NGS lecture

Isheng Jason Tsai

Chang Gung University 2020
2019.03.26

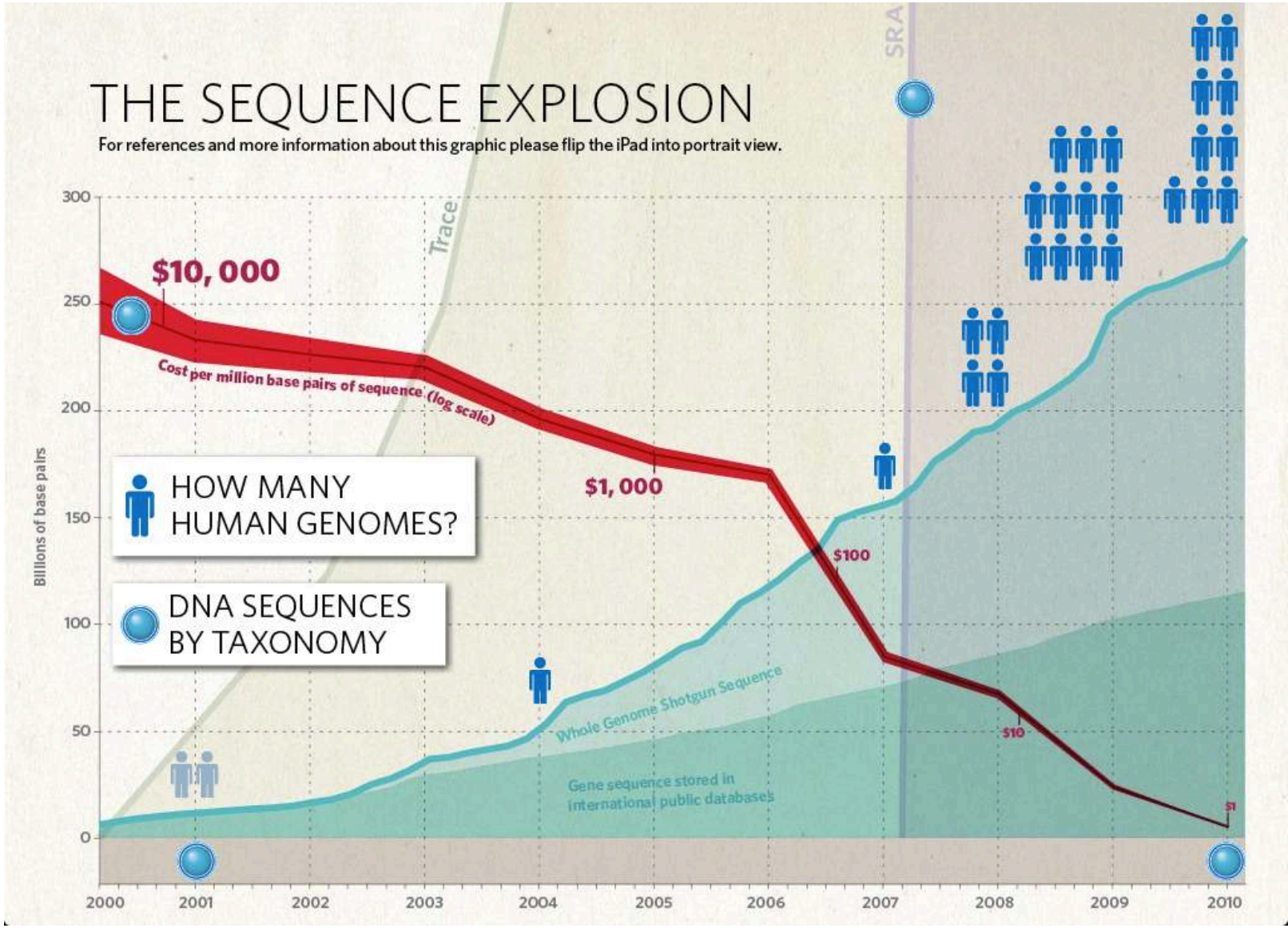
This lecture is called “NGS”

Actually

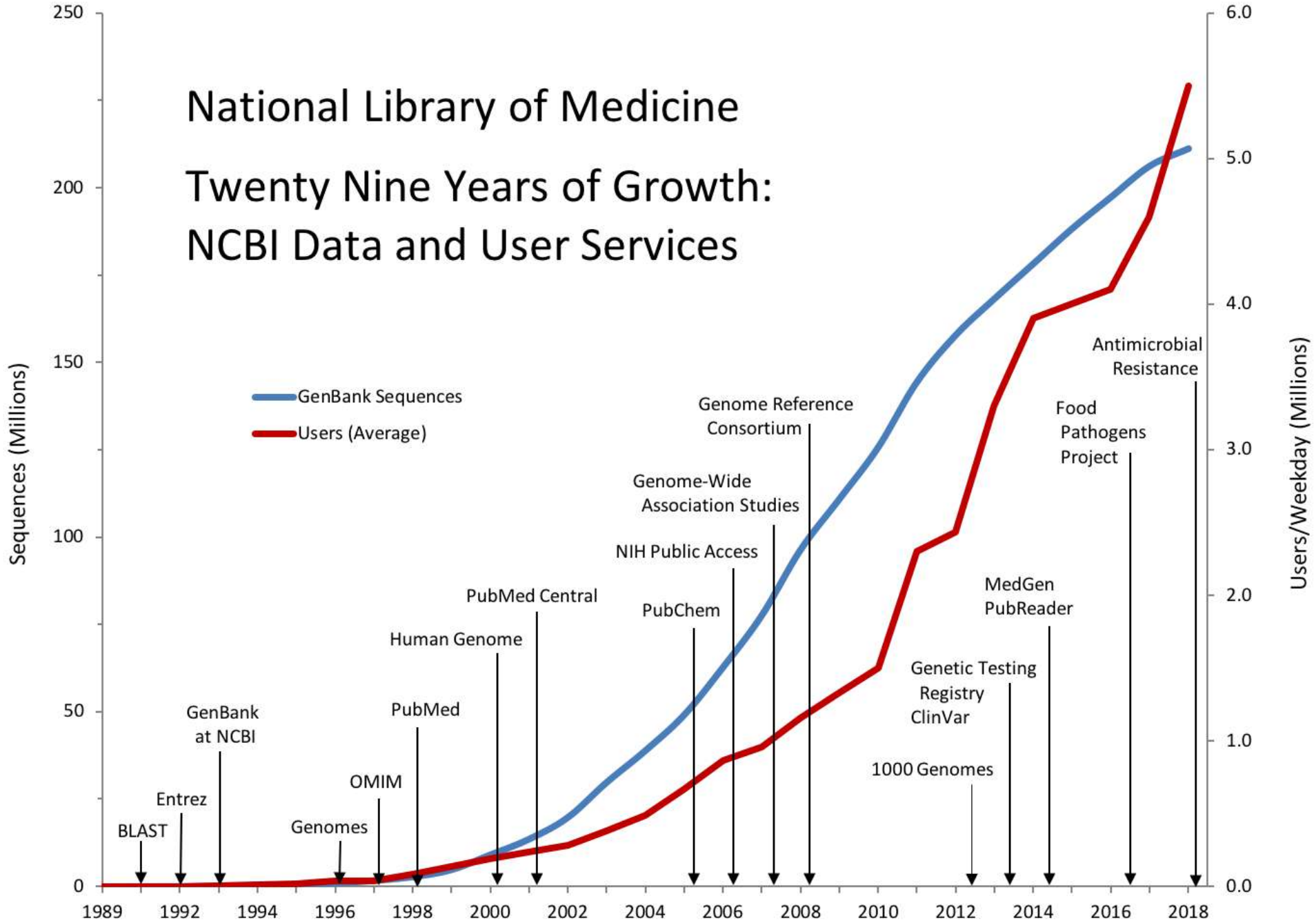
- Next Generation Sequencing is really “now” sequencing
- It won’t be so easy to tell you everything about NGS
  (it’s a bit like saying what can we do with PCR?)

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

What we will cover today

NGS: Some basics
Sequencing platforms
Data types
Analysis:

- RNAseq
- 16S
- Metagenomics

Previous questions:

- microbiota的paper要怎麼approach
- 16S sequencing region primer choice
- microbiota醫院有fecal transplantation計劃?

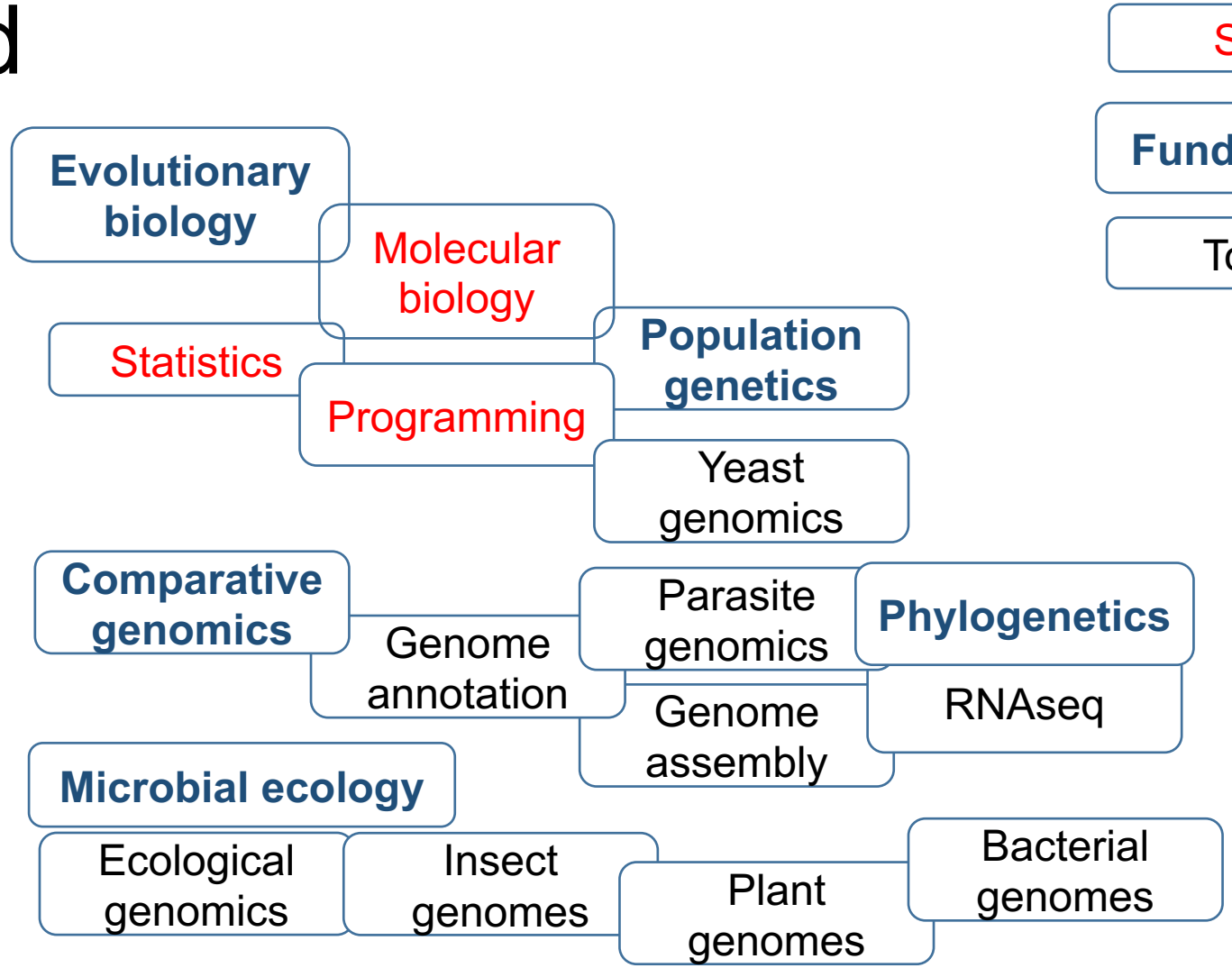
My background

Undergraduate:
Biochemistry and Genetics

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

2009-14 ; Postdoc:
Genomics & parasitology

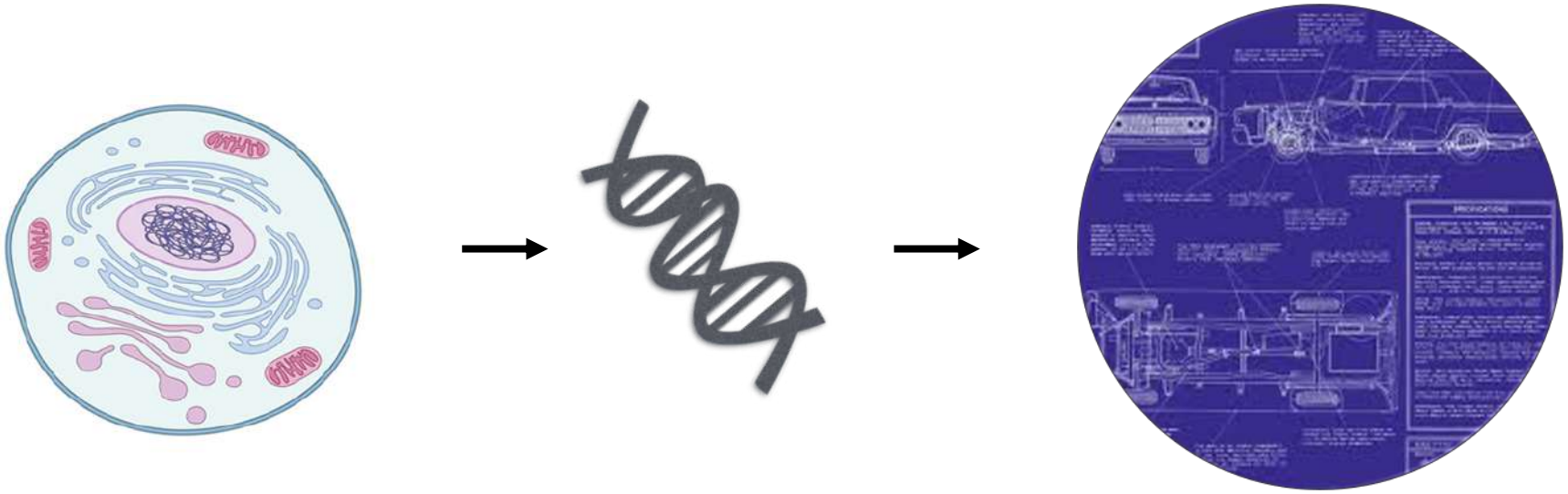
2015 - ; Academia Sinica:
Microbial diversity & Bioinformatics



- **49 publications**
  (2 Nature, 1 Science, 2 Nature Genetics, 2 PNAS,
  3 Genome Biology, 1 Nature Plant, 1 Nature Communications)

What is genomics 基因體學?

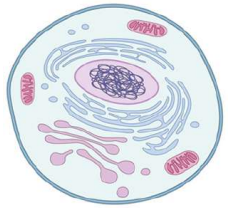
Genome



Genome = Parts list of a single genome

A genome project

Wet lab work



Extract DNA
(or RNA)

Shear DNA
(or RNA)

Sequence

Bioinformatics

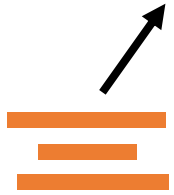
Variant
ATCG
ATGG
ATCG

Annotation

Data QC

Reads
50-500 bp

Assembly



Contigs
1kb – 100 kbp

Mapping

DNA or RNA Reads
50-500 bp

Genome reference

Three 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

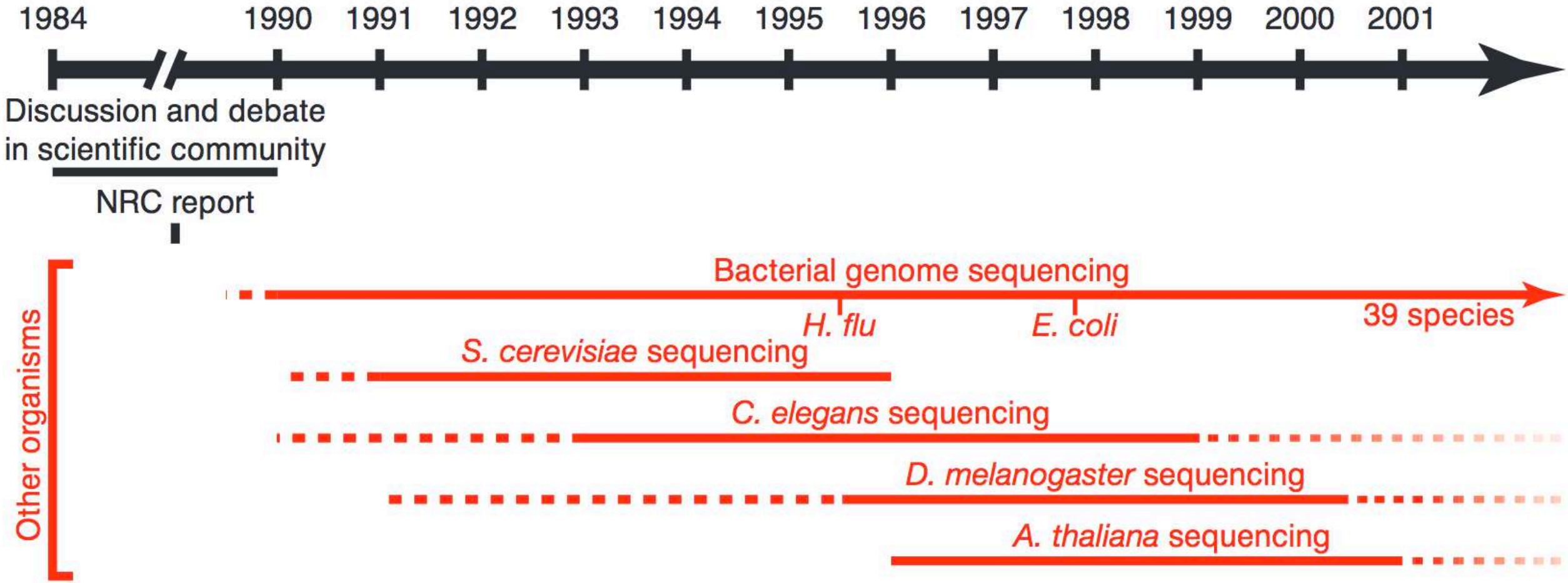
Counting:

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

Why sequence a genome?

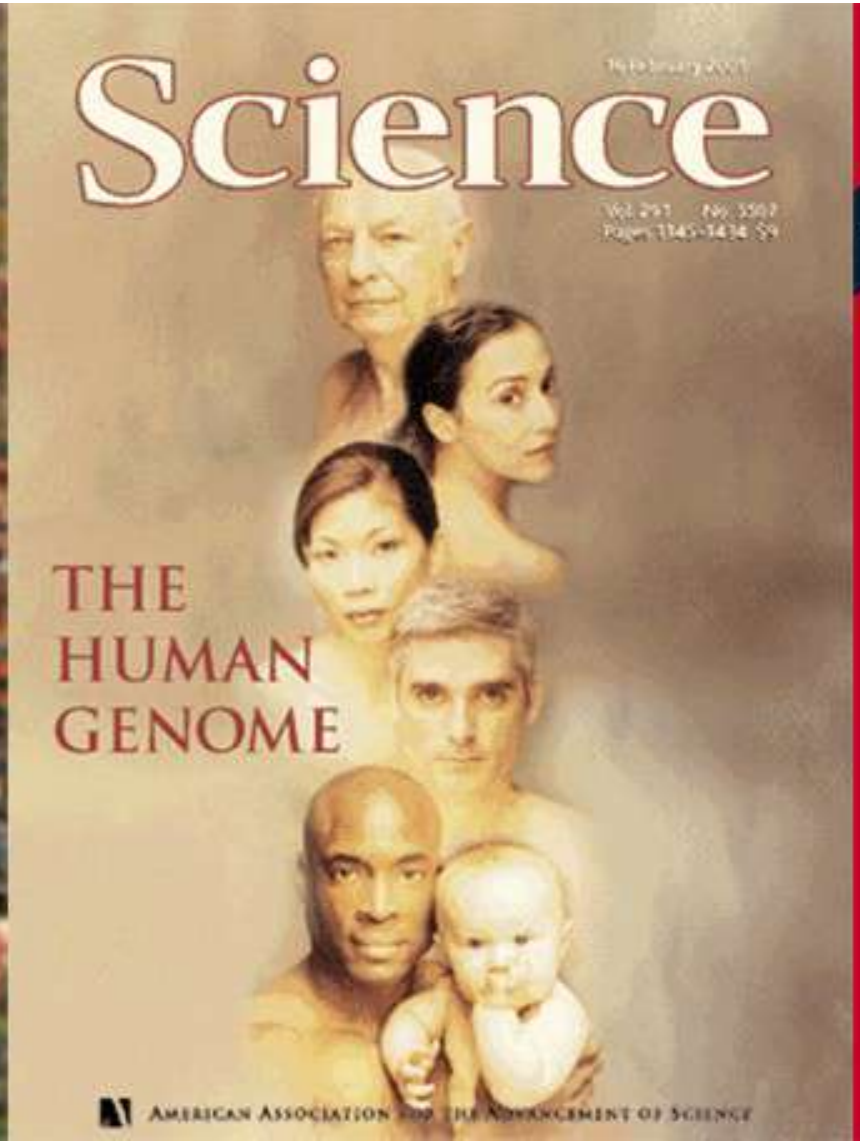
- Phylogenetic position
- Differences between species (comparative genomics)
- Variations between individuals (population genetics)
- Help to understand biology
- Of economic, agricultural, medical, ecology values

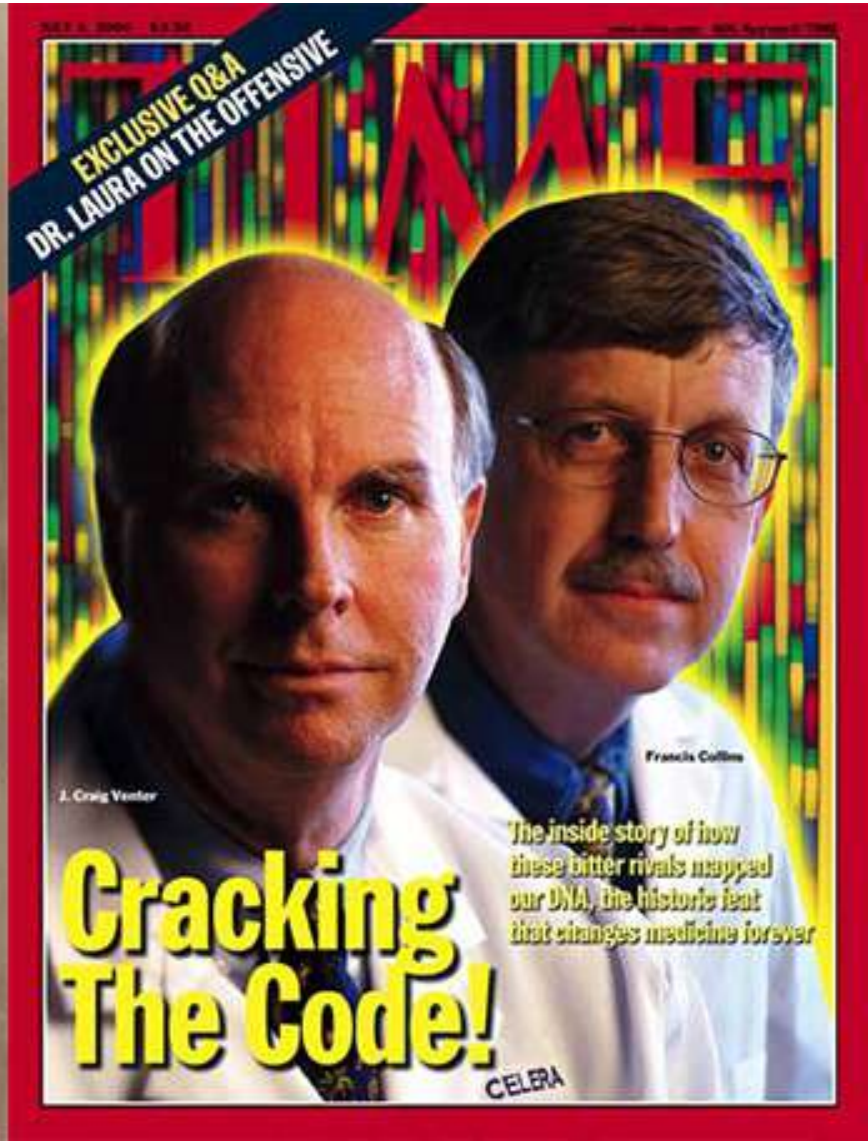
- **Help to understand biology**



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/

FOCUS ON GENOMES OF ICELANDERS

ARTICLES

nature genetics

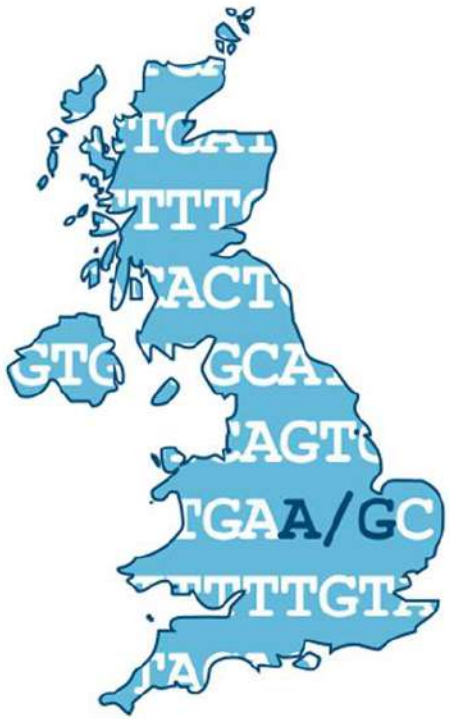
Large-scale whole-genome sequencing of the Icelandic population

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

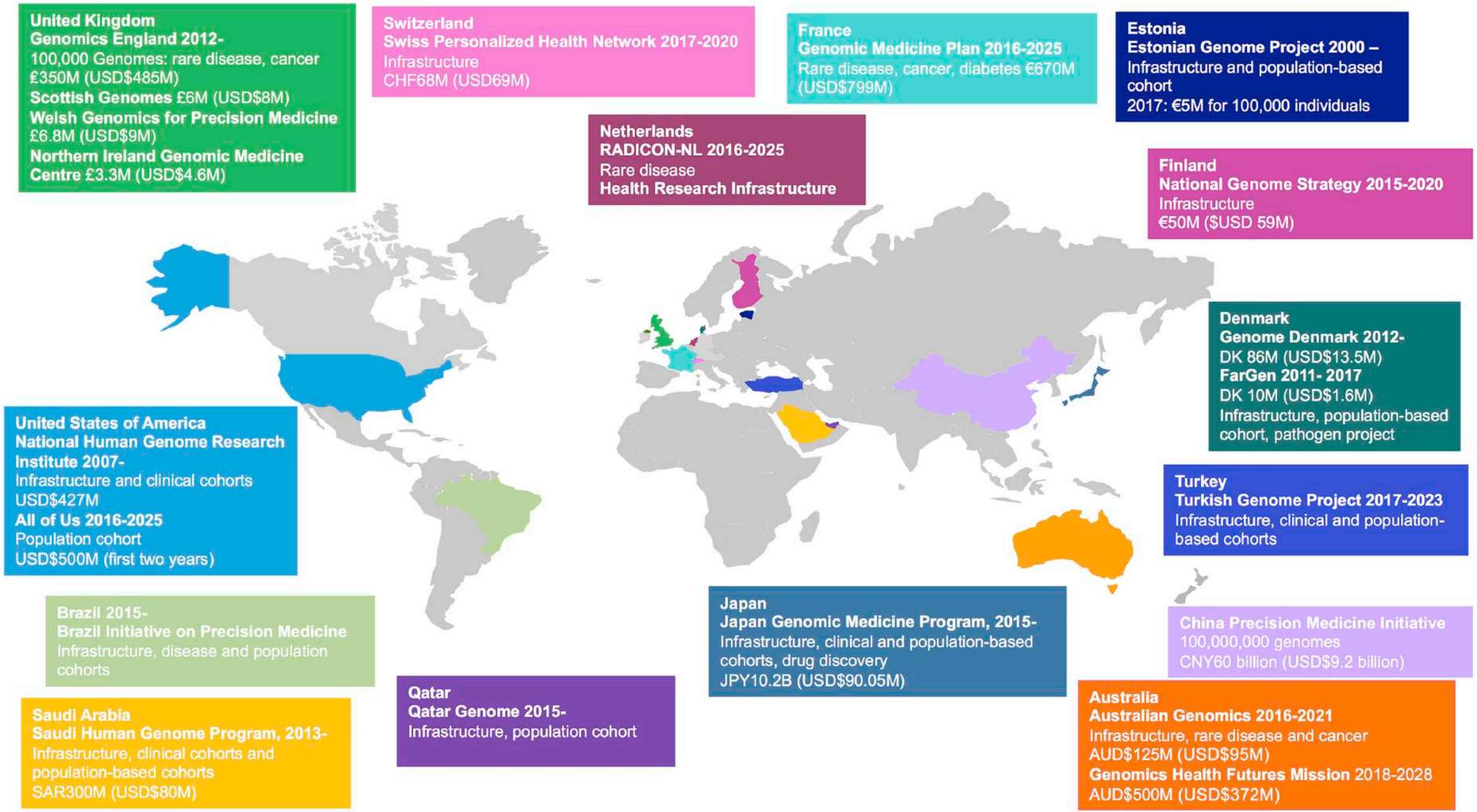
The project is taking a two-pronged approach to identify rare variants and their effects:

•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

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

Cold Spring Harbor Laboratory
bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank

Authors: Cristopher V. Van Hout[1], Ioanna Tachmazidou[2], Joshua D. Backman[1], Joshua X. Hoffman[2], Bin Ye[1], Ashutosh K. Pandey[2], Claudia Gonzaga-Jauregui[1], Shareef Khalid[1], Daren Liu[1], Nilanjana Banerjee[1], Alexander H. Li[1], Colm O'Dushlaine[1], Anthony Marcketta[1], Jeffrey Staples[1], Claudia Schurmann[1], Alicia Hawes[1], Evan Maxwell[1], Leland Barnard[1], Alexander Lopez[1], John Penn[1], Lukas Habegger[1], Andrew L. Blumenfeld[1], Ashish Yadav[1], Kavita Praveen[1], Marcus Jones[3], William J. Salerno[1], Wendy K. Chung[4], Ida Surakka[5], Cristen J. Willer[5], Kristian Hveem[6], Joseph B. Leader[7], David J. Carey[7], David H. Ledbetter[7], Geisinger-Regeneron DiscovEHR Collaboration[7], Lon Cardon[2], George D. Yancopoulos[3], Aris Economides[3], Giovanni Coppola[1], Alan R. Shuldiner[1], Suganthi Balasubramanian[1], Michael Cantor[1], Matthew R. Nelson[2,*], John Whittaker[2,*], Jeffrey G. Reid[1,*], Jonathan Marchini[1,*], John D. Overton[1,*], Robert A. Scott[2,*], Gonçalo Abecasis[1,*], Laura Yerges-Armstrong[2,*], Aris Baras[1,*] on behalf of the Regeneron Genetics Center

The UK Biobank is a prospective study of 502,543 individuals, combining extensive phenotypic and genotypic data with streamlined access for researchers around the world.

https://www.biorxiv.org/content/10.1101/572347v1

	Variants in WES, n=49,960 Participants		Median Per Participant (IQR)	
	# Variants	# Variants MAF<1%	# Variants	# Variants MAF<1%
Total	9,693,526	9,547,730	48,982 (627)	1,626 (133)
Targeted Regions[1]	4,735,722	4,665,684	24,332 (283)	780 (63)
Variant Type[1]				
SNVs	4,520,754	4,453,941	23,529 (276)	739 (61)
Indels	214,968	211,743	803 (29)	42 (10)
Multi-Allelic	591,340	580,728	3,388 (63)	117 (18)
Functional Prediction				
Synonymous	1,229,303	1,203,043	9,619 (128)	228 (28)
Missense	2,498,947	2,472,384	8,781 (137)	380 (39)
LOF (any transcript)	231,631	230,790	219 (16)	24 (8)
LOF (all transcripts)	153,903	153,441	111 (12)	15 (6)

Table 2 | Summary statistics for variants in sequenced exomes of 49,960 UKB participants

https://www.biorxiv.org/content/10.1101/572347v1

Project setup

- Sequencing a species (Comparative genomics)
  - Map, assemble

- **Sequencing multiple individuals of a species (Population genomics)**
  - **Map, count**

- Combination of (1) and (2)

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

生物資訊的開始

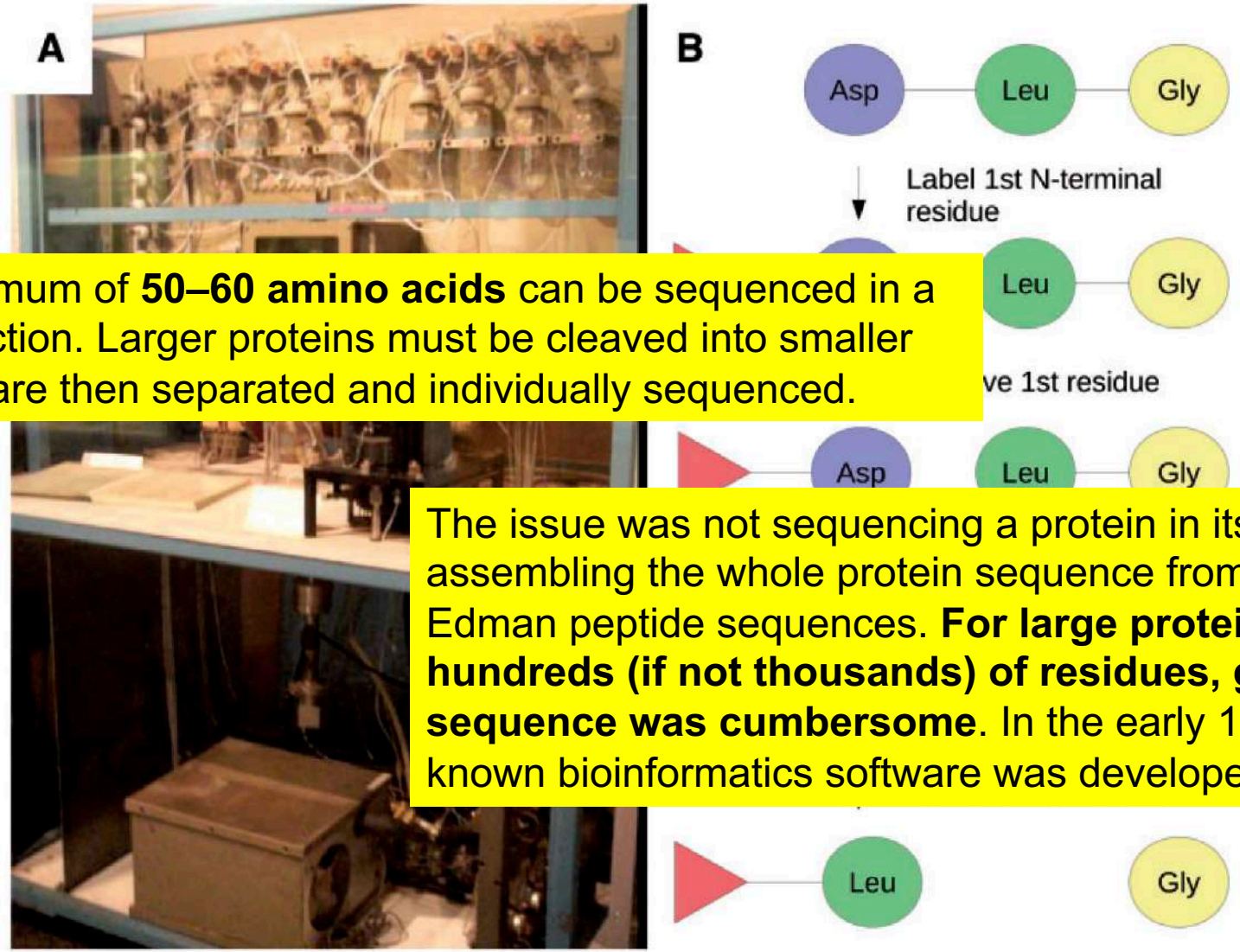
- the very beginnings of bioinformatics occurred **more than 50 years ago**, when desktop computers were still a hypothesis and DNA could not yet be sequenced."

- 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.

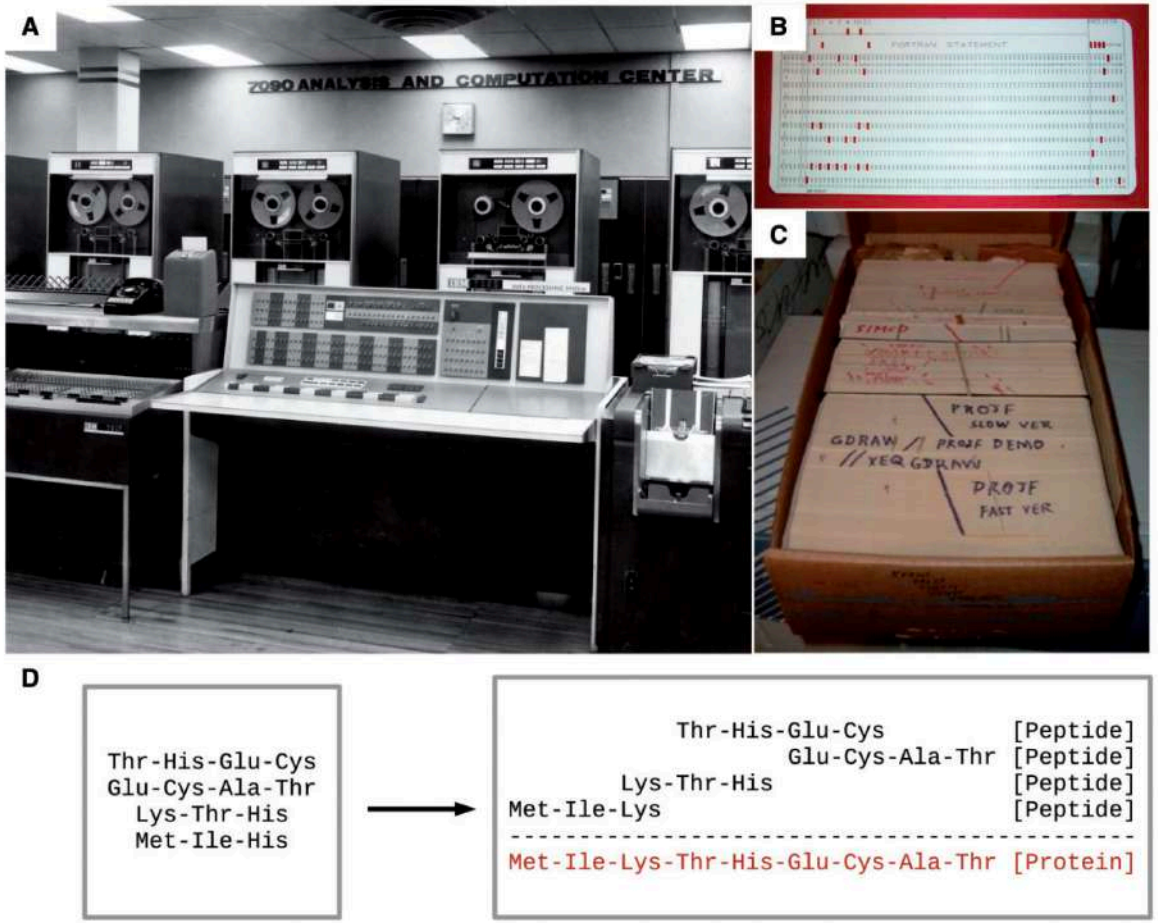


Figure 2. COMPROTEIN, the first bioinformatics software. (A) An IBM 7090 mainframe, for which COMPROTEIN was made to run. (B) A punch card containing one line of FORTRAN code (the language COMPROTEIN was written with). (C) An entire program's source code in punch cards. (D) A simplified overview of COMPROTEIN's 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)

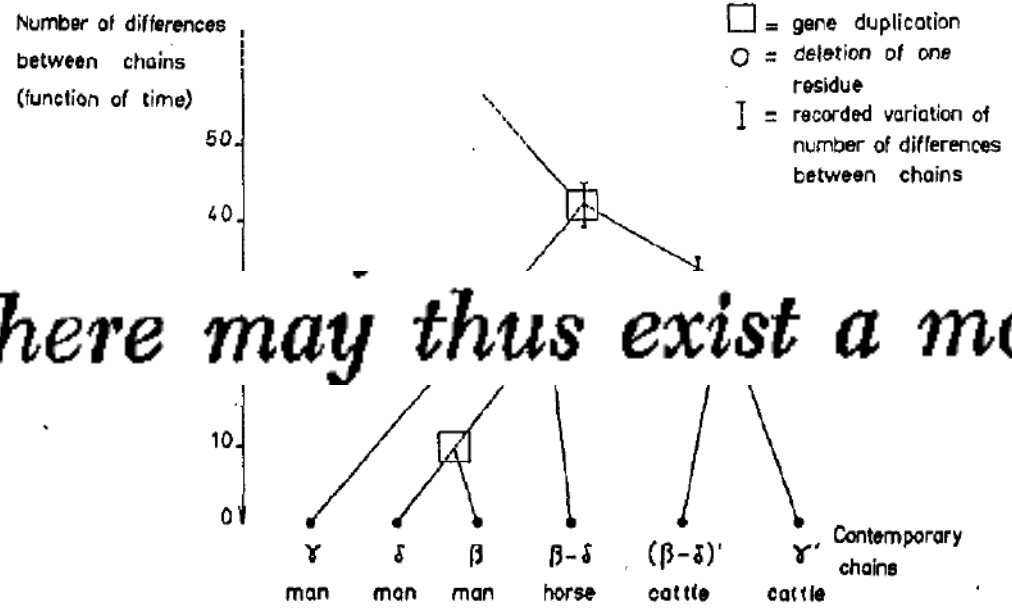
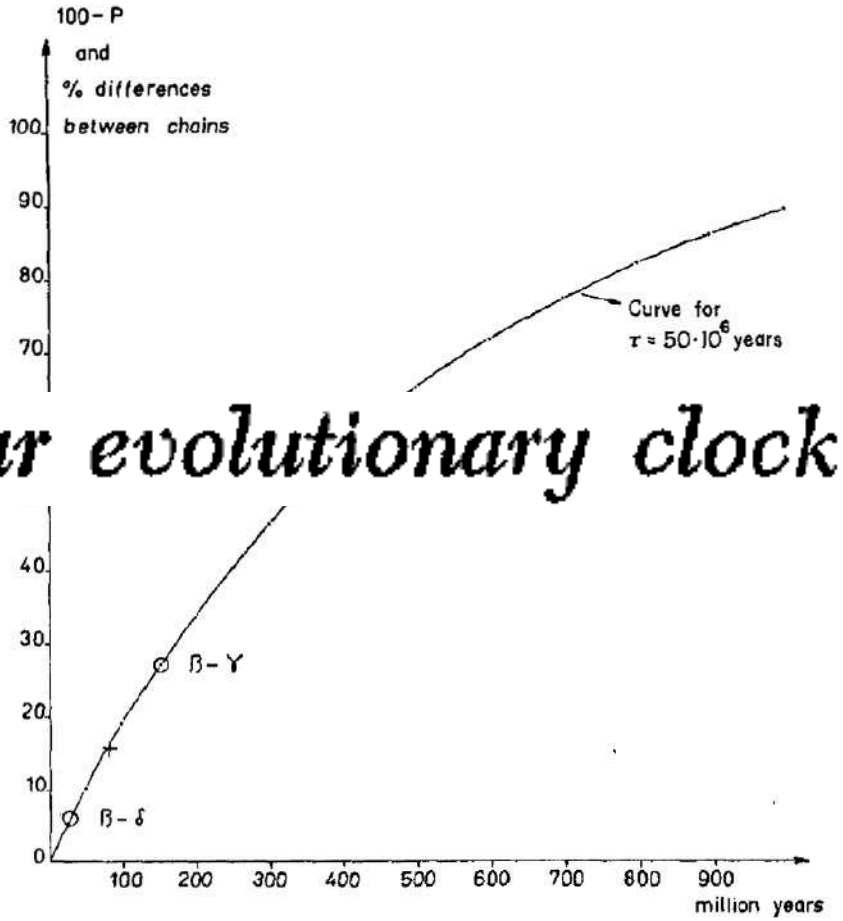


FIG. 4. Probable evolutionary relationship of some mammalian hemoglobin chains.

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

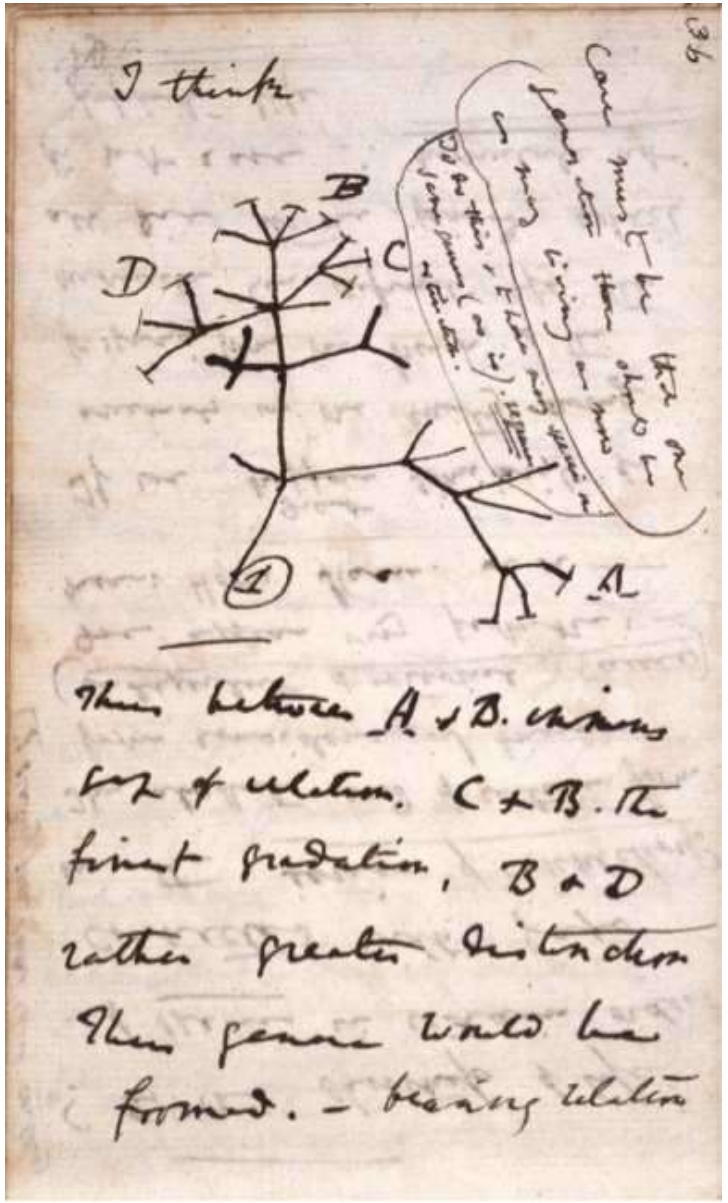


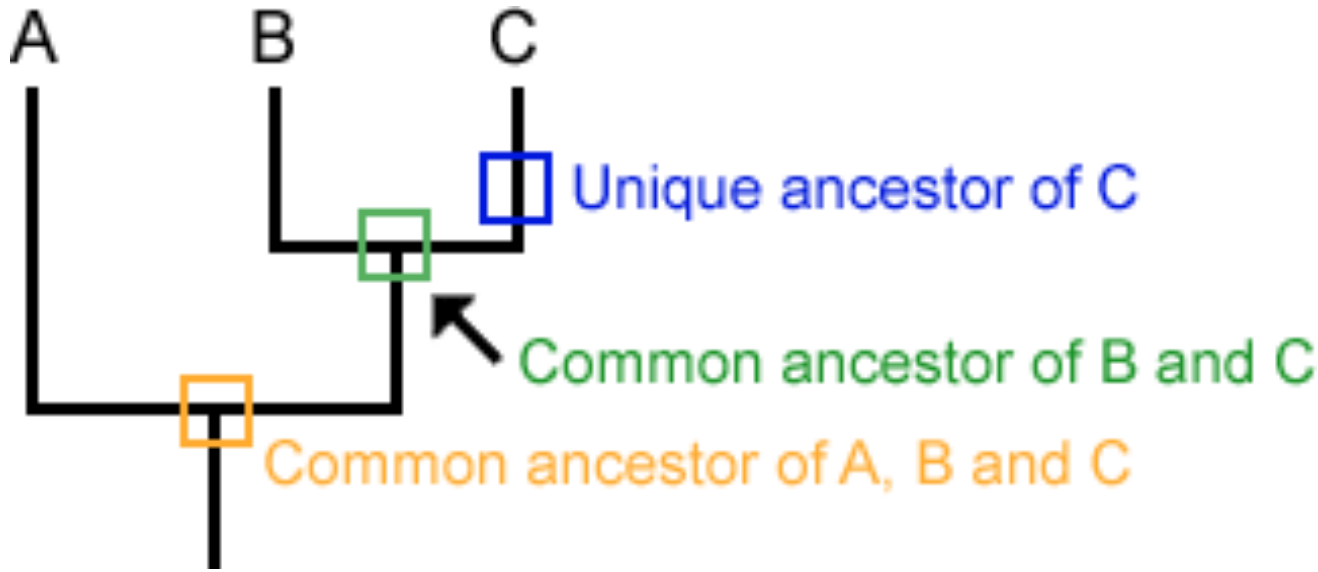
"Zuckerkandl and Pauling hypothesized that orthologous proteins evolved through divergence from a common ancestor. Consequently, by comparing the sequence of hemoglobin in currently extant organisms, it became possible to predict the 'ancestral sequences' of hemoglobin and, in the process, its evolutionary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

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

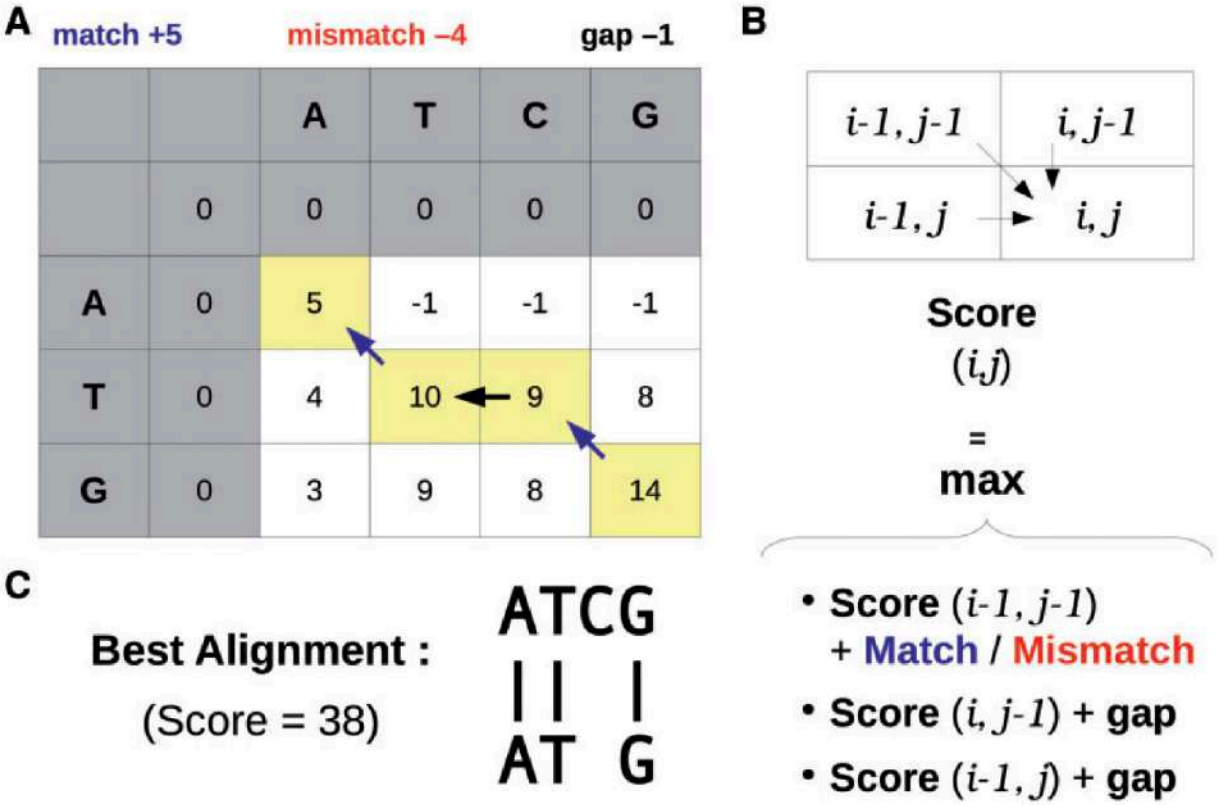
A brief history of bioinformatics
Jeff Gauthier, Antony T Vincent, Steve J Charette, Nicolas Derome
Briefings in Bioinformatics (2018) https://doi.org/10.1093/bib/bby063

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

Different sequencing platforms / History of sequencing

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

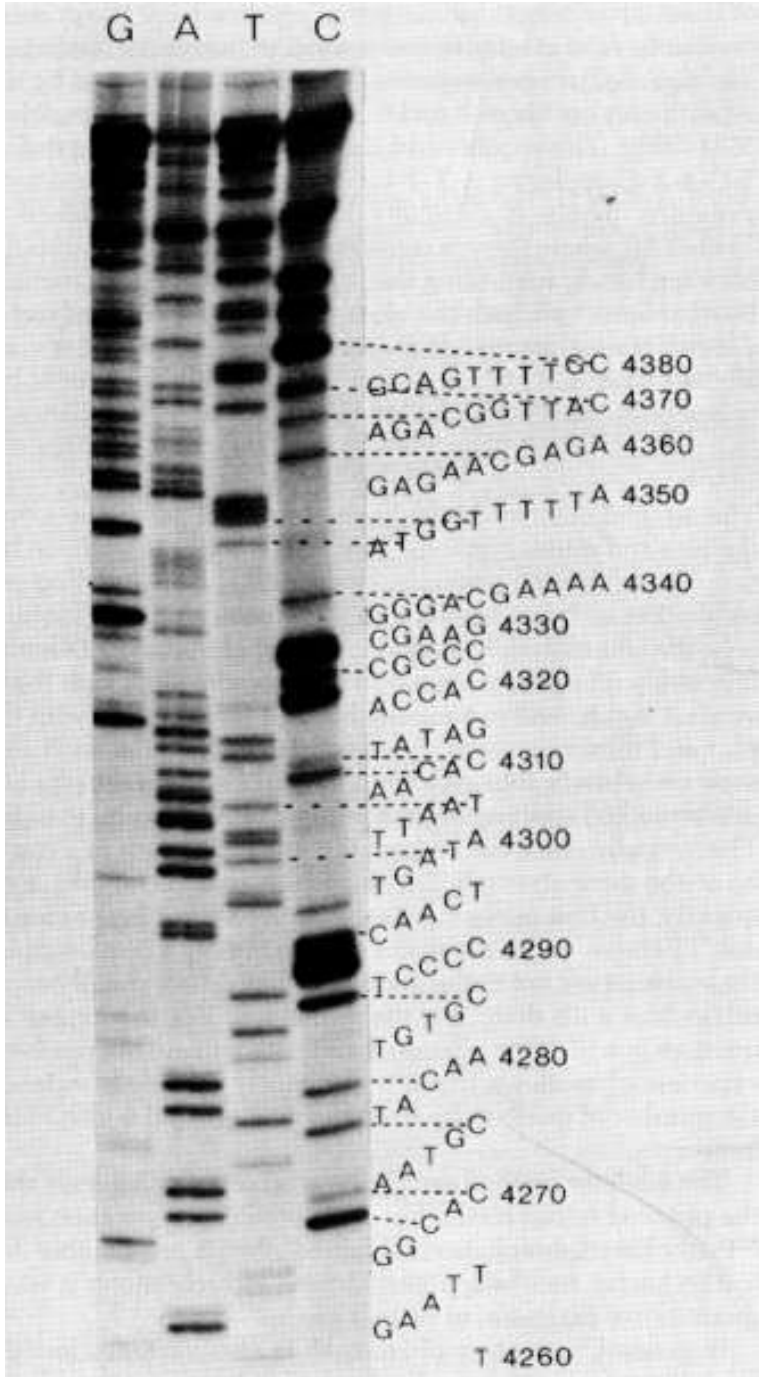
DNA sequencing with chain-terminating inhibitors

(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

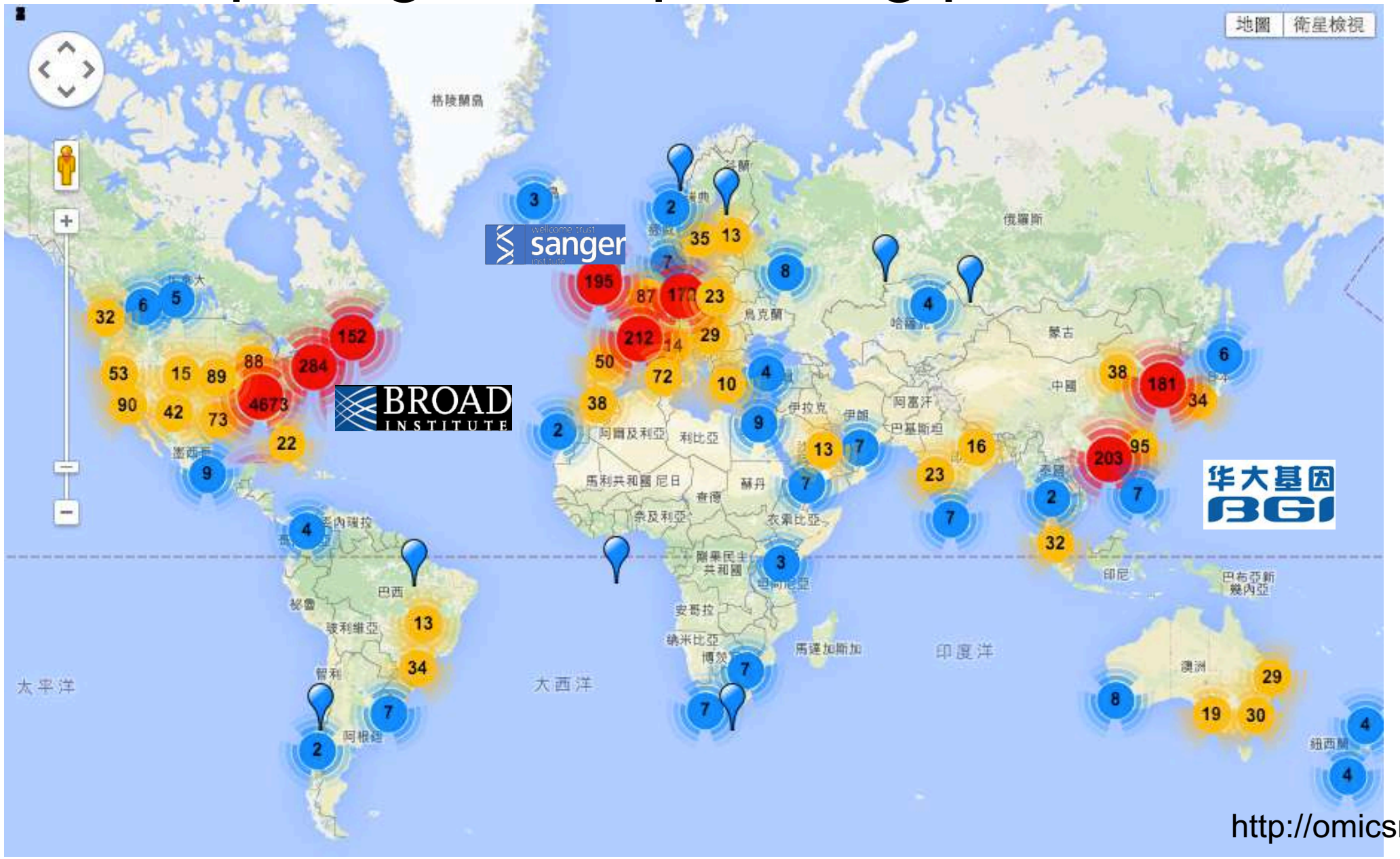


http://www.ncbi.nlm.nih.gov/pmc/articles/PMC431765/

ABI 3730xi at TIGR (1.6Mb per day)

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

World competing for sequencing power

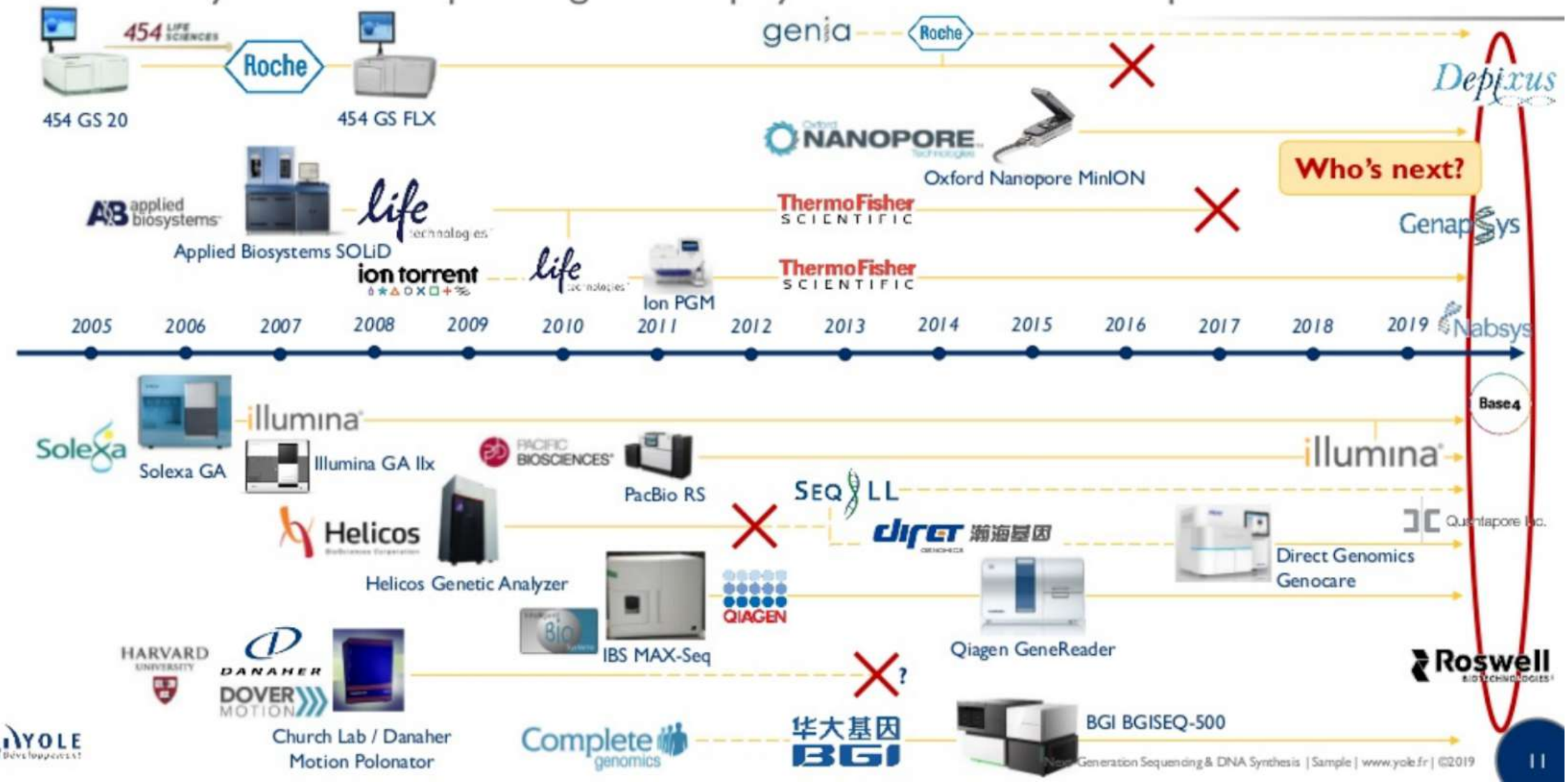


http://omicsmaps.com/

INTRODUCTION

Clip slide

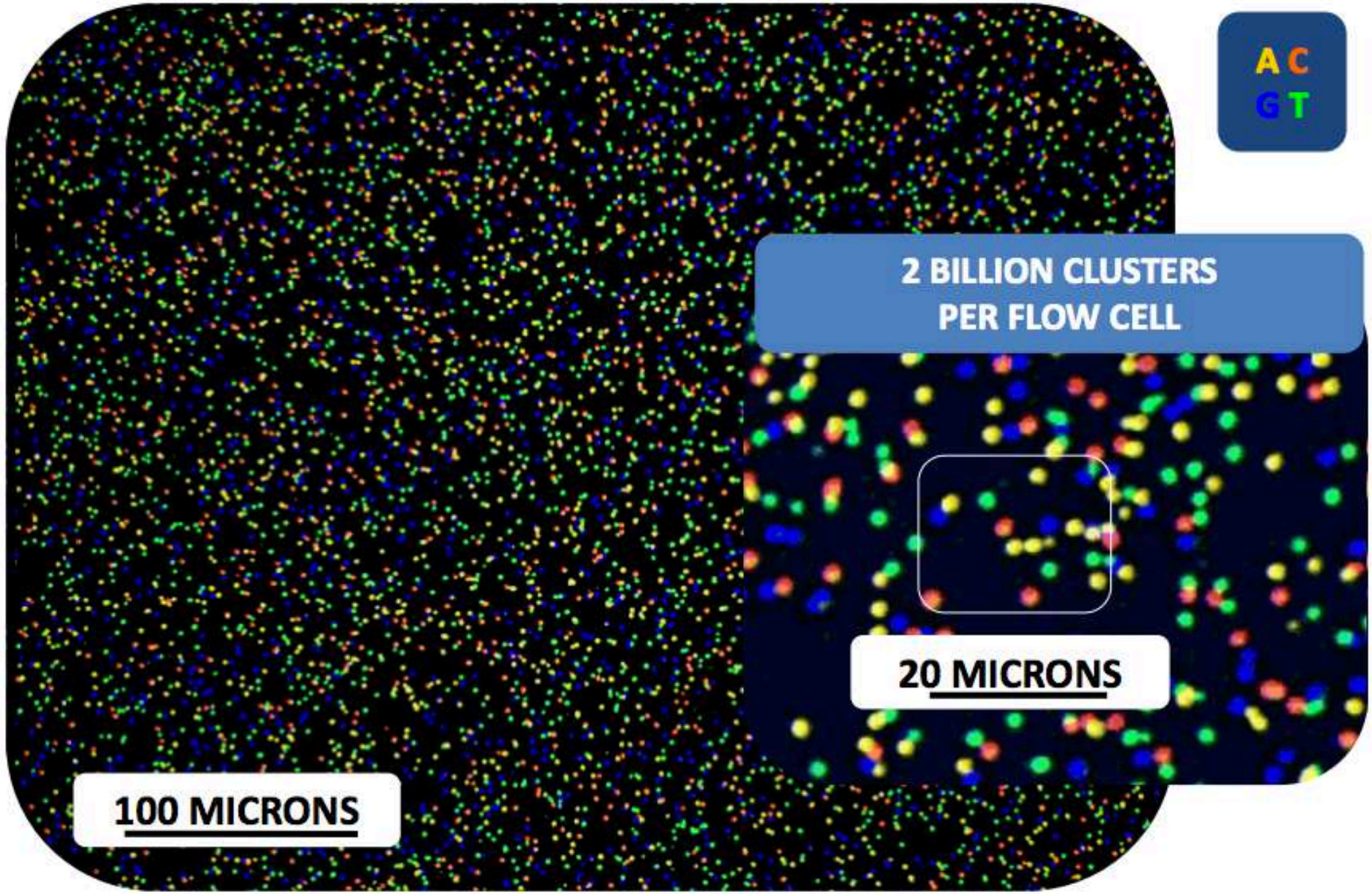
History of DNA sequencing – Main players' first commercial products and M&A

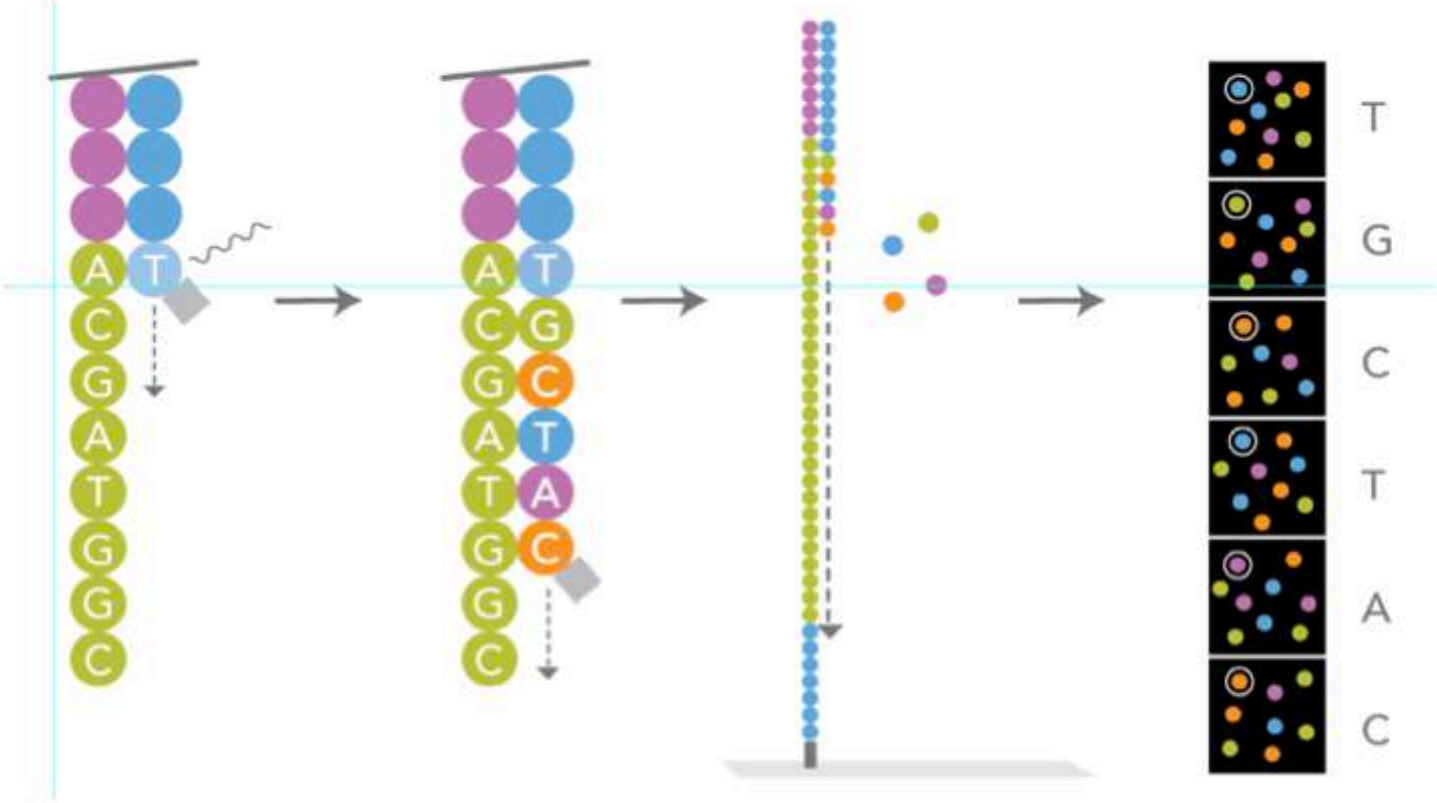


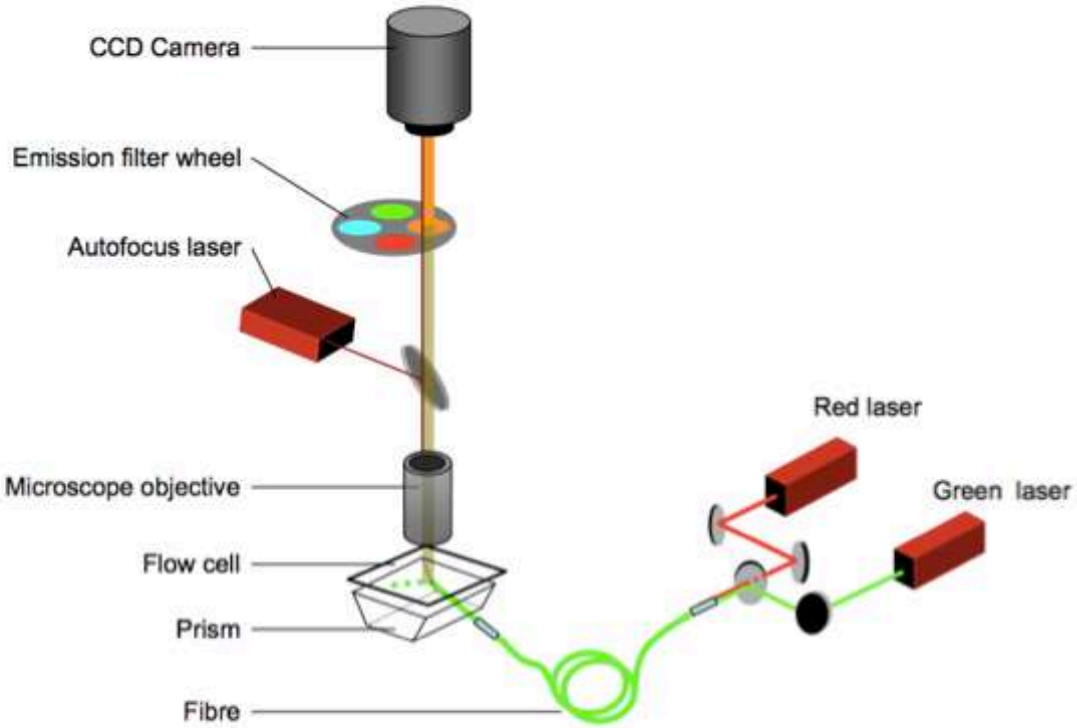
Illumina HiSeq

Sequencing by synthesis

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

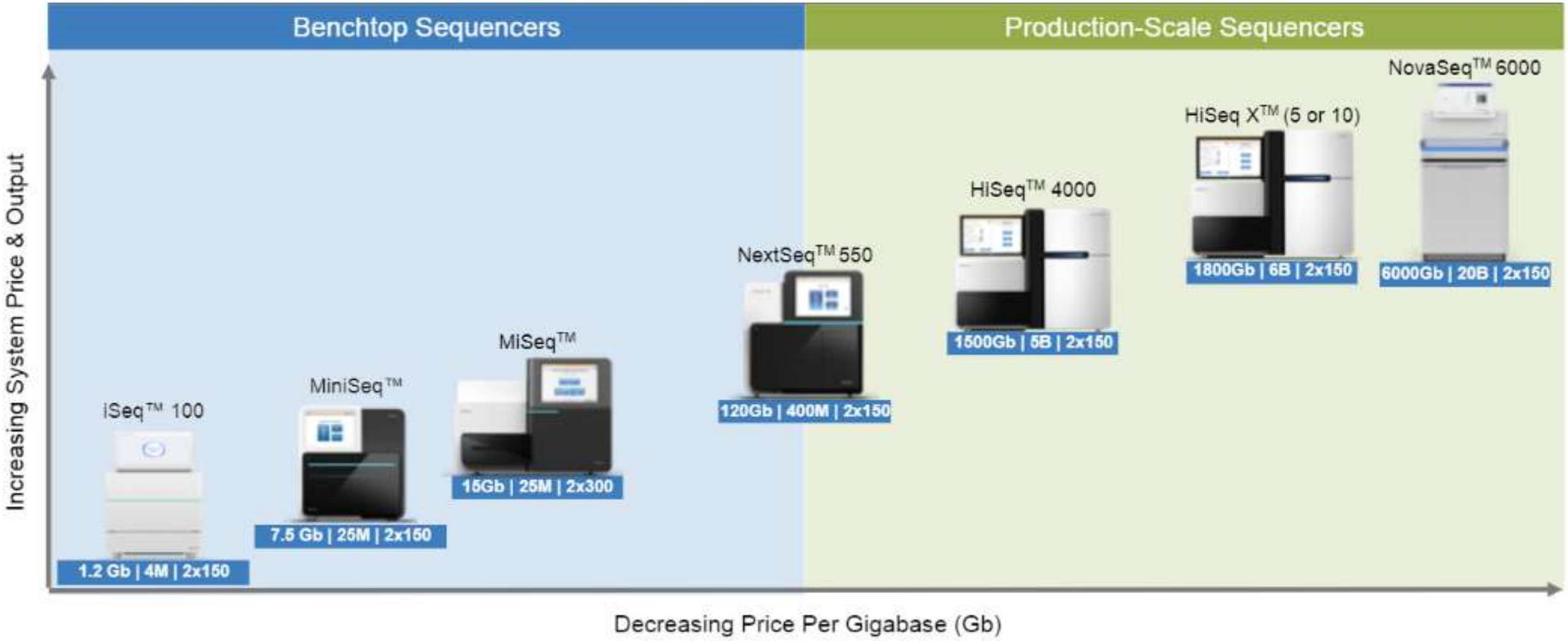

A C
G T
2 BILLION CLUSTERS
PER FLOW CELL
20 MICRONS
100 MICRONS





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

Illumina machines



And the arrival of 3rd generation sequencing…

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

PacBio (Pacific Biosciences)

RSII

Sequel II

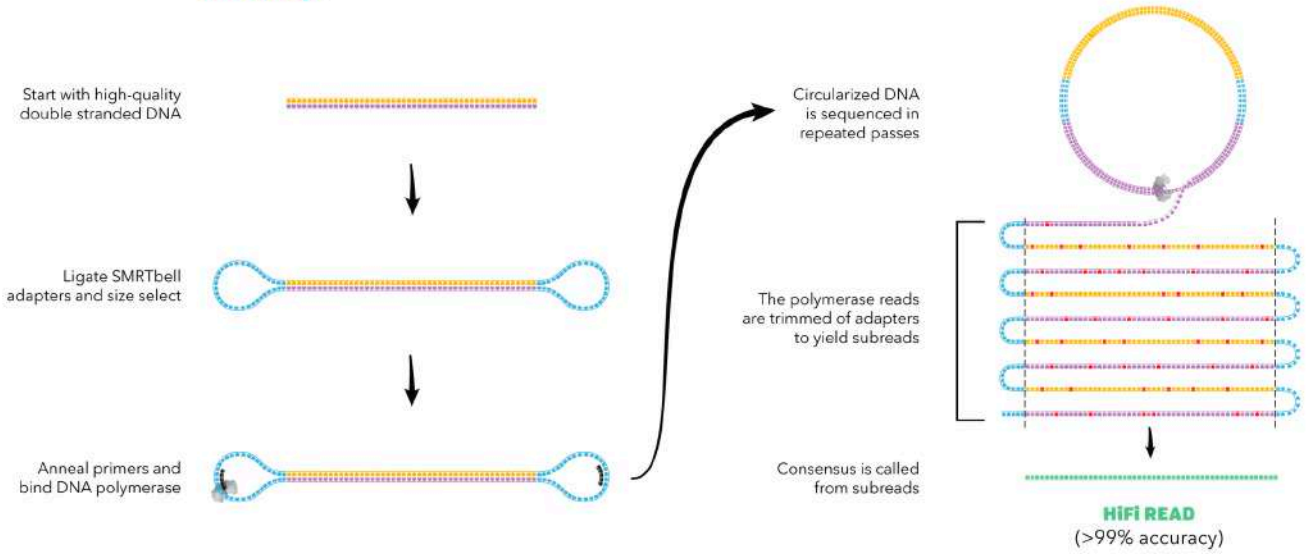
Single molecule sequencing

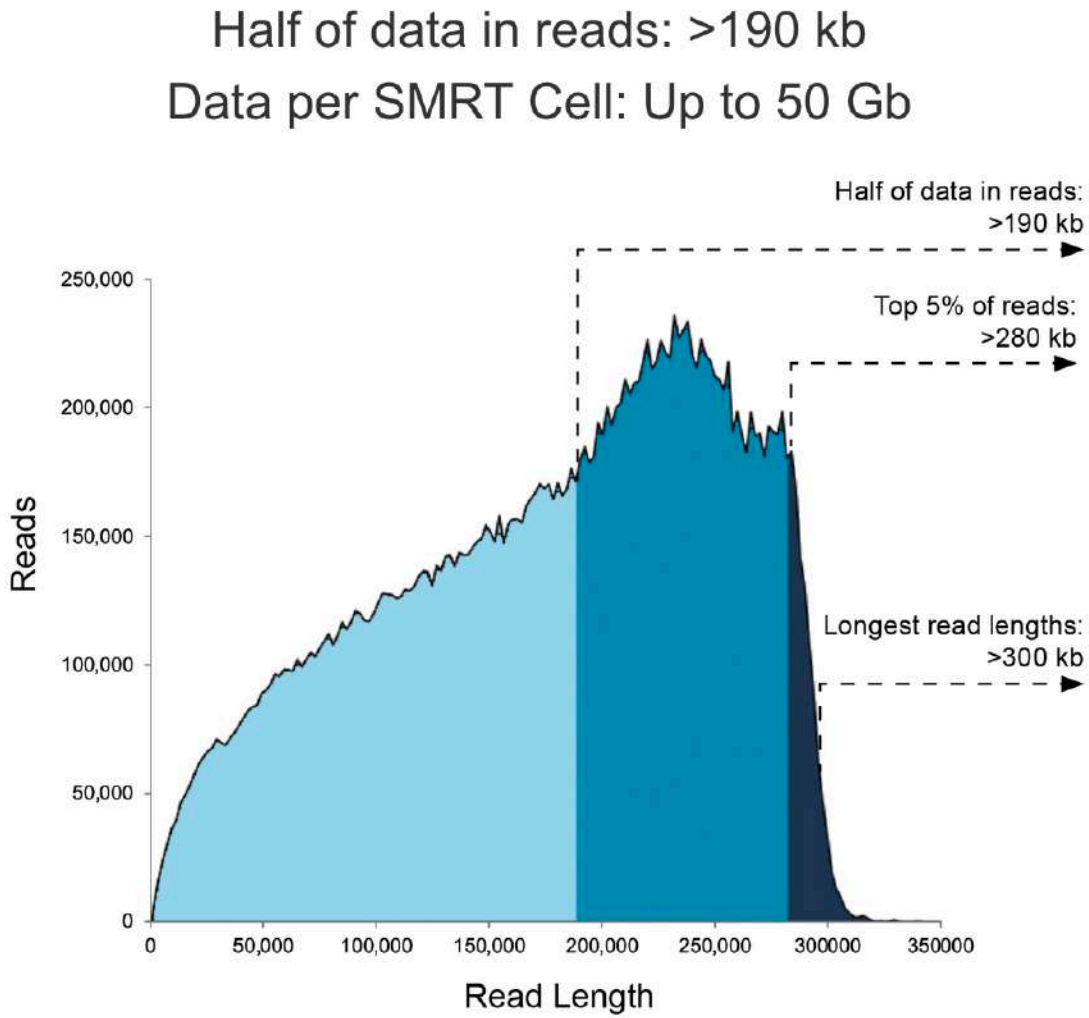
https://www.youtube.com/watch?v=NHCJ8PtYCFc

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

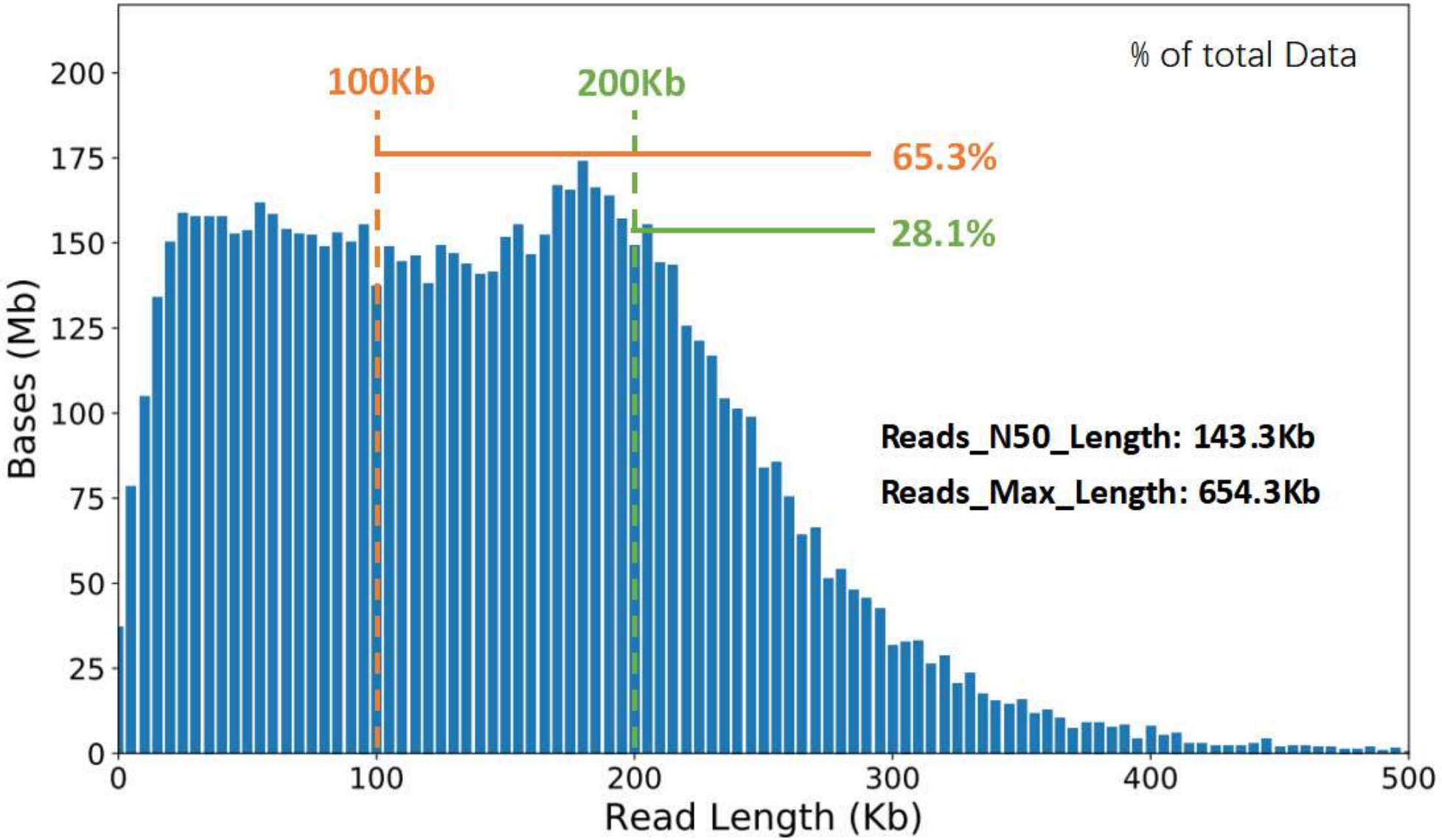
@ Oceans
https://vimeo.com/294744892

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

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

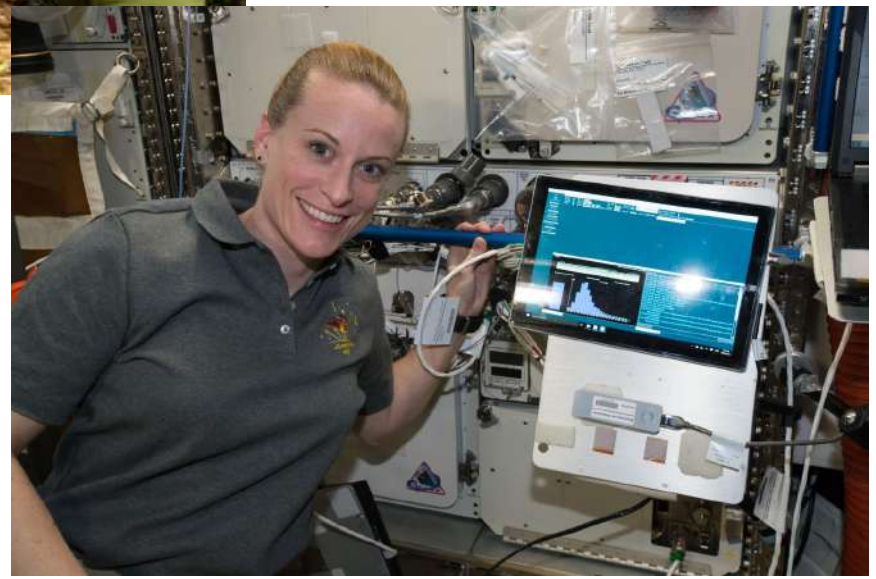
Nanopore Sequencing of Ebola Viruses Under Outbreak Conditions
https://www.youtube.com/watch?v=SYBzPEoENWI ; https://www.nature.com/articles/nature16996

Read length and capacity go beyond

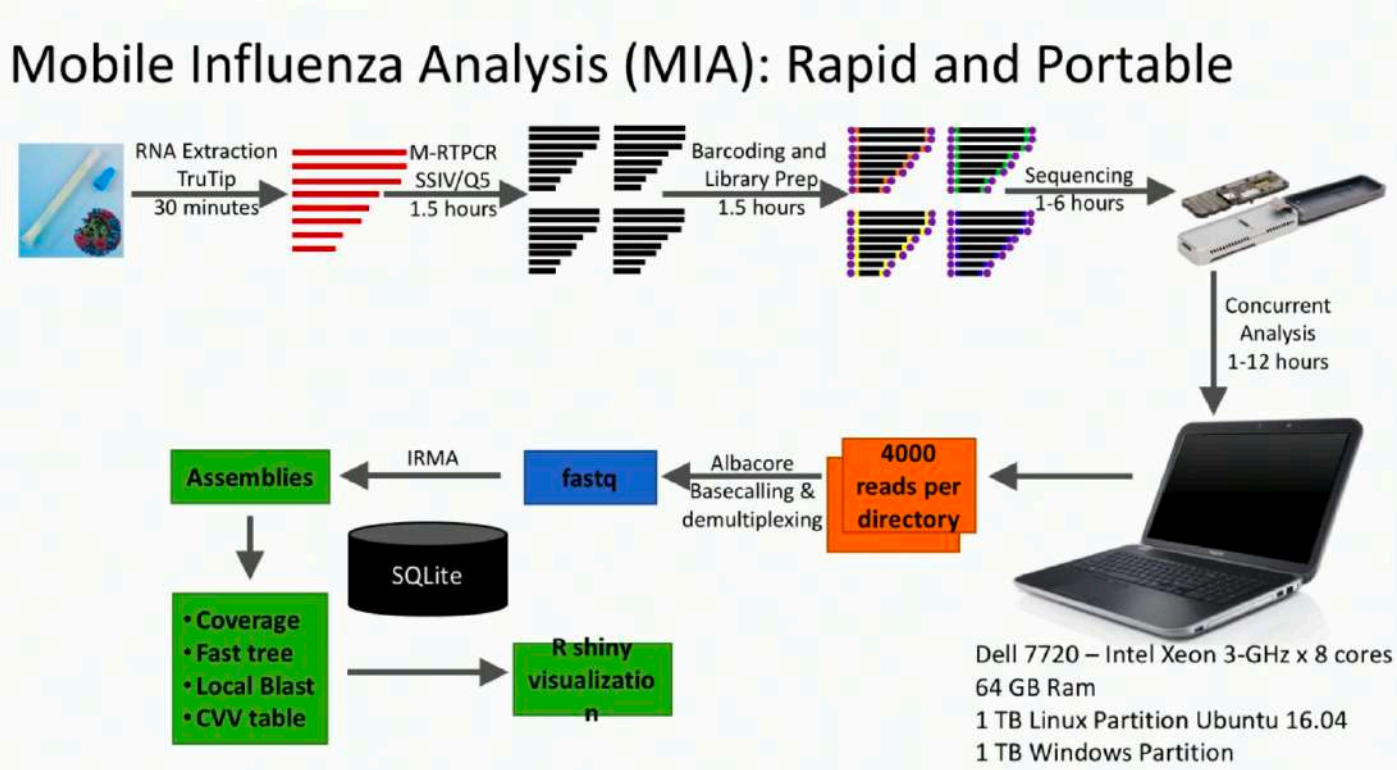


https://twitter.com/GrandOmics_Intl/status/1188724248480608257/photo/1

Mobility of sequencing

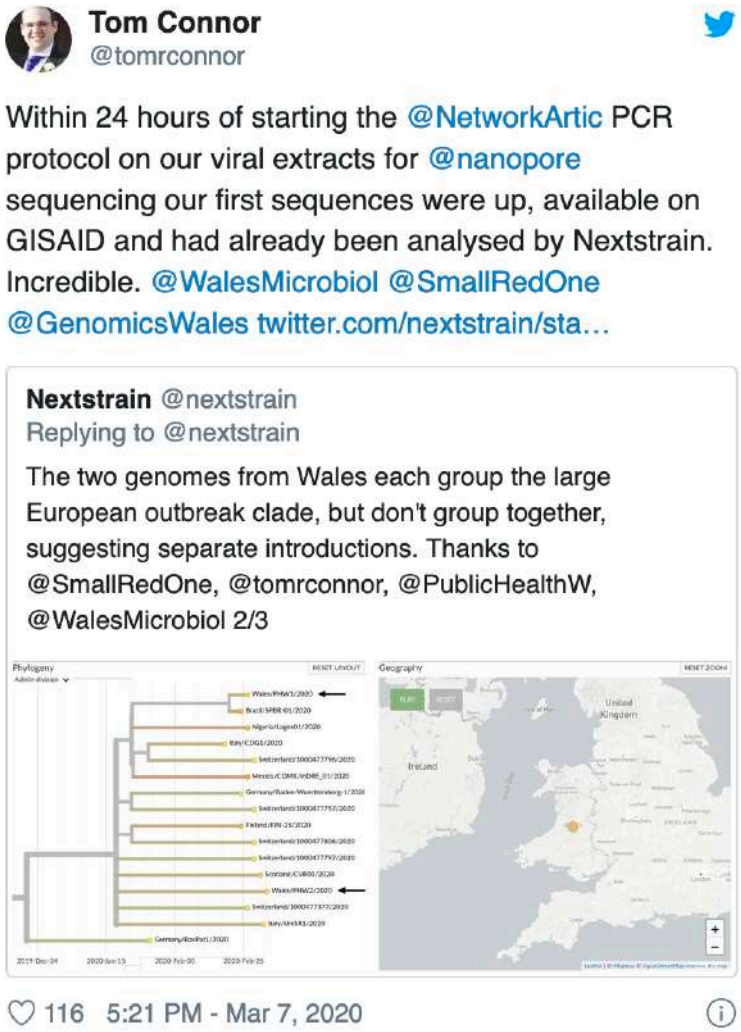


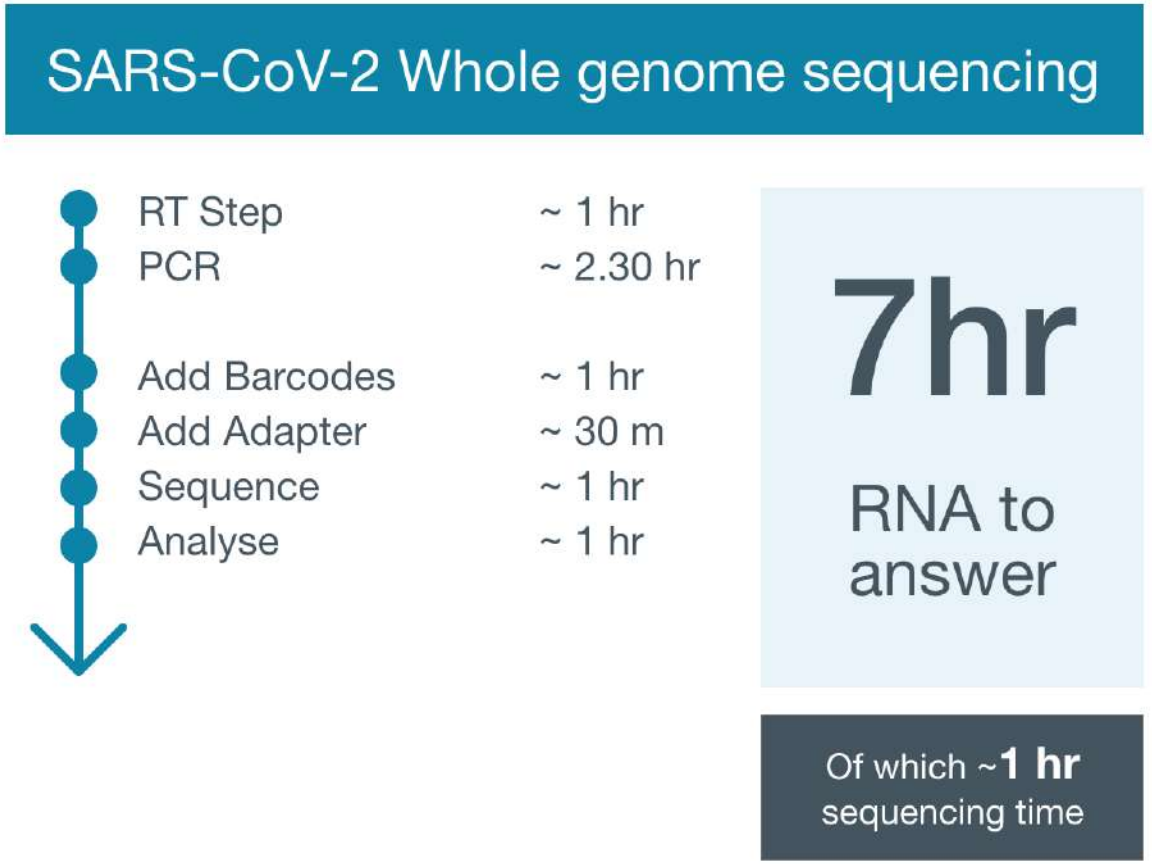
Matthew Keller: Deployable NGS for Influenza virus field surveillance and outbreak response



https://nanoporetech.com/resource-centre/matthew-keller-deployable-ngs-influenza-virus-field-surveillance-and-outbreak

Matthew Keller: Deployable NGS for Influenza virus field surveillance and outbreak response





https://nanoporetech.com/about-us/news/novel-coronavirus-covid-19-information-and-updates

https://nanoporetech.com/about-us/news/covid19-community

ARTIC NETWORK

About

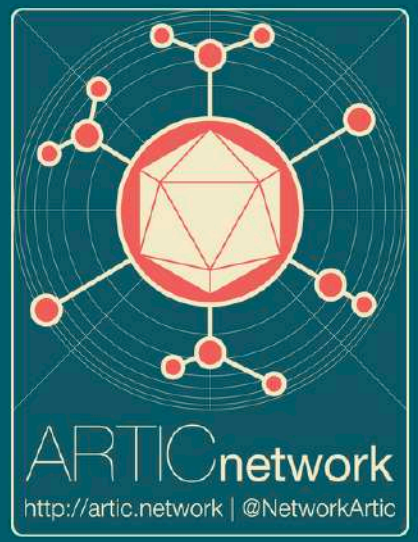
The Project

This project is developing an end-to-end system for processing samples from viral outbreaks to generate real-time epidemiological information that is interpretable and actionable by public health bodies. Fast evolving RNA viruses (such as Ebola, MERS, SARS, influenza etc) continually accumulate changes in their genomes that can be used to reconstruct the epidemiological processes that drive the epidemic. Based around a recently developed, single-molecule portable sequencing instrument, the Oxford Nanopore Technology MinION, we are creating a 'lab-in-a-suitcase' that can be deployed to remote and resource-limited locations. Targeting a wide-range of emerging viral diseases, the sequencing generation will be closely linked to the analysis platform to integrate these data and associated epidemiological knowledge to reveal the processes of transmission, virus evolution and epidemiological linkage with extremely rapid turn-around. This real-time approach will provide actionable epidemiological insights within days of samples being taken from patients.

https://artic.network/ncov-2019

ARTIC NETWORK

nCoV-2019



There is a pressing need to understand more about the short-term genomic epidemiology and evolution of the recently described novel coronavirus (nCoV-2019). Initial cases were in Wuhan City, Hubei Province, China but now cases have been confirmed both more widely in China and internationally.

Viral genome data generated prospectively during outbreaks can help provide information about relatedness to other viruses, mode and tempo of evolution, geographical spread and adaptation to human hosts. This information can be used to assist in epidemiological investigations, particularly when combined with other types of data (e.g. case counts).

The ARTIC network is making available a set of materials (see below) to assist groups in sequencing the virus including a set of primers, laboratory protocols, bioinformatics tutorials and datasets. These are mainly focused around the use of the portable Oxford Nanopore MinION sequencer, although aspects of the protocol such as the primer scheme and sample amplification may be generalised to other sequencing platforms.

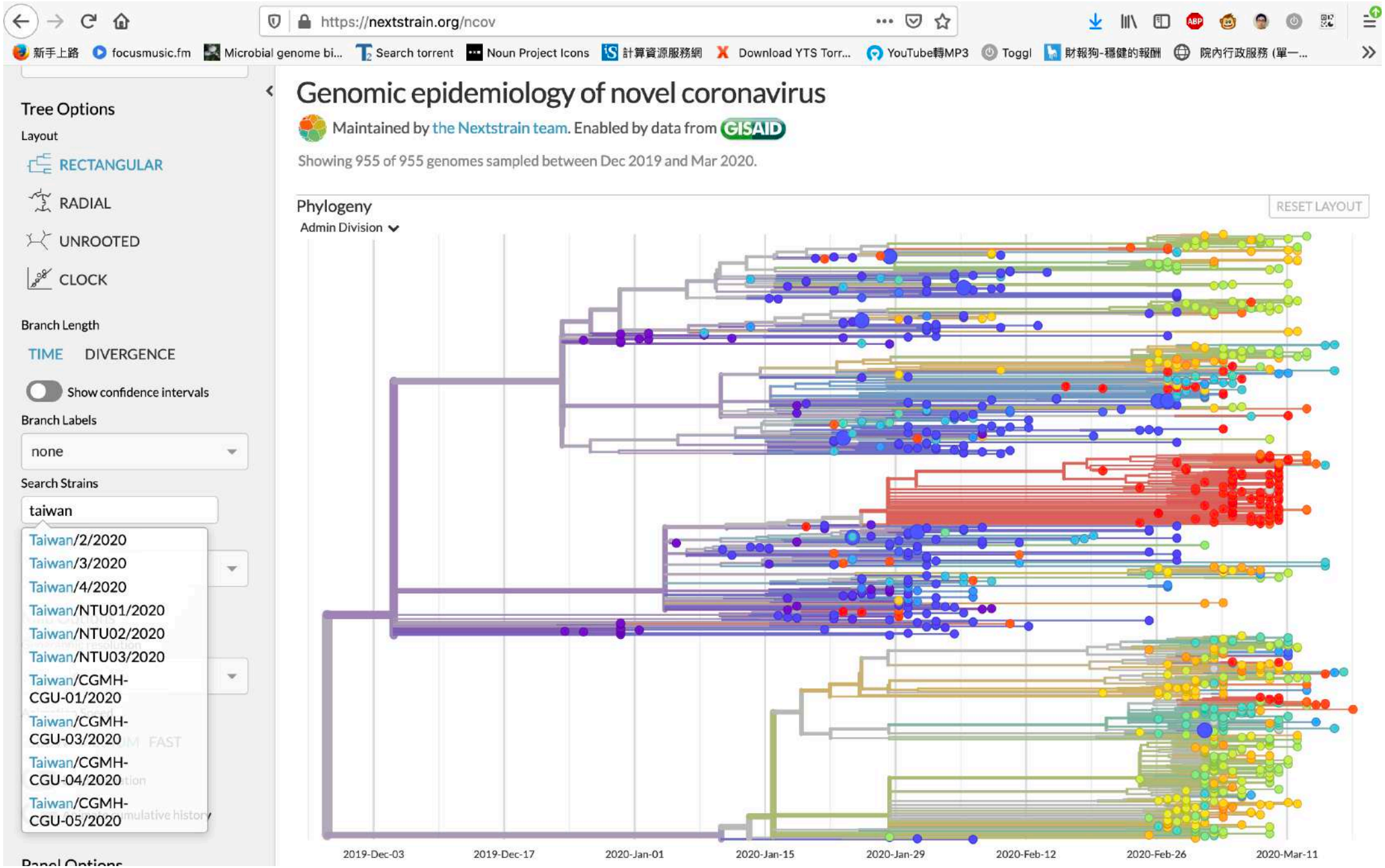
https://artic.network/ncov-2019

Nextstrain

Real-time tracking of pathogen evolution

Nextstrain is an open-source project to harness the scientific and public health potential of pathogen genome data. We provide a continually-updated view of publicly available data alongside powerful analytic and visualization tools for use by the community. Our goal is to aid epidemiological understanding and improve outbreak response. If you have any questions, or simply want to say hi, please give us a shout at hello@nextstrain.org.

https://nextstrain.org/



https://nextstrain.org/

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

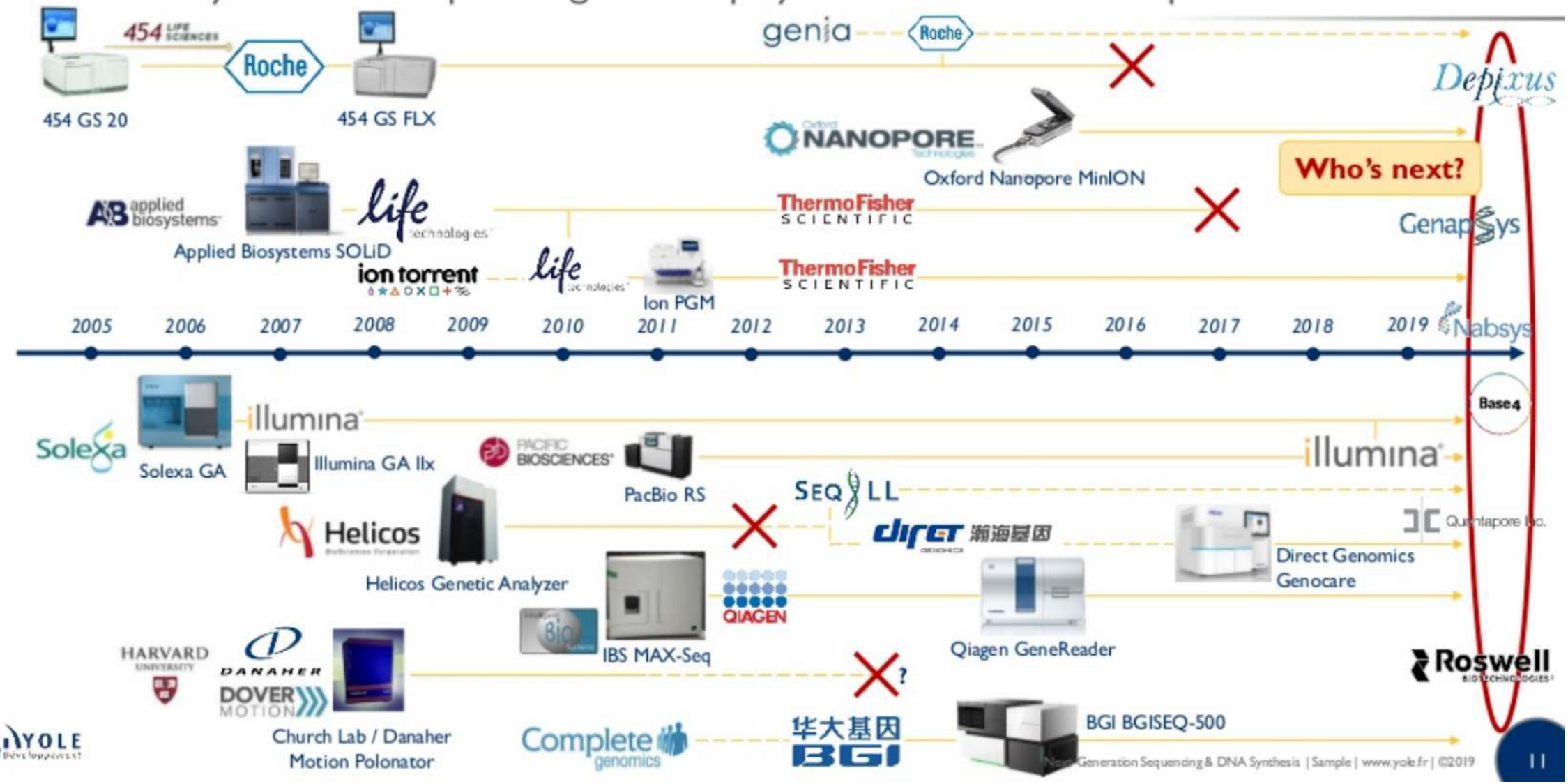
1. [lab/hospital/mountain/sea] Collect samples -> report

Minute

INTRODUCTION

Clip slide

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



Break here

Three situations you are most likely to encounter

Genome reference is available (for example, humans):

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

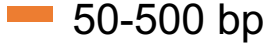
Genome reference is NOT available

- **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

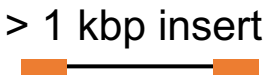


Read

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

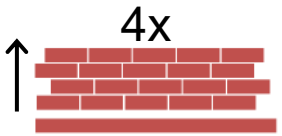
The length of the DNA fragment

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? (mapping)

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

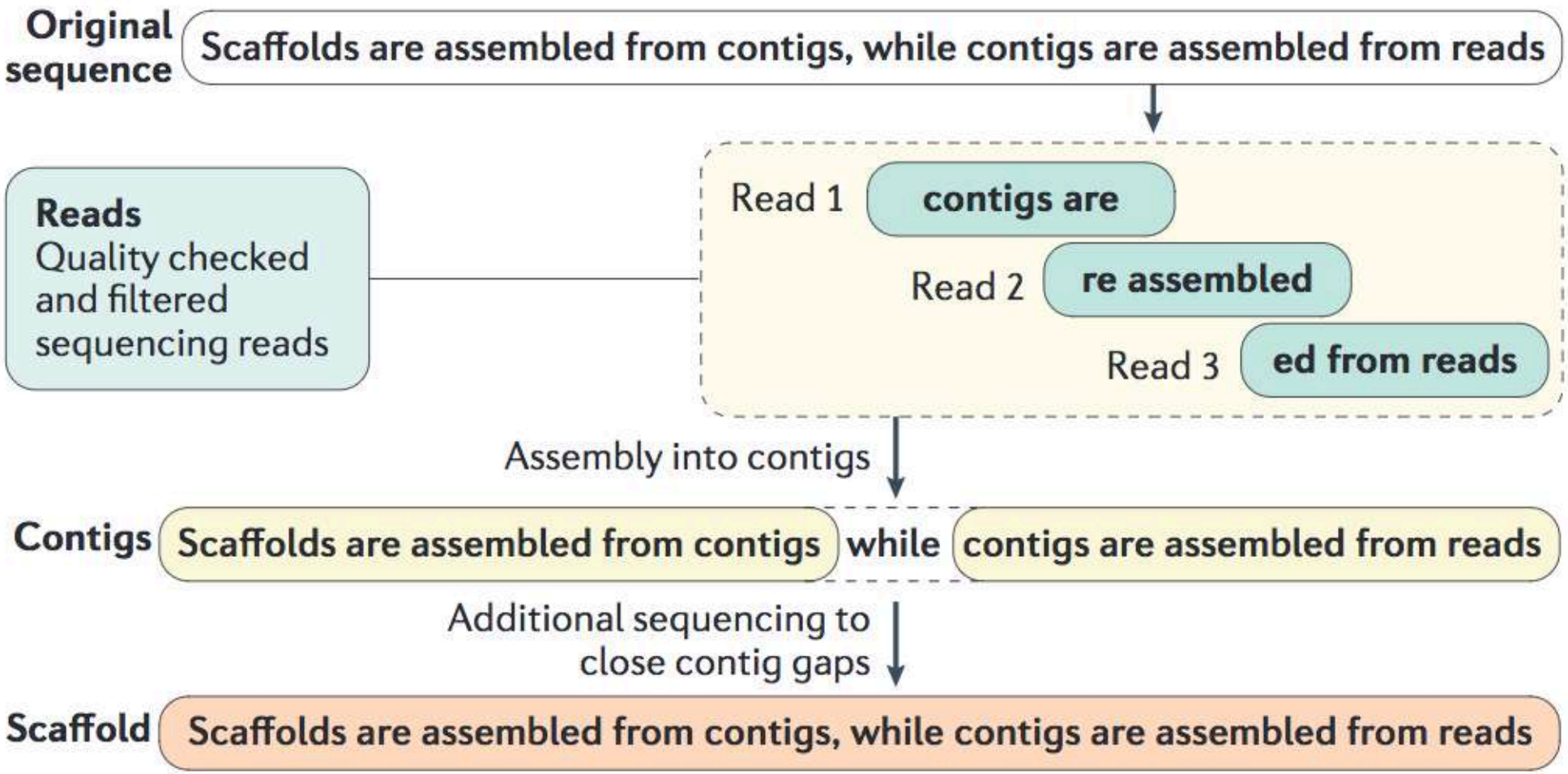
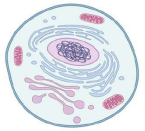


Figure 2 | **Sequence read assembly.** A mock example explaining bioinformatic sequence assembly along with the terms sequence, reads, contigs and scaffolds.

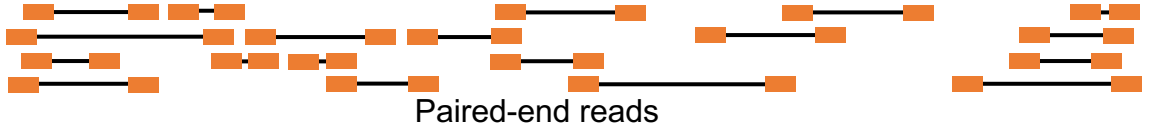
Assembly



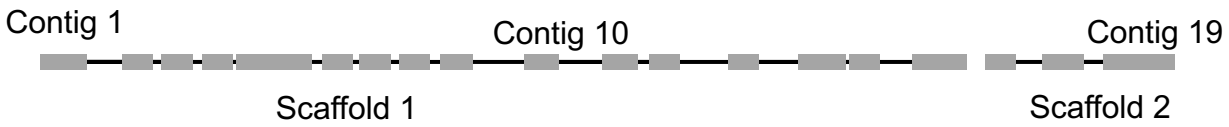
Genome

Fragment

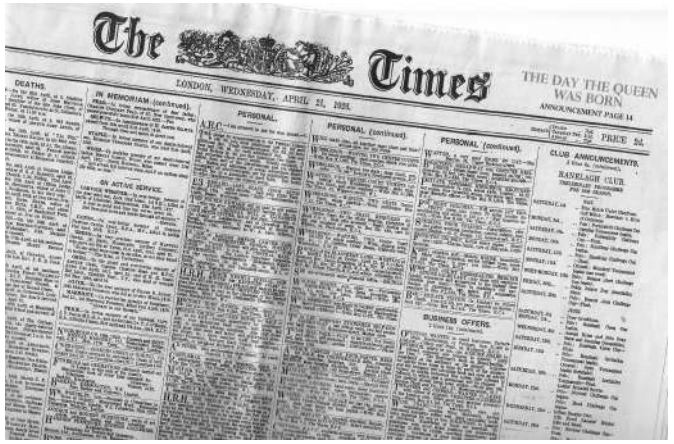
Sequence



Assemble



Assembly

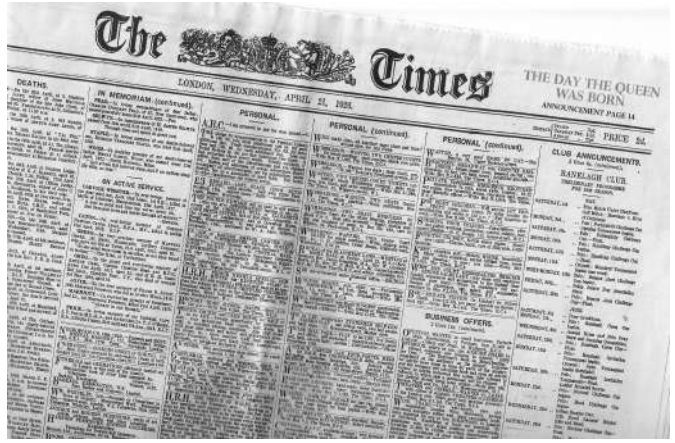


Genome
(3.000.000 letters)

Sequencing →

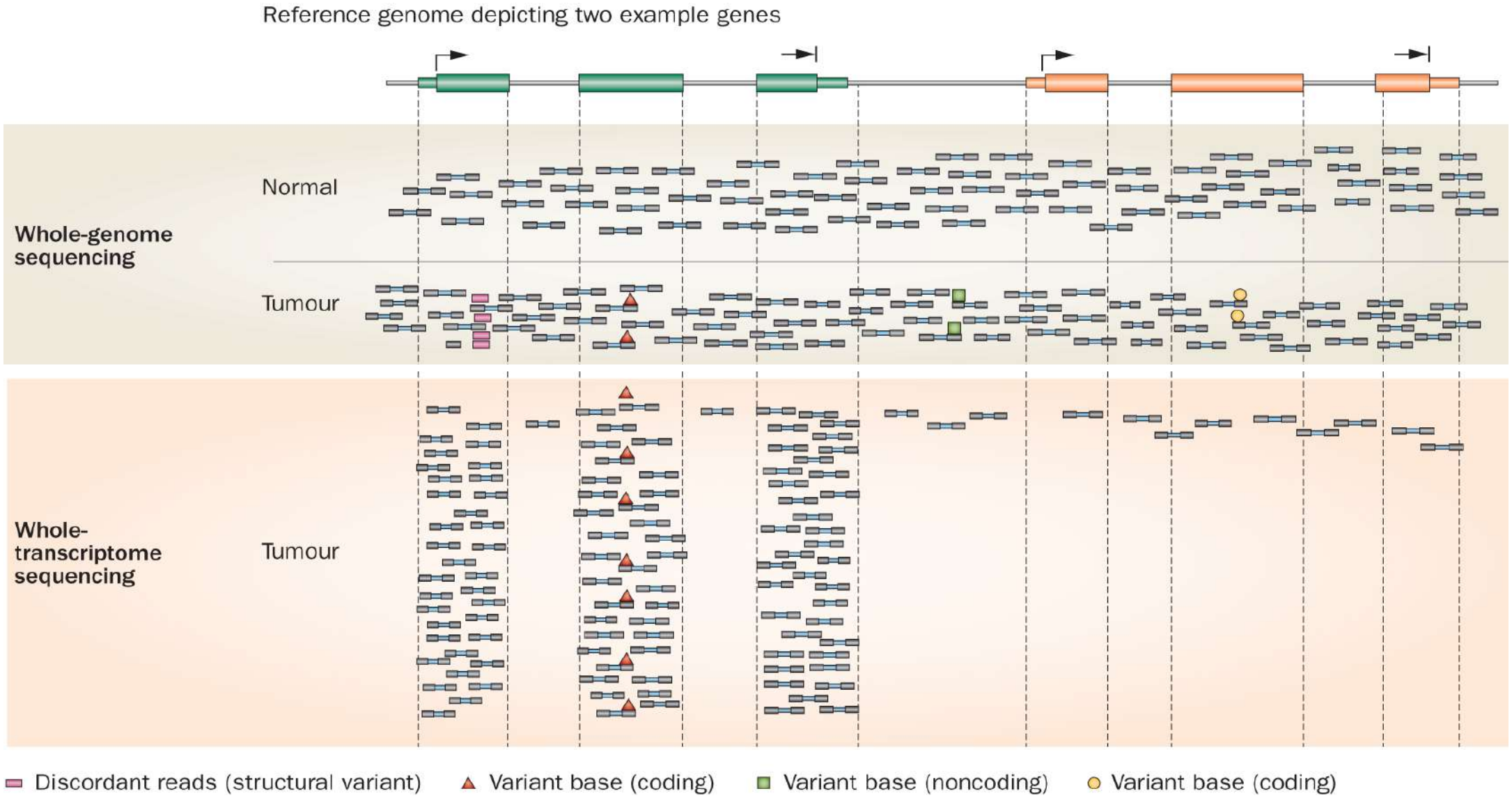
Reads
(50-500 letters each)

Assembly →



Genome
(3.000.000 letters)

Mapping



doi:10.1038/nrgastro.2012.126

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.

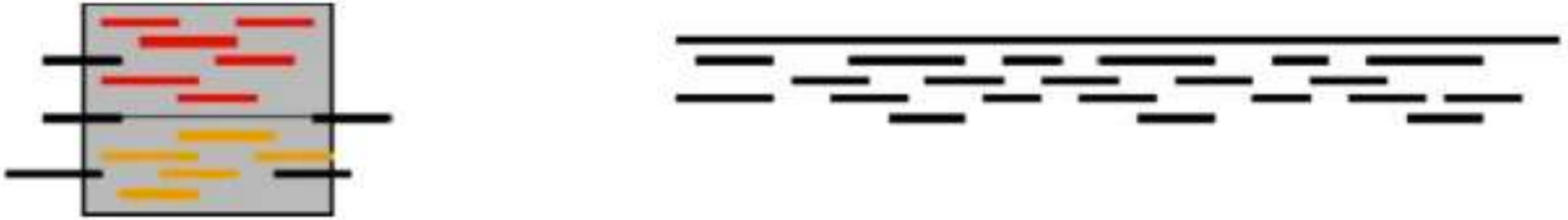
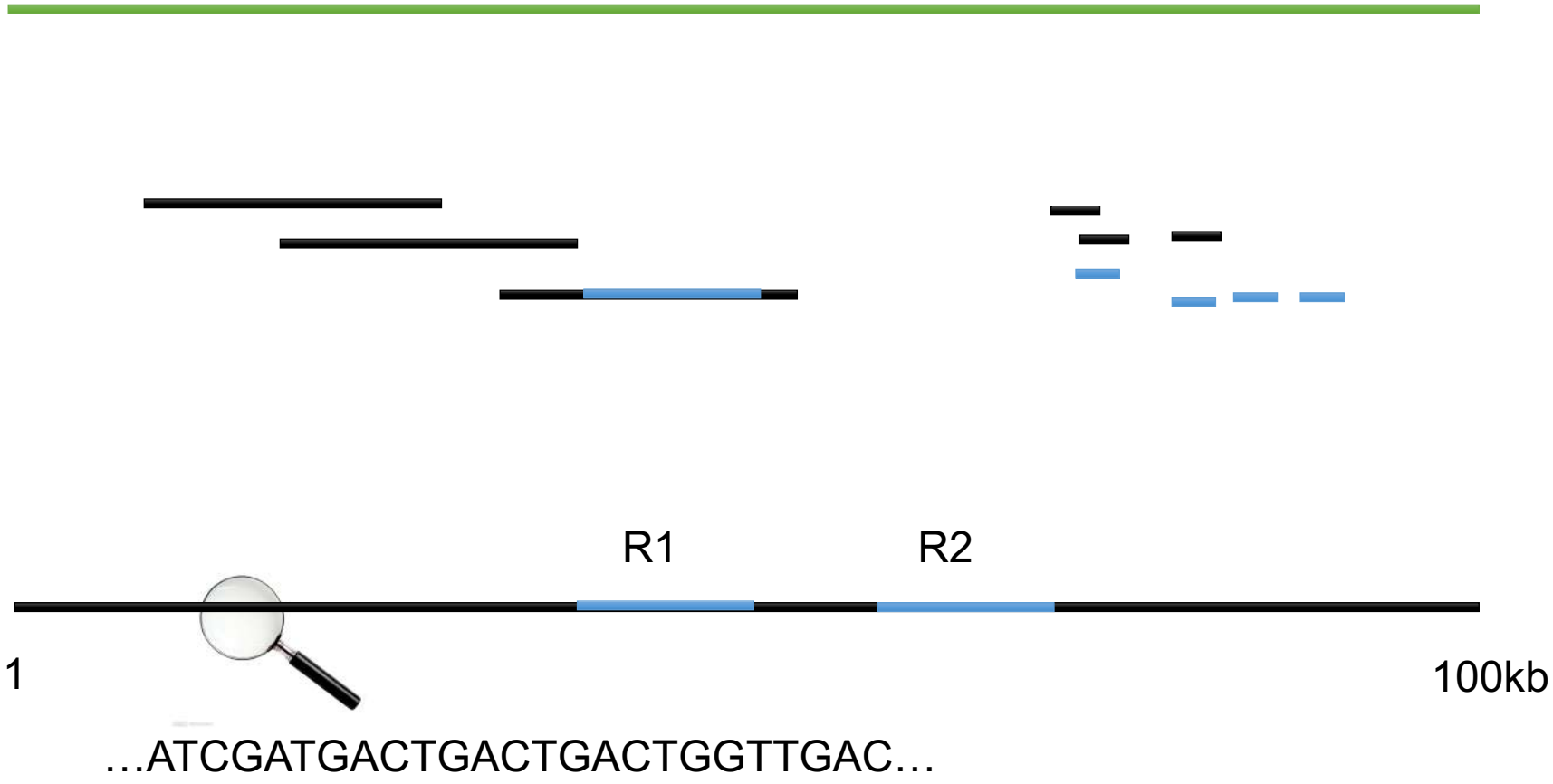


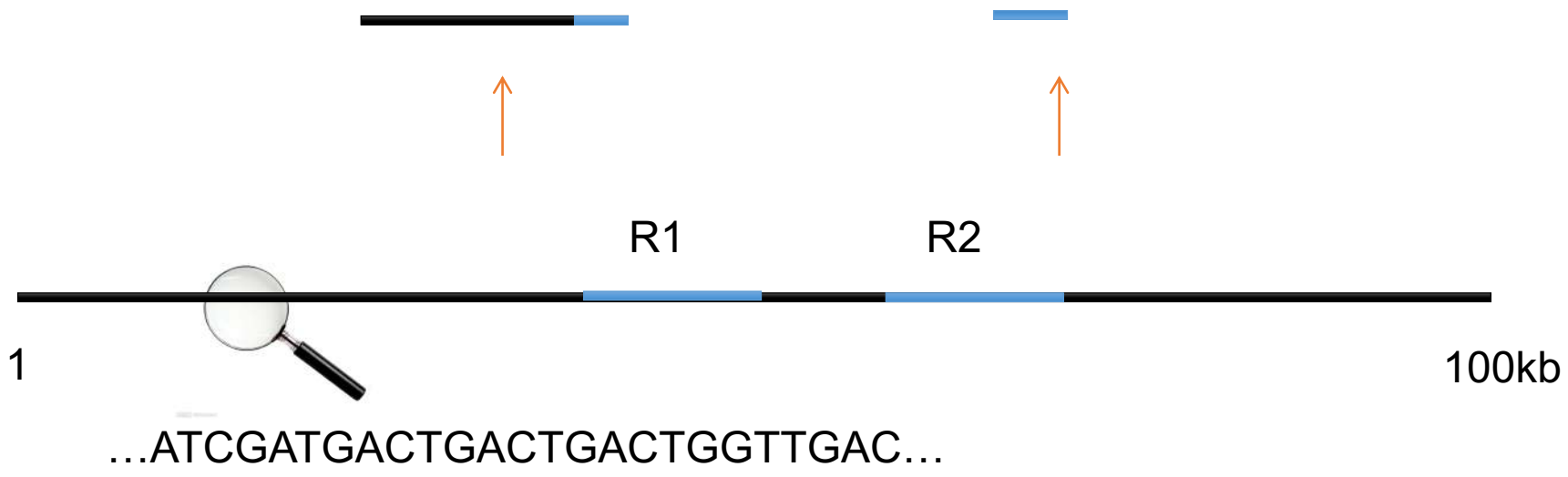
Figure 6. Genome mis-assembled due to a repeat. The assembly program incorrectly combined the reads from the two copies of the repeat leading to the creation of two separate contigs

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

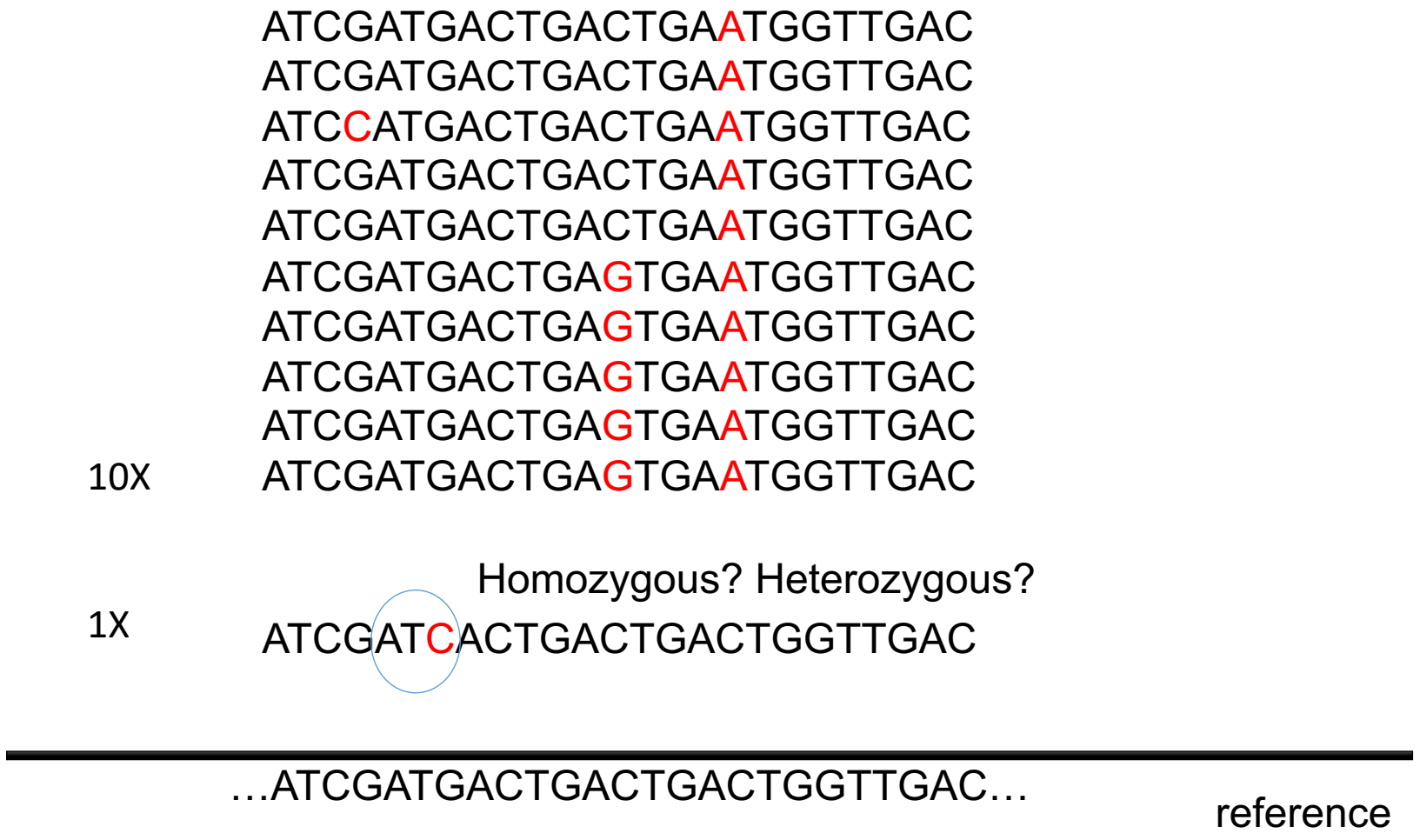
Read length matters in sequencing



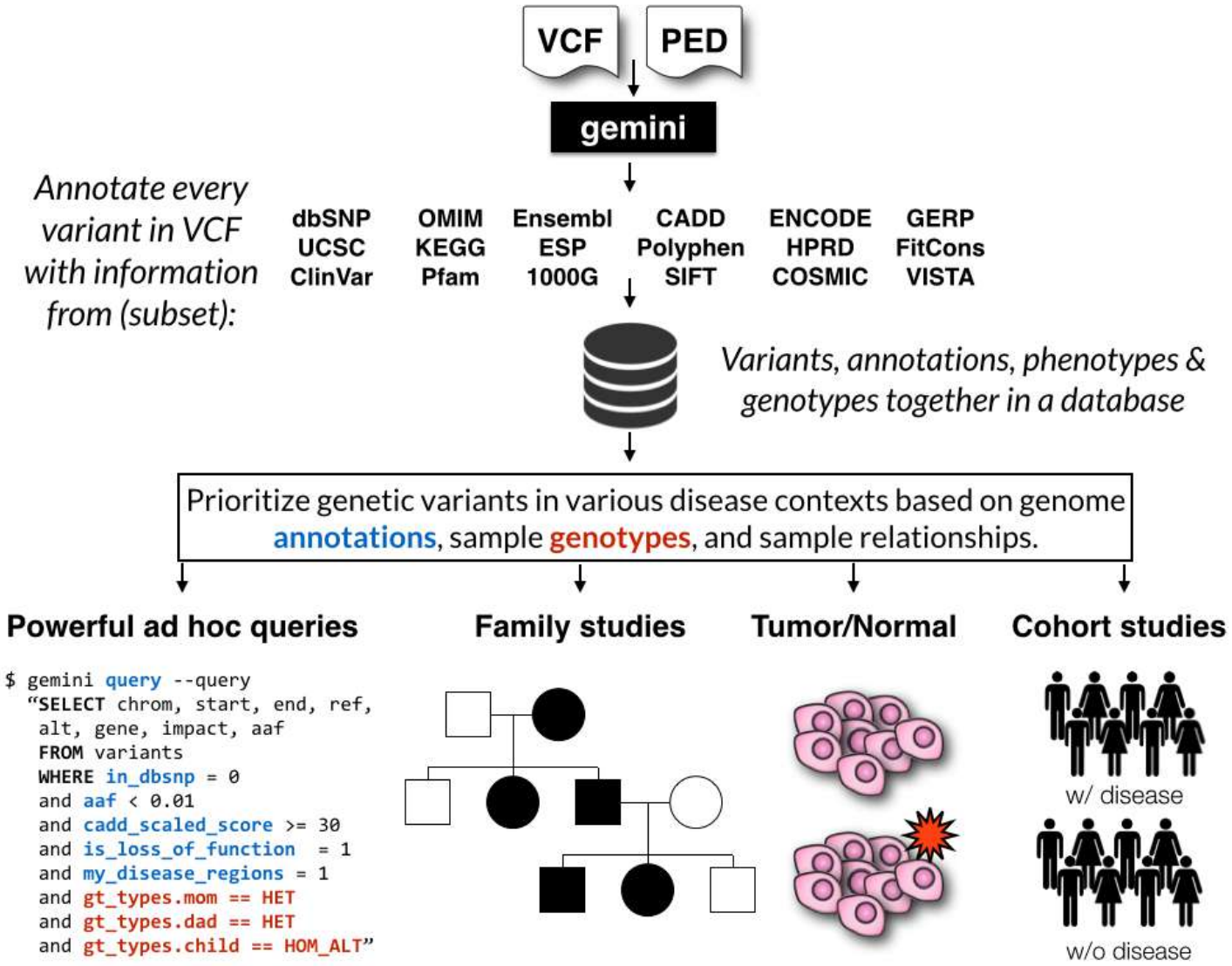
Paired end and insert size matter in sequencing



Depth matters in sequencing



Filtering and annotating variants




```
$ gemini query --query
  "SELECT chrom, start, end, ref,
   alt, gene, impact, aaf
   FROM variants
   WHERE in_dbsnp = 0
   and aaf < 0.01
   and cadd_scaled_score >= 30
   and is_loss_of_function  = 1
   and my_disease_regions = 1
   and gt_types.mom == HET
   and gt_types.dad == HET
   and gt_types.child == HOM_ALT"
```

ClinVar
SIFT
Function
MAF (Minor Allele frequency)

https://gemini.readthedocs.io/en/latest/

Case studies

Classical genetics

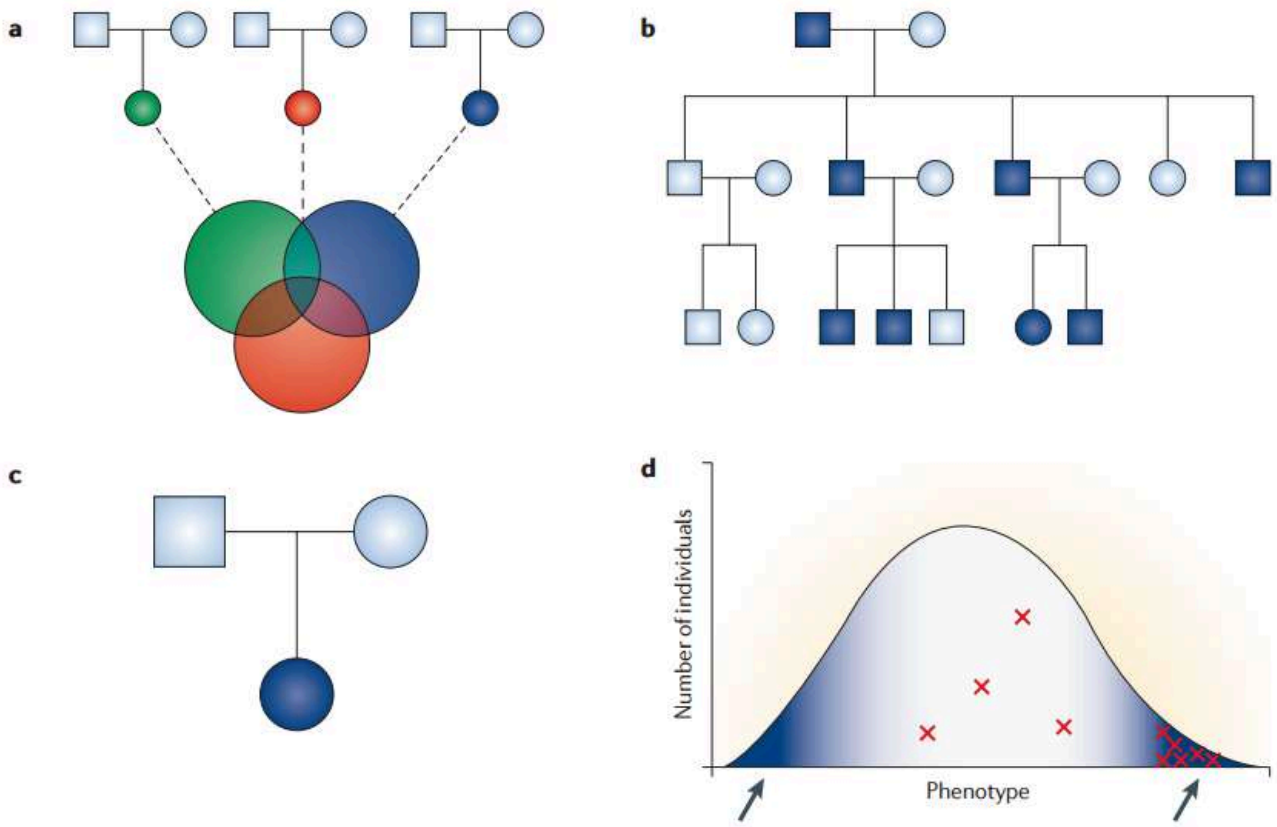
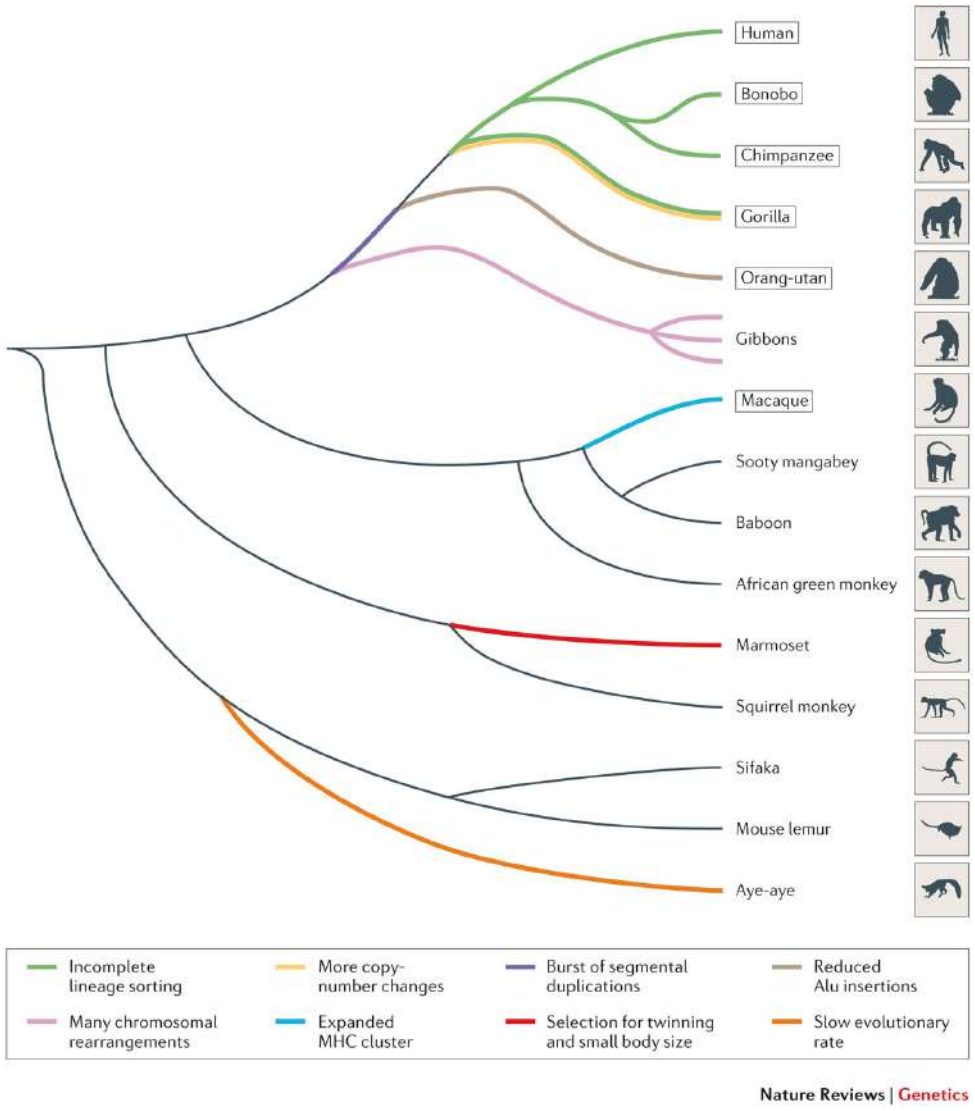
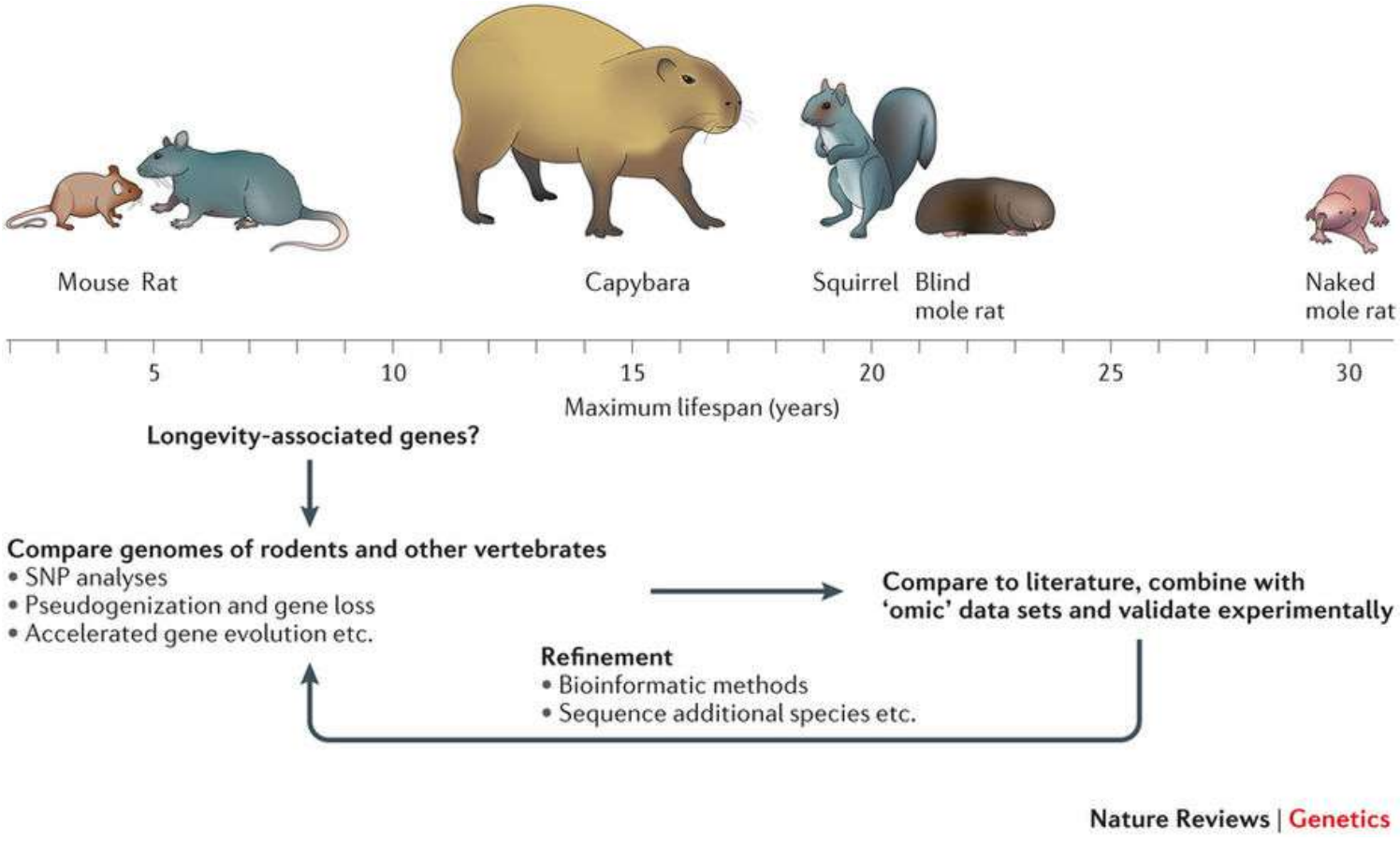


Figure 2 | **Strategies for finding disease-causing rare variants using exome sequencing.** Four main strategies are illustrated. **a** | Sequencing and filtering across multiple unrelated, affected individuals (indicated by the three coloured circles). This approach is used to identify novel variants 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

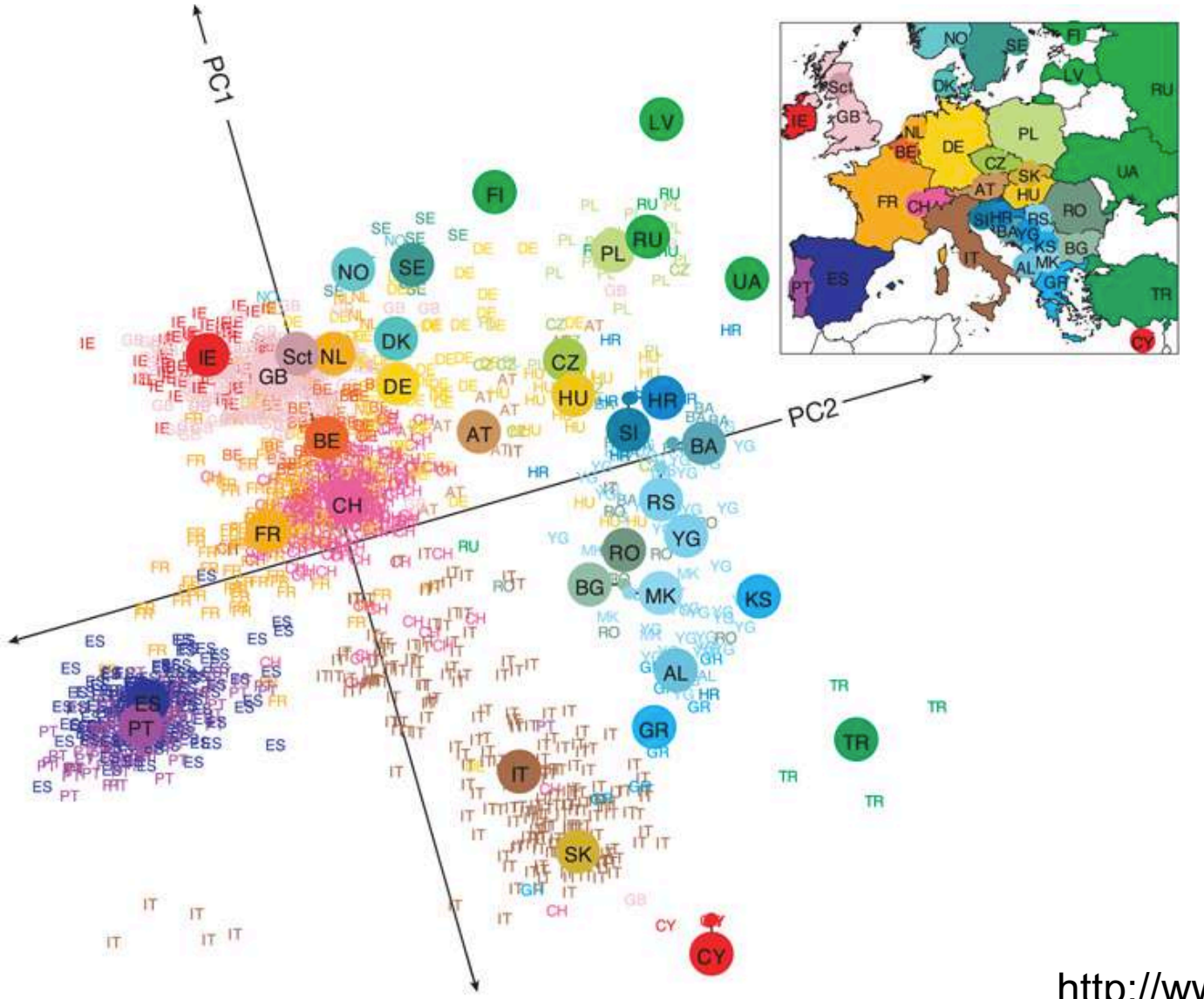




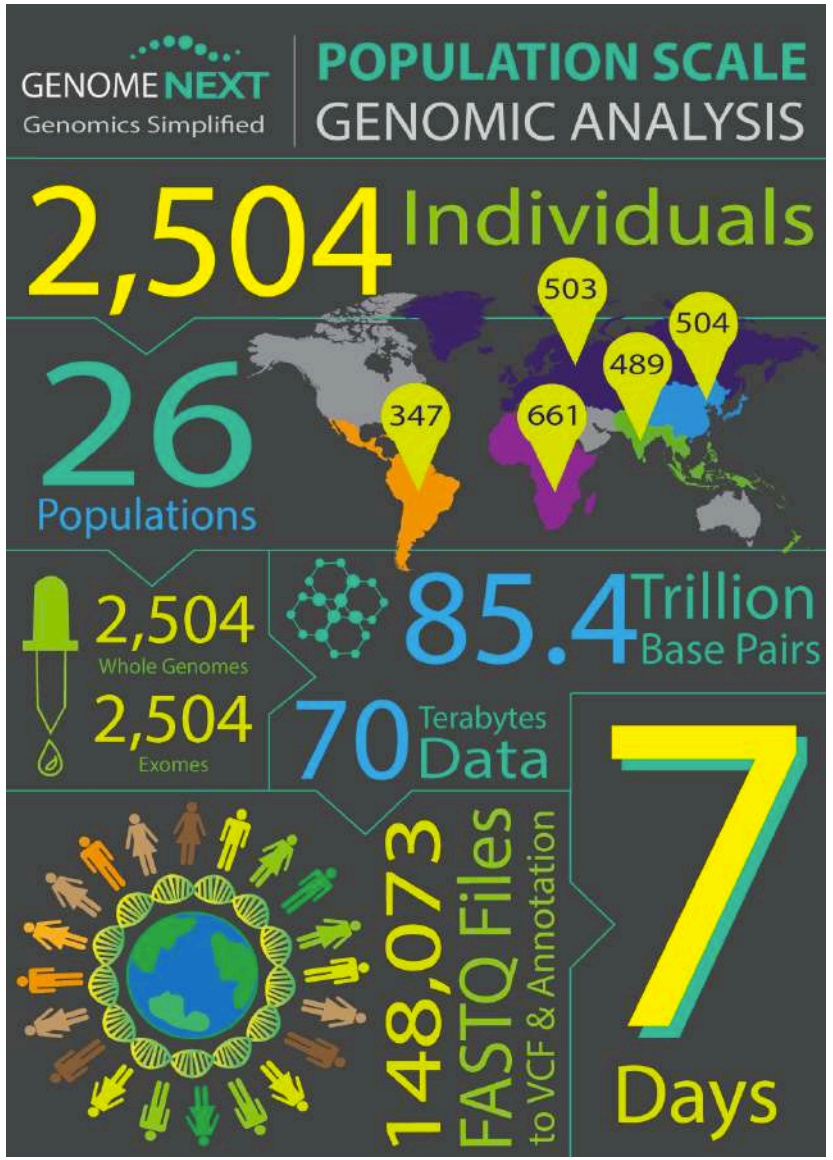
Roger & Gibbs Nature Reviews Genetics (2014)

doi:10.1038/nrg3728

Population genomics

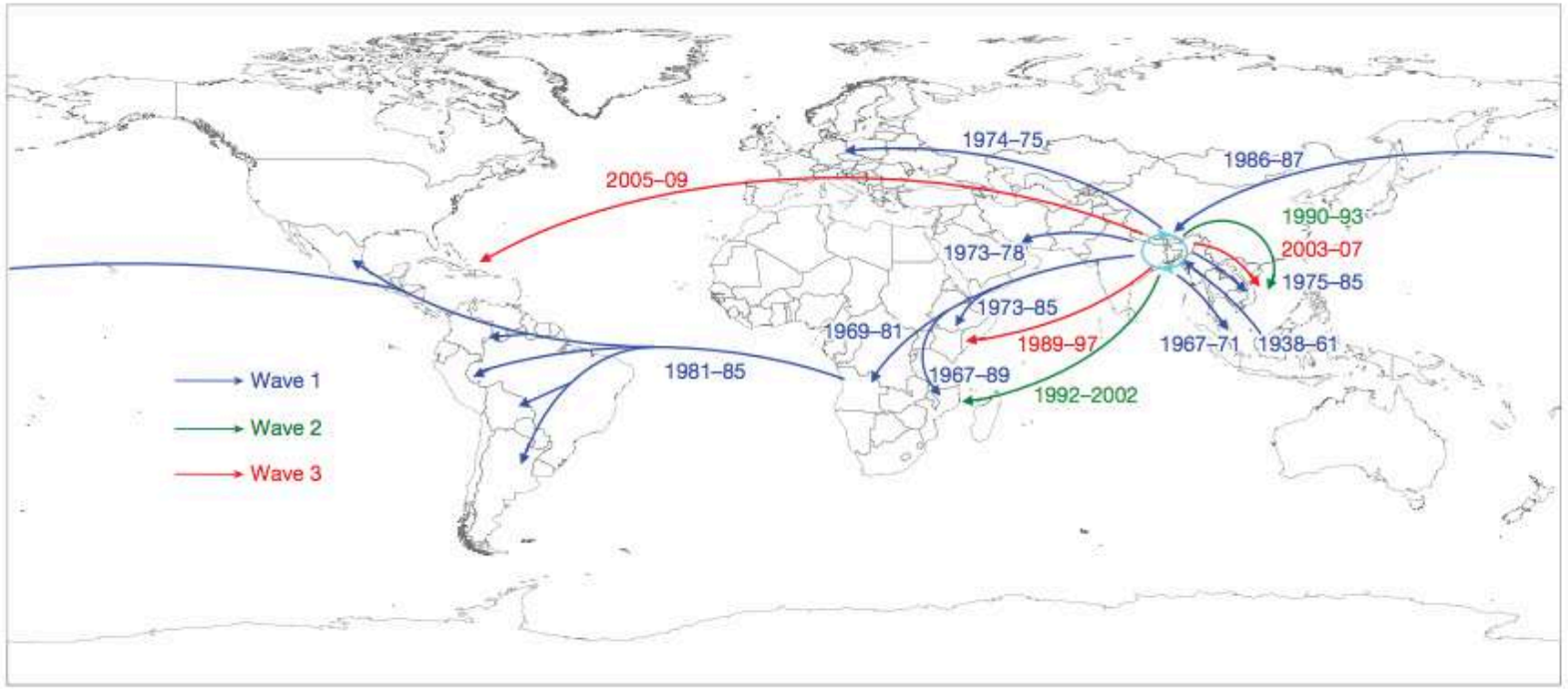


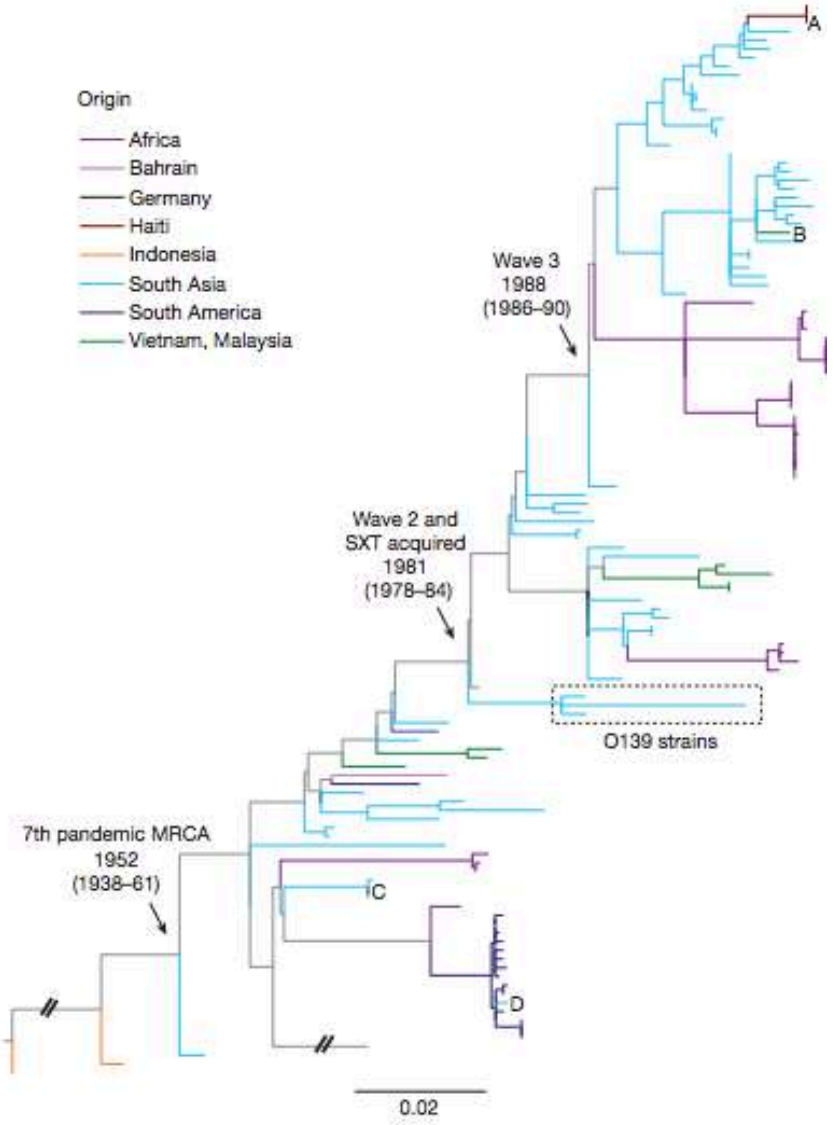
Novembre et al Nature (2008)



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。

UK launches whole genome sequence alliance to map spread of coronavirus

The Wellcome Sanger Institute will collaborate with expert groups across the country to analyse the genetic code of COVID-19 samples circulating in the UK, providing public health agencies with a unique tool to combat the virus

COVID-19 Genomics UK Consortium - comprised of the NHS, Public Health Agencies, Wellcome Sanger Institute, and numerous academic institutions - will deliver large scale, rapid sequencing of the cause of the disease and share intelligence with hospitals, regional NHS centres and the Government.

Samples from patients with confirmed cases of COVID-19 will be sent to a network of sequencing centres which currently includes Belfast, Birmingham, Cambridge, Cardiff, Edinburgh, Exeter, Glasgow, Liverpool, London, Norwich, Nottingham, Oxford and Sheffield.

The Wellcome Sanger Institute, one of the world's most advanced centres of genomes and data, will collaborate with expert groups across the country to analyse the genetic code of COVID-19 samples circulating in the UK and in doing so, give public health agencies and clinicians a unique, cutting-edge tool to combat the virus.

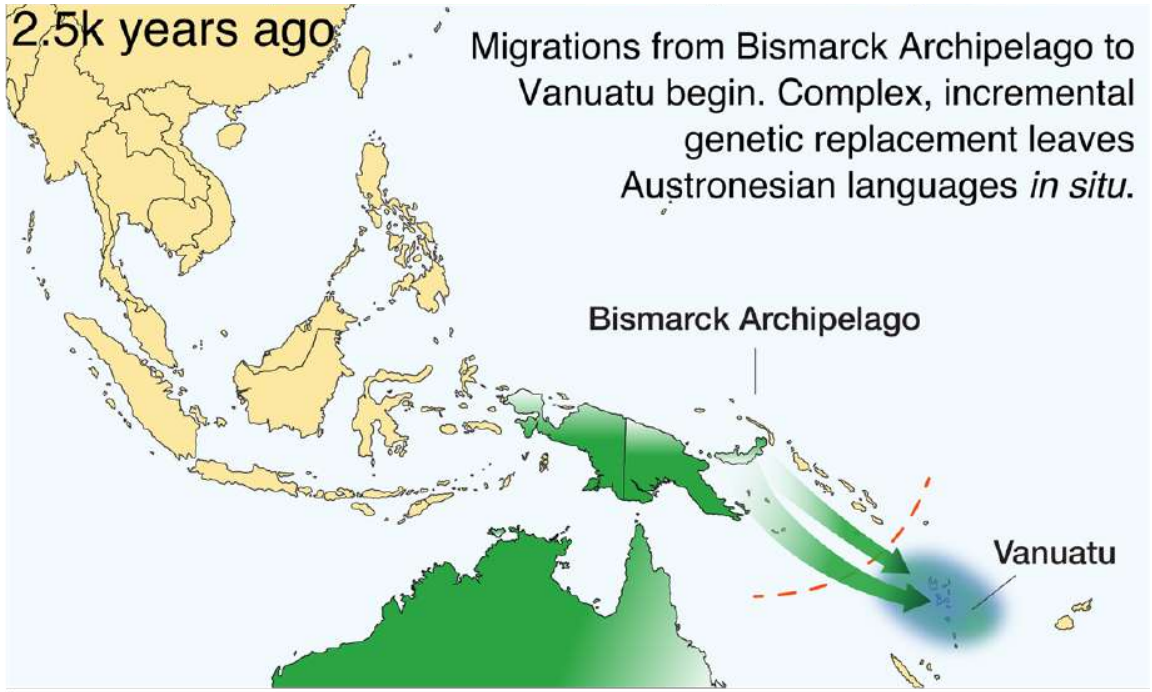
By looking at the whole virus genome in people who have had confirmed cases of COVID-19, scientists can monitor changes in the virus at a national scale to understand how the virus is spreading and whether different strains are emerging. This will help clinical care of patients and save lives.

https://www.sanger.ac.uk/news/view/uk-launches-whole-genome-sequence-alliance-map-spread-coronavirus

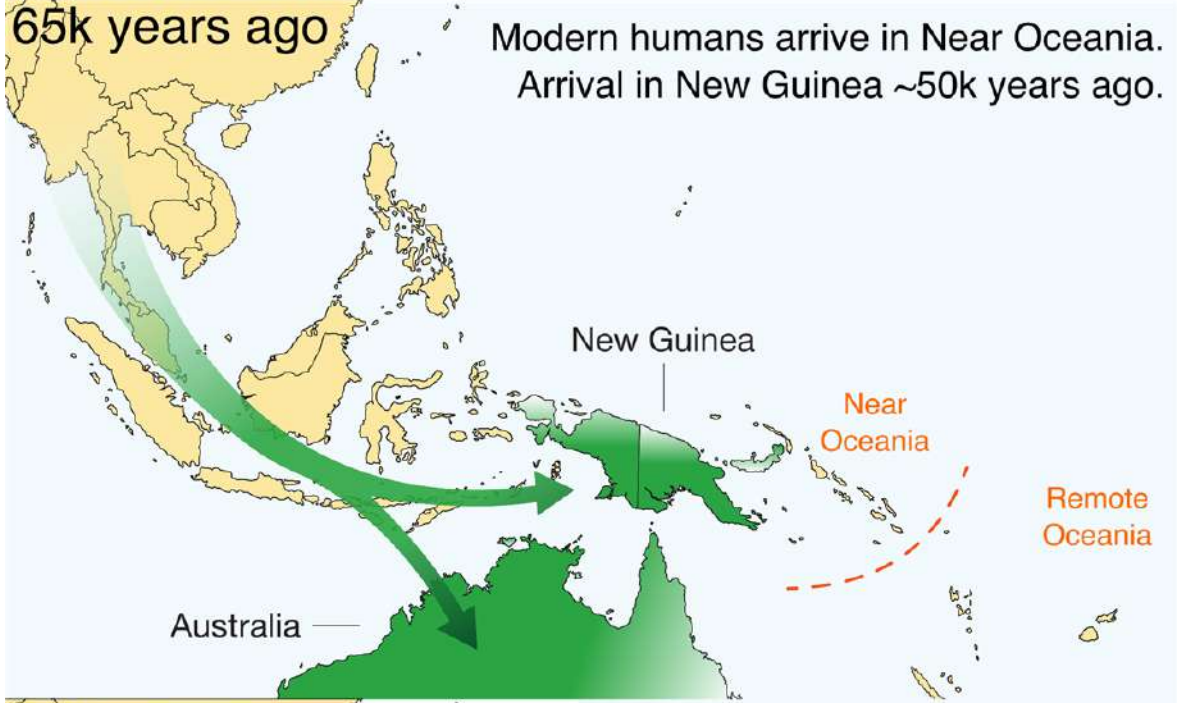
https://www.ncbi.nlm.nih.gov/pubmed/28102248

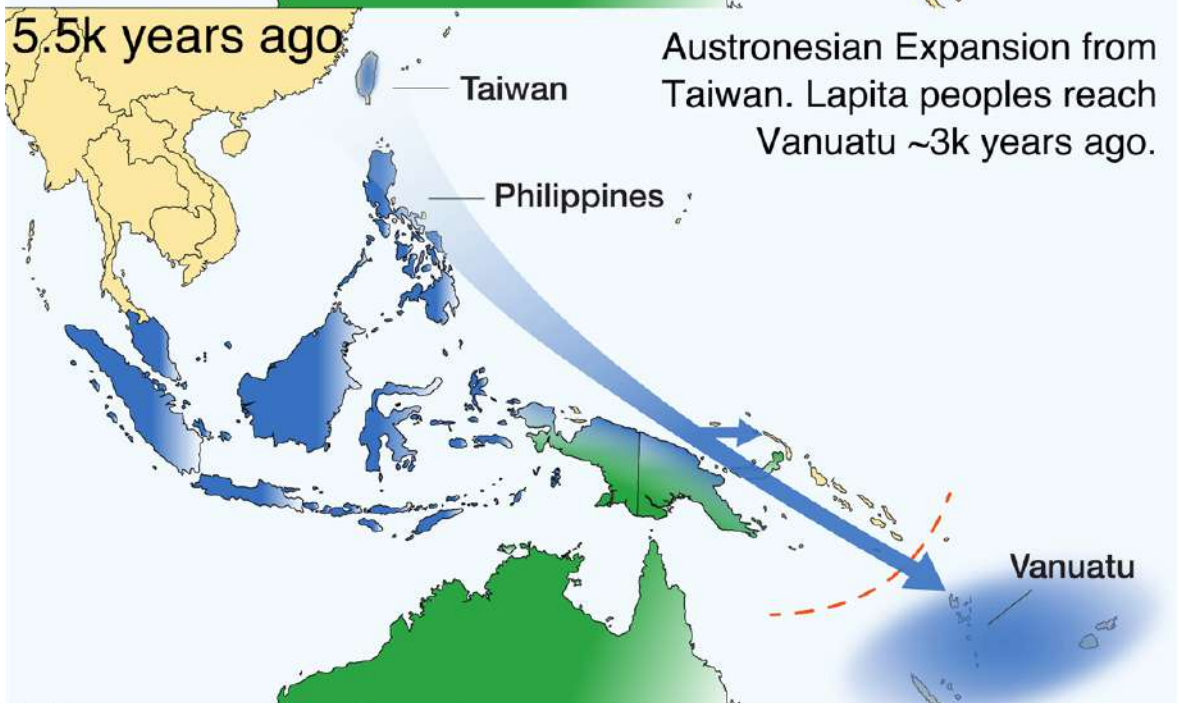
Population genomics

There is a puzzling mismatch in the population history of the Pacific: why do most people across Remote Oceania (the vast area stretching from Vanuatu to Easter Island-Rapa Nui) speak languages of the Austronesian language family that expanded into this region only 3,000 years ago, yet carry a component of genetic ancestry from a much older source population in Near Oceania (the area including New Guinea and its surrounding islands, see the top of Figure 1)?



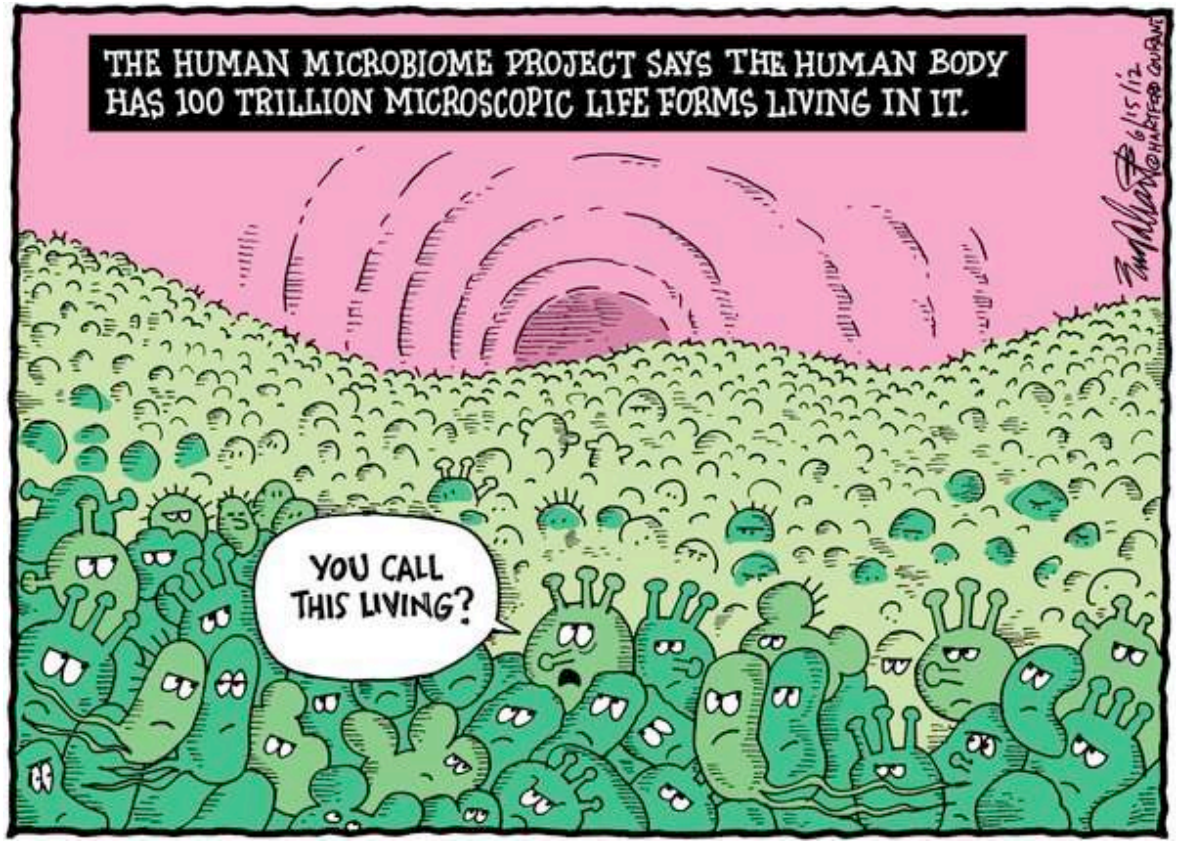
Credit: Hans Sell, MPI-SHH, adapted from Skoglund *et al.* 2016 *Nature*

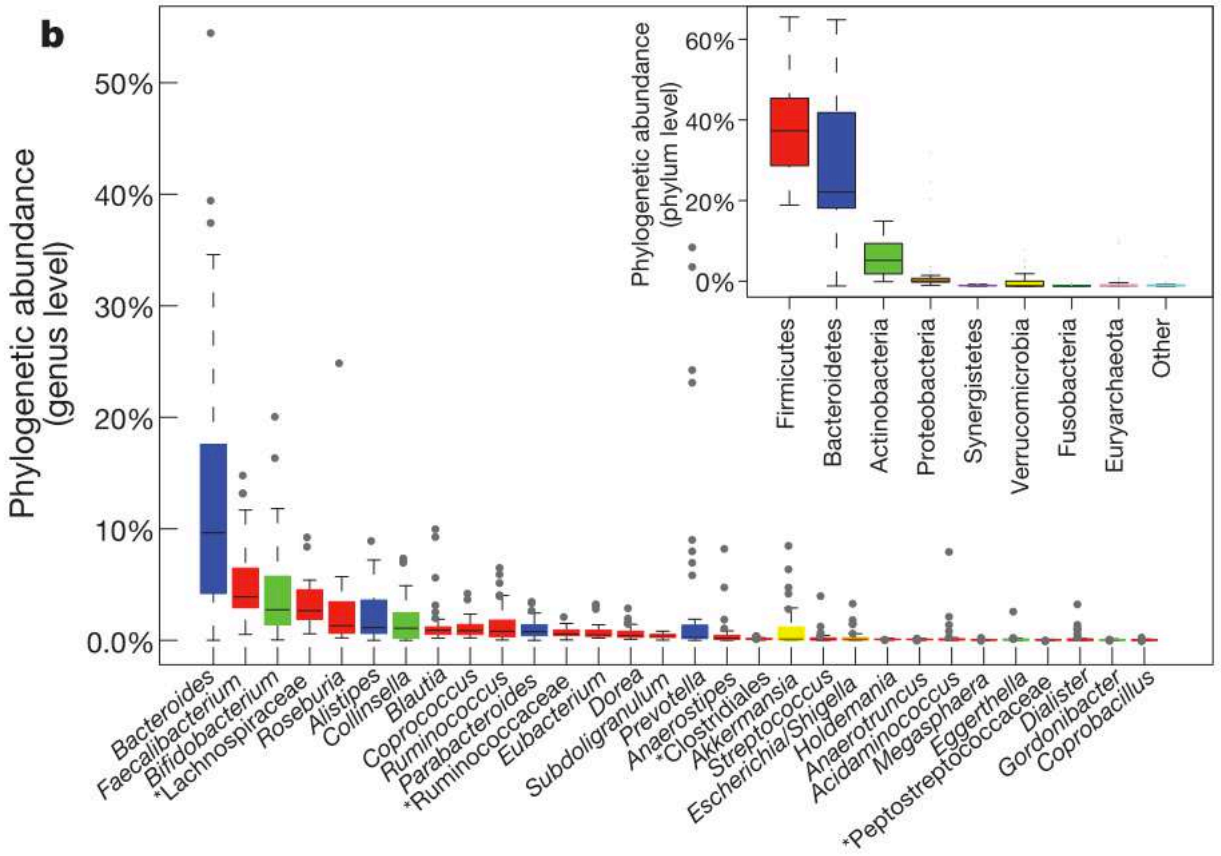




doi:10.1038/s41559-018-0498-2

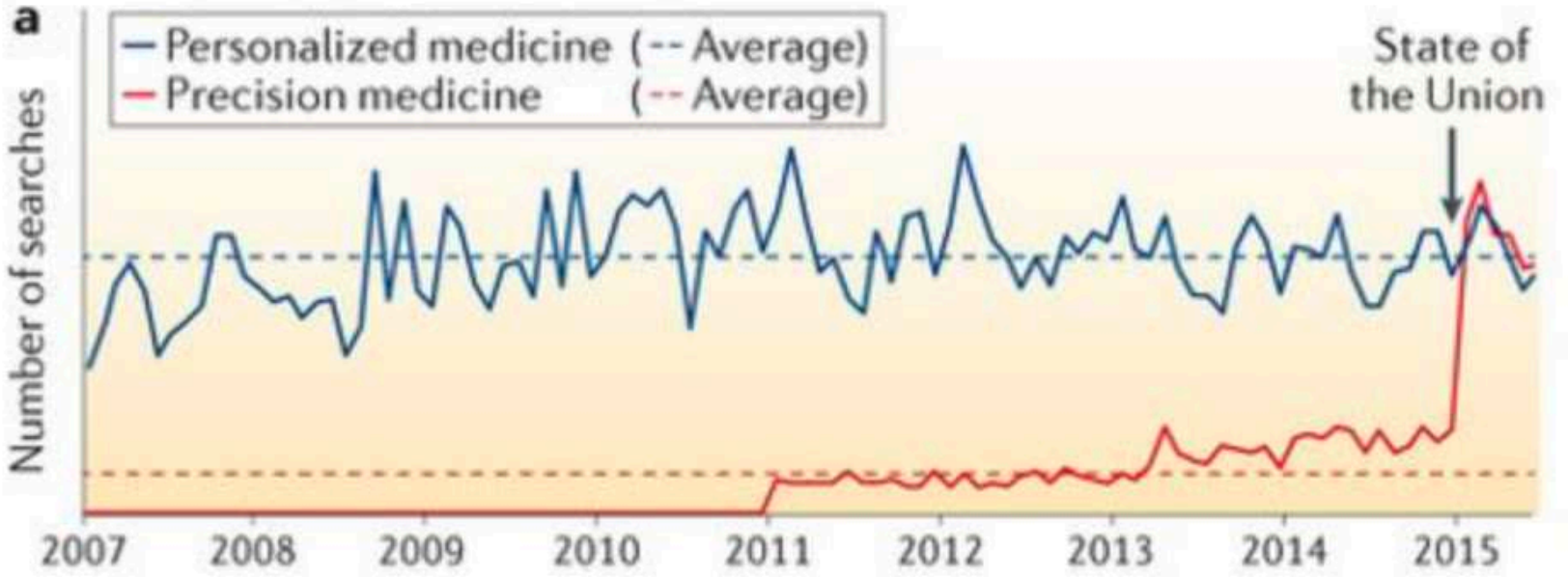
Metagenomics





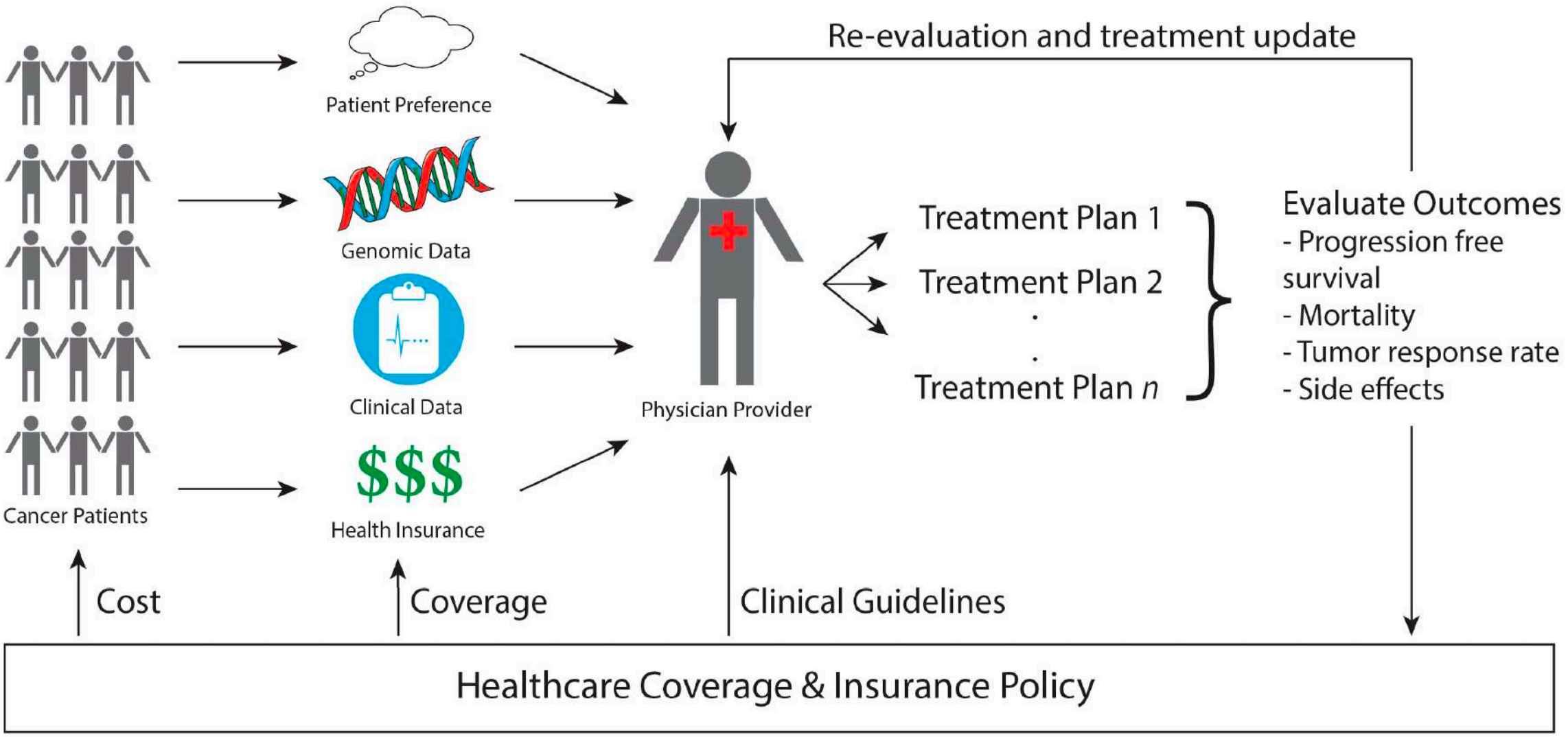
doi:10.1038/nature09944

Precision medicine



Ashley (2016) Nature Review Genetics

Outline of precision medicine



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

PM examples

Table 1 | **Examples of precision medicine**

Condition	Gene	Action
Mendelian disease		
Cystic fibrosis	*CFTR*	Specific therapies such as ivacaftor and a combination of lumacaftor and ivacaftor
Long QT syndrome	*KCNQ1*, *KCNH2* and *SCN5A*	Specific therapy for patients with *SCN5A* mutations
Duchenne muscular dystrophy	*DMD*	Ongoing phase III clinical trials of exon-skipping therapies
Malignant hyperthermia susceptibility	*RYR1*	Avoid volatile anaesthetic agents; avoid extremes of heat
Familial hypercholesterolaemia (FH)	*PCSK9*, *APOB* and *LDLR*	• Heterozygous FH (HeFH): eligible for PCSK9 inhibitor drugs • Homozygous FH (HoFH): eligible for PCSK9 inhibitor drugs in addition to lomitapide and mipomersen
Dopa-responsive dystonia	*SPR*	Therapy with dopamine precursor L-dopa and the serotonin precursor 5-hydroxytryptophan
Thoracic aortic aneurysm	*SMAD3*, *ACTA2*, *TGFBR1*, *TGFBR2* and *FBN1*	Customization of surgical thresholds based on patient genotype
Left ventricular hypertrophy	*MYH7*, *MYBPC3*, *GLA* and *TTR*	Sarcomeric cardiomyopathy, Fabry disease and transthyretin cardiac amyloid disease have specific therapies
Precision oncology		
Lung adenocarcinoma	*EGFR* and *ALK*	Targeted kinase inhibitors, such as gefitinib and crizotinib
Breast cancer	*HER2*	HER2 (also known as ERBB2)-targeted treatment, such as trastuzumab and pertuzumab
Gastrointestinal stromal tumour	*KIT*	Targeted KIT kinase activity inhibitors, such as imatinib
Melanoma	*BRAF*	BRAF inhibitors, such as vemurafenib and dabrafenib
Pharmacogenomics		
Warfarin sensitivity	*CYP2C9* and *VKORC1*	Adjust dosage of warfarin or consider alternative anticoagulant
Clopidogrel sensitivity, post-stent procedure	*CYP2C19*	Consider alternative antiplatelet therapy (for example, prasugrel or ticagrelor)
Thiopurine sensitivity	*TPMT*	Reduce thiopurine dosage or consider alternative agent
Codeine sensitivity	*CYP2D6*	Avoid use of codeine; consider alternatives such as morphine and non-opioid analgesics
Simvastatin sensitivity	*SLCO1B1*	Reduce dose of simvastatin or consider an alternative statin; consider routine creatine kinase surveillance

Ashley (2016) Nature Review Genetics

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.

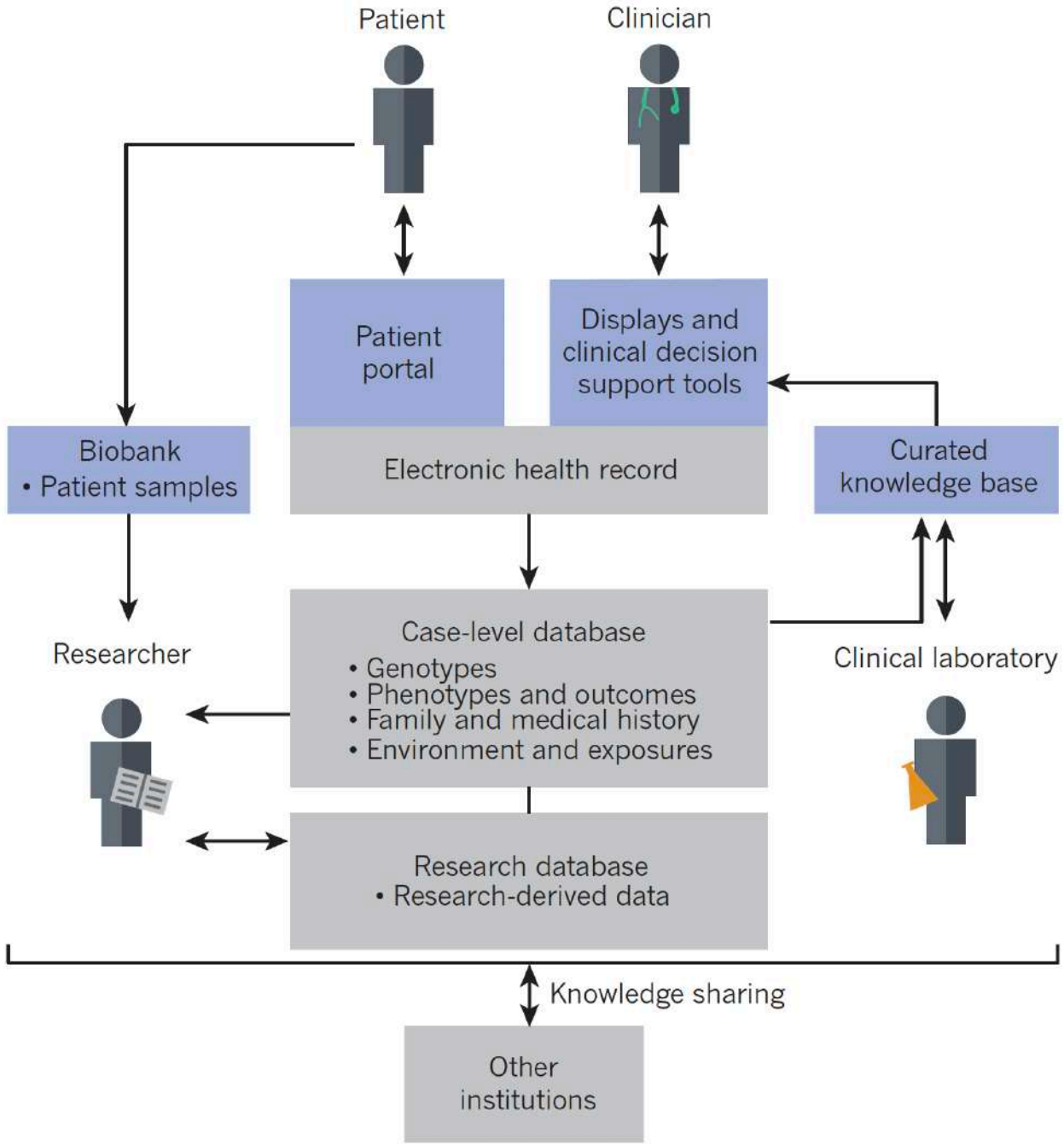
Morash *et al* (2018) Journal of Personalized Medicine

REVIEW

doi

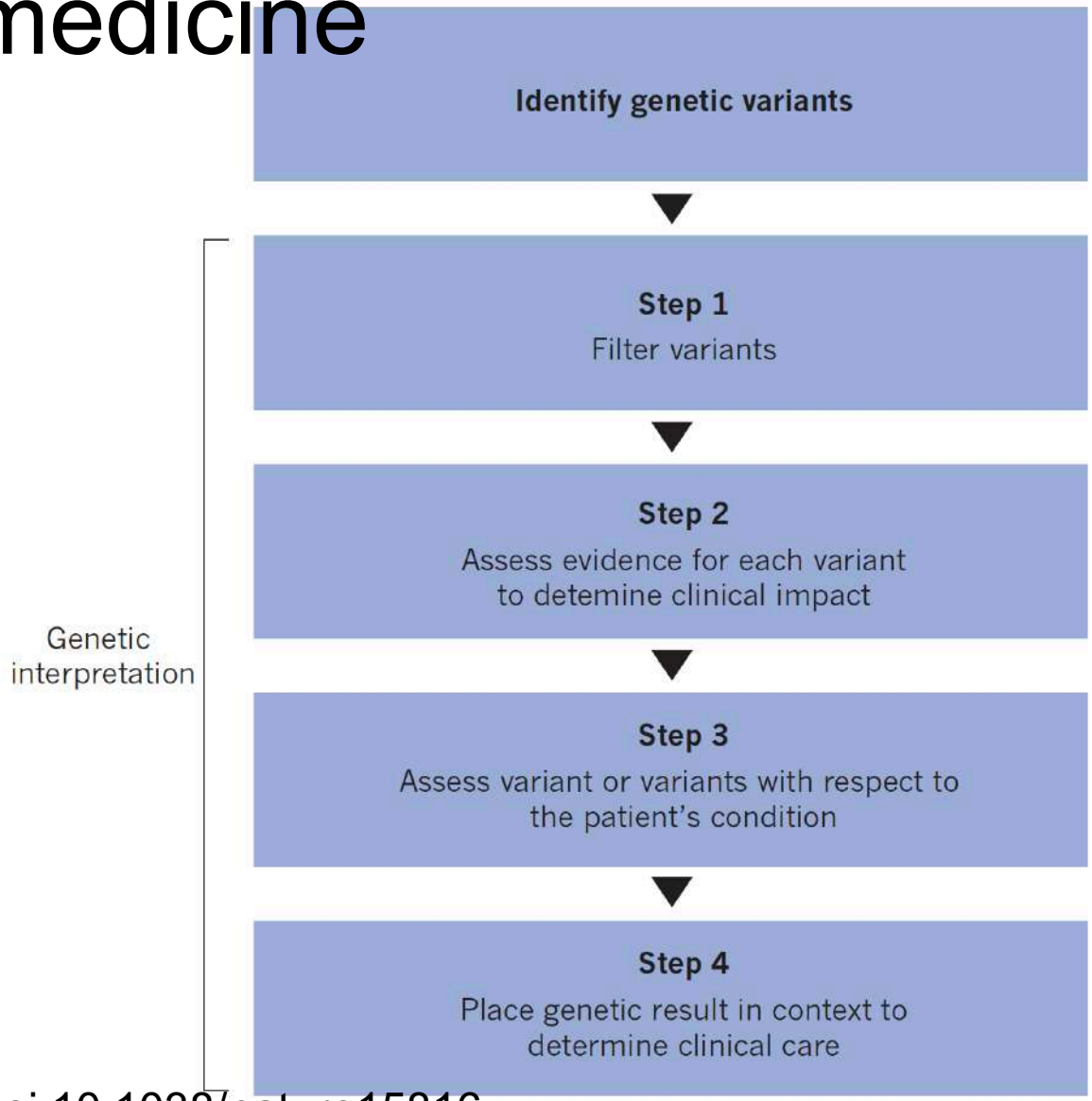
Building the foundation for genomics in precision medicine

Samuel J. Aronson[1,2] & Heidi L. Rehm[1,3,4,5]



doi:10.1038/nature15816

Building the foundation for genomics in precision medicine



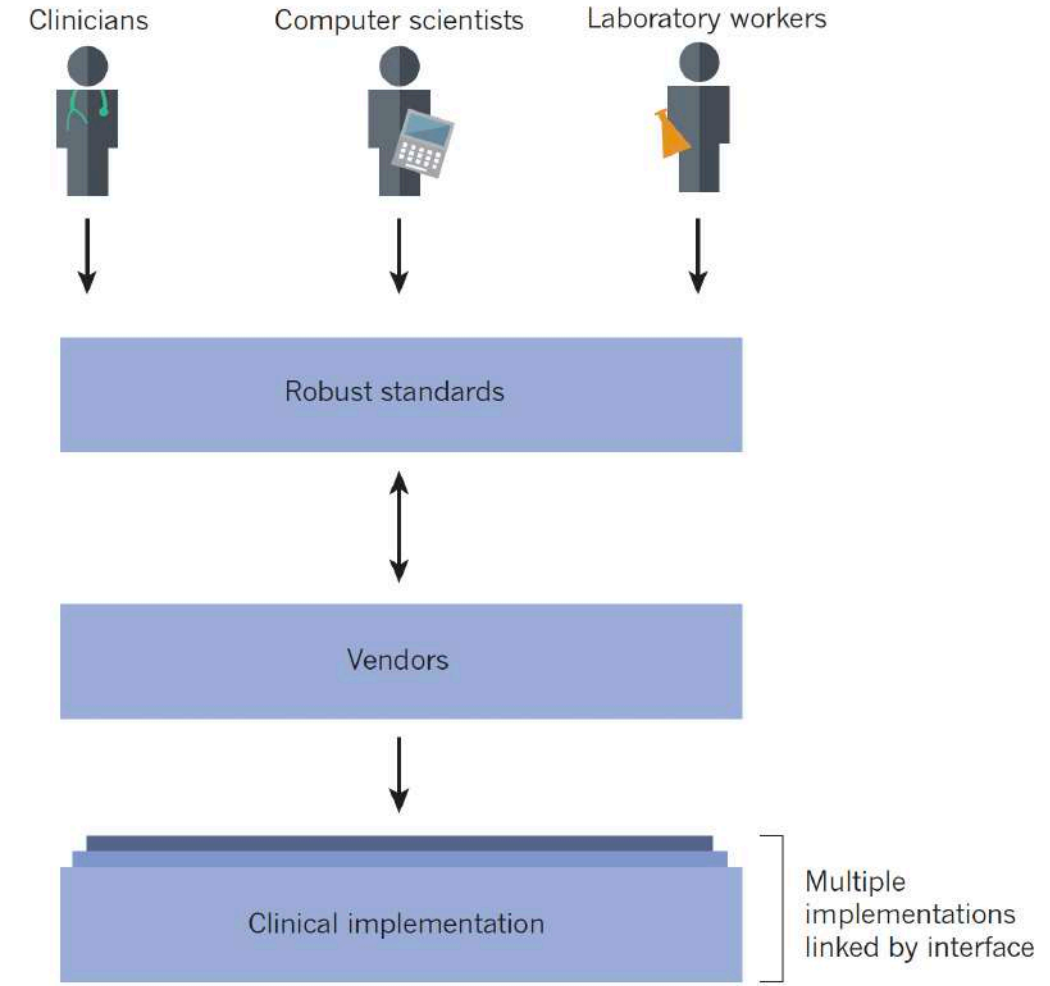
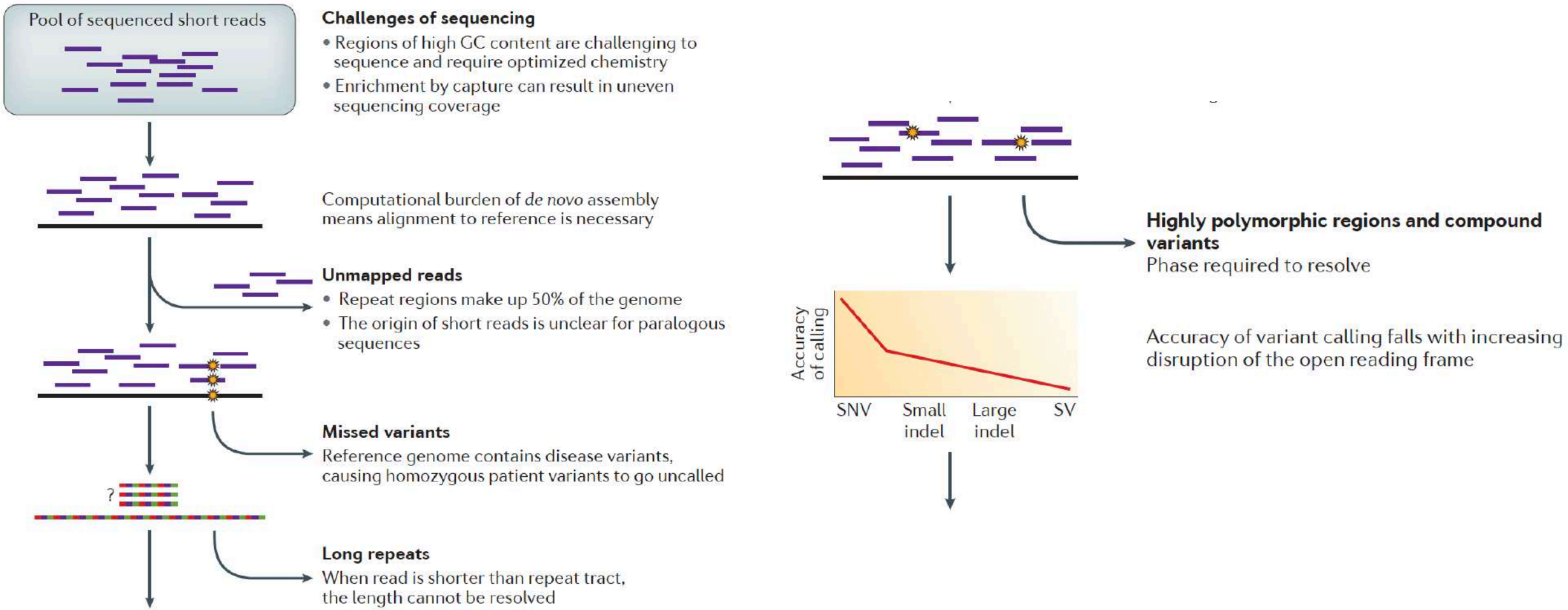


Figure 3 | Creating and implementing robust standards for the description and structuring of data in laboratory processing and patient-care systems.

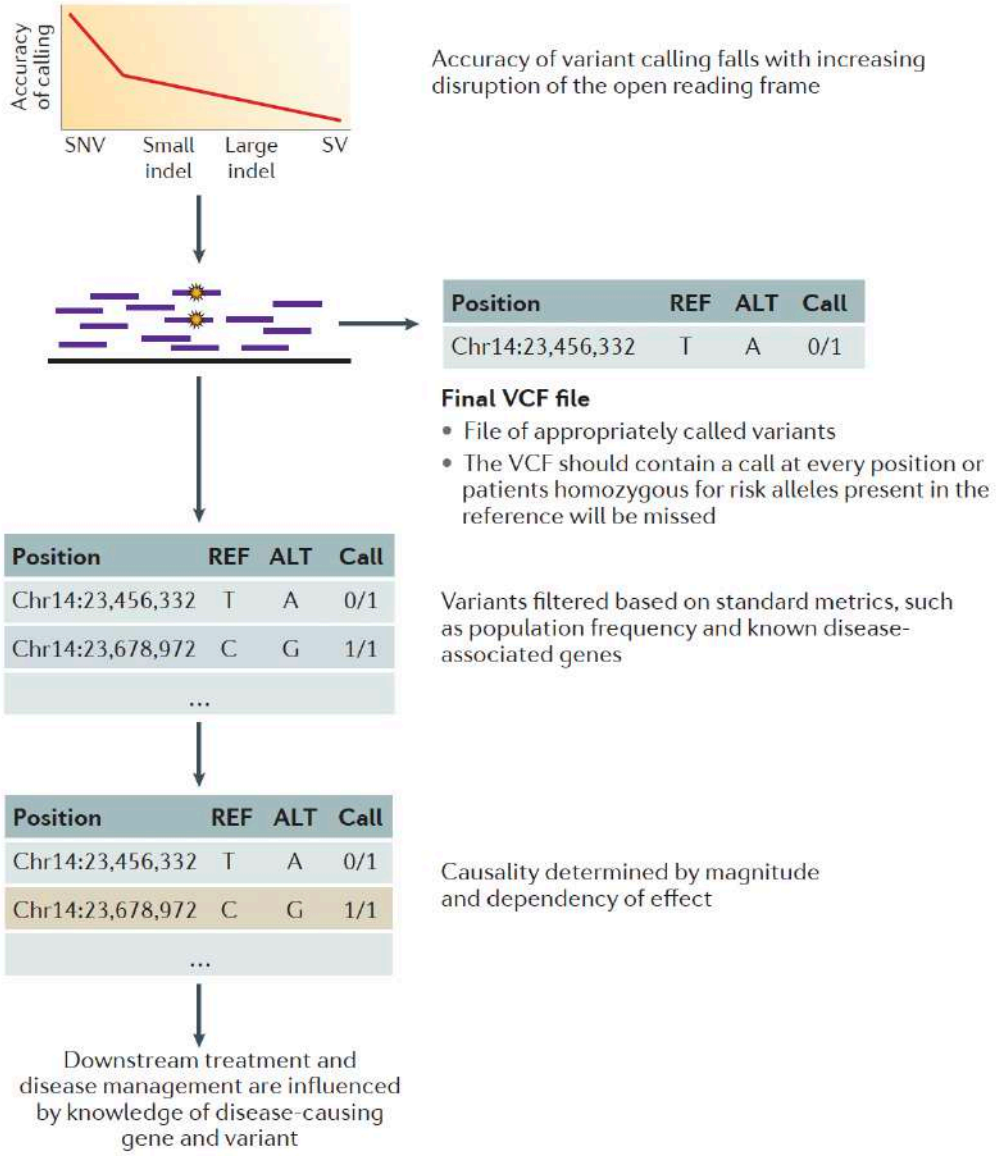
doi:10.1038/nature15816

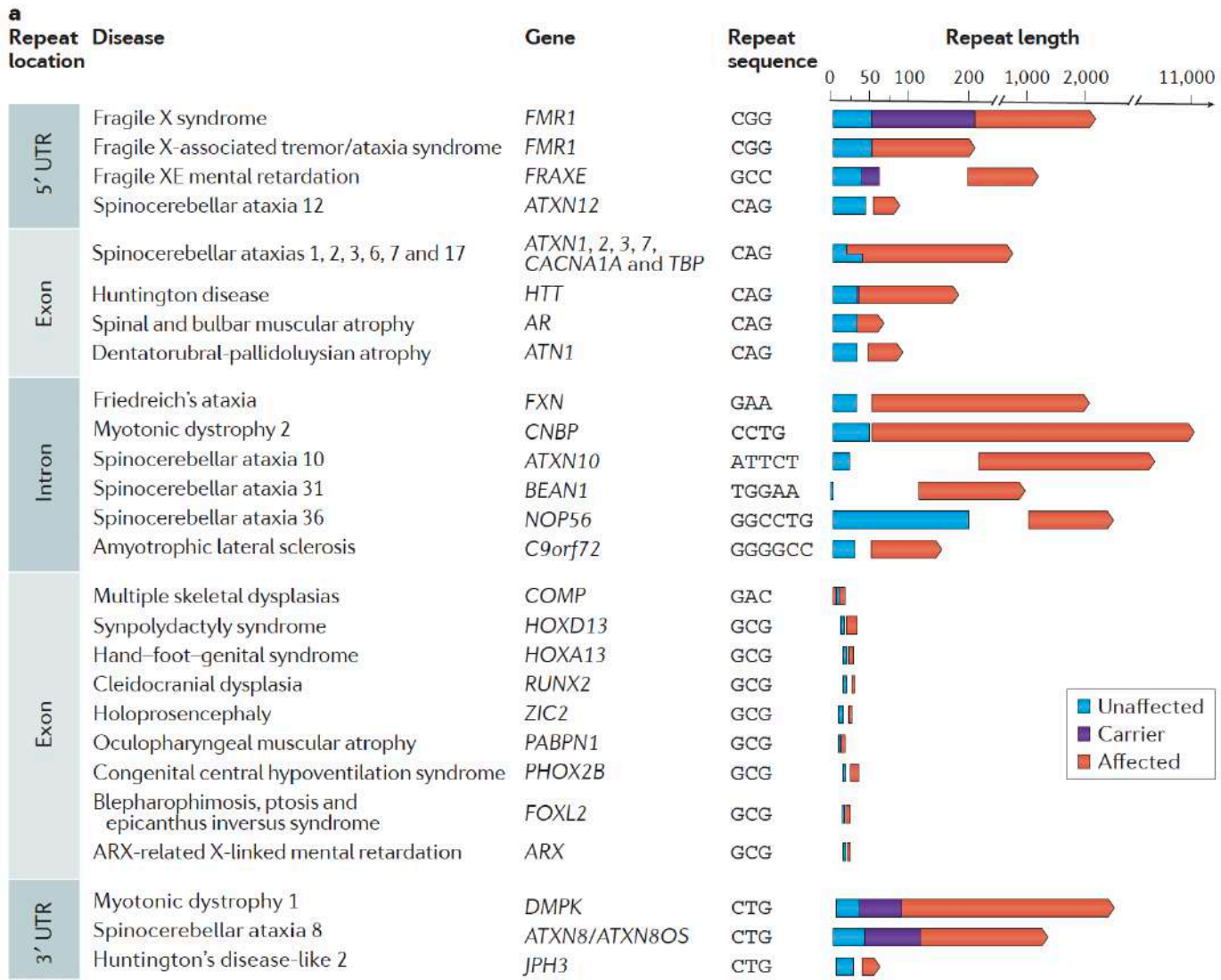
Challenges and reduced accuracies



Ashley (2016) Nature Review Genetics

Challenges and reduced accuracies

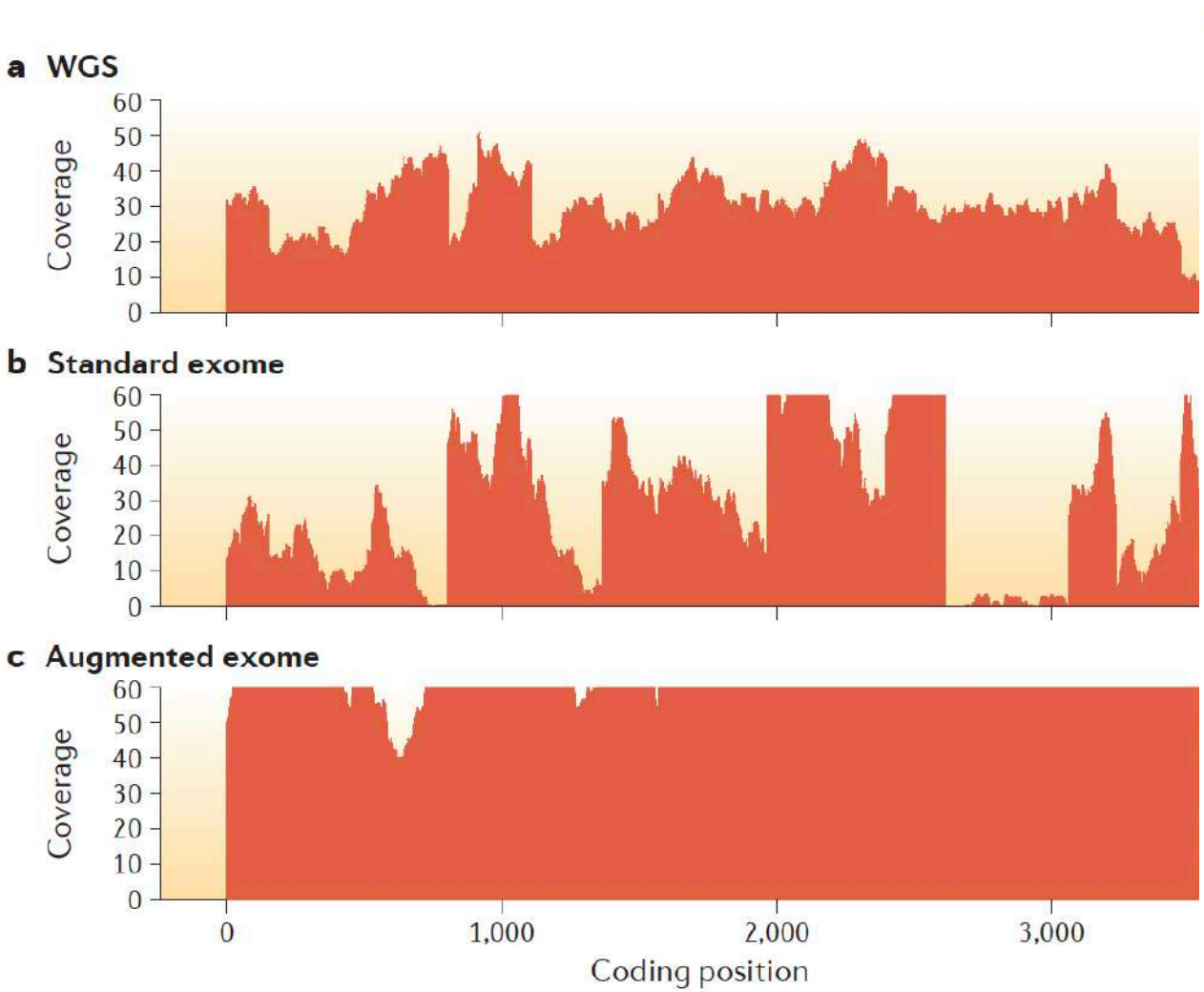


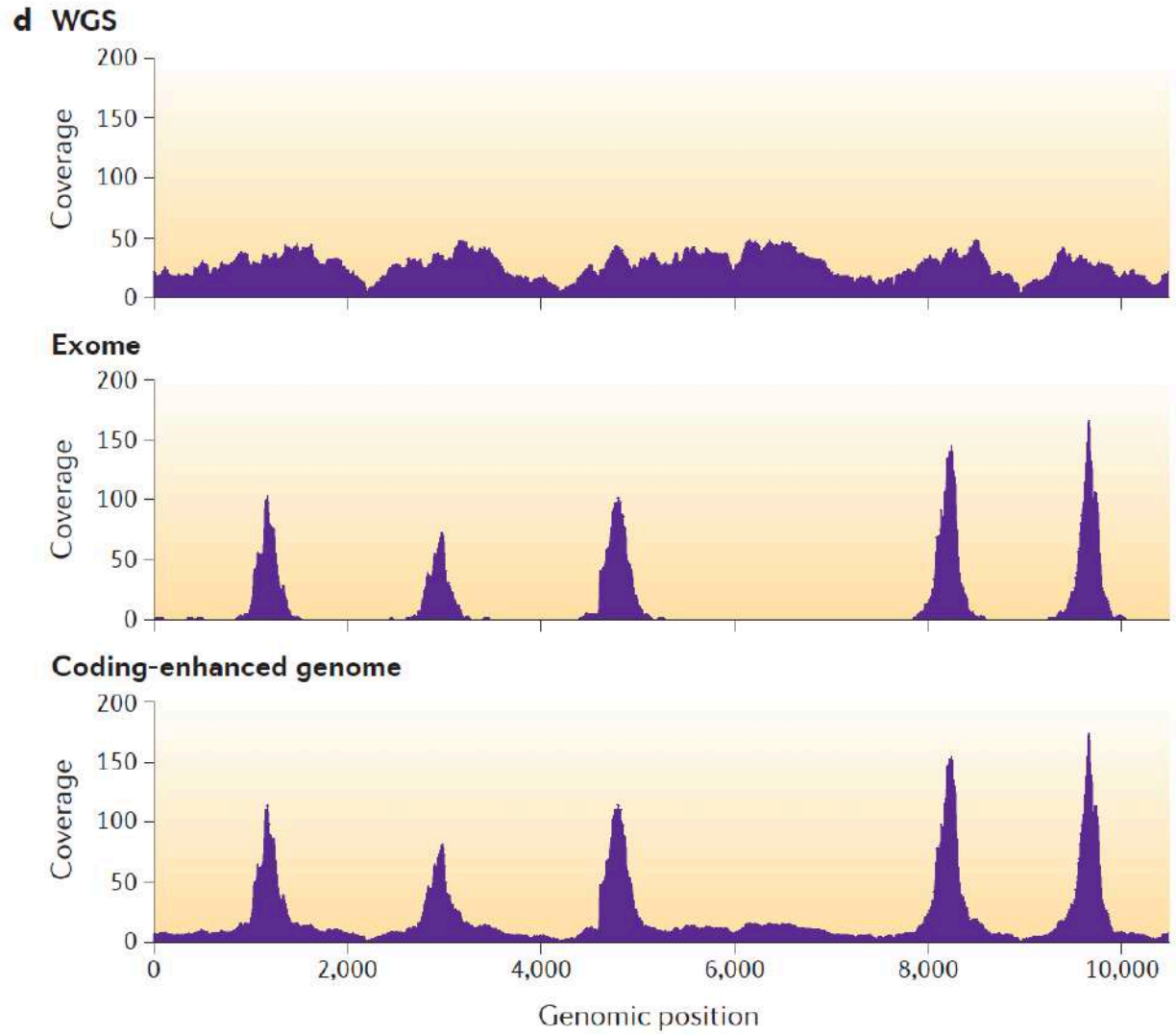


Ashley (2016) Nature Review Genetics

Resolution

Zoomed out





Ashley (2016) Nature Review Genetics

Transcriptomics / RNAseq

Applications of RNAseq

Discovery / Annotation

- Find new genes
- Find new transcripts
- Find new ncRNAs, xxx, xxx ….
- Gene fusion

Comparison / Quantification : **given X conditions, find the effect of Y on**

- **expression**
- **Isoform abundance, splice patterns, transcript boundaries**

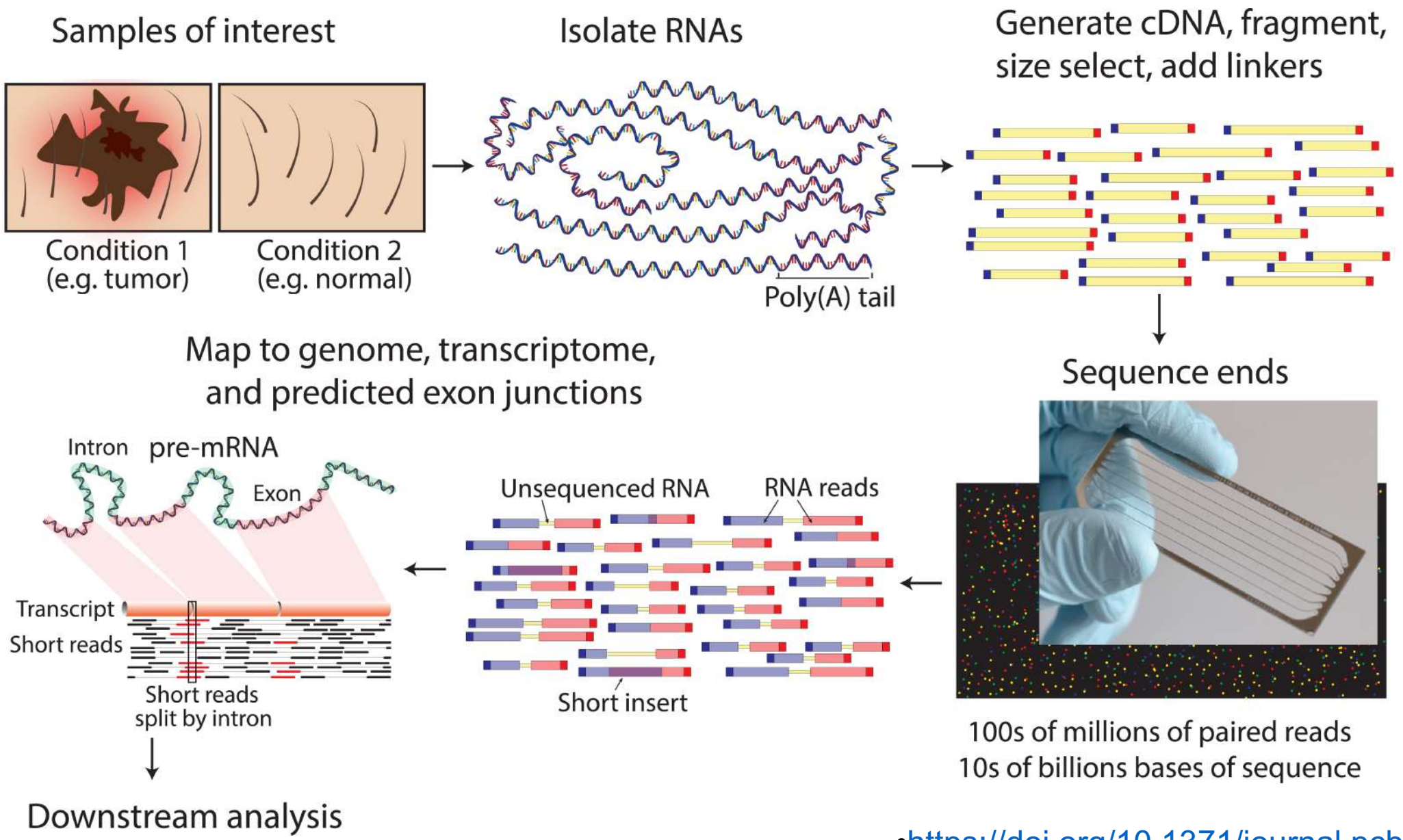
Introduction to differential gene expression analysis using RNA-seq

Written by Friederike Dündar, Luce Skrabanek, Paul Zumbo

September 2015
updated March 20, 2018

http://chagall.med.cornell.edu/RNASEQcourse/Intro2RNAseq.pdf

RNA-seq data generation



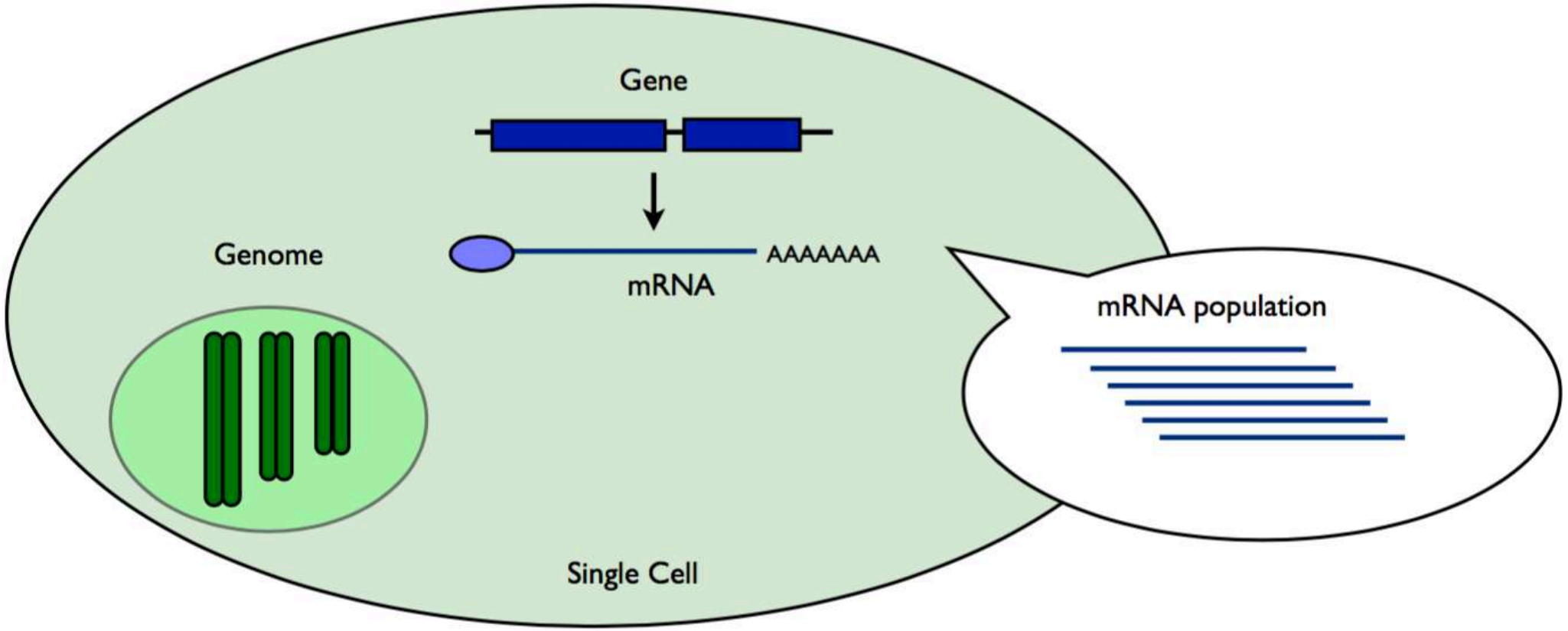
•https://doi.org/10.1371/journal.pcbi.1004393

Types of experiments

Transcriptome Complexity:

Simple System:

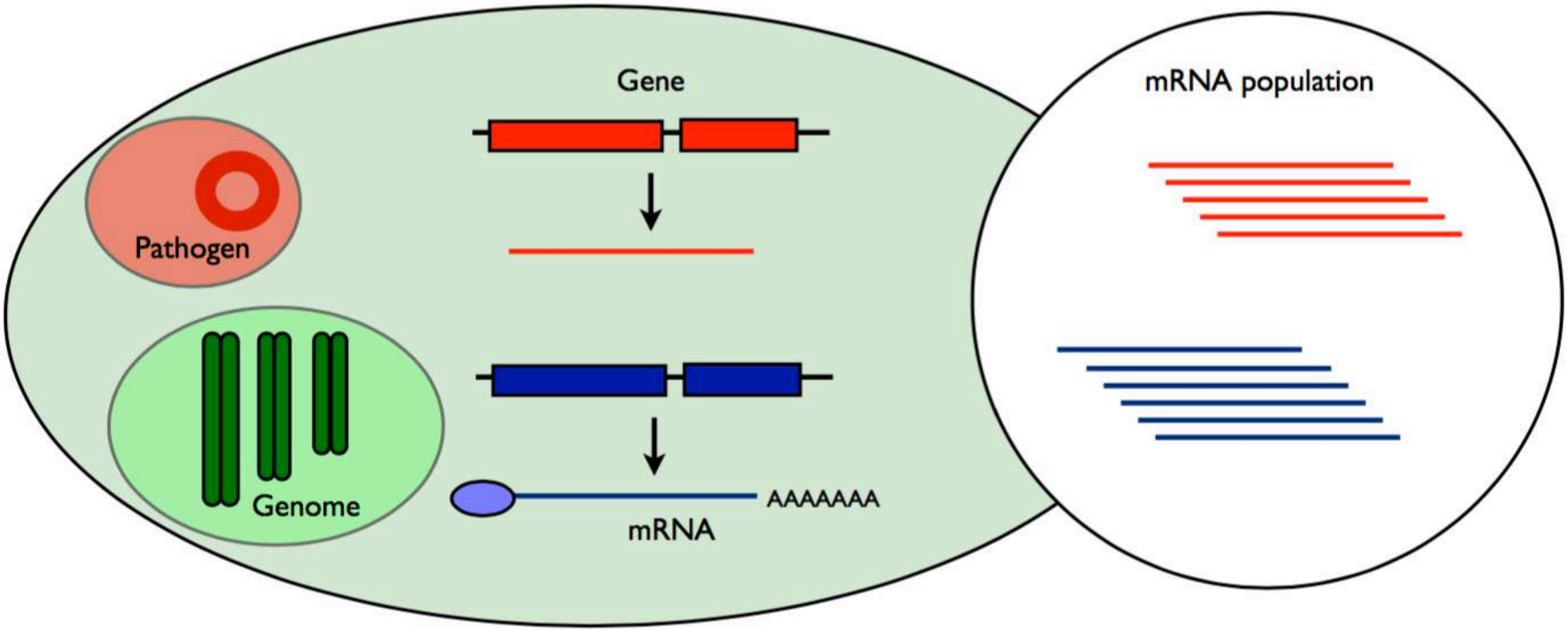
One Genome => Gene 1 copy => Single mRNA



How many species we are analyzing ?

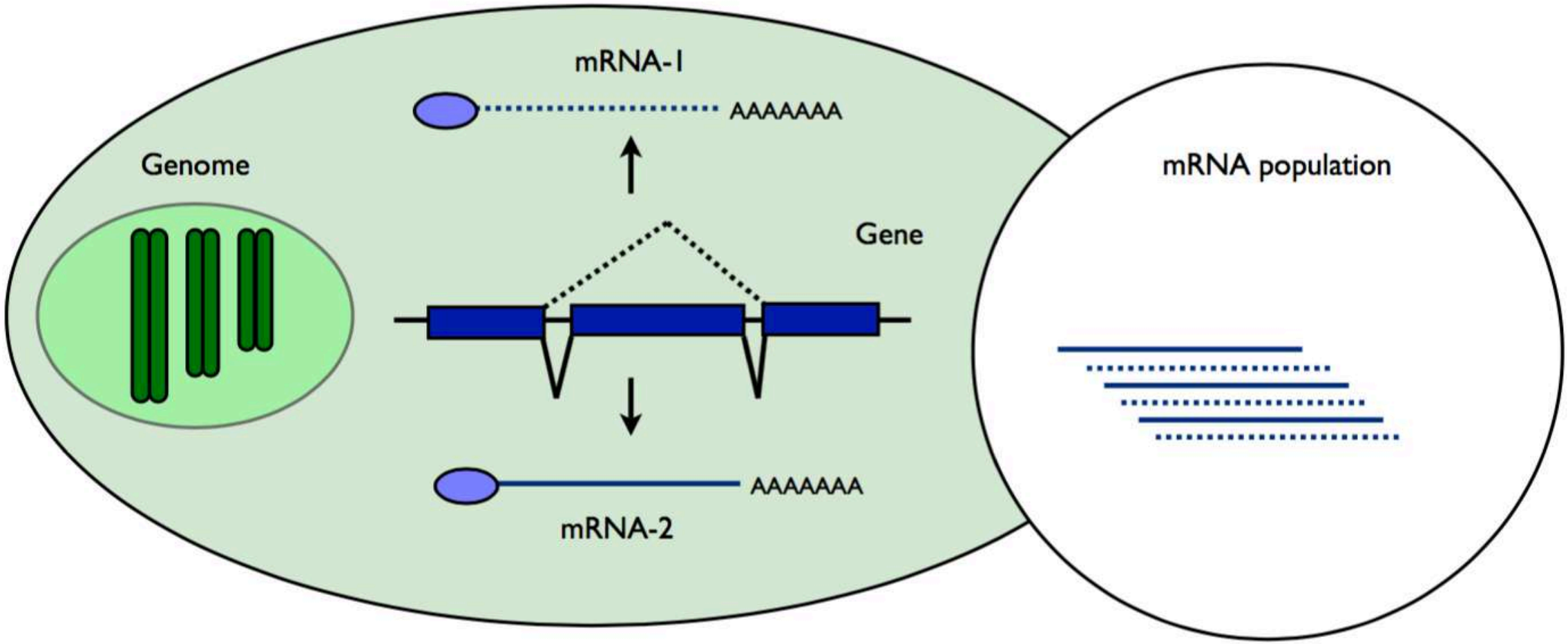
1) Problems to isolate a single species (rhizosphere)

2) Species interaction study (plant-pathogen)



How many isoforms we expect for each allele ?

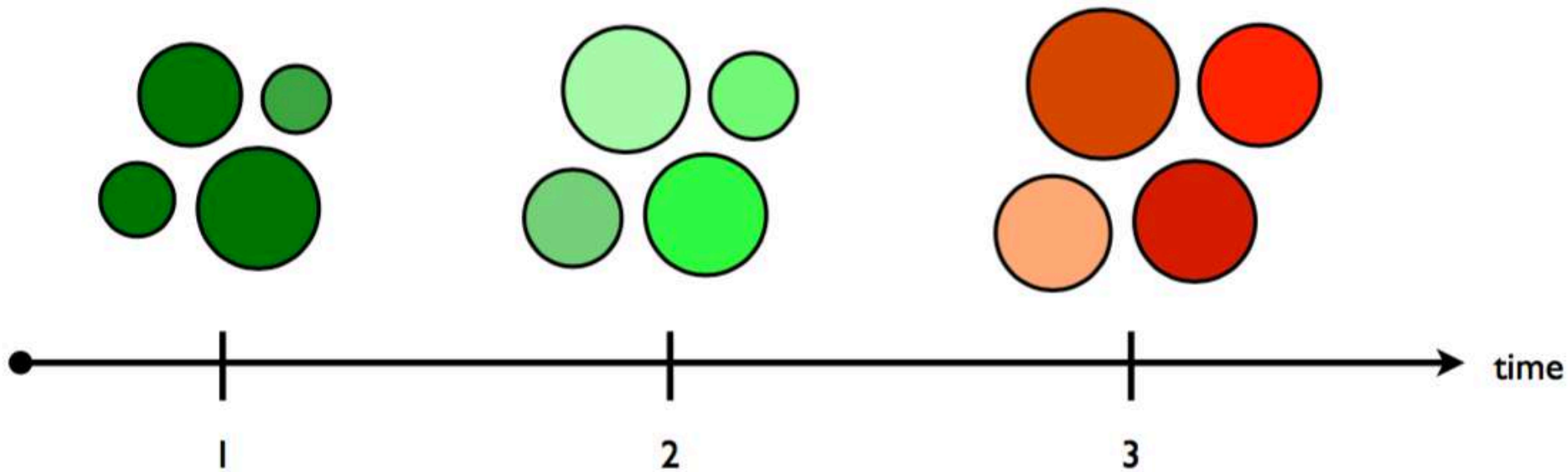
1) Alternative splicings



Is the study performed at different time points?

1) Developmental stages (difficult to select the same)

2) Response to a treatment

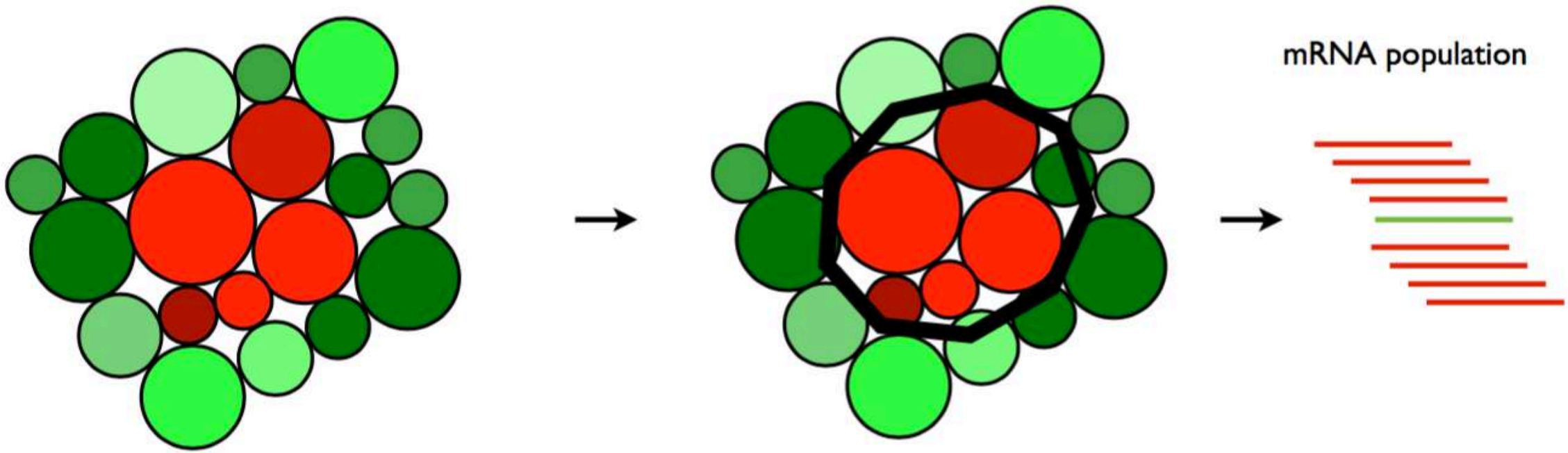


Is the study performed with different parts?

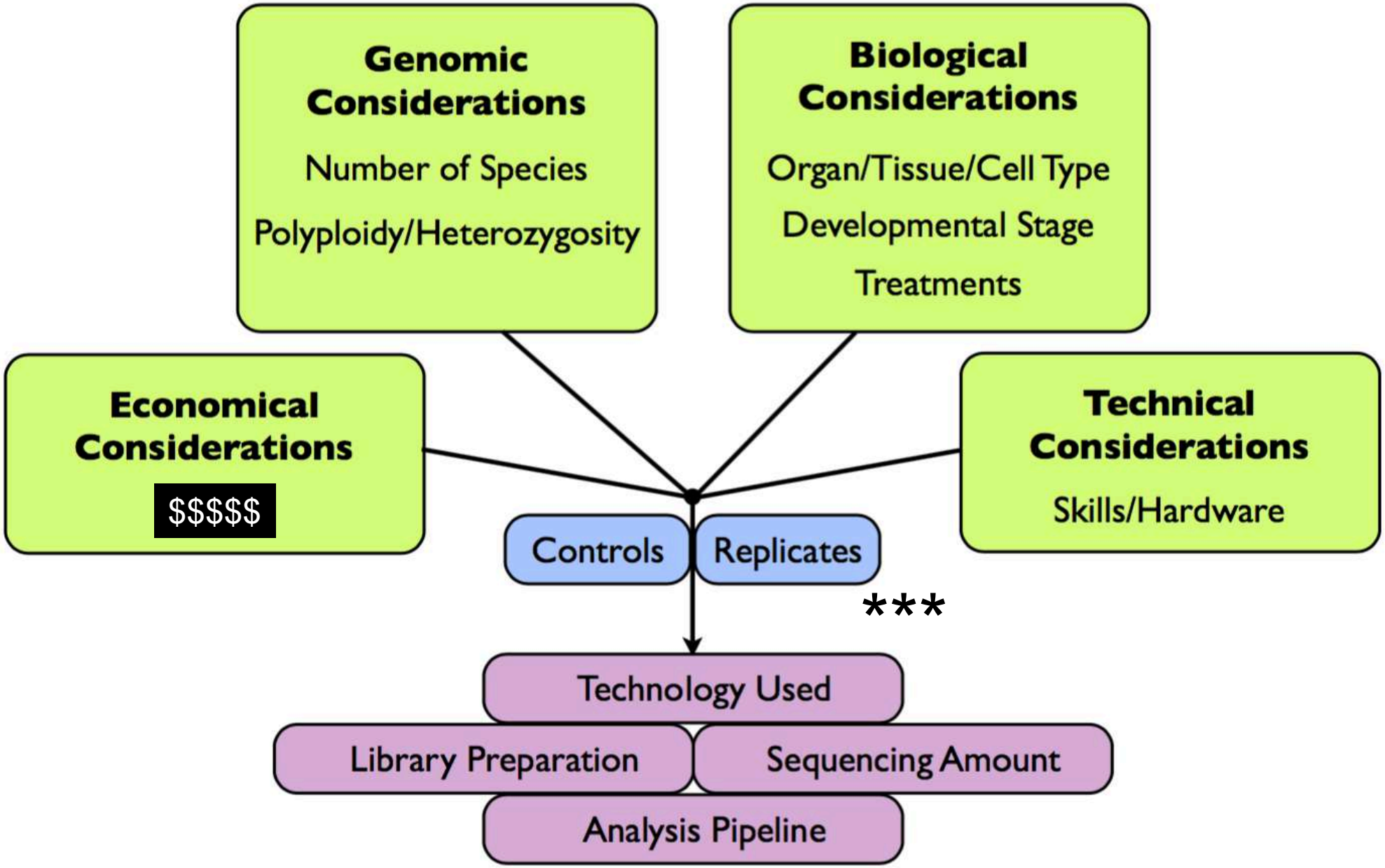
1) Organ specific

2) Tissue/Cell type specific

(Laser Capture Microdissection, LCM)



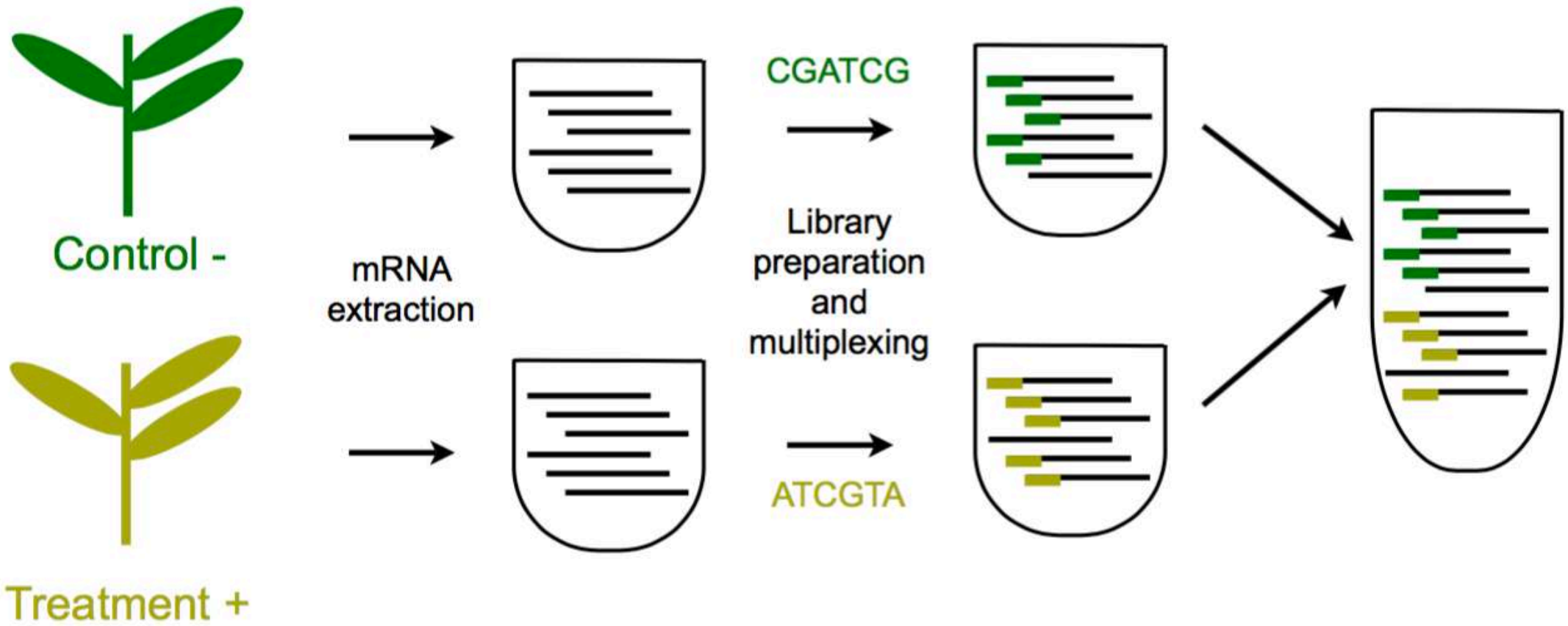
Experimental design



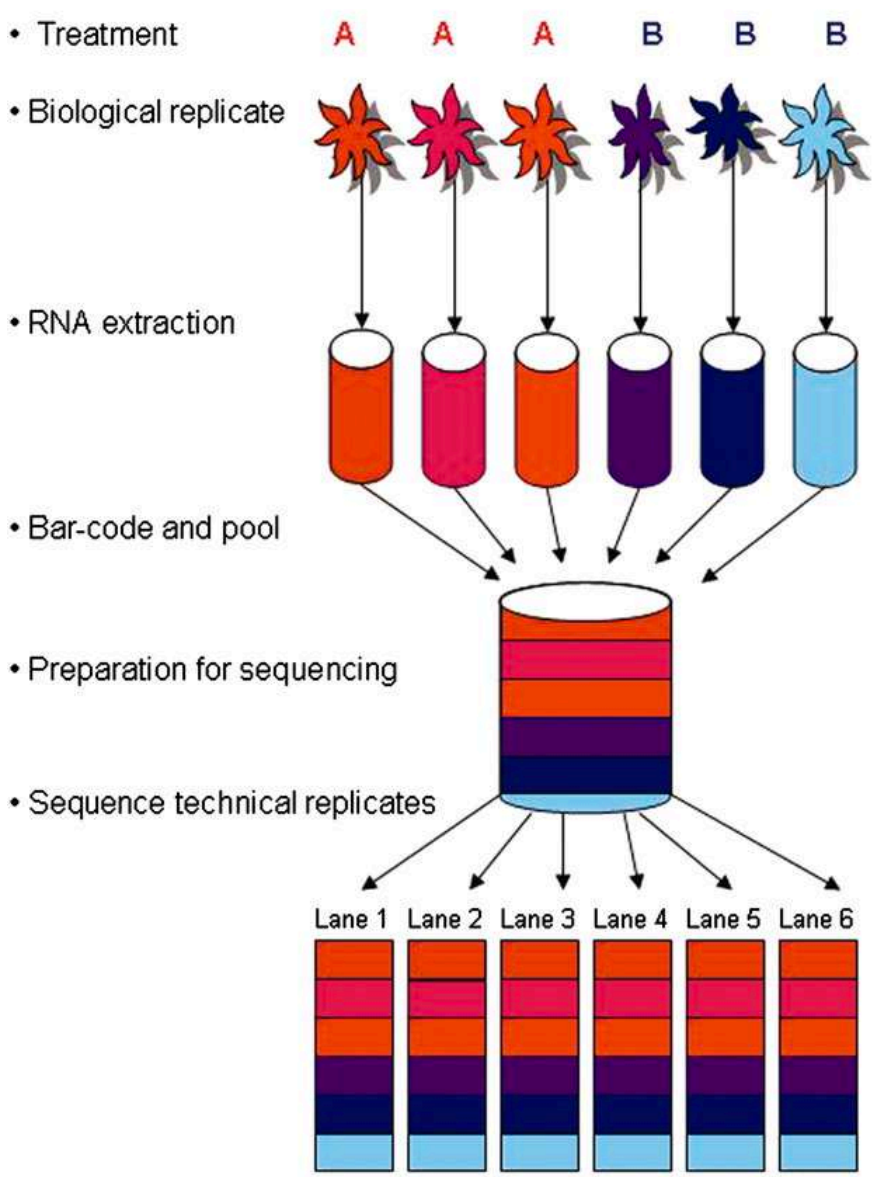
Prep and treatment

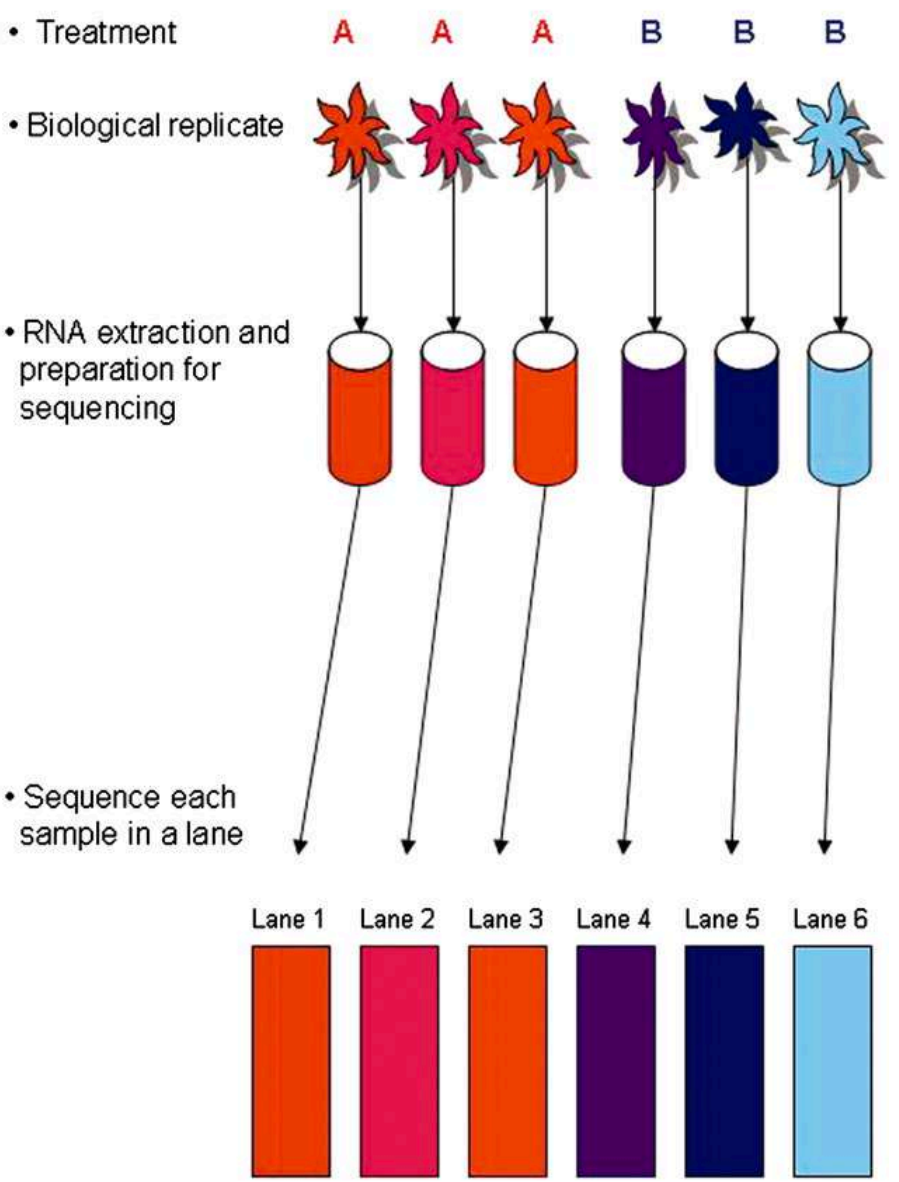
Sequencing of multiple samples can be performed using **multiplexing.**

The multiplexing add a tag/**barcode** of 4-6 nucleotides during the library preparation to identify the sample. Common kits can add up to 96 different tags.



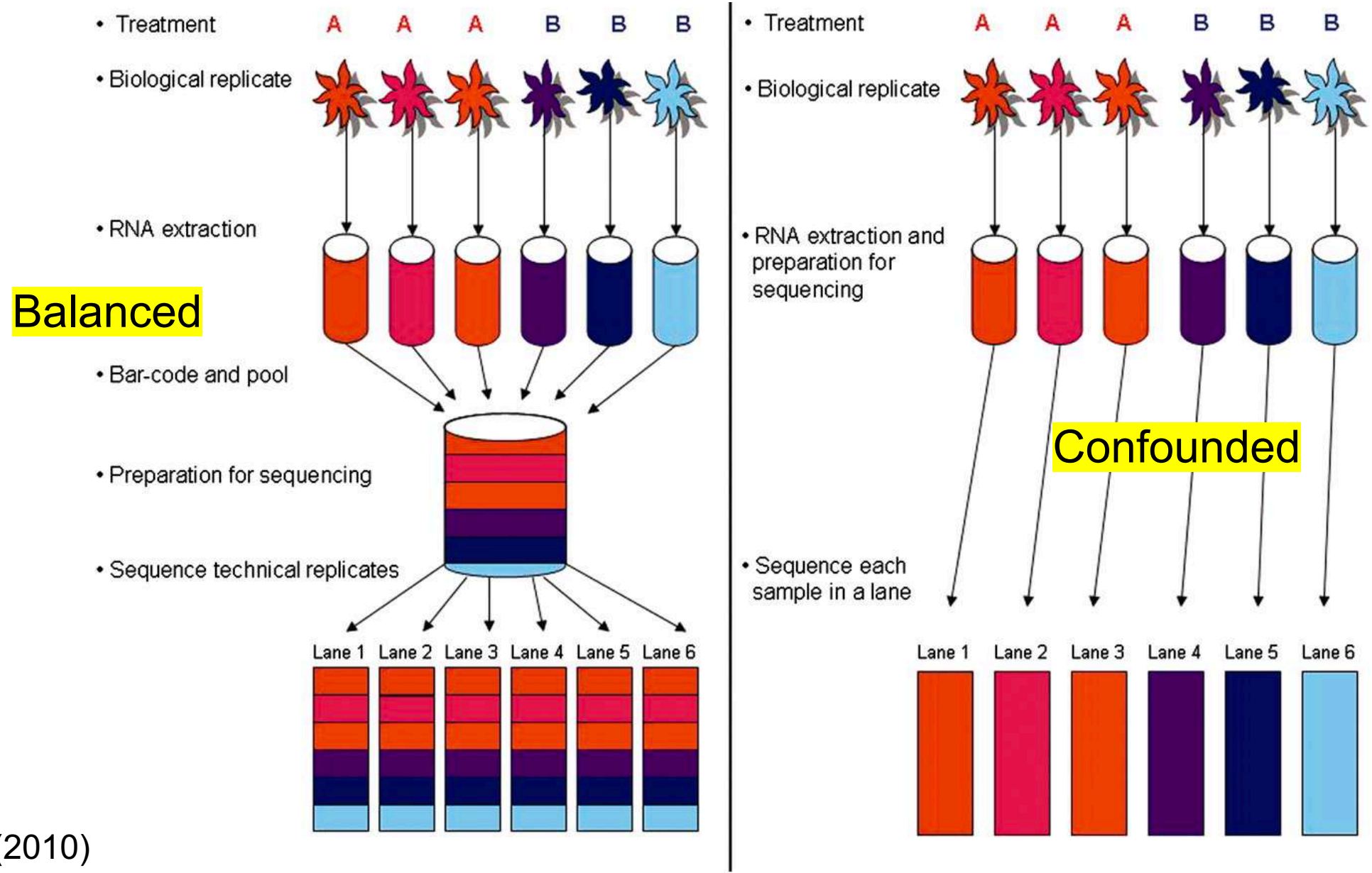
Which of the following designs is correct?





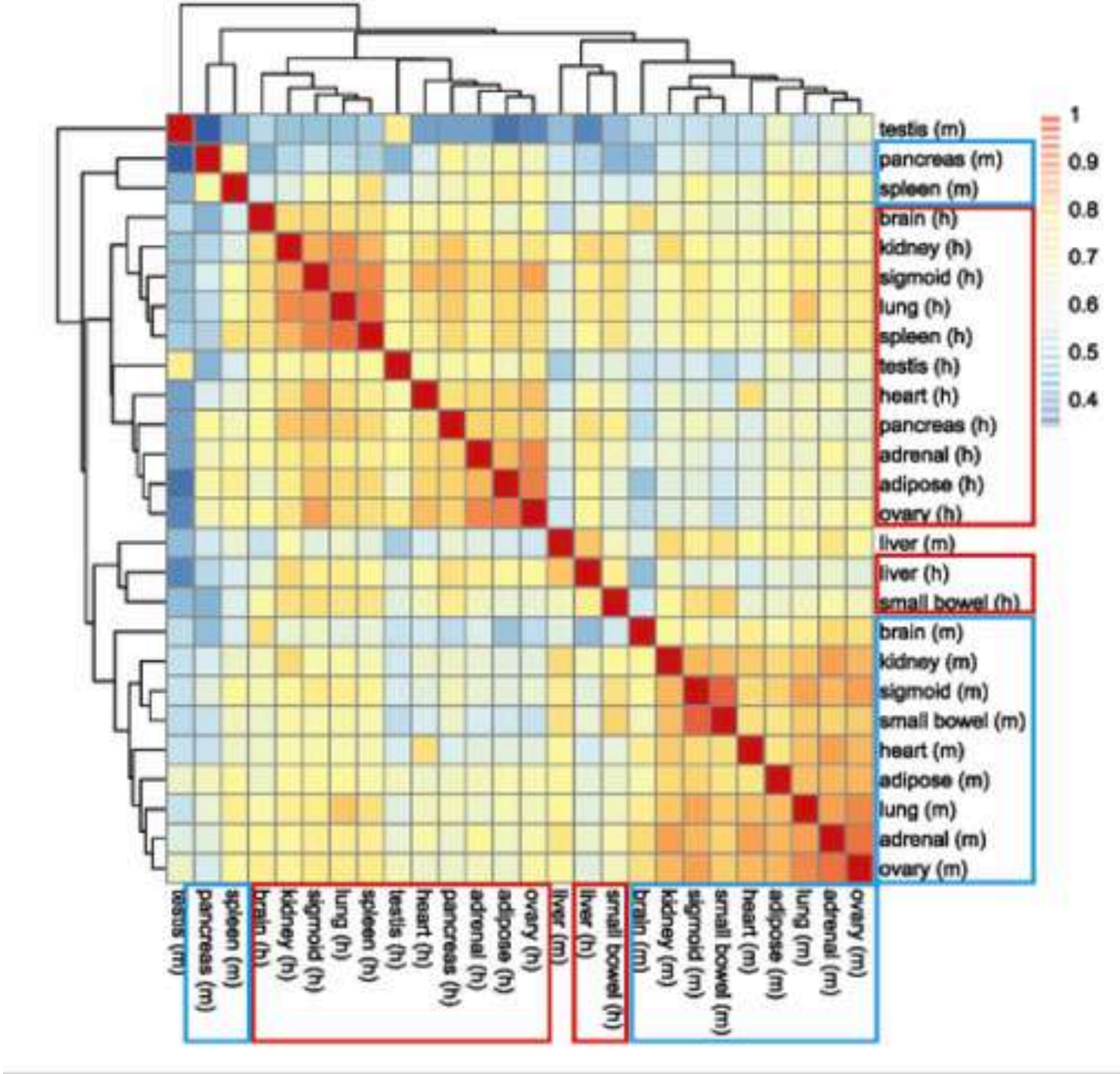
Auer *et al.,* Genetics (2010)

Which of the following designs is correct?



Auer *et al.,* Genetics (2010)

Example of batch effect:



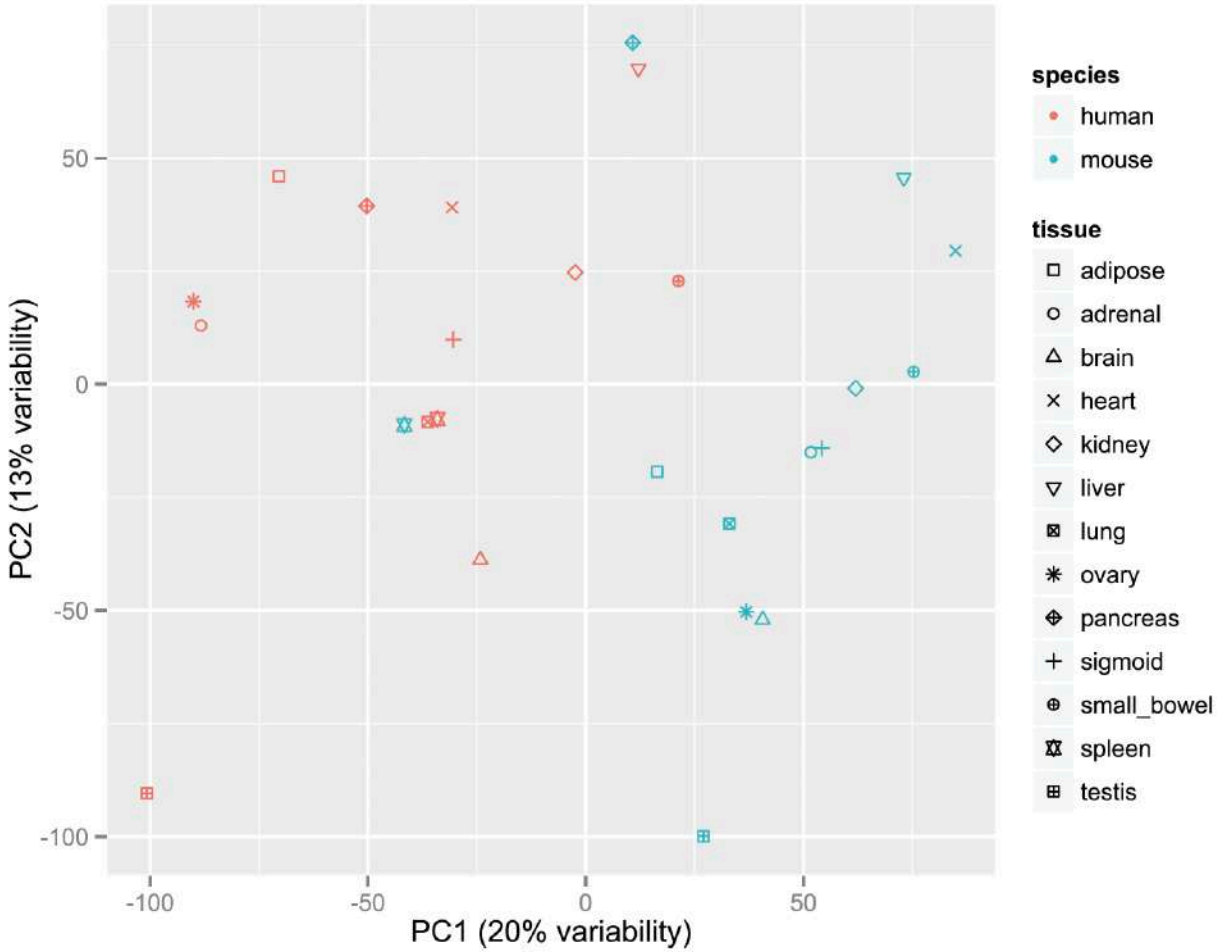
Lin *et al.,* (PNAS) 2014

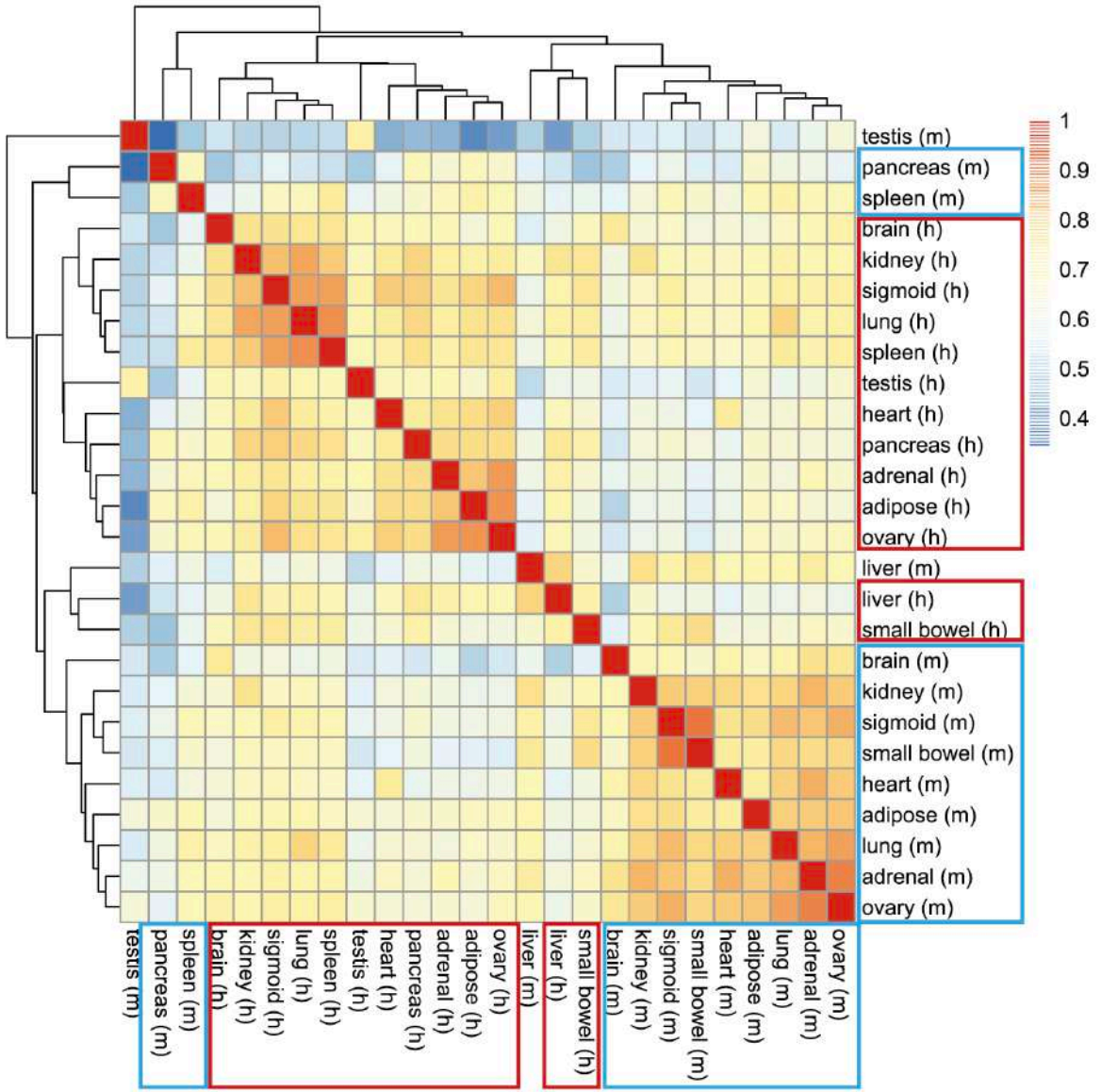
Example of batch effect:

D87PMJN1 (run 253, flow cell D2GUAACXX, lane 7)	D87PMJN1 (run 253, flow cell D2GUAACXX , lane 8)	D4LHBFN1 (run 276, flow cell C2HKJACXX , lane 4)	MONK (run 312, flow cell C2GR3ACXX , lane 6)	HWI-ST373 (run 375, flow cell C3172ACXX , lane 7)
heart	adipose	adipose	heart	brain
kidney	adrenal	adrenal	kidney	pancreas
liver	sigmoid colon	sigmoid colon	liver	brain
small bowel	lung	lung	small bowel	spleen
spleen	ovary	ovary	testis	Human
testis		pancreas		Mouse

http://f1000research.com/articles/4-121/v1

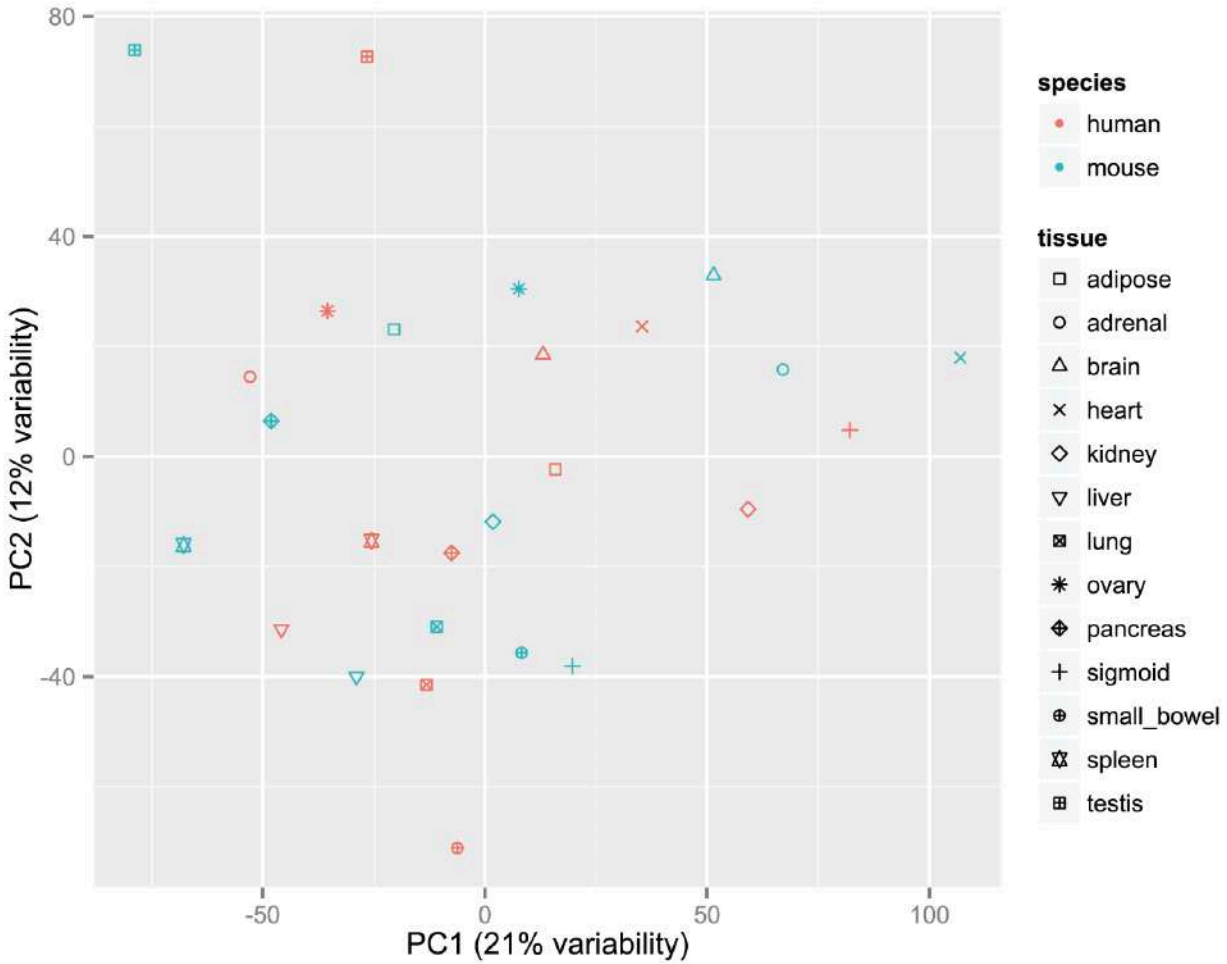
Recapitulating the patterns reported by the mouse ENCODE papers

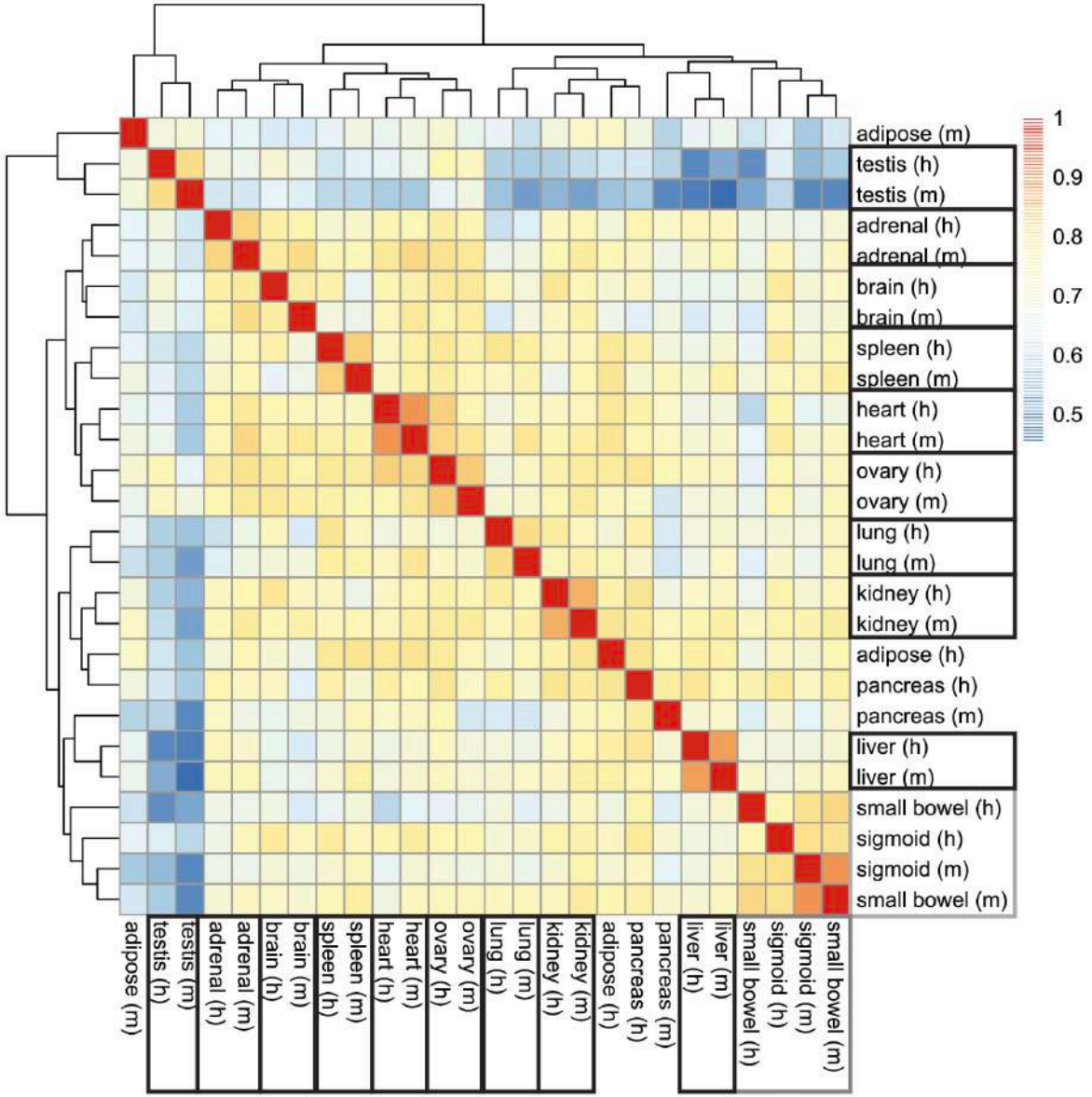




http://f1000research.com/articles/4-121/v1

Clustering of data once batch effects are accounted for

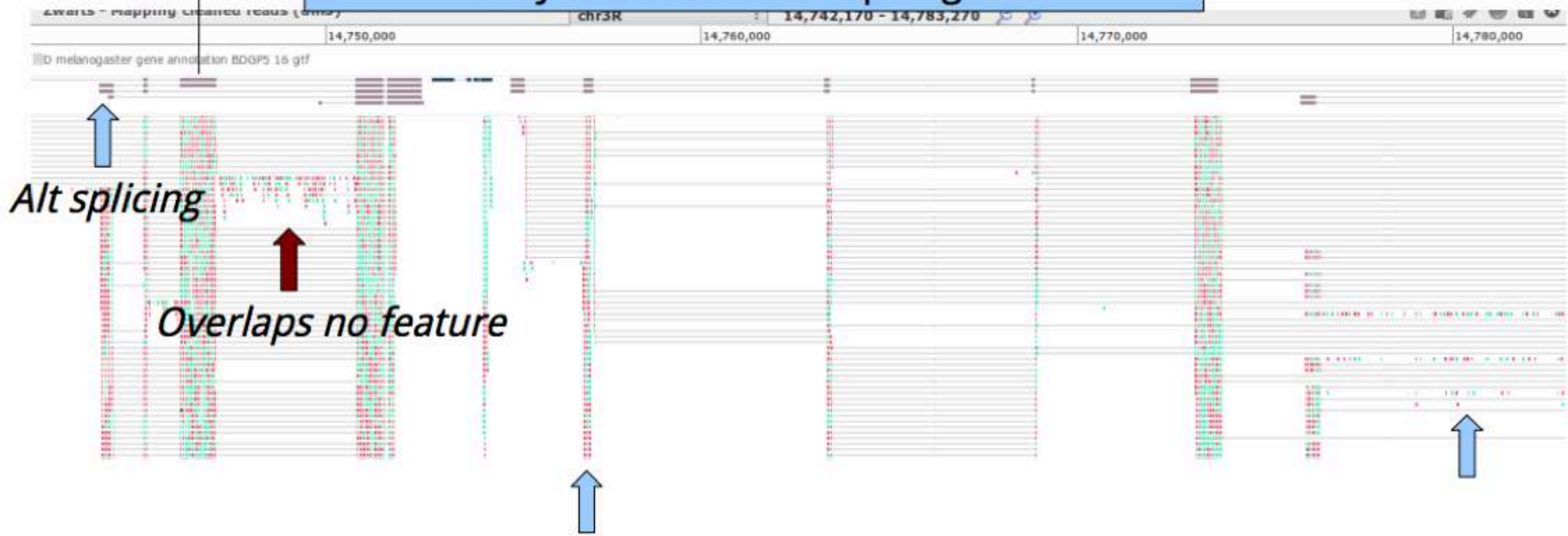




http://f1000research.com/articles/4-121/v1

Once you have mappings, you can start counting

'Exons' are the type of *features* used here.
They are summarized per 'gene'



Concept:

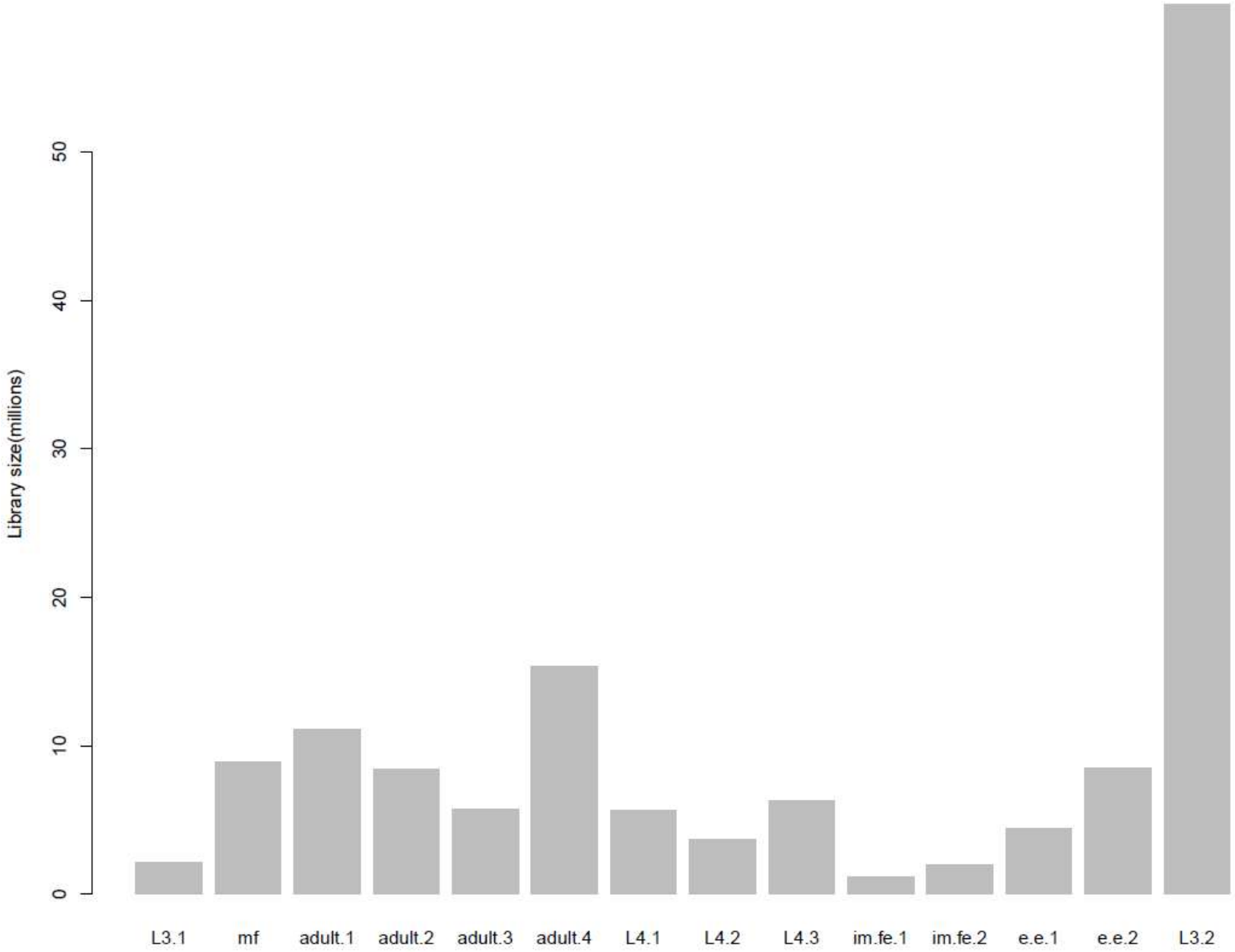
GeneA = exon 1 + exon 2 + exon 3 + exon 4 = 215 reads

GeneB = exon 1 + exon 2 + exon 3 = 180 reads

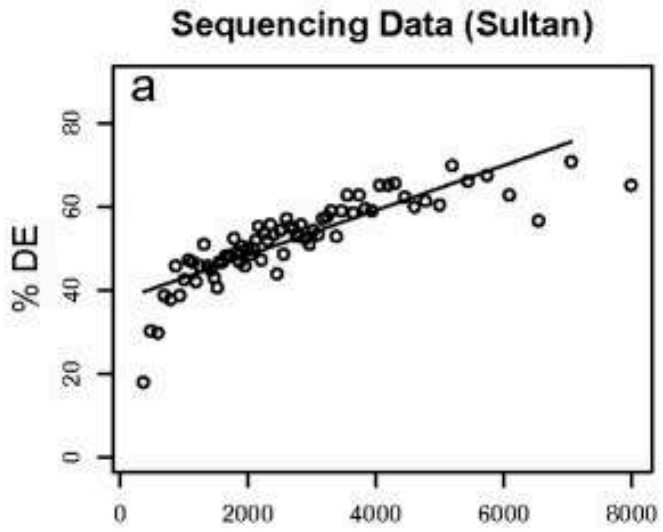
This is the bit we care about!

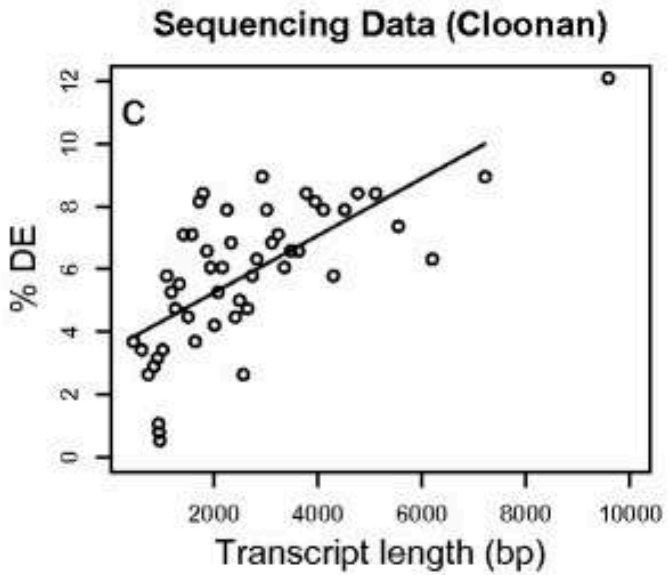
Counts of the gene depends on **expression** ,transcript length ,sequencing depth and simply chance

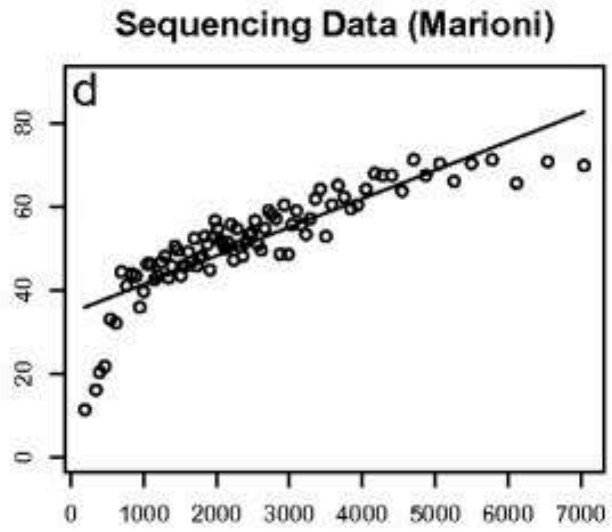
Higher sequencing depth equals more counts

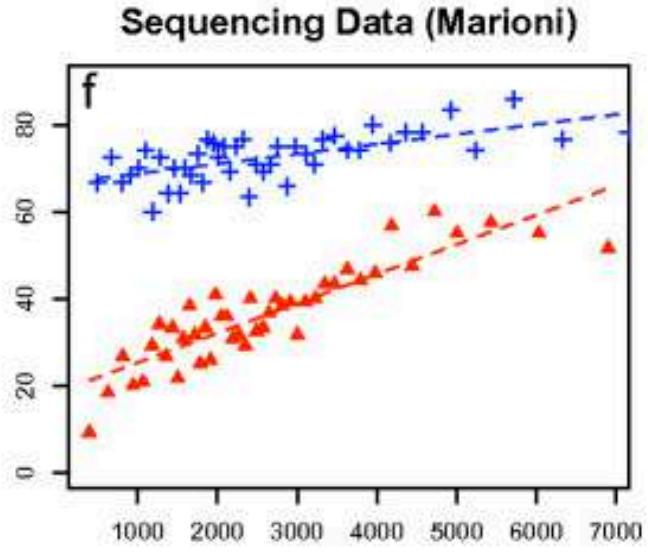


Counts are proportional to the transcript length x mRNA expression level









33% of highest expressed genes
33% of lowest expressed genes

Oshlack and Wakefield (2009) Biology Direct.

Normalization: different goals

- **R/FPKM:** (Mortazavi et al. 2008)
  - Correct for: differences in sequencing depth and transcript length
  - Aiming to: compare a gene across samples and diff genes within sample

- **TMM**: (Robinson and Oshlack 2010)
  - Correct for: differences in transcript pool composition; extreme outliers
  - Aiming to: provide better across-sample comparability

- **TPM**: (Li et al 2010, Wagner et al 2012)
  - Correct for: transcript length distribution in RNA pool
  - Aiming to: provide better across-sample comparability

- **Limma voom (logCPM)**: (Lawet al 2013)
  - Aiming to: stabilize variance; remove dependence of variance on the mean

Optimal Scaling of Digital Transcriptomes

Gustavo Glusman, Juan Caballero, Max Robinson, Burak Kutlu, Leroy Hood

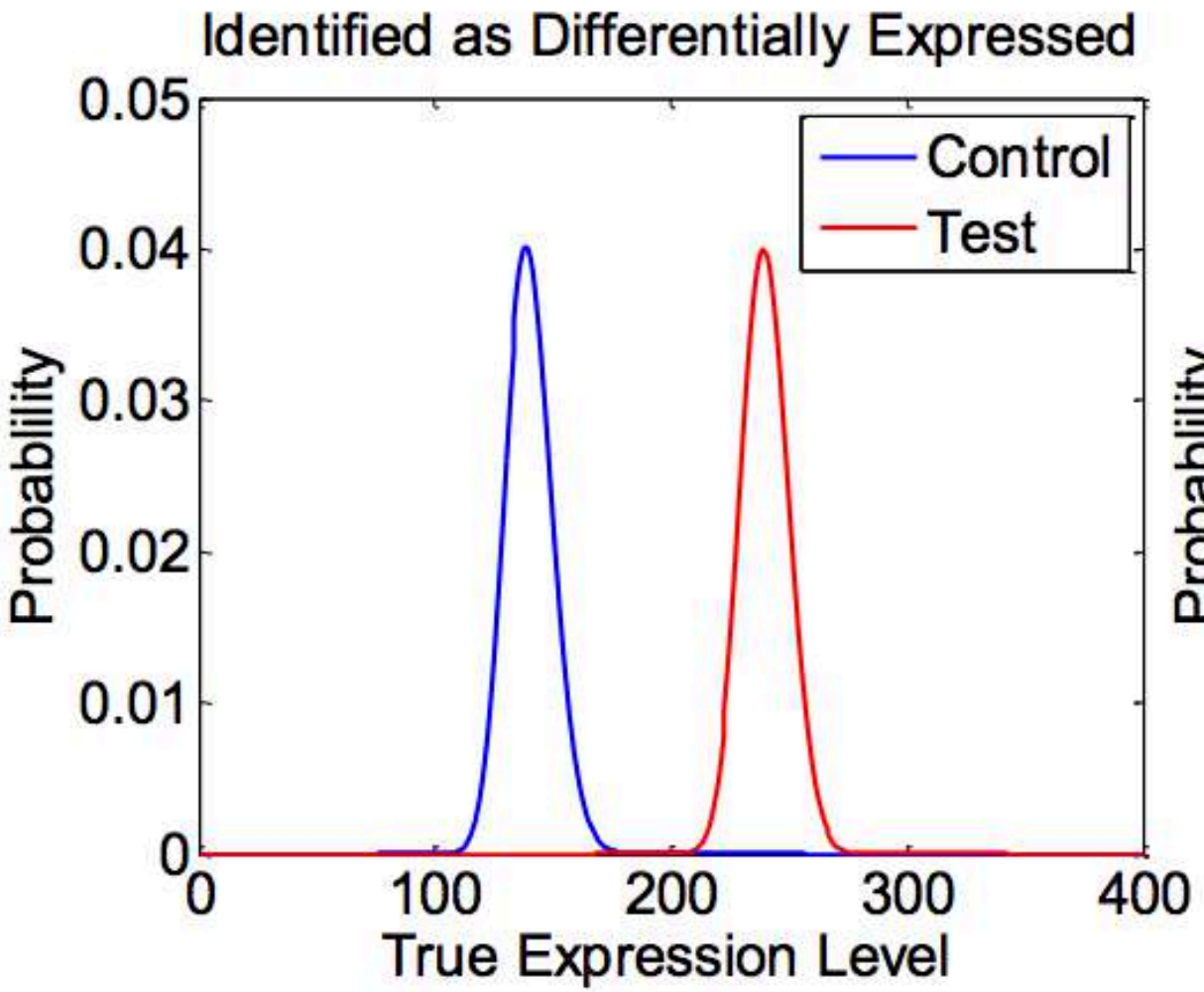
Published: Nov 06, 2013 • DOI: 10.1371/journal.pone.0077885

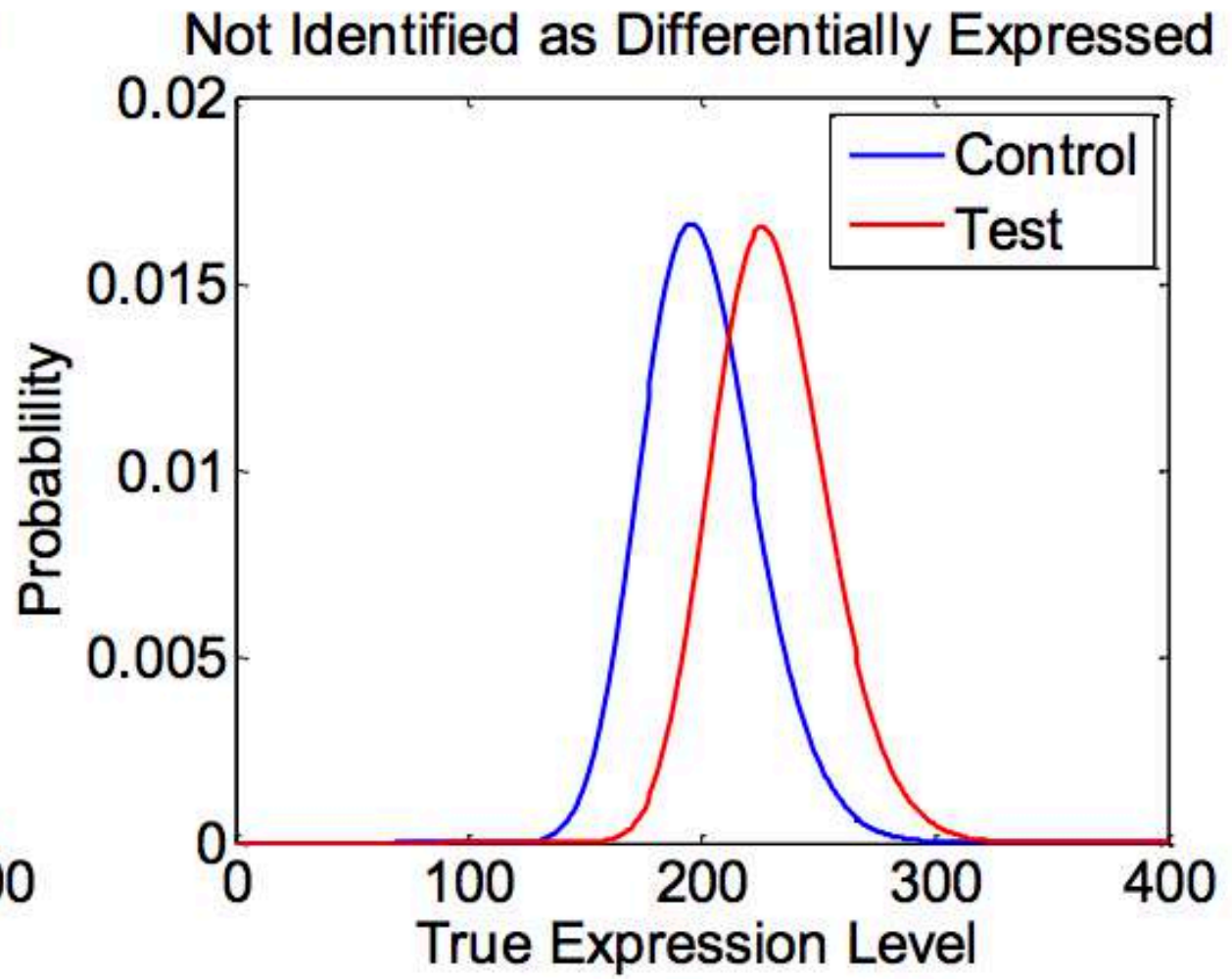
But how do you know your count = 2 is really 2?

- Differentially expressed genes = counts of genes change between conditions **more systematically** than expected by chance
- Need **biological and technical replicates** to detect differential expression

Differential expression

Fitting a distribution for every gene for **DE**

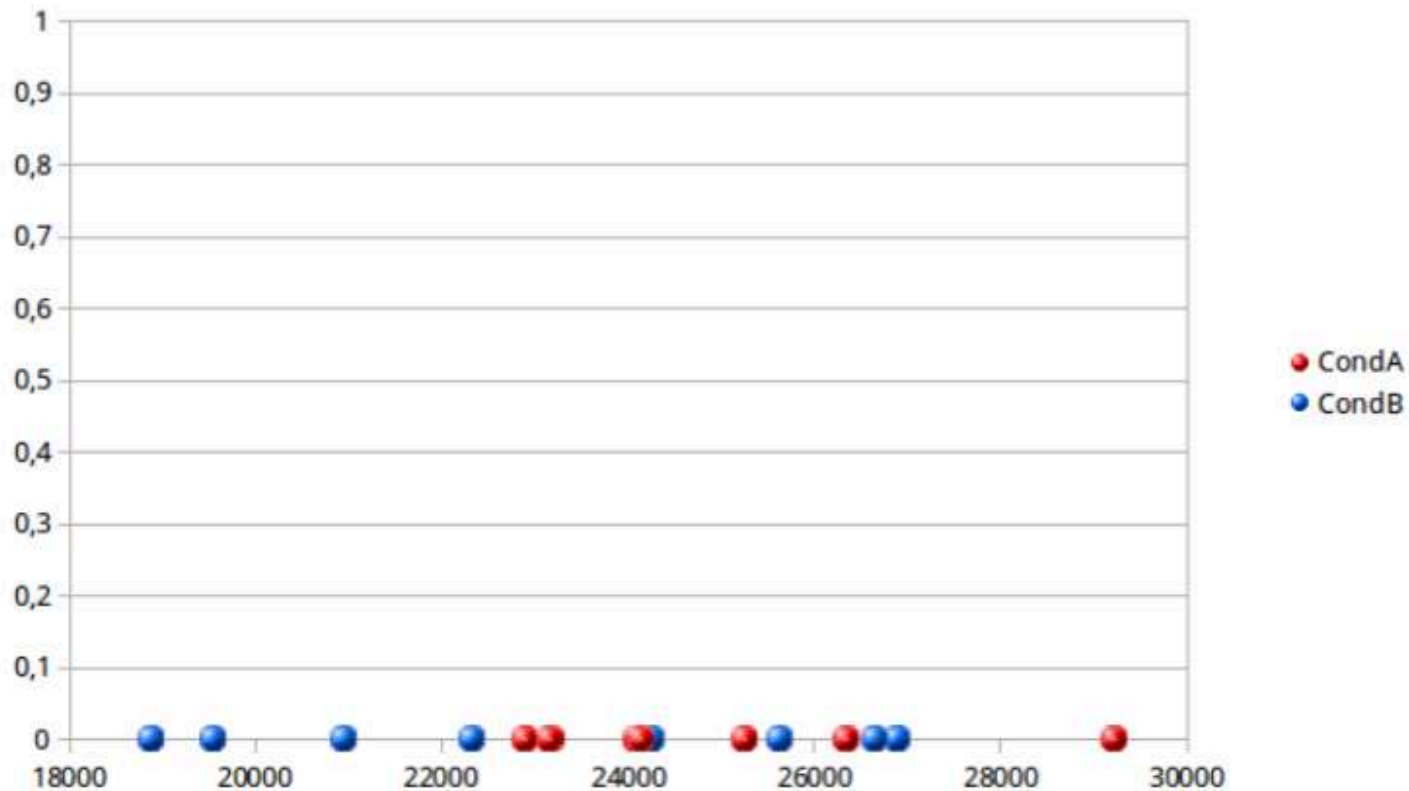




. Busby *et al.,* Bioinformatics (2013)

Scenario

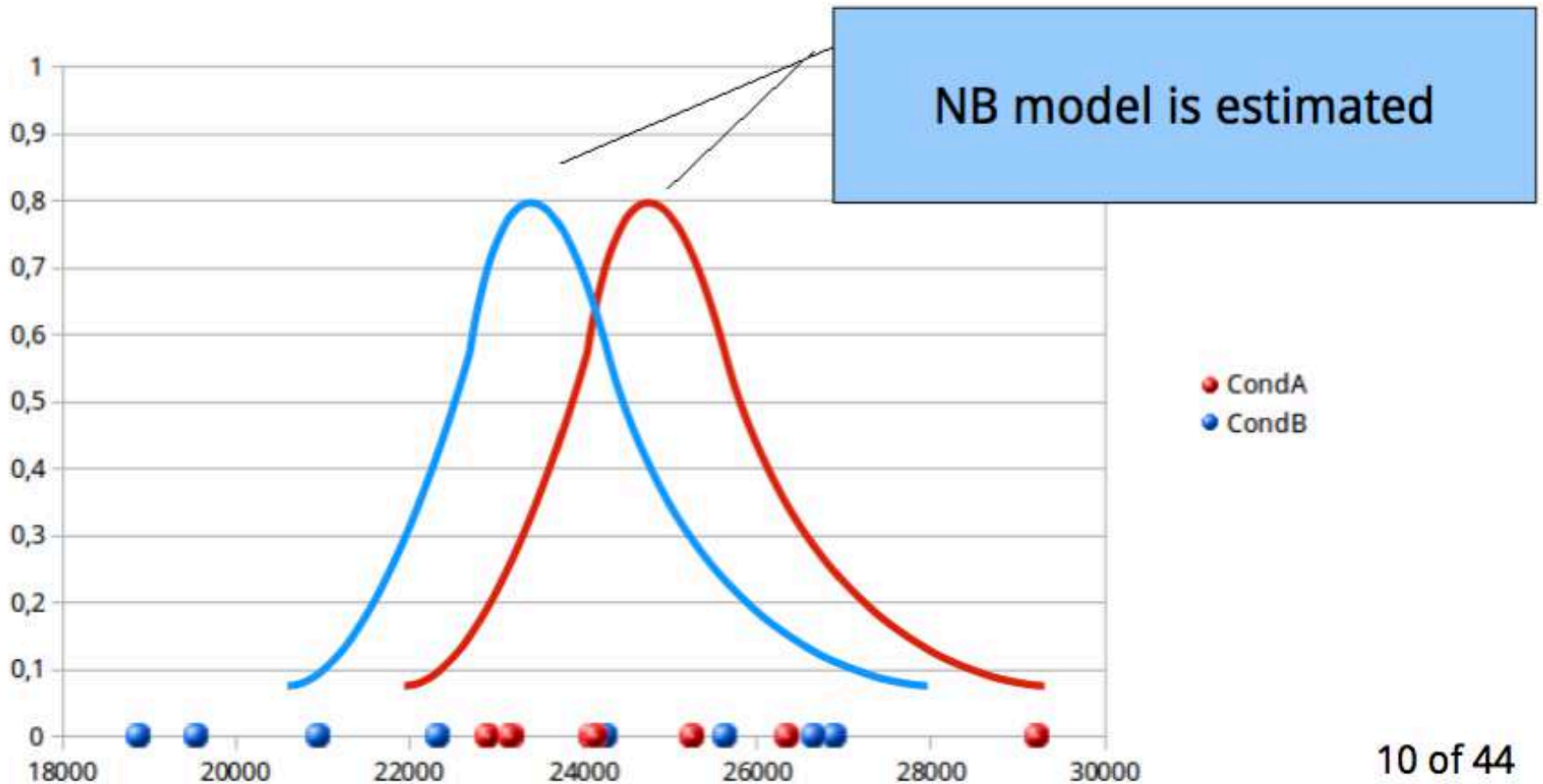
gene_id	CAF0006876							
Condition A	sample1 23171	sample2 22903	sample3 29227	sample4 24072	sample5 23151	sample6 26336	sample7 25252	sample8 24122
Condition B	Sample9 19527	sample10 26898	sample11 18880	sample12 24237	sample13 26640	sample14 22315	sample15 20952	sample16 25629



Scenario

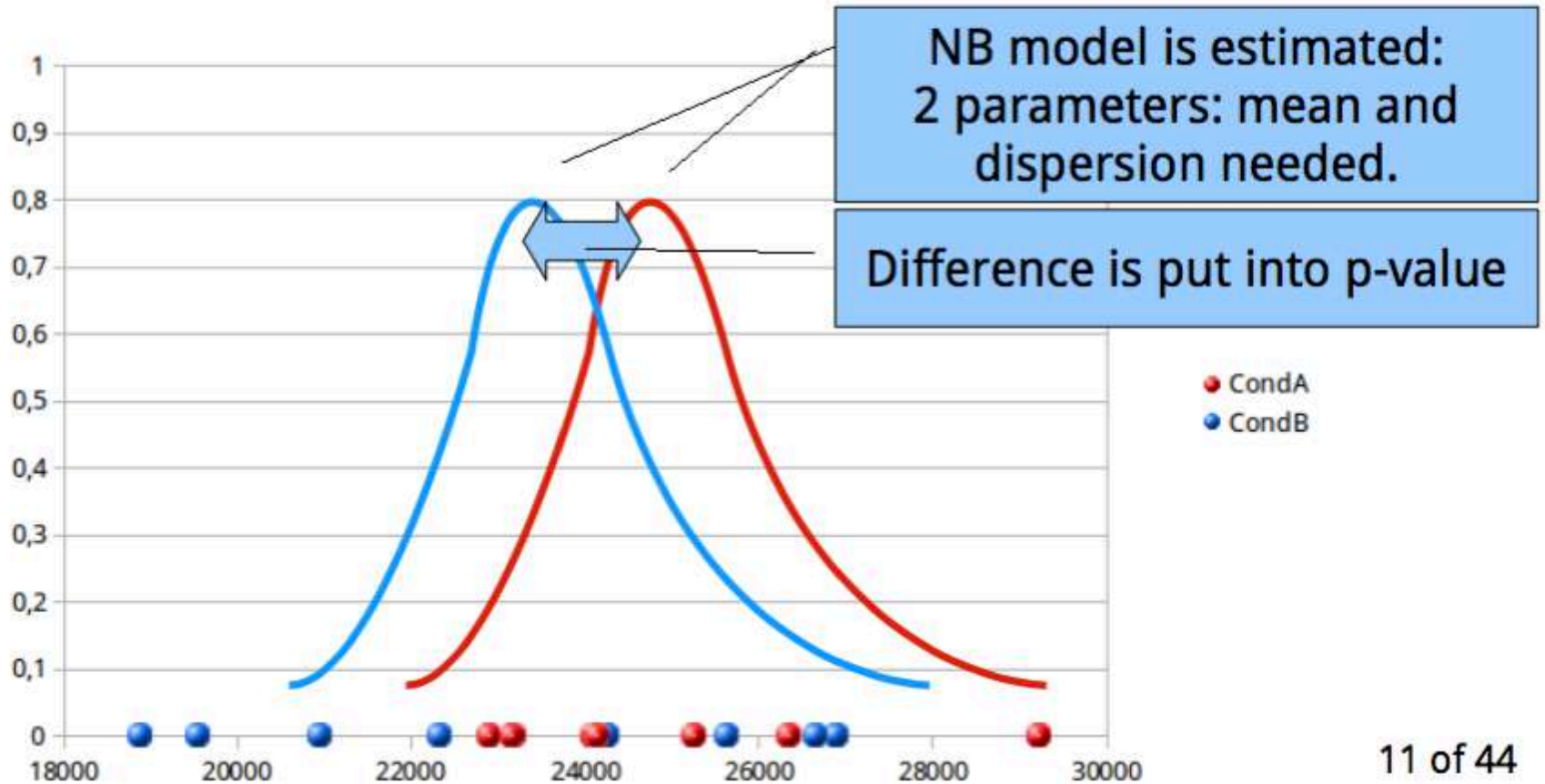
gene_id CAF0006876

Condition A	sample1 23171	sample2 22903	sample3 29227	sample4 24072	sample5 23151	sample6 26336	sample7 25252	sample8 24122
Condition B	Sample9 19527	sample10 26898	sample11 18880	sample12 24237	sample13 26640	sample14 22315	sample15 20952	sample16 25629



Scenario

gene_id	CAF0006876							
Condition A	sample1 23171	sample2 22903	sample3 29227	sample4 24072	sample5 23151	sample6 26336	sample7 25252	sample8 24122
Condition B	Sample9 19527	sample10 26898	sample11 18880	sample12 24237	sample13 26640	sample14 22315	sample15 20952	sample16 25629



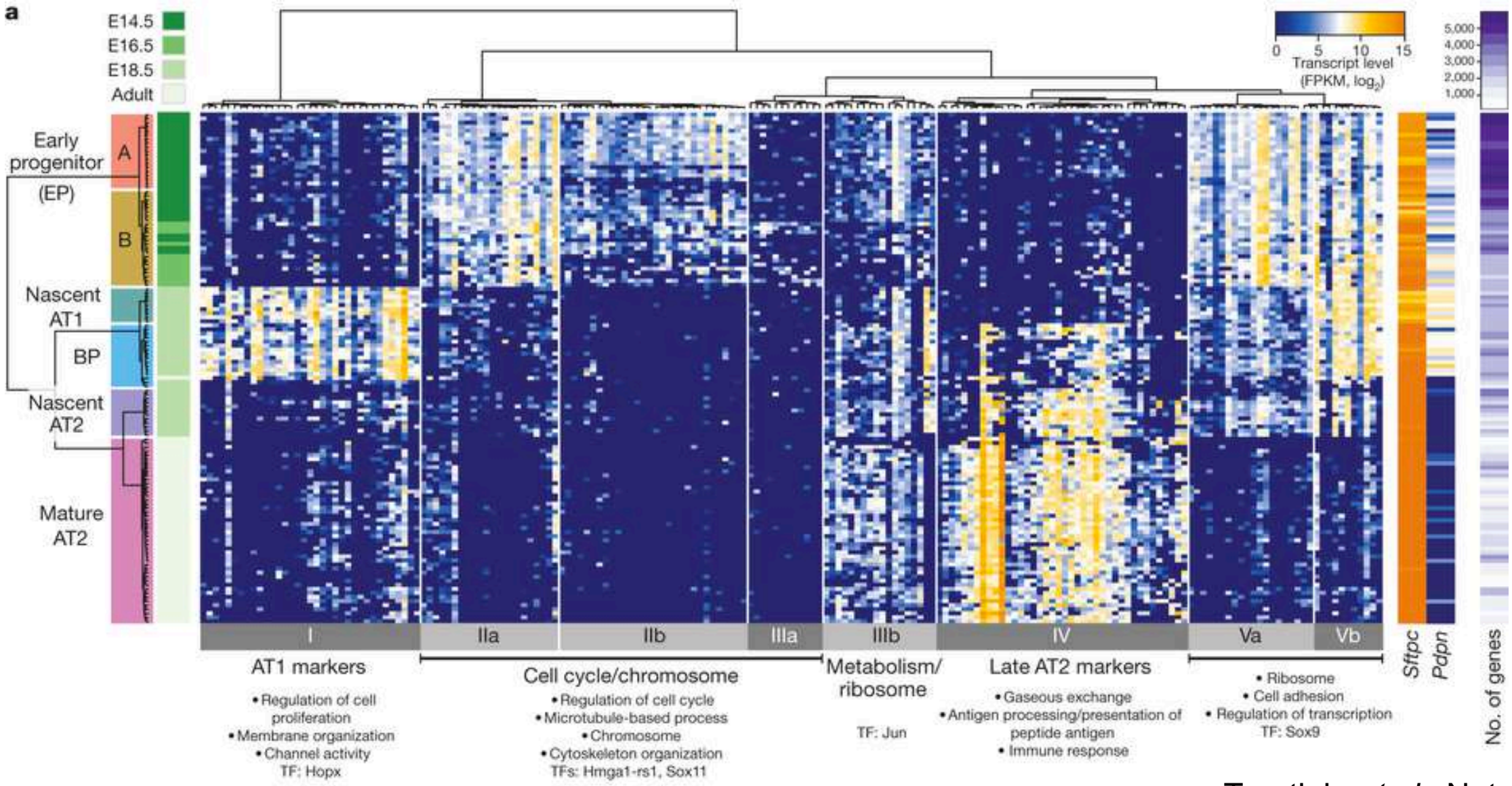
NB model is estimated:
2 parameters: mean and dispersion needed.

Difference is put into p-value

VIB JAKONIX IRCU

Once you have set of differentially expressed genes

Summarization visualizing the expression data through heatmap ; Classification using Gene Ontology terms and metabolic annotations



Treutlein *et al.,* Nature (2014)

Amplicon / Metagenomics: An Intro

A clinician's guide to microbiome analysis

Marcus J. Claesson[1,2], Adam G. Clooney[1–3]* and Paul W. O'Toole[1,2]*

Abstract | Microbiome analysis involves determining the composition and function of a community of microorganisms in a particular location. For the gastroenterologist, this technology opens up a rapidly evolving set of challenges and opportunities for generating novel insights into the health of patients on the basis of microbiota characterizations from intestinal, hepatic or extraintestinal samples. Alterations in gut microbiota composition correlate with intestinal and extraintestinal disease and, although only a few mechanisms are known, the microbiota are still an attractive target for developing biomarkers for disease detection and management as well as potential therapeutic applications. In this Review, we summarize the major decision points confronting new entrants to the field or for those designing new projects in microbiome research. We provide recommendations based on current technology options and our experience of sequencing platform choices. We also offer perspectives on future applications of microbiome research, which we hope convey the promise of this technology for clinical applications.

https://www.nature.com/articles/nrgastro.2017.97.pdf

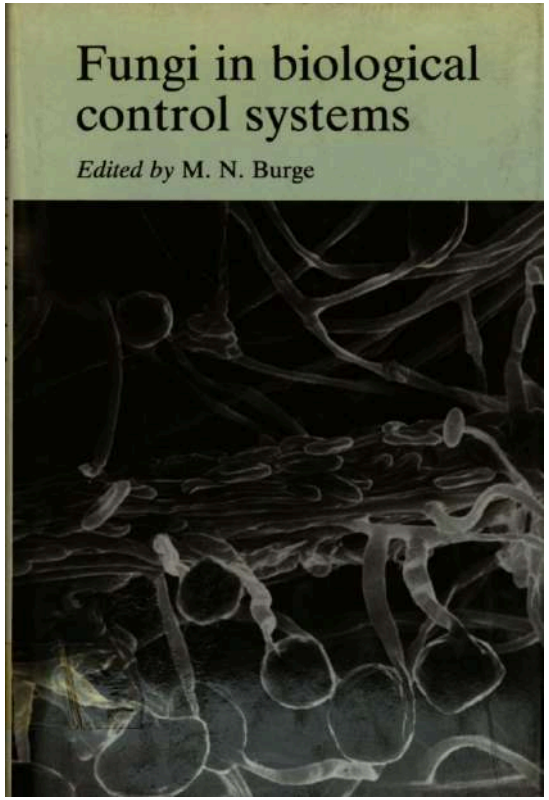
Key points

- Complex communities of microorganisms live on and in the human body, and variations in the composition and function of these communities are increasingly linked to various conditions and diseases
- Although it is not known if microbiome changes are causative or consequential in most pathophysiologies, they might provide biomarkers for disease detection or management
- Microbiome analysis is likely to become a routine component of secondary health care and is emerging as a modifiable environmental risk factor in multifactorial diseases that could be targeted by novel therapeutics
- Technology advancements are leading to a range of powerful methods for microbiome analysis becoming available and affordable for clinical studies
- Judicious choice of sample type and sequencing platform are required to maximize the clinical utility of microbiome data

https://www.nature.com/articles/nrgastro.2017.97.pdf

What is the microbiome?

Fungi in Biological Control Systems (1988)



A convenient ecological framework in which to examine biocontrol systems is that of the microbiome. This may be defined as a characteristic microbial community occupying a reasonably well defined habitat which has distinct physico-chemical properties. The term thus not only refers to the microorganisms involved but also encompasses their theatres of activity. In relation to fungal diseases of crops and their control, major microbiomes are the phylloplane, spermosphere, rhizosphere and rhizoplane, and numerous kinds of plant residues persisting on or in the soil. Mention should also be made of the wood of standing or felled trees as microbiomes where biocontrol of forest diseases using fungi has been achieved. However, in most cases competitive interactions other than mycoparasitism seem to be of greater importance.

http://microbe.net/2015/04/08/what-does-the-term-microbiome-mean-and-where-did-it-come-from-a-bit-of-a-surprise/

And then what is the metagenome?

Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products

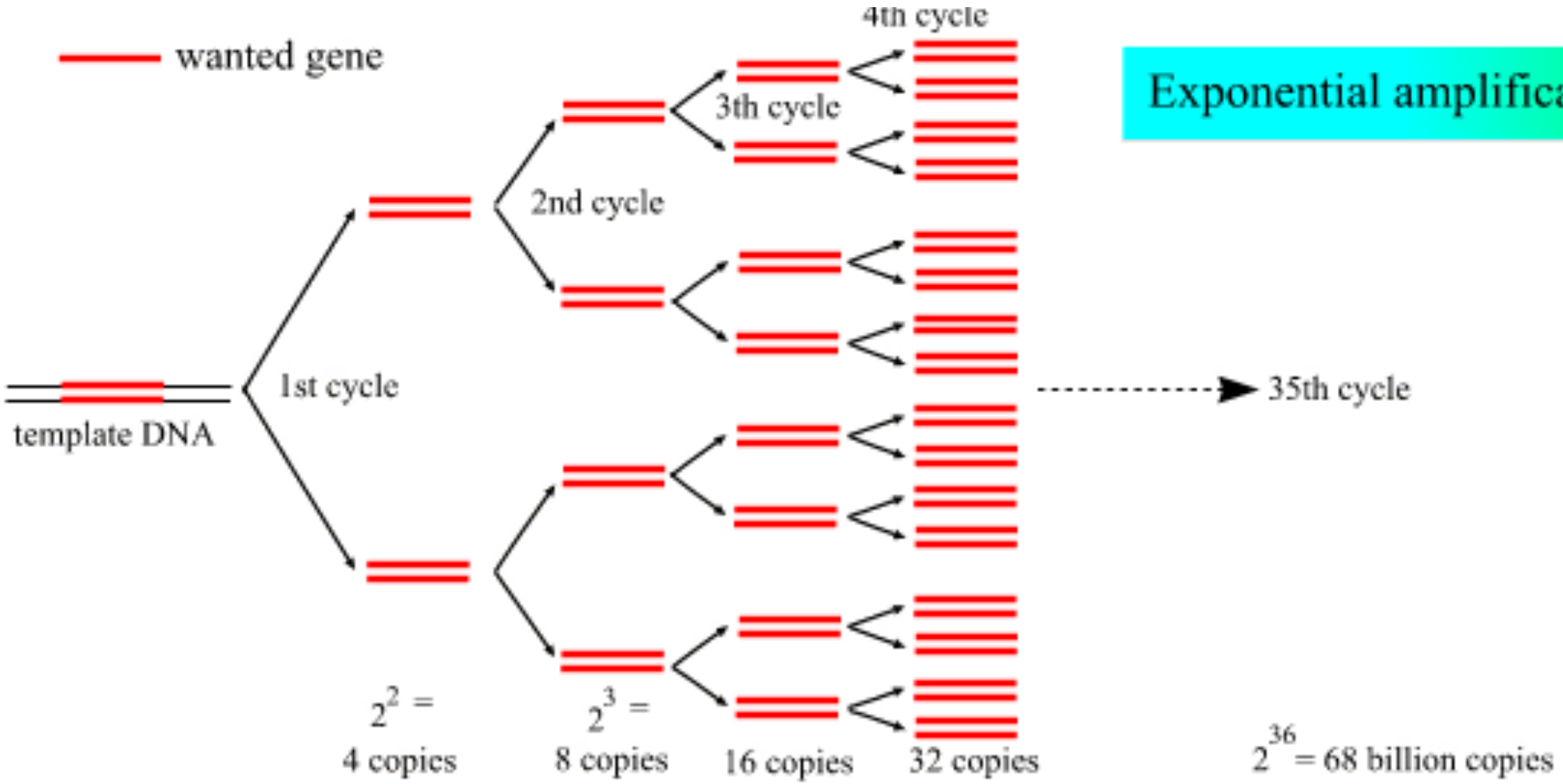
Jo Handelsman[1], Michelle R Rondon[1], Sean F Brady[2], Jon Clardy[2] and Robert M Goodman[1]

Chemistry & Biology October 1998, 5:R245–249
http://biomednet.com/elecref/10745521005R0245

… This approach involves directly accessing the genomes of soil organisms that cannot be, or have not been, cultured by isolating their DNA

What is amplicon sequencing?

Anything that requires PCR-based amplification of a specific target gene (locus)



Exponential amplification

And then what is the metagenome?

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

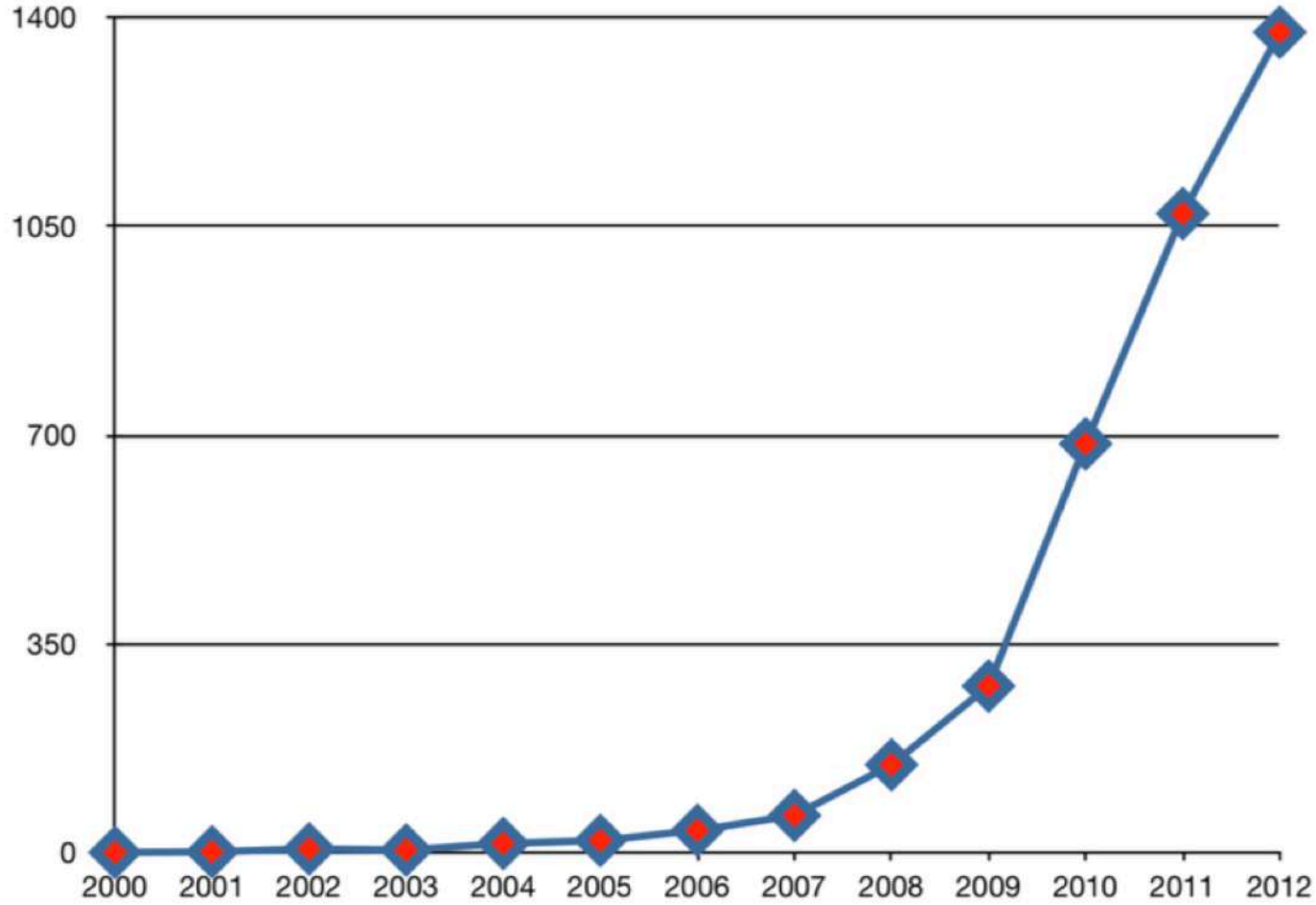
Review

Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities

Kevin Chen*, Lior Pachter*

Metagenomics is the application of modern genomics techniques to the study of communities of microbial organisms directly in their natural environments, bypassing the need for isolation and lab cultivation of individual species. The field has its roots in the culture-independent retrieval of **16S rRNA** genes, pioneered by Pace and colleagues two decades ago.

Pubmed hits for “Microbiome”



Jonathan Eisen, Slideshare

Metagenomics ≠ Amplicon sequencing

Metagenomics is undergoing a crisis

Please don't make things worse ☺

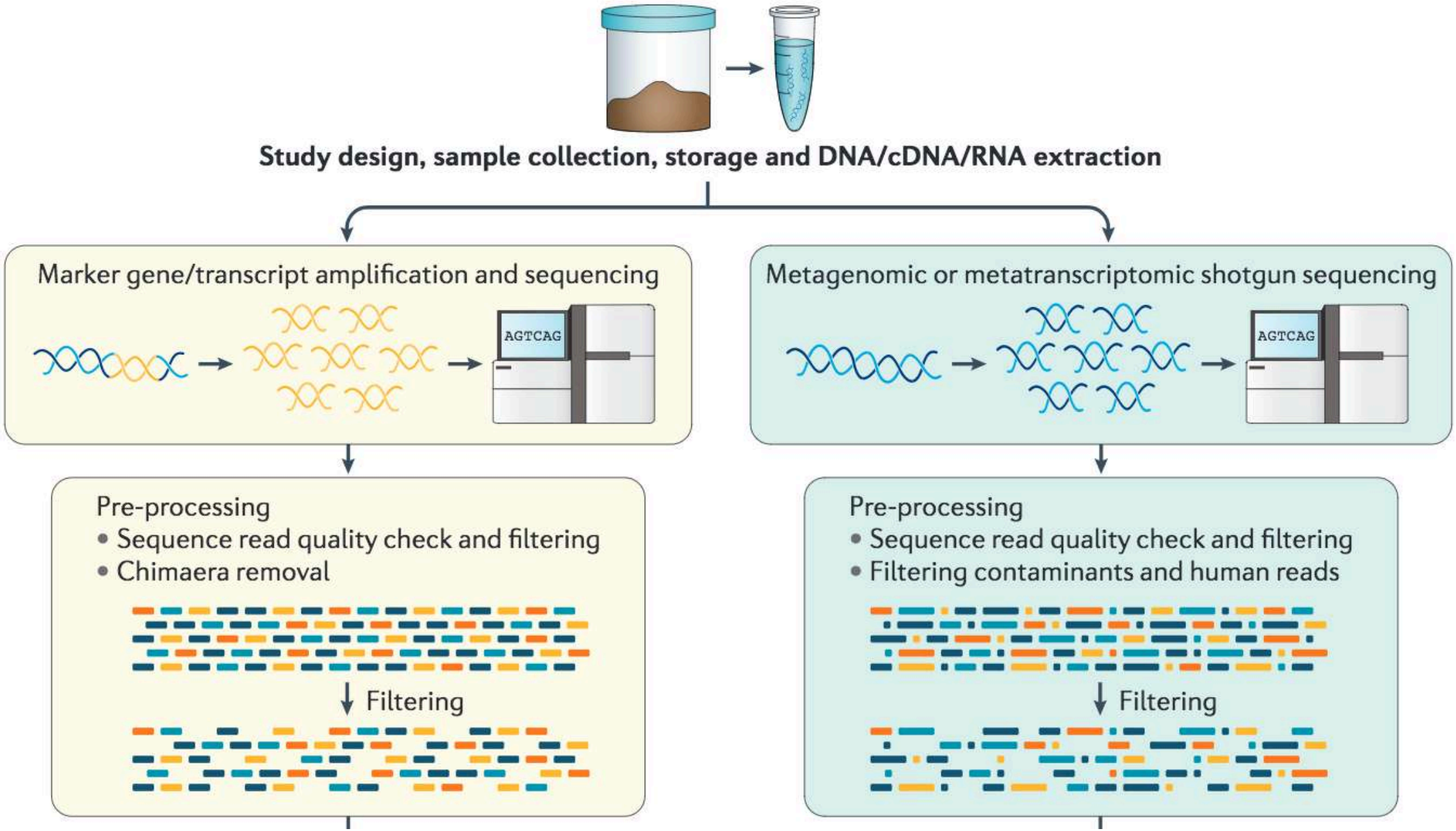
- Crisis 1
  - **The correlation/causation fallacy**. For example….
  - Patients with type II diabetes have a different gut microbiome compared to healthy patients
  - Does the microbiome cause diabetes?
  - Or do they have a different microbiome because they have diabetes? (therefore different diet)
- Crisis 2
  - A lot of people want to do it, but don't know how
  - Errors, bad experimental design, incorrect conclusions

Mick Watson, Slideshare

Basic Purpose

Characteristics of (microbial) community

Who are they?
Where do they come from?
Are their similarities (at what level)
between communities
of different conditions
of similar conditions?
within a community?
What are they doing?
How are they doing?


Study design, sample collection, storage and DNA/cDNA/RNA extraction
Marker gene/transcript amplification and sequencing
AGTCAG
Metagenomic or metatranscriptomic shotgun sequencing
AGTCAG
Pre-processing
• Sequence read quality check and filtering
• Chimaera removal
Filtering
Pre-processing
• Sequence read quality check and filtering
• Filtering contaminants and human reads
Filtering

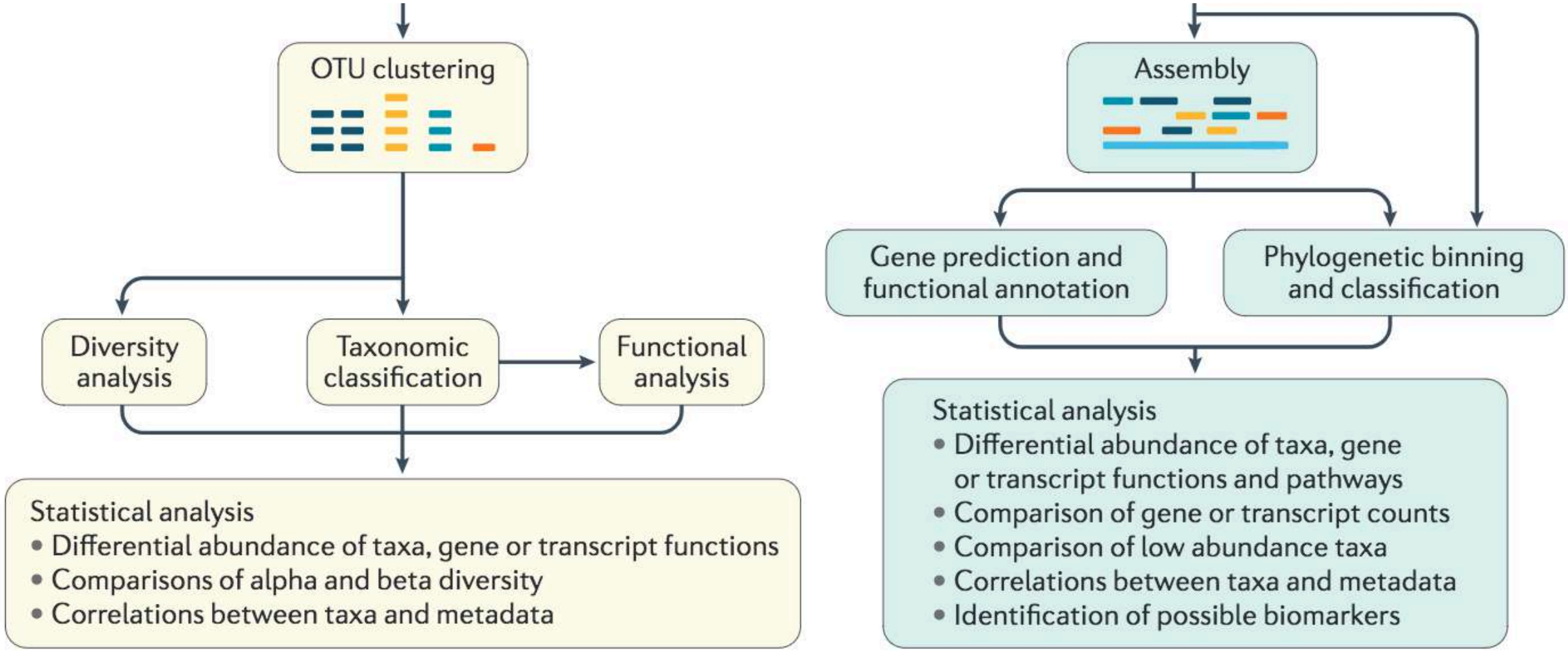


Figure 1 | **Flowchart of the major steps involved in bioinformatic analysis of the microbiome.** The analysis is divided into two sections depending on the type of sequencing. This schematic describes the basic steps and might vary depending on the aim of the analysis. OTU, operational taxonomic unit.

Applications

What have metagenomics been used for?

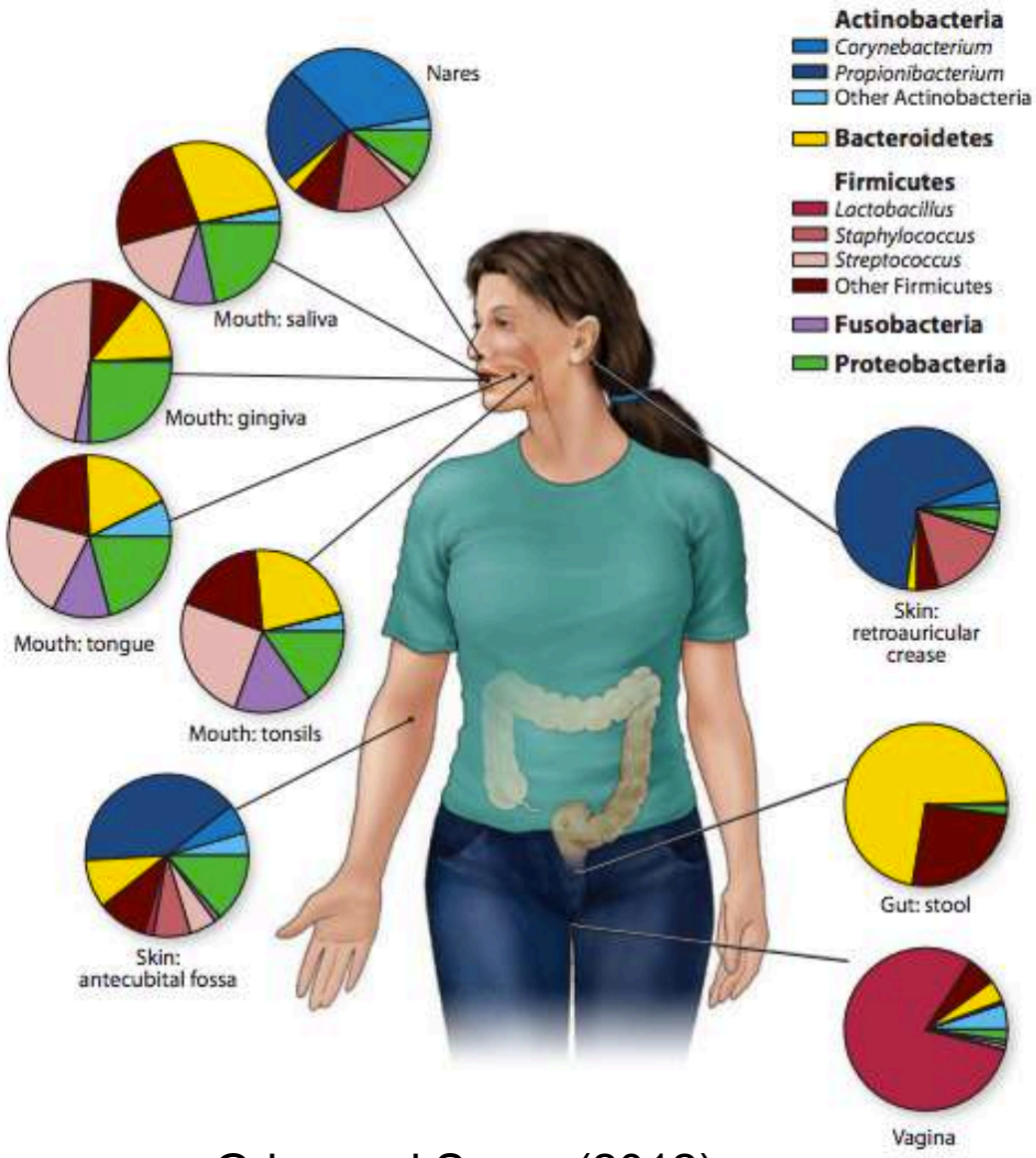
Exploration and categorisation

Rusch *et al*., 2007 **Plos Biology**

- 6.3 Gbp of sequence (2x Human genomes, 2000 x Bacterial genomes)
- Most sequences were novel compared to the databases

Qin *et al*., 2010 **Nature**

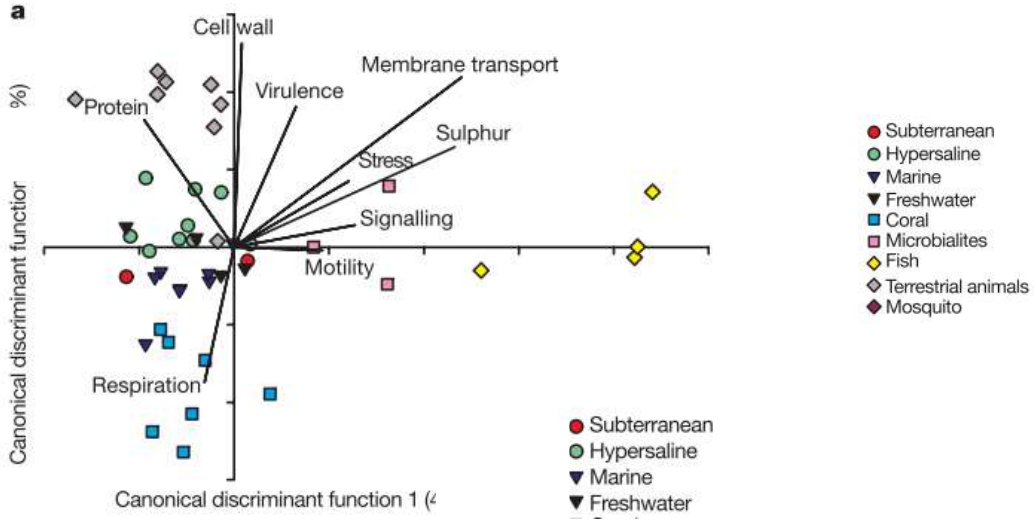
- 127 Human gut metagenomes
- 600 Gbp sequence (200 x Human genomes)
- 3.3 million genes identified
- Minimal gut metagenome defined



Grice and Segre (2012)

What have metagenomics been used for?

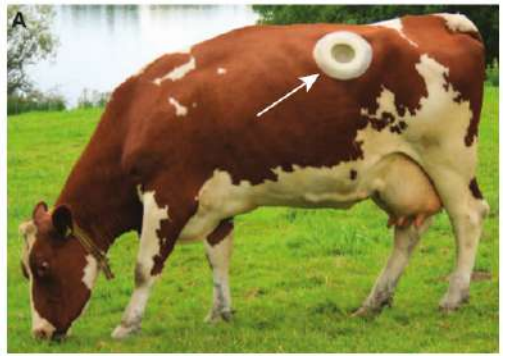
Comparative



Dinsdale *et al*., 2008 **Nature**

- A characteristic microbial fingerprint for each of the nine different ecosystem types

Specific functions

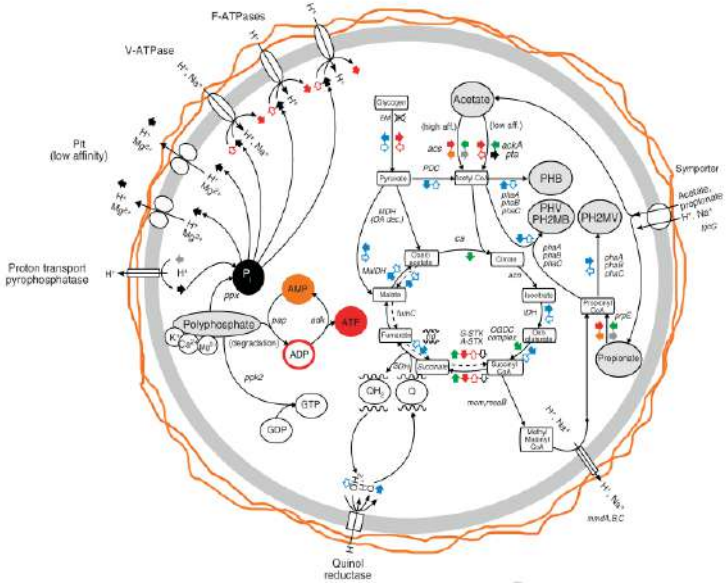


Hess *et al*., 2011 **Science**

- Identified 27.755 putative carbohydrate-active genes from a cow rumen metagenome
- Expressed 90 candidates of which 57% had enzymatic activity against cellulosic substrates

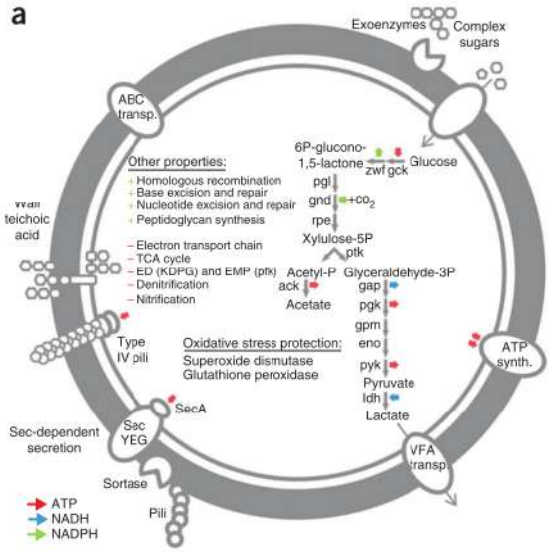
What have metagenomics been used for?

Extracting genomes



Garcia Martin *et al*., 2006 **Nat. Biotechnol.**

- Genome extraction from low complexity metagenome
- *Candidatus* Accumulibacter phosphatis
- The first genome of a polyphosphate accumulating organism (PAO) with a major role en enhanced biological phosphorus removal



Albertsen *et al*., 2013 **Nat. Biotechnol.**

- Genome extraction of low abundant species (< 0.1%) from metagenomes
- First complete TM7 genome
- Access to genomes of the ”uncultured majority”

Concept: OTU (Operational Taxonomic Unit)

OTU for Ecology

Operational Taxonomic Unit: a grouping of similar sequences that can be treated as a single “species”

Strengths

- Conceptually simple
- Mask effect of poor quality data
  - Sequencing error
  - in vitro recombination

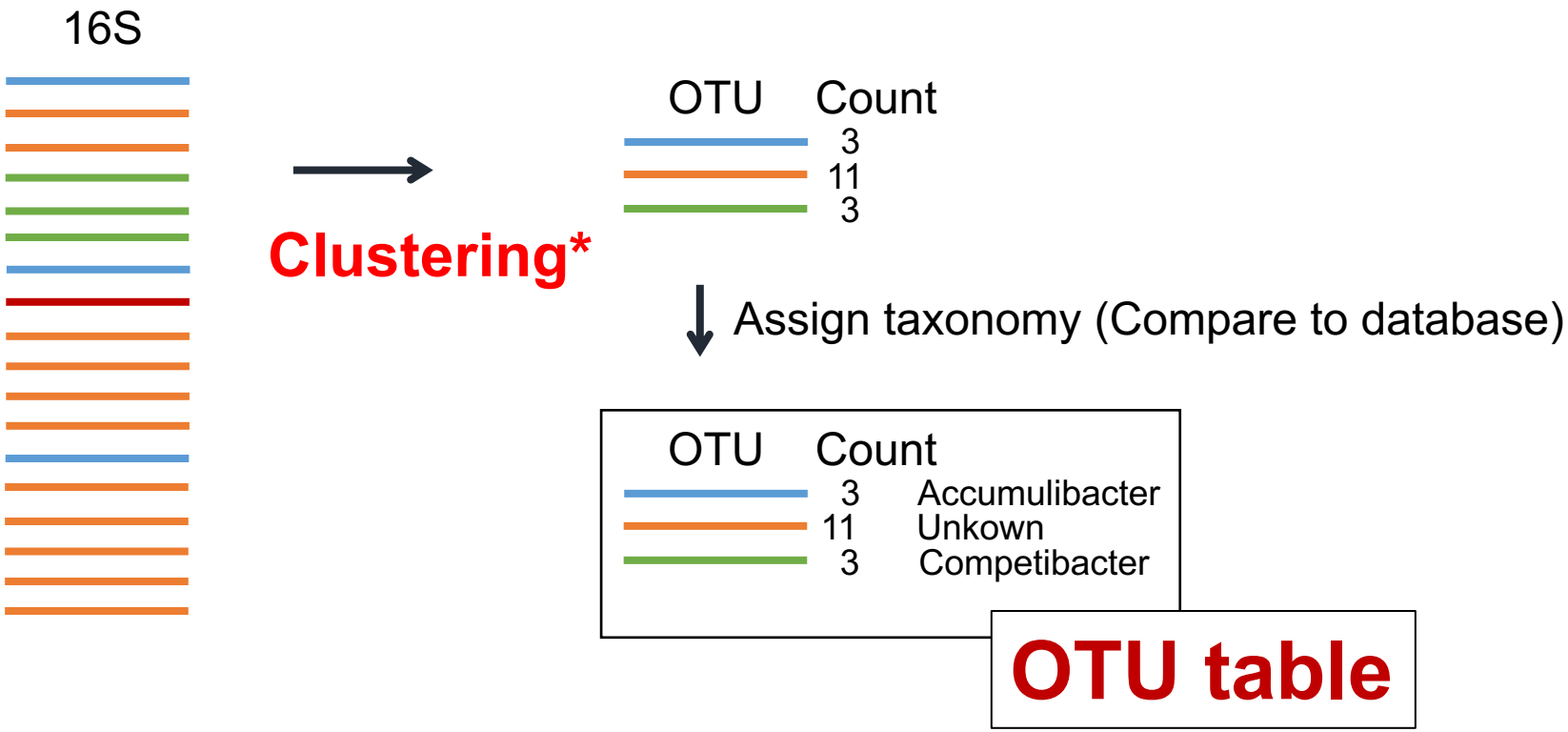
Weaknesses

- Limited resolution
- Logically inconsistent definition

Slide of Aaron Darlin

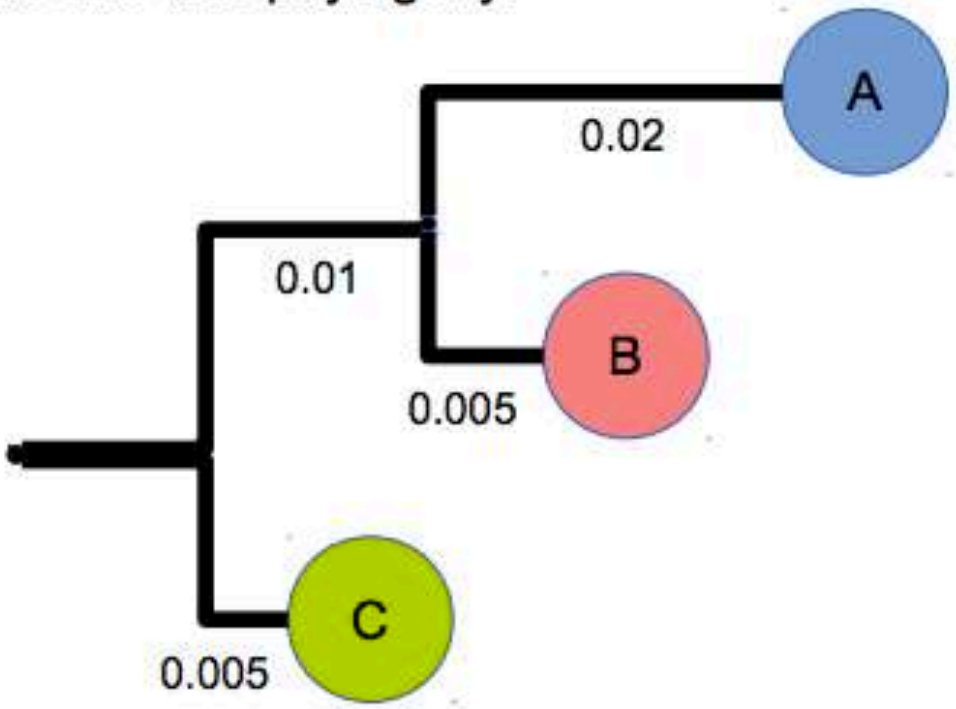
Assign OTU

- Cluster by their similarity to other sequences in the sample (operations taxonomic units → OTU)
- 95% genus level, **97% species level**, 99% strain level



Logical inconsistency: OTUs at 97% ID

Assume the true phylogeny:



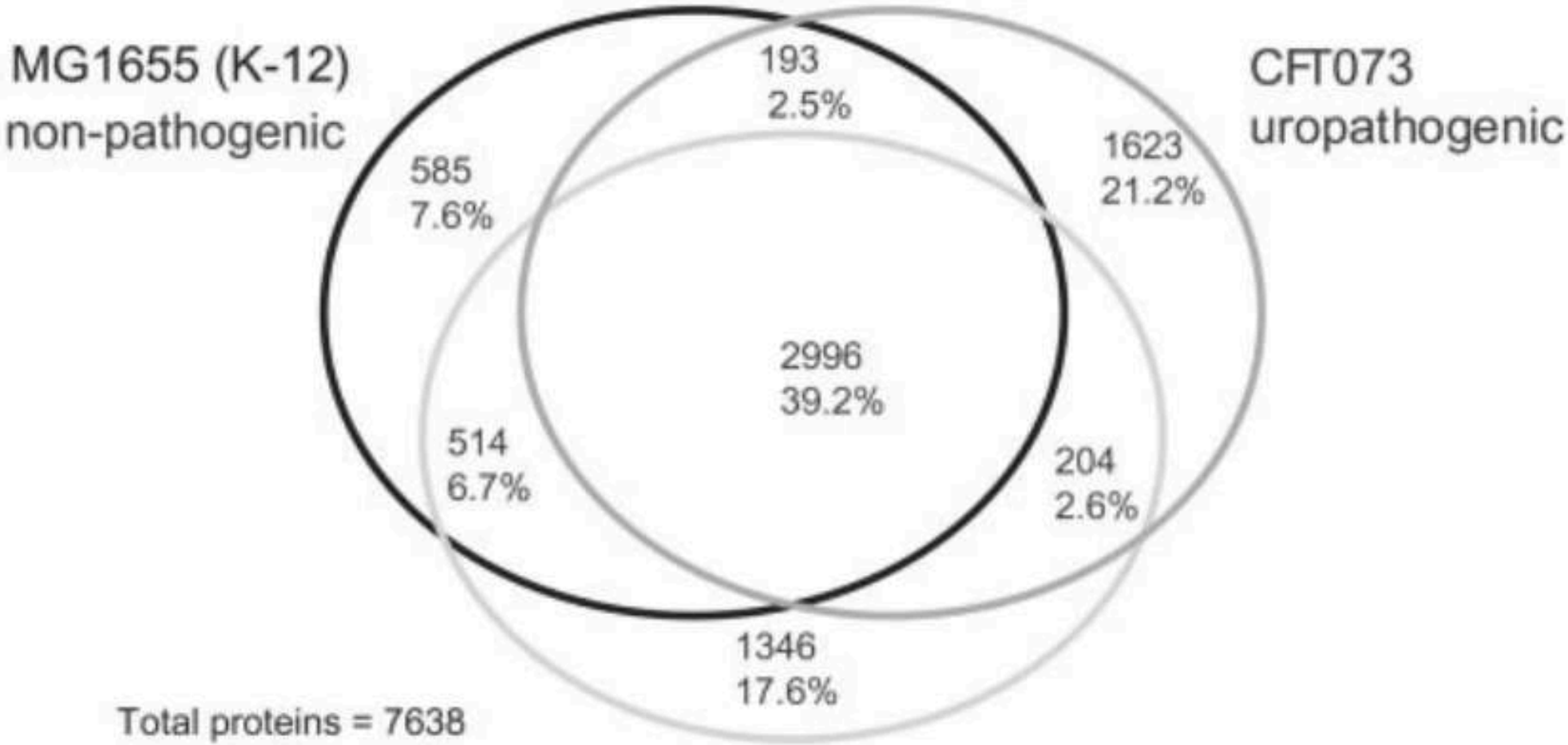
A, B > 97% identity
B, C > 97% identity
A and C not > 97% ID

Possible valid OTUs:
AB, C (with A & C centroids)
A, BC (with A & C centroids)
ABC (with B centroid)

OTU pipelines will arbitrarily pick one of the three solutions.
Is this actually a problem??

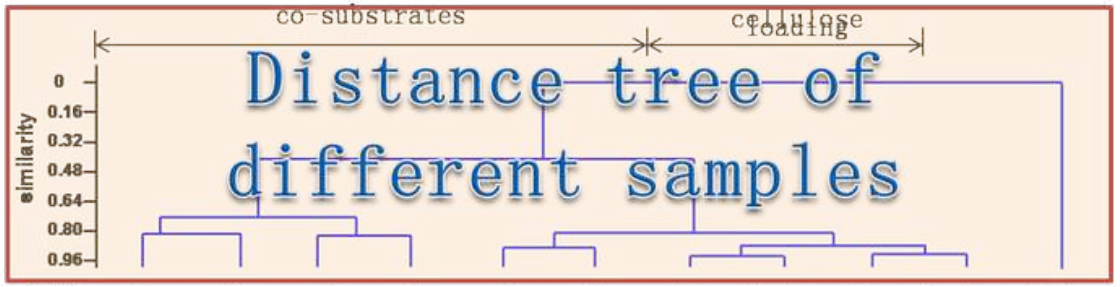
Slide of Aaron Darlin

Same species (16S): Different genomes



Welch et al (2002)

Tree way plot with top OTUs abundance and classification



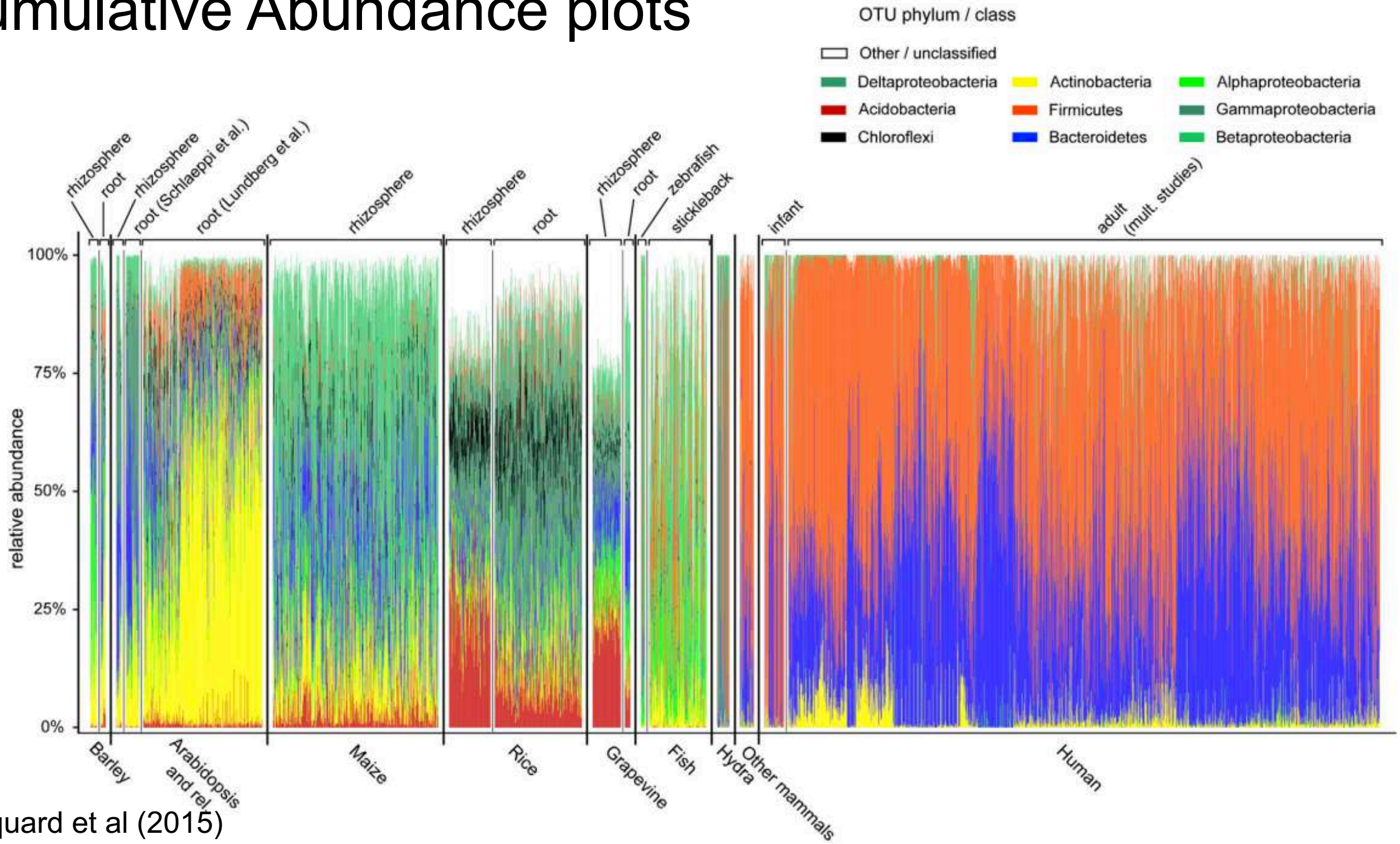
OUT NO.	G1	X1	G2	X2	S1	S2	High	Medium	Low	Inoculum	Seed
1	16.394%	18.740%	9.072%	10.154%	31.778%	26.680%	32.332%	32.312%	40.620%	39.410%	0.040%
2	7.282%	6.904%	4.352%	4.994%	12.884%	13.464%	13.726%	13.334%	15.630%	13.790%	0.006%
3	7.108%	7.174%	4.070%	4.610%	13.604%	12.230%	13.612%	13.276%	15.874%	14.348%	0.006%
4	5.264%	5.418%	2.988%	3.376%	10.168%	9.328%	10.632%	11.062%	11.600%	10.104%	0.004%
5	0.002%	0.002%	1.136%	0.000%	0.666%	2.178%	0.000%	0.000%	0.006%	0.000%	0.000%
6	0.338%	0.296%	0.382%	0.244%	0.564%	0.866%	0.006%	0.036%	0.070%	0.816%	0.000%
7	2.460%	3.044%	1.586%	2.646%	4.904%	5.052%	2.068%	2.434%	0.708%	1.532%	0.002%
8	1.508%	1.246%	0.980%	1.338%	2.340%	3.154%	1.282%	1.542%	0.460%	0.814%	0.002%
9	1.192%	1.142%	0.784%	1.022%	2.026%	2.516%	1.066%	1.258%	0.386%	0.750%	0.000%
10	1.098%	0.[illegible]%	0.626%	[illegible]	[illegible]42%	[illegible]22%	0.870%	1.192%	0.336%	0.582%	0.000%
11	3.686%	2.[illegible]%	[illegible]%	[illegible]%	[illegible]218%	[illegible]%	[illegible]%	[illegible]%	[illegible]%	0.000%	0.000%
12	1.862%	0.904%	[illegible]%	0.864%	0.042%	[illegible]%	[illegible]%	[illegible]%	[illegible]%	0.002%	0.002%
13	1.756%	0.946%	1.302%	0.698%	0.068%	0.078%	0.002%	0.002%	0.000%	0.000%	0.000%
14	5.978%	12.532%	16.236%	16.356%	[illegible]%	[illegible]%	[illegible]008%	0.006%	0.004%	0.008%	0.010%
15	10.494%	6.218%	8.710%	4.958%	[illegible]%	[illegible]%	[illegible]002%	0.002%	0.010%	0.000%	0.002%
16	3.152%	5.022%	7.460%	7.656%	0.404%	1.214%	0.004%	0.004%	0.006%	0.004%	0.010%
17	8.504%	4.474%	6.390%	3.668%	0.256%	0.332%	0.000%	0.002%	0.008%	0.004%	0.002%
18	2.764%	5.346%	6.940%	7.068%	0.446%	1.230%	0.006%	0.002%	0.008%	0.000%	0.004%
19	4.642%	2.480%	3.866%	2.190%	0.162%	0.210%	0.000%	0.002%	0.002%	0.002%	0.002%
20	2.098%	3.808%	4.594%	4.858%	0.384%	0.986%	0.000%	0.008%	0.002%	0.000%	0.004%
21	1.486%	0.276%	4.892%	0.530%	0.106%	0.396%	0.000%	0.002%	0.002%	0.000%	0.000%
22	0.018%	0.300%	0.042%	3.894%	0.016%	0.064%	0.000%	0.006%	0.000%	0.002%	0.002%
23	0.014%	0.362%	0.064%	5.142%	0.018%	0.134%	0.002%	0.002%	0.000%	0.000%	0.004%
24	0.146%	0.216%	0.084%	0.112%	0.040%	0.018%	1.244%	1.048%	0.662%	1.742%	0.004%
25	0.138%	0.140%	0.062%	0.054%	0.002%	0.006%	0.922%	0.838%	0.536%	1.036%	0.000%
26	0.012%	0.016%	0.002%	0.002%	0.038%	0.020%	0.004%	0.006%	0.004%	0.050%	3.950%
27	0.000%	0.002%	0.000%	0.004%	0.004%	0.004%	0.002%	0.008%	0.002%	0.000%	4.204%
28	0.108%	0.122%	0.018%	0.044%	1.392%	0.316%	0.614%	1.012%	0.880%	0.440%	0.000%
29	0.166%	0.078%	0.014%	0.026%	0.210%	0.102%	1.744%	0.870%	0.914%	0.152%	0.000%
30	0.002%	0.010%	0.008%	0.004%	0.072%	0.016%	0.000%	5.428%	0.128%	0.478%	0.006%
Coverage	89.7%	90.4%	91.1%	89.0%	86.9%	87.6%	80.1%	85.7%	88.9%	86.1%	8.3%

Top OTU abundance matrix

OTU	classification
1	*Clostridium sp.* 4-2a
2	*Clostridium sp.* 4-2a
3	*Clostridium sp.* 4-2a
4	*Clostridium sp.* 4-2a
5	*Anoxybacillus sp.* HT8
6	*Clostridiaceae bacterium* 37-7-2Cl
7	*Clostridium sp.* YN5
8	*Clostridium sp.* YN6
9	*Clostridium sp.* YN7
10	*Clostridium sp.* YN8
11	*Thermoanaerobacterium xylanolyticum* LX-11
12	*Thermoanaerobacterium xylanolyticum* LX-11
13	*Thermoanaerobacterium xylanolyticum* LX-11
14	*Thermoanaerobacterium aotearoense* SCUT27
15	*Thermoanaerobacterium xylanolyticum* LX-11
16	*Thermoanaerobacterium aotearoense* SCUT27
17	*Thermoanaerobacterium xylanolyticum* LX-11
18	*Thermoanaerobacterium aotearoense* SCUT27
19	*Thermoanaerobacterium xylanolyticum* LX-11
20	*Thermoanaerobacterium aotearoense* SCUT27
21	*Thermoanaerobacterium saccharolyticum* D22
22	*Thermoanaerobacterium thermosaccharolyticum* CECT5853T
23	*Thermohydrogenium kirishiense* DSM11055T
24	*Uncultured Bacillus sp.* clone 93
25	*Uncultured Bacillus sp.* clone 94
26	*Mycobacterium gilvum* B4
27	*Nocardioides terrae* VA15
28	*Anaerolinea thermolimosa* IMO-1
29	*Uncultured bacterium* clone 49c
30	*Clostridium cellobioparum* JCM 1422

http://www.shuixia100.com/my-blog/mothur-tutorial-1

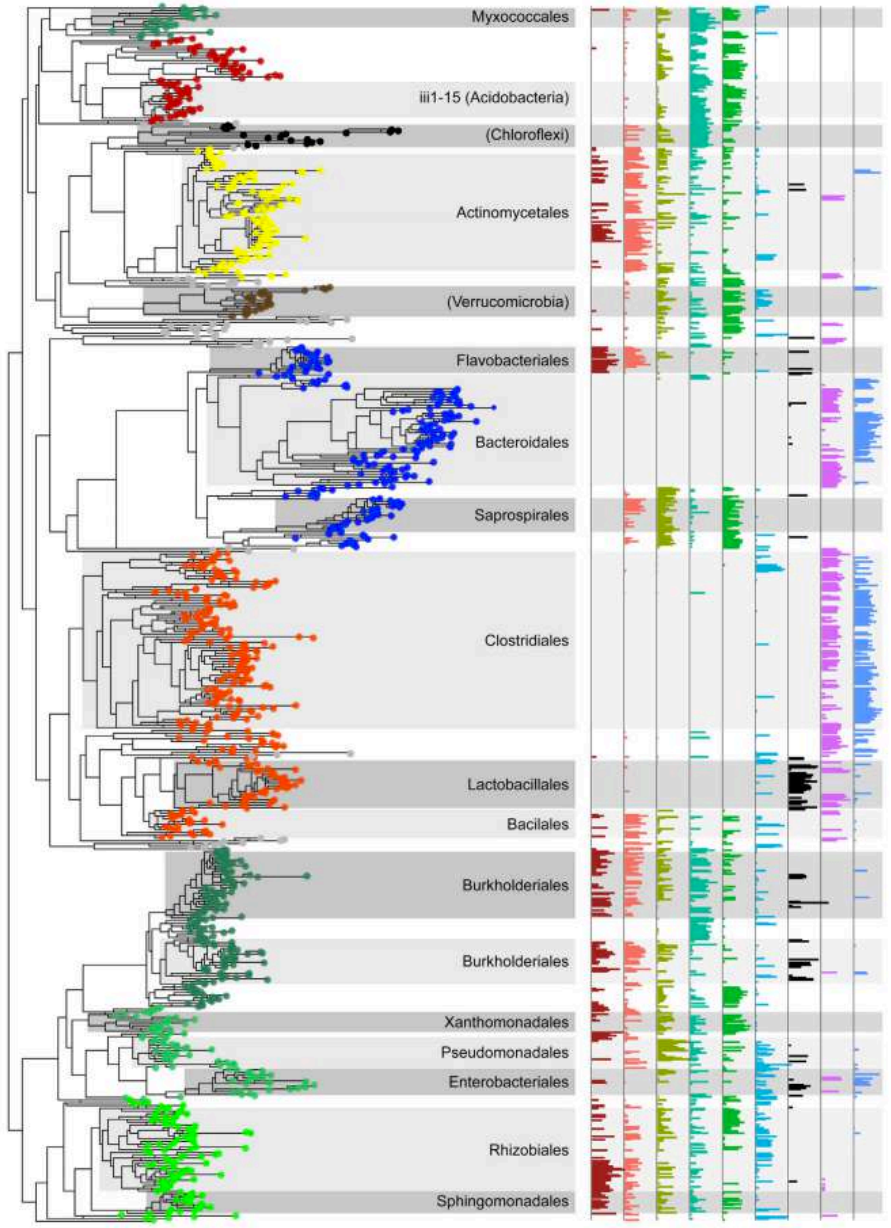
Cumulative Abundance plots



Hacquard et al (2015)

Phylogenetic Analysis of OTU abundances

Relationship between OTUs



How do we compare between different samples?

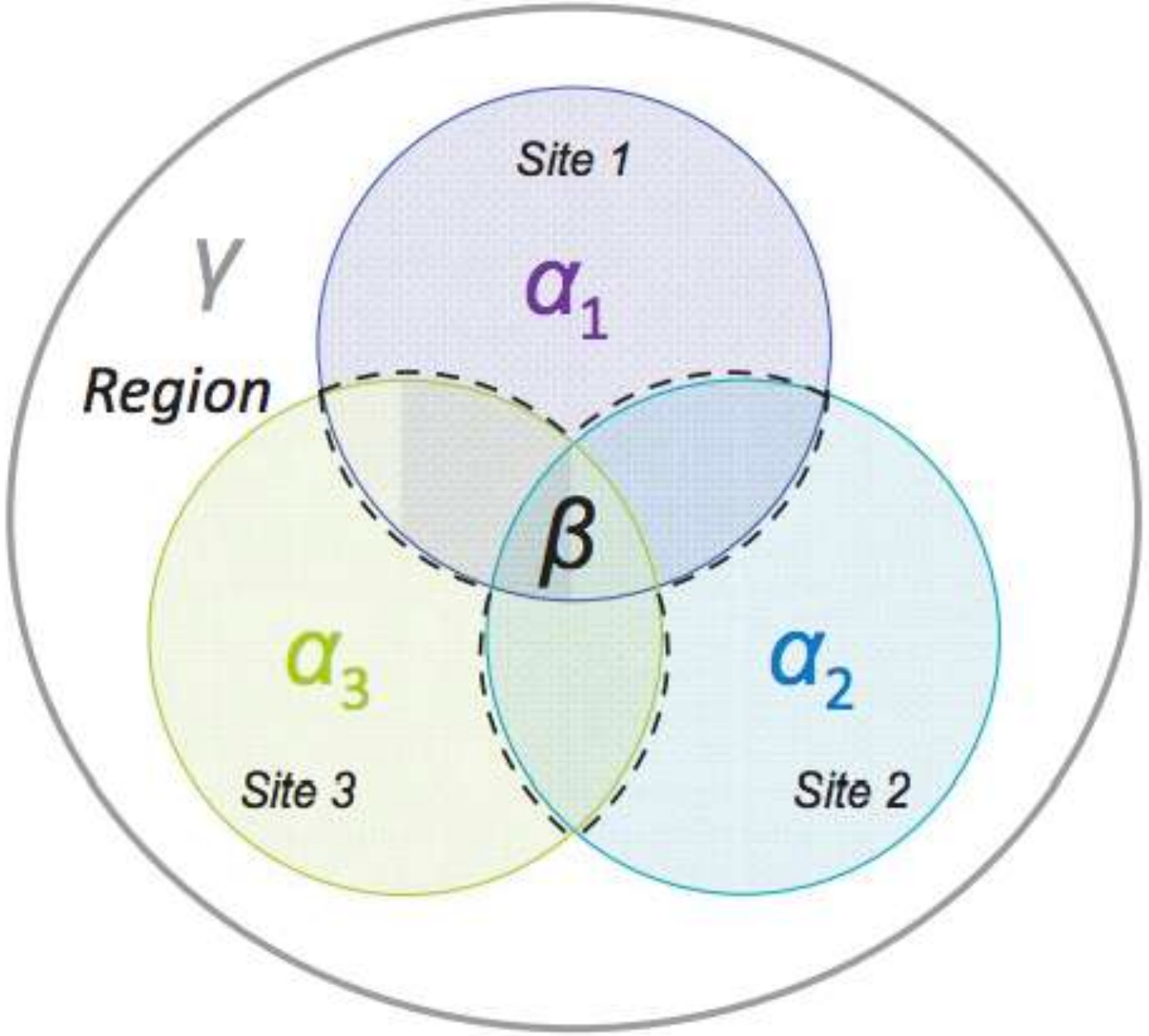
Hacquard et al (2015)

Concept: Diverisity measures

Measures of biodiversity

Zinger et al (2012)

"… measuring biodiversity consists of characterizing the **number**, **composition** and **variation** in taxonomic or functional units **(OTU)** over a wide range of biological organizations"



Measures of biodiversity

Zinger et al (2012)

Alpha diversity refers to the diversity within one location or sample. It is often measured as species richness (i.e. number of species), seldom as species evenness (extent of species dominance). Species richness is strongly sensitive to sampling effort, and requires standardized samples, or the use of estimators that corrects undersampling biases, such as Chao1 or ACE. Evenness is less affected by undersampling biases and is usually assessed with Simpson's or Pielou's indices or rank abundance curves (review in Magurran 2004).

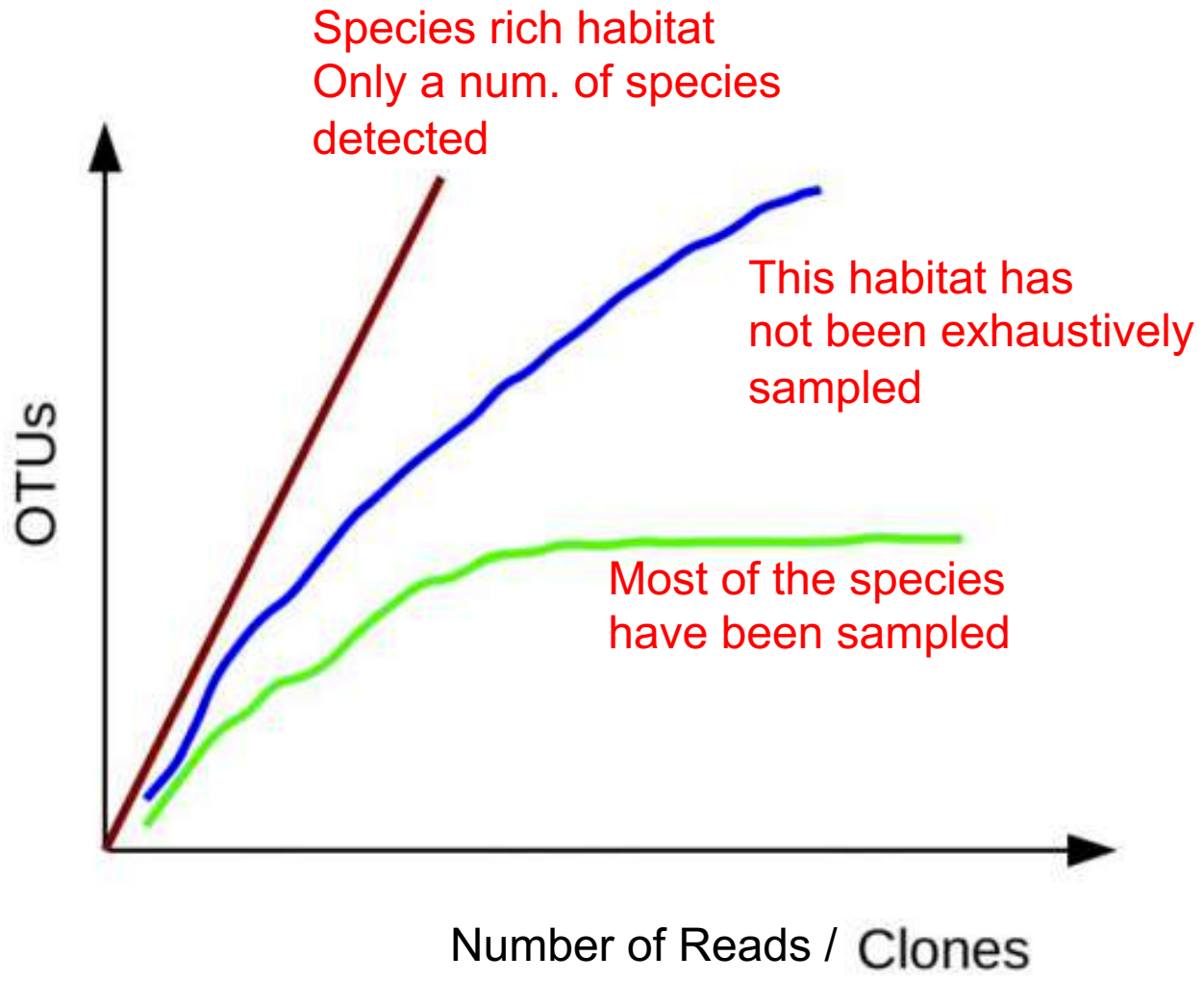
Beta diversity consists in determining the difference in diversity or community composition between two or more locations or samples (i) by considering species composition only, and use incidence data with associated metrics such as Jaccard or Sorensen similarity indices or (ii) by taking species relative abundances into account, and use Bray–Curtis or Morisita–Horn dissimilarity measures (Anderson *et al.* 2011). Using abundance data is, however, strongly discussed among microbiologists when dealing with rRNA gene data because of variations in gene copy number among strains (Acinas *et al.* 2004b; Zhu *et al.* 2005) as well as PCR artefacts.

Gamma diversity, or regional diversity, is similar to alpha diversity but applies for a larger area that encompasses the units under study.

Finally, the spatial scale of investigation can produce very different results and should be consistent in cross-study comparisons (Magurran 2004).

Species sampling and Rarefaction

Rarefaction allows the calculation of **species richness** for a given number of individual samples, based on the construction of so-called **rarefaction curves**. This curve is a plot of the number of species as a function of the number of samples



Wooley et al (2010)

Alpha diversity

a measure of the diversity within a single sample

Types of alpha diversity

Total # of species = **richness**

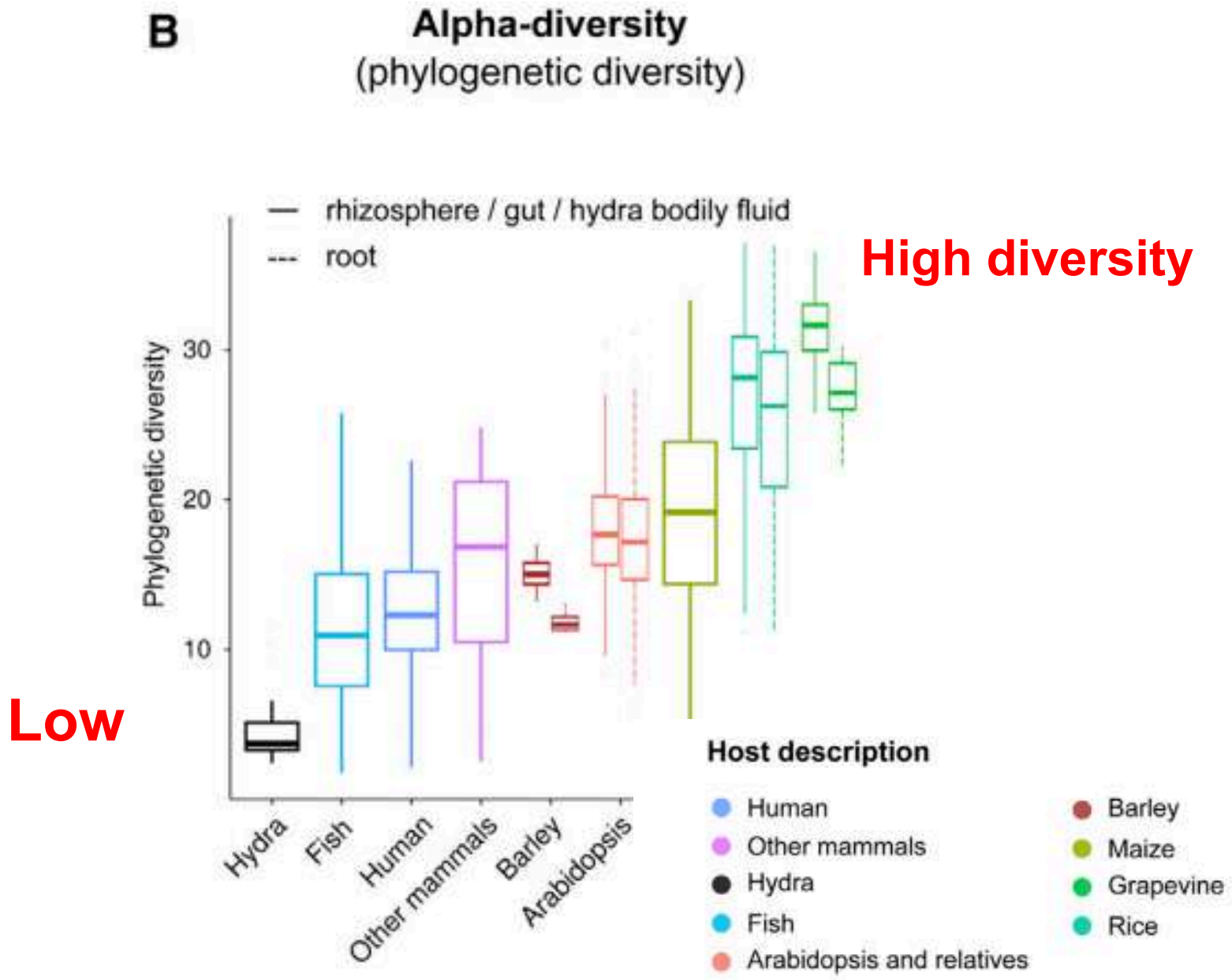
How many OTUs?

Total # of genes = genetic richness

Phylogenetic diversity of genes = genetic PD

Eveness = What is the distribution of abundance in the community?

How many OTUs at high abundance and how many OTU at low abundance?



High diversity

Low

Hacquard et al (2015)

Beta diversity

a measure of **the similarity in diversity between samples**

Types of beta diversity

- Species presence/absence
- Shared phylogenetic diversity
- Gene presence / absence
- Shared phylogenetic diversity of genes

Frequently used as values for PCA of PCoA analysis

Beta diversity

A. Membership:
shared OTU occurences across communities
1 = present, 0 = below detection

List of observed OTUs	Occurences in community **A**	Occurences in community **B**	Shared occurences **A & B**
OTU 1	1	0	
OTU 2	0	1	
OTU 3	1	1	X
OTU 4	1	1	X
OTU 5	1	1	X

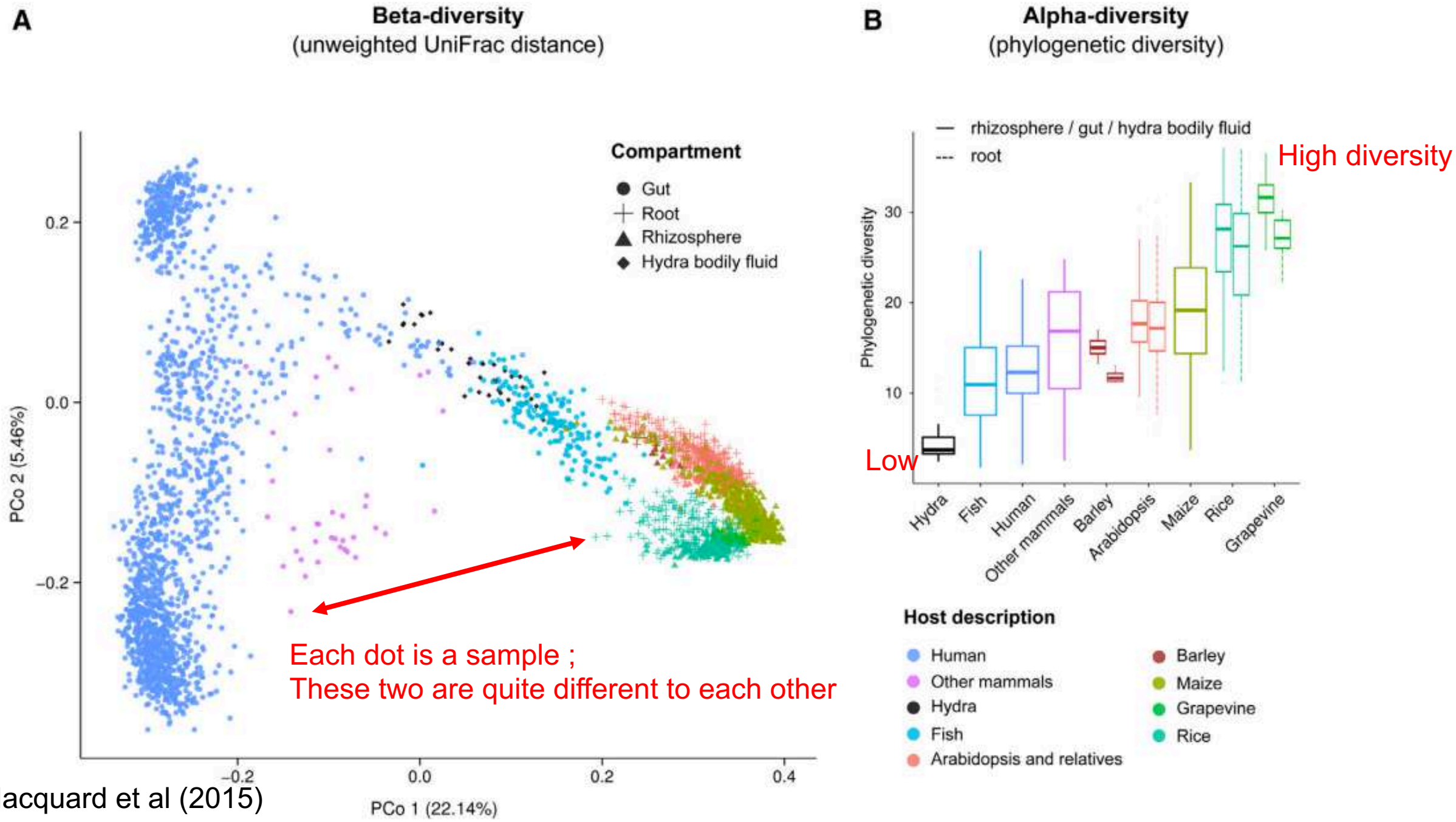
B. Composition:
similar OTU abundances across communities

List of observed OTUs	Abundances community **A**	Abundances community **B**	Similar abundances **A & B**
OTU 1	0.4	0	
OTU 2	0	0.1	
OTU 3	0.1	0.1	X
OTU 4	0.2	0.5	
OTU 5	0.3	0.3	X

Phylogeny:
shared OTU lineages across communities

	Abundances community **A**	Abundances community **B**	Similar abundances **A & B**
OTU 1	0.4	0	⌋ X
OTU 2	0	0.1	
OTU 3	0.1	0.1	⌋ X
OTU 4	0	0.8	
OTU 5	0.5	0	

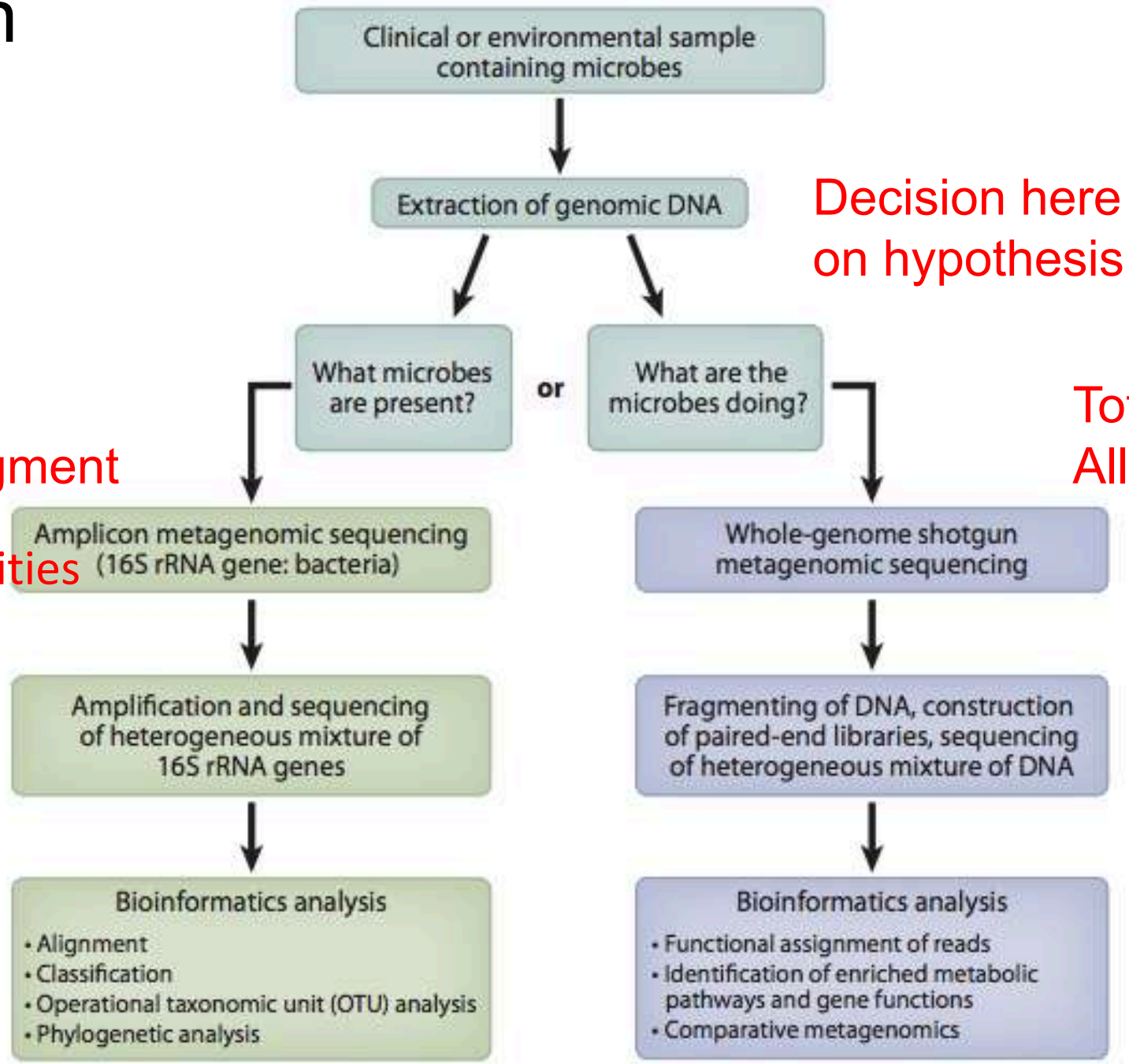
Shade and Handelsman (2012)



Hacquard et al (2015)

Amplicon sequencing or metagenomes?

Workflow decision



Decision here based on hypothesis and budget

PCR product
One gene fragment

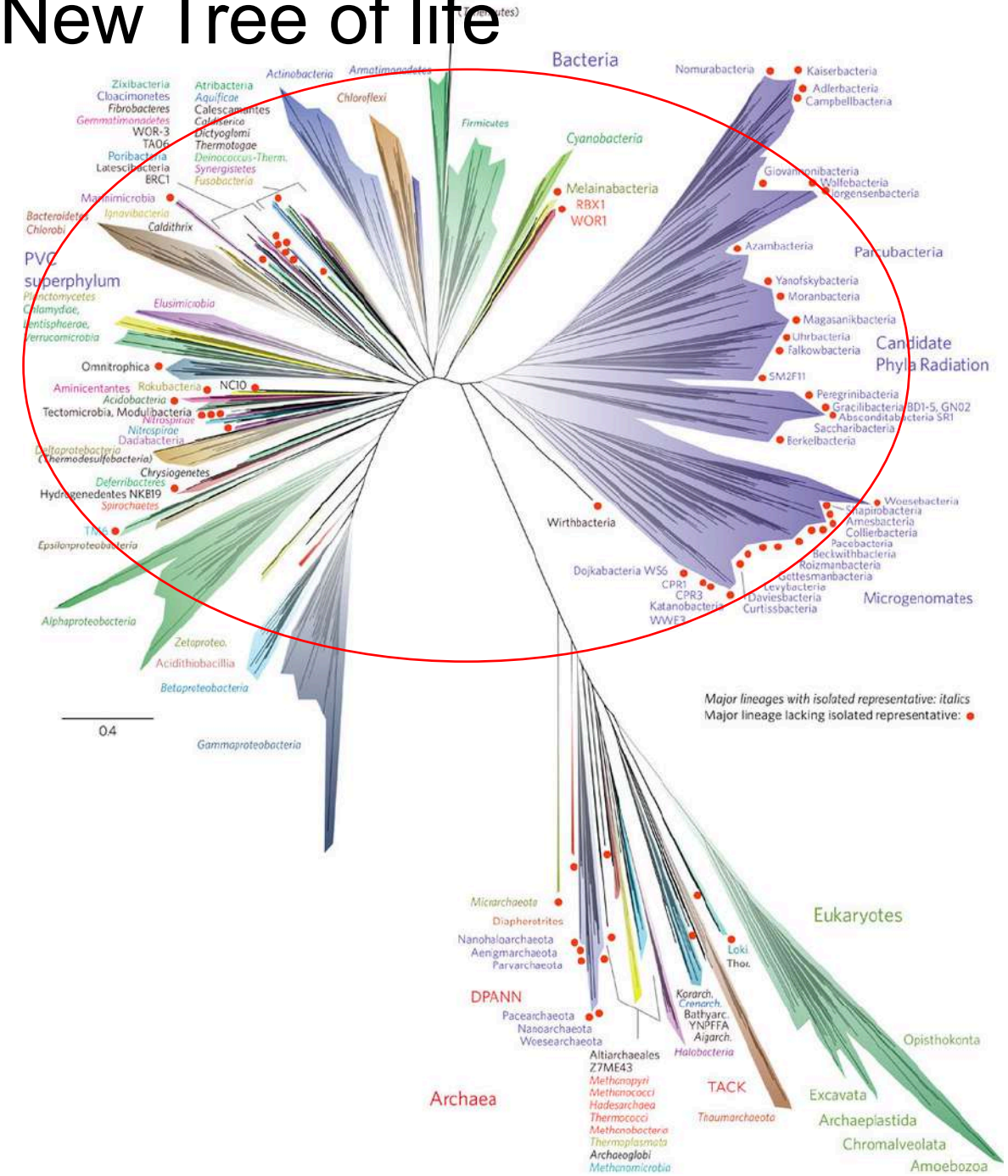
Total DNA
All genes

Fast survey of large communities
Good for first round survey

Grice and Segre (2012)

Amplicon sequencing

New Tree of life

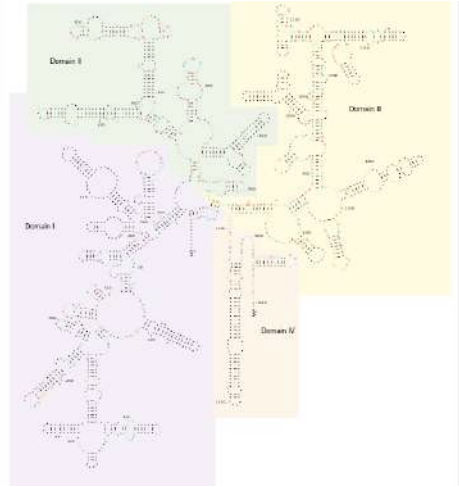


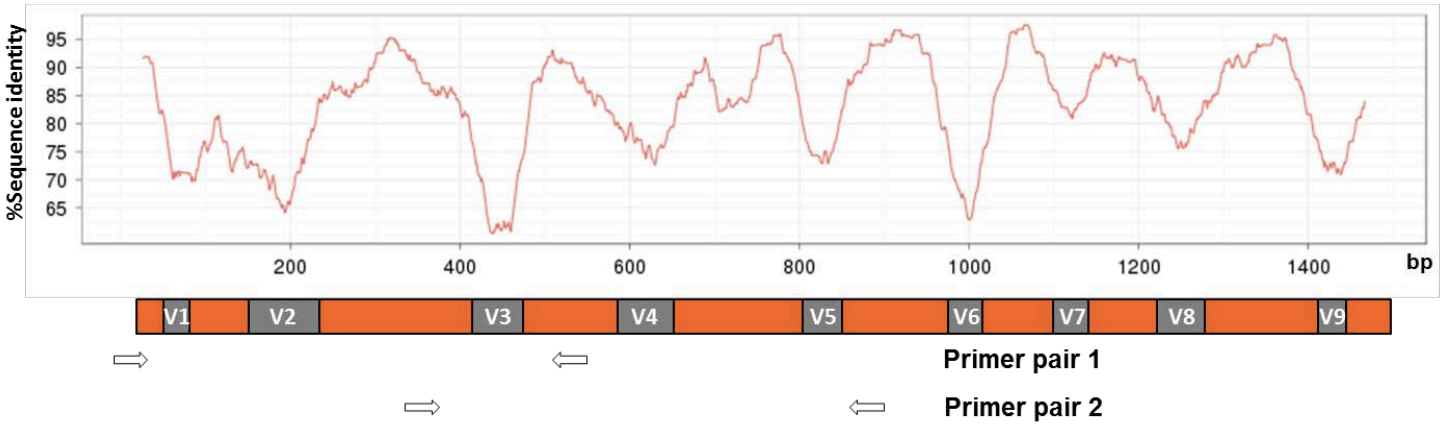
What do they have in common?

Hug et al (2016)

http://www.nytimes.com/2016/04/12/science/scientists-unveil-new-tree-of-life.html?_r=0

16S





- **Advantages**:
  - Universal: Every bacterial and archea species has this gene
  - Conserved regions (for primer design)
  - **Variable regions (to distinguish different species)**
  - Great databases and alignments (for human related species)
  - Mainly used for taxonomical classification

- Problems:
  - Variable copy number in each species
  - No universal (unbiased) primers
  - (Not directly correlated with activity)
  - (Lack of functional information)

http://en.wikipedia.org/wiki/16S_ribosomal_RNA

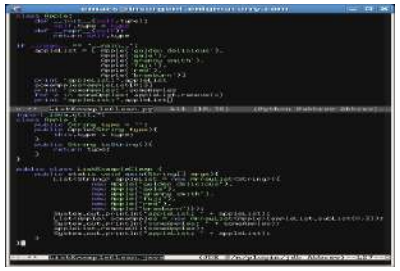
Typical workflow

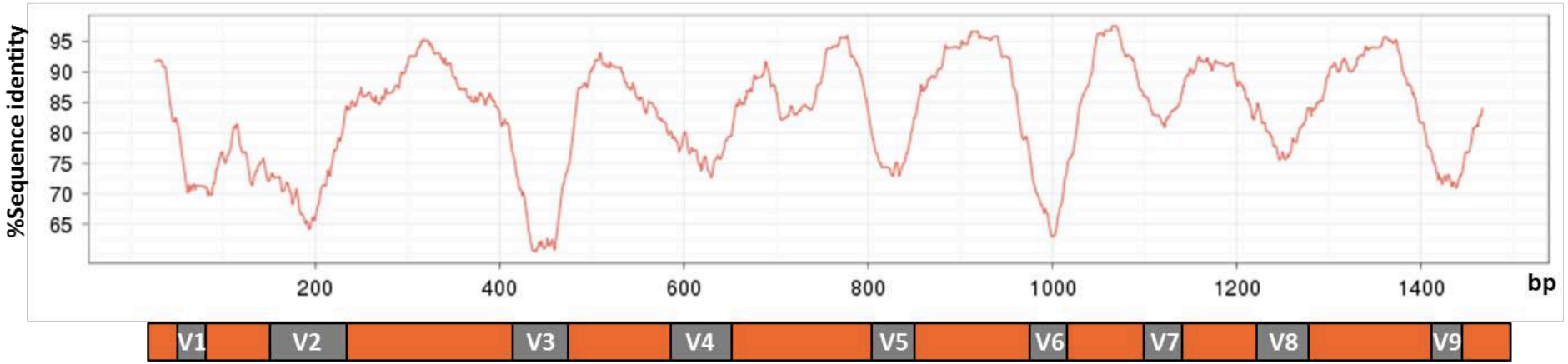
Sampling

Extraction

Sequencing

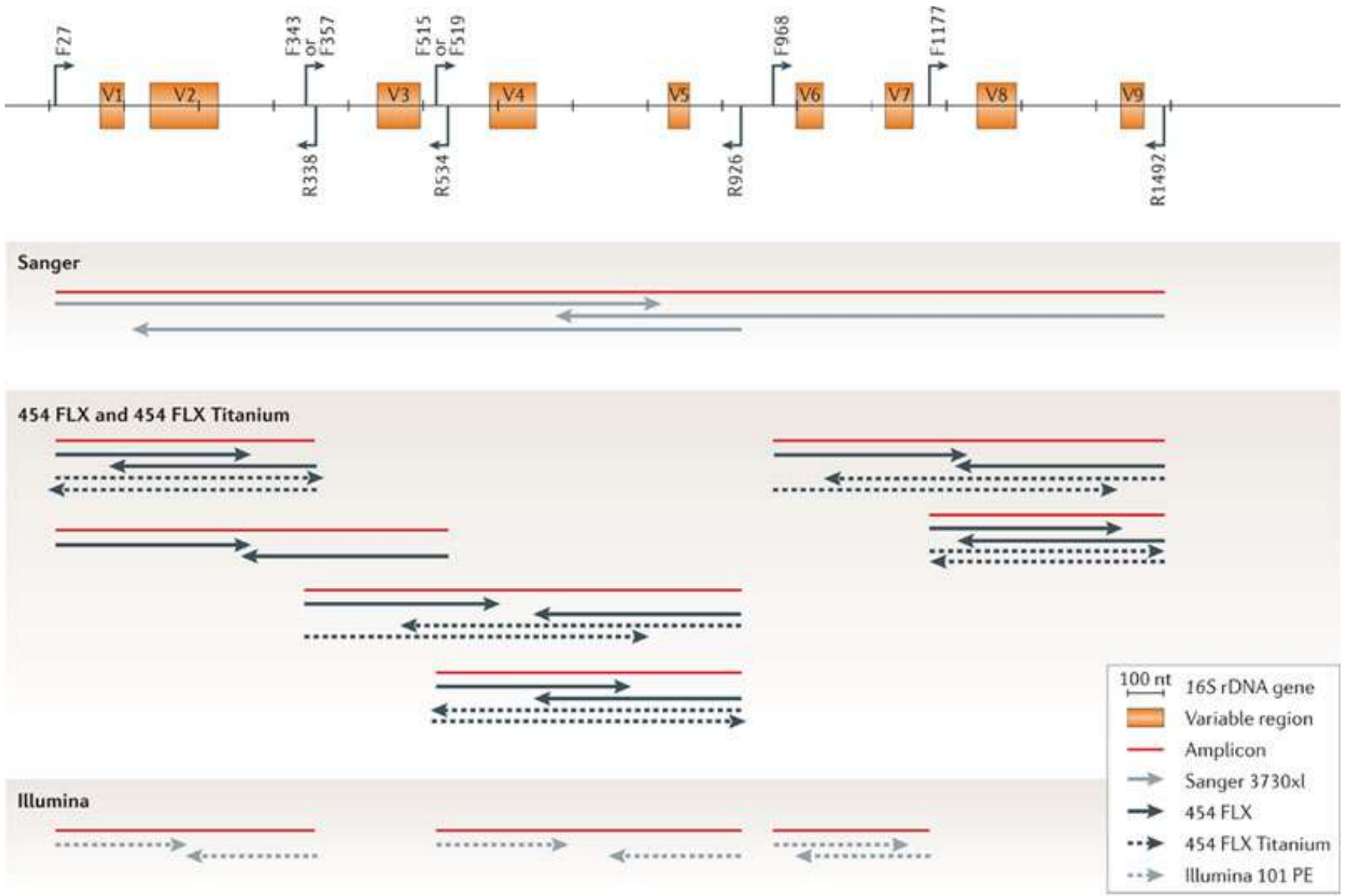
Analysis





Which region to sequence?

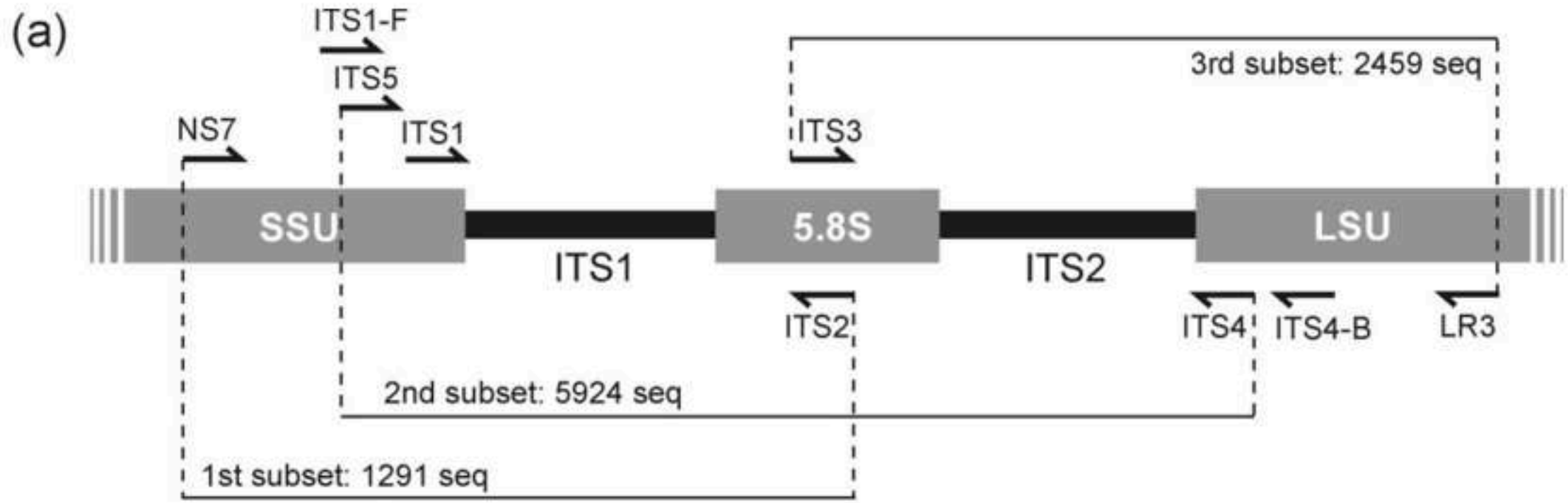
16S amplified region



Kuczynski et al (2011)

Nature Reviews | Genetics

ITS for characterization of fungi species



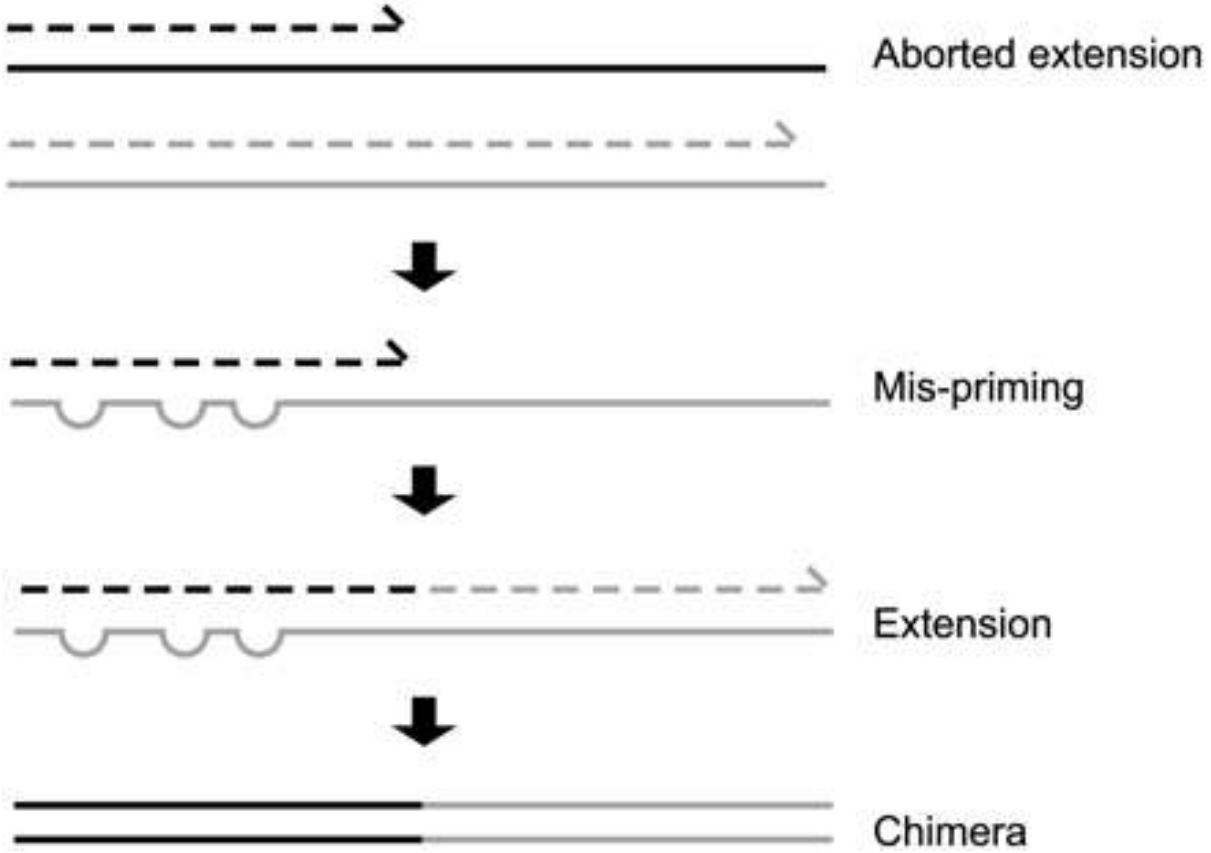
Bellemain et al (2010)

Potential problem

- Lack of tools for processing ITS/Fungal microbiome data sets
- Amplification bias effects accuracy and replication
- Use of short reads prevents disambiguation of similar strains
- 16S or ITS may not differentiate between similar strains –
  - Clustering is done at 97%
  - Regions may be >99% similar
- Sequencing error inflates number of OTUs
- Chloroplast 16S sequences can get amplified in plant metagenomes

Slide of Surya Saha

Chimeric 16S (Artificial sequences formed during PCR amplification



"Chimeras were found to reproducibly form among independent amplifications and contributed to false perceptions of sample diversity and the false identification of novel taxa, **with less-abundant species exhibiting chimera rates exceeding 70%**"

Haas et al (2011)

Metagenomics

Advantage of metagenomics approach

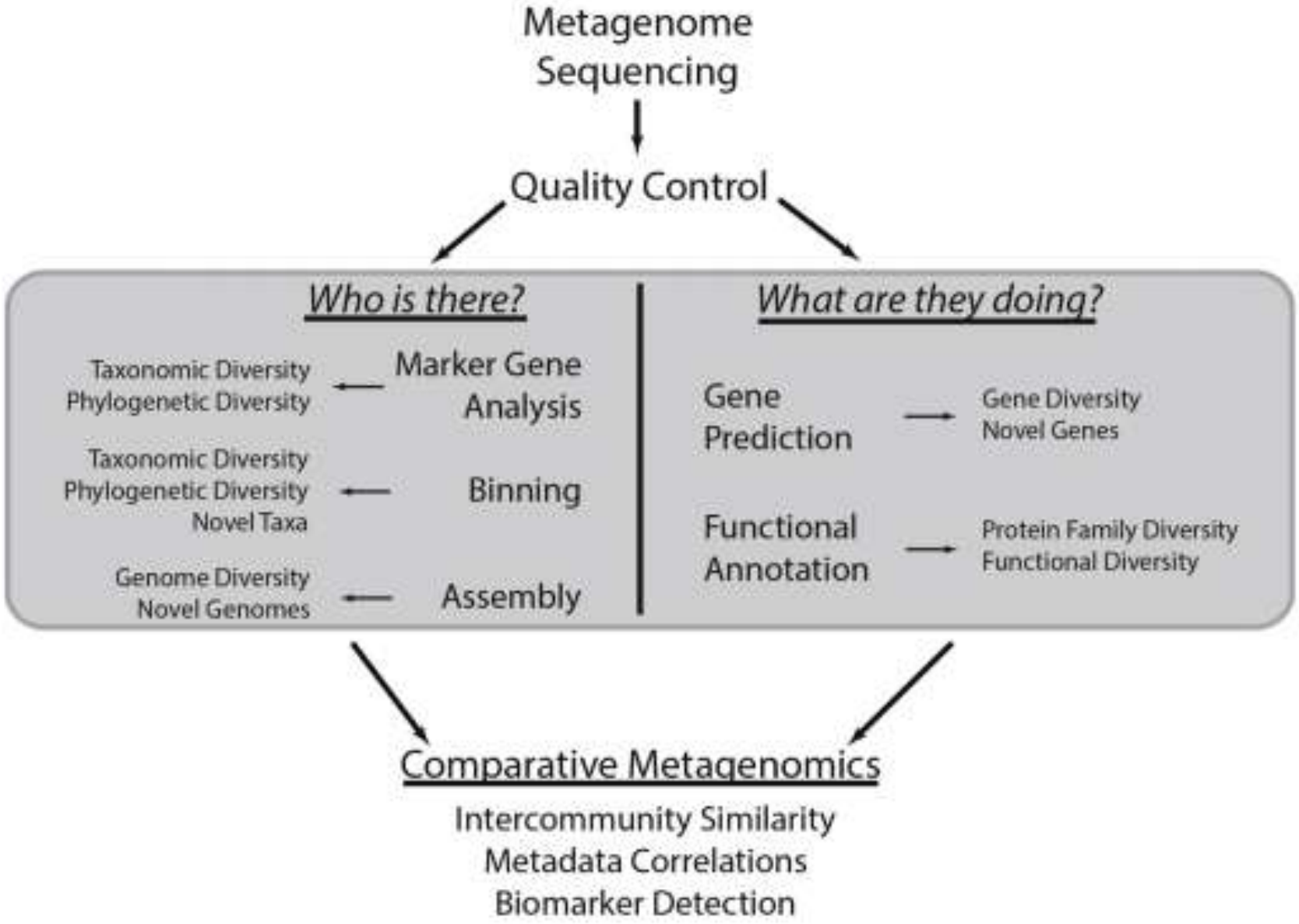
Better classification with Increasing number of complete genomes
Focus on whole genome based phylogeny (whole genome phylotyping)

•Advantages
No amplification bias like in 16S/ITS

Issues
Poor sampling beyond eukaryotic diversity
Assembly of metagenomes is **challenging** due to uneven coverage
Requires **high** depth of coverage

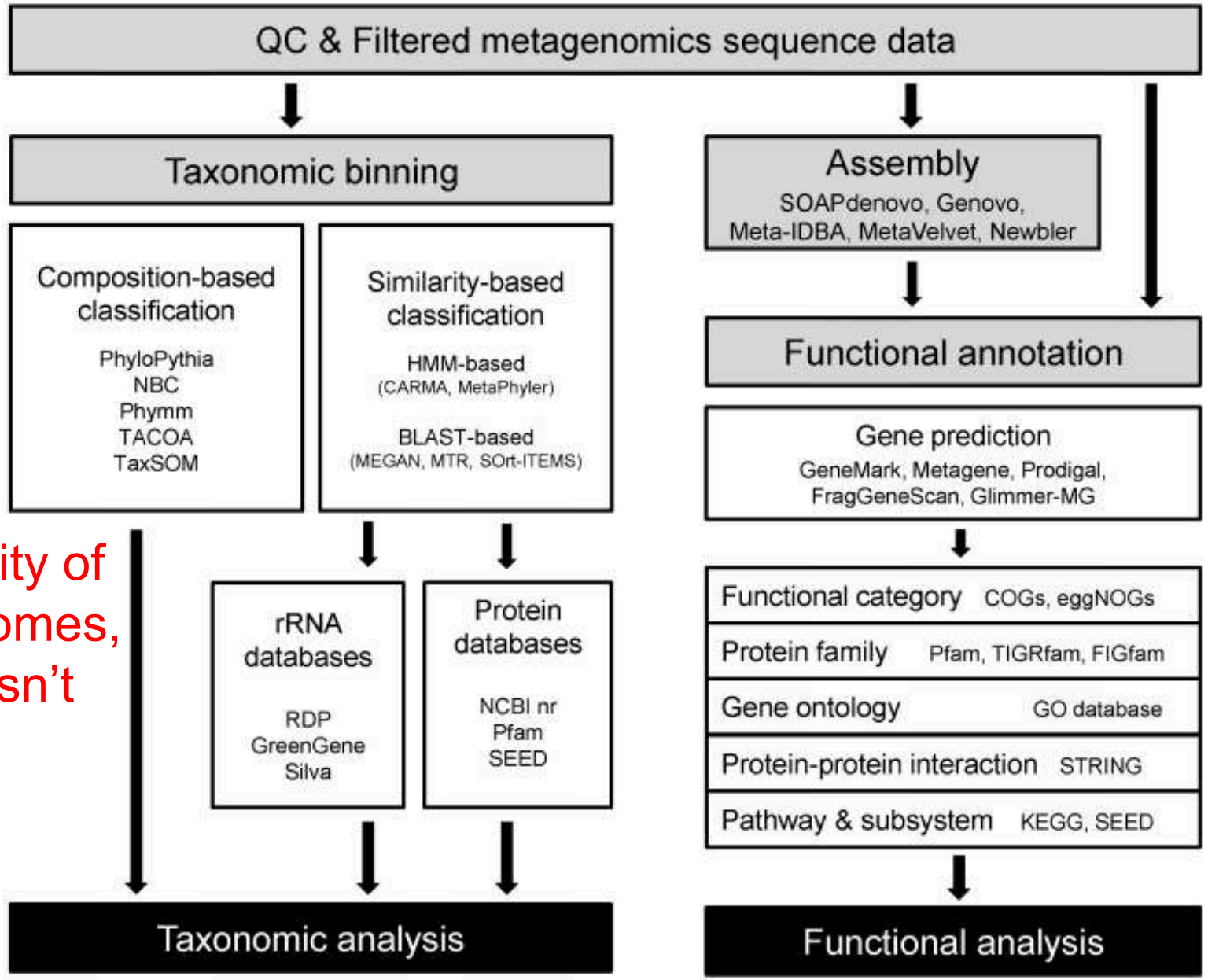
Slide of Surya Saha (slideshare)

Overall workflow



Sharpton (2014)

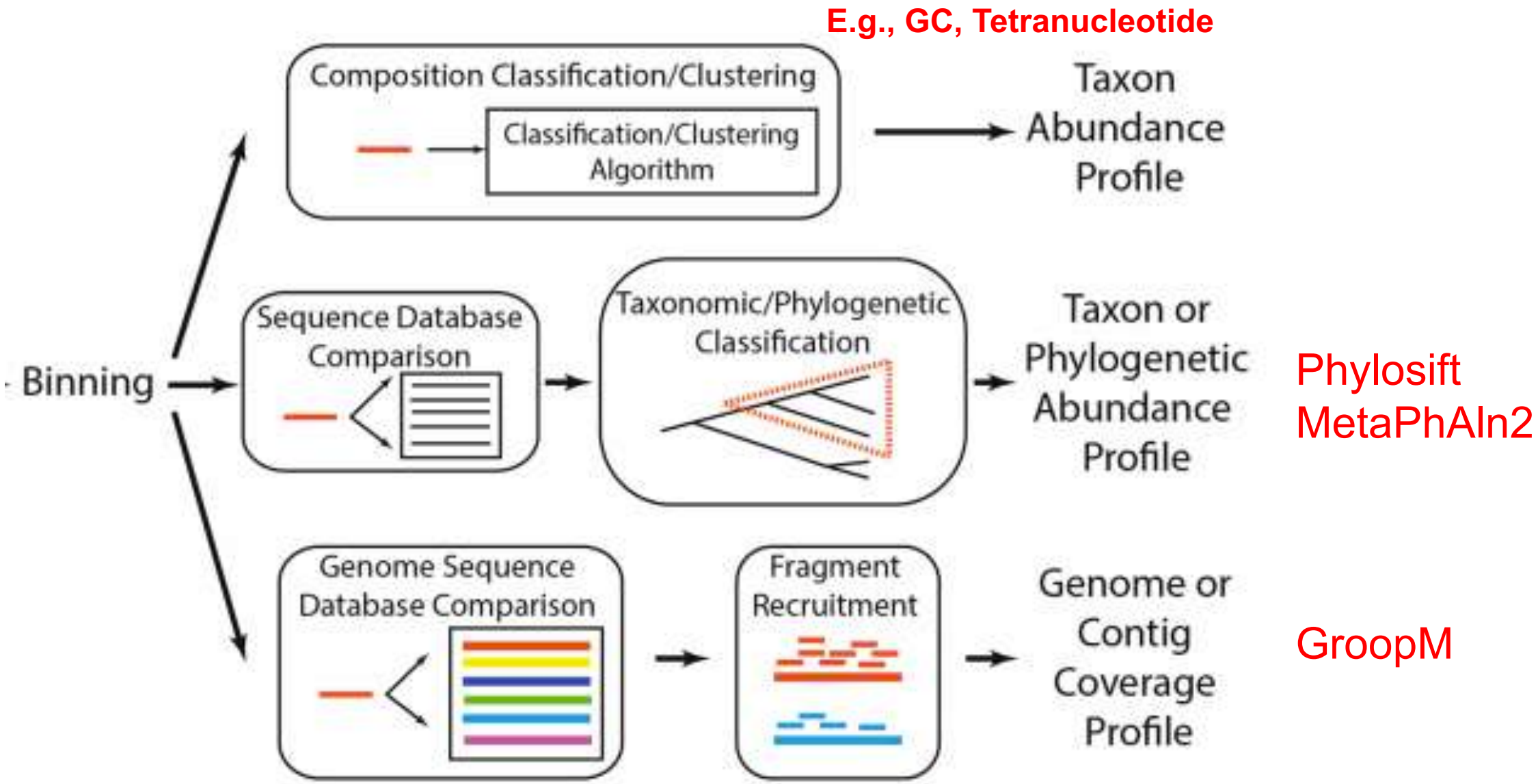
Overall workflow



With the increase availability of reference sequenced genomes, probably one day one doesn't require assembly of metagenomes

Kim et al (2013)

Binning methods



Sharpton (2014)

Binning methods: A combination of

Classification based on **sequence composition:**

Advantage : all reads can be categorised into bins

Disadvantage: no taxonomy / function of the bins.

Classification based on **sequence similarity (of known genes)**

Advantage: One can determine taxonomy and function of reads.

Disadvantage: reads with similarity can not be classified .

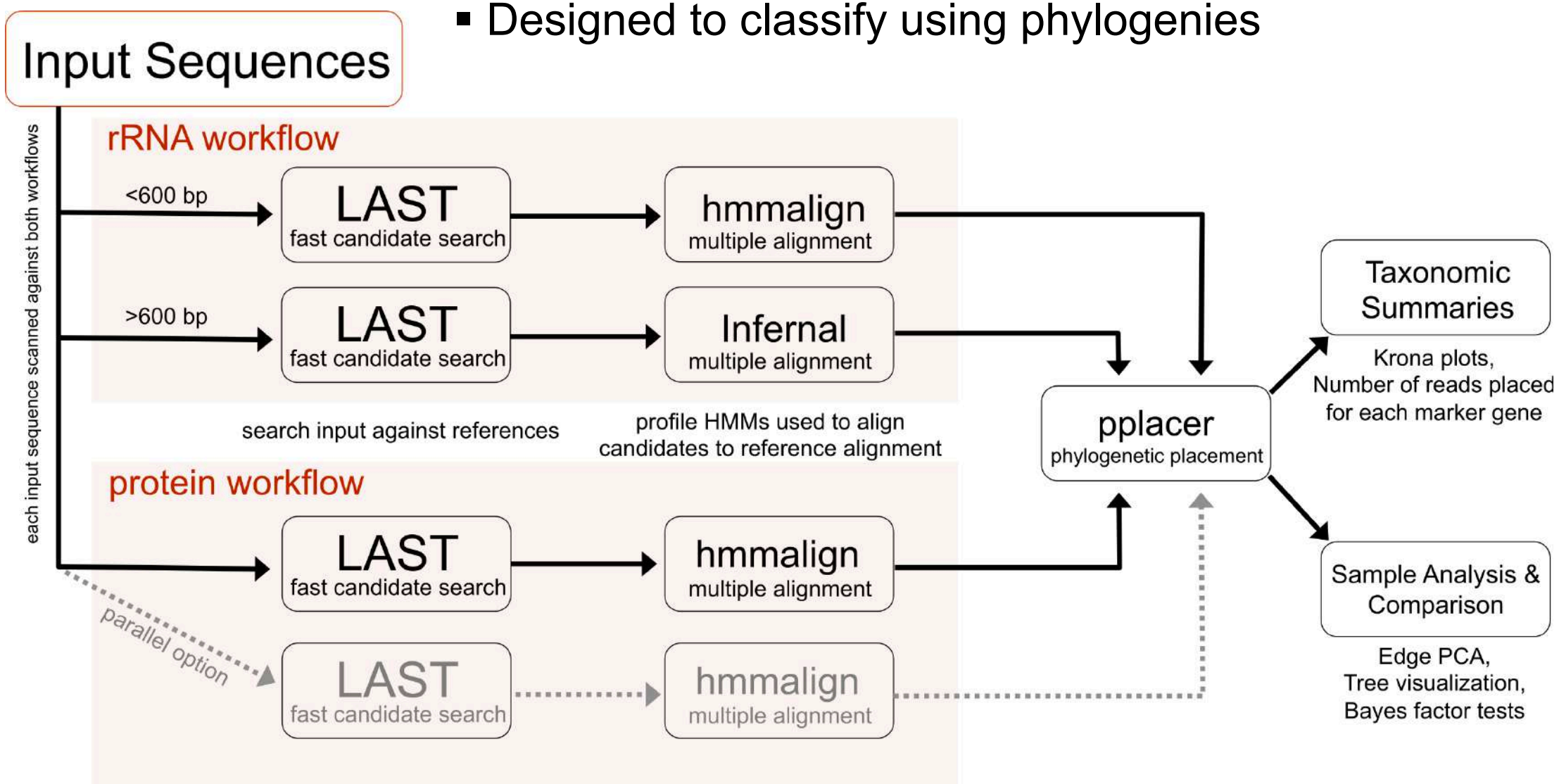
Sharpton (2014)

PhyloSift

mining the global metagenome

https://phylosift.wordpress.com/

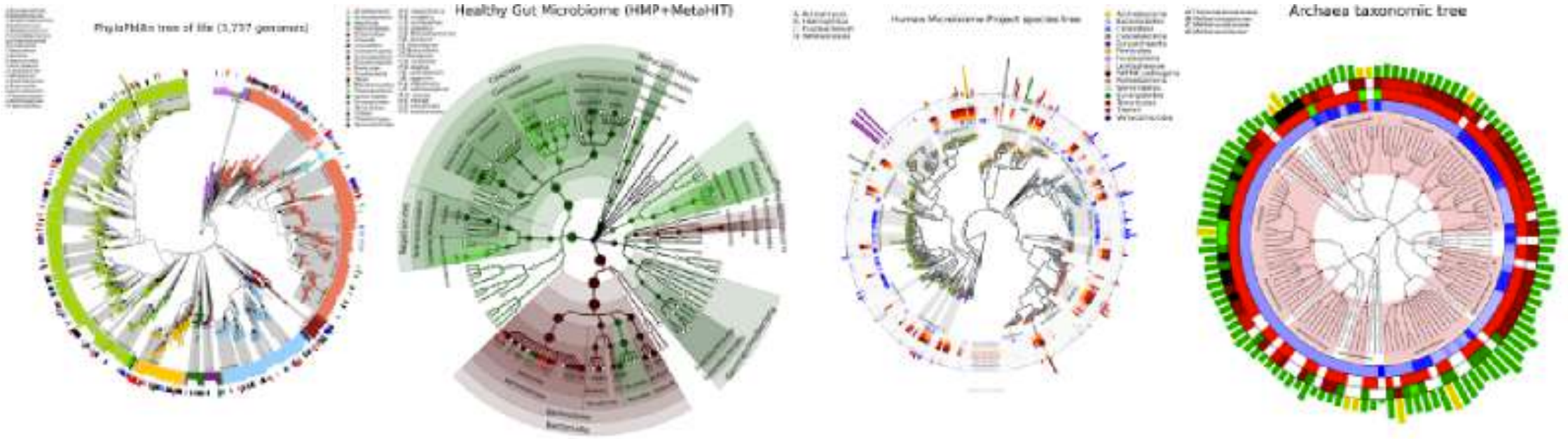
- Uses a database of 37 universal proteins & rRNA genes.
- Designed to classify using phylogenies



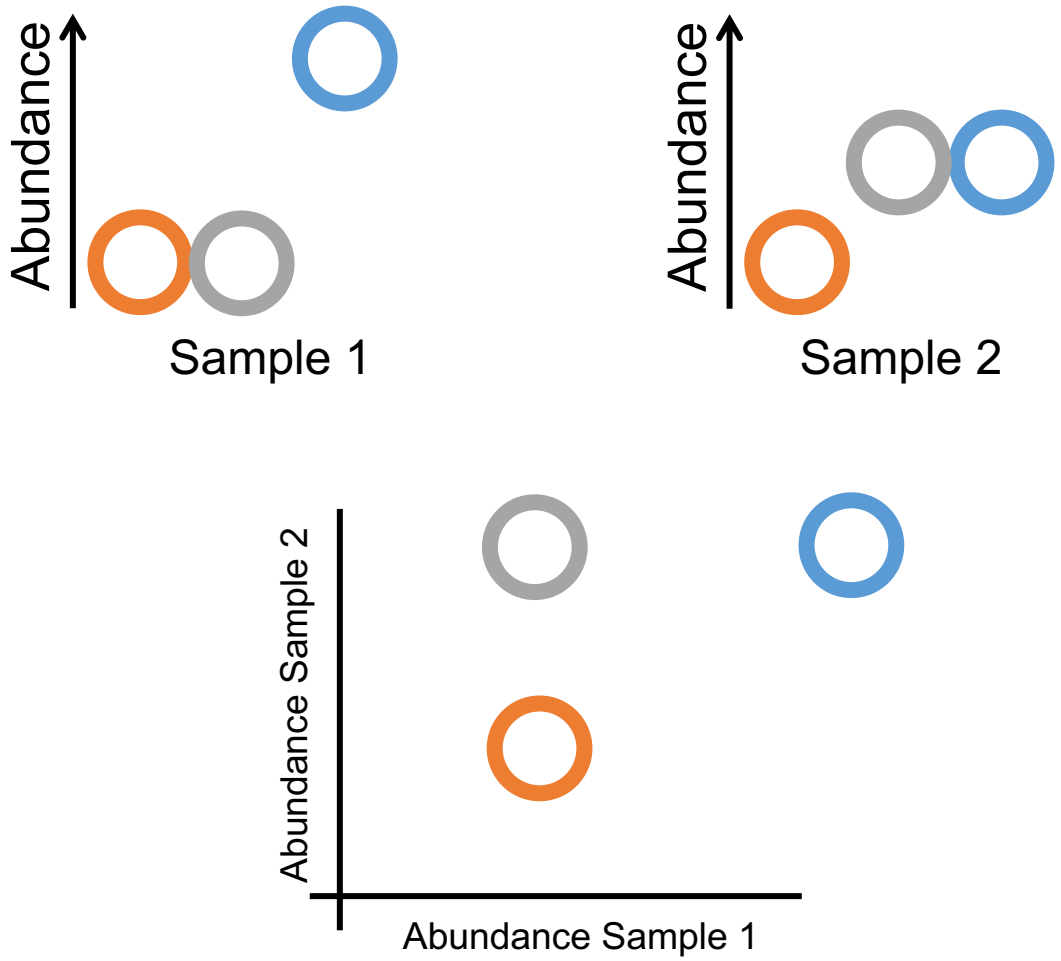
MetaPhAln2 – enhanced metagenomic taxonomic profiling

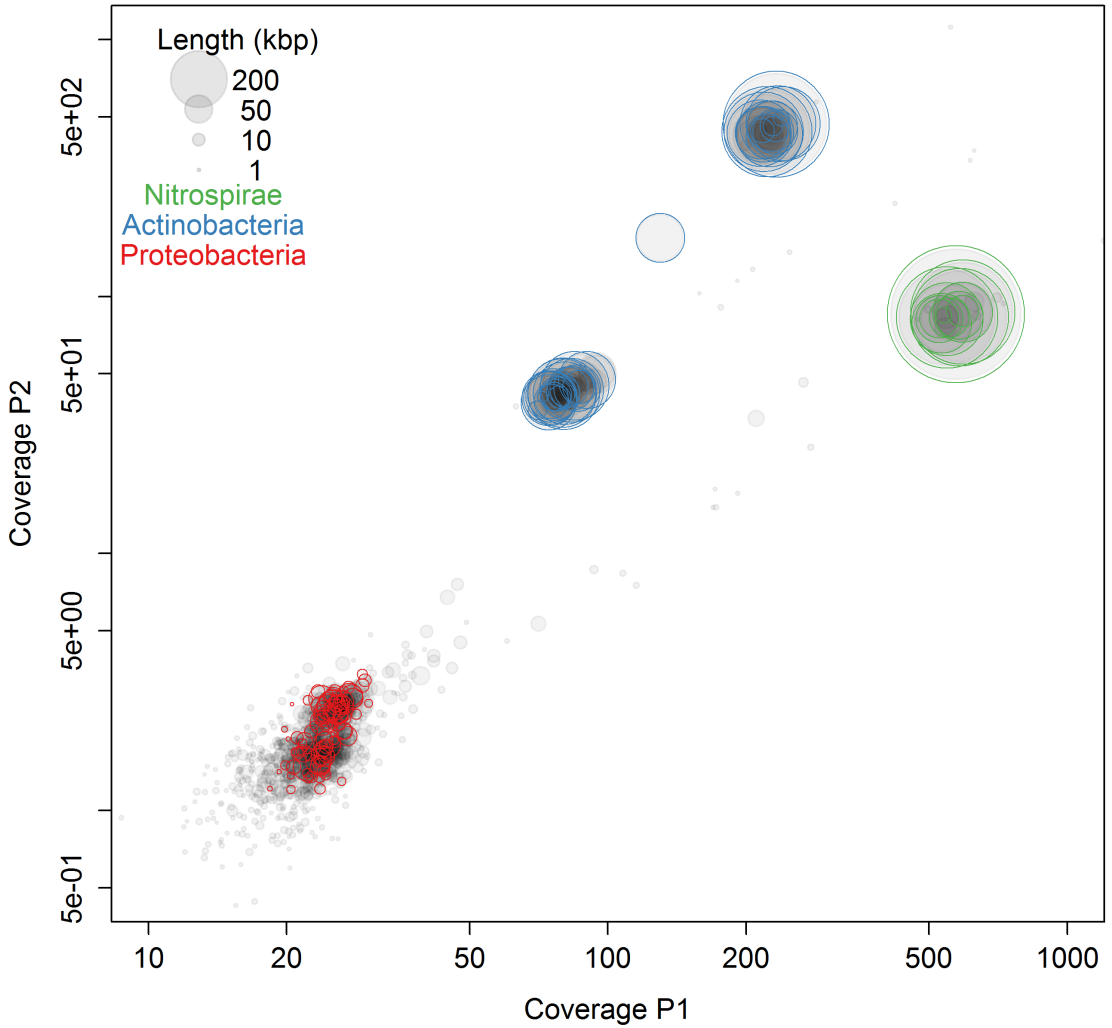
relies **on ~1M unique clade-specific marker genes** identified from ~17,000 reference genomes (~13,500 bacterial and archaeal, ~3,500 viral, and ~110 eukaryotic), allowing:

Species level resolution
Good visualisation with **GraphAln**
(So it's useful with known ecosystems)



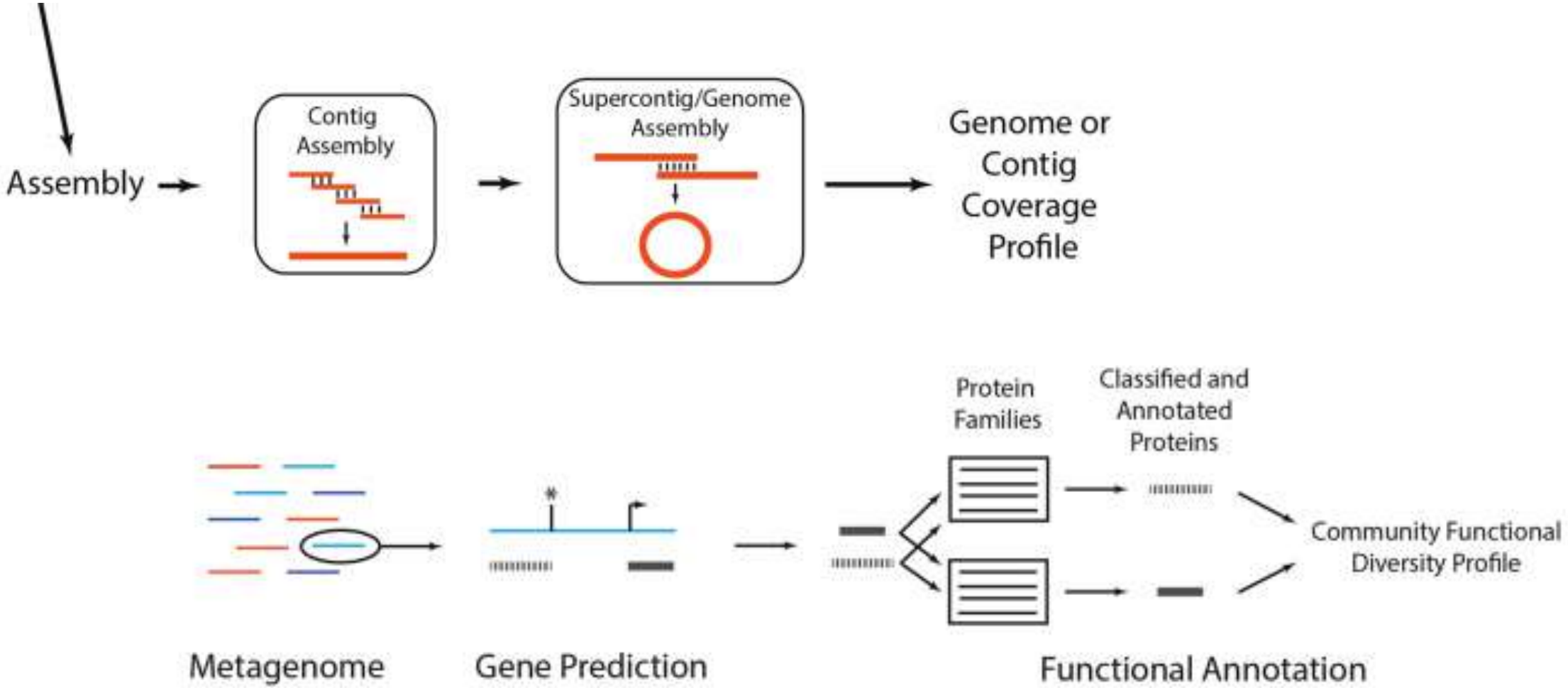
Example of binning based on differential coverage





H. Daims & C. Dorninger, DOME, University of Vienna

Actual assembly



Sharpton (2014)

Algorithm advancements lead to recovery of genomes

nature microbiology

ARTICLES

DOI: 10.1038/s41564-017-0012-7

OPEN

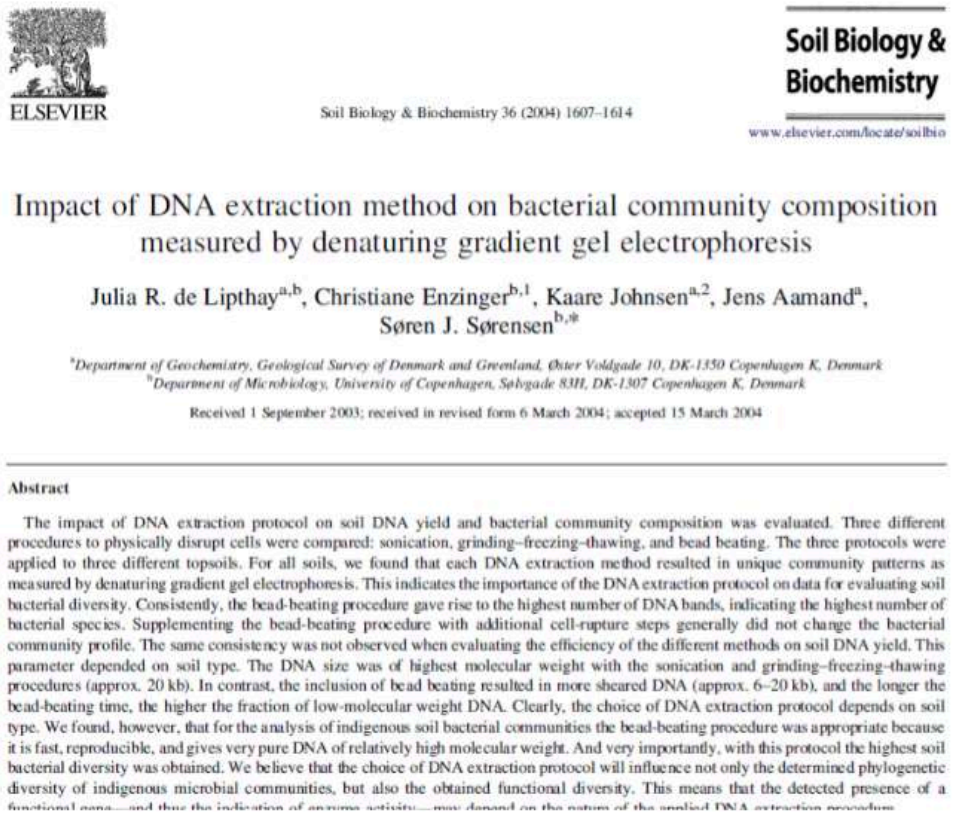
Corrected: Author correction

Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life

Donovan H. Parks, Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J. Woodcroft, Paul N. Evans, Philip Hugenholtz* and Gene W. Tyson*

Biases

Extraction protocol matters



ELSEVIER

Soil Biology & Biochemistry 36 (2004) 1607–1614

Soil Biology & Biochemistry

www.elsevier.com/locate/soilbio

Impact of DNA extraction method on bacterial community composition measured by denaturing gradient gel electrophoresis

Julia R. de Lipthay[a,b], Christiane Enzinger[b,1], Kaare Johnsen[a,2], Jens Aamand[a], Søren J. Sørensen[b,*]

[a]Department of Geochemistry, Geological Survey of Denmark and Greenland, Øster Voldgade 10, DK-1350 Copenhagen K, Denmark
[b]Department of Microbiology, University of Copenhagen, Sølvgade 83H, DK-1307 Copenhagen K, Denmark

Received 1 September 2003; received in revised form 6 March 2004; accepted 15 March 2004

Abstract

The impact of DNA extraction protocol on soil DNA yield and bacterial community composition was evaluated. Three different procedures to physically disrupt cells were compared: sonication, grinding–freezing–thawing, and bead beating. The three protocols were applied to three different topsoils. For all soils, we found that each DNA extraction method resulted in unique community patterns as measured by denaturing gradient gel electrophoresis. This indicates the importance of the DNA extraction protocol on data for evaluating soil bacterial diversity. Consistently, the bead-beating procedure gave rise to the highest number of DNA bands, indicating the highest number of bacterial species. Supplementing the bead-beating procedure with additional cell-rupture steps generally did not change the bacterial community profile. The same consistency was not observed when evaluating the efficiency of the different methods on soil DNA yield. This parameter depended on soil type. The DNA size was of highest molecular weight with the sonication and grinding–freezing–thawing procedures (approx. 20 kb). In contrast, the inclusion of bead beating resulted in more sheared DNA (approx. 6–20 kb), and the longer the bead-beating time, the higher the fraction of low-molecular weight DNA. Clearly, the choice of DNA extraction protocol depends on soil type. We found, however, that for the analysis of indigenous soil bacterial communities the bead-beating procedure was appropriate because it is fast, reproducible, and gives very pure DNA of relatively high molecular weight. And very importantly, with this protocol the highest soil bacterial diversity was obtained. We believe that the choice of DNA extraction protocol will influence not only the determined phylogenetic diversity of indigenous microbial communities, but also the obtained functional diversity. This means that the detected presence of a functional gene—and thus the indication of enzyme activity—may depend on the nature of the applied DNA extraction procedure.

Wesolowska-Andersen et al. Microbiome 2014, 2:19
http://www.microbiomejournal.com/content/2/1/19

Microbiome

RESEARCH Open Access

Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis

Agata Wesolowska-Andersen[1], Martin Iain Bahl[2], Vera Carvalho[2], Karsten Kristiansen[3], Thomas Sicheritz-Pontén[1], Ramneek Gupta[1*] and Tine Rask Licht[2*]

Abstract

Background: In recent years, studies on the human intestinal microbiota have attracted tremendous attention. Application of next generation sequencing for mapping of bacterial phylogeny and function has opened new doors to this field of research. However, little attention has been given to the effects of choice of methodology on the output resulting from such studies.

Results: In this study we conducted a systematic comparison of the DNA extraction methods used by the two major collaborative efforts: The European MetaHIT and the American Human Microbiome Project (HMP). Additionally, effects of homogenizing the samples before extraction were addressed. We observed significant differences in distribution of bacterial taxa depending on the method. While eukaryotic DNA was most efficiently extracted by the MetaHIT protocol, DNA from bacteria within the Bacteroidetes phylum was most efficiently extracted by the HMP protocol.

Conclusions: Whereas it is comforting that the inter-individual variation clearly exceeded the variation resulting from choice of extraction method, our data highlight the challenge of comparing data across studies applying different methodologies.

"we found that each DNA extraction method resulted in unique community patterns"

"We observed significant differences in distribution of bacterial taxa depending on the method."

Alpha diversity is always overestimated

Table 1. Effect of quality filtering and clustering on diversity estimates (OTU number), error rate and data loss of pyrotags amplified from two regions of E. coli MG1655 16S rRNA genes.

	Number of OTUs at percentage identity thresholds							
Read filtering	100	99	98	97	95	90	% errorless reads	% reads used
5′ forward (V1 and V2)								
Theoretical number	5	4	3	1	1	1		
No quality filtering	643	95	31	16	5	3	68.7	77.9
Reads with N's removed	600	85	29	14	4	3	69.8	76.7
Quality score-based filtering (% per-base error probability)								
3	638	92	31	13	3	3	68.9	77.7
2	632	90	30	14	3	3	69.0	77.6
1	609	79	24	9	3	3	69.1	77.3
0.5	562	66	15	7	3	3	70.7	75.3
0.2	469	30	6	**3**	3	3	73.2	70.8
0.1	372	26	5	3	3	3	77.8	57.8
3′ reverse (V8)								
Theoretical number	1	1	1	1	1	1		
No quality filtering	385	43	13	7	5	4	84.6	94.4
Reads with N's removed	361	40	12	6	4	3	85.3	93.6
Quality score-based filtering (% per-base error probability)								
3	378	40	12	7	5	4	84.8	94.2
2	368	32	10	6	5	4	85.1	93.8
1	342	25	9	6	5	4	85.3	93.3
0.5	310	20	8	6	5	4	87.5	89.5
0.2	236	7	2	**2**	2	2	89.6	82.1
0.1	196	4	2	2	2	2	90.7	70.6

Diversity estimates should be considered relative to the theoretical number of OTUs from E. coli.

Kunin et al (2010)

Reagent and laboratory contamination

RESEARCH ARTICLE Open Access

Reagent and laboratory contamination can critically impact sequence-based microbiome analyses

Susannah J Salter[1*], Michael J Cox[2], Elena M Turek[2], Szymon T Calus[3], William O Cookson[2], Miriam F Moffatt[2], Paul Turner[4,5], Julian Parkhill[1], Nicholas J Loman[3] and Alan W Walker[1,6*]

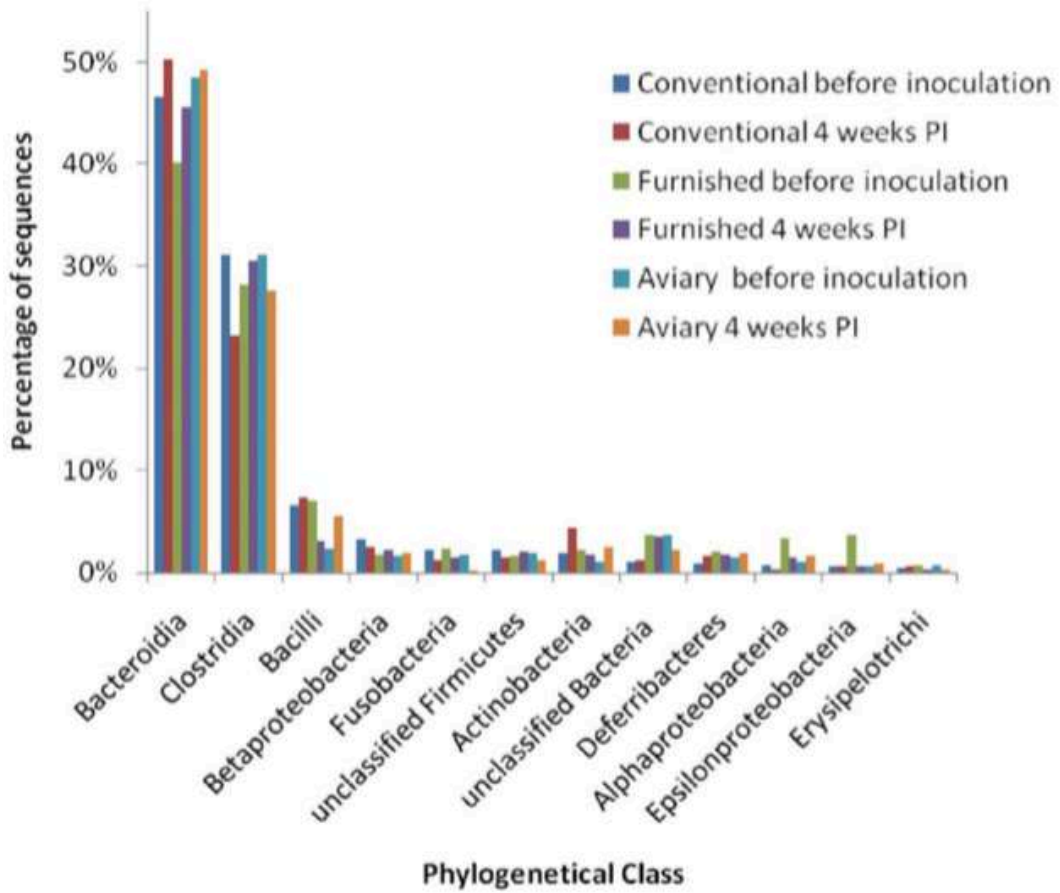
RESEARCH HIGHLIGHT

Tracking down the sources of experimental contamination in microbiome studies

Sophie Weiss[1], Amnon Amir[2], Embriette R Hyde[2], Jessica L Metcalf[2], Se Jin Song[2] and Rob Knight[2,3,4*]

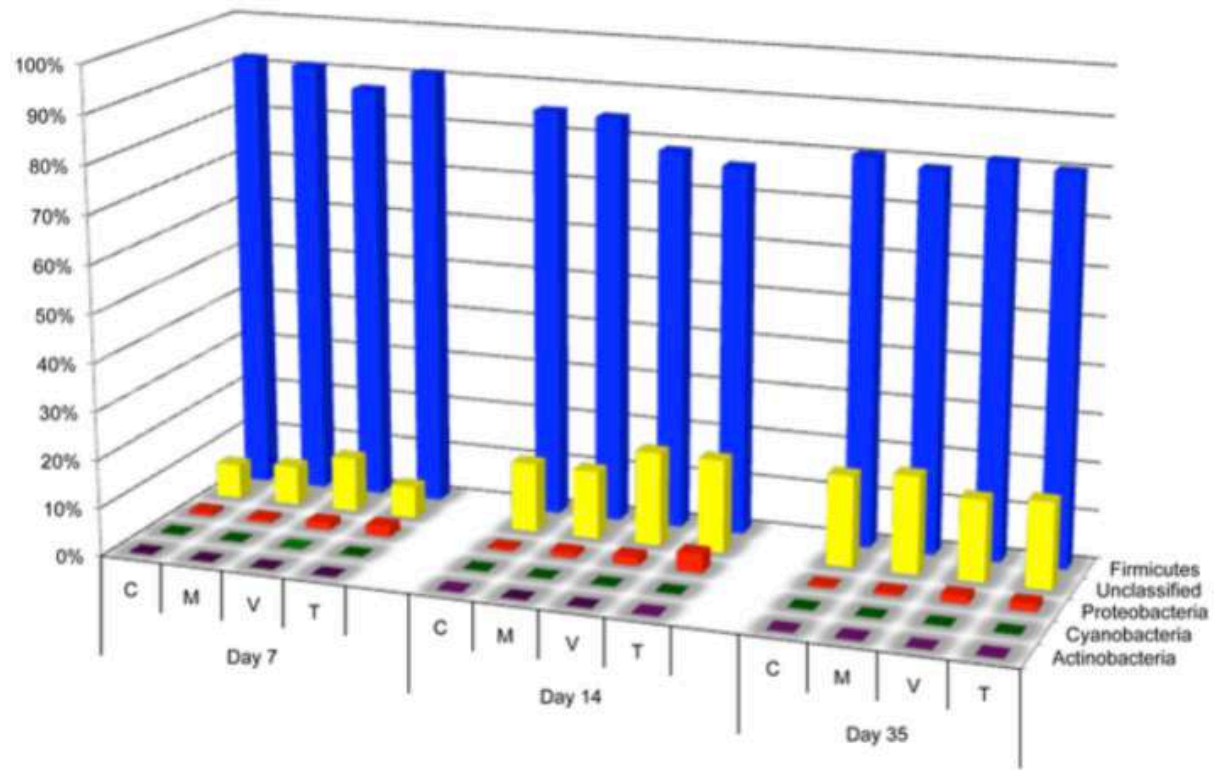
2 papers with different results at the same year

Bacteroidetes >>> rest



Nordentoft S *et al* (2011) BMC Microbiology

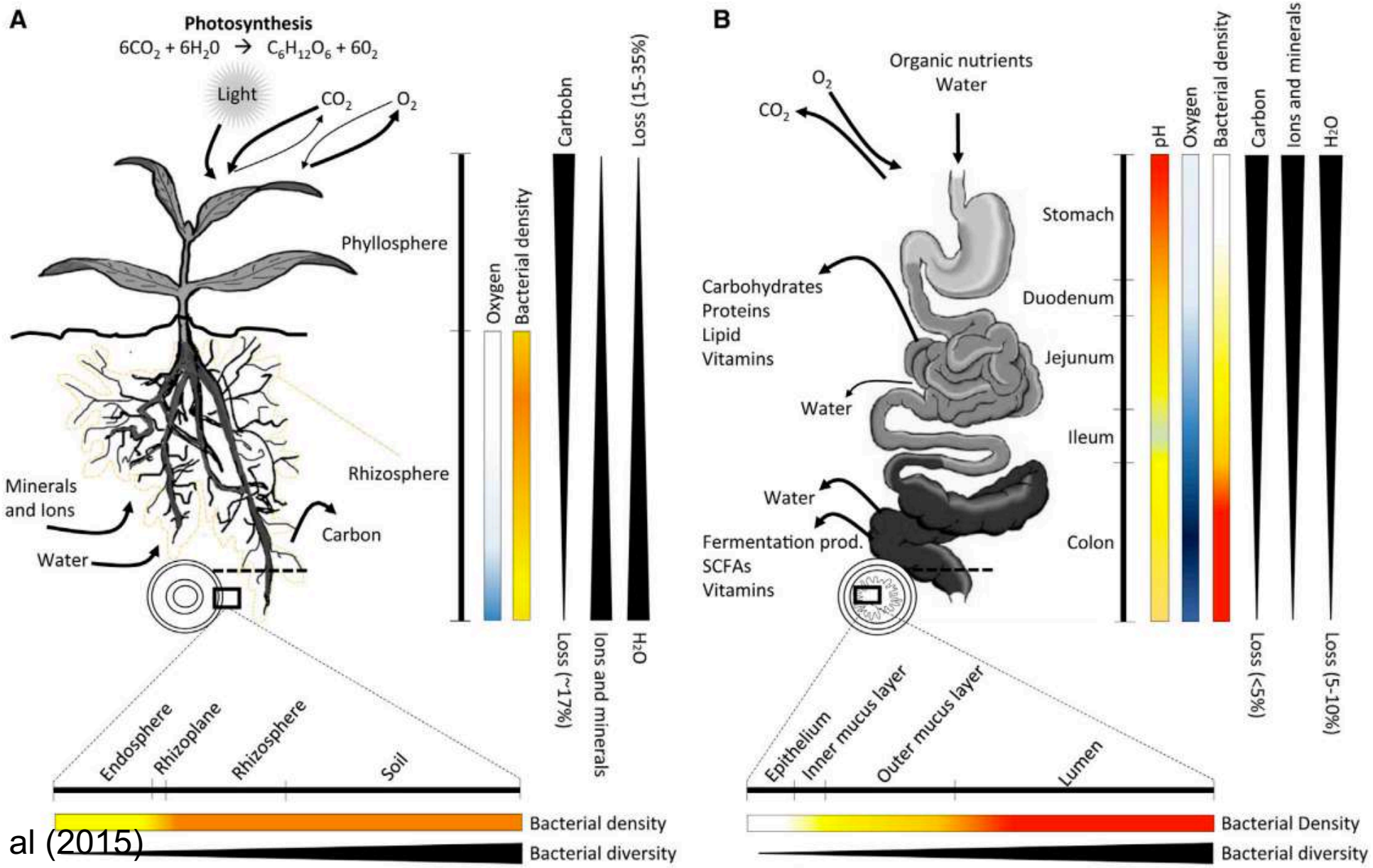
firmicutes >>> rest > bacteroidetes



Danzeisen JL *et al* (2011) PLOS one

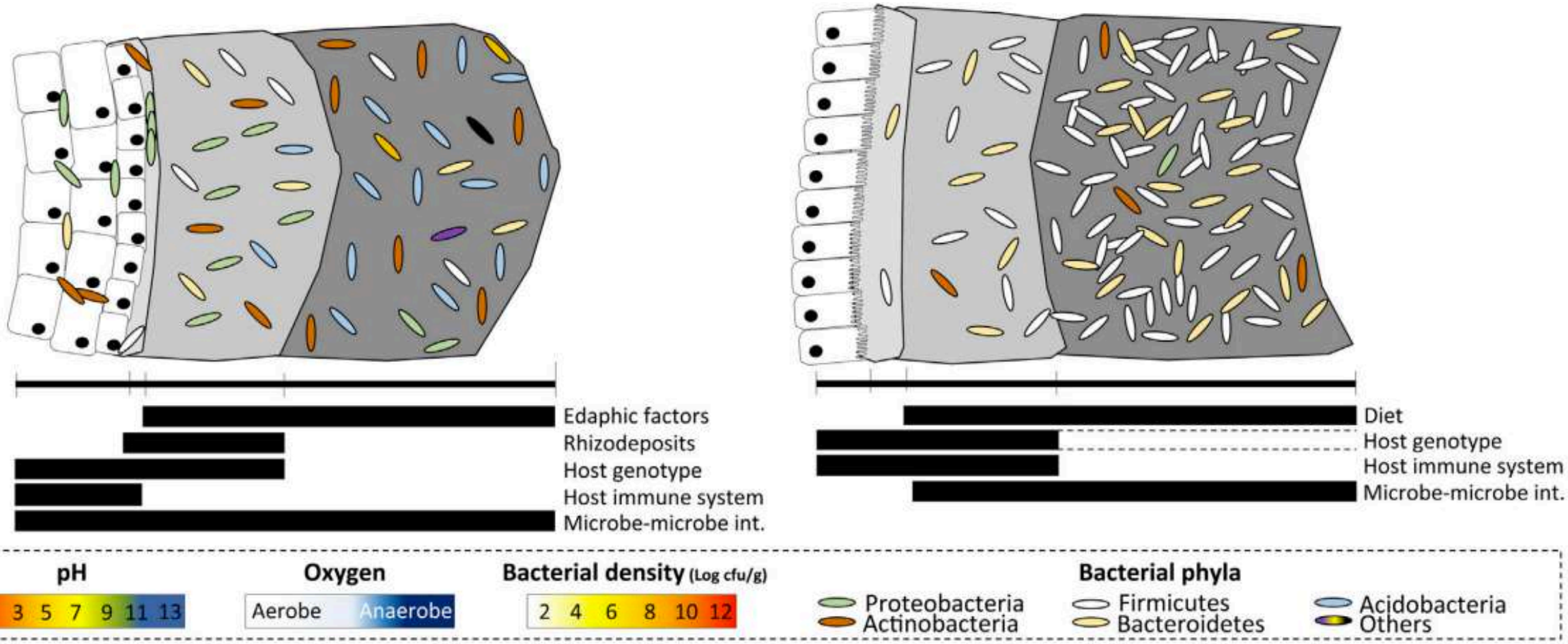
Case studies

Two most common systems



Hacquard et al (2015)

Two most common systems



Hacquard et al (2015)

Two most common systems

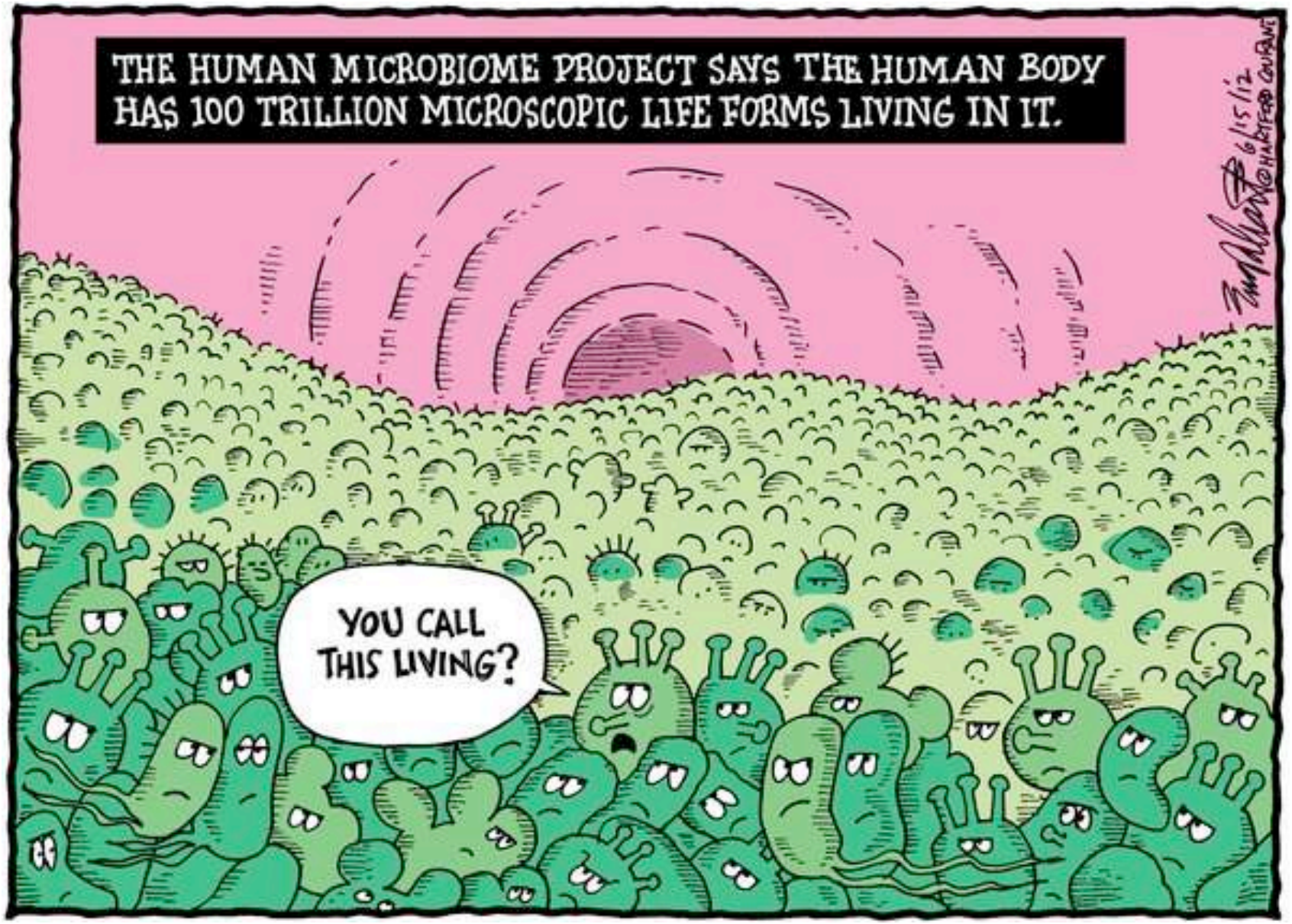
Table 1. Percentage of Shotgun Metagenome Reads Assigned to Each Kingdom of Life across Metagenome Studies

	Cucumber[a]	Wheat[a]	Soybean[b]	Wheat[c]	Oat[c]	Pea[c]	Barley[d]	Gut[e]
Bacteria	99.36	99.45	96	88.5	77.3	73.7	94.04	99.1
Archaea	0.02	0.02	<1	<0.5	<0.5	<0.5	0.054	
Eukaryotes	0.54	0.48	3	3.3	16.6	20.7	5.90	<0.1

[a]Ofek-Lalzar et al. (2014) (metagenomics of rhizoplane samples).
[b]Mendes et al. (2014) (metagenomics of rhizosphere samples).
[c]Turner et al. (2013) (metatranscriptomics of rhizosphere samples).
[d]Bulgarelli et al. (2015) (metagenomics of rhizosphere samples).
[e]Qin et al. (2010) (metagenomics of gut samples).

Hacquard et al (2015)

Human gut microbiome



Human gut microbiome

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

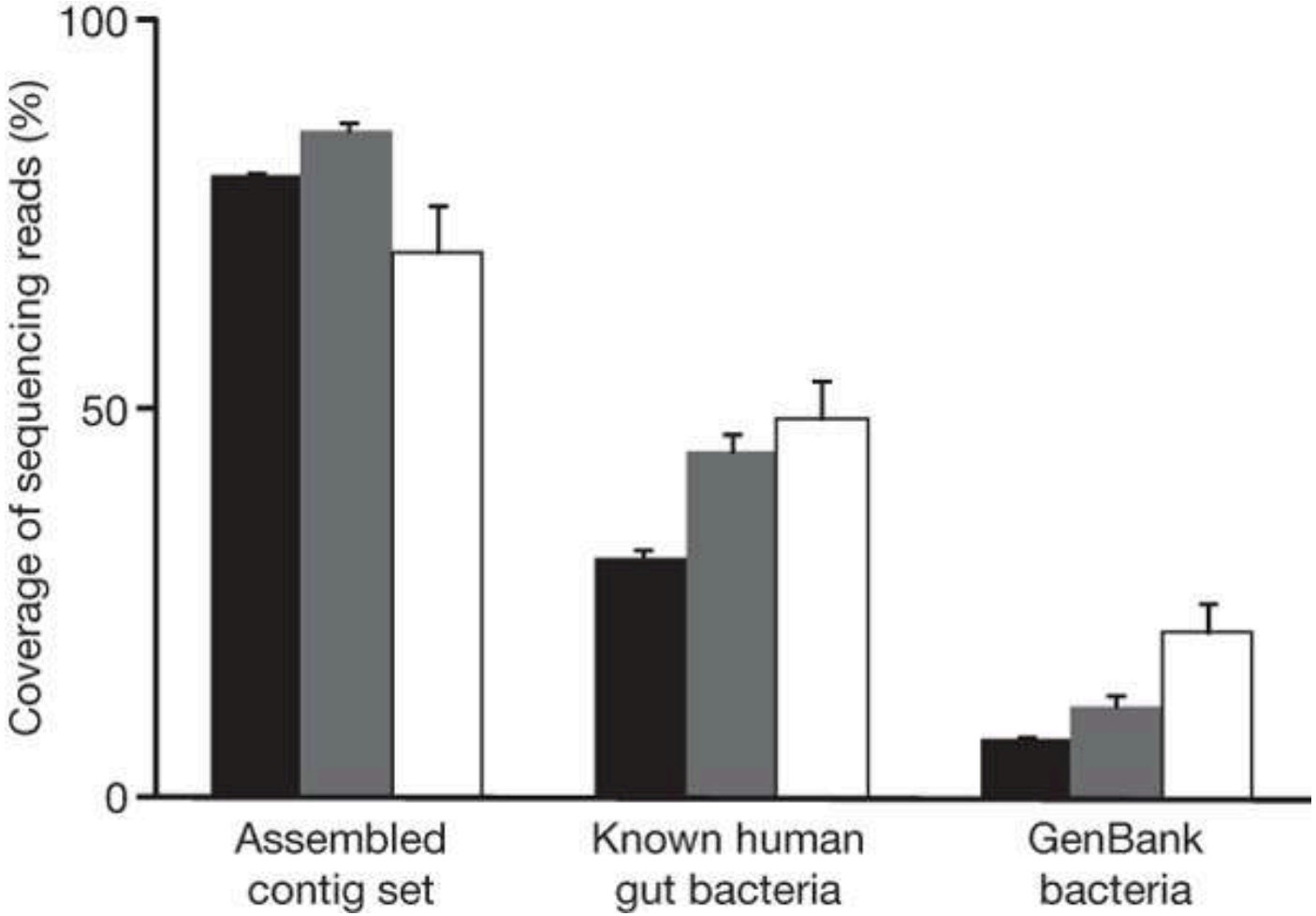
ARTICLES

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]

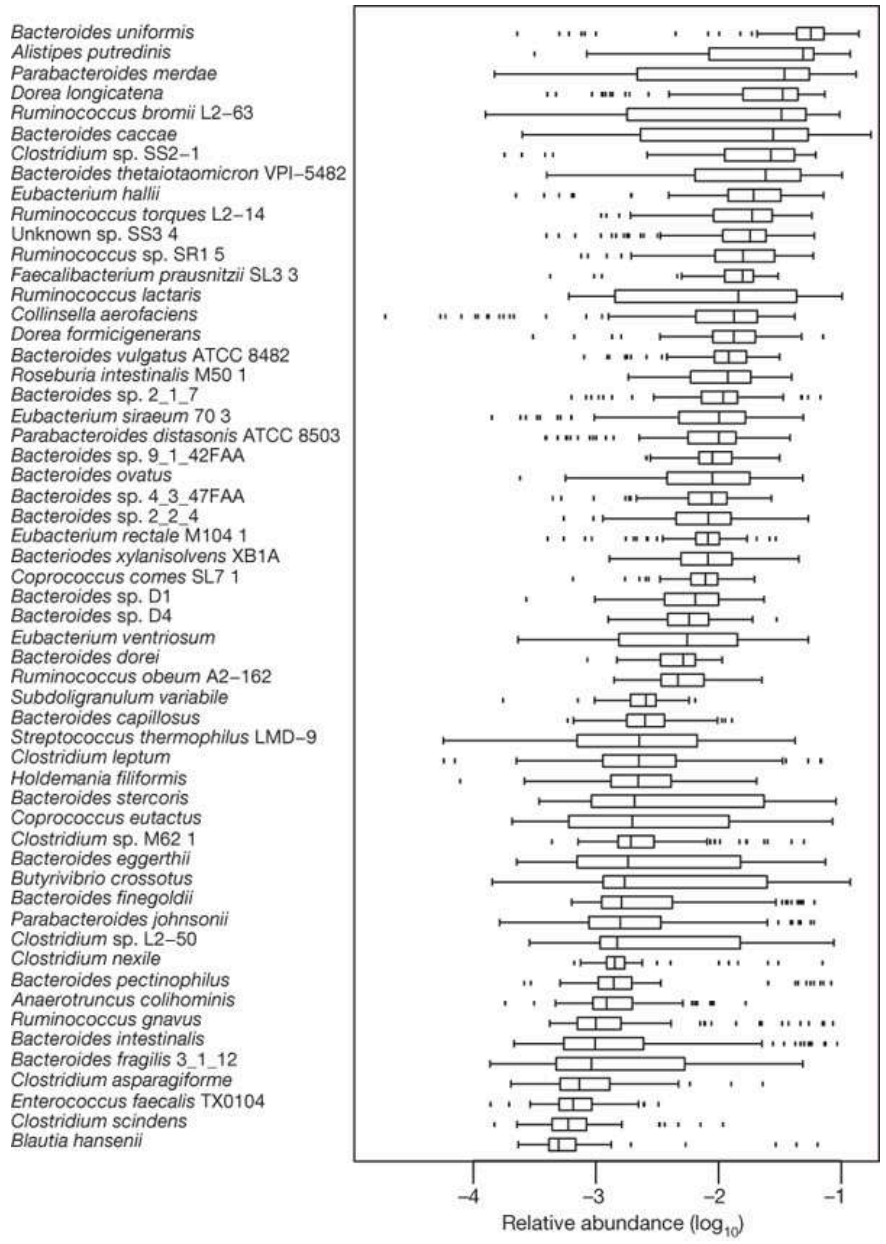
doi:10.1038/nature08821

Human gut microbiome



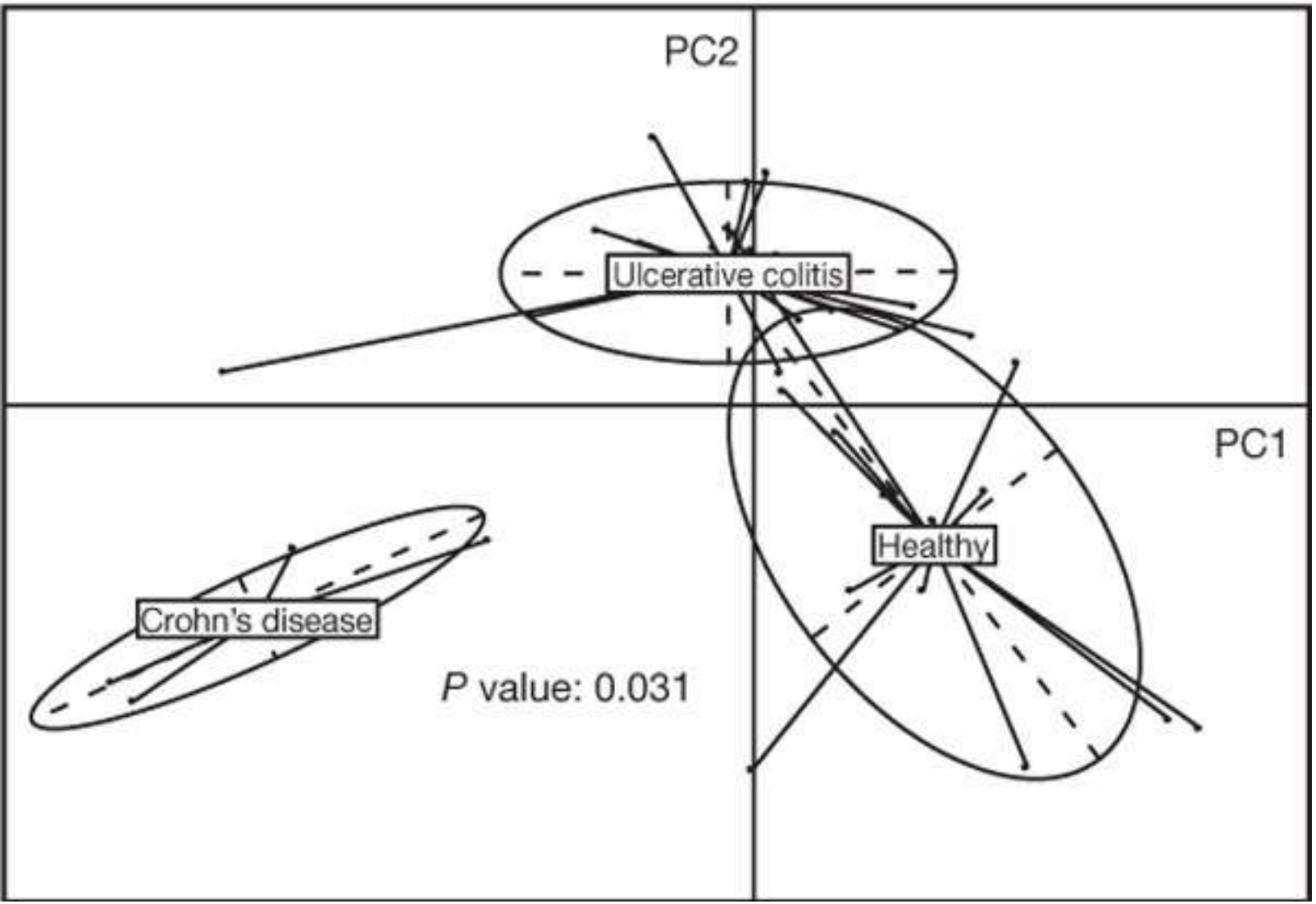
doi:10.1038/nature08821

Human gut microbiome



doi:10.1038/nature08821

Human gut microbiome

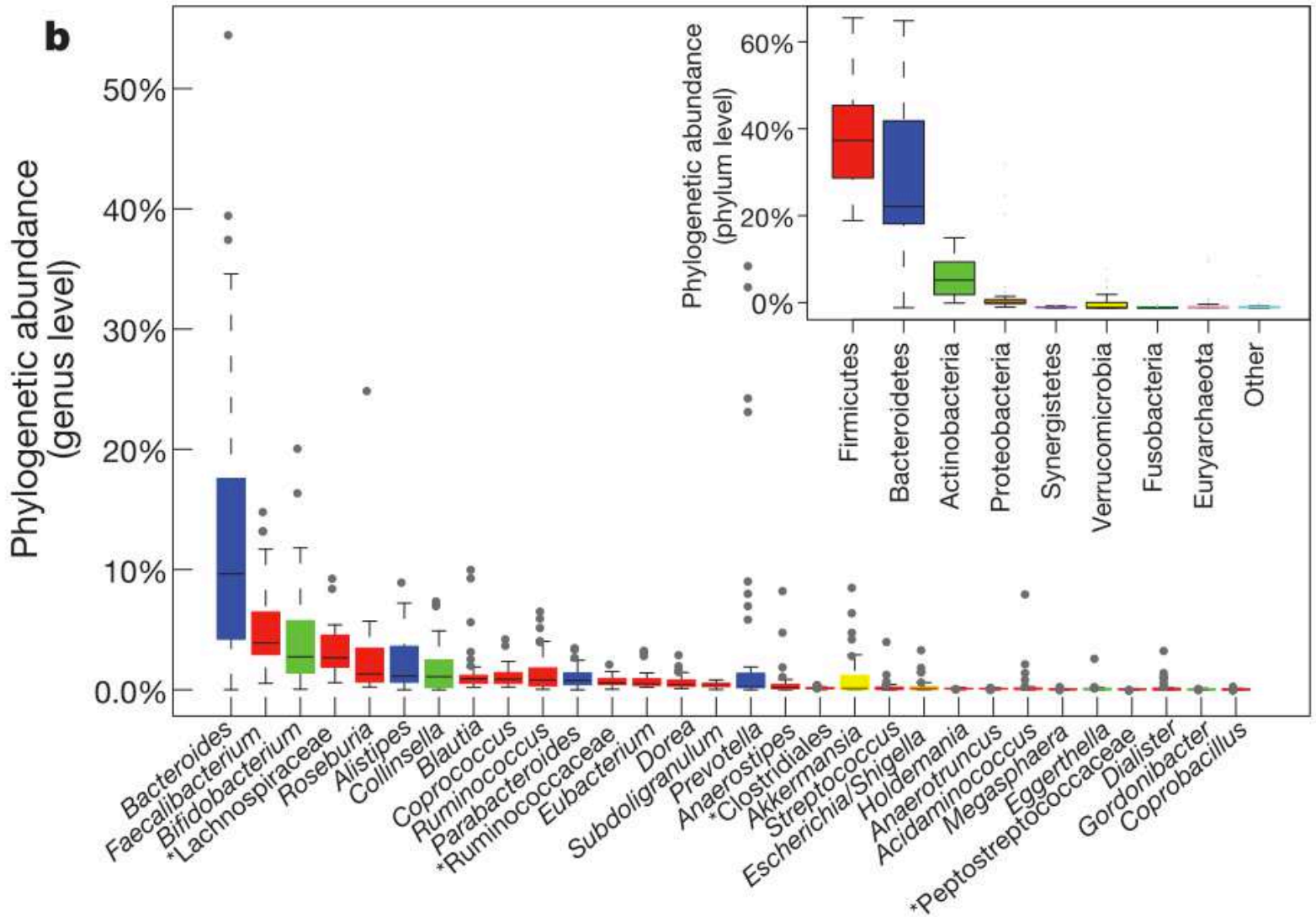


We can check which OTUs constitute the clustering (and separation) patterns

-> Biology
-> Biomarkers

doi:10.1038/nature08821

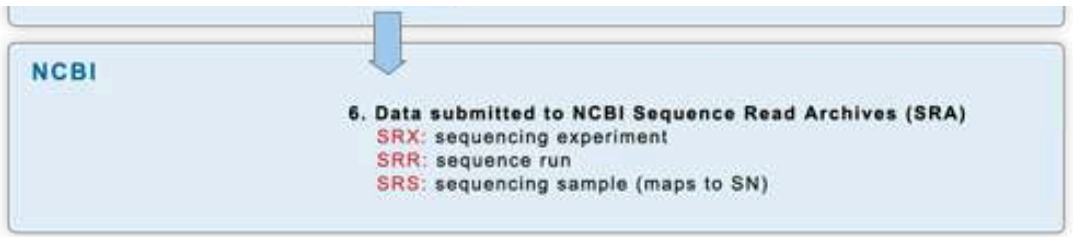
Human gut microbiome



doi:10.1038/nature09944

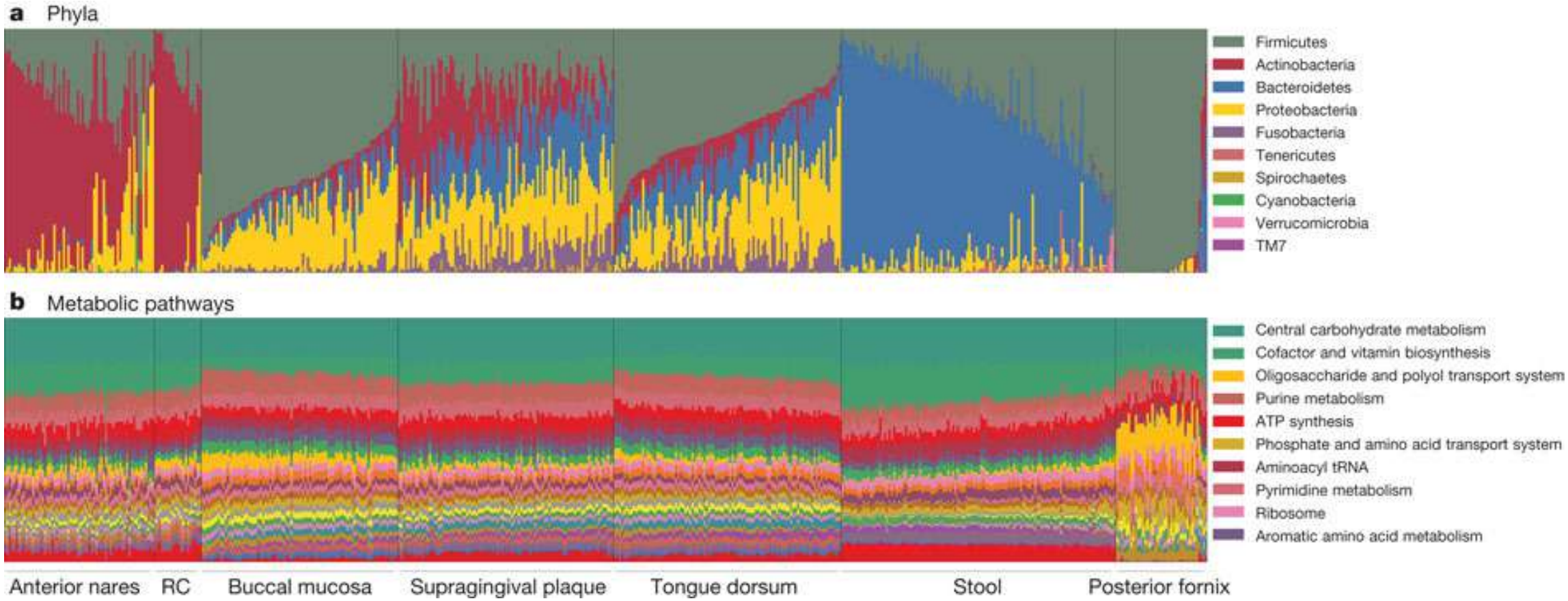
Table 1 | HMP donor samples examined by 16S and WGS

Body region	Body site	Total samples	Total 16S samples	V13 samples	V13 read depth (M)*	V35 samples	V35 read depth (M)*	Samples V13 and V35	Total WGS samples	Total read depth (G)†	Filtered reads (%)‡	Human reads (%)§	Remaining read depth (G)†	Samples 16S and WGS
Gut	Stool	352	337	193	1.4	328	2.4	184	136	1,720.7	15	1	1,450.6	124
Oral cavity	Buccal mucosa	346	330	184	1.3	314	1.7	168	107	1,438.0	9	82	136.7	91
	Hard palate	325	325	179	1.2	310	1.7	164	1	10.9	20	25	5.9	1
	Keratinized gingiva	335	329	183	1.3	319	1.7	173	6	72.3	5	47	34.4	0
	Palatine tonsils	337	332	189	1.2	315	1.9	172	6	74.8	2	80	13.5	1
	Saliva	315	310	166	0.9	292	1.5	148	5	55.7	1	91	4.2	0
	Subgingival plaque	334	328	186	1.2	314	1.8	172	7	92.1	5	79	15.3	1
	Supragingival plaque	345	331	192	1.3	316	1.9	177	115	1,500.7	15	40	674.8	101
	Throat	331	325	176	1.0	312	1.7	163	7	78.8	4	79	13.6	1
	Tongue dorsum	348	332	193	1.3	320	2.0	181	122	1,620.1	15	19	1,084.3	106
Airway	Anterior nares	316	302	169	1.0	283	1.2	150	84	1,129.9	3	96	14.3	70
Skin	Left antecubital fossa	269	269	158	0.7	221	0.5	110	0	NA	NA	NA	0	NA
	Left retroauricular crease	313	312	188	1.6	295	1.5	171	9	126.3	9	73	22.1	8
	Right antecubital fossa	274	274	158	0.7	229	0.5	113	0	NA	NA	NA	0	NA
	Right retroauricular crease	319	316	190	1.4	304	1.6	178	15	181.9	18	59	42.4	12
Vagina	Mid-vagina	145	143	91	0.6	140	1.0	88	2	22.6	0	99	0.2	0
	Posterior fornix	152	142	89	0.6	136	1.0	83	53	702.1	6	90	25.2	43
	Vaginal introitus	142	140	87	0.6	131	0.9	78	3	36.5	1	98	0.6	1
	Total	5,298	5,177	2,971	19	4,879	26.3	2,673	681	8,863.3	11	49	3,538.1	560



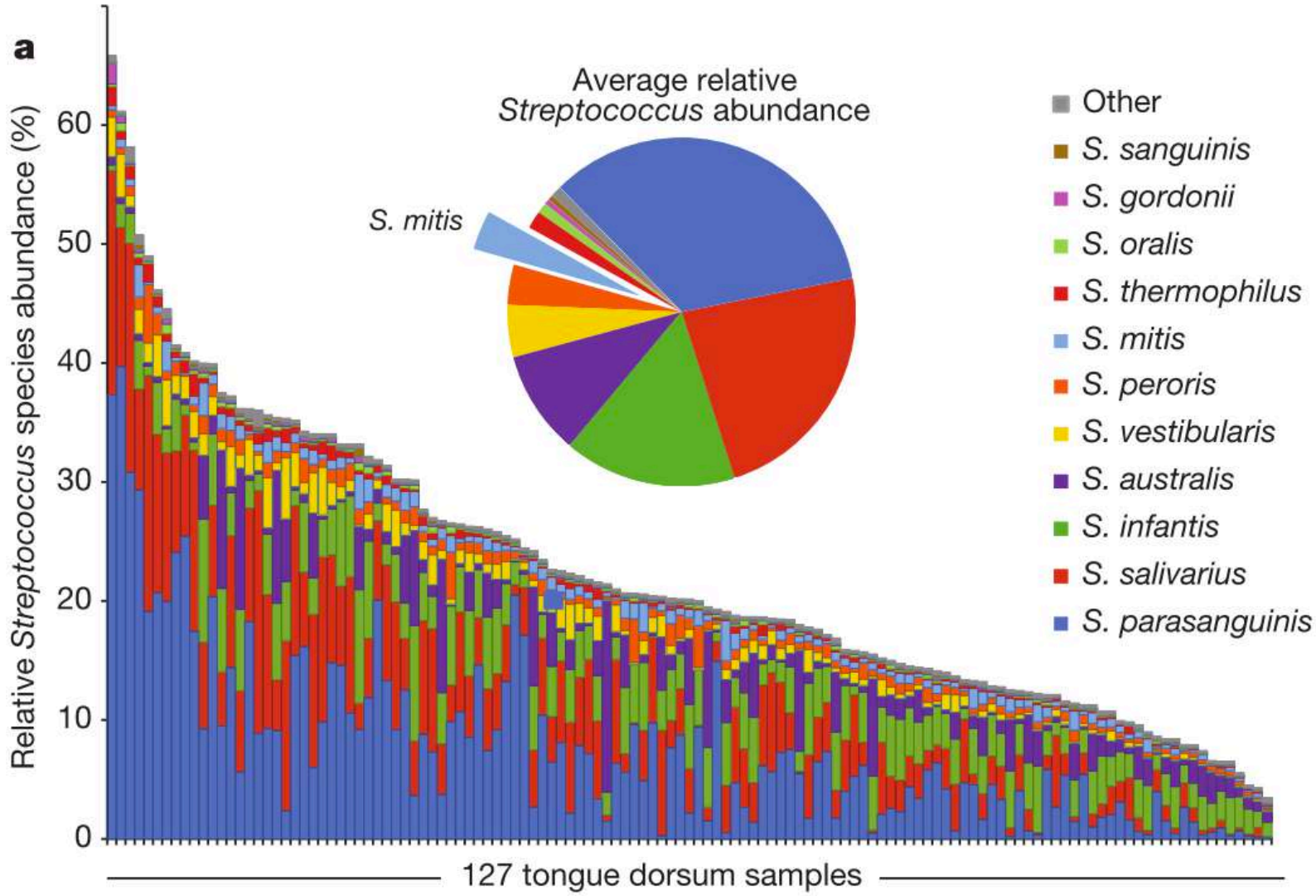
doi:10.1038/nature11209

Human microbiome



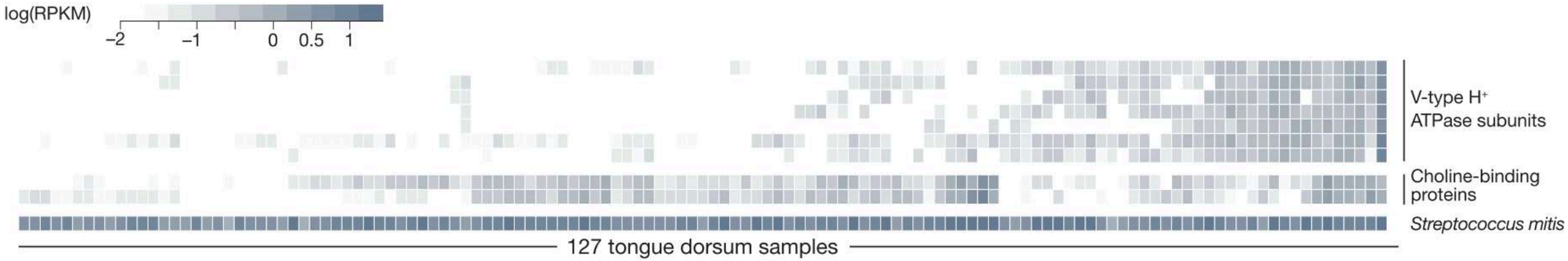
doi:10.1038/nature11234

Inter-individual variation in the microbiome proved to be specific, functionally relevant and personalized



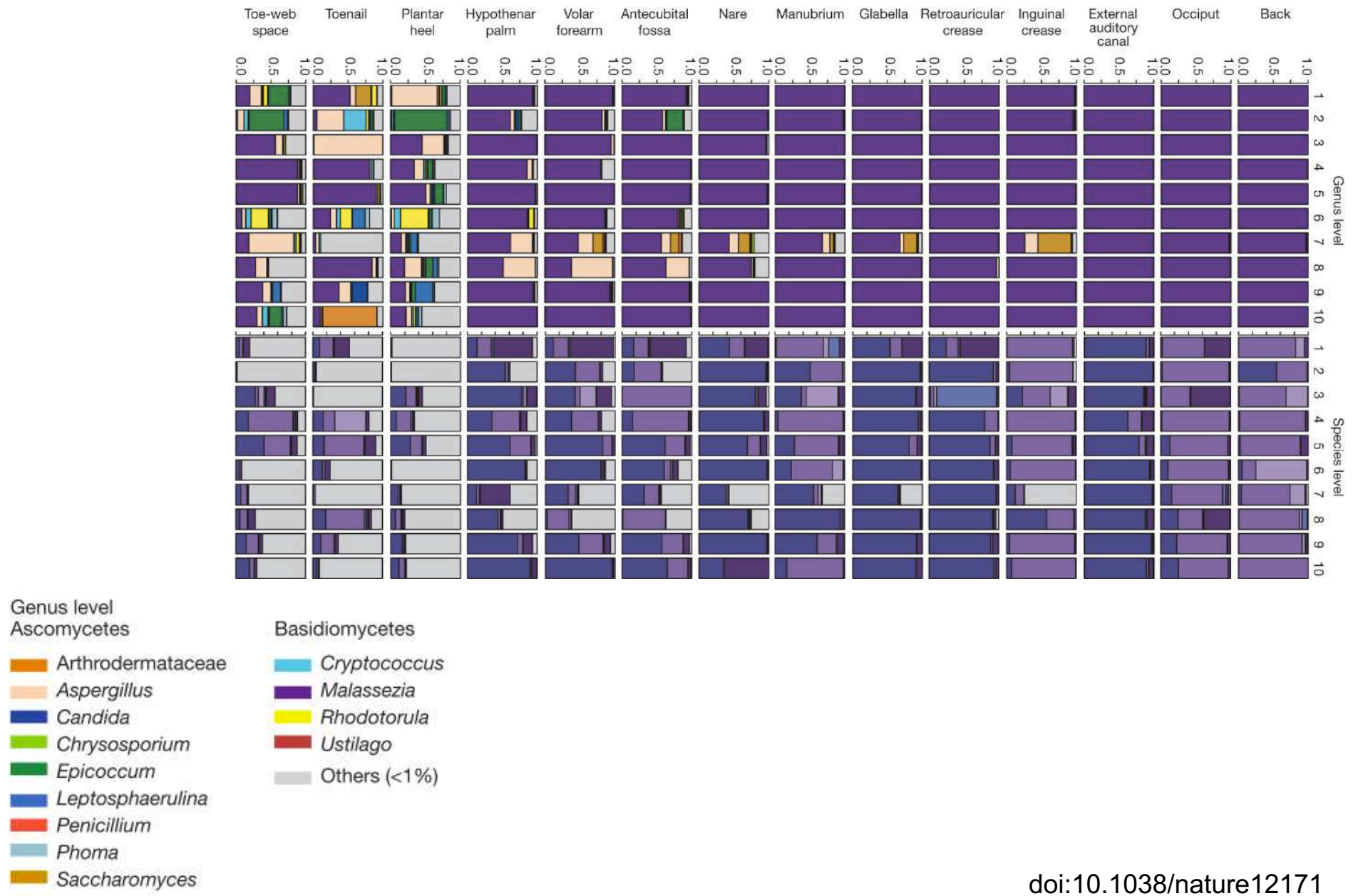
doi:10.1038/nature11234

Gene loss & Structural variants are common



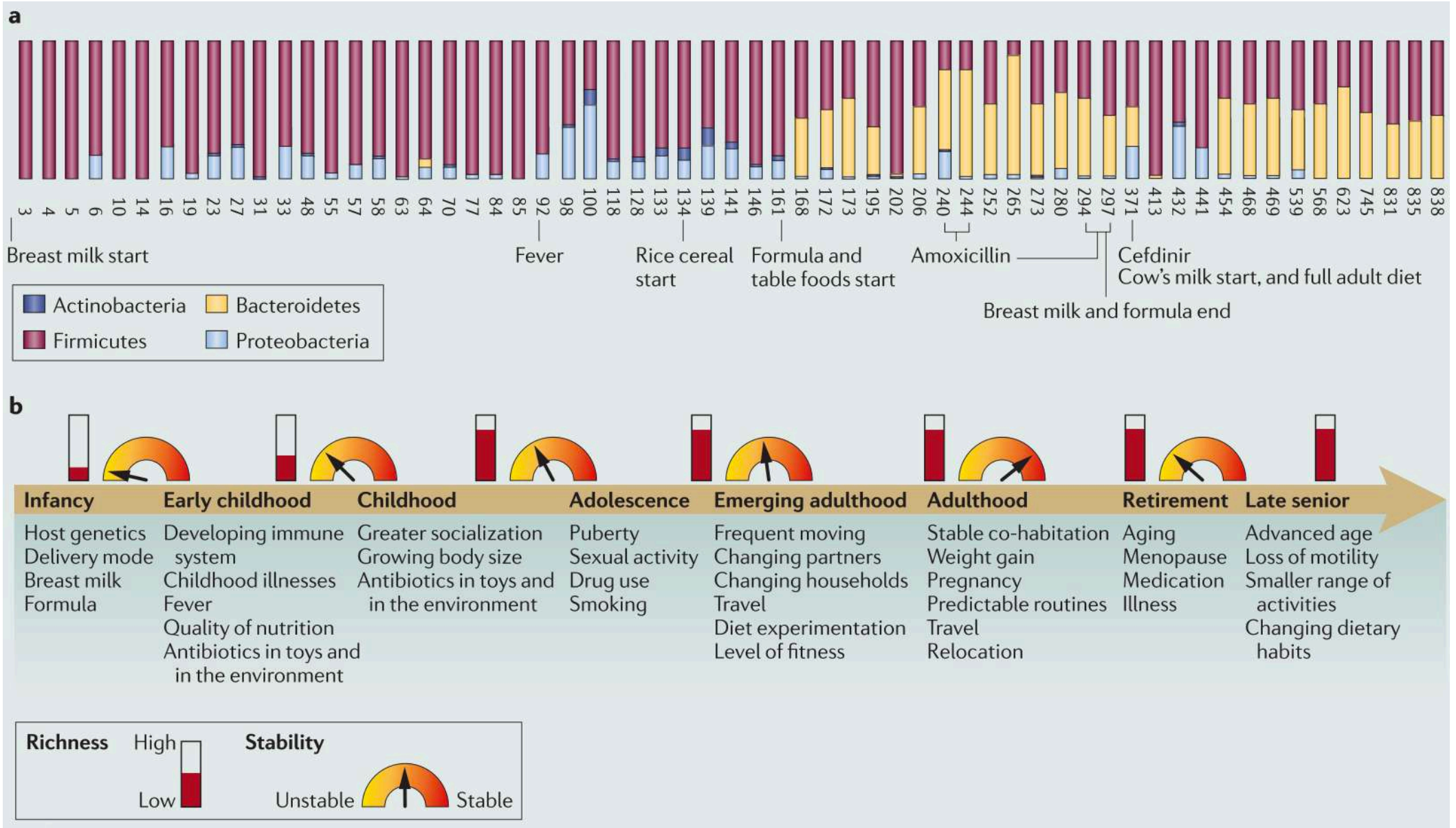
doi:10.1038/nature11234

Skins



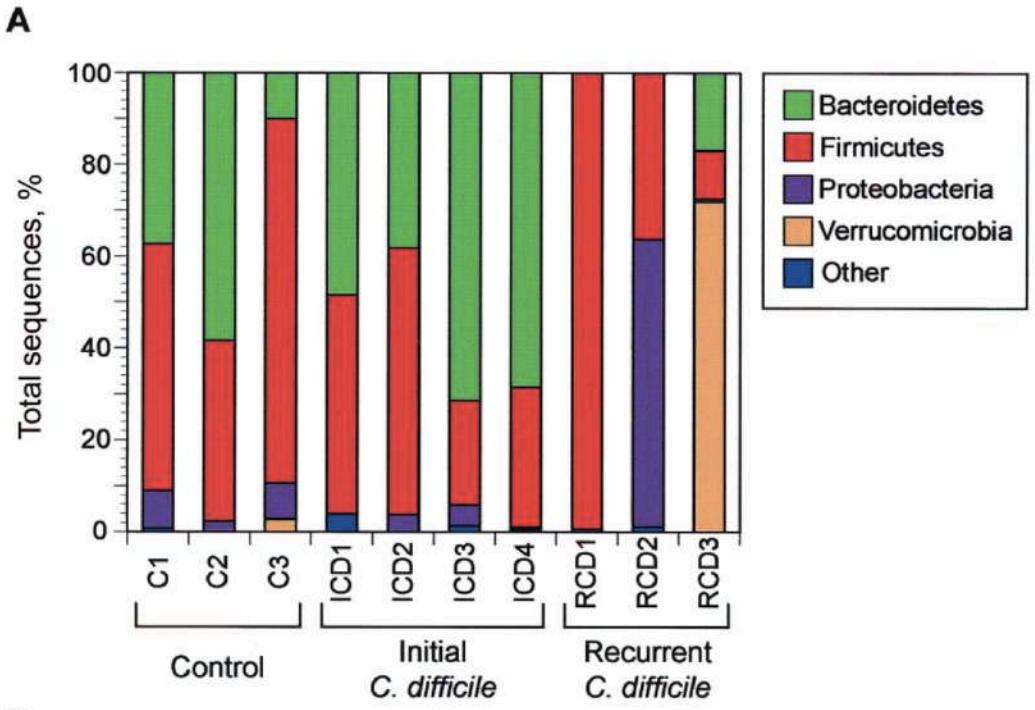
doi:10.1038/nature12171

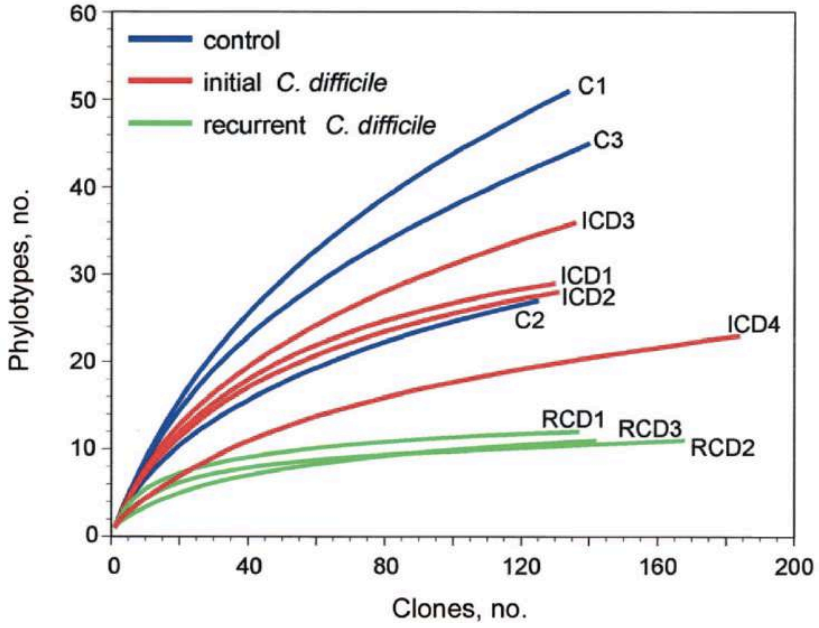
The gut microbiome during life



doi:10.1038/nrmicro2540

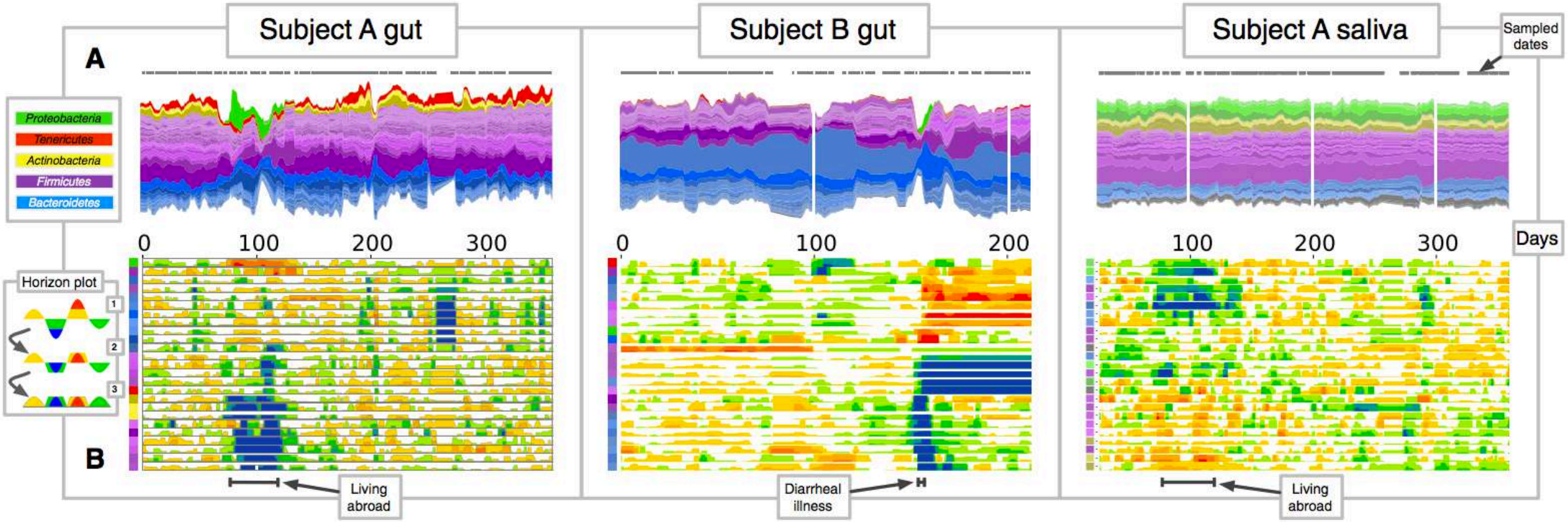
Decreased diversity with *Clostridium difficile* – assciated diarrhea





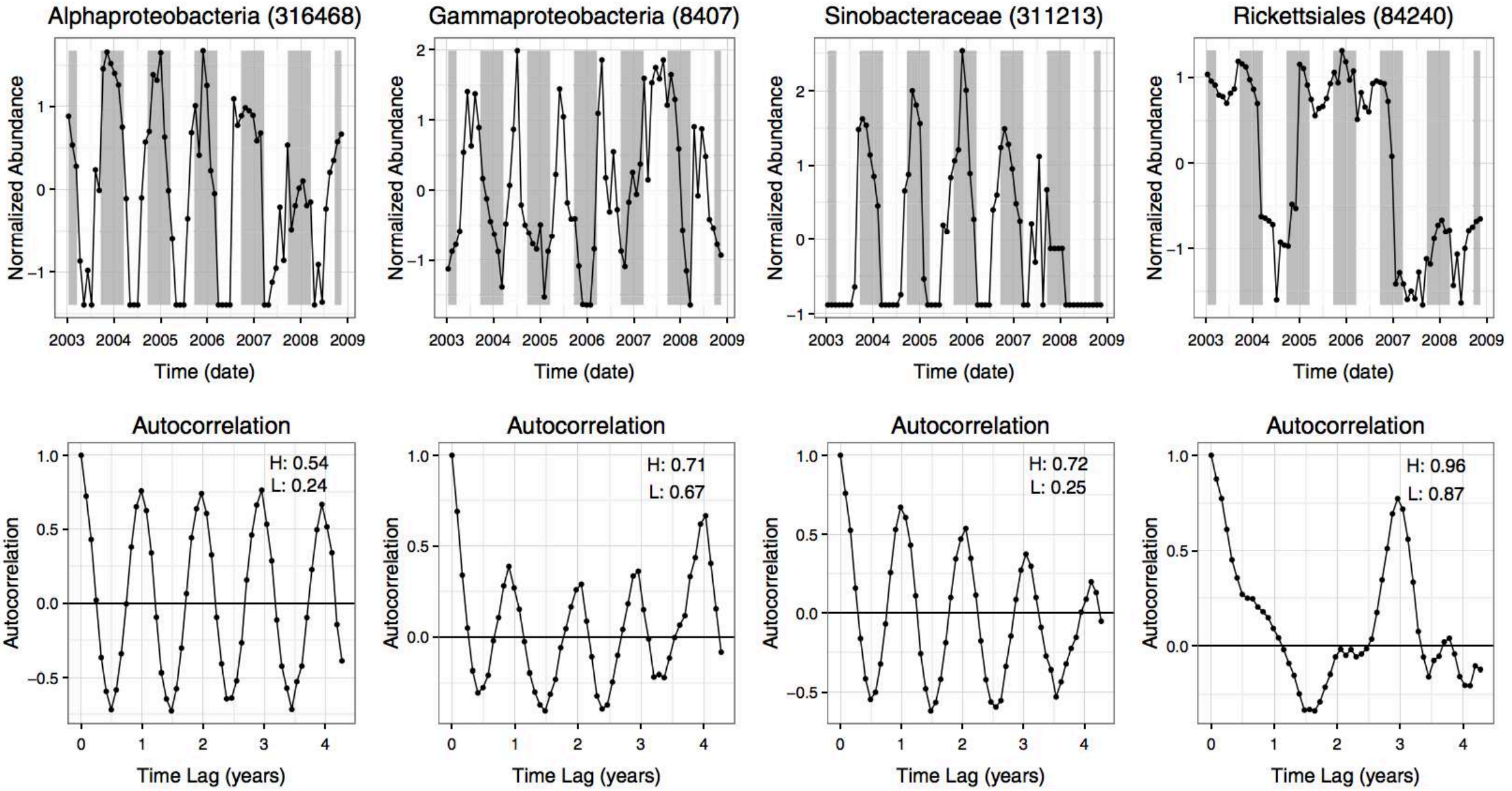
doi:10.1086/525047

Tracking microbiome on a daily scale



David et al. Genome Biology 2014, 15:R89

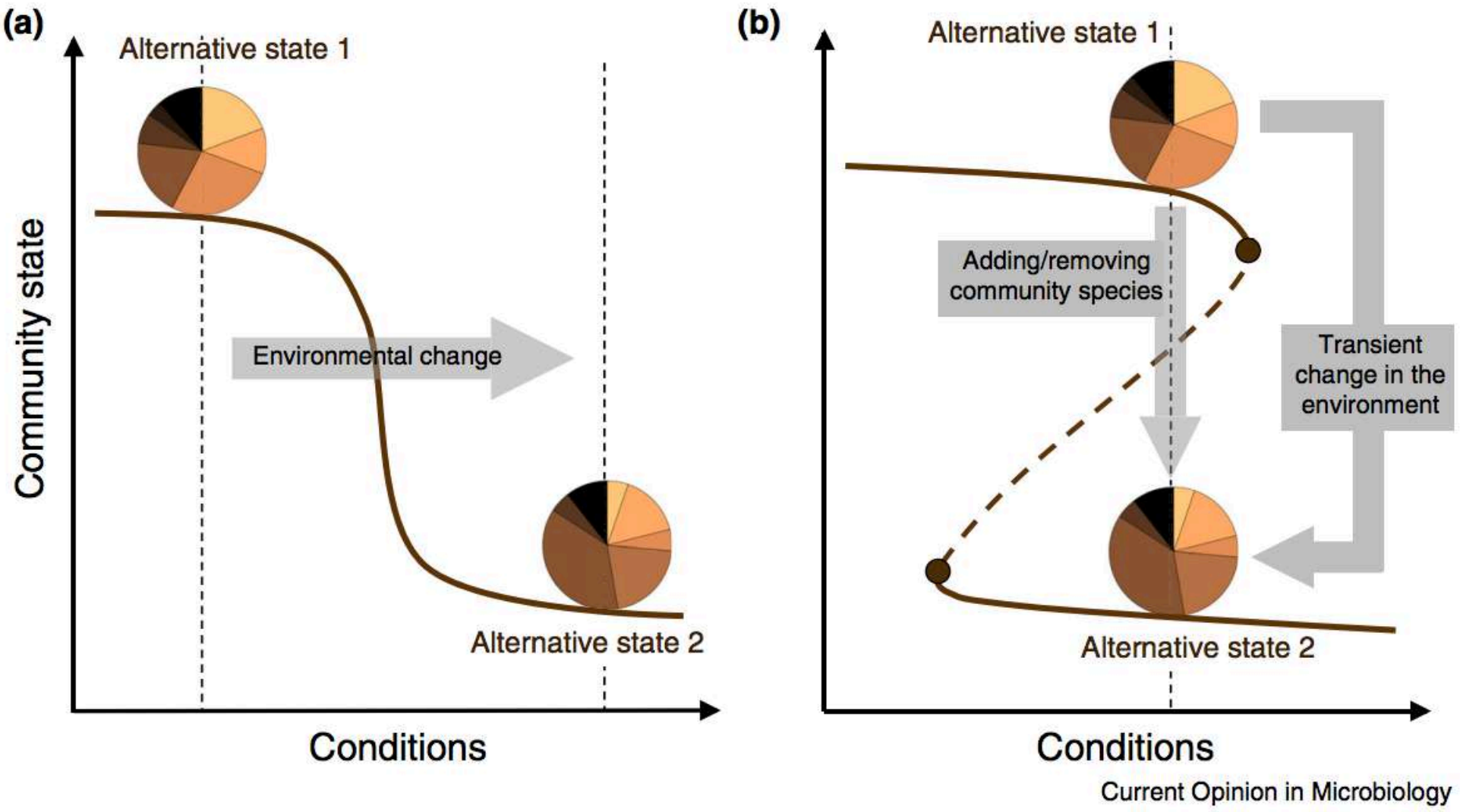
Tracking microbiome spanning 6 years



Faust et al 2015

Current Opinion in Microbiology

Tracking microbiome on a daily scale



Faust et al 2015

Question: What community gets reset and what don't?

Question: What community gets reset and what don't?

A. Shade, J.S. Read, N.D. Youngblut, N. Fierer, R. Knight, T.K. Kratz, N.R. Lottig, E.E. Roden, E.H. Stanley, J. Stombaugh, et al.
Lake microbial communities are resilient after a whole-ecosystem disturbance
ISME J, 6 (2012), pp. 2153–2167

Yes

L. Dethlefsen, D.A. Relman
Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation
Proc Natl Acad Sci U S A, 108 (2011), pp. 4554–4561

No

L.A. David, A.C. Materna, J. Friedman, M.I. Campos-Baptista, M.C. Blackburn, A. Perrotta, S.E. Erdman, E.J. Alm
Host lifestyle affects human microbiota on daily timescales
Genome Biol, 15 (2014), p. R89

Yes and No

Faust et al 2015

ARTICLE

OPEN

doi:10.1038/nature23889

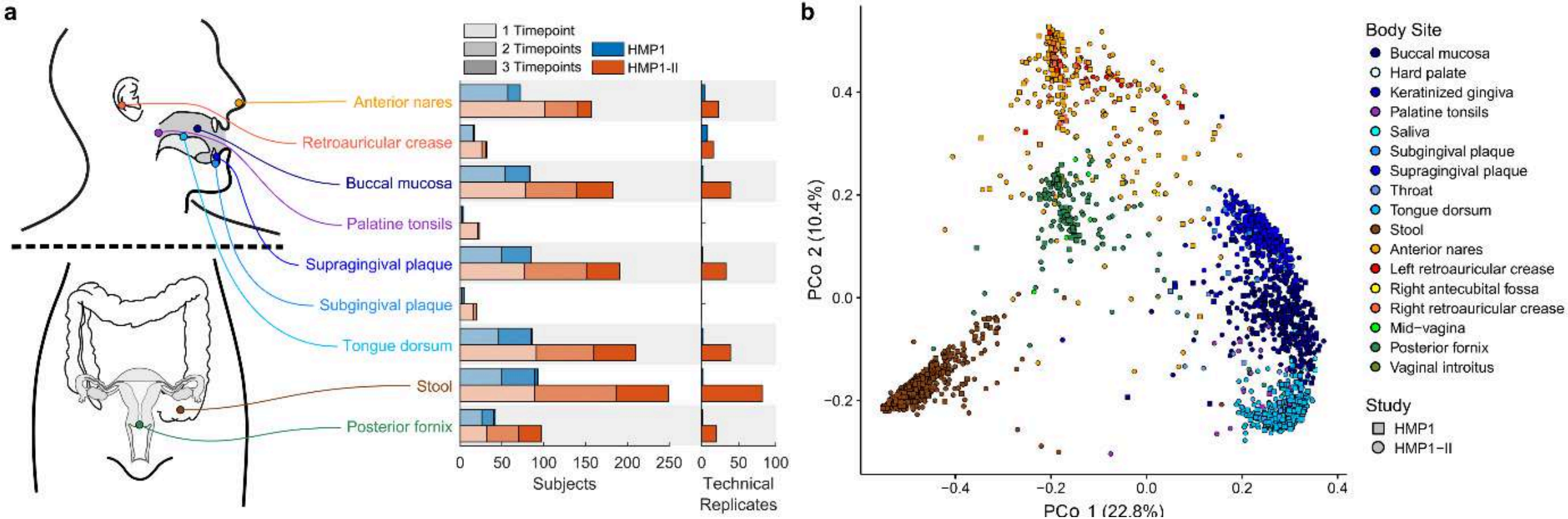
Strains, functions and dynamics in the expanded Human Microbiome Project

Jason Lloyd-Price[1,2]*, Anup Mahurkar[3]*, Gholamali Rahnavard[1,2], Jonathan Crabtree[3], Joshua Orvis[3], A. Brantley Hall[2], Arthur Brady[3], Heather H. Creasy[3], Carrie McCracken[3], Michelle G. Giglio[3], Daniel McDonald[4], Eric A. Franzosa[1,2], Rob Knight[4,5], Owen White[3] & Curtis Huttenhower[1,2]

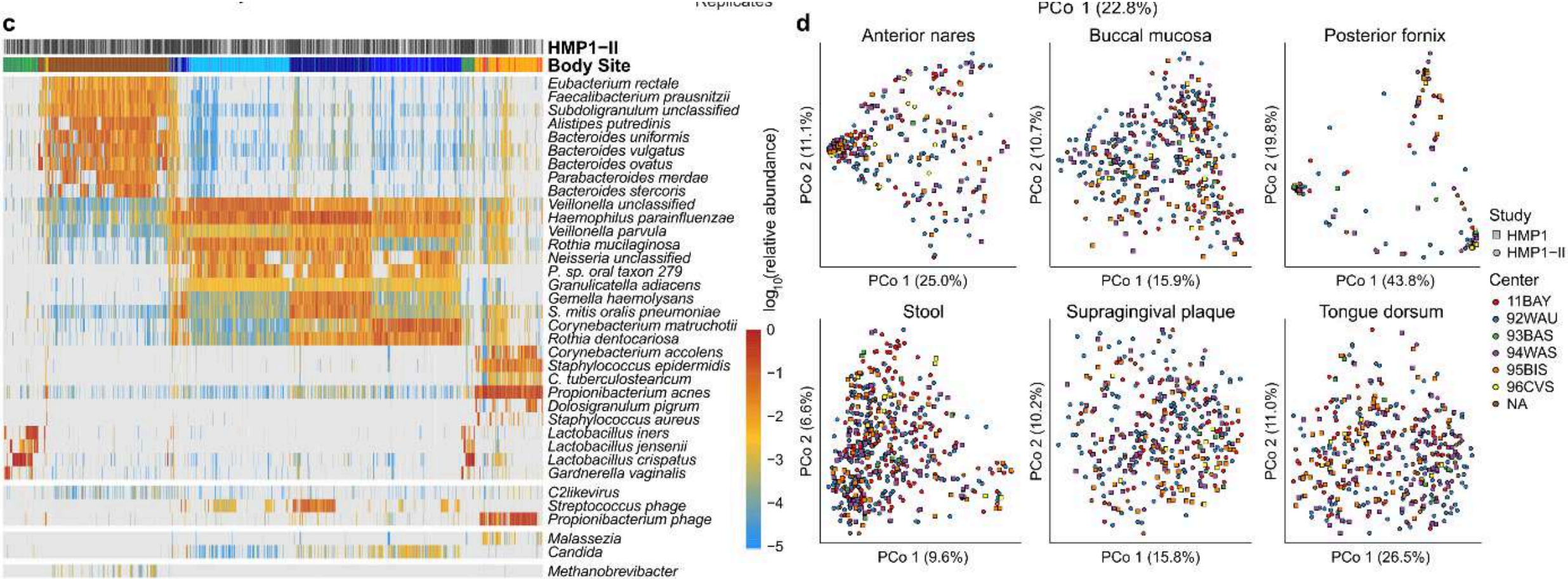
The characterization of baseline microbial and functional diversity in the human microbiome has enabled studies of microbiome-related disease, diversity, biogeography, and molecular function. The National Institutes of Health Human Microbiome Project has provided one of the broadest such characterizations so far. Here we introduce a second wave of data from the study, comprising 1,631 new metagenomes (2,355 total) targeting diverse body sites with multiple time points in 265 individuals. We applied updated profiling and assembly methods to provide new characterizations of microbiome personalization. Strain identification revealed subspecies clades specific to body sites; it also quantified species with phylogenetic diversity under-represented in isolate genomes. Body-wide functional profiling classified pathways into universal, human-enriched, and body site-enriched subsets. Finally, temporal analysis decomposed microbial variation into rapidly variable, moderately variable, and stable subsets. This study furthers our knowledge of baseline human microbial diversity and enables an understanding of personalized microbiome function and dynamics.

Received 30 November 2016; accepted 8 August 2017.
Published online 20 September 2017.

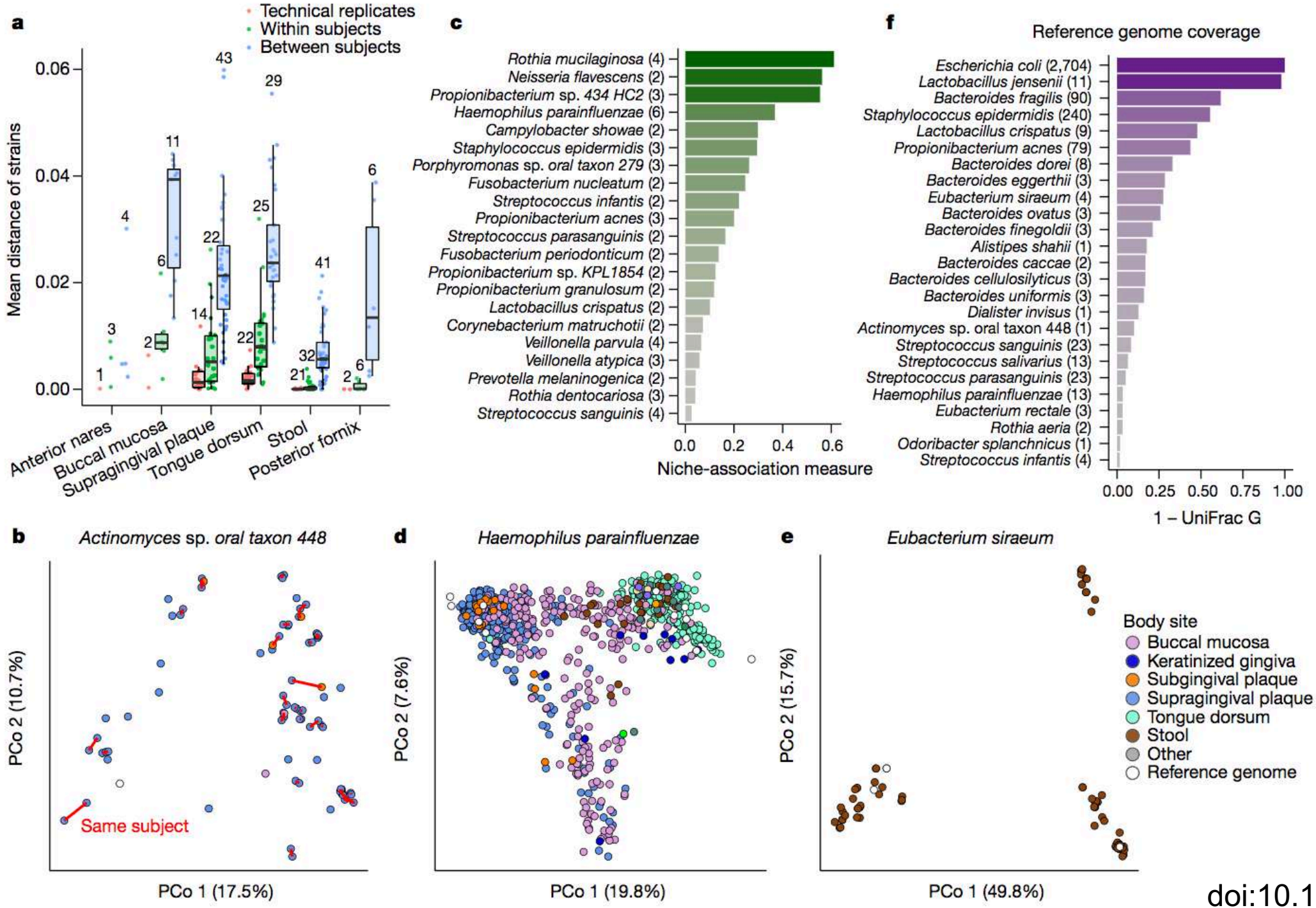
doi:10.1038/nature23889



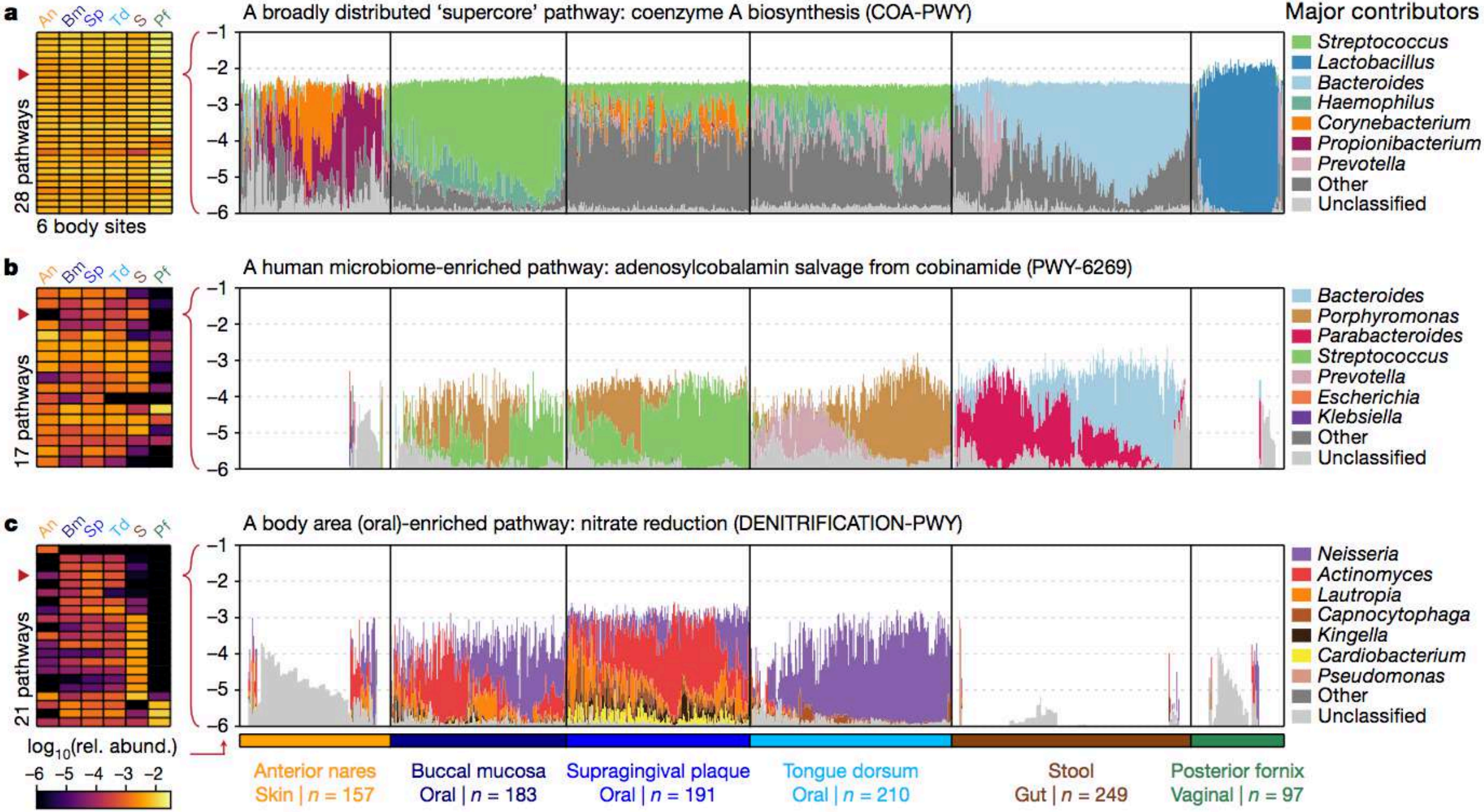
Extended Data Figure 1 | Extended body-wide metagenomic taxonomic profiles in HMP1-II. a, The combined HMP1-II datasets include a total of 2,355 metagenomes (724 previously published and 1,631 new, including 252 technical replicates). These span the project's six targeted body sites (anterior nares, buccal mucosa, supragingival plaque, tongue dorsum, stool, and posterior fornix) in addition to at least 20 samples each from 3 additional sites, of the 18 total sampled sites: retroauricular crease, palatine tonsils, and subgingival plaque. Metagenomes are now available for at least one body site for a total of 265 individuals. **b**, PCoA using Bray–Curtis distances among all microbes at the species level. **c**, Relative abundances of the most prevalent and abundant microbes (bacterial, viral, eukaryotic, and archaeal) among all body sites, as profiled by MetaPhlAn2[20]. Prevalent eukaryotic microbes are shown at the genus level. **d**, Taxonomic profiles do not vary more between sequencing centres, batches, or clinical centres than they do among individuals within body sites. Ordinations show Bray–Curtis principal coordinates of species-level abundances at each body site. Within-site ecological structure is as expected[1], with no divergence associated with technical variables along the first two ordination axes.



Extended Data Figure 1 | Extended body-wide metagenomic taxonomic profiles in HMP1-II. a, The combined HMP1-II datasets include a total of 2,355 metagenomes (724 previously published and 1,631 new, including 252 technical replicates). These span the project's six targeted body sites (anterior nares, buccal mucosa, supragingival plaque, tongue dorsum, stool, and posterior fornix) in addition to at least 20 samples each from 3 additional sites, of the 18 total sampled sites: retroauricular crease, palatine tonsils, and subgingival plaque. Metagenomes are now available for at least one body site for a total of 265 individuals. **b**, PCoA using Bray–Curtis distances among all microbes at the species level. **c**, Relative abundances of the most prevalent and abundant microbes (bacterial, viral, eukaryotic, and archaeal) among all body sites, as profiled by MetaPhlAn2[20]. Prevalent eukaryotic microbes are shown at the genus level. **d**, Taxonomic profiles do not vary more between sequencing centres, batches, or clinical centres than they do among individuals within body sites. Ordinations show Bray–Curtis principal coordinates of species-level abundances at each body site. Within-site ecological structure is as expected[1], with no divergence associated with technical variables along the first two ordination axes.



doi:10.1038/nature23889


a
An Bm Sp Td S Pf
28 pathways
6 body sites
A broadly distributed 'supercore' pathway: coenzyme A biosynthesis (COA-PWY)
−1 −2 −3 −4 −5 −6
Major contributors
Streptococcus
Lactobacillus
Bacteroides
Haemophilus
Corynebacterium
Propionibacterium
Prevotella
Other
Unclassified
b
An Bm Sp Td S Pf
17 pathways
A human microbiome-enriched pathway: adenosylcobalamin salvage from cobinamide (PWY-6269)
−1 −2 −3 −4 −5 −6
Bacteroides
Porphyromonas
Parabacteroides
Streptococcus
Prevotella
Escherichia
Klebsiella
Other
Unclassified
c
An Bm Sp Td S Pf
21 pathways
log10(rel. abund.)
−6 −5 −4 −3 −2
A body area (oral)-enriched pathway: nitrate reduction (DENITRIFICATION-PWY)
−1 −2 −3 −4 −5 −6
Neisseria
Actinomyces
Lautropia
Capnocytophaga
Kingella
Cardiobacterium
Pseudomonas
Other
Unclassified
Anterior nares
Skin | n = 157
Buccal mucosa
Oral | n = 183
Supragingival plaque
Oral | n = 191
Tongue dorsum
Oral | n = 210
Stool
Gut | n = 249
Posterior fornix
Vaginal | n = 97

ARTICLE

doi:10.1038/nature25973

Environment dominates over host genetics in shaping human gut microbiota

Daphna Rothschild[1,2*], Omer Weissbrod[1,2*], Elad Barkan[1,2*], Alexander Kurilshikov[3], Tal Korem[1,2], David Zeevi[1,2], Paul I. Costea[1,2], Anastasia Godneva[1,2], Iris N. Kalka[1,2], Noam Bar[1,2], Smadar Shilo[1,2], Dar Lador[1,2], Arnau Vich Vila[3,4], Niv Zmora[5,6,7], Meirav Pevsner-Fischer[5], David Israeli[8], Noa Kosower[1,2], Gal Malka[1,2], Bat Chen Wolf[1,2], Tali Avnit-Sagi[1,2], Maya Lotan-Pompan[1,2], Adina Weinberger[1,2], Zamir Halpern[7,9], Shai Carmi[10], Jingyuan Fu[3,11], Cisca Wijmenga[3,12], Alexandra Zhernakova[3], Eran Elinav[5]§ & Eran Segal[1,2]§

Human gut microbiome composition is shaped by multiple factors but the relative contribution of host genetics remains elusive. Here we examine genotype and microbiome data from 1,046 healthy individuals with several distinct ancestral origins who share a relatively common environment, and demonstrate that the gut microbiome is not significantly associated with genetic ancestry, and that host genetics have a minor role in determining microbiome composition. We show that, by contrast, there are significant similarities in the compositions of the microbiomes of genetically unrelated individuals who share a household, and that over 20% of the inter-person microbiome variability is associated with factors related to diet, drugs and anthropometric measurements. We further demonstrate that microbiome data significantly improve the prediction accuracy for many human traits, such as glucose and obesity measures, compared to models that use only host genetic and environmental data. These results suggest that microbiome alterations aimed at improving clinical outcomes may be carried out across diverse genetic backgrounds.

Received 7 June 2017; accepted 16 January 2018.
Published online 28 February 2018.

doi:10.1038/nature25973

- 1,046 healthy Israeli adults
- 16S rRNA + metagenomics
- Genotyping 712,540 SNPs
- Questionnaires

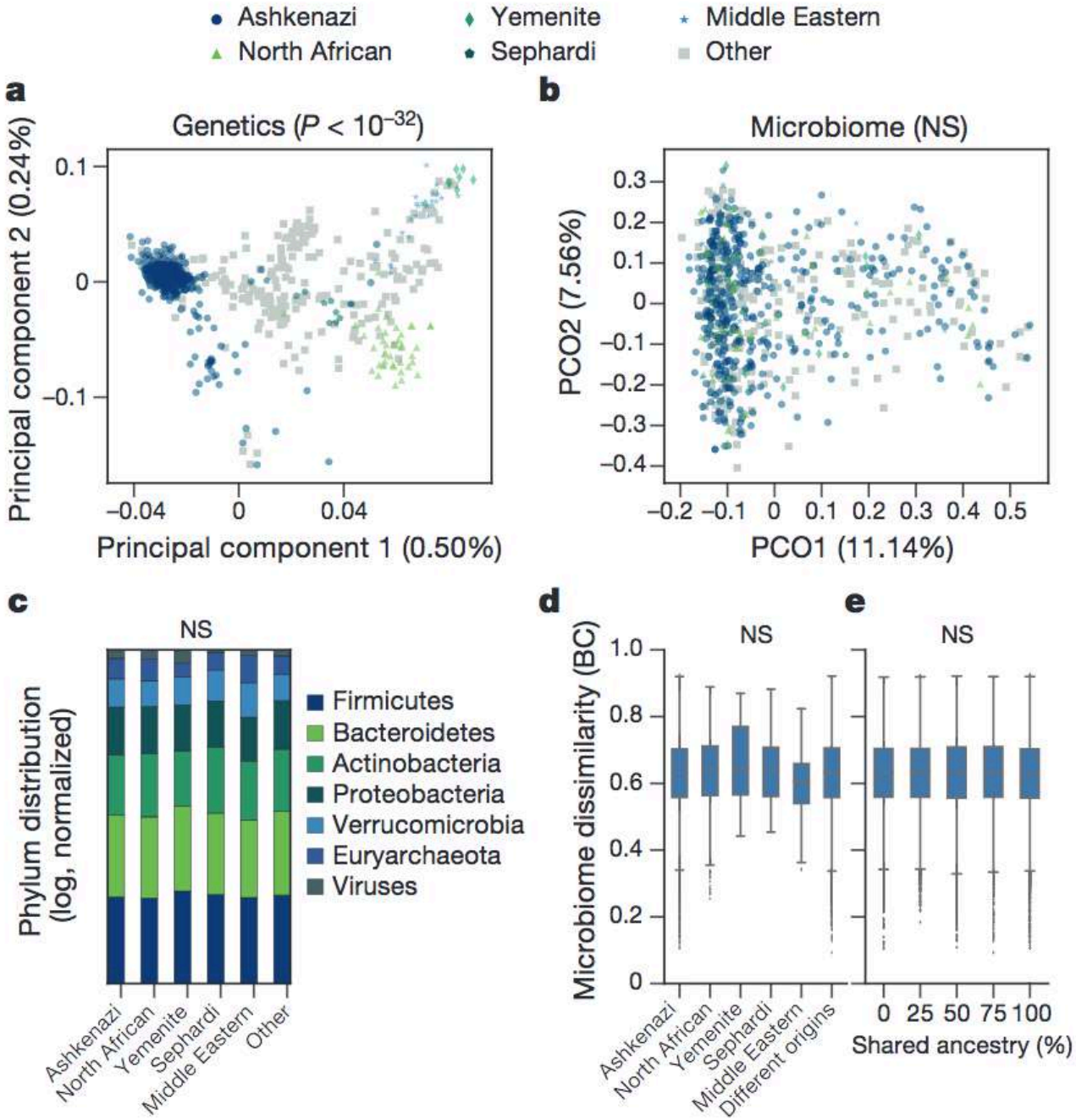


Figure 1 | Genetic ancestry is not significantly associated with microbiome composition. **a**, Genetic principal components are strongly associated with self-reported ancestry, with Ashkenazi ($n = 345$), North African ($n = 42$), Middle Eastern ($n = 24$), Sephardi ($n = 10$), Yemenite ($n = 8$) and admixed/other (other) ($n = 286$) ancestries ($P < 10^{-32}$; Kruskal–Wallis). **b**, As in **a**, but for microbiome principal coordinate analysis ($P > 0.08$; Kruskal–Wallis). **c**, The distribution of average phylum abundance among 582 non-admixed individuals (in log scale, normalized to sum to 1.0) is not associated with ancestry ($P > 0.05$; Kruskal–Wallis). NS, not significant. **d**, Box plots of Bray–Curtis (BC) dissimilarities across all pairs of 737 individuals for whom the ancestries of all grandparents are known, demonstrating that microbiome composition is not associated with ancestry ($P > 0.06$; Kruskal–Wallis test for the top five Bray–Curtis PCOs). $n = 105{,}570$ (Ashkenazi), 1,711 (North African), 528 (Middle Eastern), 136 (Sephardi) and 78 (Yemenite) same ancestry pairs; $n = 61{,}048$ different ancestry pairs. The lower and upper limits of the boxes represent the 25% and 75% percentiles, respectively, and the top and bottom whiskers represent the 5% and 95% percentiles, respectively. **e**, Box plots of Bray–Curtis dissimilarities across pairs of 946 individuals (including admixed individuals), organized according to shared ancestry fraction (the fraction of grandparents of the same ancestry), for pairs with 0% ($n = 167{,}618$), 25% ($n = 33{,}119$), 50% ($n = 100{,}163$), 75% ($n = 34{,}187$) and 100% ($n = 111{,}898$) shared ancestry fractions. The lower and upper limits of the boxes represent the 25% and 75% percentiles, respectively, and the top and bottom whiskers represent the 5% and 95% percentiles, respectively. The figure demonstrates that microbiome similarity is not associated with ancestral similarity ($P = 0.73$; Mantel test).

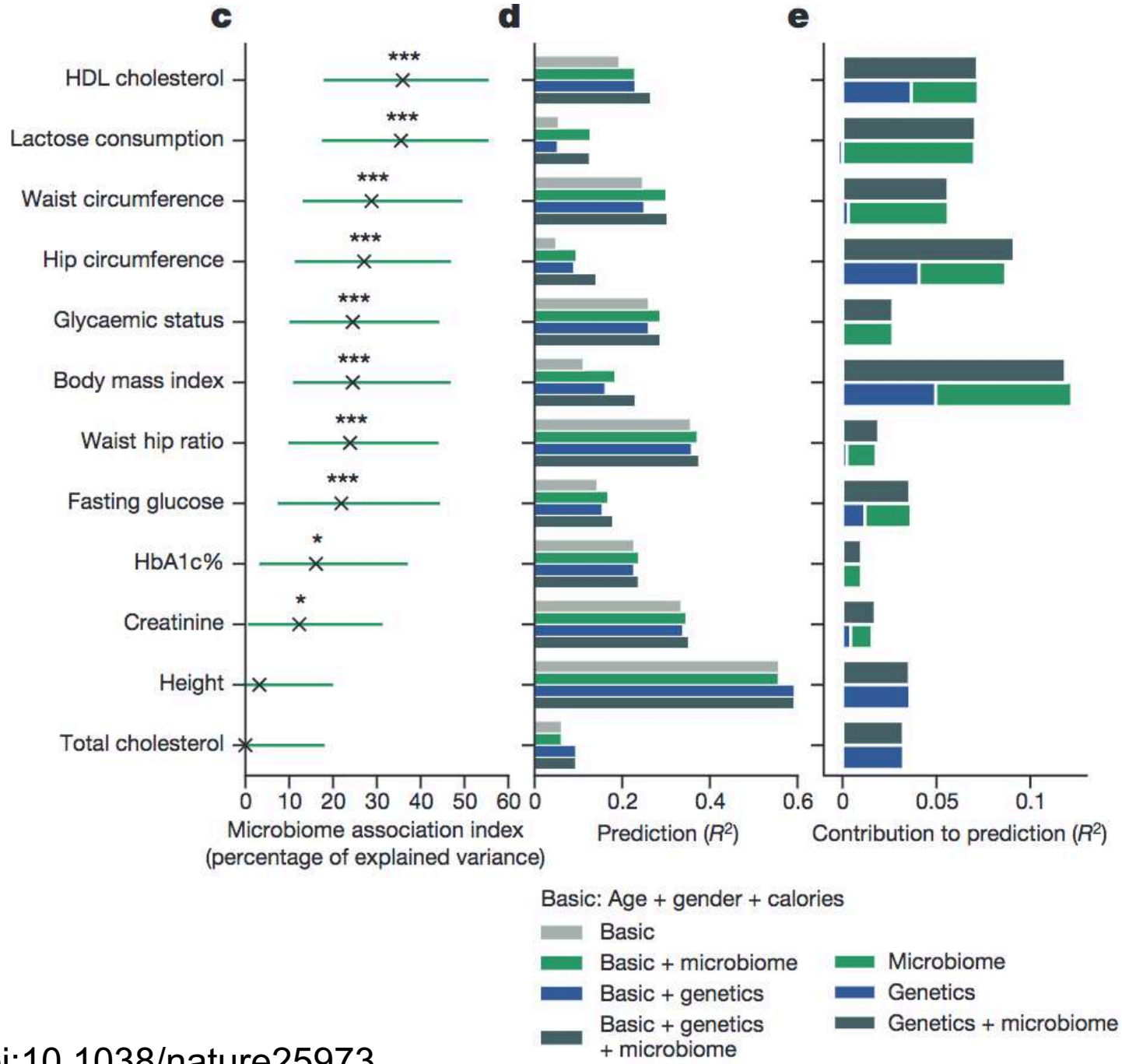
doi:10.1038/nature25973

b

Phenotype	Microbiome association index Israeli cohort	LLD cohort	Genetic heritability (literature)
HDL	35.9%***	27.9%***	23.9%–48%
Lactose cons.	35.5%***	N/A	N/A
Waist circ.	28.8%***	26%***	15%–24%
Hip circ.	27.1%***	28%***	10.6%–27%
Glycaemic status	24.5%***	N/A	N/A
BMI	24.5%***	27.8%***	14%–32%
WHR	23.9%***	6.9%*	12%–14%
Fasting glucose	21.9%***	8%**	9%–33%
HbA1c%	16.1%*	8.4%	21%–32%
Creatinine	12.3%*	6.7%	19%–25%
Height	3.2%	25.9%***	33%–68%
Total cholesterol	0%	13.5%	14%–53%

indicate a greater confidence in the estimation. **b**, b^2 estimates from the analysis of 715 individuals with measured genotyped and gut microbiomes from the Israeli cohort (left column) and of 836 individuals from the LLD cohort (middle column) are comparable to previous genetic heritability estimates[27–34] (right column). *FDR < 0.05, **FDR < 0.01 and ***FDR < 0.001. Cons., consumption, circ., circumference. **c**, b^2 estimates

doi:10.1038/nature25973



***FDR < 0.001. Cons., consumption, circ., circumference. **c**, b^2 estimates of several human phenotypes and their 95% confidence intervals, evaluated using 715 individuals. *FDR < 0.05, **FDR < 0.01 and ***FDR < 0.001. **d**, Phenotype prediction accuracy for 715 individuals, evaluated using a LMM under different sets of predictive features (measured using coefficient of determination (R^2)), using four different models for each phenotype: (i) 'Basic', age, gender and diet features; (ii) 'Basic + microbiome', basic features and relative abundances of bacterial genes; (iii) 'Basic + genetics', basic features and host genotypes; and (iv) 'Basic + genetics + microbiome': basic features, relative abundances of bacterial genes and host genotypes. **e**, The additive contribution of microbiome and genetics to prediction performance evaluated using a LMM across 715 individuals, over a model that includes only basic features. The joint contribution of microbiome and genetics is similar to the sum of the individual contributions, suggesting these are independent contributions.

doi:10.1038/nature25973

Box 1 | **Ten areas of microbiome inquiry that should be pursued**

- Understanding microbiome characteristics in relation to families: which features are inherited and which are not?*
- Understanding secular trends in microbiome composition: which taxonomic groups have been lost or gained?‡
- For diseases that have changed markedly in incidence in recent decades, do changes in the microbiome have a role? Notable examples include childhood-onset asthma, food allergies, type 1 diabetes, obesity, inflammatory bowel disease and autism.*‡
- Do particular signatures of the metagenome predict risks for specific human cancers and other diseases that are associated with ageing? Can these signatures be pursued to better understand oncogenesis? (Work on *Helicobacter pylori* provides a clear example of this.)*
- How do antibiotics perturb the microbiome, both in the short-term and long-term? Does the route of administration matter?*
- How does the microbiome affect the pharmacology of medications? Can we 'micro-type' people to improve pharmacokinetics and/or reduce toxicity? Can we manipulate the microbiome to improve pharmacokinetic stability?*‡
- Can we harness knowledge of microbiomes to improve diagnostics for disease status and susceptibility?*
- Can we harness the close mechanistic interactions between the microbiome and the host to provide hints for the development of new drugs?‡
- Specifically, can we harness the microbiome to develop new narrow-spectrum antibiotics?‡
- Can we use knowledge of the microbiota to develop true probiotics (and prebiotics)?*‡

*Areas currently under investigation. ‡Proposed areas for investigation.

doi:10.1038/nrg3182

ARTICLE

OPEN

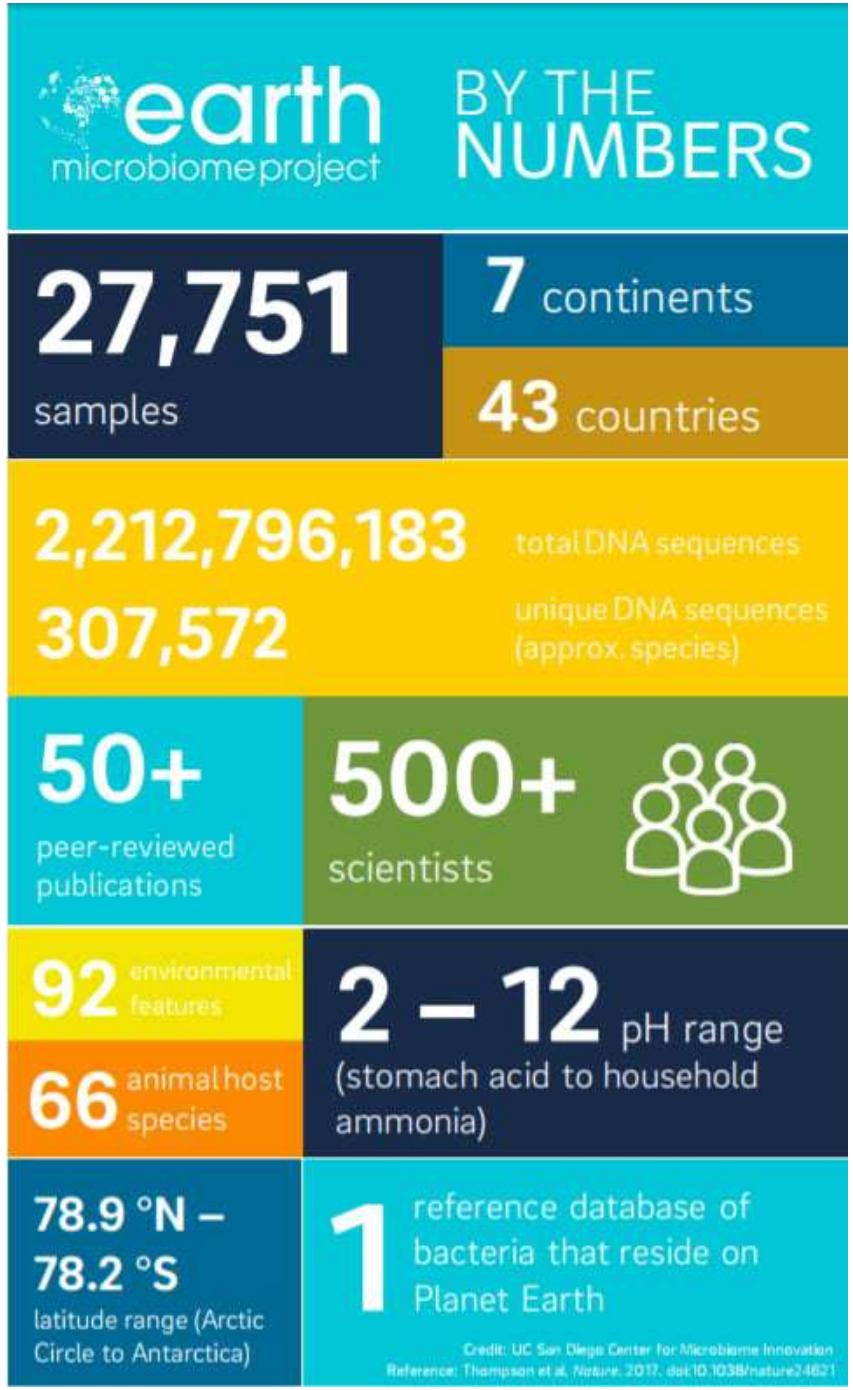
doi:10.1038/nature24621

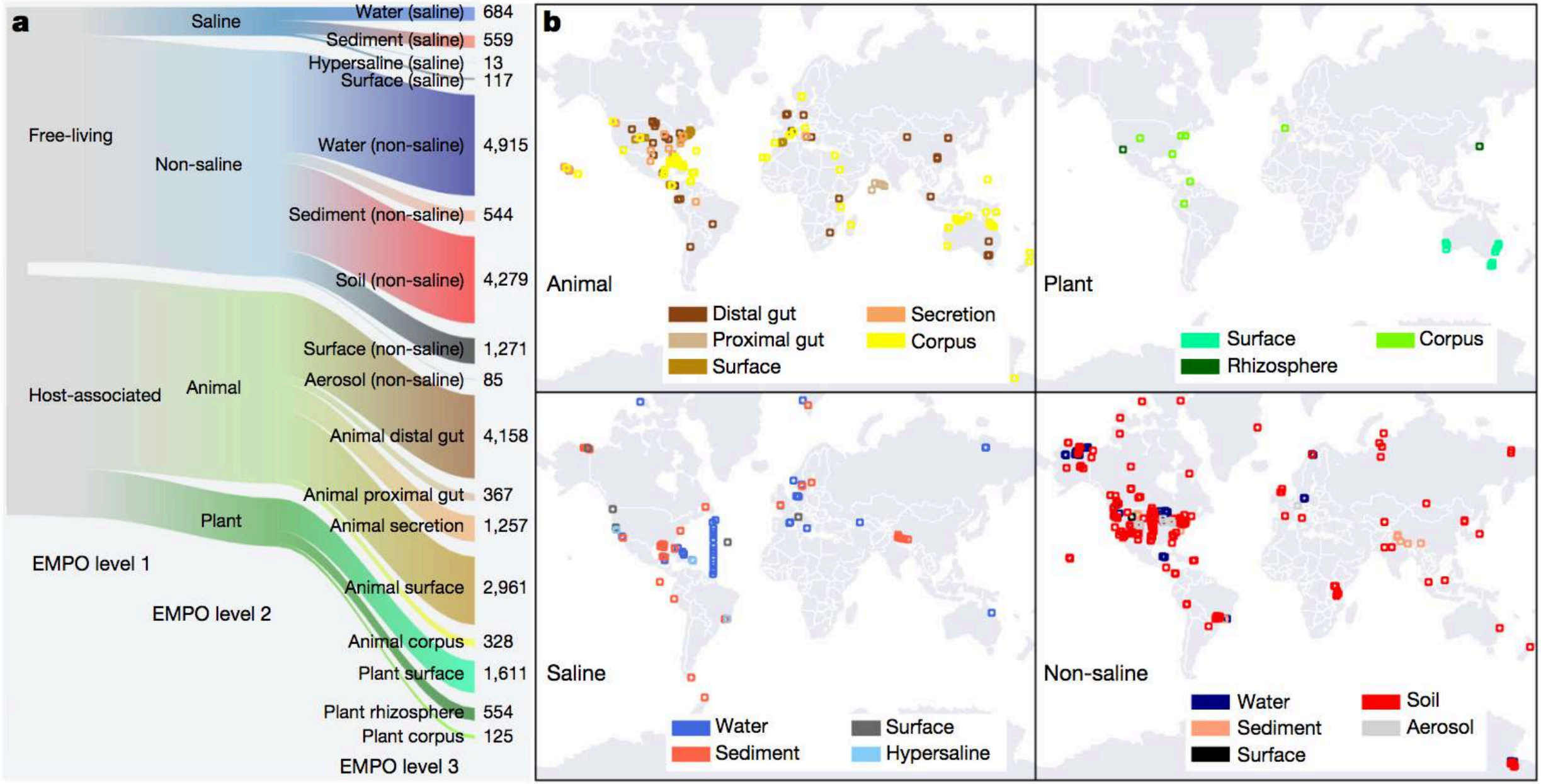
A communal catalogue reveals Earth's multiscale microbial diversity

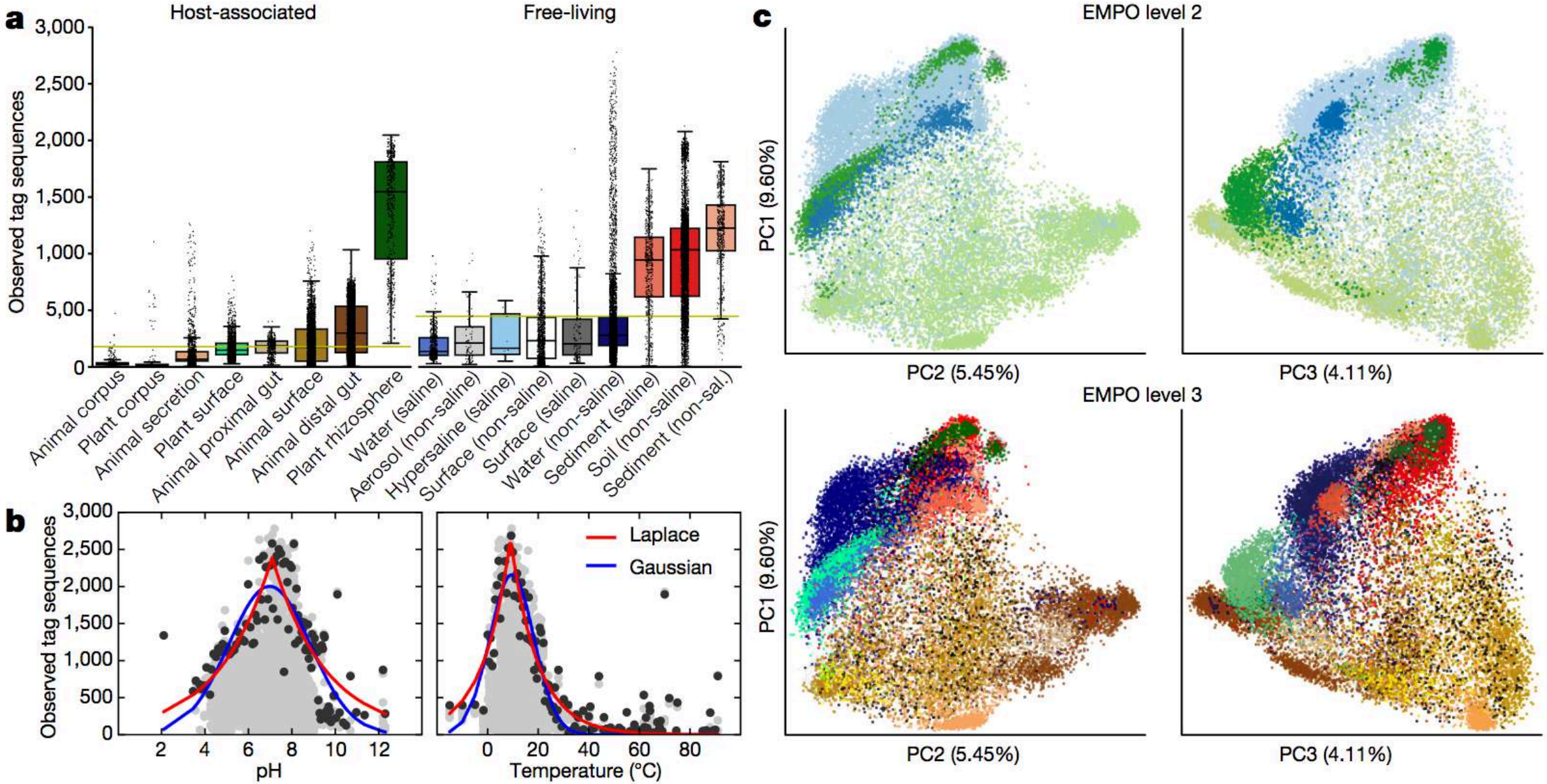
Luke R. Thompson[1,2,3], Jon G. Sanders[1], Daniel McDonald[1], Amnon Amir[1], Joshua Ladau[4], Kenneth J. Locey[5], Robert J. Prill[6], Anupriya Tripathi[1,7,8], Sean M. Gibbons[9,10], Gail Ackermann[1], Jose A. Navas-Molina[1,11], Stefan Janssen[1], Evguenia Kopylova[1], Yoshiki Vázquez-Baeza[1,11], Antonio González[1], James T. Morton[1,11], Siavash Mirarab[12], Zhenjiang Zech Xu[1], Lingjing Jiang[1,13], Mohamed F. Haroon[14], Jad Kanbar[1], Qiyun Zhu[1], Se Jin Song[1], Tomasz Kosciolek[1], Nicholas A. Bokulich[15], Joshua Lefler[1], Colin J. Brislawn[16], Gregory Humphrey[1], Sarah M. Owens[17], Jarrad Hampton-Marcell[17,18], Donna Berg-Lyons[19], Valerie McKenzie[20], Noah Fierer[20,21], Jed A. Fuhrman[22], Aaron Clauset[19,23], Rick L. Stevens[24,25], Ashley Shade[26,27,28], Katherine S. Pollard[4], Kelly D. Goodwin[3], Janet K. Jansson[16], Jack A. Gilbert[17,29], Rob Knight[1,11,30] & The Earth Microbiome Project Consortium*

Our growing awareness of the microbial world's importance and diversity contrasts starkly with our limited understanding of its fundamental structure. Despite recent advances in DNA sequencing, a lack of standardized protocols and common analytical frameworks impedes comparisons among studies, hindering the development of global inferences about microbial life on Earth. Here we present a meta-analysis of microbial community samples collected by hundreds of researchers for the Earth Microbiome Project. Coordinated protocols and new analytical methods, particularly the use of exact sequences instead of clustered operational taxonomic units, enable bacterial and archaeal ribosomal RNA gene sequences to be followed across multiple studies and allow us to explore patterns of diversity at an unprecedented scale. The result is both a reference database giving global context to DNA sequence data and a framework for incorporating data from future studies, fostering increasingly complete characterization of Earth's microbial diversity.

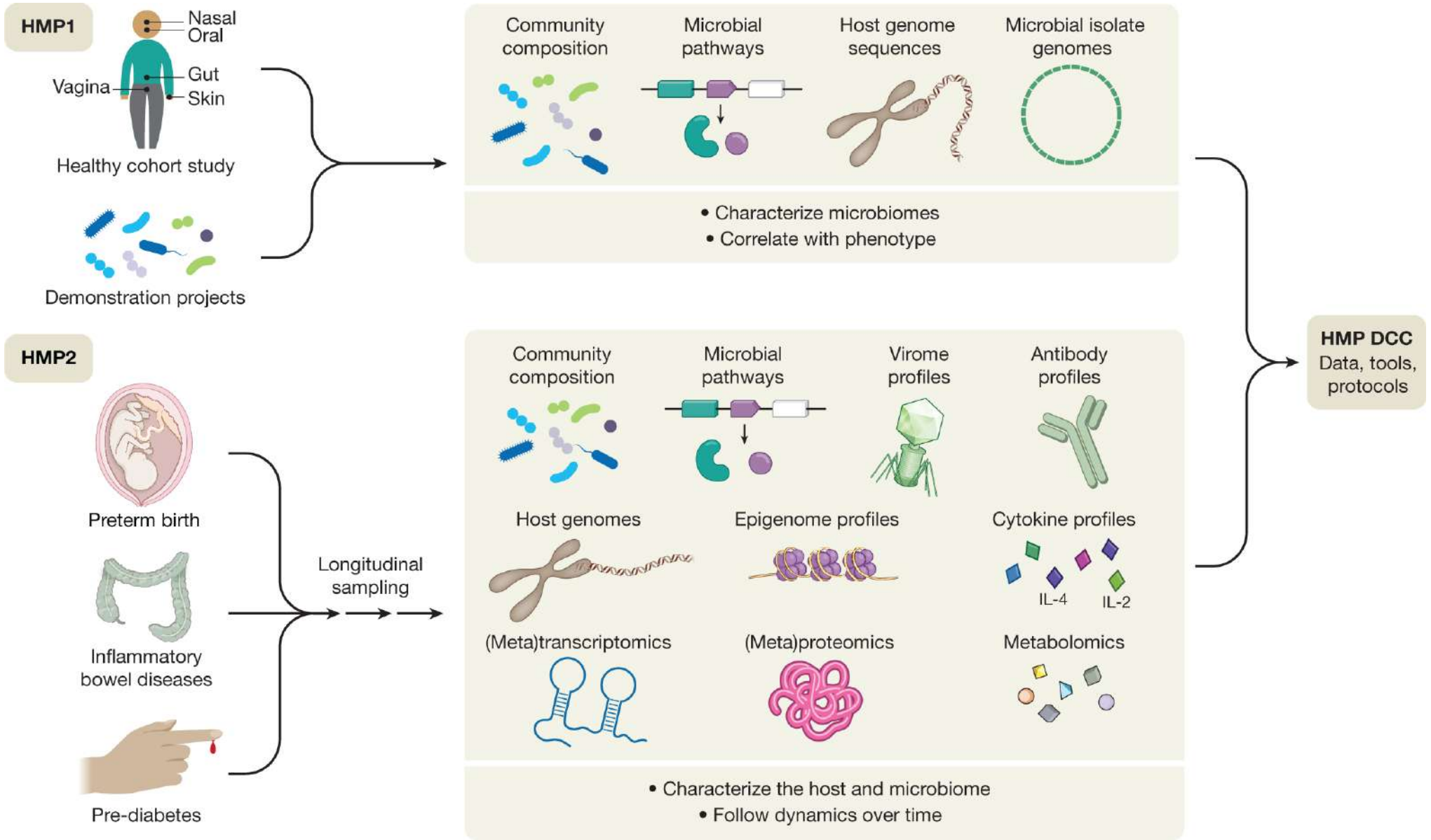
Received 13 March; accepted 10 October 2017.
Published online 1 November 2017.




a
Saline
Free-living
Non-saline
Host-associated
Animal
Plant
EMPO level 1
EMPO level 2
EMPO level 3
Water (saline) 684
Sediment (saline) 559
Hypersaline (saline) 13
Surface (saline) 117
Water (non-saline) 4,915
Sediment (non-saline) 544
Soil (non-saline) 4,279
Surface (non-saline) 1,271
Aerosol (non-saline) 85
Animal distal gut 4,158
Animal proximal gut 367
Animal secretion 1,257
Animal surface 2,961
Animal corpus 328
Plant surface 1,611
Plant rhizosphere 554
Plant corpus 125
b
Animal
Distal gut
Proximal gut
Surface
Secretion
Corpus
Plant
Surface
Rhizosphere
Corpus
Saline
Water
Sediment
Surface
Hypersaline
Non-saline
Water
Sediment
Surface
Soil
Aerosol


a
Host-associated
Free-living
Observed tag sequences
3,000
2,500
2,000
1,500
1,000
5,00
0
Animal corpus
Plant corpus
Animal secretion
Plant surface
Animal proximal gut
Animal surface
Animal distal gut
Plant rhizosphere
Water (saline)
Aerosol (non-saline)
Hypersaline (saline)
Surface (non-saline)
Surface (saline)
Water (non-saline)
Sediment (saline)
Soil (non-saline)
Sediment (non-sal.)
b
Observed tag sequences
3,000
2,500
2,000
1,500
1,000
500
0
2
4
6
8
10
12
pH
0
20
40
60
80
Temperature (°C)
Laplace
Gaussian
c
EMPO level 2
PC1 (9.60%)
PC2 (5.45%)
PC3 (4.11%)
EMPO level 3
PC1 (9.60%)
PC2 (5.45%)
PC3 (4.11%)

“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