencing multiple individuals of a species (Population genomics)
- Count
  - Sequencing every gene’s expression in a condition (RNAseq)
- Combination of (1), (2) and (3)

(Important!)

Nothing makes sense in the light of evolution

Theodosius Dobzhansky 1973



Different sequencing platforms / History of sequencing

Proc. Natl. Acad. Sci. USA
Vol. 74, No. 12, pp. 5463–5467, December 1977
Biochemistry

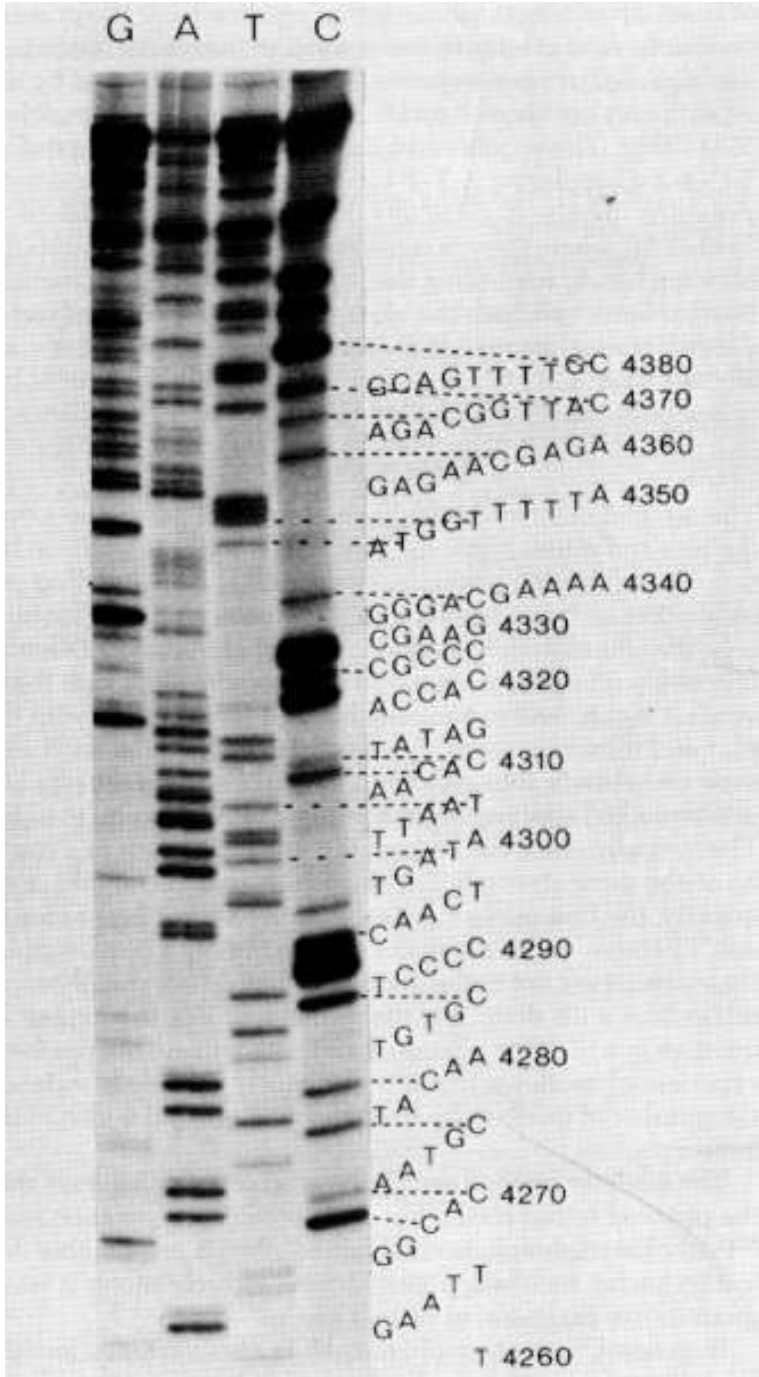
DNA sequencing with chain-terminating inhibitors

(DNA polymerase/nucleotide sequences/bacteriophage φX174)

F. SANGER, S. NICKLEN, AND A. R. COULSON

Medical Research Council Laboratory of Molecular Biology, Cambridge CB2 2QH, England

Contributed by F. Sanger, October 3, 1977

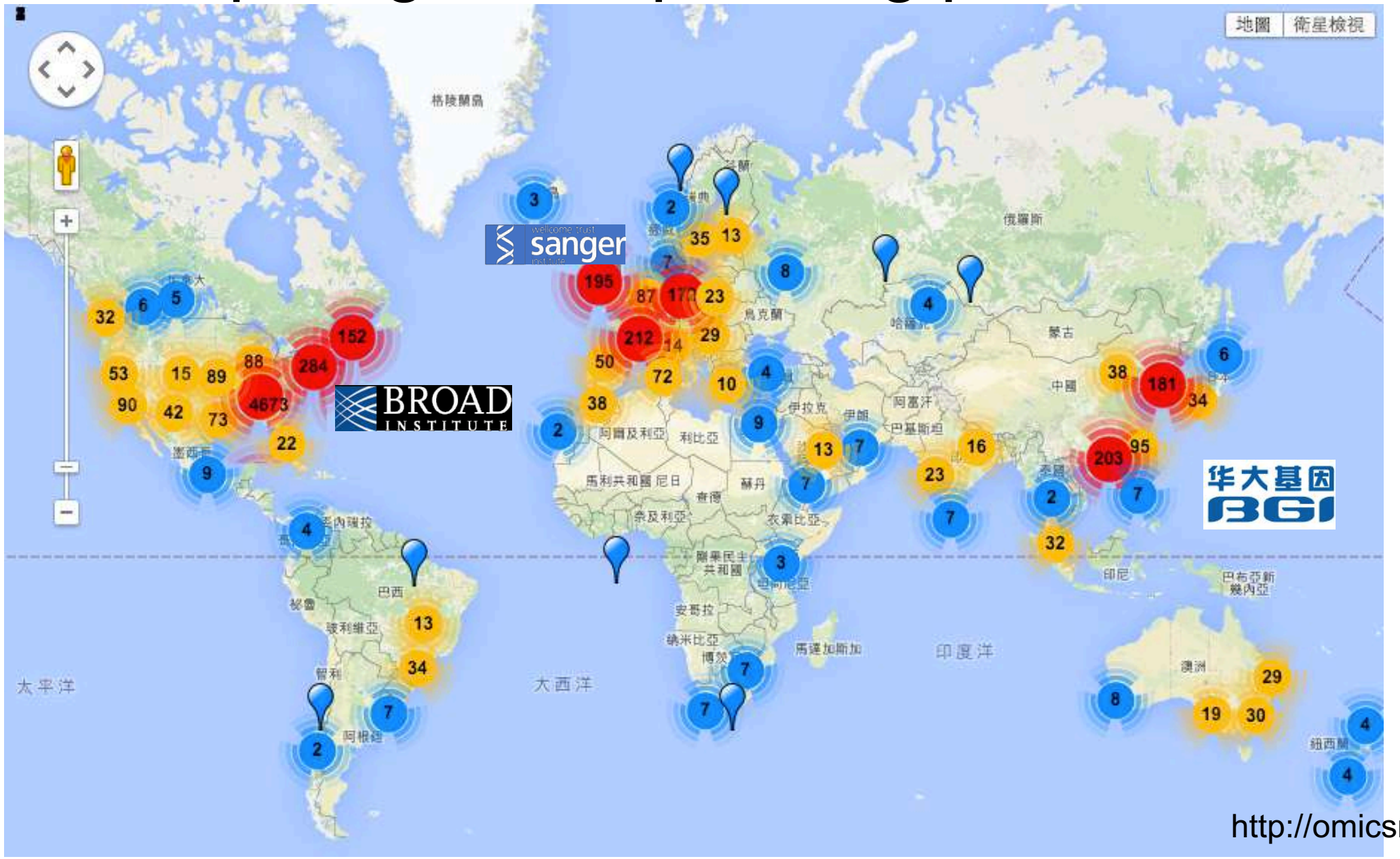


http://www.ncbi.nlm.nih.gov/pmc/articles/PMC431765/

ABI 3730xi at TIGR (1.6Mb per day)

https://www.flickr.com/photos/jurvetson/57080968

World competing for sequencing power

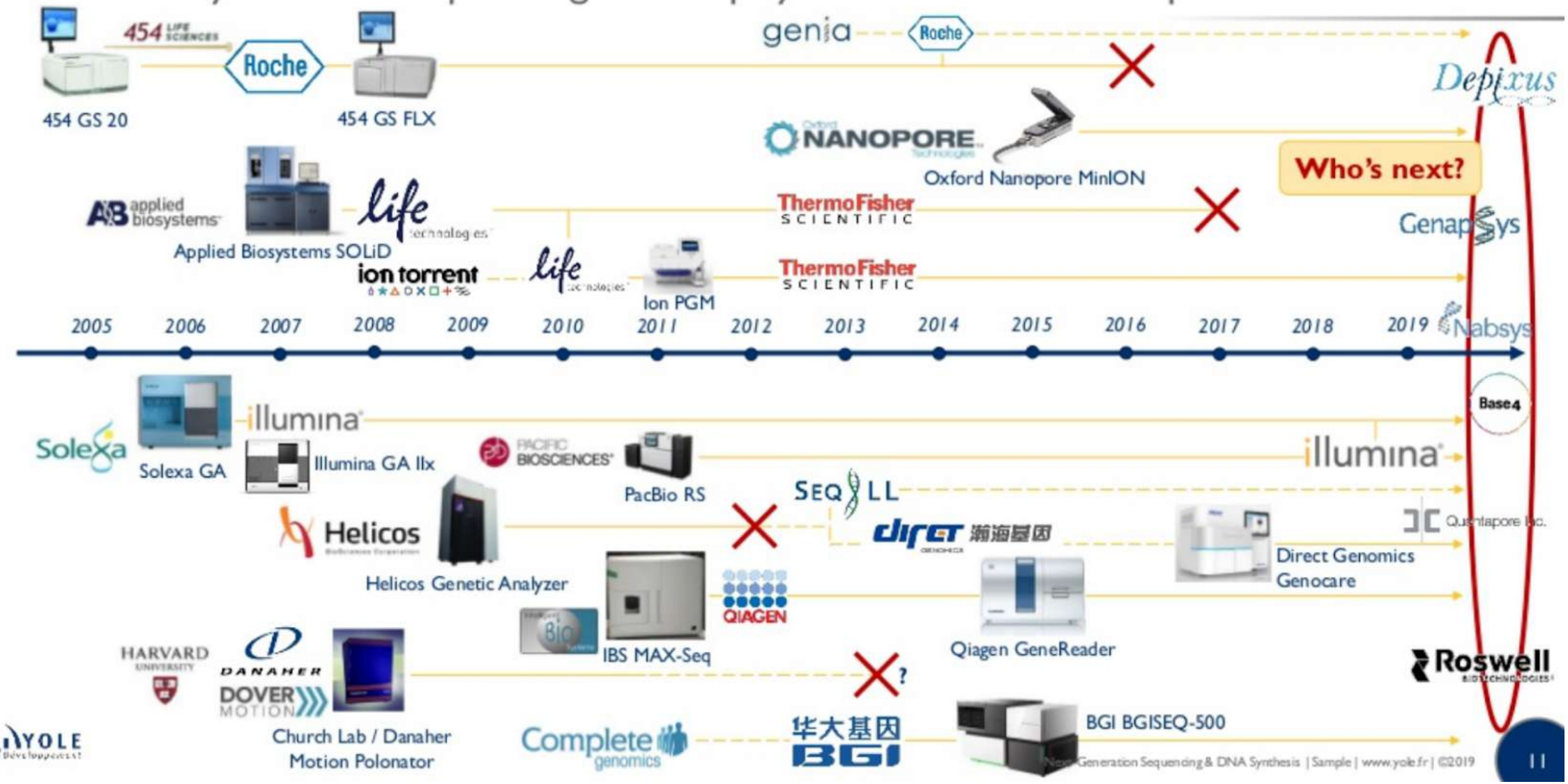


http://omicsmaps.com/

INTRODUCTION

Clip slide

History of DNA sequencing – Main players' first commercial products and M&A



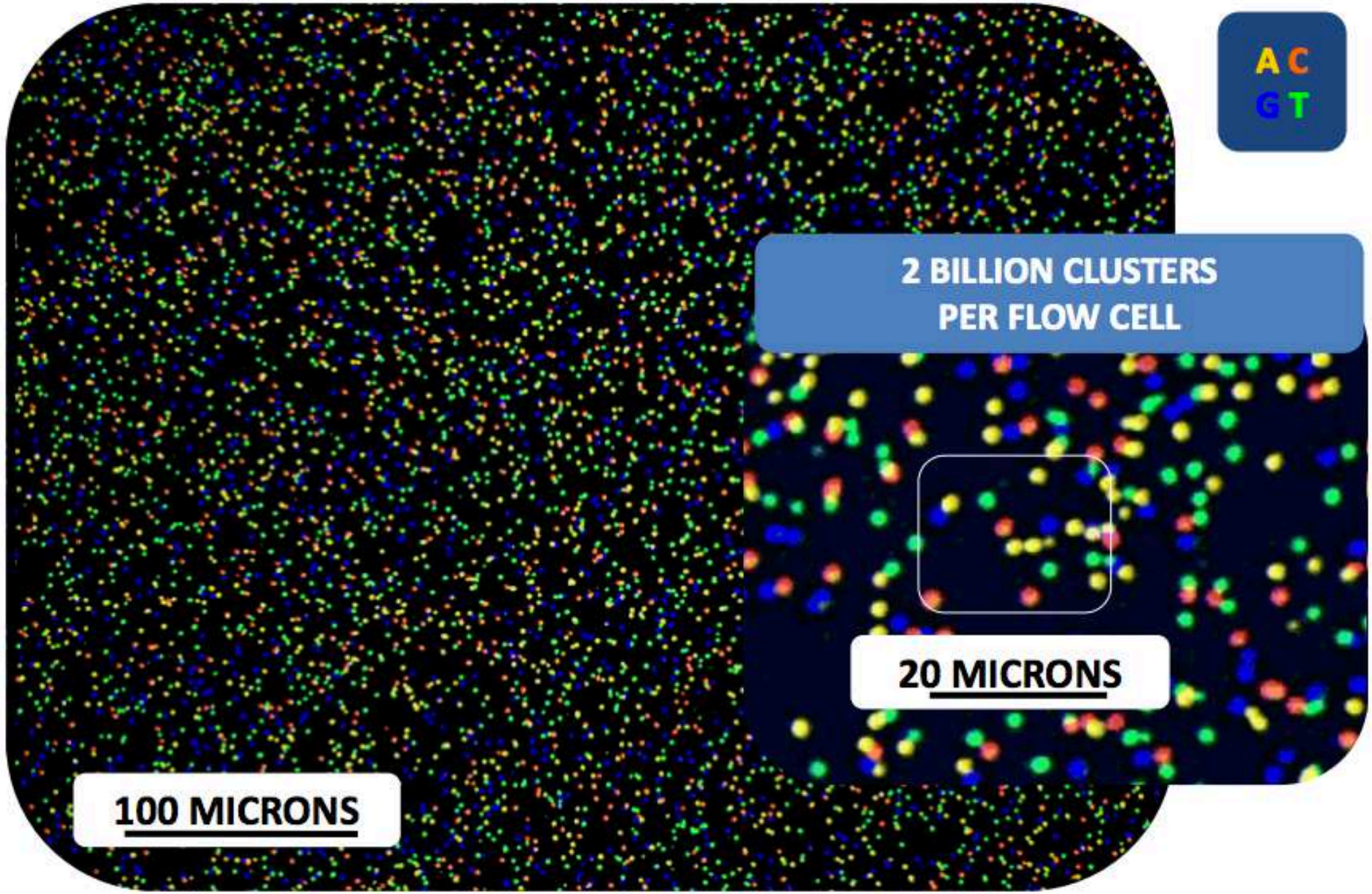
Sequencing Platforms

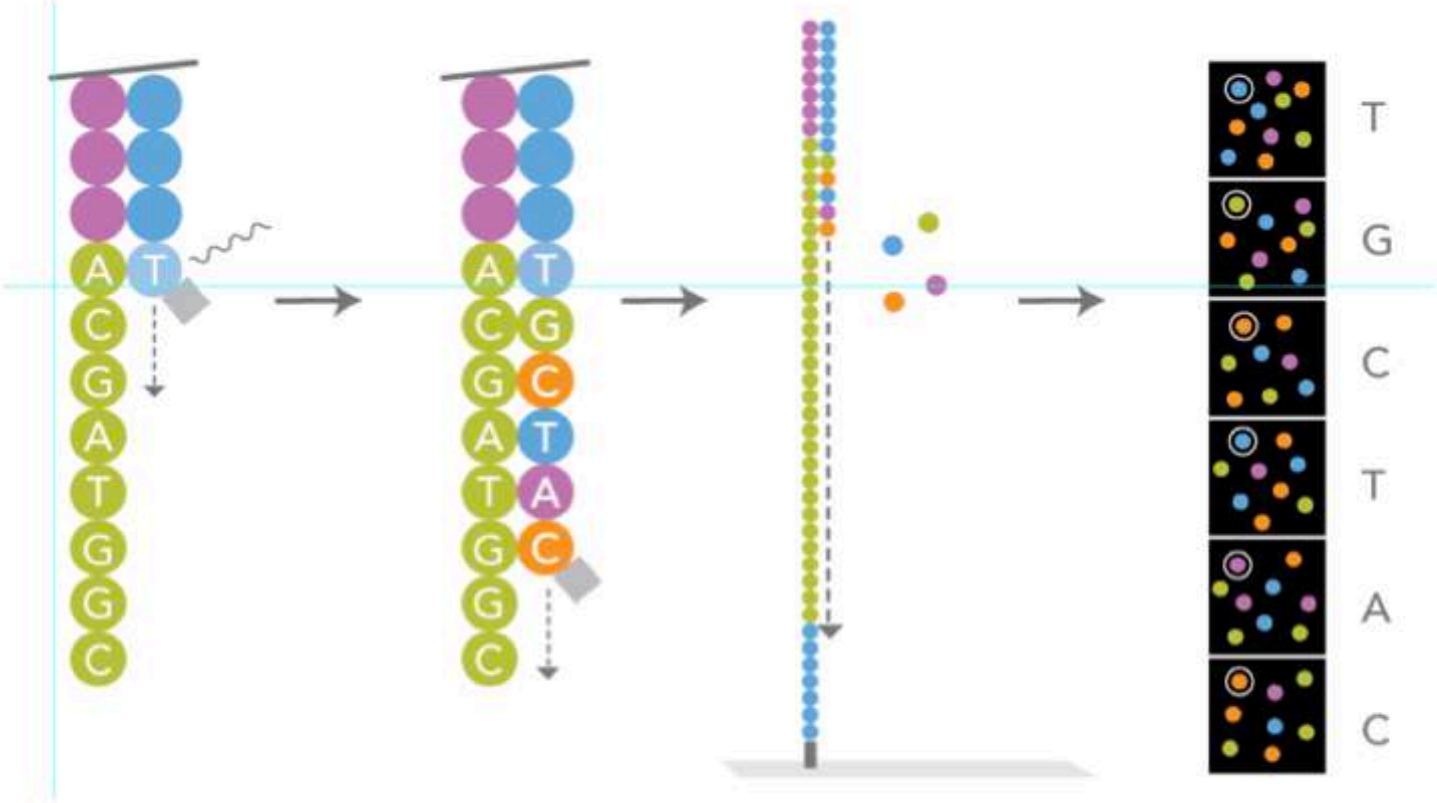
- Short reads
  1. ~~Genome Analyzer IIx (GAIIx)~~ – Illumina
  2. HiSeq, MiSeq, Novaseq – Illumina

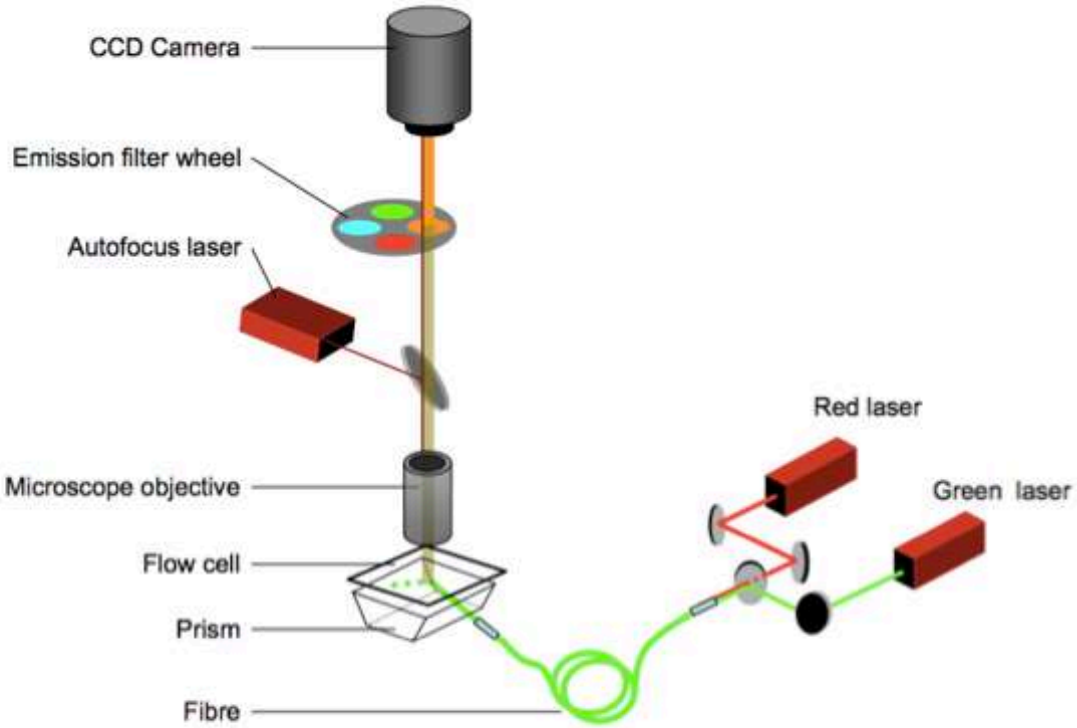
- Long reads
  1. ~~Genome Sequencer FLX System (454) – Roche~~
  2. Pacific Bioscience
  3. Oxford Nanopore

Illumina: sequencing by synthesis

https://www.youtube.com/watch?v=fCd6B5HRaZ8

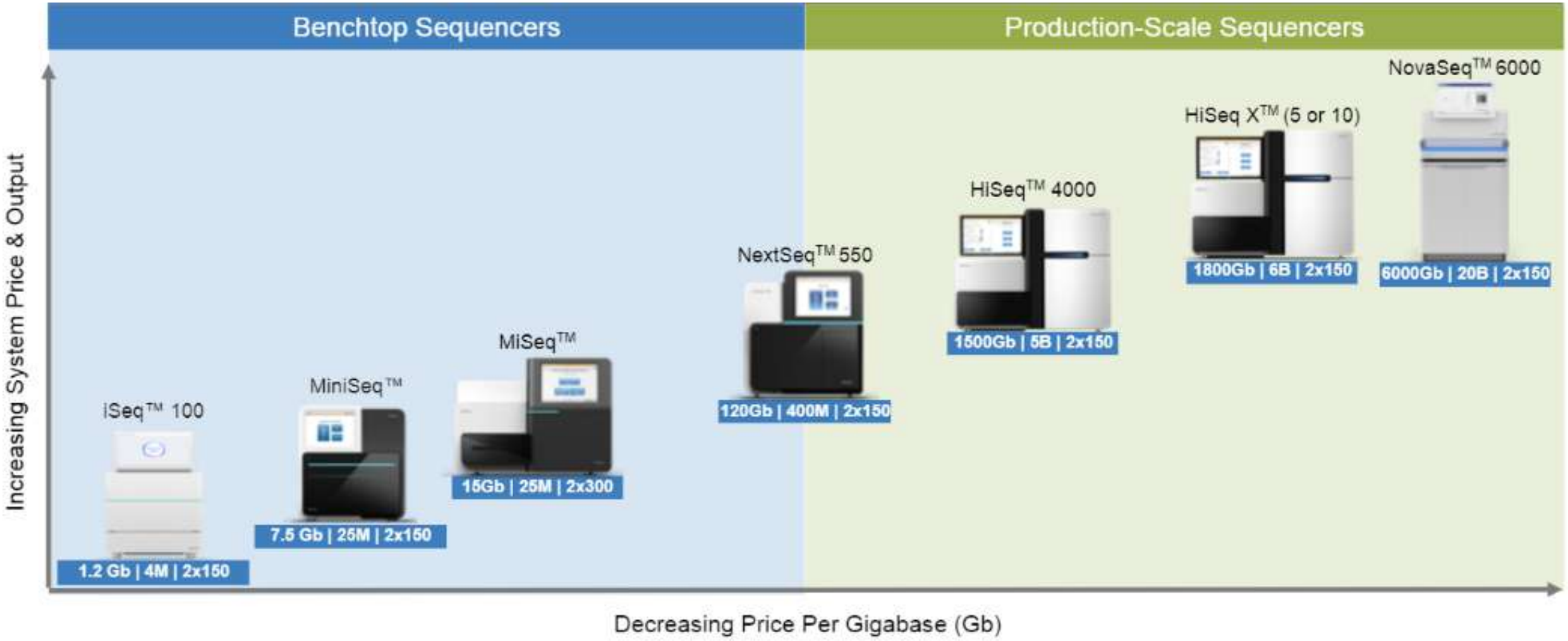

A C
G T
2 BILLION CLUSTERS
PER FLOW CELL
20 MICRONS
100 MICRONS





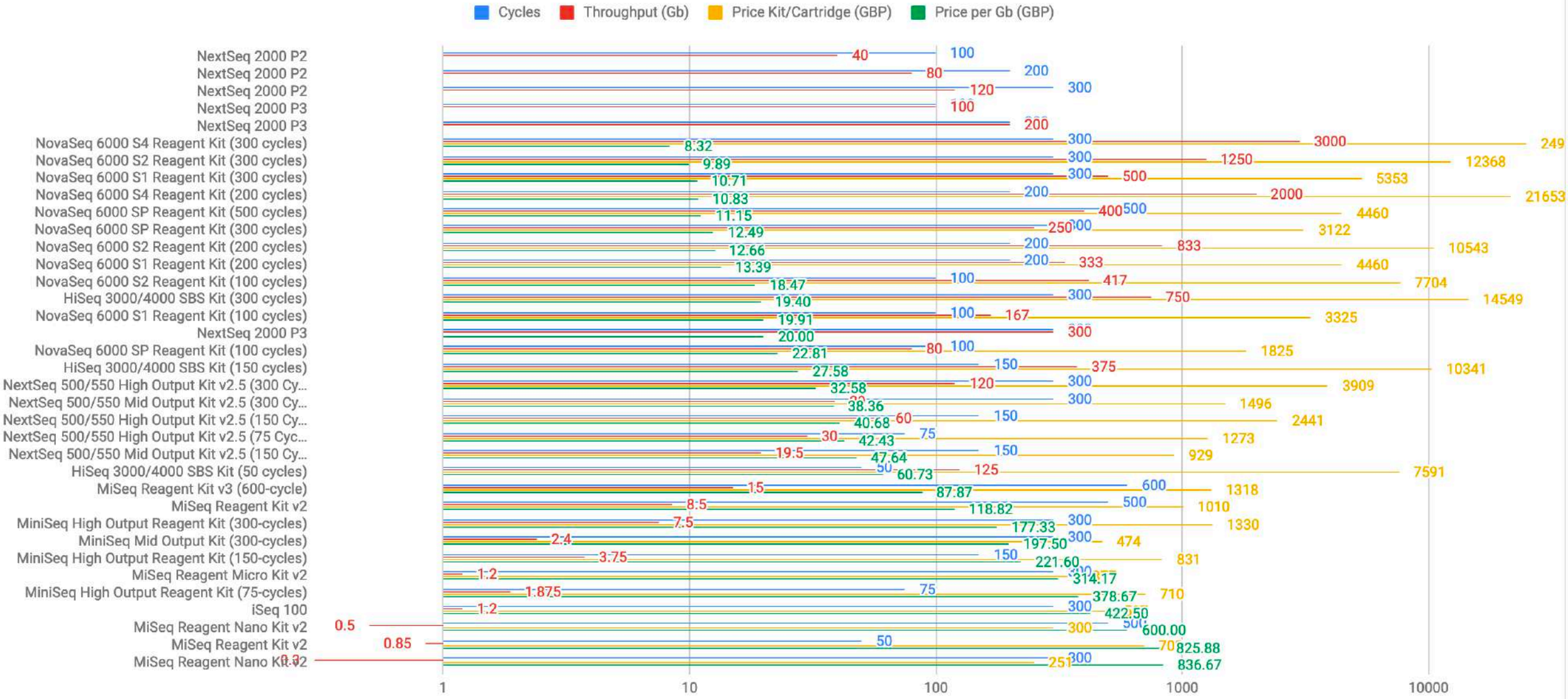
Nature **456**, 53-59 (6 November 2008)
| doi:10.1038/nature07517

Illumina machines



Illumina HiSeq

Illumina platform comparison



And the arrival of 3rd generation sequencing…

(much longer read lengths and not so bad yield!!)

PacBio (Pacific Biosciences)

RSII

Sequel II

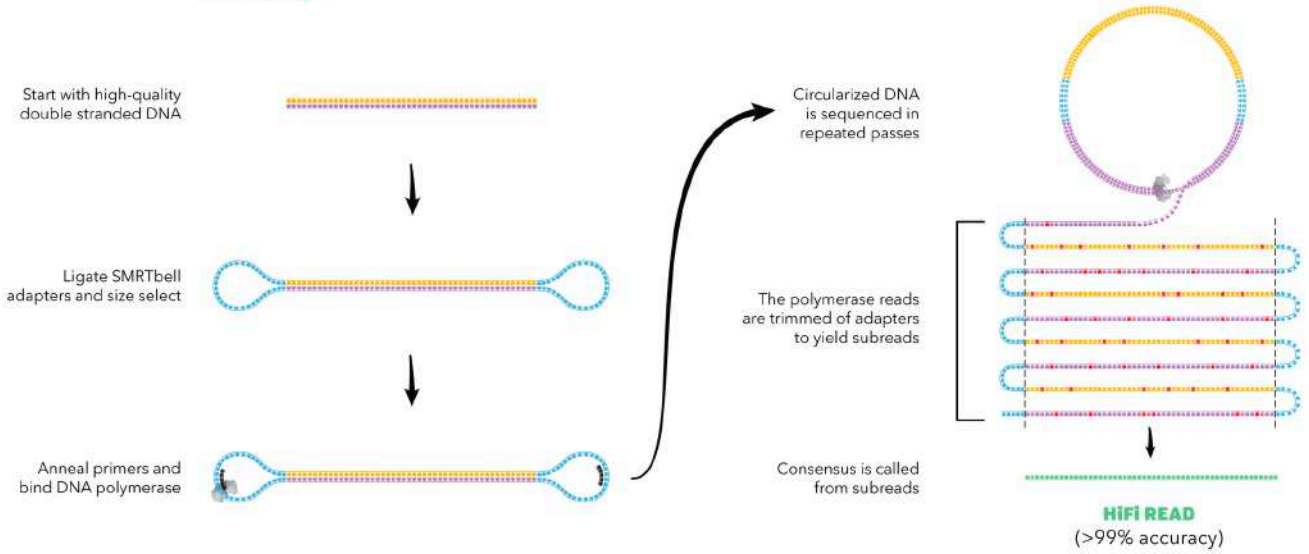
Single molecule sequencing

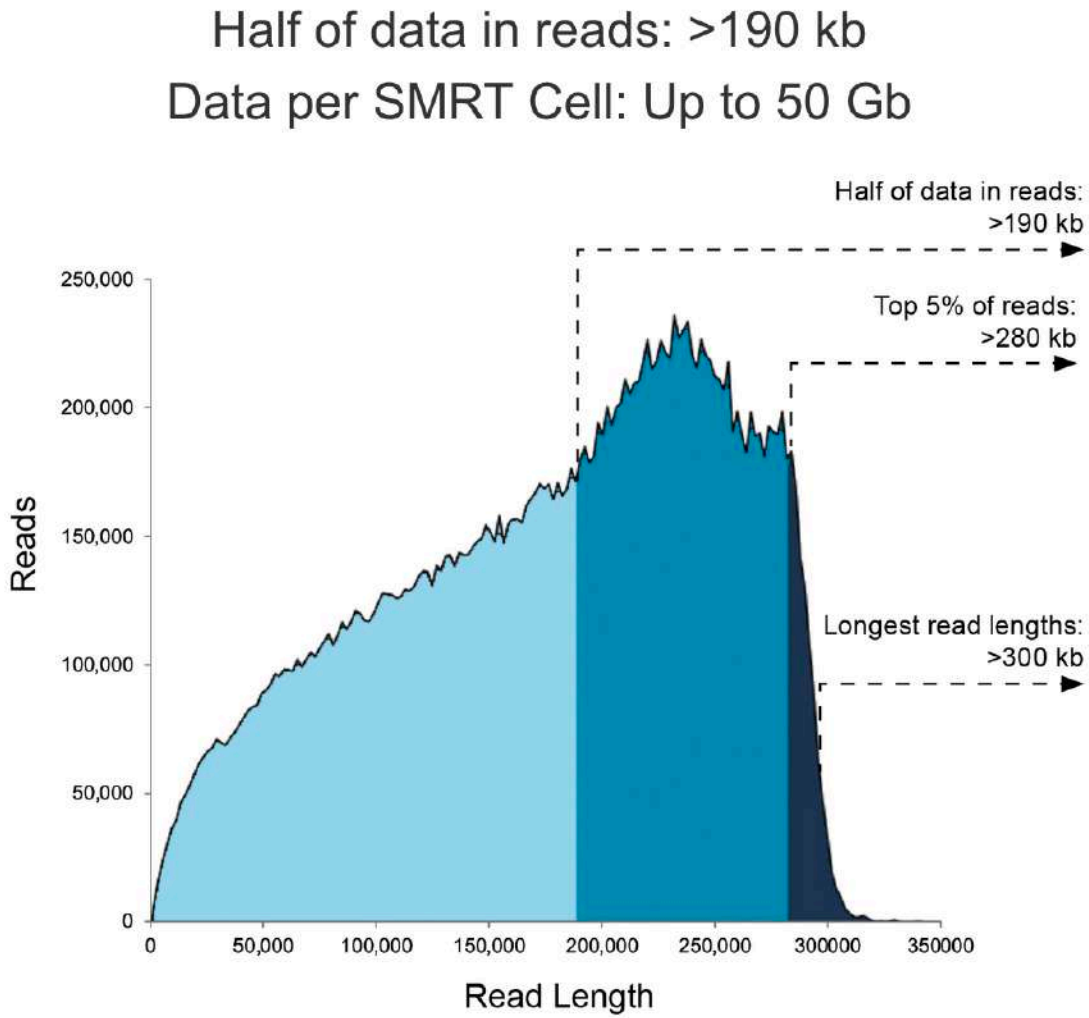
https://www.youtube.com/watch?v=NHCJ8PtYCFc

PacBio (Pacific Biosciences)

HiFi READS

Produce HiFi reads using the circular consensus sequencing (CCS) mode to provide base-level resolution with >99% single-molecule read accuracy for the detection of all variant types from single nucleotide to structural variants. Learn more about the advantages of long reads with high accuracy.





https://www.pacb.com/smrt-science/smrt-sequencing/smrt-sequencing-modes/

Oxford Nanopore

Key	SmidgION	Flongle	MinION	GridION	PromethION
					
System Price	TBC	Included in $5K Starter Pack	Included in $1K Starter Pack	Included in $50K Starter Pack	Included in $135K Starter Pack
Number of channels	200 channels	128 channels	512 channels	5 x 512 = 2,560*	48 x 3,000* = 144,000
Per flow cell Current Data – Max Data	TBC	1 - 3.3 Gb	17 - 40 Gb	17 - 40 Gb	125 - 311 Gb
Per Device Current Data – Max Data				85 - 200 Gb	3/6 - 20 Tb
Price per Gb Current Data – Max Data	**TBC**	**$90 - $30**	**$30 - $12.5**	**$17.5 - $7.5**	**$5 - $2**

Oxford Nanopore – how it works

Introduction to nanopore
https://vimeo.com/297106166

Voltrax
https://vimeo.com/297106291

Sequencing for farmers
https://vimeo.com/294216876

Rainforest
https://www.youtube.com/watch?v=6RRSxWtJPUw

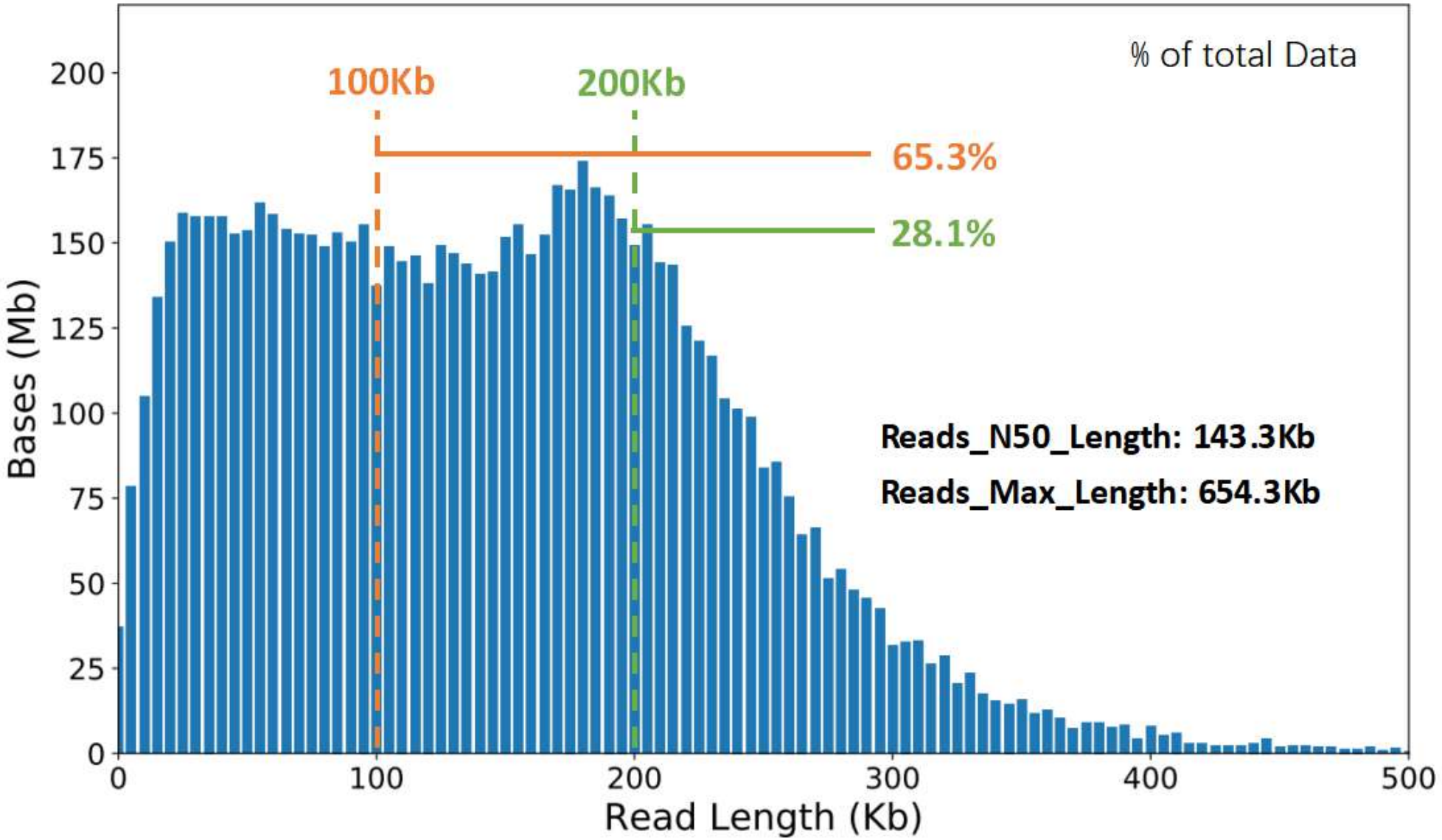
@ Oceans
https://vimeo.com/294744892

From Extreme to everyday
https://www.youtube.com/watch?v=tQ_oo7_36r8

Reference
https://nanoporetech.com/how-it-works

Nanopore Sequencing of Ebola Viruses Under Outbreak Conditions
https://www.youtube.com/watch?v=SYBzPEoENWI ; https://www.nature.com/articles/nature16996

Read length and capacity go beyond



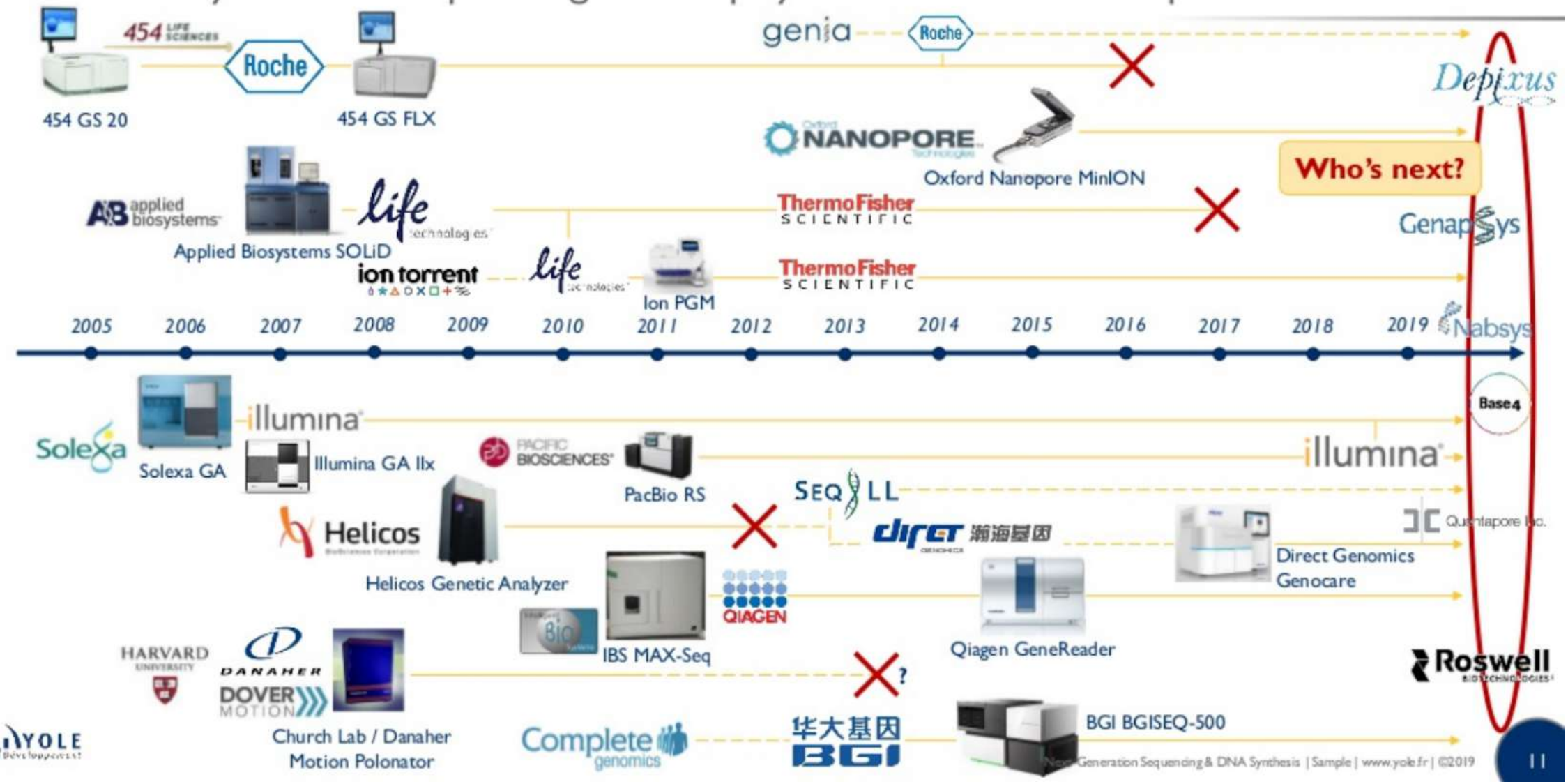
https://twitter.com/GrandOmics_Intl/status/1188724248480608257/photo/1

Come and go of technologies

INTRODUCTION

Clip slide

History of DNA sequencing – Main players' first commercial products and M&A



Break here

A lot of data

- We biologists generate a lot of data
  - Experiments, sequencing
  - Everything is more high throughput, but not necessarily less noisy

- Different data types
  - Images, Sequences, Signals, Locations, Linkage, Frequencies…

- How do we
  - analyse them?
  - store them?
  - publish them?
  - reuse them?

A small project's typical output

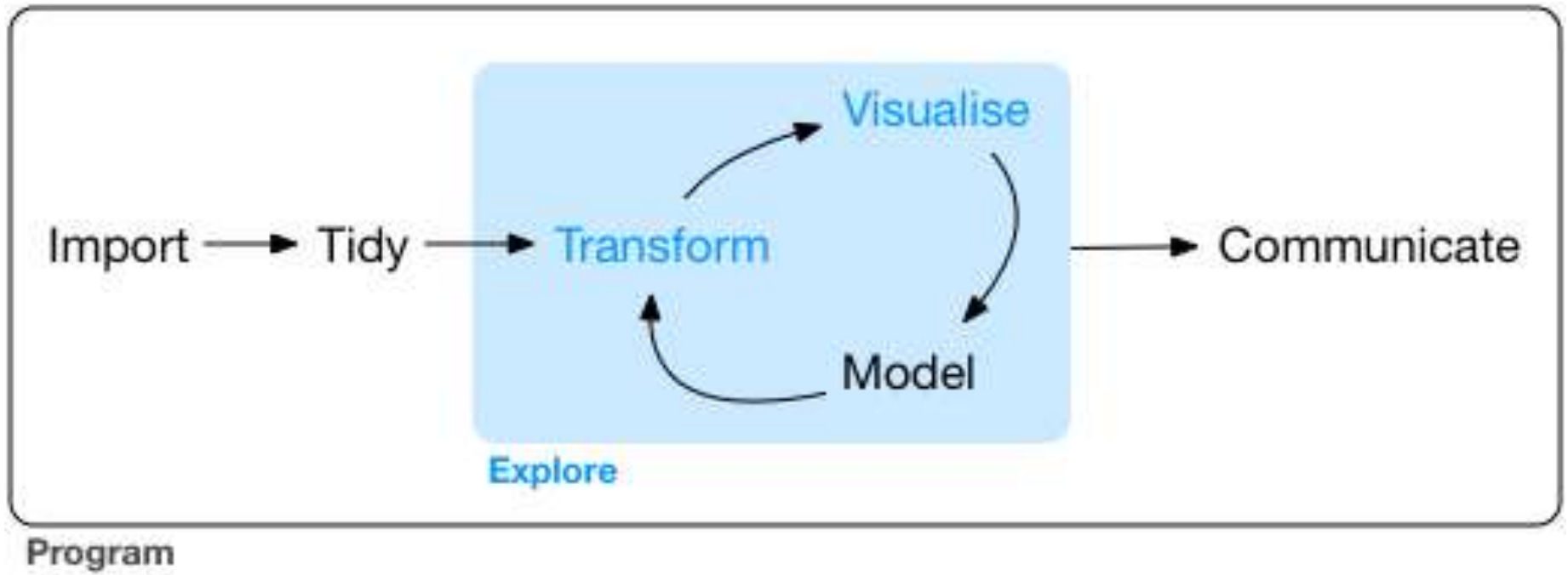
Sample Name	Sample ID	Lane ID	Yield (Mb)	# of Reads
F2-1	SG-IB01	1	11,435	75,729,838
F2-2	SG-IB02		12,014	79,561,504
F2-3	SG-IB03		11,577	76,666,714
F3-2	SG-IB05		11,119	73,638,446
F3-4	SG-IB07		10,399	68,870,380
F3-5	SG-IB08		11,671	77,292,976
F3-1	SG-IB09		12,474	82,610,516
F3-3	SG-IB10		11,916	78,915,536
F2-1	SG-IB01	2	11,366	75,271,724
F2-2	SG-IB02		11,920	78,940,010
F2-3	SG-IB03		11,481	76,031,166
F3-2	SG-IB05		11,054	73,203,066
F3-4	SG-IB07		10333	68,429,564
F3-5	SG-IB08		11550	76,488,178
F3-1	SG-IB09		12328	81,640,878
F3-3	SG-IB10		11812	78,225,876

8 exome samples ;
2 Illumina Hiseq lanes with **184GB** of data

~100X of human exome to detect disease causing SNP

Higher yield at lower cost = More samples can be barcoded into one lane

More samples = more replicates (power) in statistical analysis to pick up real biological difference



http://r4ds.had.co.nz/explore-intro.html

More data but less people with informatics skills

- Sequencing is the result of many types of experiment

- Everyone wants to make use of this technology
- Not everyone will be able analyse them
  - You can’t just open the file in Microsoft office anymore
- Collaborate or learn yourself

- **Bottleneck is bioinformatics analysis**

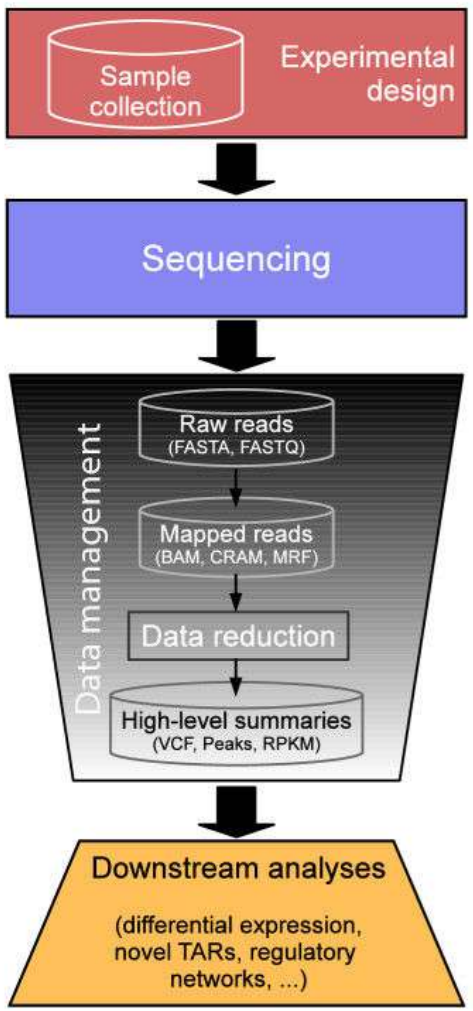
Sboner et al. Genome Biology 2011, 12:125
http://genomebiology.com/2011/12/8/125

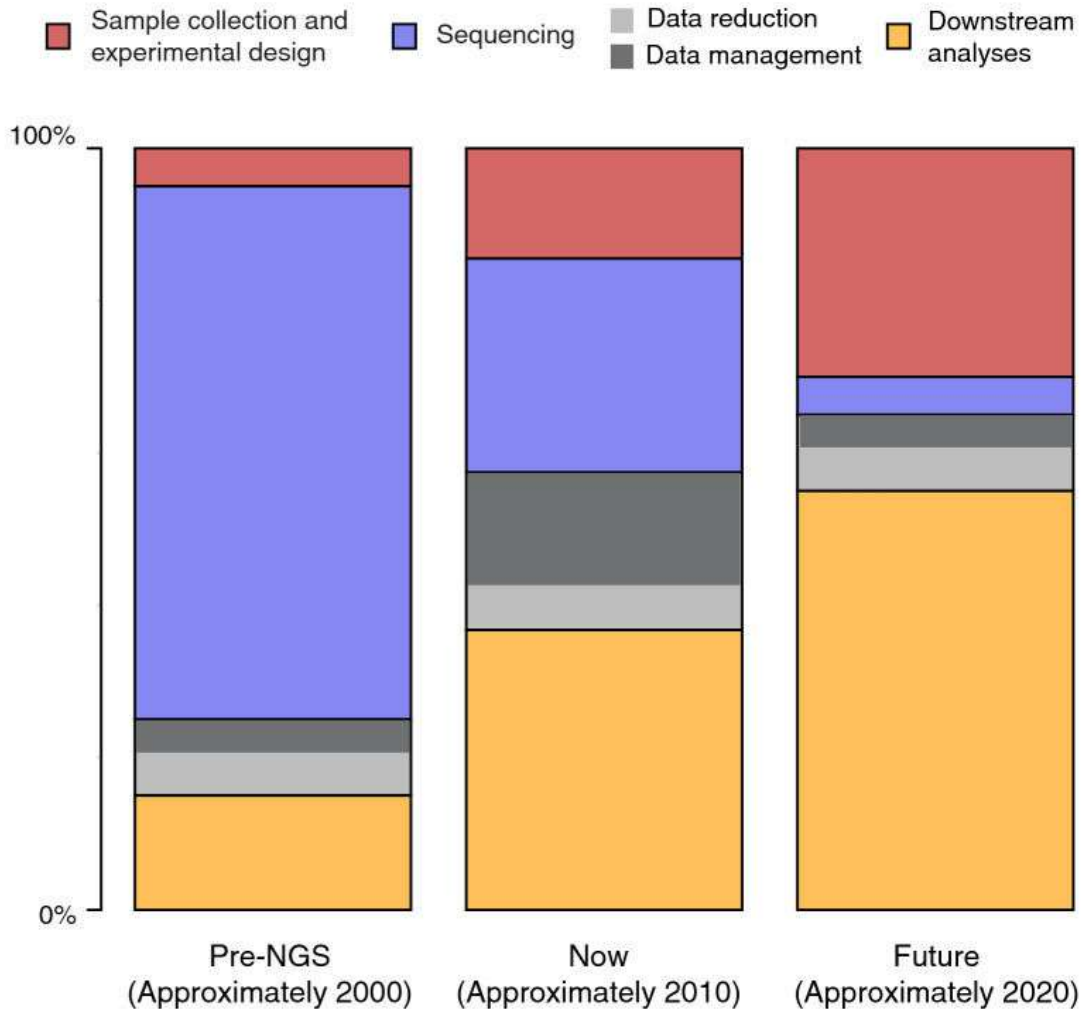
Genome Biology

OPINION

The real cost of sequencing: higher than you think!

Andrea Sboner[1,2], Xinmeng Jasmine Mu[1], Dov Greenbaum[1,2,3,4,5], Raymond K Auerbach[1] and Mark B Gerstein*[1,2,6]





You will end up with an analysis **pipeline**

Run **multiple programs** to analyse / get the results

Important problems:

- Which program to use?
- Which parameter to use for each program?
- How do you get results of program A to feed into program B?
- How do you know if the program finishes correctly?
- Is there ever going to be a correct answer? (most likely no)

No ’perfect’ pipeline – learn through experience

Always understand your data / programs

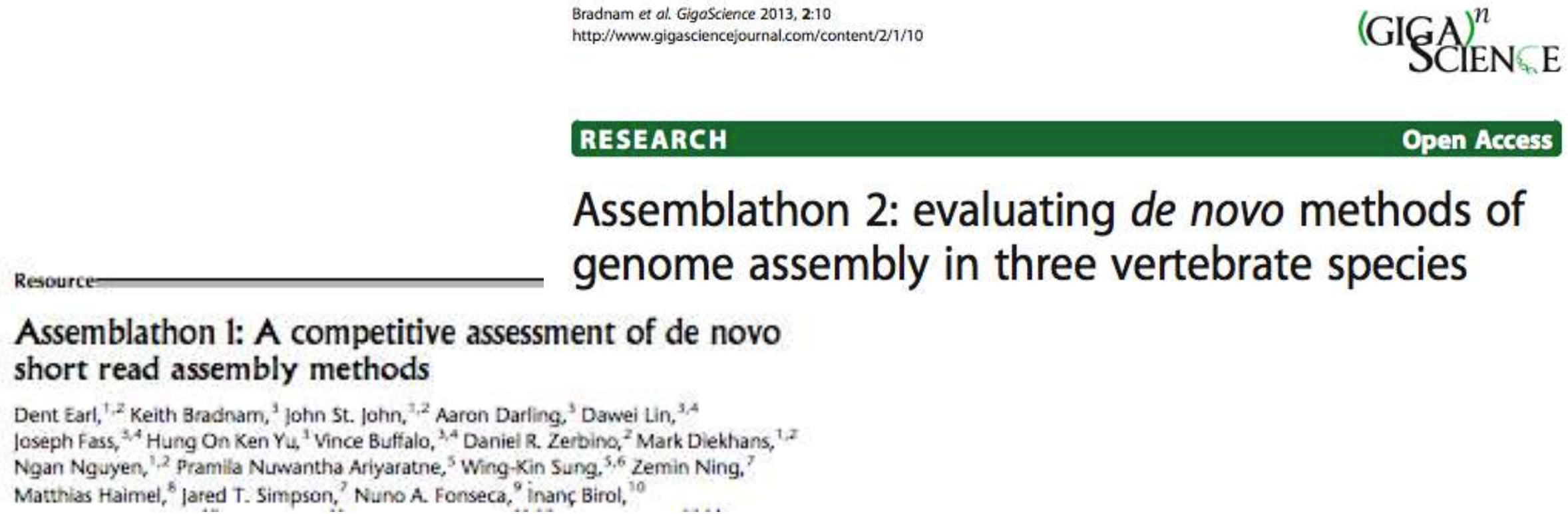
- Understand:
  - Data format
  - The nature of your data

- Please don’t
  - assume data you are given is ’correct’
    - Scenario 1: We got the assmblies and analysis from company XXXX, and we don’t know what to do with it

  - assume everything’s correct online
  - Run everything in ‘default’ mode

If unsure – always check **benchmark** studies

- Don’t run programs that you are not sure the concepts
- Programs need to be **benchmarked**
- **Always look for most recent (and fair) benchmarks**



Bradnam *et al. GigaScience* 2013, **2**:10
http://www.gigasciencejournal.com/content/2/1/10

(GIGA)^n SCIENCE

RESEARCH Open Access

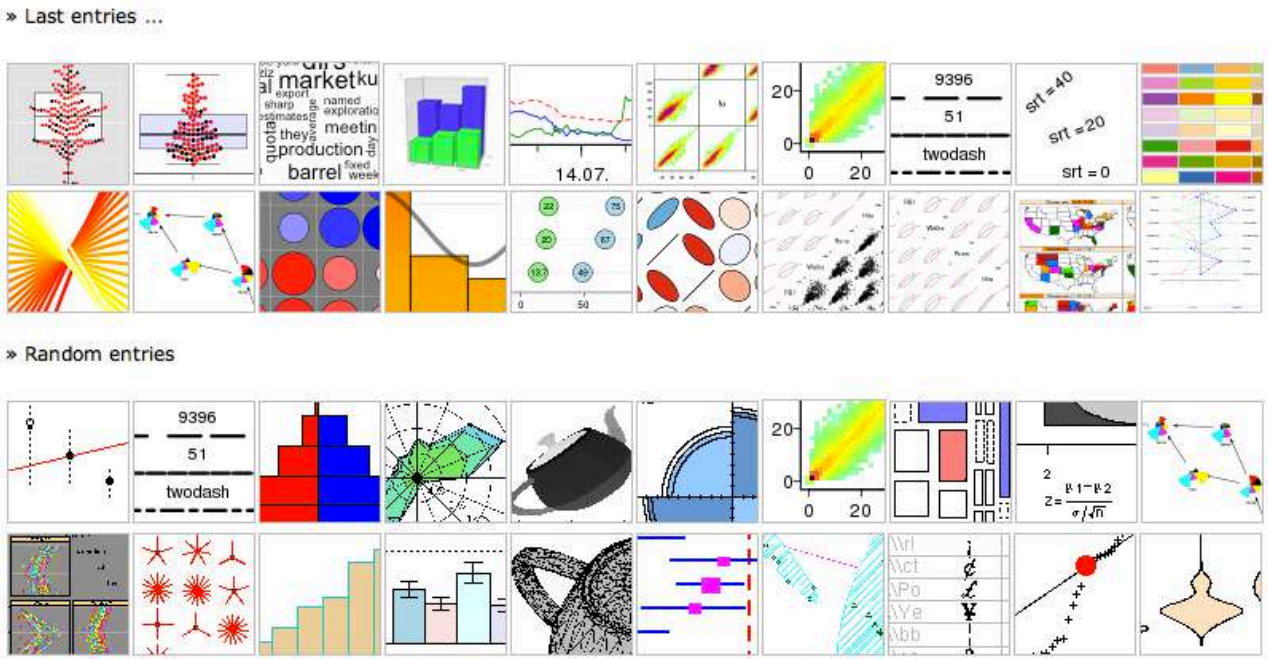
Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species

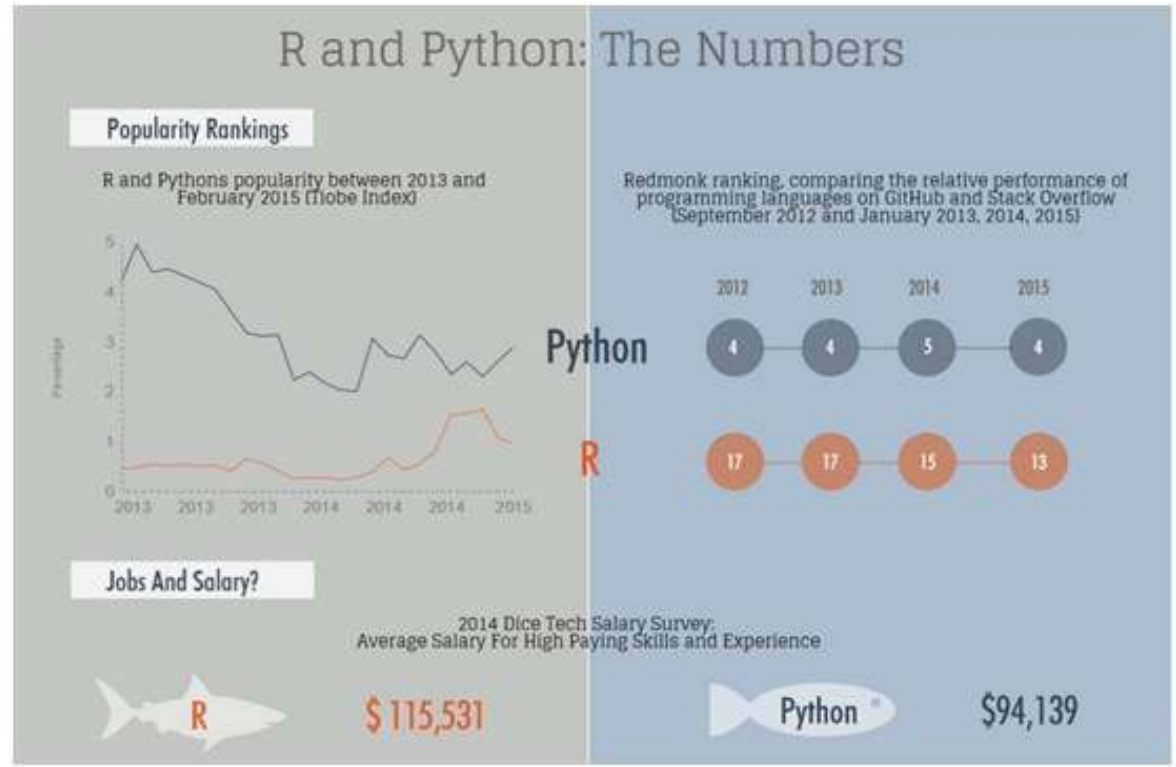
Resource

Assemblathon 1: A competitive assessment of de novo short read assembly methods

Dent Earl,[1,2] Keith Bradnam,[3] John St. John,[1,2] Aaron Darling,[3] Dawei Lin,[3,4] Joseph Fass,[3,4] Hung On Ken Yu,[3] Vince Buffalo,[3,4] Daniel R. Zerbino,[2] Mark Diekhans,[1,2] Ngan Nguyen,[1,2] Pramila Nuwantha Ariyaratne,[5] Wing-Kin Sung,[5,6] Zemin Ning,[7] Matthias Haimel,[8] Jared T. Simpson,[7] Nuno A. Fonseca,[9] İnanç Birol,[10]

Python and R


» Last entries ...
14.07.
9396
51
twodash
srt = 40
srt = 20
srt = 0
» Random entries
9396
51
twodash


R and Python: The Numbers
Popularity Rankings
R and Pythons popularity between 2013 and February 2015 (Tiobe Index)
Redmonk ranking, comparing the relative performance of programming languages on GitHub and Stack Overflow (September 2012 and January 2013, 2014, 2015)
2012
2013
2014
2015
Python
4
4
5
4
R
17
17
15
13
Jobs And Salary?
2014 Dice Tech Salary Survey:
Average Salary For High Paying Skills and Experience
R
$ 115,531
Python
$94,139

FASTA format

>Name_of_sequence
GCGGGCATCCGCTGCGTGCTGGGCAAAGTCTGTGGCGA
CCGTATGAAACCCTGAACCCGGCAATACGGTTGCAGATG
CTGGAGGCGGCGGCAACGGCAATCAGCTTGATTGAGGT
GAGGTACATAAGCCCGCCAGGCGAGCAGGCAAAGCTGT
TCTGGTCCGAGCCCCAAACAGGGTTCACCAGTGGCCTG
CCGACGAAAGCGCCGAAGCCCGAACCCATCAATCACTG
GCAGCGTGCAGTCCAGGCCATCGACGAGGCCATCATTGA
AGCGCGGTACGACCCCGAAACGGCACGCTCATTGTTTGC
GTTGGCTTCCTATGGTCGGCGCGACCCAGCTTCCCTGGA
ACAGTTGCGCGCCACCTTCGCGAAGGAAGGCATTCCCC
CGGAATTTCTGTCACATTATGAGCCTGACGGACCCTTTGC

Alignment format

- Some programs need slightly modified format

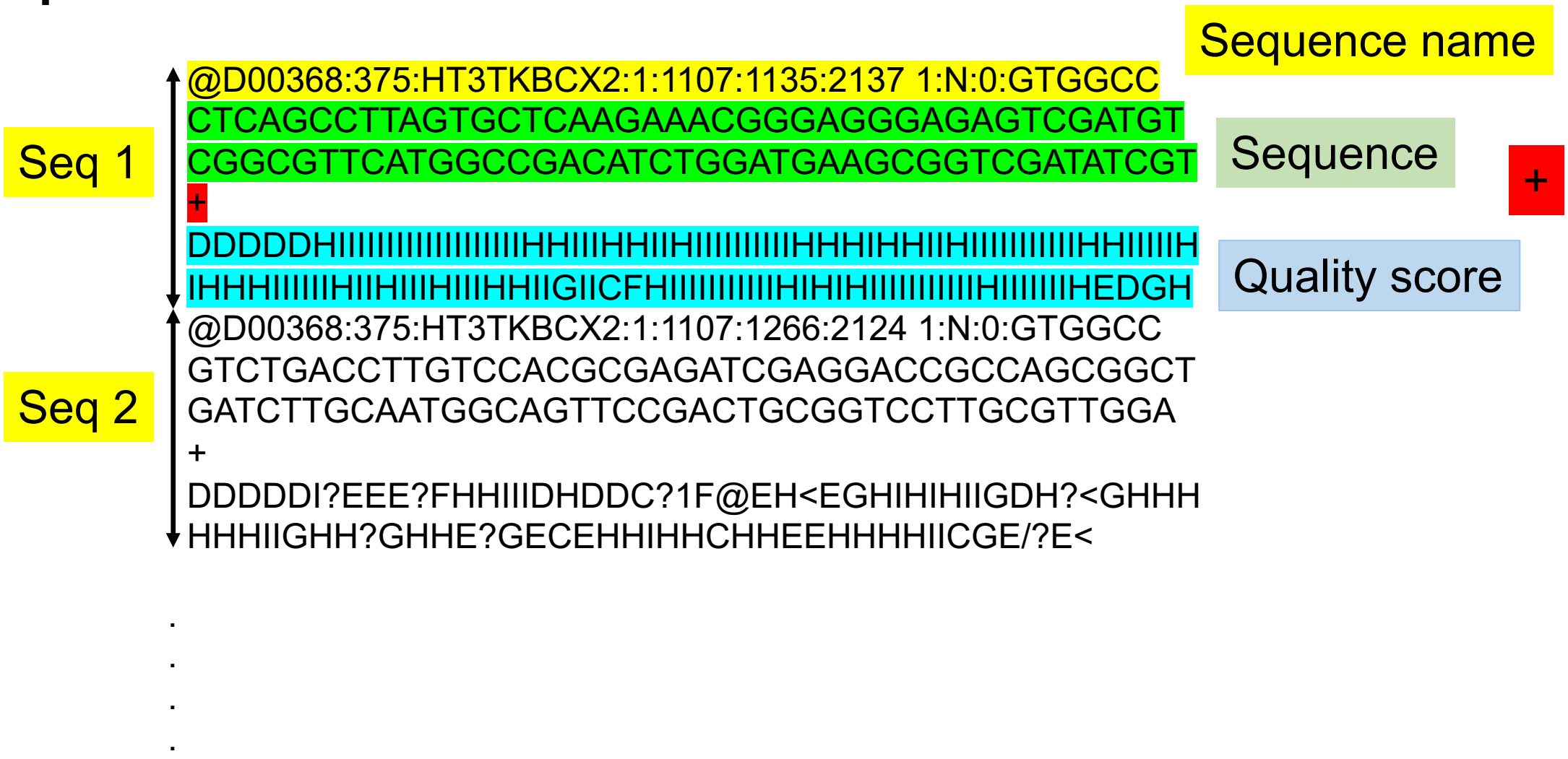
```
>Name_of_sequence_1
GCGGGCATCCGCTGCGTGCTGGGCAAAGTCTGTGGCGA
CCGTATGAAACCCTGAACCCGGCAATACGGTTGCAGATG
CTGGAGGCGGCGGCAACGGCAATCAGCTTGATTGAGGT
GAGGTACATAAGCCCGCCAGGCGAGCAGGCAAAGCTGT
TCTGGTCCGAGCCCCAAACAGGGTTCACCAGTGGCCTG
CCGACGAAAGCGCCGAAGCCCG
>Name_of_sequence_2
GCGGGCATCCGCTGCGTGCTGGGCAAAGTCTGTGGCGA
CCGTATGAAACCCTGAACCCGGCAATACGGTTGCAGATG
CTGGAGGCGGCGGCAACGCAATCAGCTTGATTGAGGTG
AGGTACATAAGCCCGCCAGGCGAGCAGGCAAAGCTGTTC
TGGTCCGAGCCCCAAACAGGGTTCACCAGTGGCCTGCC
GACGAAAGCGCCGAAGCCCG
```

Data type keep evolving

- Very first fastq file was invented in 2007?
- Obviously will become problematic in storage later on…

>Name_of_sequence_1
GCGGGTA
>Name_of_sequence_1
20 30 33 30 20 33 19

Fastq file


Sequence name
Seq 1
@D00368:375:HT3TKBCX2:1:1107:1135:2137 1:N:0:GTGGCC
CTCAGCCTTAGTGCTCAAGAAACGGGAGGGAGAGTCGATGT
CGGCGTTCATGGCCGACATCTGGATGAAGCGGTCGATATCGT
Sequence
+
+
DDDDDHIIIIIIIIIIIIIIIIIIHHIIIHHIIHIIIIIIIIIHHHIHHIIHIIIIIIIIIIHHIIIIIH
IHHHIIIIIIHIIHIIIHIIIHHIIGIICFHIIIIIIIIIIHIHIHIIIIIIIIIIHIIIIIIIHEDGH
Quality score
Seq 2
@D00368:375:HT3TKBCX2:1:1107:1266:2124 1:N:0:GTGGCC
GTCTGACCTTGTCCACGCGAGATCGAGGACCGCCAGCGGCT
GATCTTGCAATGGCAGTTCCGACTGCGGTCCTTGCGTTGGA
+
DDDDDI?EEE?FHHIIIDHDDC?1F@EH<EGHIHIHIIGDH?<GHHH
HHHIIGHH?GHHE?GECEHHIHHCHHEEHHHHIICGE/?E<

.
.
.
.

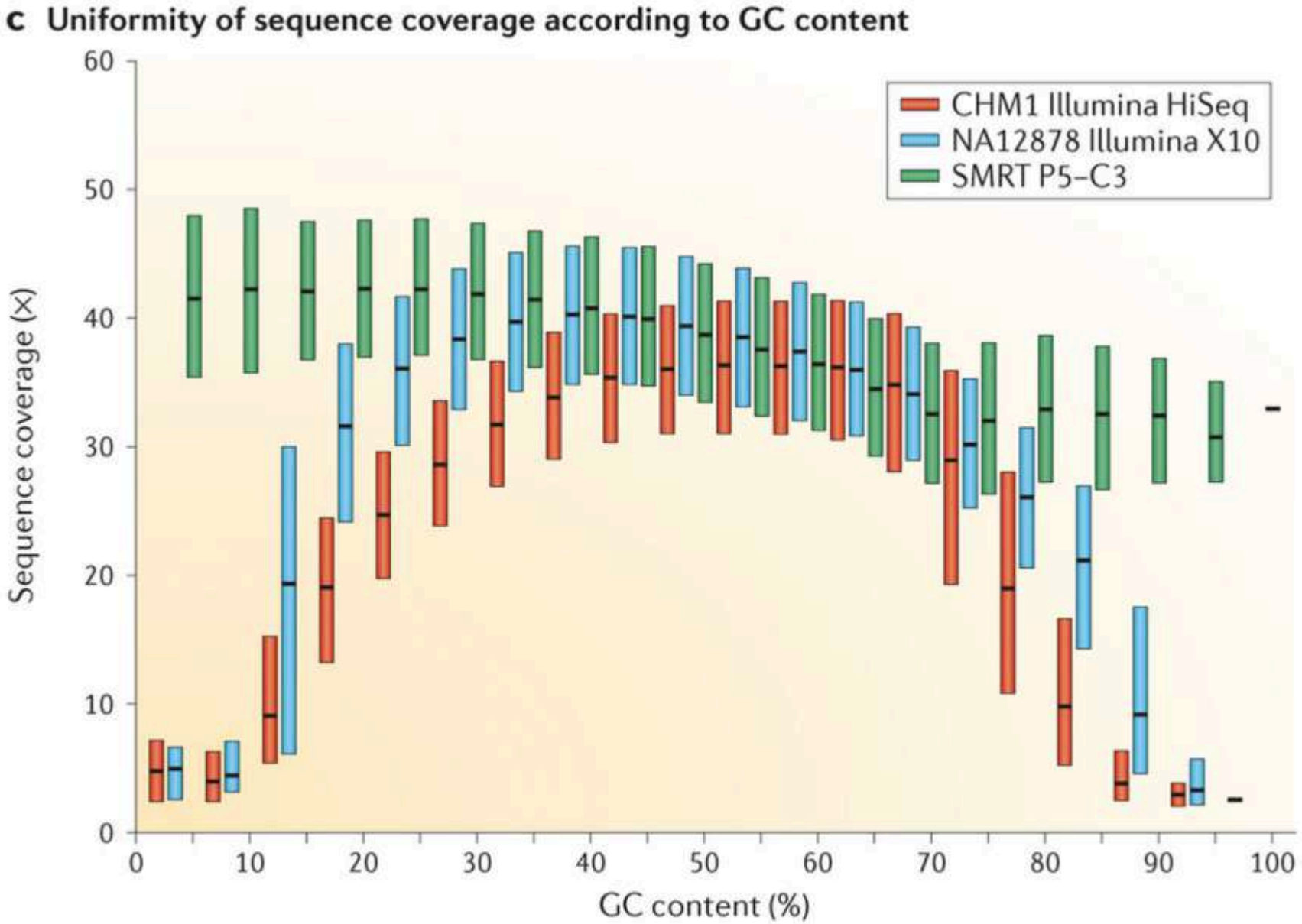
Analysis and interpretation

Is your data good enough?

(garbage in, garbage out)

https://sequencing.qcfail.com

Sequencing Biases

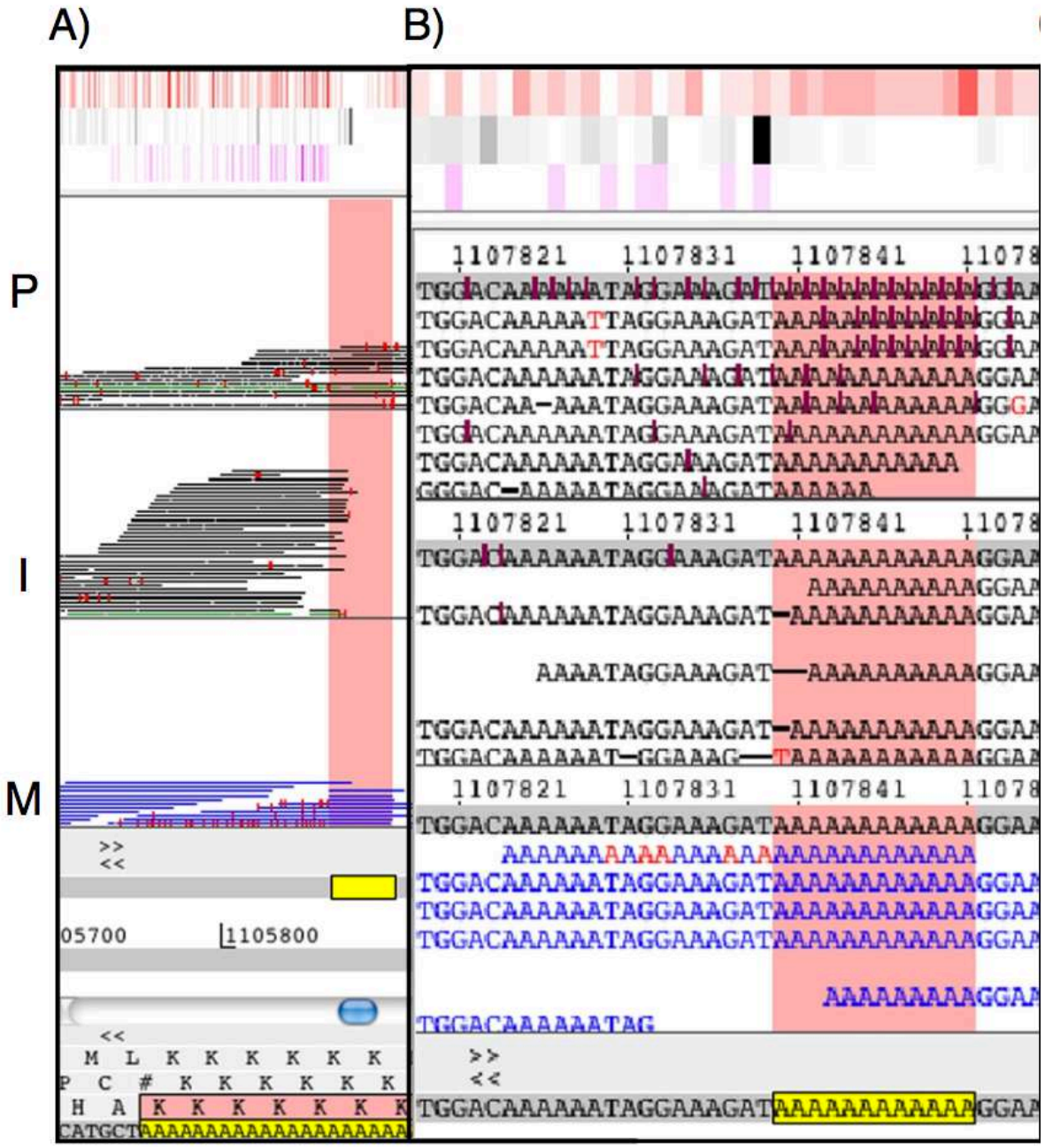


http://www.nature.com/nrg/journal/v16/n11/fig_tab/nrg3933_F2.html

Sequencing Errors

A) Illustration of errors in Illumina data after a long homopolymer tract. Ion torrent data has a drop of coverage and multiple indels are visible in PacBio data.

B) Example of errors associated with short homopolymer tracts. Multiple insertions are visible in the PacBio Data… MiSeq sequences read generally correct through the homopolymer
tract.



Quail *et al.*, BMC Genomics (2012) **13**:341

Three situations you are most likely to encounter

Genome reference is available (for example, humans):

- Re-sequence (DNA, RNA)
- **Map** sequence to the genome

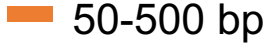
Genome reference is NOT available

- **Assemble** the reads to get the genome

Counting:

- For a given region (gene) we want to know how much.→ gene expression or metagenomics

More Definition

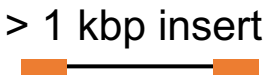


Read

A sequenced piece of DNA

Paired-end read

Sequencing both ends of a short DNA fragment



Mate-pair read

Sequencing both ends of a long DNA fragment

Insert size

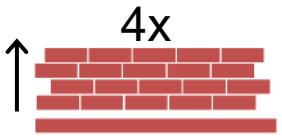
The length of the DNA fragment

Contig

A set of overlapping DNA segments that represents a consensus region of DNA

Scaffold

Contigs separated by gaps of known length



Coverage

The number of times a specific position in the genome is covered by reads

What is an alignment?

Align the following two sequences:

```
ATTGAAAGCTA
GAAATGAAAAGG
1:
--ATTGAAA-GCTA
   | ||||| |
GAAATGAAAAGG--

2:
ATTGAAA-GCTA---
   |||| |   |
---GAAATGAAAAGG
```

Scoring scheme is needed:

1 for match

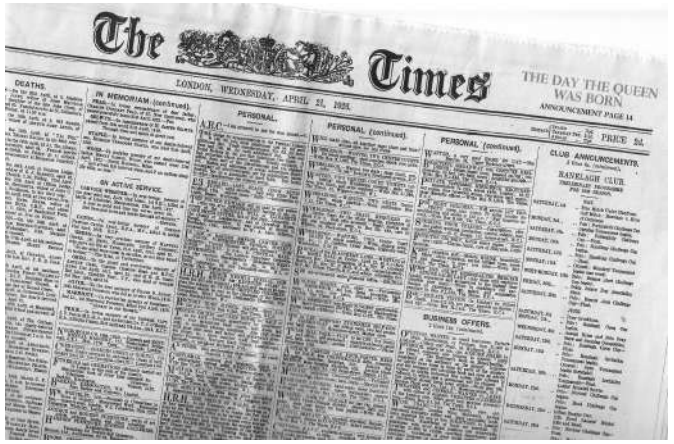
-1 for mismatch

-2 for gap

insertions / deletions (indels) mismatches

Which alignment is better?

Assembly

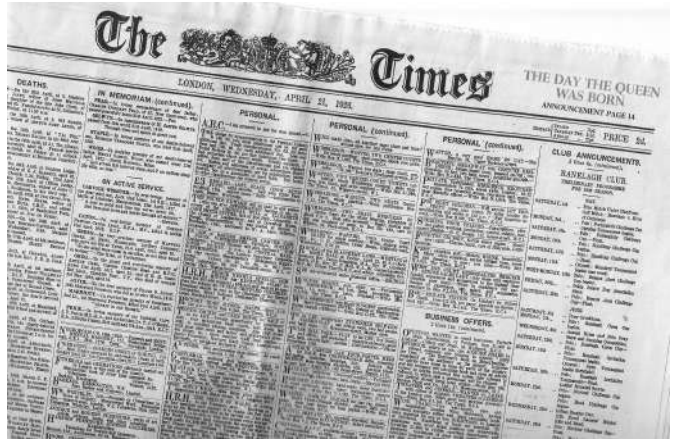


Genome
(3.000.000 letters)

Sequencing →

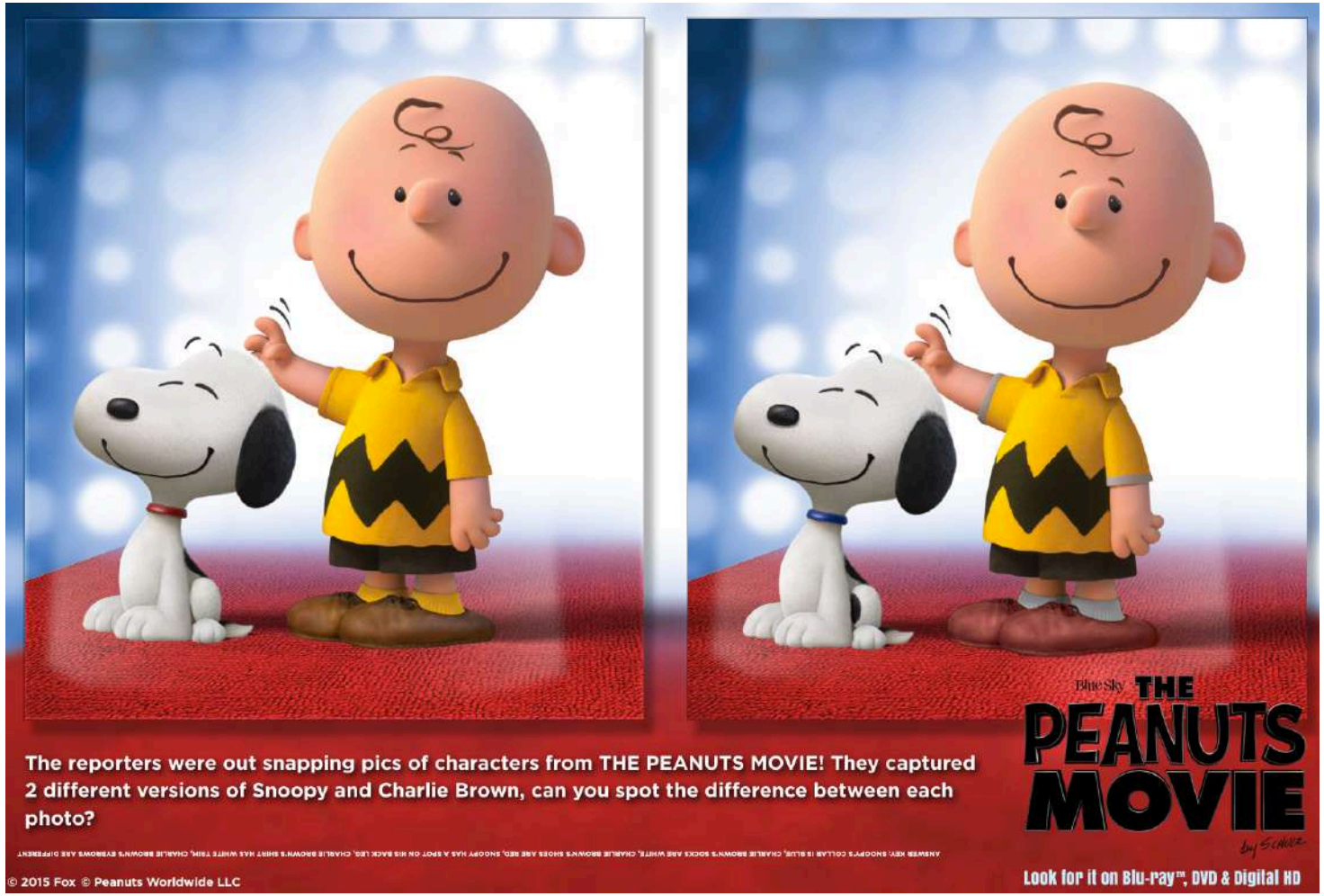
Reads
(50-500 letters each)

Assembly →

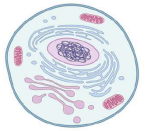


Genome
(3.000.000 letters)

Depending on nature of data, assembly can be different (wrong or?)



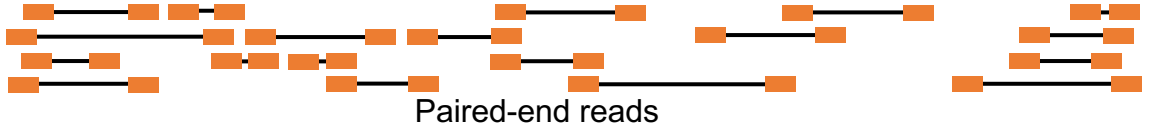
Assembly



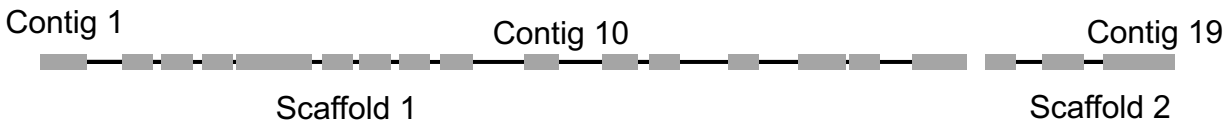
Genome

Fragment

Sequence



Assemble

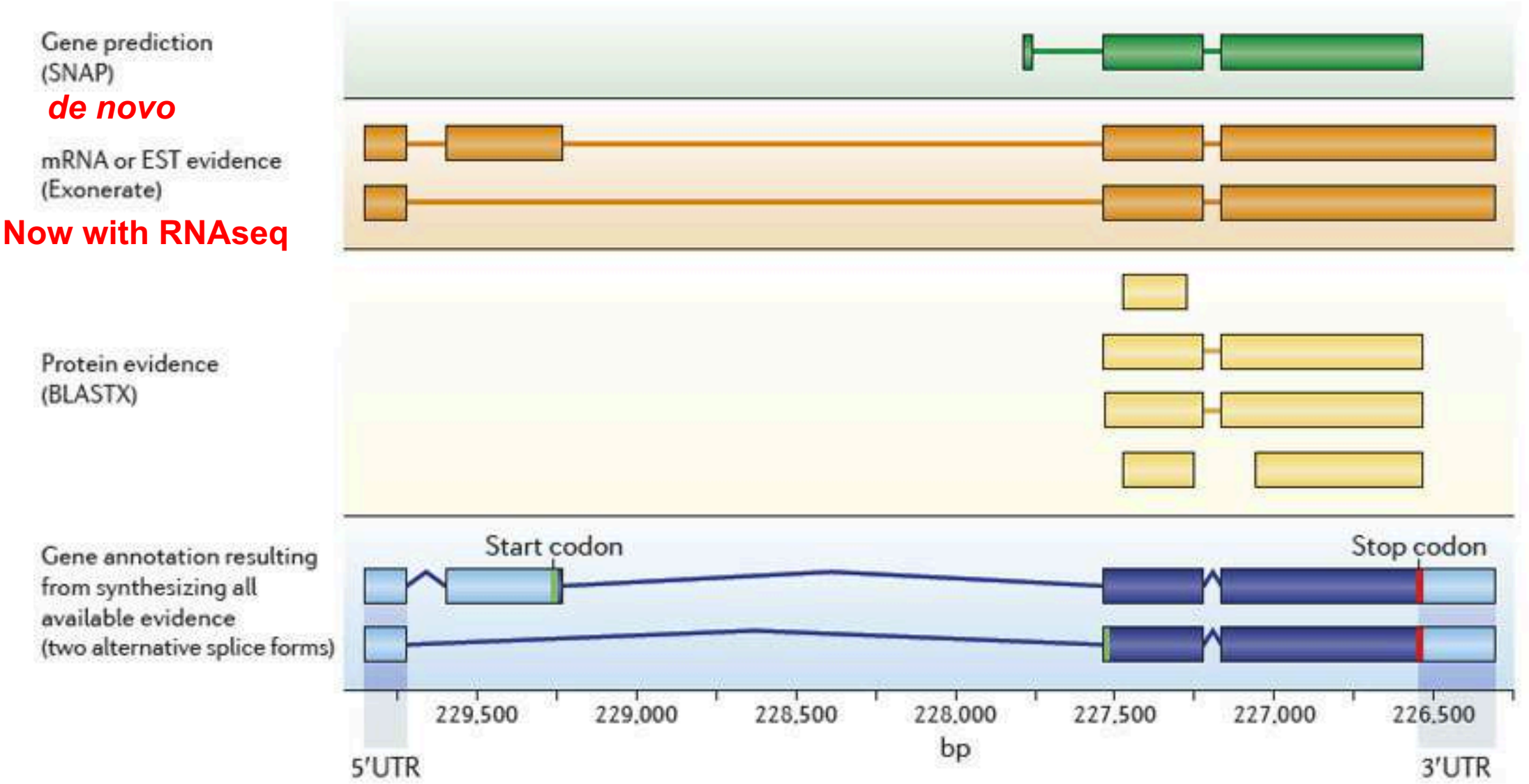


After assembly

- Say you have an assembly with 200 contigs and 34 scaffolds. What do you do next?

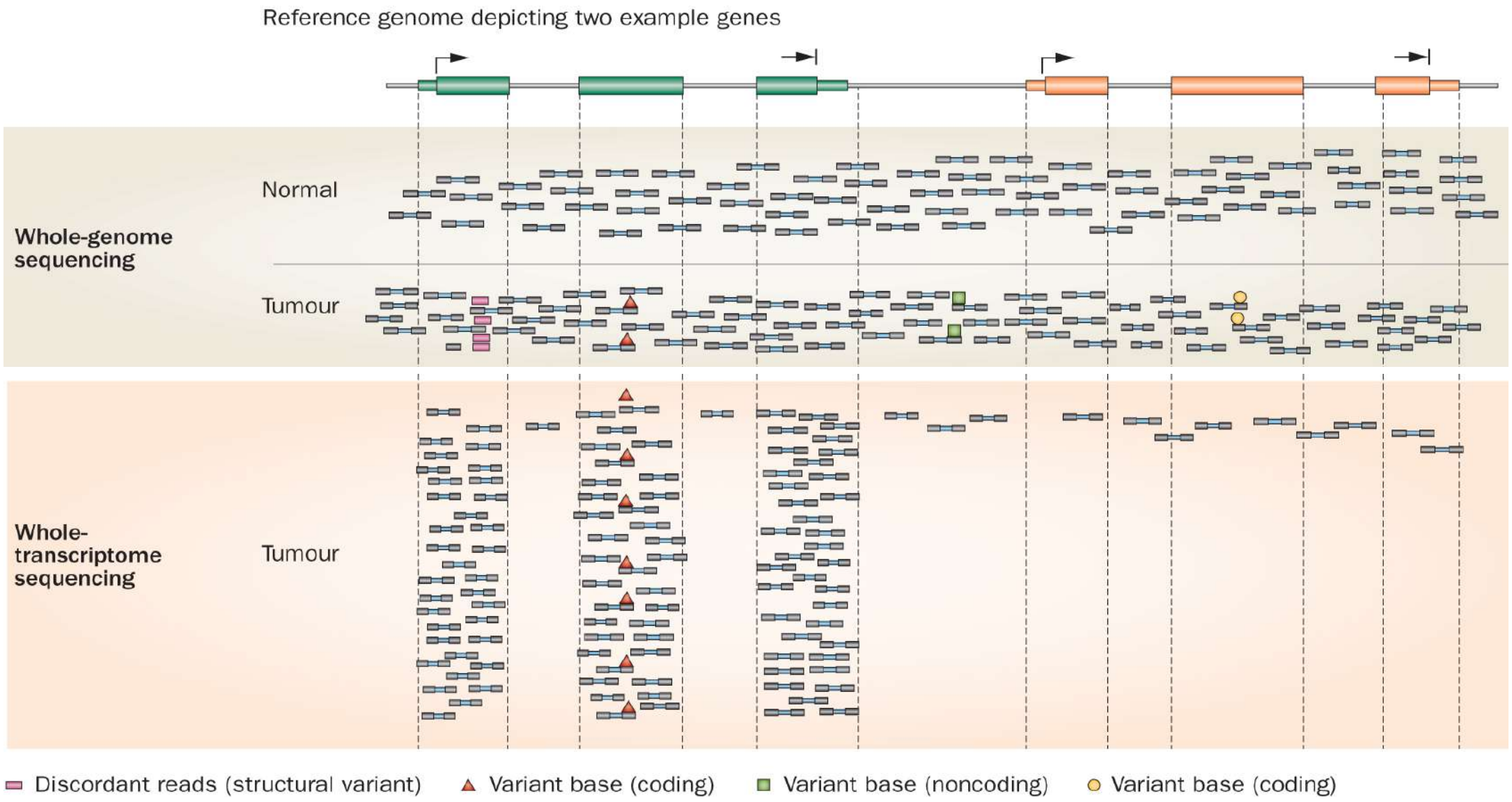
- How accurate is it?
- Have you tried different assemblers?
- Can you improve with additional data or diminishing returns?
- Is there contamination?
- How does it compare to other species?

Annotation



Yandell and Ence Nature Genetics Review (2012)

Mapping



doi:10.1038/nrgastro.2012.126

How?

Brute force comparison
Smith-Waterman
Suffix Tree
Burrows-Wheeler Transform

Brute force (is there a better one?)

1.

2.

3.

4.

Credit: Mike Zody

Read length matters in sequencing

Figure 5. Two copies of a repeat along a genome. The reads colored in red and those colored in yellow appear identical to the assembly program.

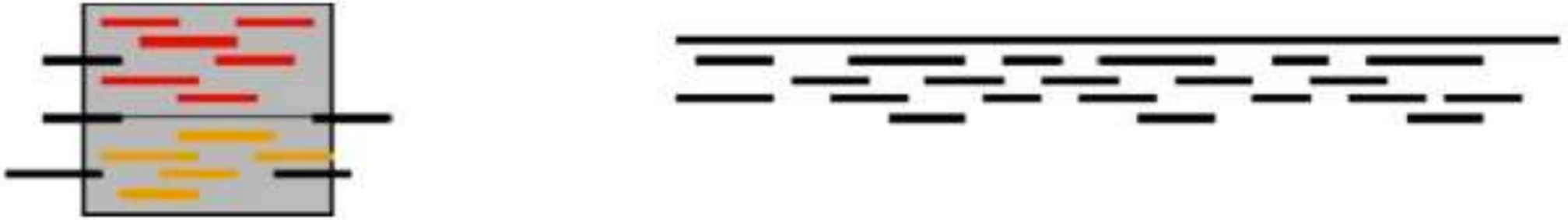
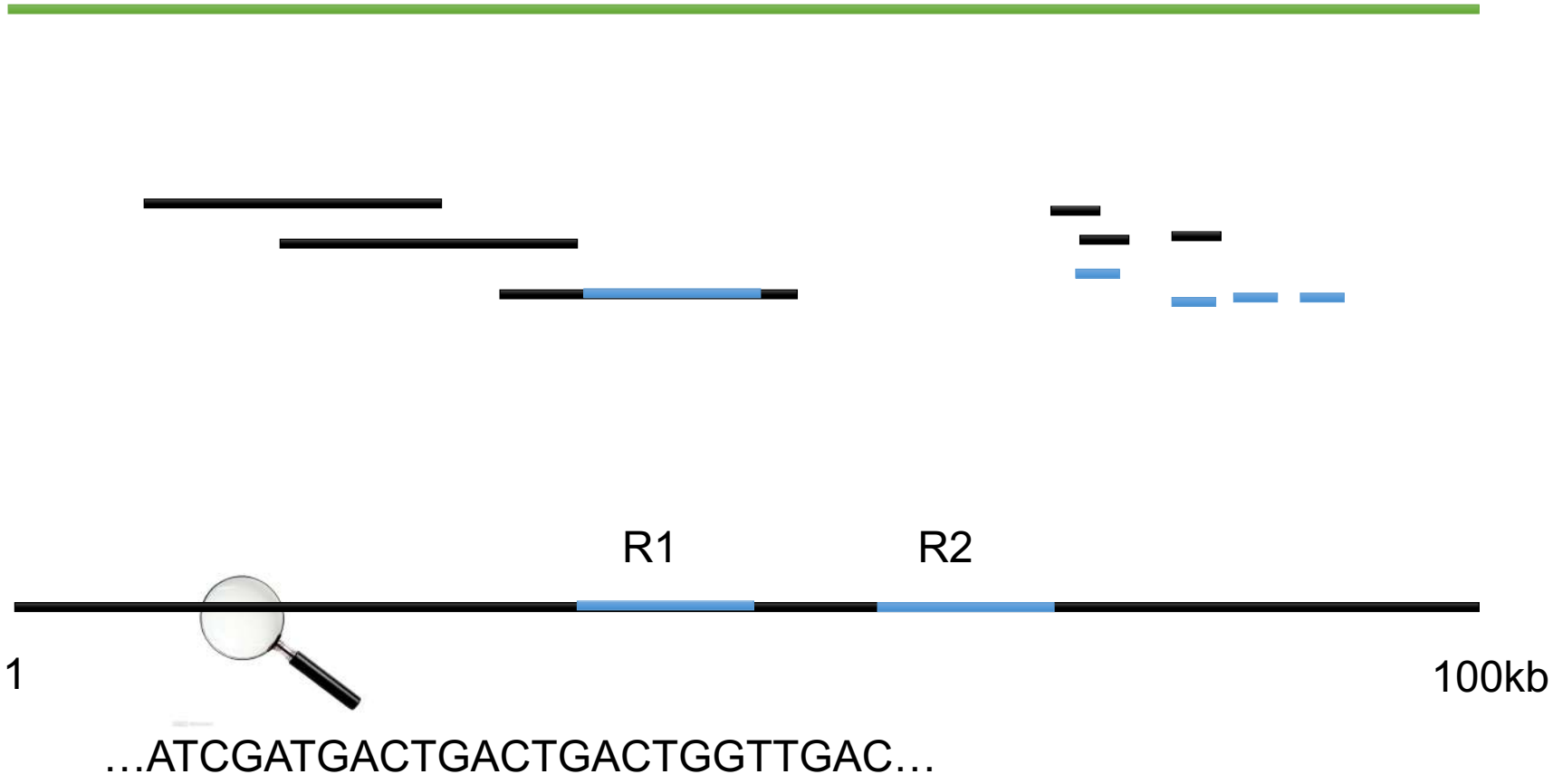


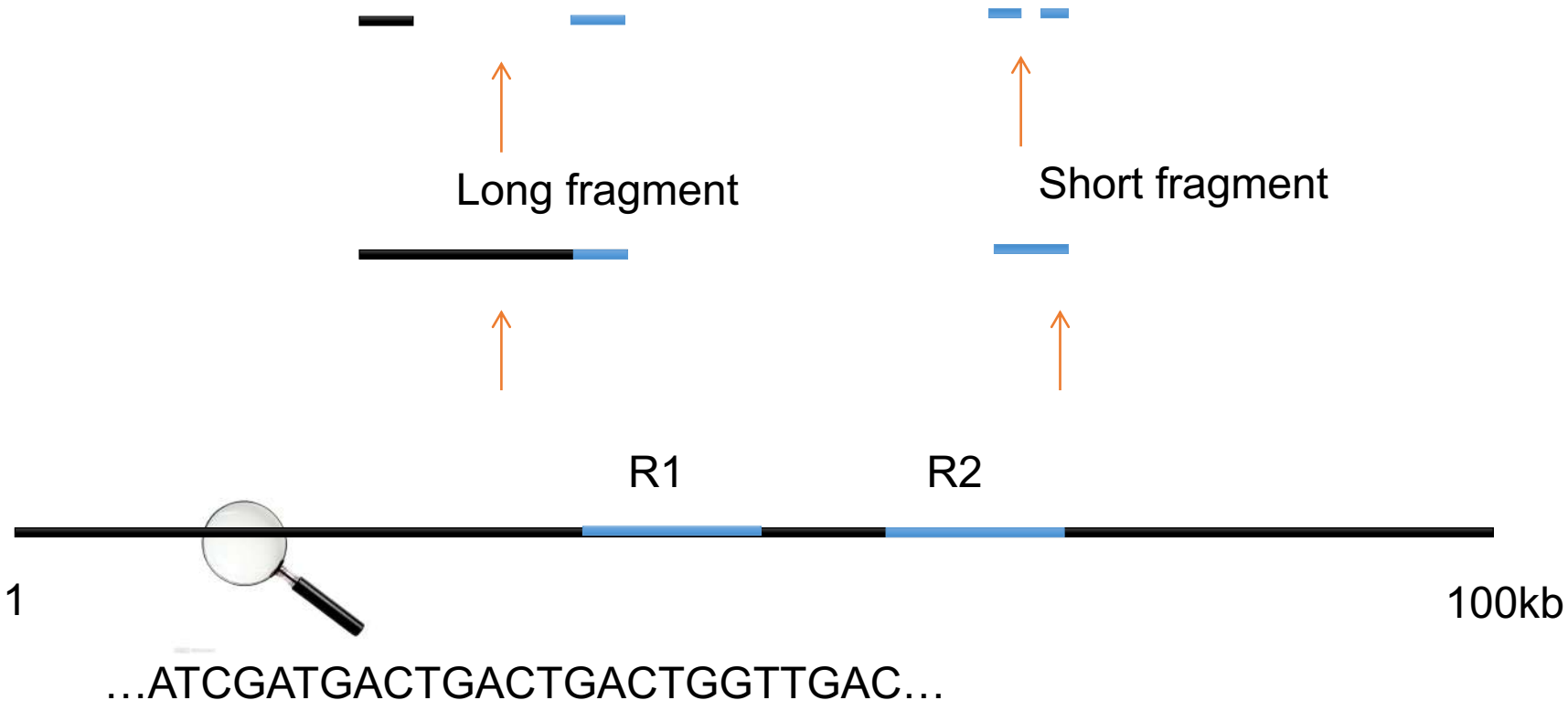
Figure 6. Genome mis-assembled due to a repeat. The assembly program incorrectly combined the reads from the two copies of the repeat leading to the creation of two separate contigs

https://www.cbcb.umd.edu/research/assembly_primer

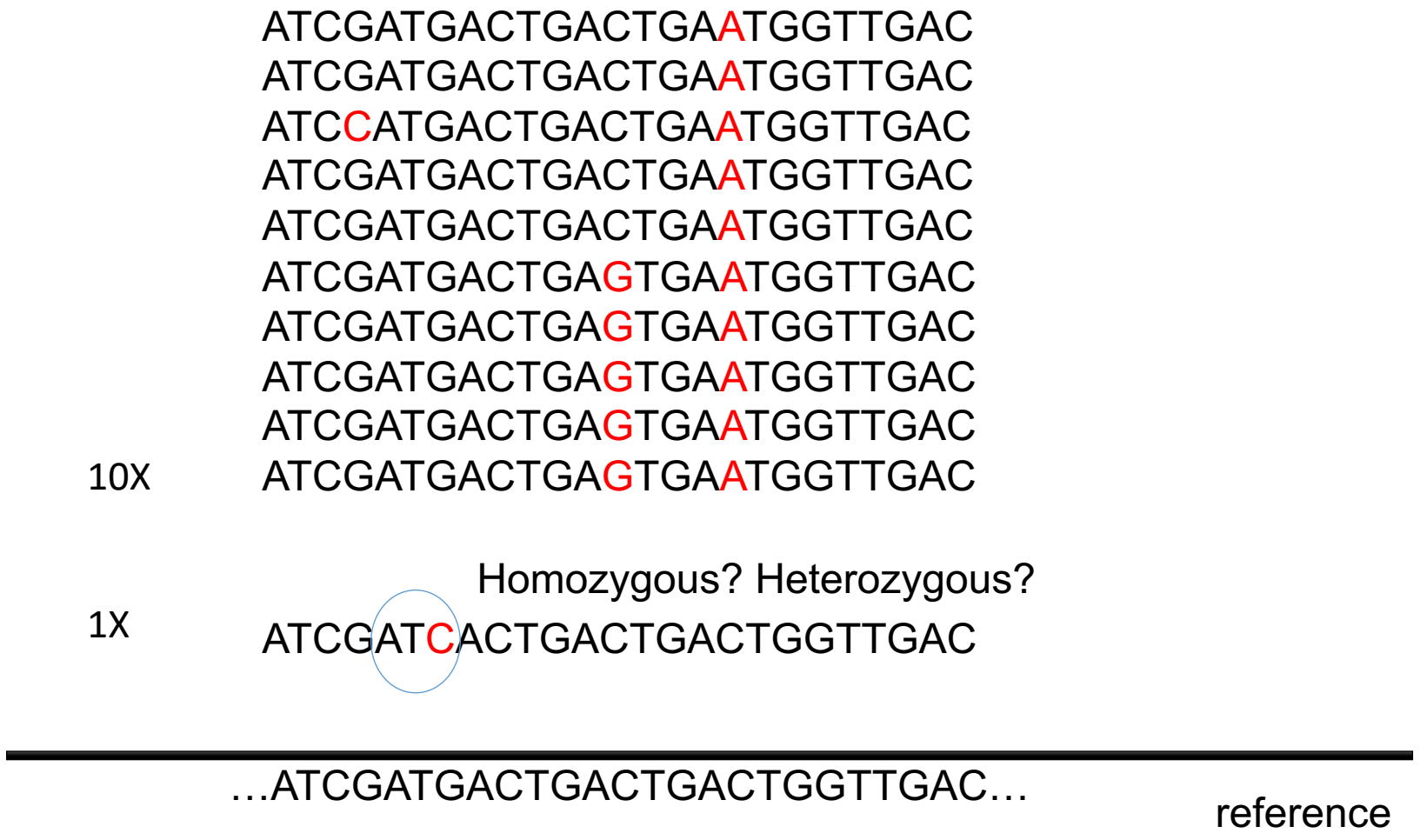
Read length matters in sequencing



Paired end and insert size matter in sequencing

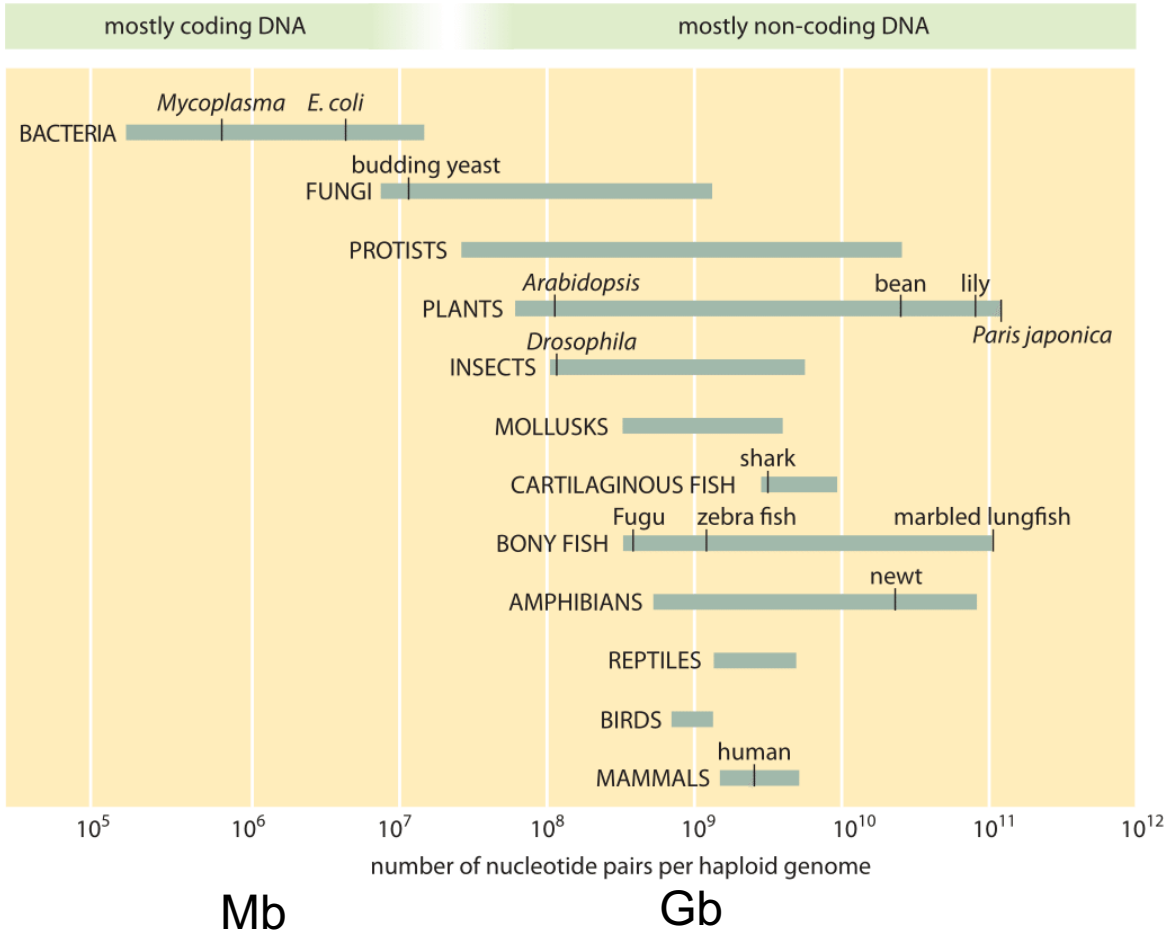


Depth matters in sequencing



Current perspective and challenges

現狀與挑戰



organism	genome size (base pairs)	protein coding genes	number of chromosomes
model organisms			
model bacteria *E. coli*	4.6 Mbp	4,300	1
budding yeast *S. cerevisiae*	12 Mbp	6,600	16
fission yeast *S. pombe*	13 Mbp	4,800	3
amoeba *D. discoideum*	34 Mbp	13,000	6
nematode *C. elegans*	100 Mbp	20,000	12 (2n)
fruit fly *D. melanogaster*	140 Mbp	14,000	8 (2n)
model plant *A. thaliana*	140 Mbp	27,000	10 (2n)
moss *P. patens*	510 Mbp	28,000	27
mouse *M. musculus*	2.8 Gbp	20,000	40 (2n)
human *H. sapiens*	3.2 Gbp	21,000	46 (2n)
viruses			
hepatitis D virus (smallest known animal RNA virus)	1.7 Kb	1	ssRNA
HIV-1	9.7 kbp	9	2 ssRNA (2n)
influenza A	14 kbp	11	8 ssRNA
bacteriophage λ	49 kbp	66	1 dsDNA
Pandoravirus salinus (largest known viral genome)	2.8 Mbp	2500	1 dsDNA
organelles			
mitochondria - *H. sapiens*	16.8 kbp	13 (+22 tRNA +2 rRNA)	1
mitochondria – *S. cerevisiae*	86 kbp	8	1
chloroplast – *A. thaliana*	150 kbp	100	1
bacteria			
C. ruddii (smallest genome of an endosymbiont bacteria)	160 kbp	182	1
M. genitalium (smallest genome of a free living bacteria)	580 kbp	470	1
H. pylori	1.7 Mbp	1,600	1
Cyanobacteria S. elongatus	2.7 Mbp	3,000	1
methicillin-resistant *S. aureus* (MRSA)	2.9 Mbp	2,700	1
B. subtilis	4.3 Mbp	4,100	1
S. cellulosum (largest known bacterial genome)	13 Mbp	9,400	1

http://book.bionumbers.org/how-big-are-genomes/

Why sequence a genome?

- Differences between species (comparative genomics)
- Variations between individuals (population genetics)
- Of economic, agricultural, medical, ecology values

- **Help to understand biology**

Case studies

Classical genetics

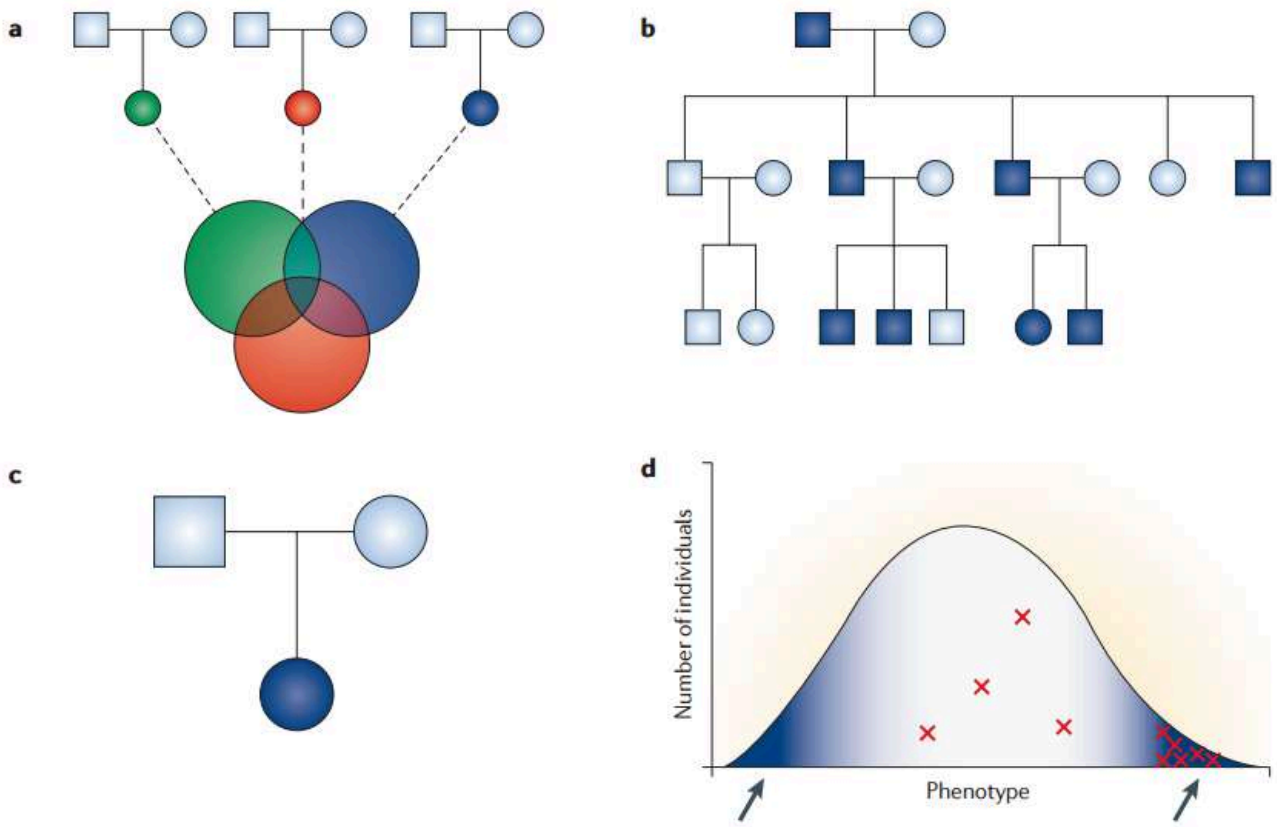
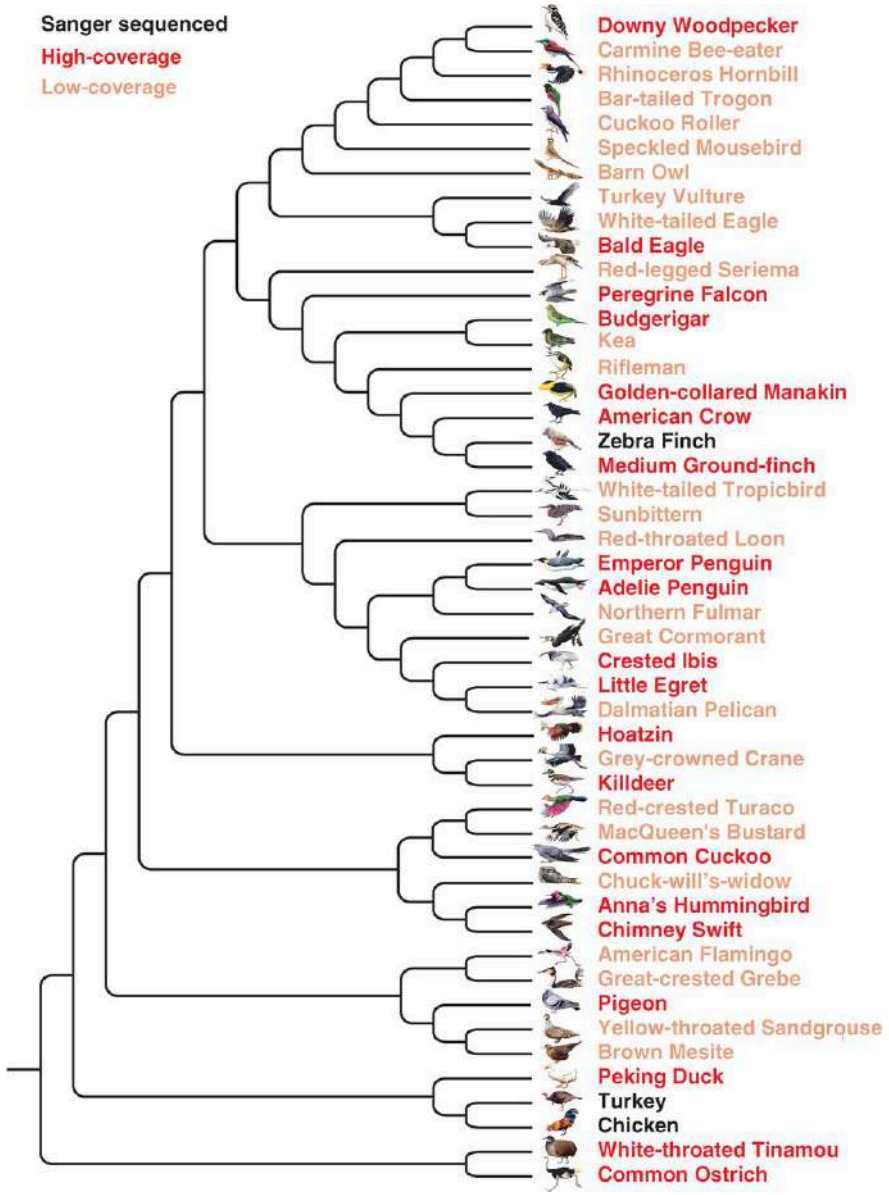


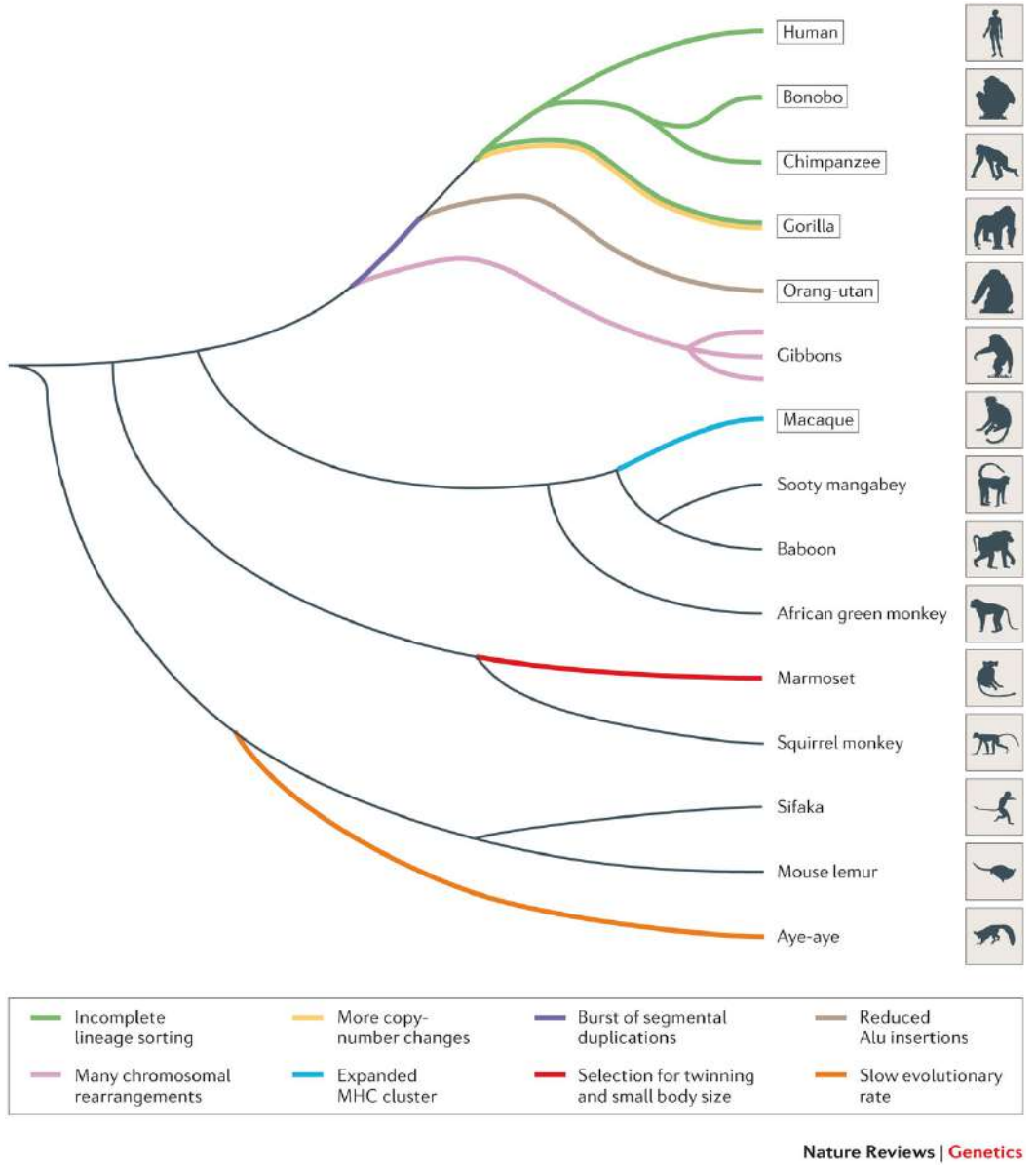
Figure 2 | **Strategies for finding disease-causing rare variants using exome sequencing.** Four main strategies are illustrated. **a** | Sequencing and filtering across multiple unrelated, affected individuals (indicated by the three coloured circles). This approach is used to identify novel variants in the same gene (or genes), as indicated by the shaded region that is shared by the three individuals in this example. **b** | Sequencing and filtering among multiple affected individuals from within a pedigree (shaded circles and squares) to identify a gene (or genes) with a novel variant in a shared region of the genome. **c** | Sequencing parent–child trios for identifying *de novo* mutations. **d** | Sampling and comparing the extremes of the distribution (arrows) for a quantitative phenotype. As shown in panel **d**, individuals with rare variants in the same gene (red crosses) are concentrated in one extreme of the distribution.

http://www.nature.com/nrg/journal/v12/n11/pdf/nrg3031.pdf

Comparative genomics / Phylogenomics



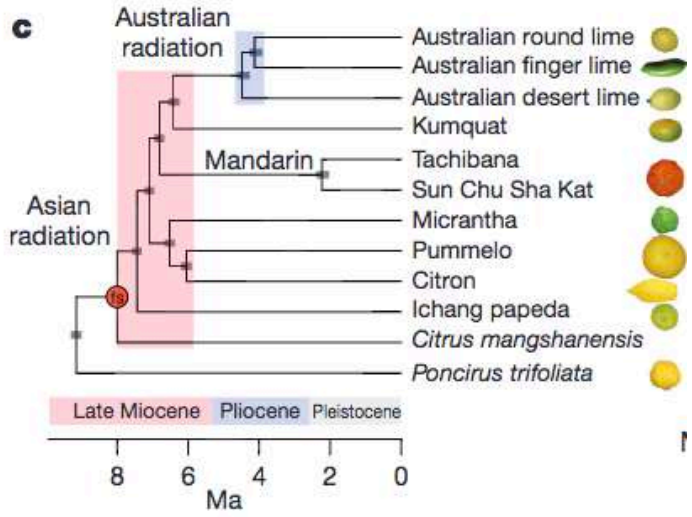
Guojie Zhang et al. Science (2014)

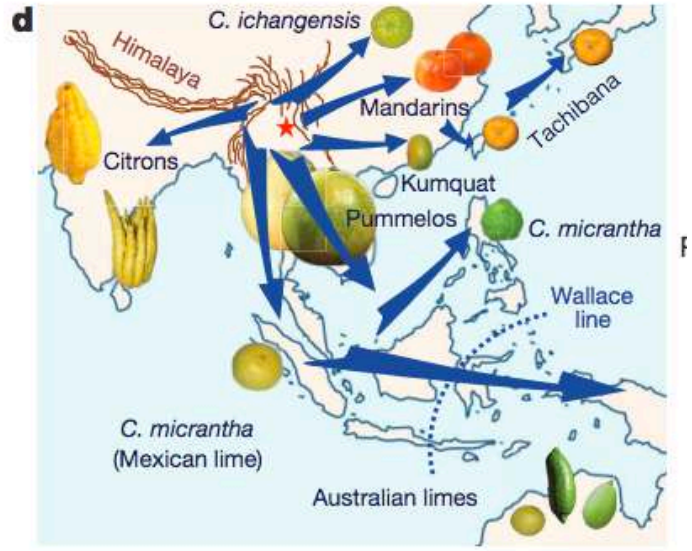


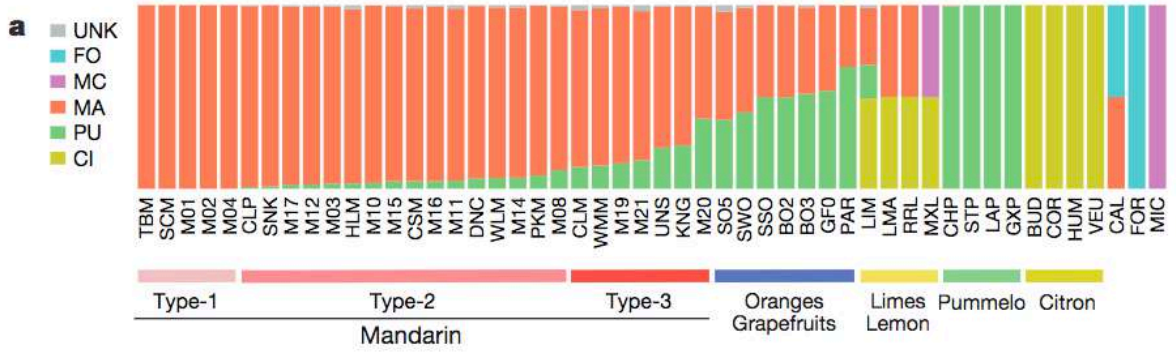
Roger & Gibbs Nature Reviews Genetics (2014)

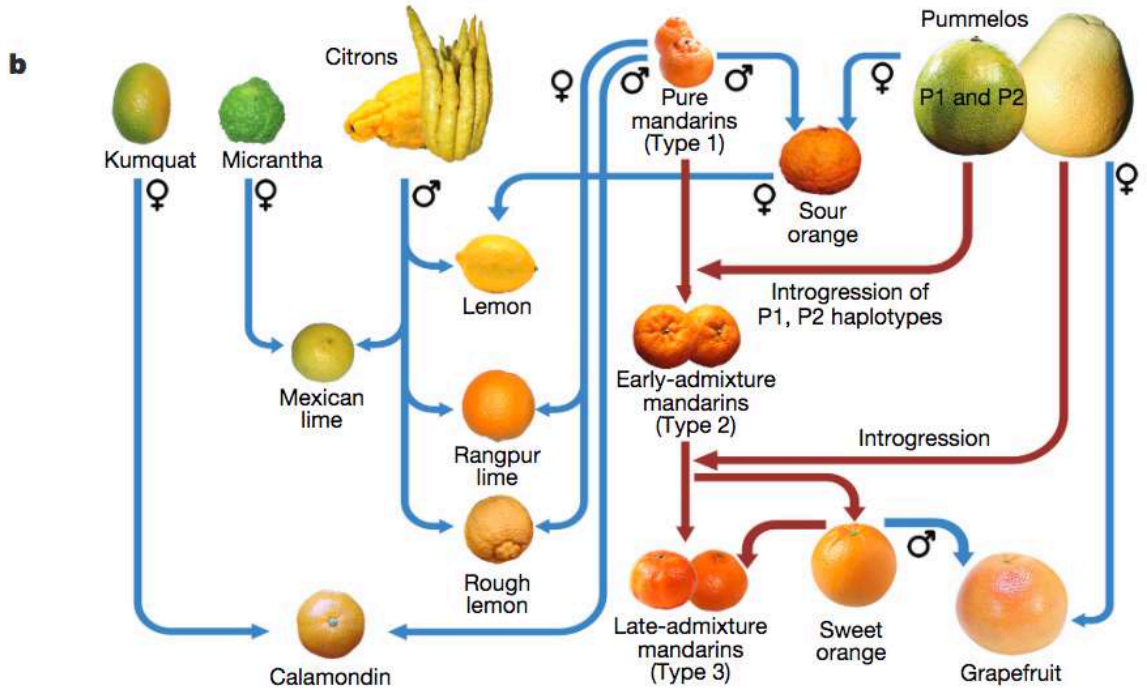
Comparative genomics

Genomics of the origin and evolution of Citrus









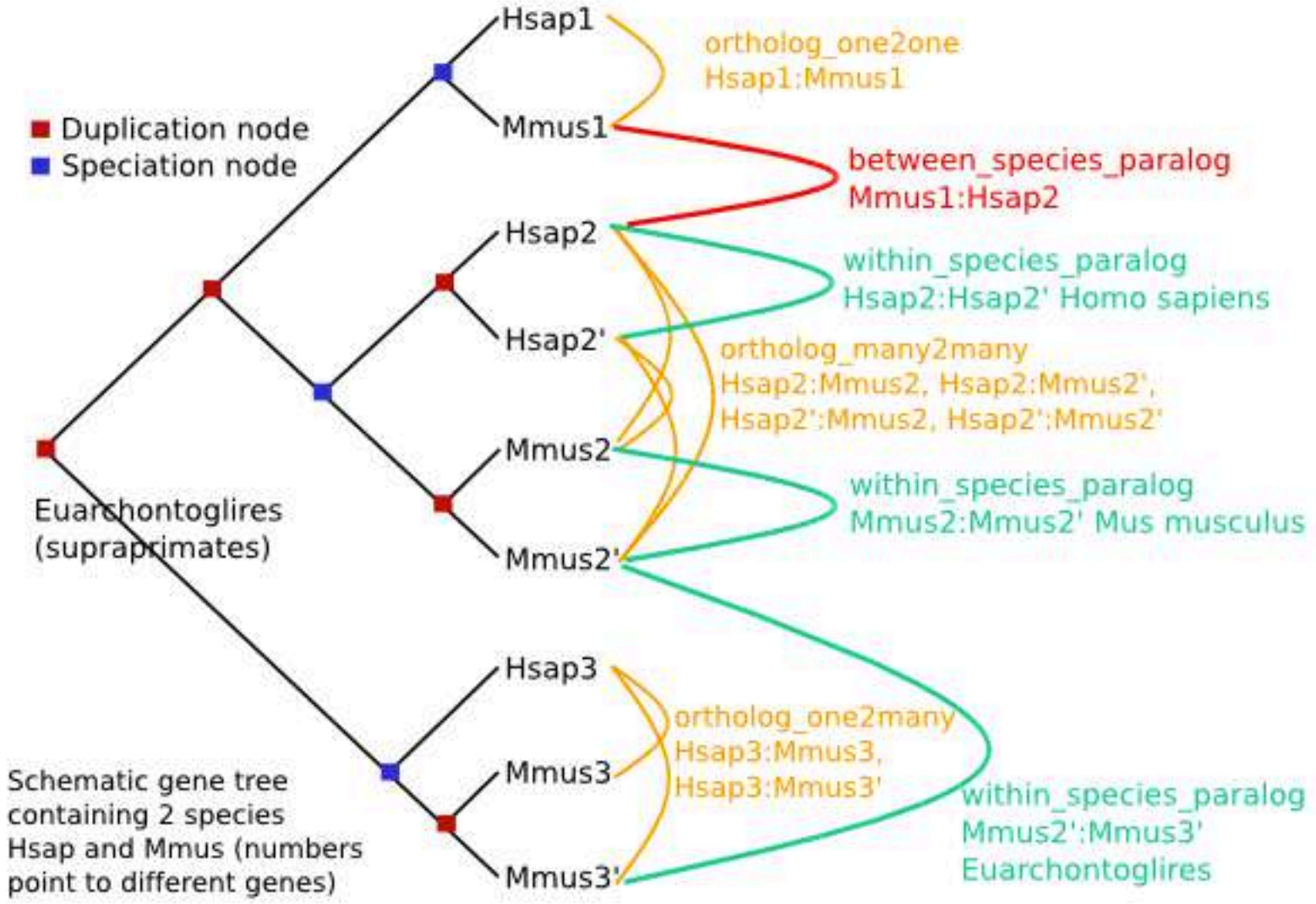
Wu *et al.*, Nature (2018)

Homologs: Orthologs and paralogs

Genes in different species and related by a speciation event are defined as **orthologs**.

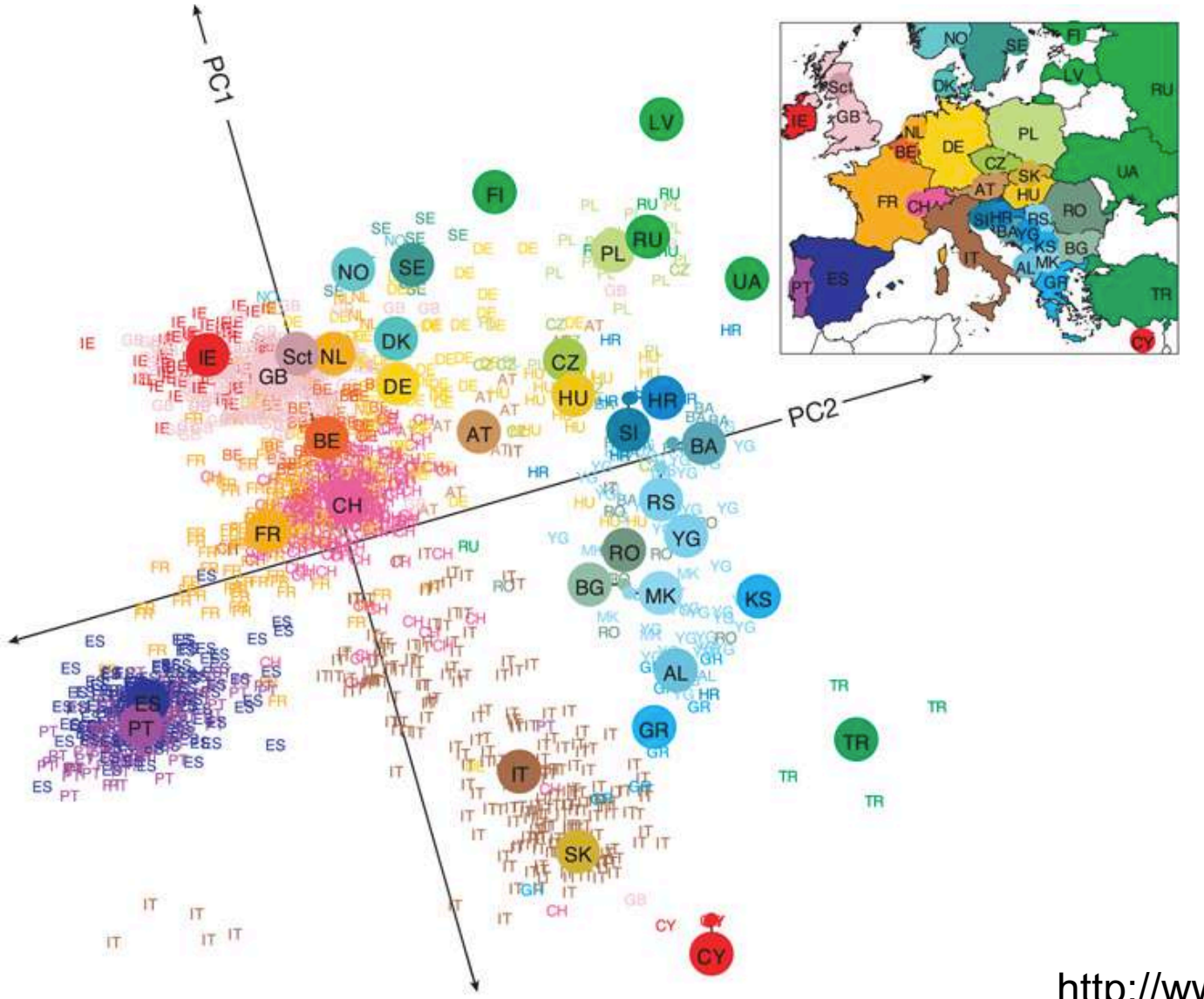
Depending on the number of genes found in each species, we differentiate among 1:1, 1:many and many:many relationships.

Genes of the same species and related by a duplication event are defined as **paralogs**.

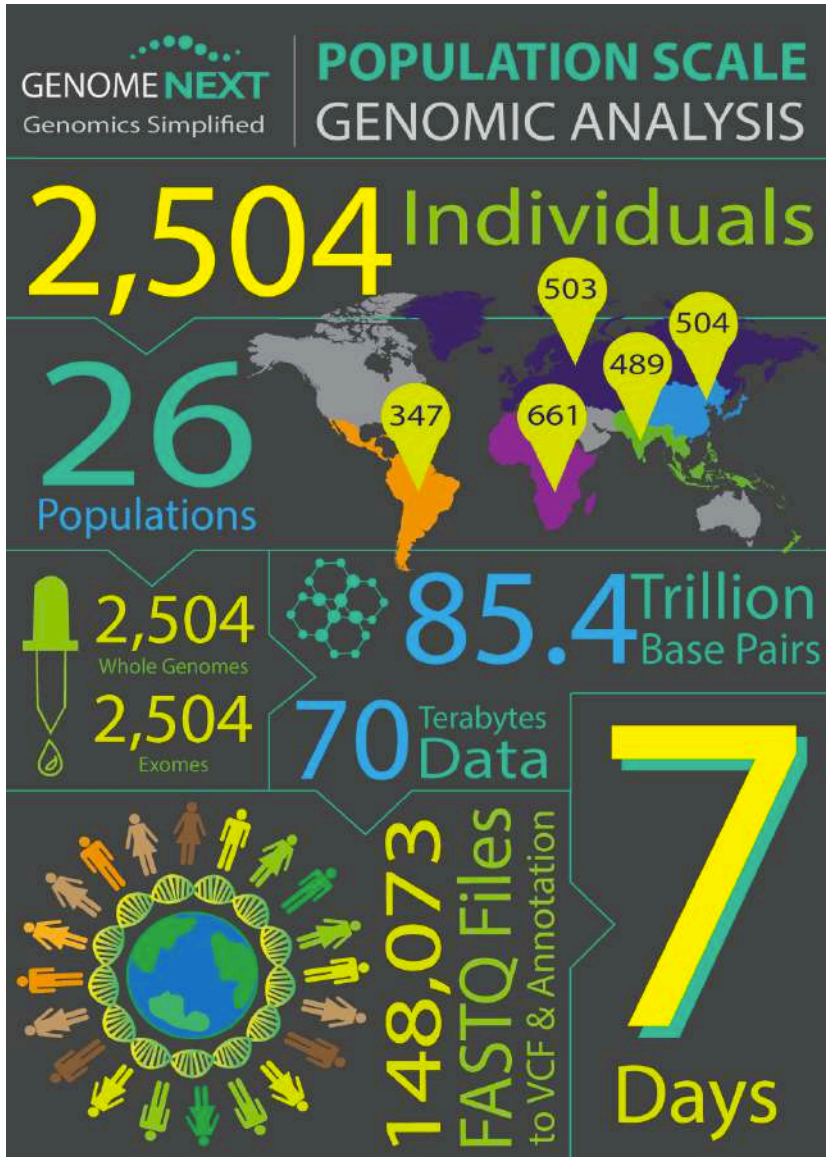


http://asia.ensembl.org/info/genome/compara/homology_method.html?redirect=no

Population genomics

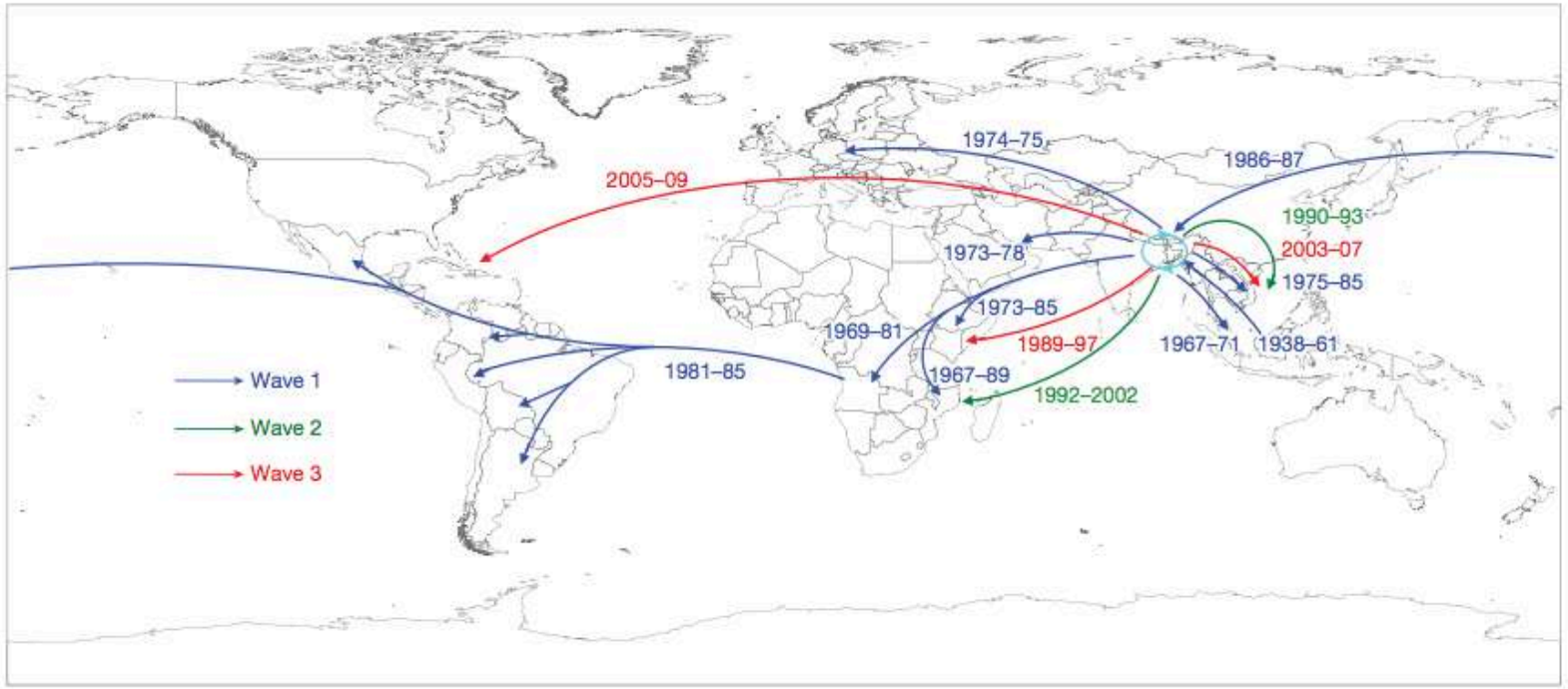


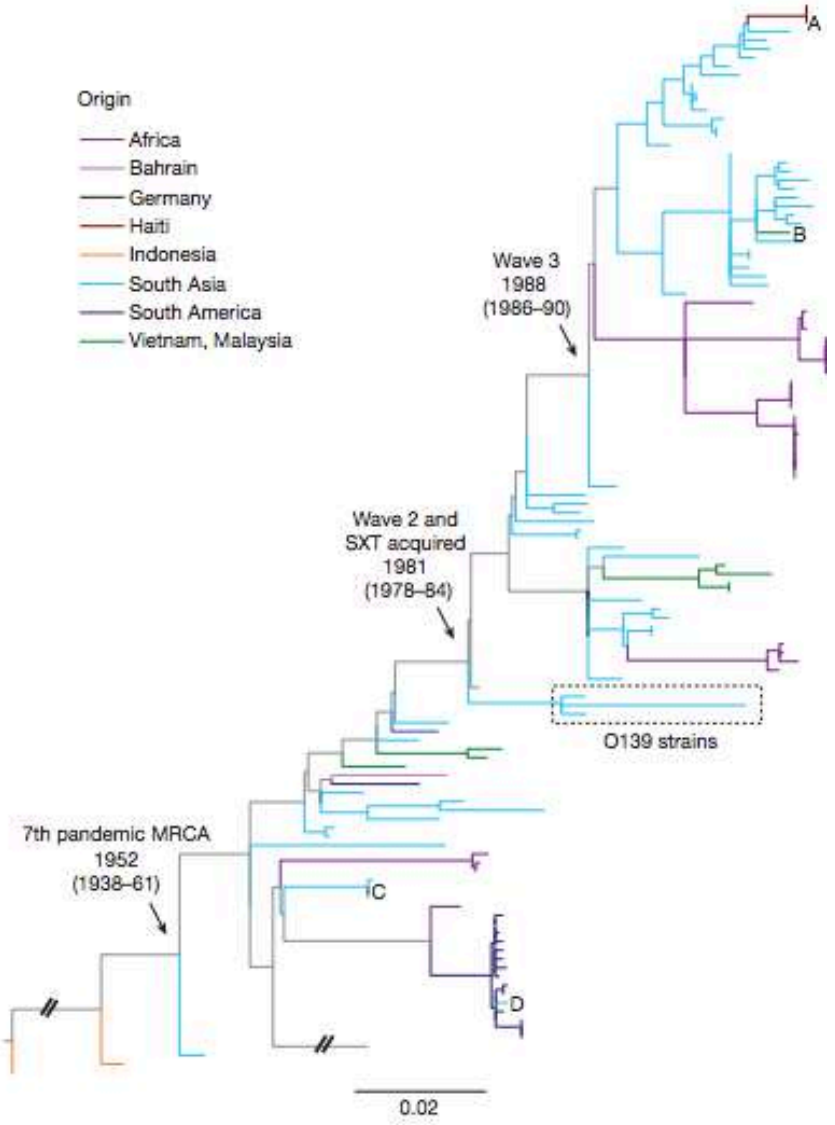
Novembre et al Nature (2008)



http://www.genomenext.com/casestudies_post/population-scale-analysis-genomic-samples-analyzed-from-2504-individuals-in-1-week/

Population genomics





Mutreja *et al.*, Nature Genetics (2011)

2013

8000家庭破碎 聯合國遭控傳染霍亂

2013-10-11 by：阿咖 5725

一直以來，聯合國給世人的印象多是促進世界永續發展的正面印象，但對海地居民來說，聯合國卻成了當地人最恐懼的劊子手，最新報導就指出，因聯合國駐軍而散佈的霍亂已經造成8,000人死亡。

BBC綜合報導，聯合過派駐的維和部隊(UN peacekeepers)意外將細菌帶到海地境內，在當地造成霍亂大流行，自2010年爆發至今，霍亂已經在海地造成8,000人病死，這也讓海地成為目前全世界霍亂疫病最嚴重的地區。

聯合國是兇手

儘管許多調查指出聯合國就是霍亂源頭，但海地數度請願要求補償未果，現在海地的代表律師團就上訴紐約法院，控告聯合國是造成海地霍亂疫情的元凶。

2016

聯合國坦承：我們將霍亂帶進了海地

2016-08-19 by：泥仔 15040

將近六年的時間，聯合國終於承認海地的霍亂疫情與他們有關。到目前為止，已經有數十萬名居民感染上霍亂、一萬名海地人因霍亂而去世。

維和部隊惹的禍？

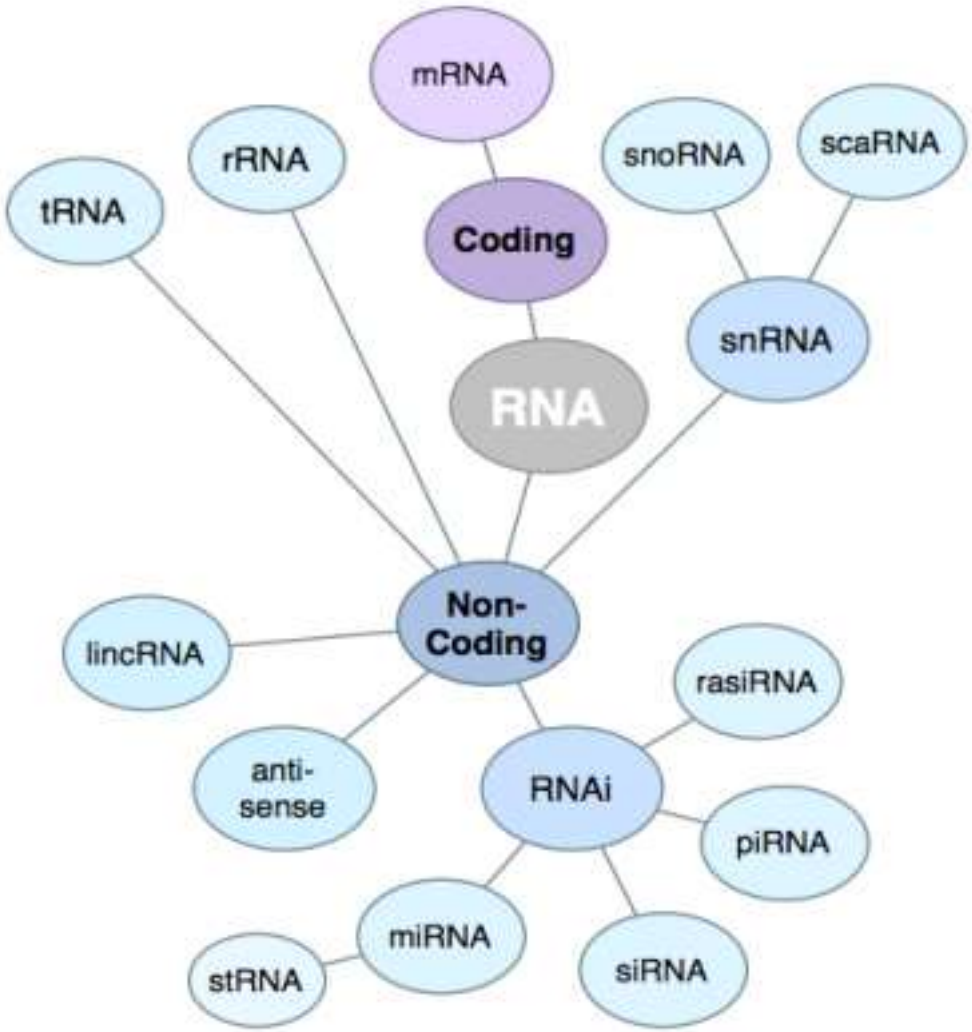
由於海地過去都沒有類似霍亂症狀的疾病，部分專家也發現海地的霍亂細菌種類與尼泊爾的種類是一樣的，因此懷疑是聯合國在尼泊爾的維和部隊將霍亂弧菌帶進海地。但將近六年來，聯合國一直都否認這樣的指控。

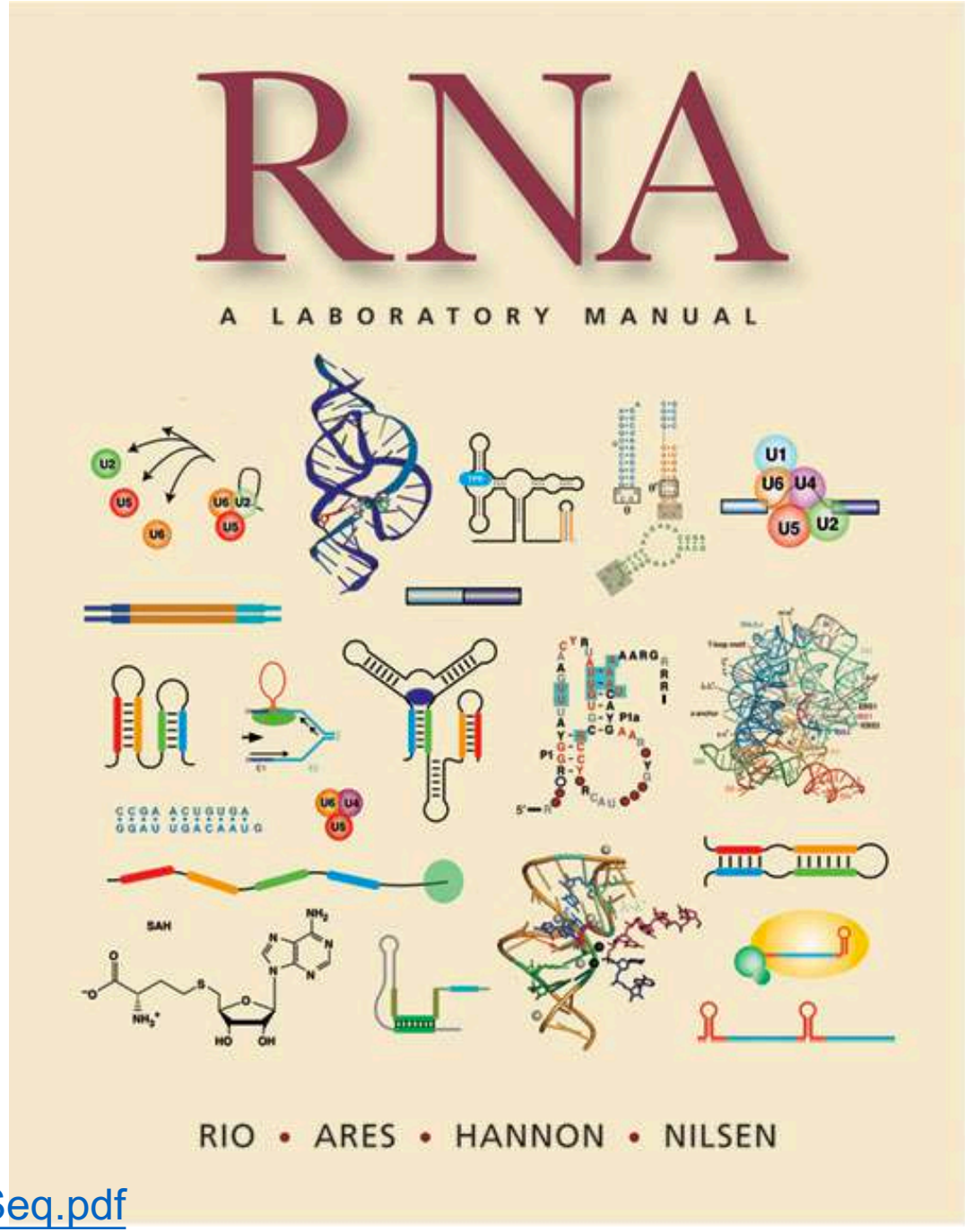
聯合國坦承與疫情爆發有關

在本周三(17)，聯合國副發言人哈奇(Farhan Haq)聲明：「過去幾年來，聯合國有鑑於海地初期的瘟疫爆發與我們有些關係，聯合國決定要多做些什麼。」他也強調聯合國會在接下來兩個月內有所行動。

Transcriptomics / RNAseq

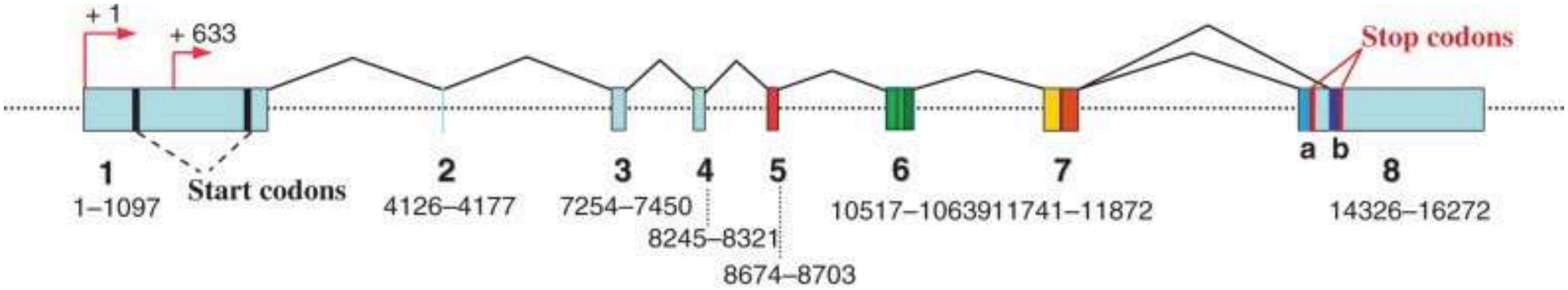
Types of RNA

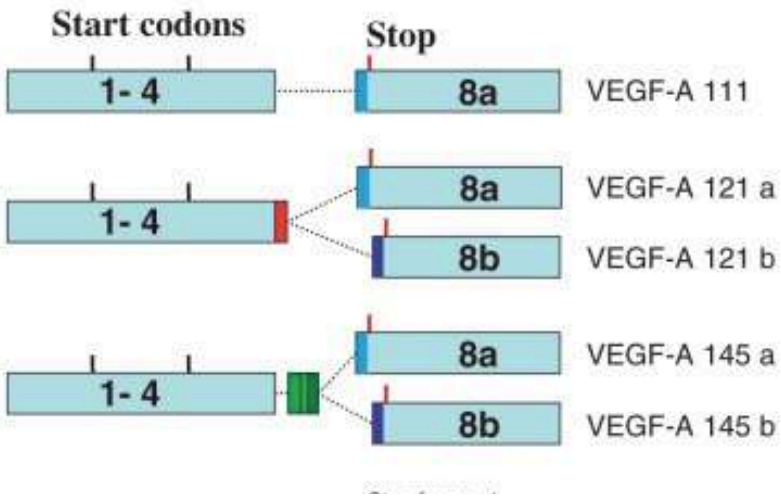


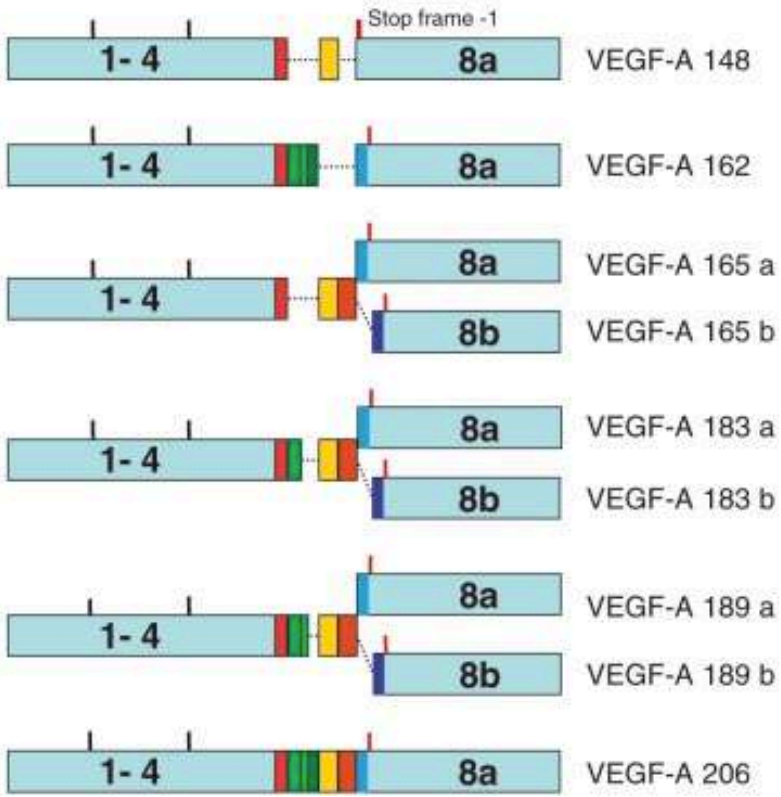


http://jura.wi.mit.edu/bio/education/hot_topics/RNAseq/RNA_Seq.pdf

Gene and isoforms

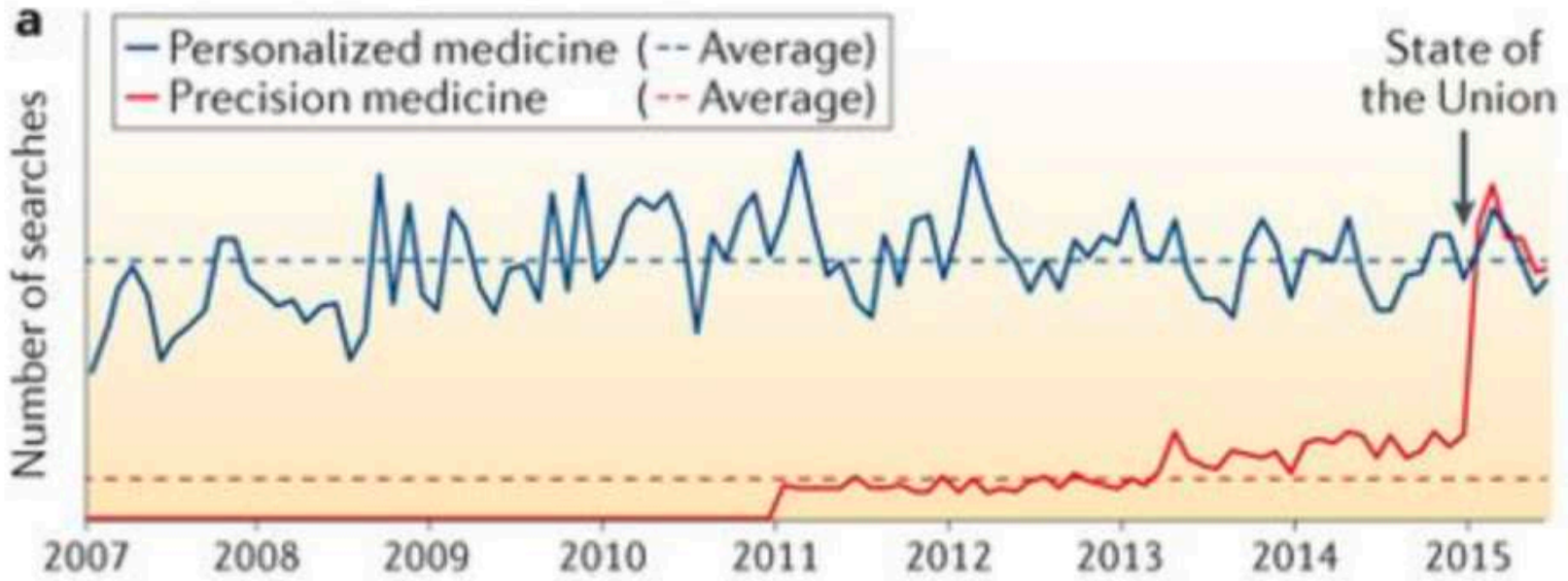






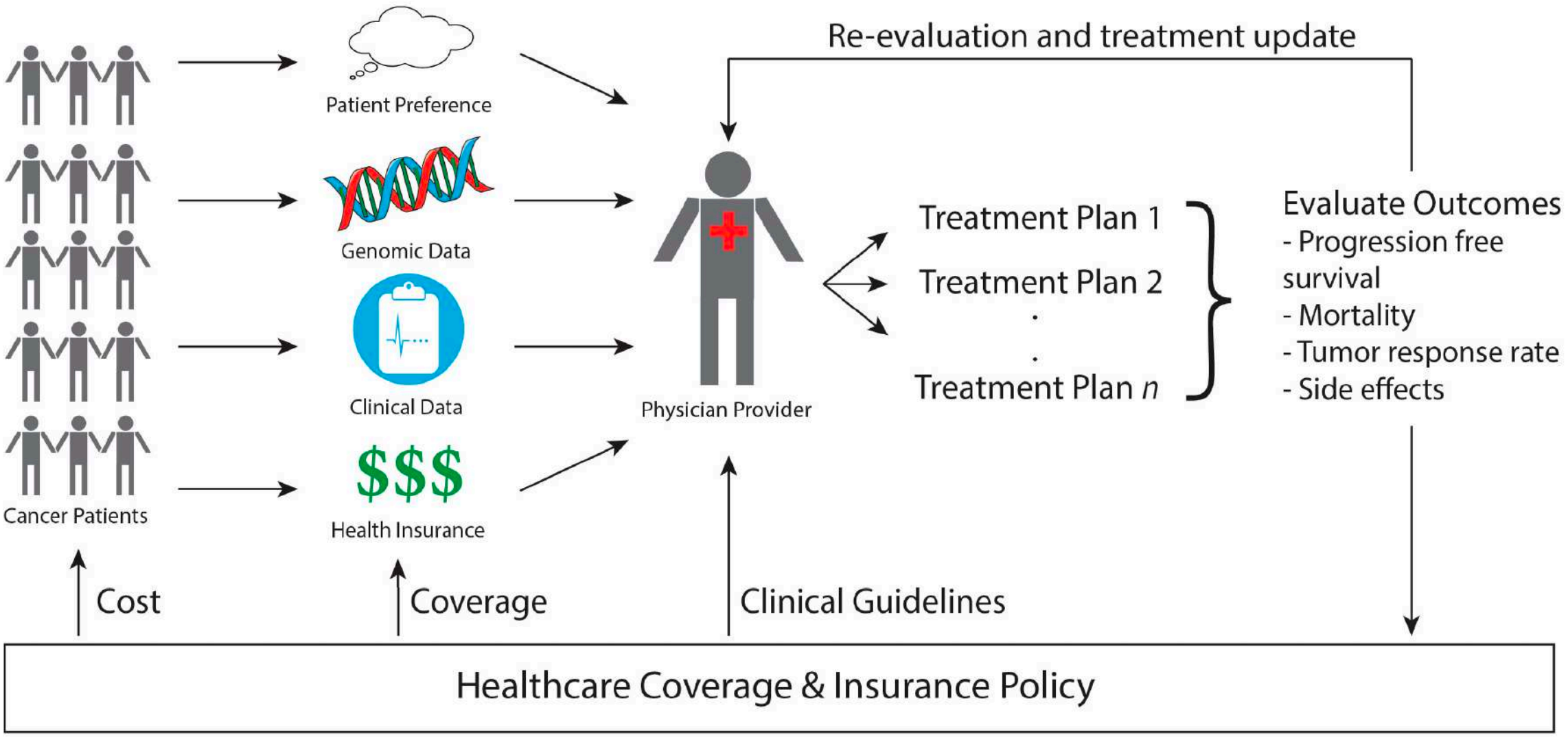
Arcondeguy et al (2015)

Precision medicine
精準醫學



Ashley (2016) Nature Review Genetics

Outline of precision medicine



Morash *et al* (2018) Journal of Personalized Medicine

Summary of outcomes in Oncology PM Studies

Study	Sample Size	Most Prevalent Tumor Types	Outcomes Reported
Tsimberidou et al. *Clin. Cancer Res.* **2012** [5]	291 patients with one molecular aberration (175 treated with matched therapy, 116 control)	Colorectal, melanoma, lung, ovarian	Matched group had improved ORR (27% vs. 5%), TTF (median 5.2 vs. 2.2 month), OS (median 13.4 vs. 9.0 month)
Radovich et al. *Oncotarget* **2016** [6]	101 patients with sequencing and follow up (44 treated with matched therapy, 57 control)	Soft tissue sarcoma, breast, colorectal	Matched group had improved PFS (86 vs. 49 days)
Schwaederle et al. *Mol. Cancer Ther.* **2016** [7]	180 patients with sequencing and follow up (87 treated with matched therapy, 93 control)	Gastrointestinal, breast, brain	Matched group had improved PFS (4.0 vs. 3.0 month), TRR (34.5% vs. 16.1% achieving SD/PR/CR)
Kris et al. *JAMA* **2014** [8]	578 patients with oncogenic driver and followup (260 with matched therapy, 318 control)	Lung only	Matched group had improved survival (median 3.5 vs. 2.4 years)
Aisner et al. *J. Clin. Oncol.* **2016** [9]	187 patients with targetable alteration and follow up (112 with matched therapy, 74 control)	Lung only	Matched group had improved survival (median 2.8 vs. 1.5 years)
Stockley et al. *Genome Med.* **2016** [10]	245 patients with sequencing matched to clinical trials (84 on matched trial, 161 control)	Gynecological, lung, breast	Matched group had improved ORR (19% vs. 9%)
LeTourneau et al. *Lancet Oncol.* **2015** [11]	RCT with 195 patients with molecular aberration (99 treated with matched therapy, 96 control)	Gastrointestinal, breast, brain	No difference in PFS between groups

ORR = overall response rate, TTF = time to treatment failure, OS = overall survival, PFS = progression free survival, TRR = tumor response rate, SD = stable disease, PR = partial response, CR = complete response, RCT = randomized controlled trial. Matched group indicates patients matched to a therapy based on sequencing results.

Morash *et al* (2018) Journal of Personalized Medicine

FOCUS ON GENOMES OF ICELANDERS

ARTICLES

nature genetics

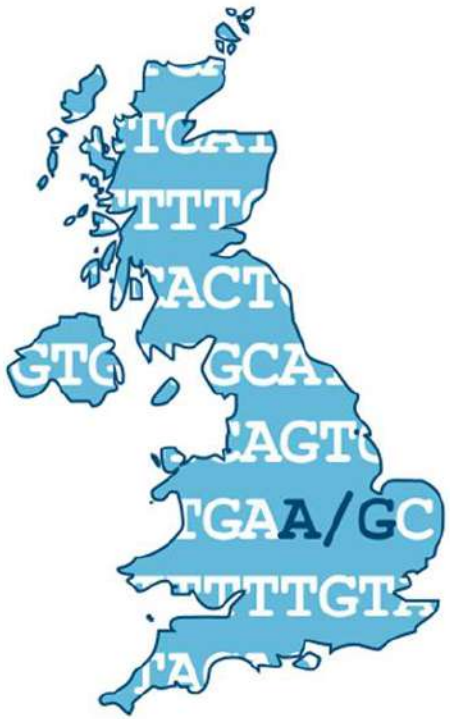
Large-scale whole-genome sequencing of the Icelandic population

Here we describe the insights gained from sequencing the whole genomes of 2,636 Icelanders to a median depth of 20×.

A collection of Icelandic genealogical records dating back to the 1700s.

The blood of a thousand Icelanders.
Photo: Chris Lund

Nature Genetics **volume 47**, pages 435–444 (2015)



UK 10K

RARE GENETIC VARIANTS IN HEALTH AND DISEASE

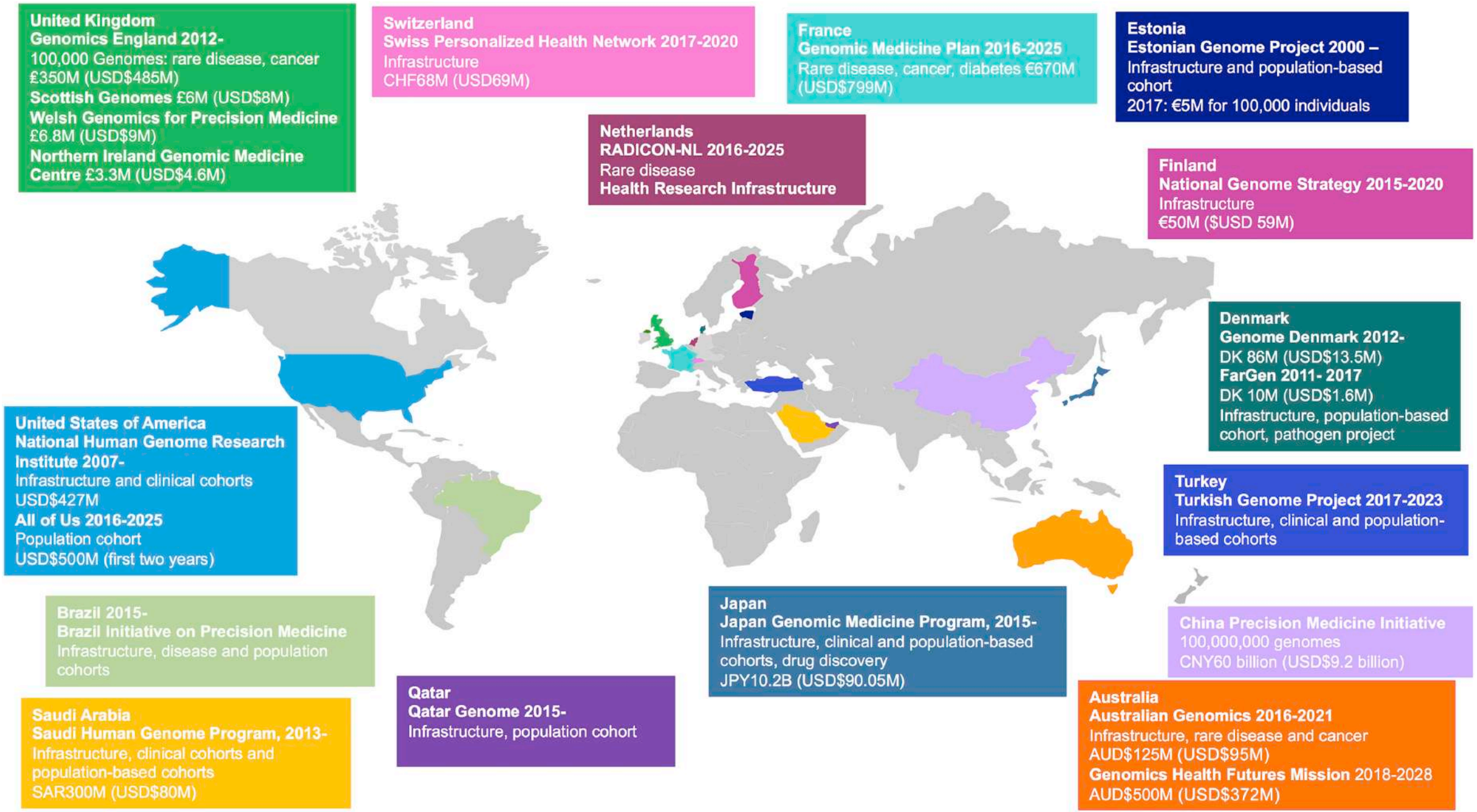
The project is taking a two-pronged approach to identify rare variants and their effects:

•by **studying and comparing the DNA of 4,000 people whose physical characteristics are well documented**, the project aims to identify those changes that have no discernible effect and those that may be linked to a particular disease;

•by **studying the changes within protein-coding areas of DNA that tell the body how to make proteins of 6,000 people with extreme health problems and comparing them with the first group**, it is hoped to find only those changes in DNA that are responsible for the particular health problems observed.

The project received a **£10.5 million** funding award from Wellcome in March 2010 and sequencing started in late 2010. For more information, please use the links on the right hand side.

https://www.uk10k.org/



Stark *et al* (2019) AJHG

臺灣
Taiwan Biobank

The Cumulative 累計收案數

統計至2019年01月31日止(請按此)

社區民眾收案數

109,059
參與個案總數

22,502
完成第一輪追蹤個案總數

醫學中心患者收案數

1,862
參與個案總數

320
完成第一輪追蹤個案總數

8
完成第二輪追蹤個案總數

The Cumulative 累計收案數

統計至2019年07月31日止(請按此)

社區民眾收案數

118,548
參與個案總數

24,936
完成第一輪追蹤個案總數

醫學中心患者收案數

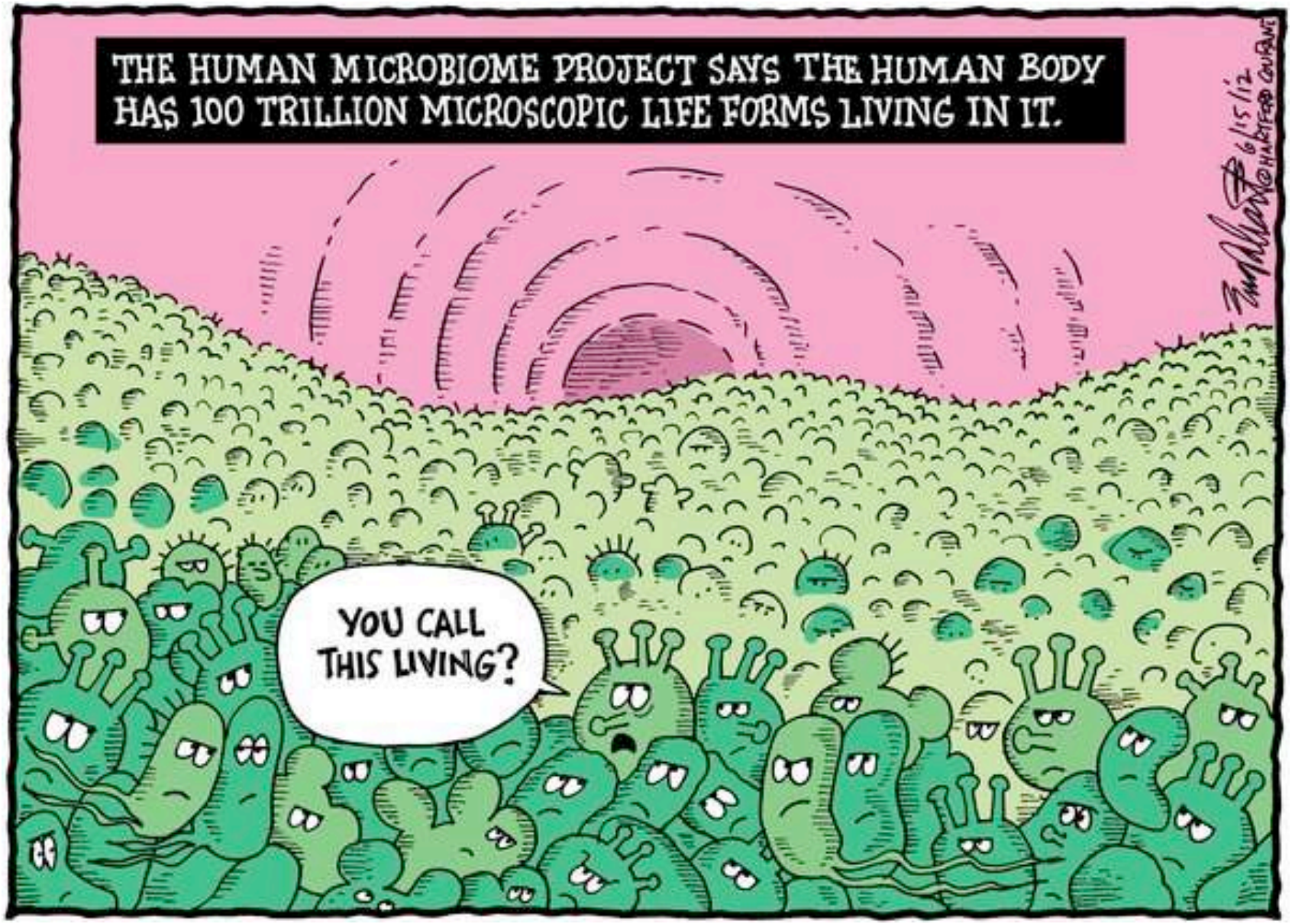
3,145
參與個案總數

659
完成第一輪追蹤個案總數

104
完成第二輪追蹤個案總數

https://www.twbiobank.org.tw/new_web/index.php

Human gut microbiome



Human gut microbiome

Vol 464|4 March 2010|doi:10.1038/nature08821 nature

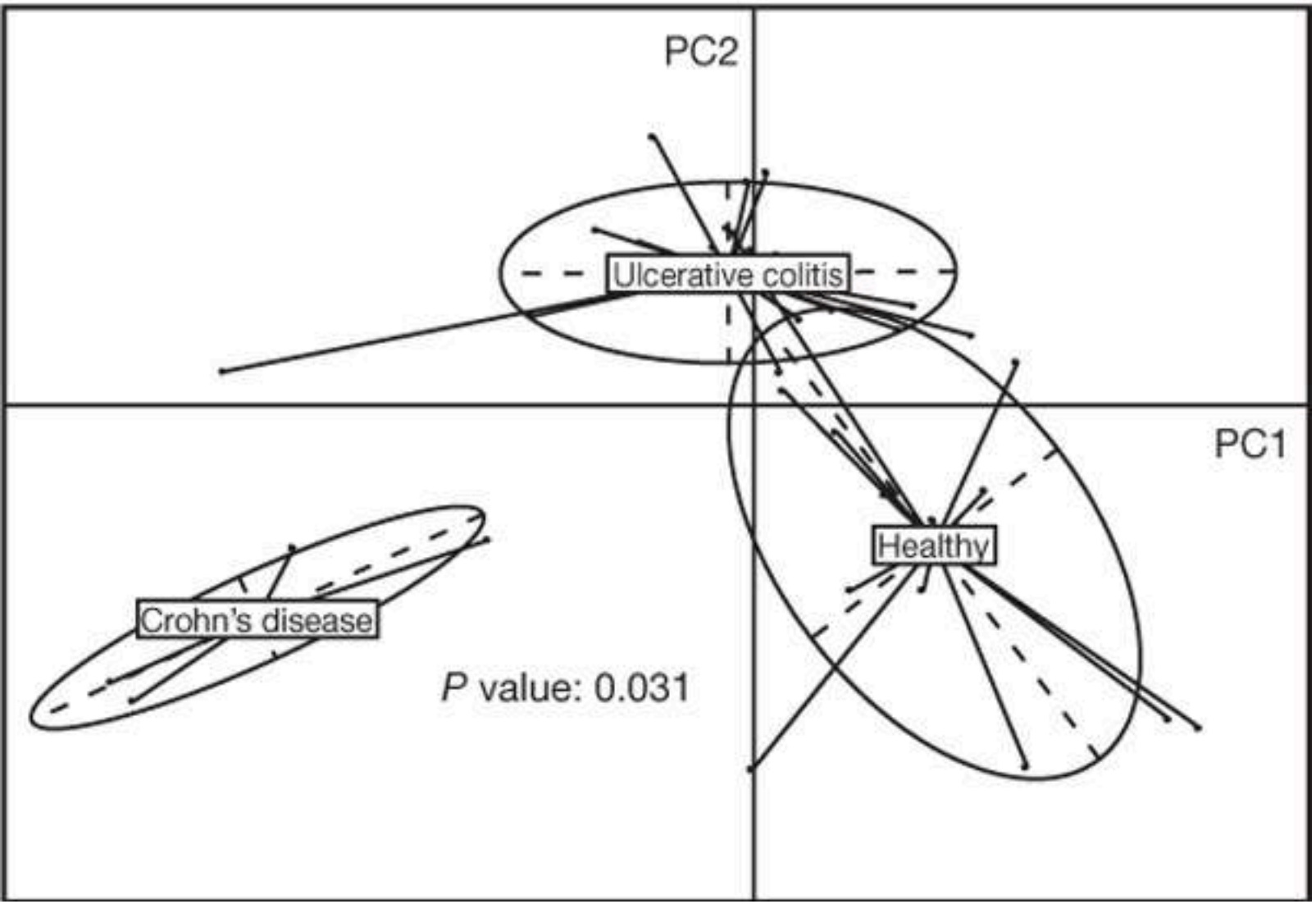
ARTICLES

A human gut microbial gene catalogue established by metagenomic sequencing

Junjie Qin[1]*, Ruiqiang Li[1]*, Jeroen Raes[2,3], Manimozhiyan Arumugam[2], Kristoffer Solvsten Burgdorf[4], Chaysavanh Manichanh[5], Trine Nielsen[4], Nicolas Pons[6], Florence Levenez[6], Takuji Yamada[2], Daniel R. Mende[2], Junhua Li[1,7], Junming Xu[1], Shaochuan Li[1], Dongfang Li[1,8], Jianjun Cao[1], Bo Wang[1], Huiqing Liang[1], Huisong Zheng[1], Yinlong Xie[1,7], Julien Tap[6], Patricia Lepage[6], Marcelo Bertalan[9], Jean-Michel Batto[6], Torben Hansen[4], Denis Le Paslier[10], Allan Linneberg[11], H. Bjørn Nielsen[9], Eric Pelletier[10], Pierre Renault[6], Thomas Sicheritz-Ponten[9], Keith Turner[12], Hongmei Zhu[1], Chang Yu[1], Shengting Li[1], Min Jian[1], Yan Zhou[1], Yingrui Li[1], Xiuqing Zhang[1], Songgang Li[1], Nan Qin[1], Huanming Yang[1], Jian Wang[1], Søren Brunak[9], Joel Doré[6], Francisco Guarner[5], Karsten Kristiansen[13], Oluf Pedersen[4,14], Julian Parkhill[12], Jean Weissenbach[10], MetaHIT Consortium†, Peer Bork[2], S. Dusko Ehrlich[6] & Jun Wang[1,13]

doi:10.1038/nature08821

Human gut microbiome

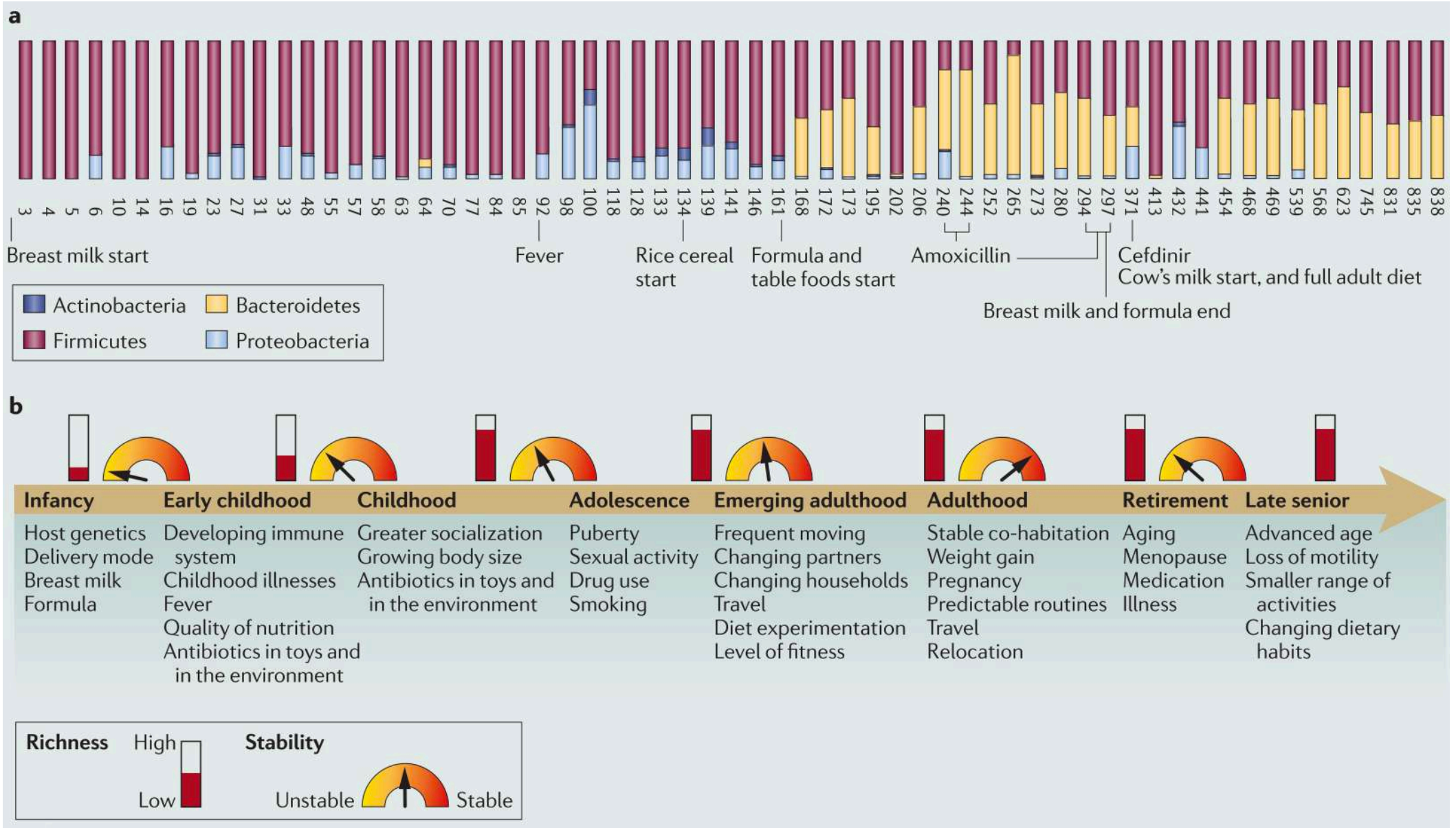


We can check which OTUs constitute the clustering (and separation) patterns

-> Biology
-> Biomarkers

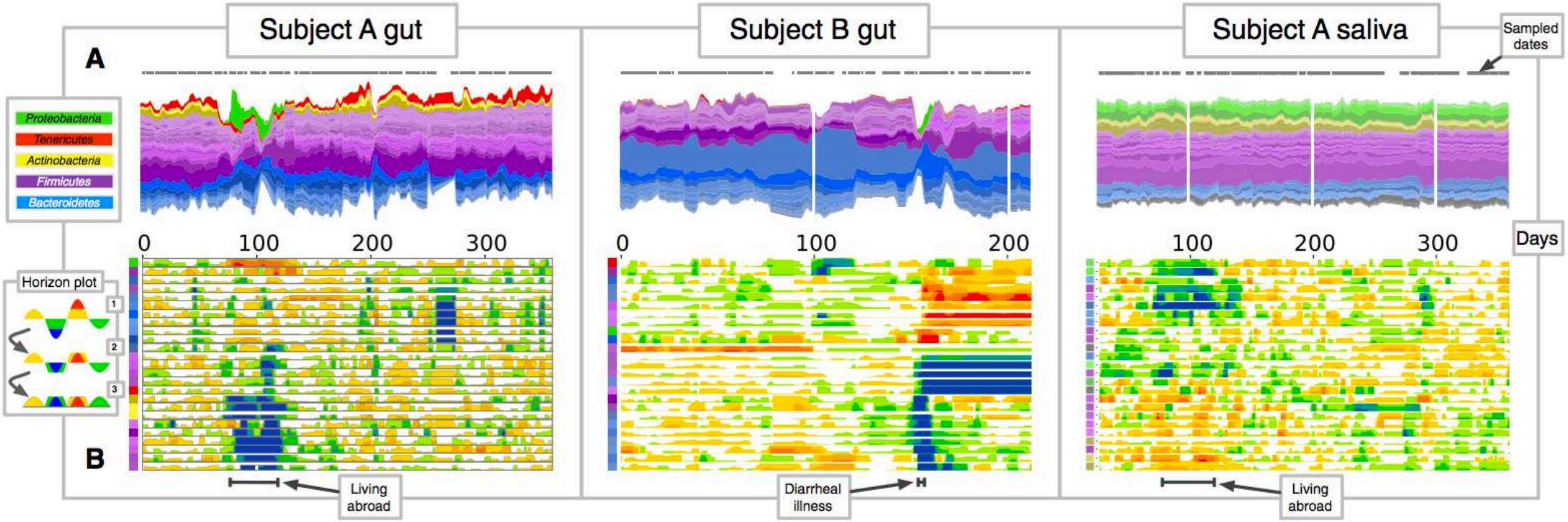
doi:10.1038/nature08821

The gut microbiome during life



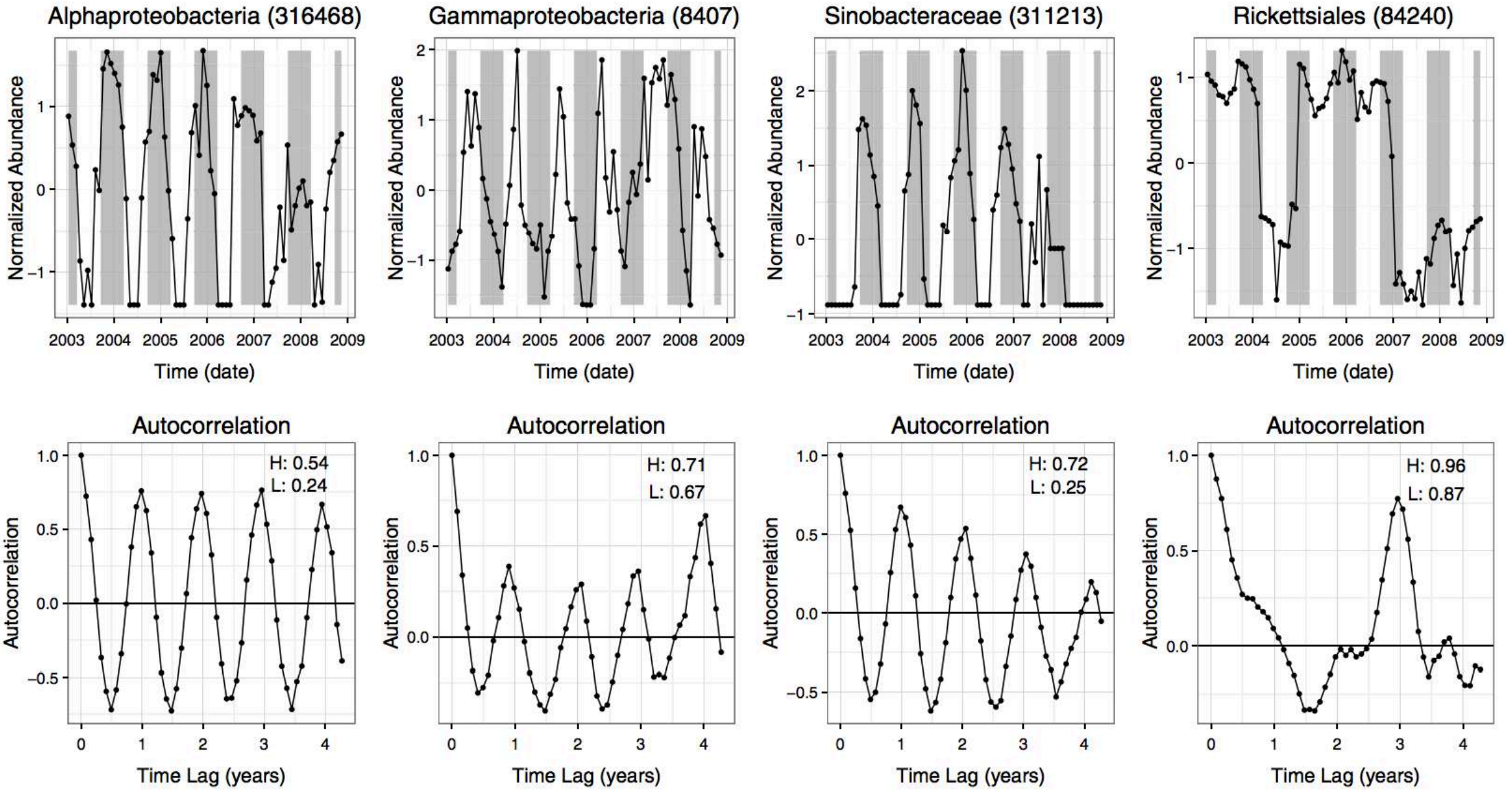
doi:10.1038/nrmicro2540

Tracking microbiome on a daily scale



David et al. Genome Biology 2014, 15:R89

Tracking microbiome spanning 6 years



Faust et al 2015

Current Opinion in Microbiology

Priority effect

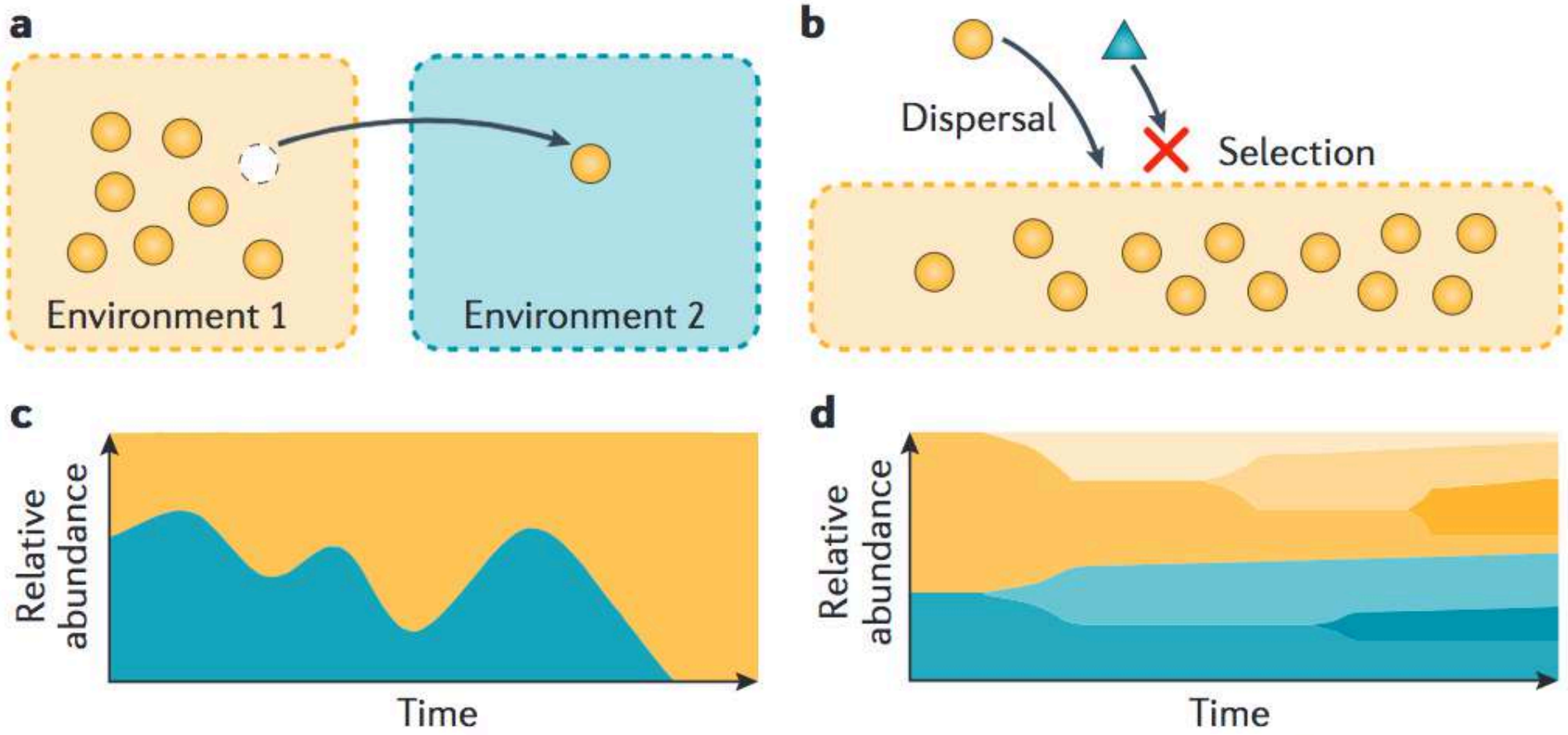


Figure 1 | **Four processes that affect ecological communities. a** | The arrow represents dispersal of an organism (orange circle) from Environment 1 (orange shading) to Environment 2 (blue shading). **b** | Deterministic fitness differences between two species (orange circle, blue triangle) cause the orange environment to select for one (orange circle) and against the other (blue triangle). **c** | Stochastic changes in the relative abundances of two species (orange area and blue area) result in changes in community structure within one environment through time. As a result, one population (blue) has gone locally extinct by the end of the time period. **d** | Mutation and/or recombination within a population (blue and orange areas) results in new genetic variation through time, leading to new strains (as denoted by different shades).

https://www.nature.com/articles/nrgastro.2017.173.pdf

Smithsonian

Institute for

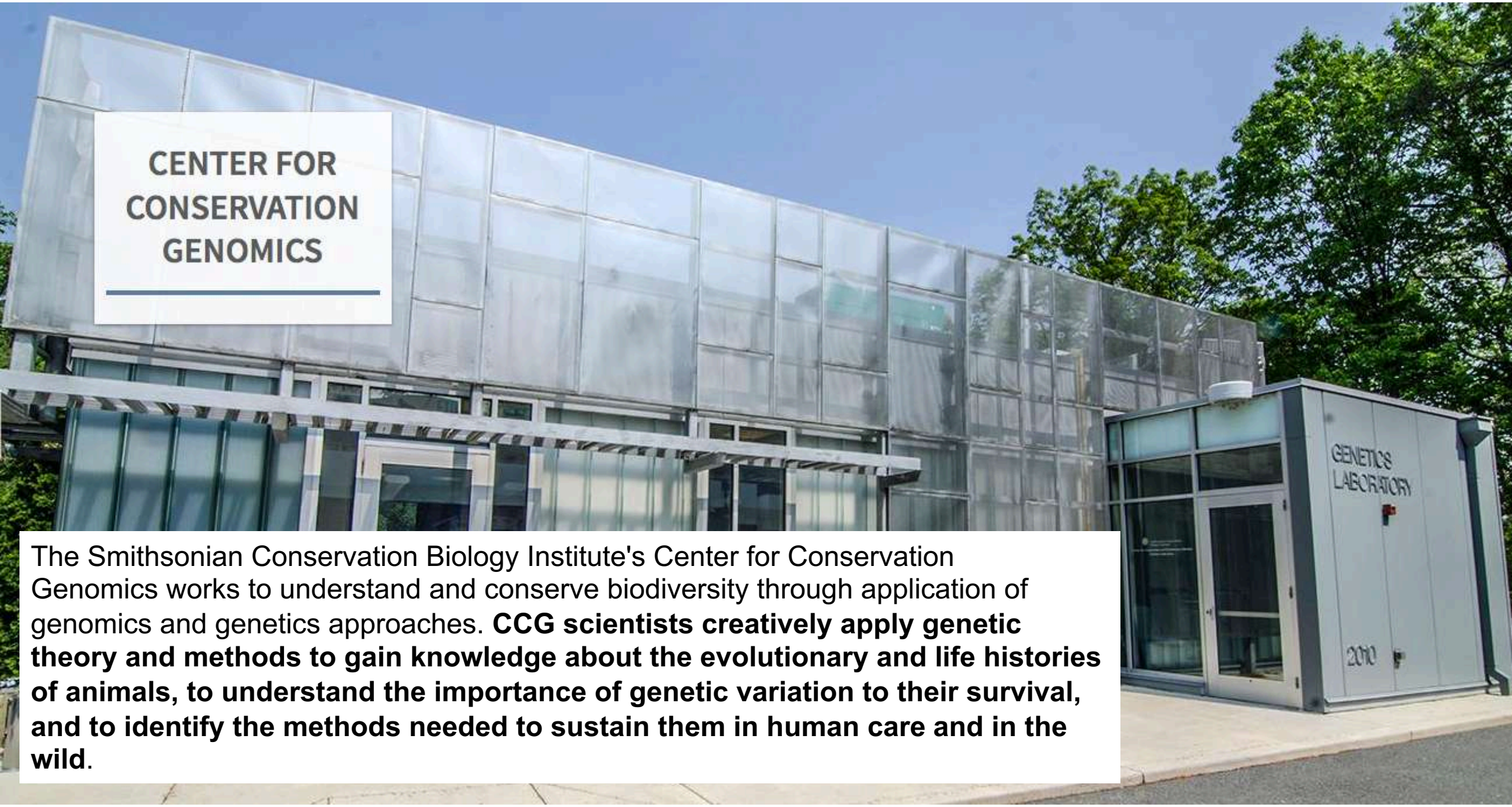
Biodiversity Genomics

How do we sustain life on our changing planet?

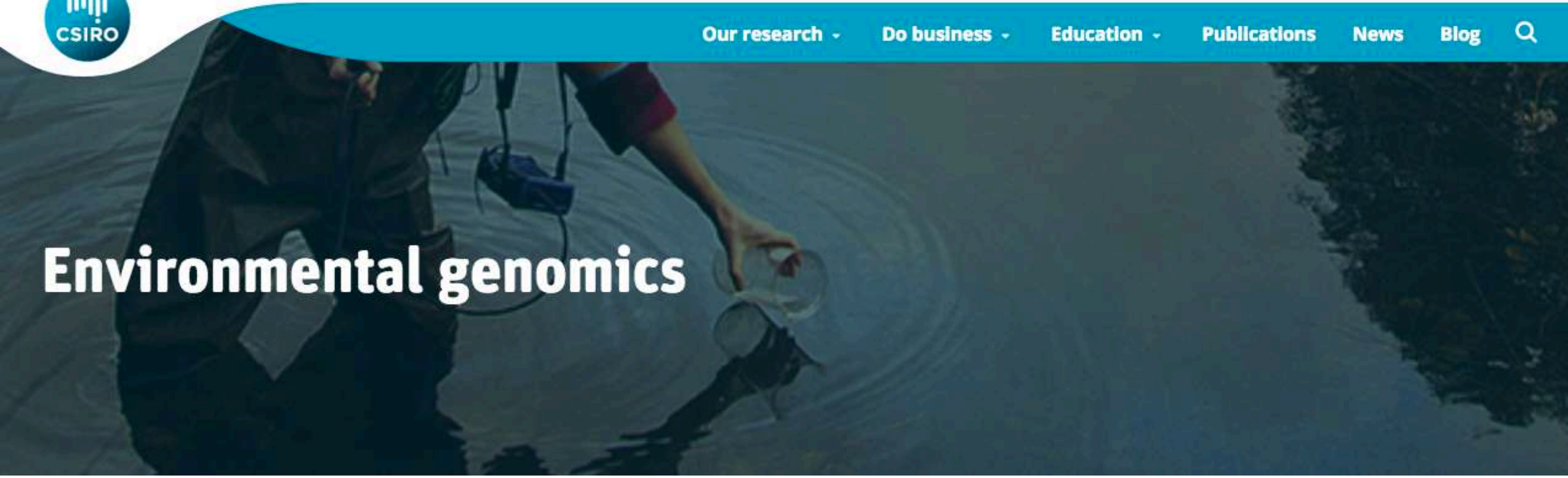
Biodiversity—our planet's complex web of interdependent species and ecosystems—is critical to our survival and includes the water we drink, the air we breathe, the food we eat, the medicines that heal, and the soils that nurture.

But our biodiversity faces serious challenges.

The emerging Institute for Biodiversity Genomics, a united effort of existing Smithsonian research entities and a suite of partners around the world, will help scientists address these challenges. By using the latest genome research and technologies, we will gain greater understanding of how life on Earth evolved, how species interact, how ecosystems function, and how to sustain the diversity of life that allows us to adapt and thrive in our changing world.



The Smithsonian Conservation Biology Institute's Center for Conservation Genomics works to understand and conserve biodiversity through application of genomics and genetics approaches. **CCG scientists creatively apply genetic theory and methods to gain knowledge about the evolutionary and life histories of animals, to understand the importance of genetic variation to their survival, and to identify the methods needed to sustain them in human care and in the wild**.



Environmental genomics

We use genomics approaches to determine **how species and communities respond to a global environment altering with land use change and development, including exposure to industrial contaminants and agricultural chemicals.**

EARTH BIOGENOME PROJECT

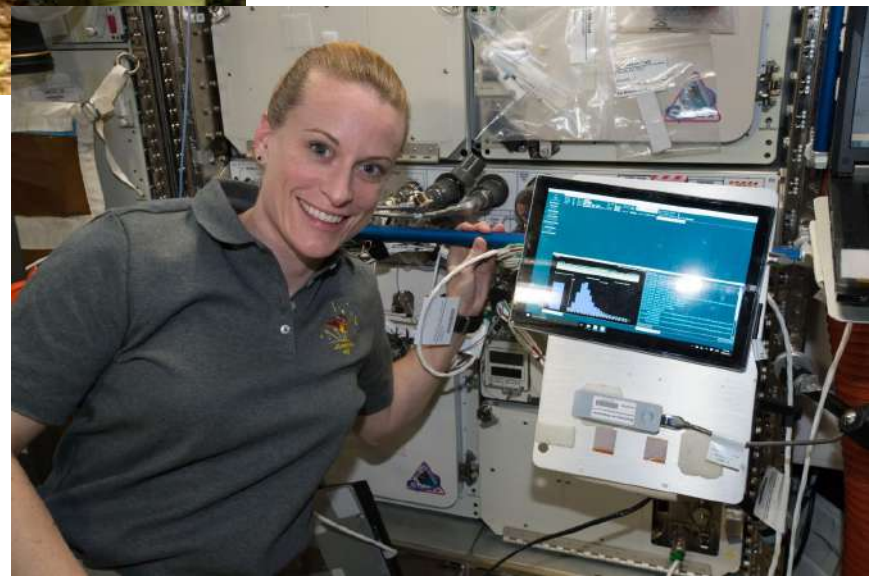
A GRAND CHALLENGE

The Earth BioGenome Project, a Moon Shot 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

The Earth BioGenome Project will create a new foundation for biology, informing a broad range of major issues facing humanity, such as the impact of climate change on biodiversity, the conservation of endangered species and ecosystems, and the preservation and enhancement of ecosystem services.

Mobility of sequencing



Scenarios now and then

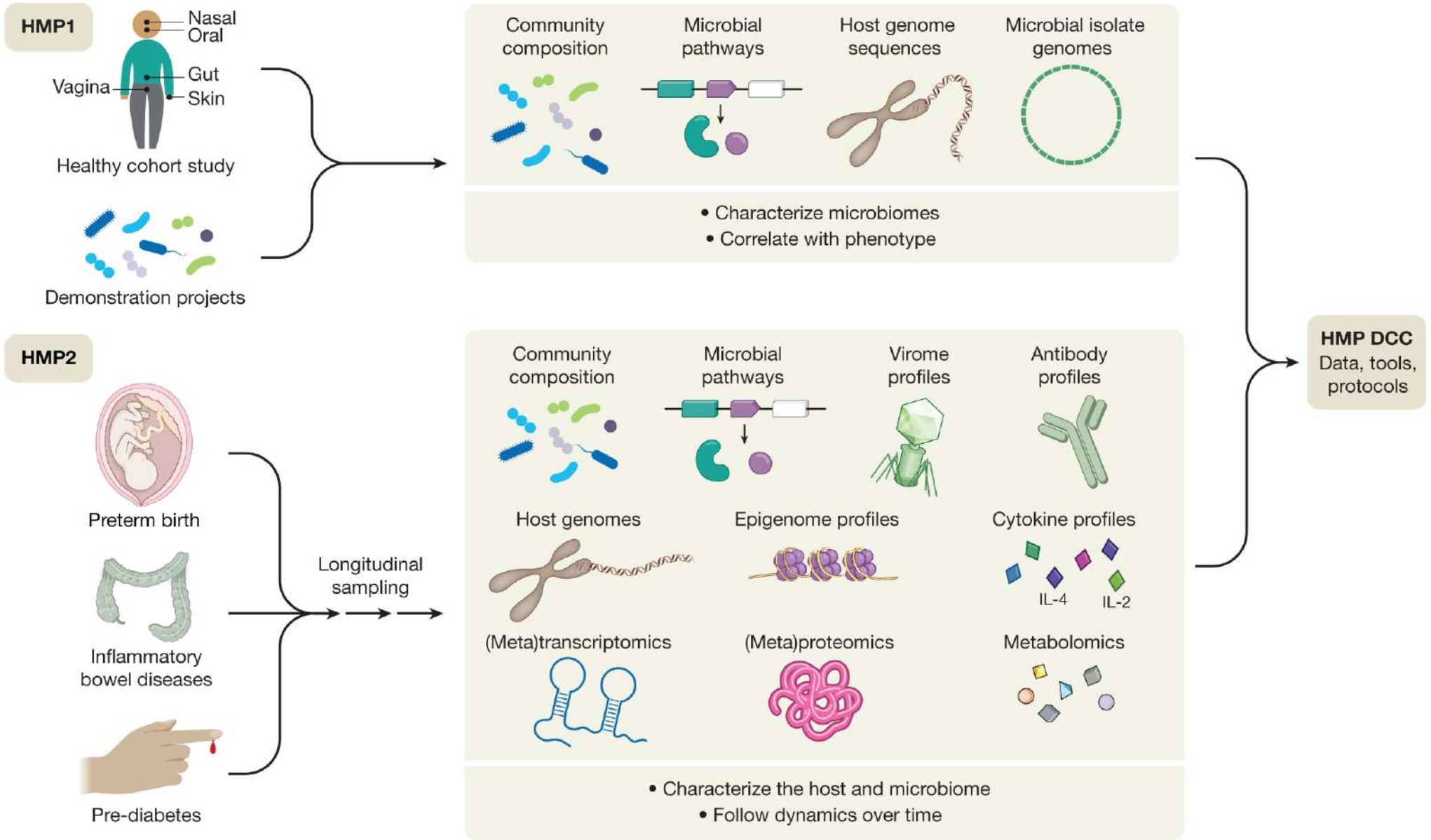
1. [lab/hospital/mountain/sea] Collect samples (1.1, 1.2, 1.3…)
2. [lab/hospital] Extract DNA (2.1, 2.2, 2.3…)
3. [lab/hospital/company] Sequencing (3.1, 3.2, 3.3…)
4. [lab/company] Analysis
5. [lab/hospital] Report

Weeks

1. [lab/hospital/mountain/sea] Collect samples -> report

Minute

“a paradigm for future multi-omic studies of the human microbiome”



Proctor, L.M., Creasy, H.H., Fettweis, J.M. *et al.* The Integrative Human Microbiome Project. *Nature* **569,** 641–648 (2019). https://doi.org/10.1038/s41586-019-1238-8

New challenges

- So much data
- Technology advancement
- **Integrating different kinds of data (multi-omic)**
- High performance
- Reproducibility crisis

- Bioinformaticians as a profession
- Only biology has a specific term to refer to the use of computers in this discipline (‘bioinformatics’)
- Proper integration into academic curriculums

Shift in paradigm 2005-2020 (My personal take)

- A genome, a few genomes are no longer “enough”
  - ~since everybody can do it reasonably well
- Genome sequencing projects are
  - being done on a per-lab basis and no longer exclusive to sequencing centers
  - moving away from exploration to question orientated.
- Data being produced on a **much faster speed** at a **much higher throughput**, and a much **cheaper scale**

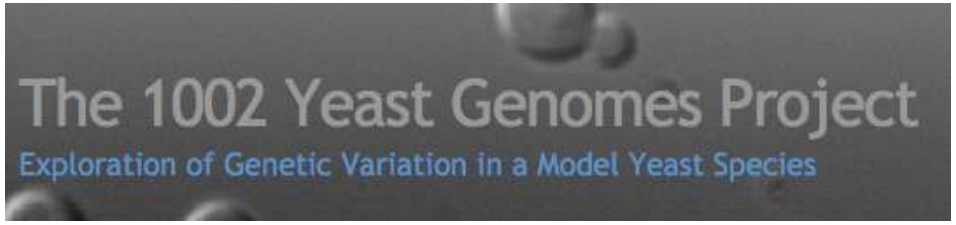
- More methods, analysis, tools, experiments…
  - Not always better

It is an exciting time to be in

Current and future

- Sequencing will still be cheaper, read will get longer
- Projects will be bigger

- Standard labs will be able to generate collections of themselves



(3 labs)

There’s so much more…

- Read, read, read
- Twitter and blogs

First written assignment

- Find a paper that has a combination of comparative, population, RNAseq or metagenomics in your field (at least 2).
- Write a protocol on how the bioinformatics part of the study was conducted (what tools, what version, input, output). As detailed as possible

- **Deadline: 25th March**
- **Email the assignment to Vanessa** (biodiv@gate.sinica.edu.tw)