

# NGS lecture

Isheng Jason Tsai

Chang Gung University 2019  
2019.03.14



This lecture is called “NGS”

## **Actually**

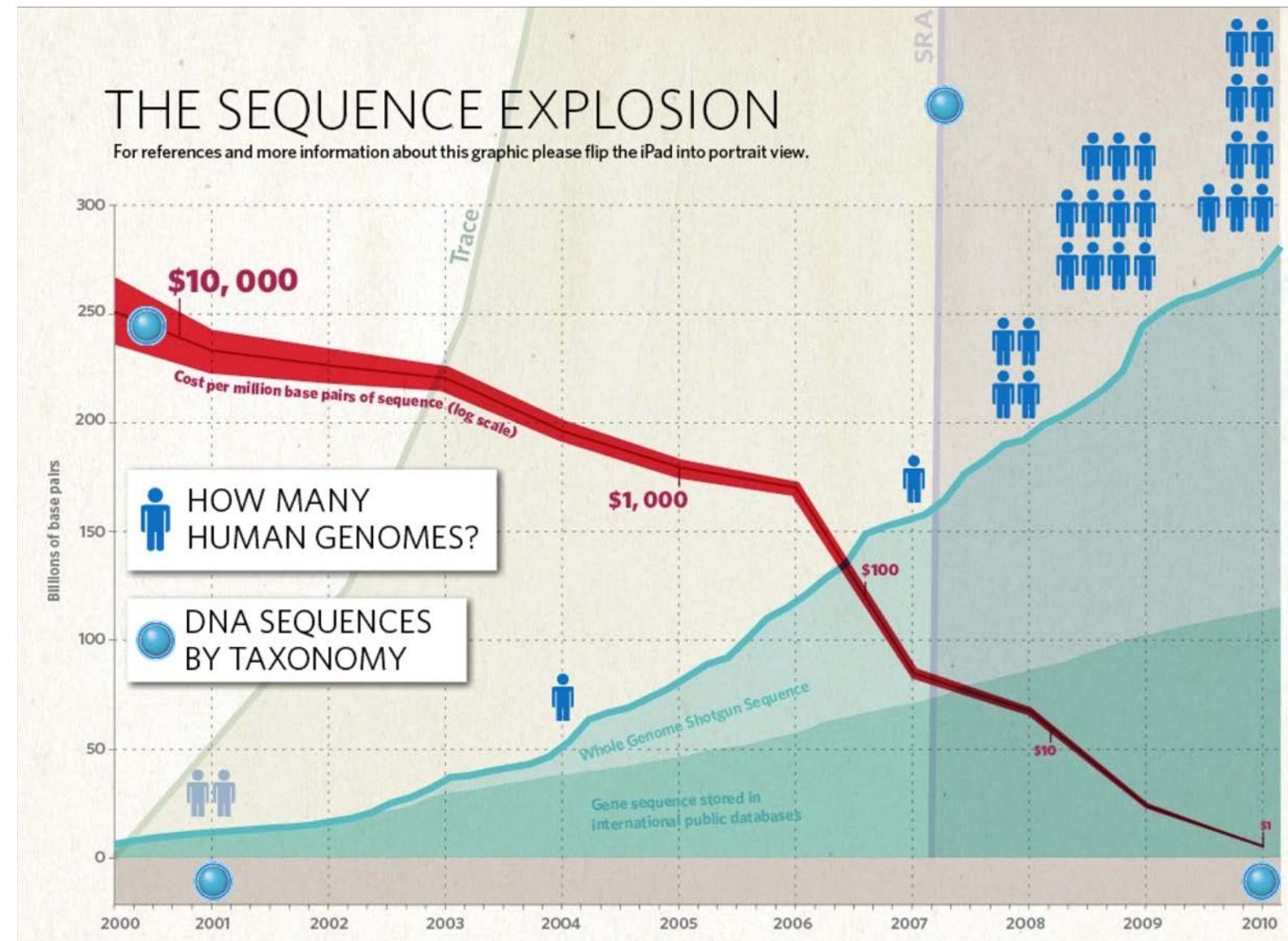
- Next Generation Sequencing is really “now” sequencing
- It won’t be so easy to tell you everything about NGS  
(it’s a bit like saying what can we do with PCR?)

## **Today we will cover**

- NGS: An introduction
- Genomics
- RNAseq
- Metagenomics

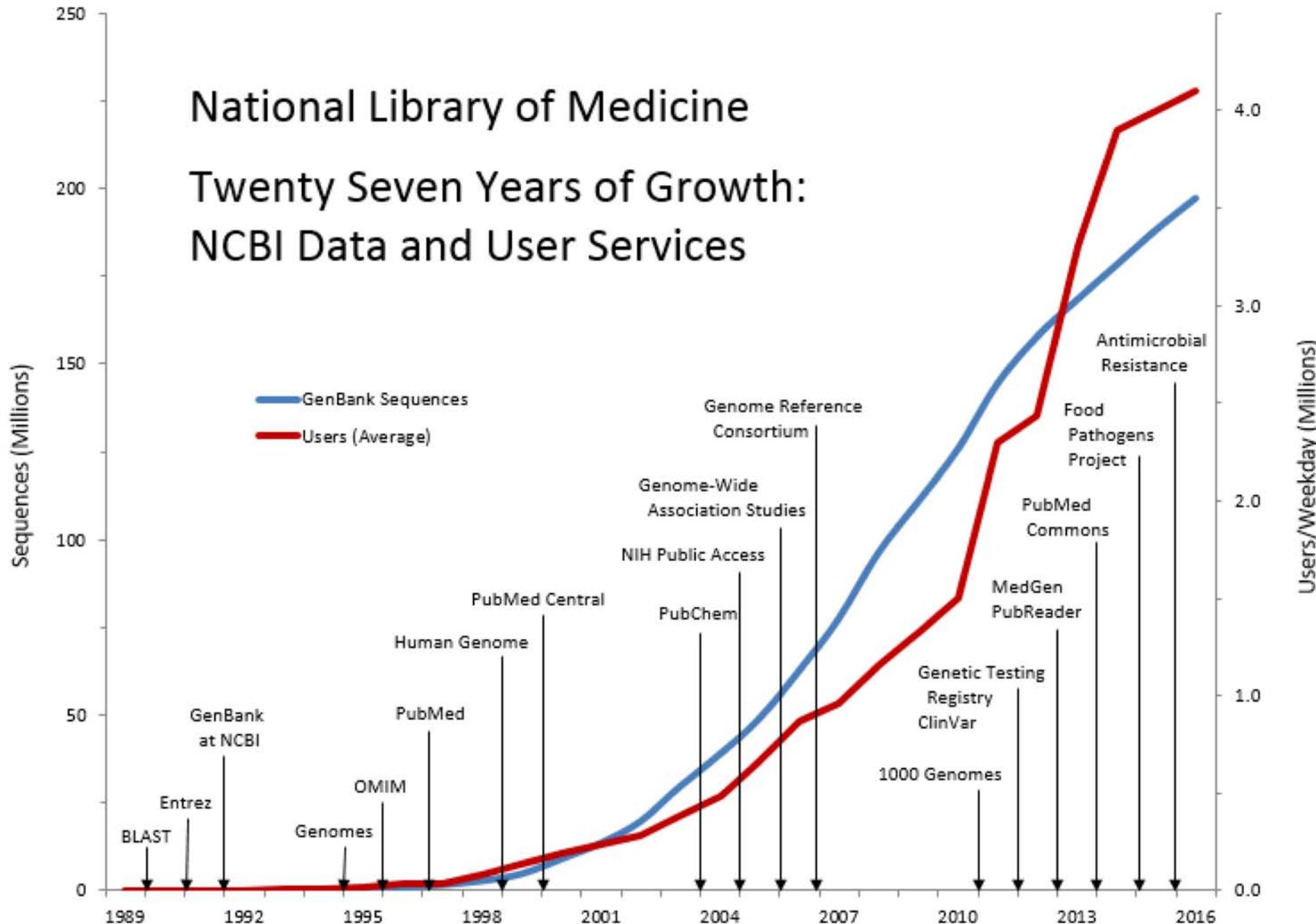
# What is NGS?

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



# National Library of Medicine

## Twenty Seven Years of Growth: NCBI Data and User Services



**NGS = sequencing made cheaper, faster and higher throughput**

# What we will cover today

NGS: Some basics

Sequencing platforms

Data types

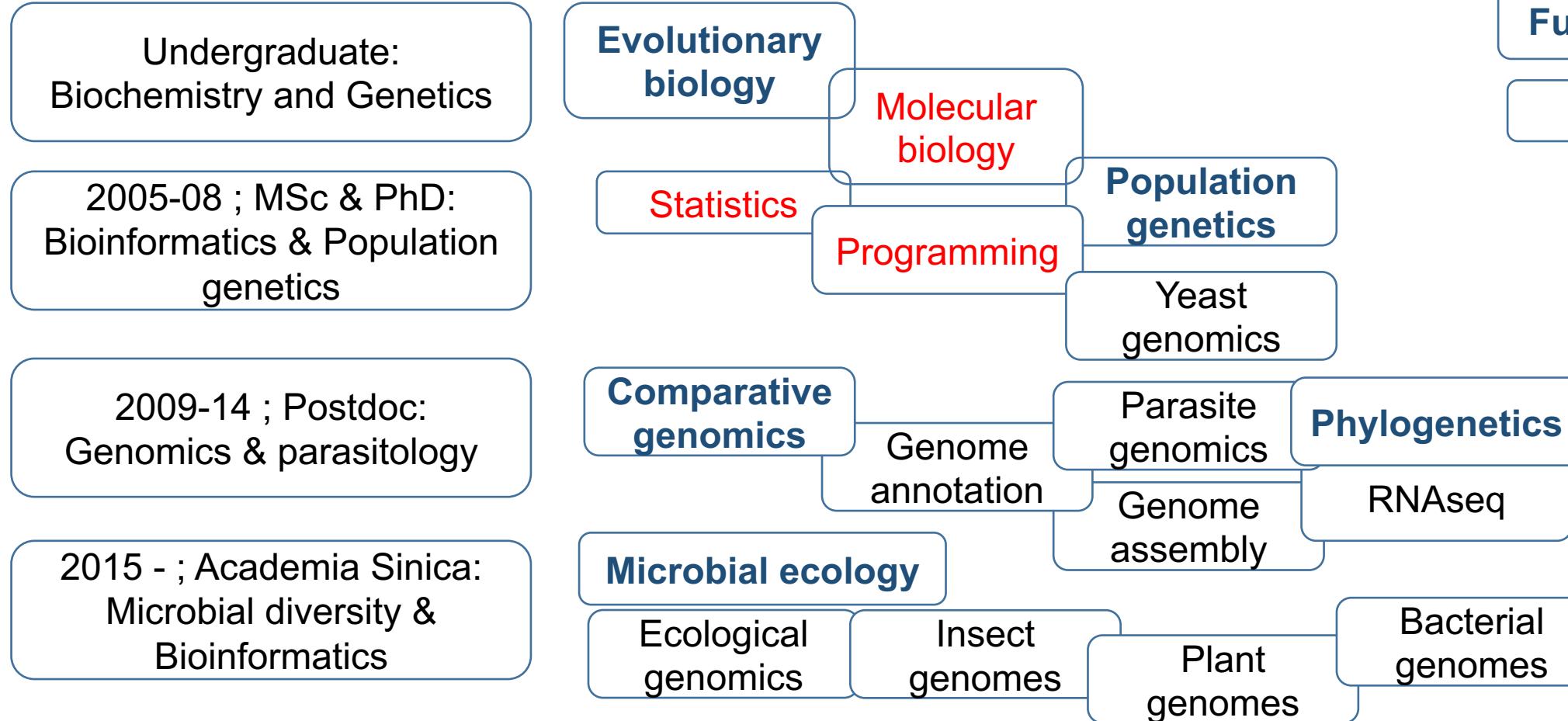
Analysis:

- RNAseq
- 16S
- Metagenomics

Previous questions:

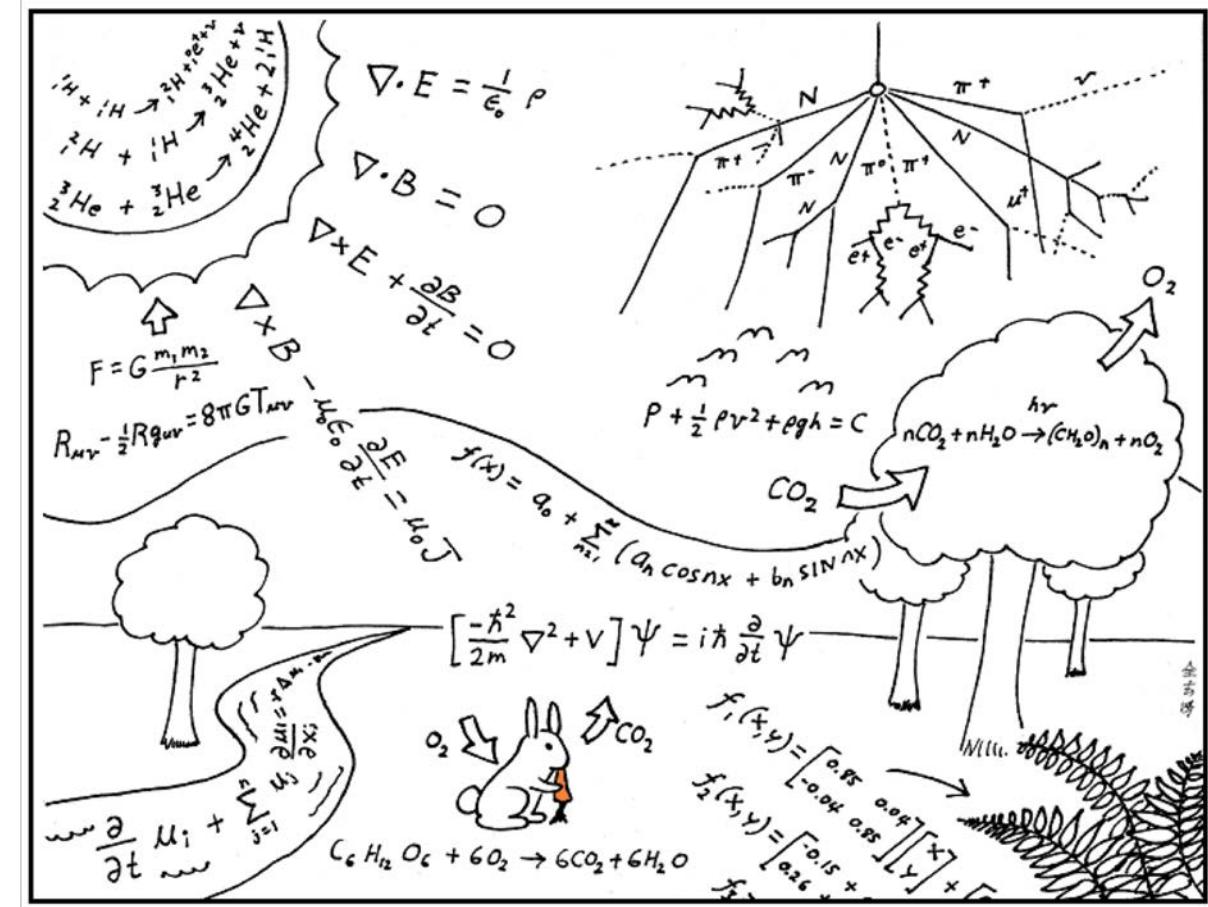
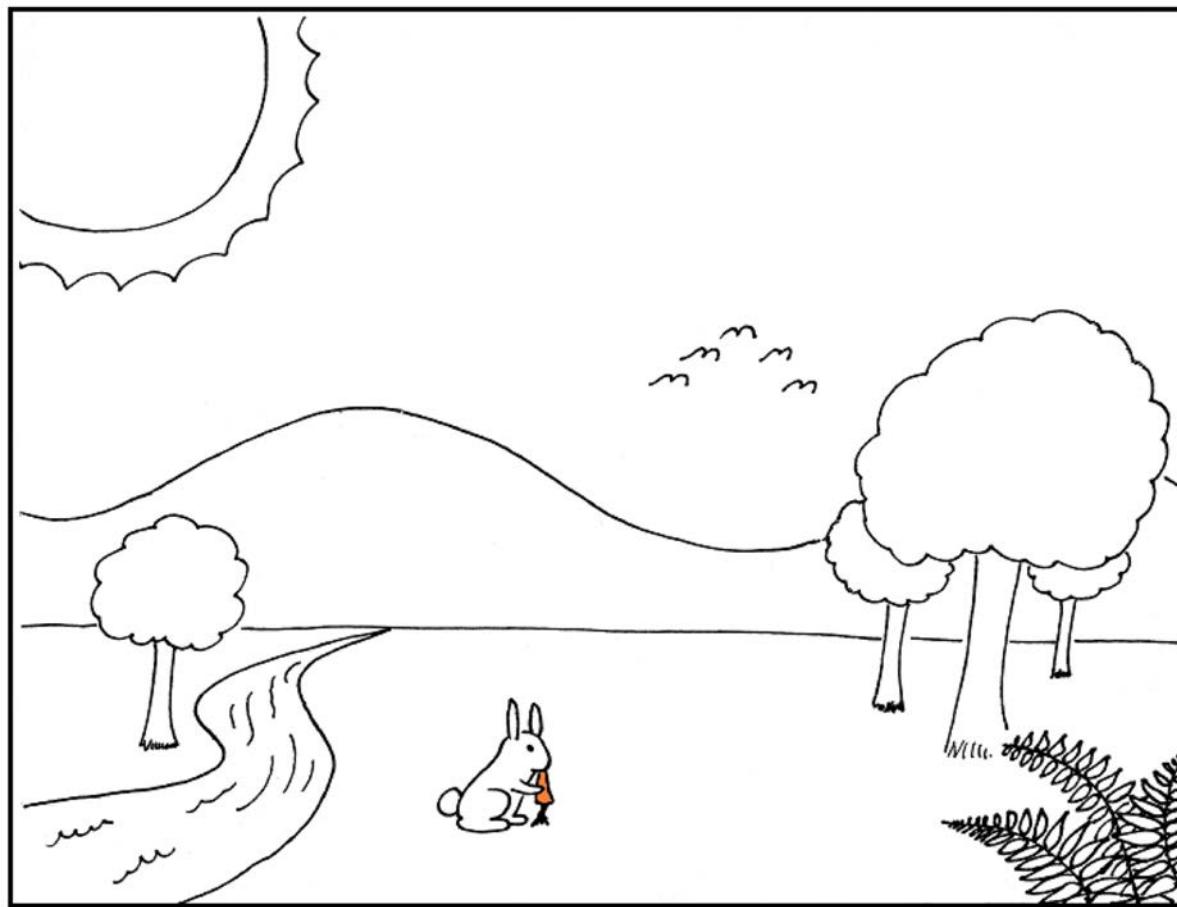
- microbiota的paper要怎麼approach
- 16S sequencing region primer choice
- microbiota醫院有fecal transplantation計劃？

# My background



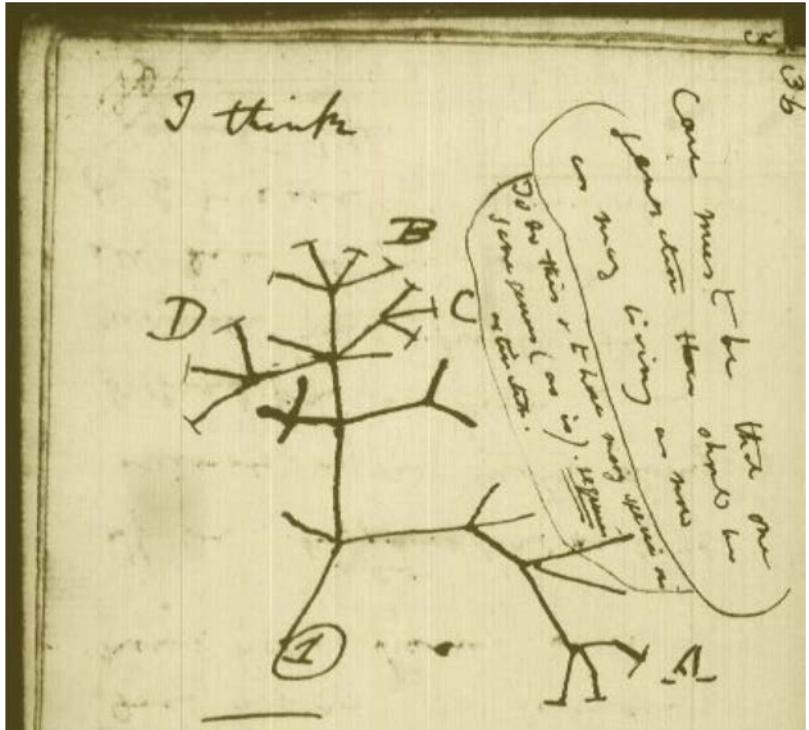
- **43 publications**  
(2 Nature, 1 Science, 2 Nature Genetics, 2 PNAS,  
3 Genome Biology, 1 Nature Plant, 1 Nature Communications)

# This is how scientists see the world



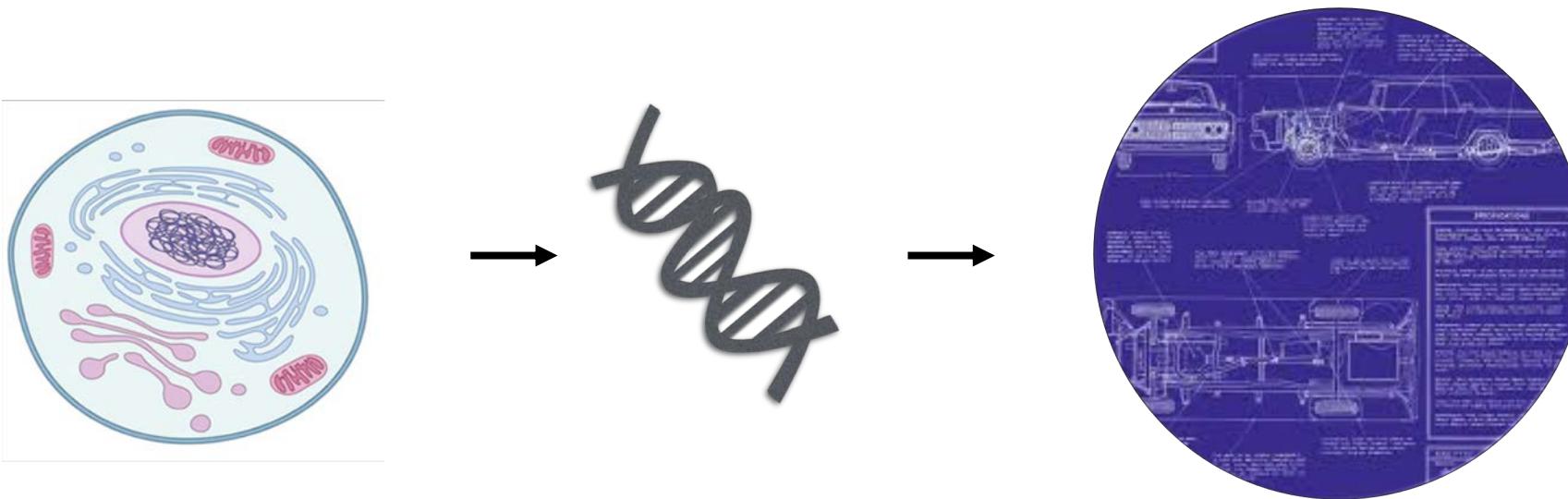
# Nothing makes sense in the light of evolution

Theodosius Dobzhansky 1973



# What is genomics?

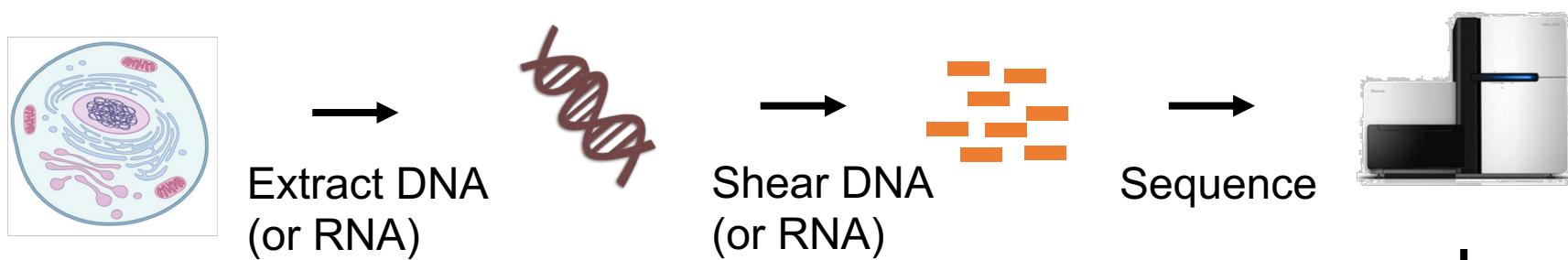
# Genome



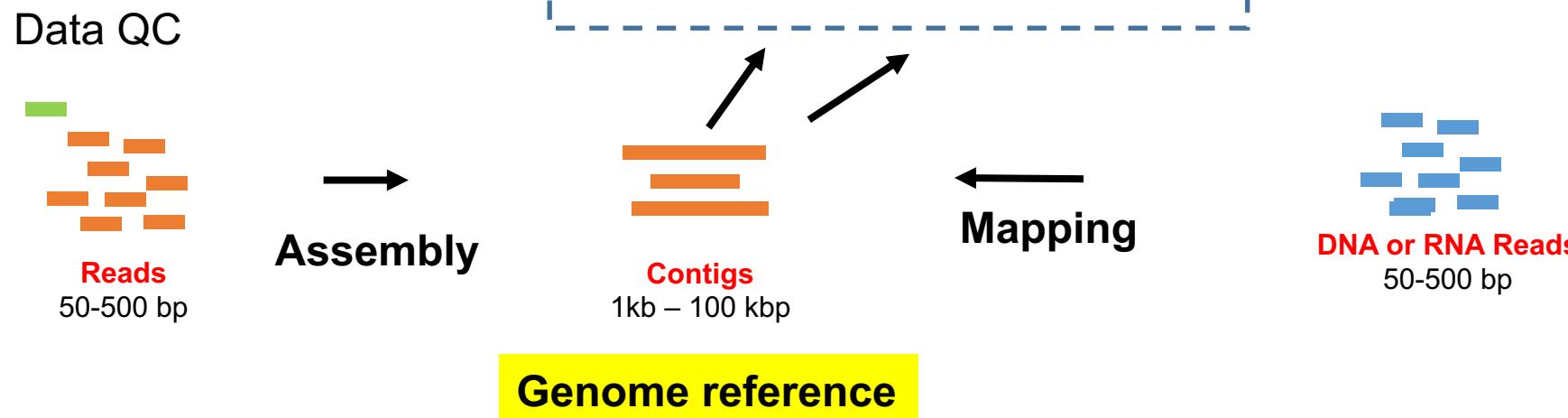
Genome = Parts list of a single genome

# A genome project

## Wet lab work



## Bioinformatics



# 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

# Many perceptions of NGS / genomics



What my parents  
think I do



What less friendly colleagues  
think we do



What more friendly colleagues  
think we do



What my friends  
think I do



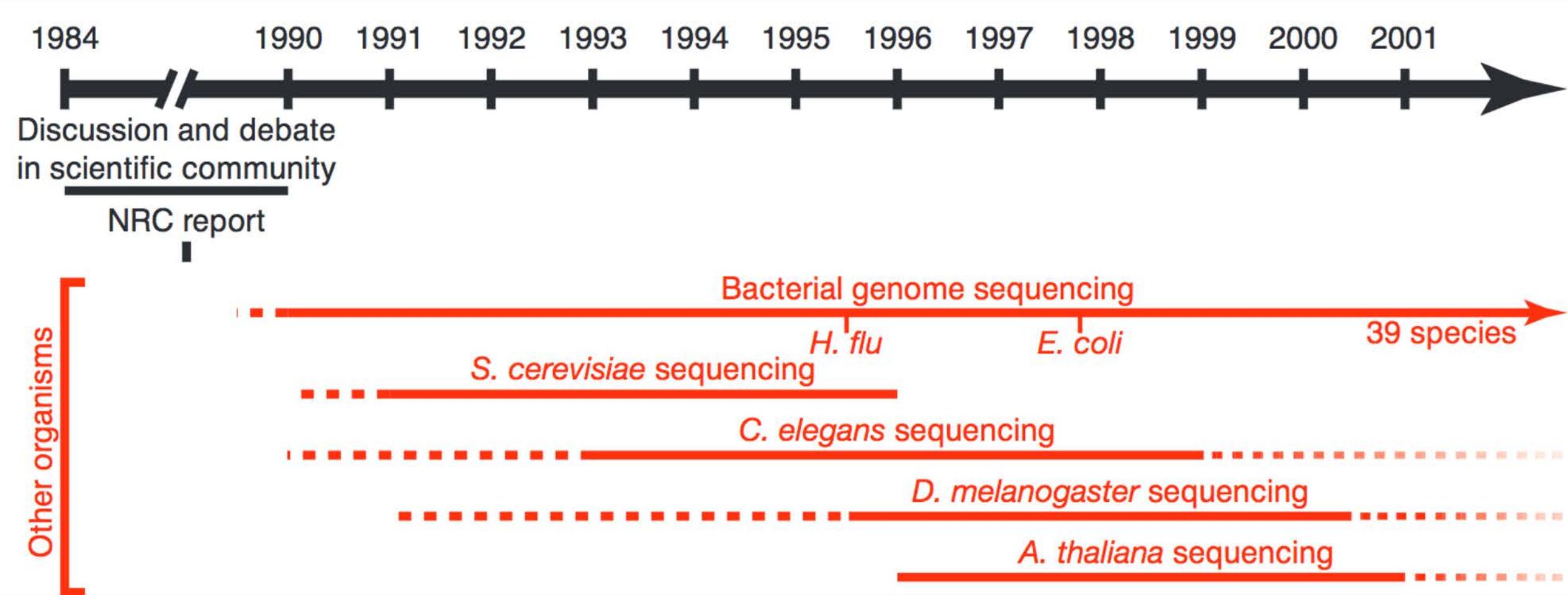
What we  
think we do

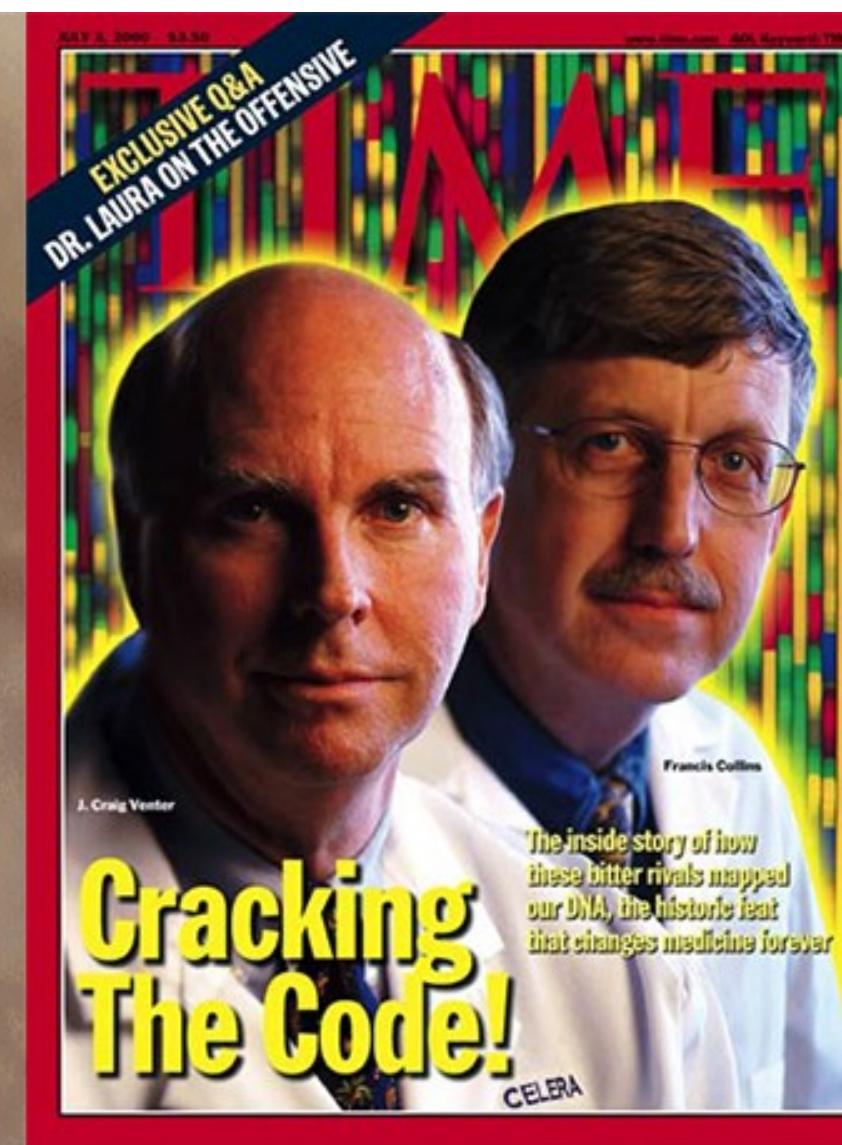


What we  
actually do

# Why sequence a genome?

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





# 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/>

## Large-scale whole-genome sequencing of the Icelandic population



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

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



The blood of a thousand Icelanders.  
Photo: Chris Lund



# 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.

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

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

	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 <sup>1</sup>	4,735,722	4,665,684	24,332 (283)	780 (63)
<b>Variant Type<sup>1</sup></b>				
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)
<b>Functional Prediction</b>				
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



# 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

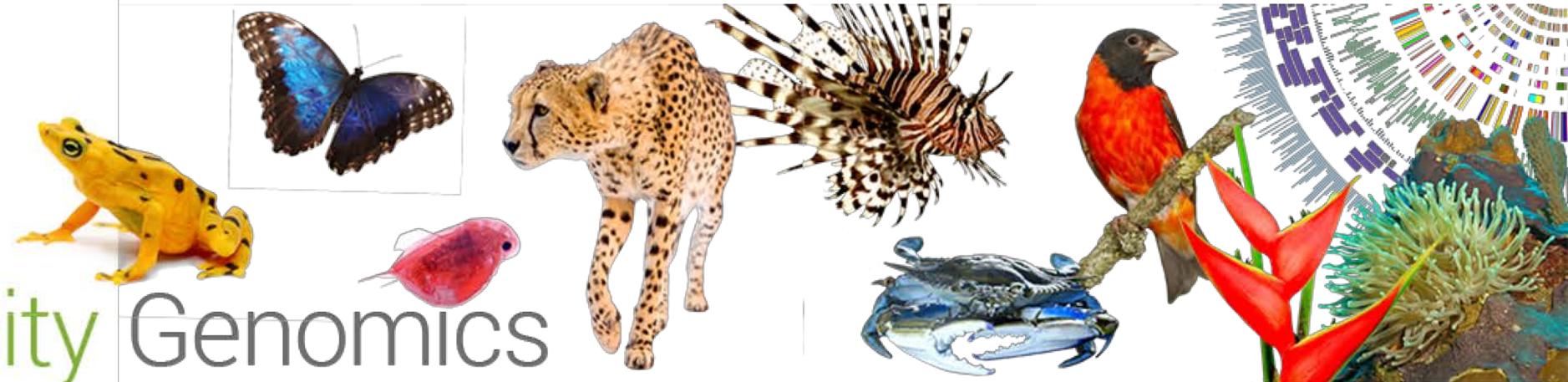
完成第二輪追蹤個案總數



Smithsonian

Institute for

Biodiversity Genomics

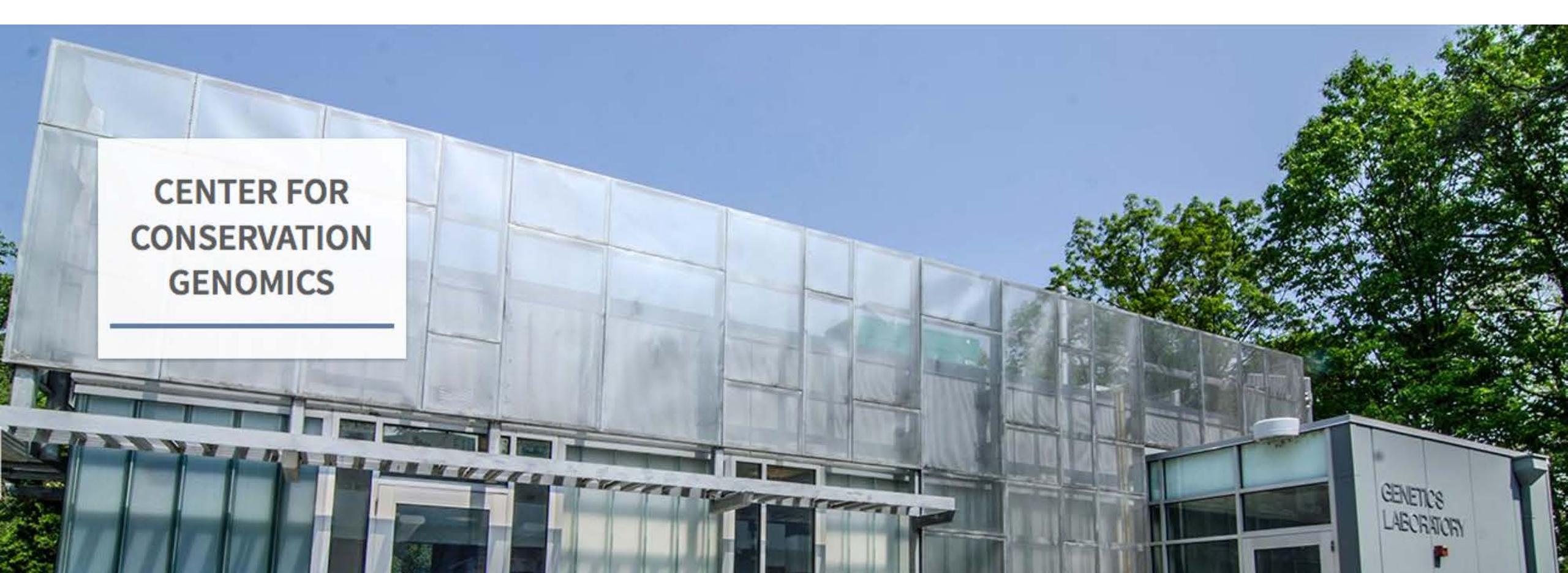


## How do we sustain life on our changing planet?

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

But our biodiversity faces serious challenges.

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



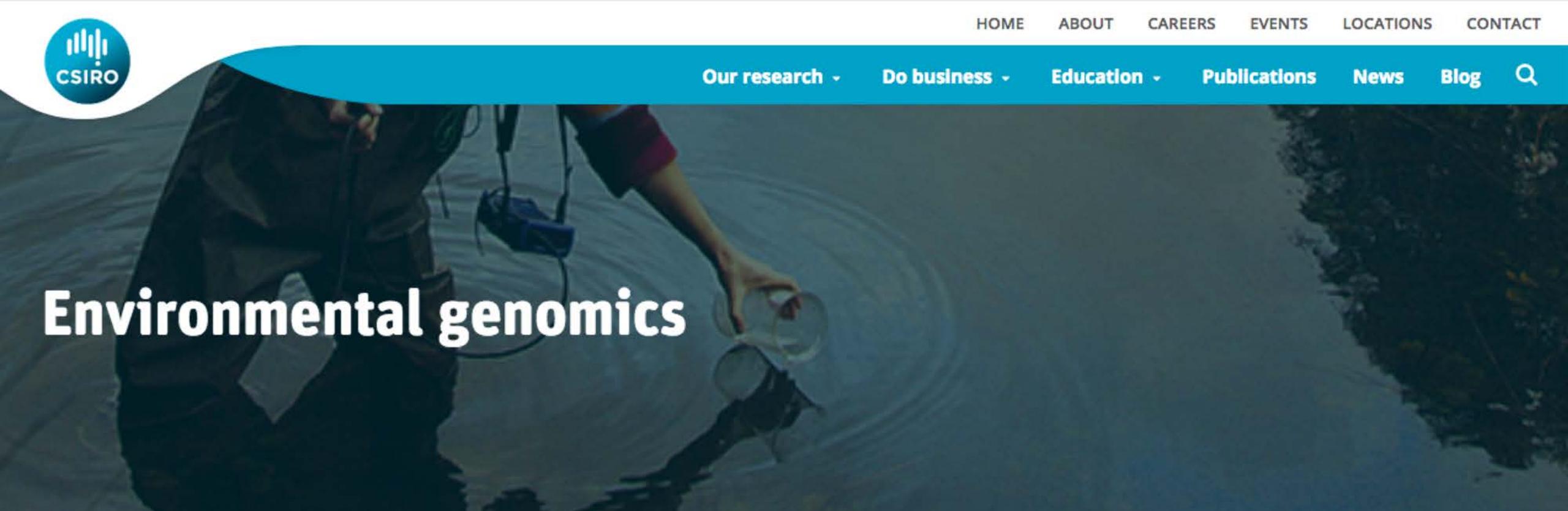
## CENTER FOR CONSERVATION GENOMICS

---

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



# Environmental genomics

A photograph showing a person from the waist down, wearing a dark wetsuit and fins, crouching near a body of water. They are holding a clear plastic container with both hands, seemingly collecting a water sample. The background is a blurred outdoor environment.

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



EARTH BIOGENOME PROJECT

# Sequencing Life for the Future of Life

## A GRAND CHALLENGE

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

## A GRAND VISION

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

# 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**

# More data but less people with informatics skills

- Sequencing is the result of many types of experiment
- Everyone wants to make use of this technology
- Not everyone will be able analyse them
  - You can't just open the file in Microsoft office anymore
- Collaborate or learn yourself
- **Bottleneck is bioinformatics analysis**

Different sequencing platforms /  
History of sequencing

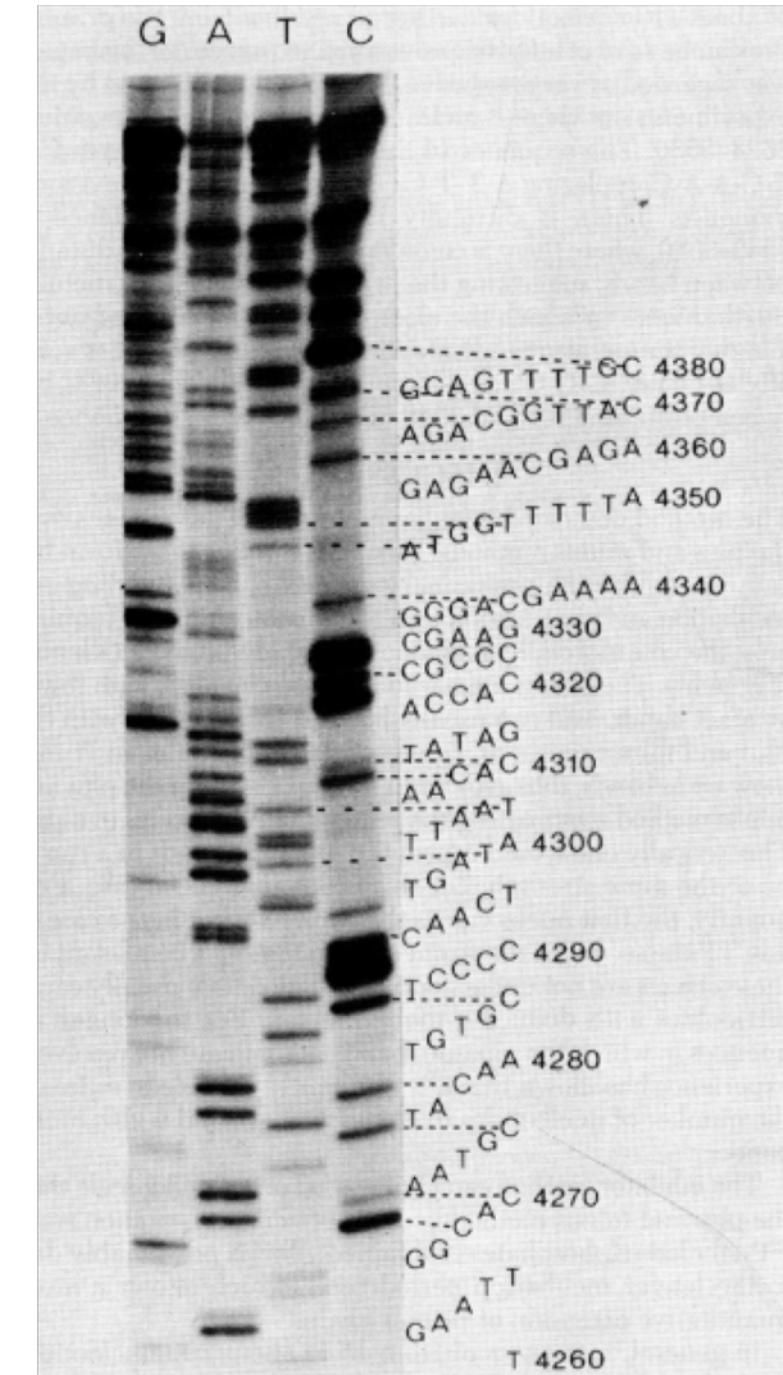
# DNA sequencing with chain-terminating inhibitors

(DNA polymerase/nucleotide sequences/bacteriophage  $\phi$ X174)

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

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

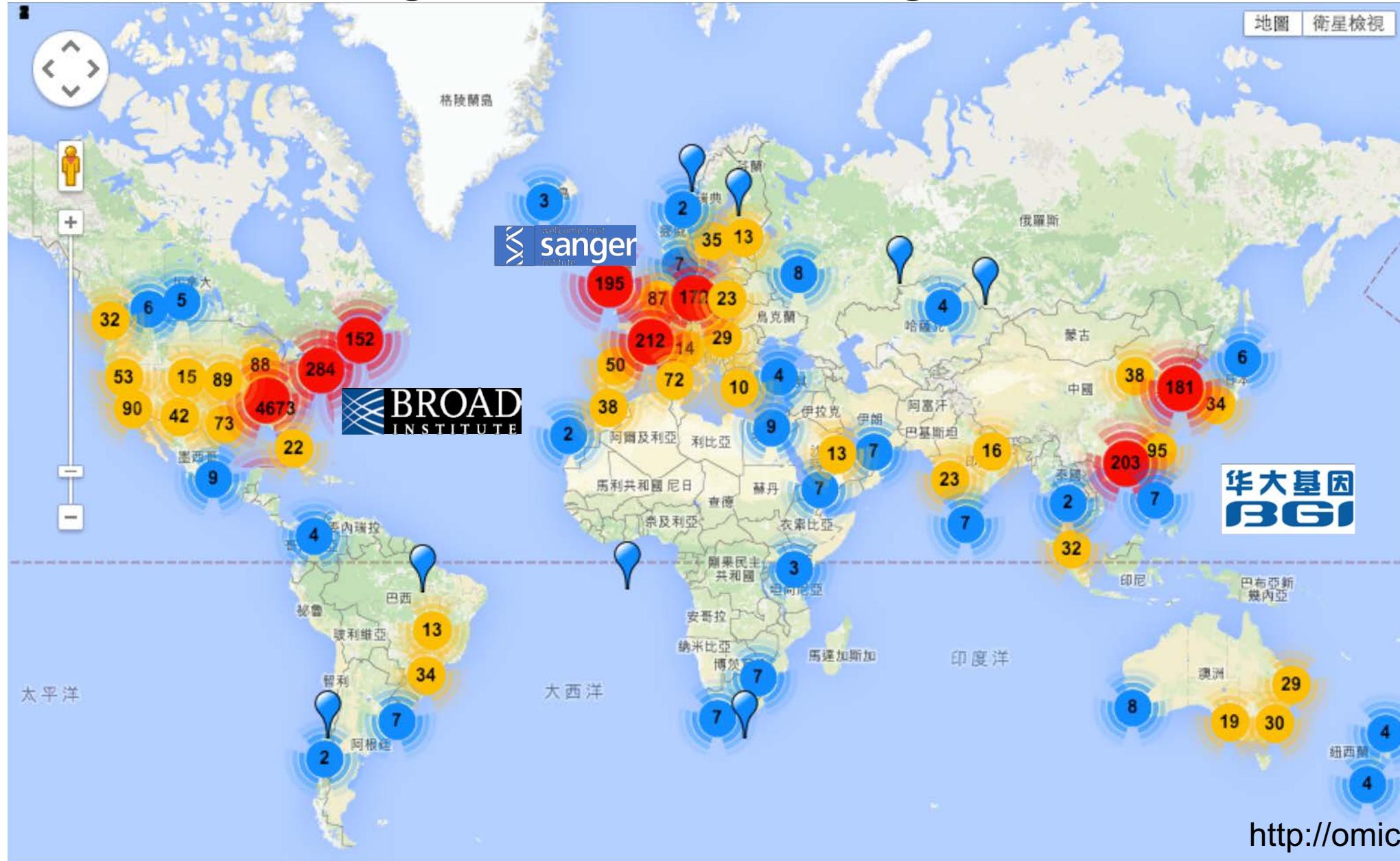
Contributed by F. Sanger, October 3, 1977



# ABI 3730xi at TIGR

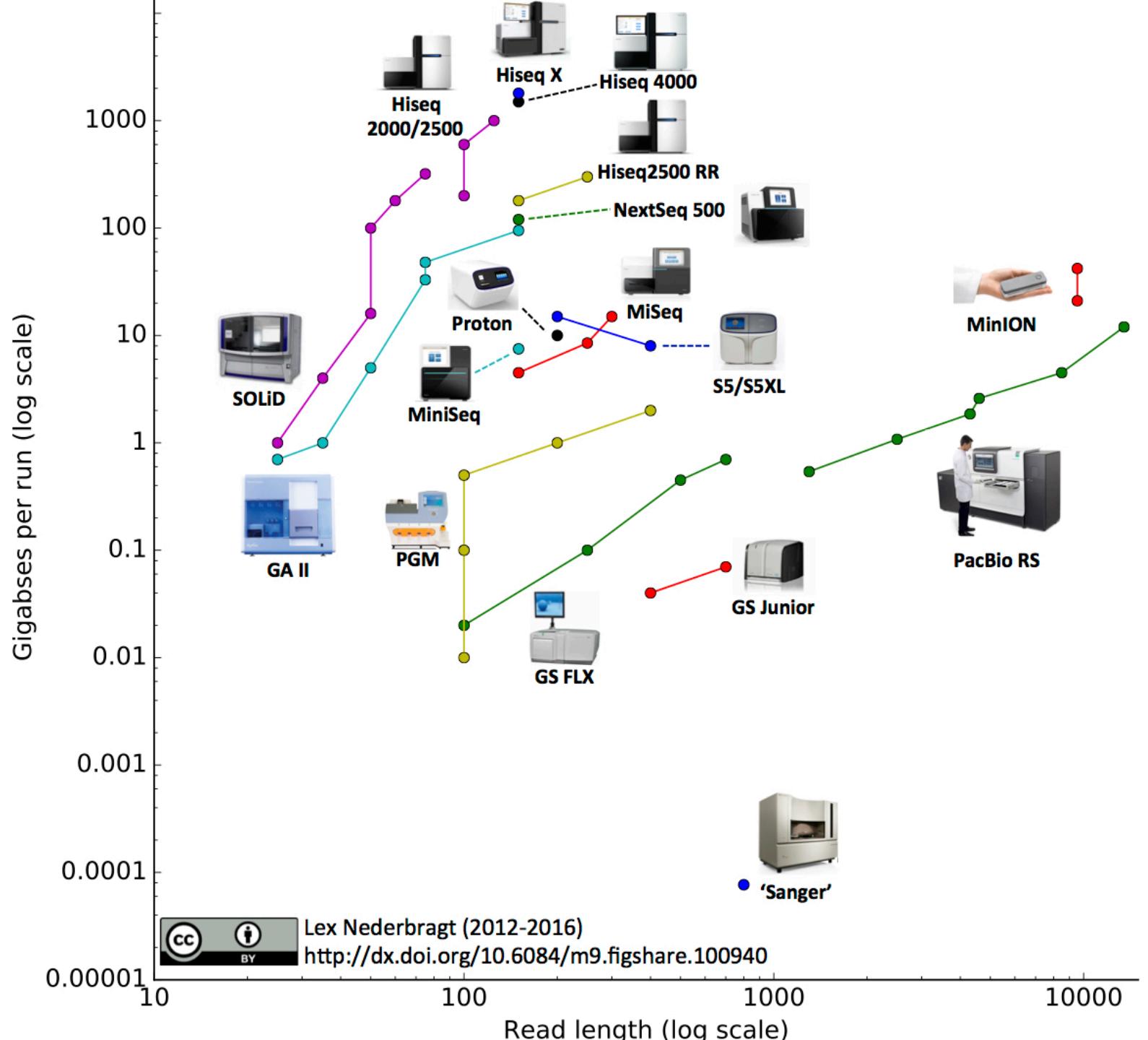


# World competing for sequencing power



# Sequencing Platforms

- Short reads
  1. ~~Genome Analyzer IIx (GAIIx) – Illumina~~
  2. HiSeq, MiSeq, Novaseq – Illumina
- Long reads
  1. ~~Genome Sequencer FLX System (454) – Roche~~
  2. Pacific Bioscience
  3. Oxford Nanopore



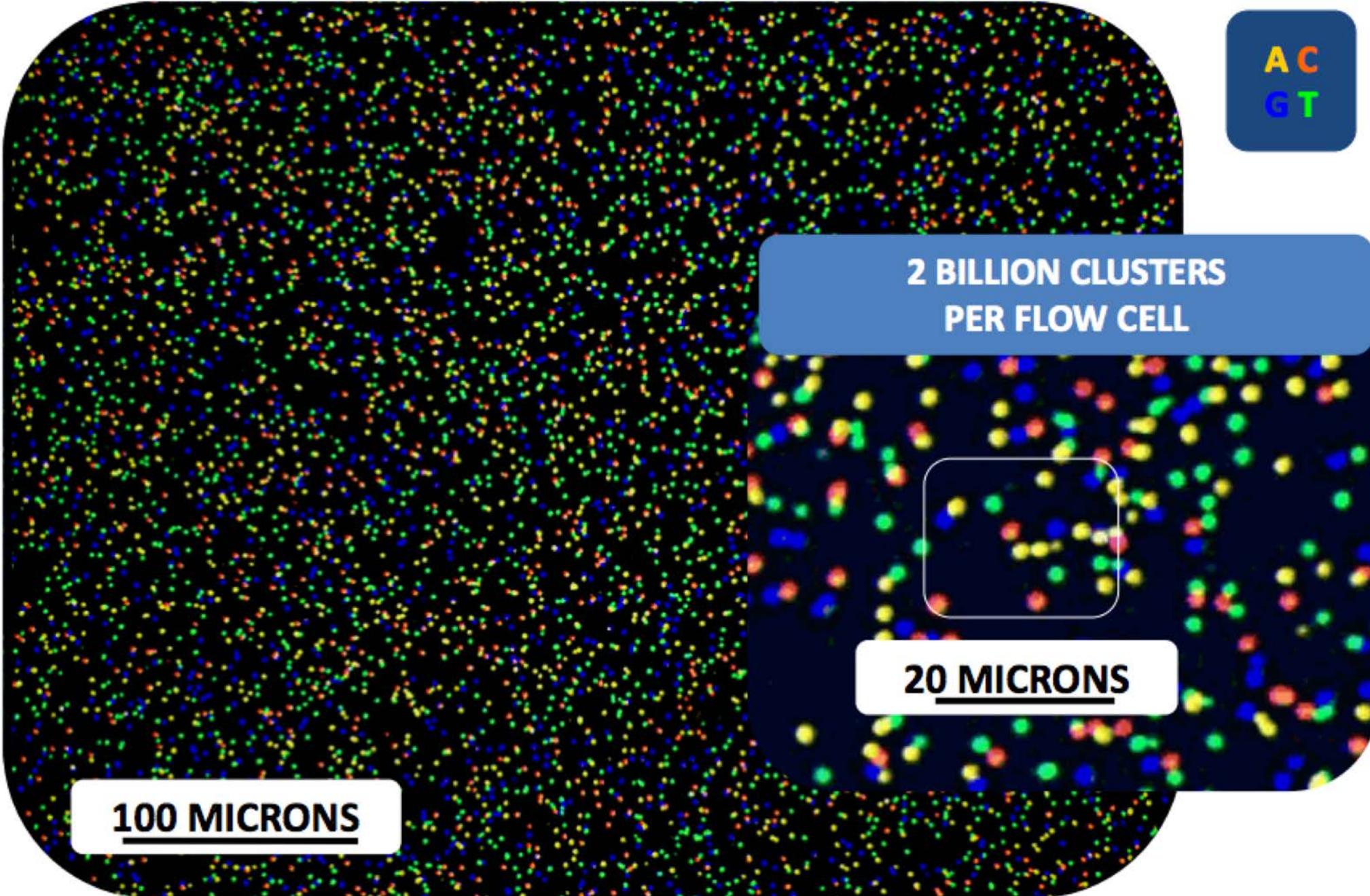
		Read length: (paired-end*, Half of data in reads**)								
Platform	Reads x run: (M)		Run time: (d)	Yield: (Gb)	Rate: (Gb/d)	Reagents: (\$K)	per-Gb: (\$)	hg-30x: (\$)	Machine: (\$)	
iSeq 100 1fcell	4	150*	0.77	1.2	1.56	0.625	521	62500	19.9K	
MiniSeq 1fcell	25	150*	1	7.5	7.5	1.75	233	28000	49.5K	
MiSeq 1fcell	25	300*	2	15	7.5	1	66	8000	99K	
NextSeq 550 1fcell	400	150*	1.2	120	100	5	50	5000	250K	
HiSeq 2500 RR 2fcells	600	100*	1.125	120	106.6	6.145	51.2	6144	740K	
HiSeq 2500 V3 2fcells	3000	100*	11	600	55	23.47	39.1	4692	690K	
HiSeq 2500 V4 2fcells	4000	125*	6	1000	166	29.9	31.7	3804	690K	
HiSeq 4000 2fcells	5000	150*	3.5	1500	400	--	20.5	2460	900K	
HiSeq X 2fcells	6000	150*	3	1800	600	--	7.08	849.6	1M	
NovaSeq S1 2018 2fcells	3300	150*	1.66	1000	600	--	18	1800	999K	
NovaSeq S2 2fcells	6600	150*	1.66	2000	1200	--	15	1564	999K	
NovaSeq S4 2fcells	20000	150*	1.66	6000	3600	64	5.8	700	999K	
5500 XL	1400	60	7	180	30	10.5	58.33	7000	595K	
Ion S5 510 1chip	2 - 3	200 400	0.21	1	4.8	0.95	950	114000	65K	
Ion S5 520 1chip	3 - 6	200 400 600	0.23	1	4.3	1	500	60000	65K	
Ion S5 530 1chip	20	200 400 600	0.29	4	13.8	1.2	150	18000	65K	
Ion S5 540 1chip	80	200	0.42	15	35.7	1.4	93.3	11196	65k	
Ion S5 550 1chip	130	200	0.5	25	50	1.67	66.8	8016	65k	
PacBio RSII P6-C4 16cells	0.88	20K**	4.3	12	2.8	2.4	200	24000	695K	
PacBio Sequel 16cells 2018	6.4	33K**	6.6	160	24.2	--	80	9600	350K	
PacBio R&D end 2018	--	32K**	--	192	--	1	6.6	1000	350K	
SmidgION 1fcell	--	--	TBC	TBC	TBC	TBC	TBC	--	--	
Flongle 1fcell	--	--	0.7	1-3.3	--	--	90-30	--	--	
MinION R9.5.1 1fcell	--	--	2	17-40	--	--	30-12.5	--	--	
GridION X5 5fcells	--	--	2	85-200	--	--	17.5-7.5	--	--	
PromethION RnD 48fcells	--	--	2	20000	--	--	43136	--	--	
QiaGen GeneReader	400	--	--	80	--	0.5	--	--	--	
BGISEQ 500	1600	100*	7	260	37.14285714	--	--	600?	500K	
BGISEQ 50	1600	50*	0.4	8	20	--	--	--	--	
MGISEQ 2000	--	100*	2	600	300	4.8	8	960	310K	
MIGSEQ 200	--	100*	--	60	--	--	--	--	150K	

# Illumina HiSeq



# Sequencing by synthesis

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

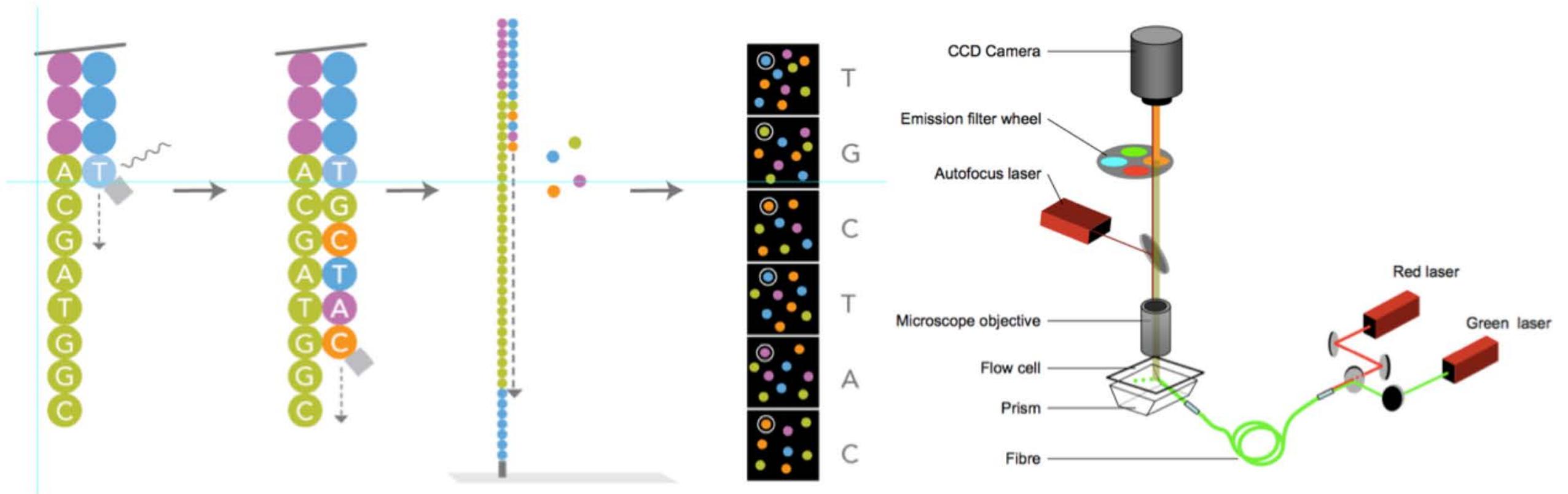


AC  
G T

**100 MICRONS**

**2 BILLION CLUSTERS  
PER FLOW CELL**

**20 MICRONS**



# HiSeq and MiSeq

## HiSeq 2000

Initially capable of up to 600Gb per run in 13 days.  
Cost of resequencing one human genome:  
30x coverage about \$6,000- \$9,000



## HiSeq 2500

Initially capable of up 100Gb per run in 27hours.  
Cost per genome - ???

## MiSeq

- Small capacity system. PE 2x250cycles in 24 hours.
- Long insert size possible: 1.5kb, 3kb
- 2x400bp in R&D



# Novaseq



# Illumina platform comparison

Platform	Reads x run: (M)	Read length: (paired-end*, Half of data in reads**)	Run time: (d)	Yield: (Gb)	Rate: (Gb/d)	Reagents: (\$K)	per-Gb: (\$)	hg-30x: (\$)	Machine: (\$)
iSeq 100 1fcell	4	150*	0.77	1.2	1.56	0.625	521	62500	19.9K
MiniSeq 1fcell	25	150*	1	7.5	7.5	1.75	233	28000	49.5K
MiSeq 1fcell	25	300*	2	15	7.5	1	66	8000	99K
NextSeq 550 1fcell	400	150*	1.2	120	100	5	50	5000	250K
HiSeq 2500 RR 2fcells	600	100*	1.125	120	106.6	6.145	51.2	6144	740K
HiSeq 2500 V3 2fcells	3000	100*	11	600	55	23.47	39.1	4692	690K
HiSeq 2500 V4 2fcells	4000	125*	6	1000	166	29.9	31.7	3804	690K
HiSeq 4000 2fcells	5000	150*	3.5	1500	400	--	20.5	2460	900K
HiSeq X 2fcells	6000	150*	3	1800	600	--	7.08	849.6	1M
NovaSeq S1 2018 2fcells	3300	150*	1.66	1000	600	--	18	1800	999K
NovaSeq S2 2fcells	6600	150*	1.66	2000	1200	--	15	1564	999K
NovaSeq S4 2fcells	20000	150*	1.66	6000	3600	64	5.8	700	999K

# Third generation sequencing



# PacBio (Pacific Biosciences)



RSII

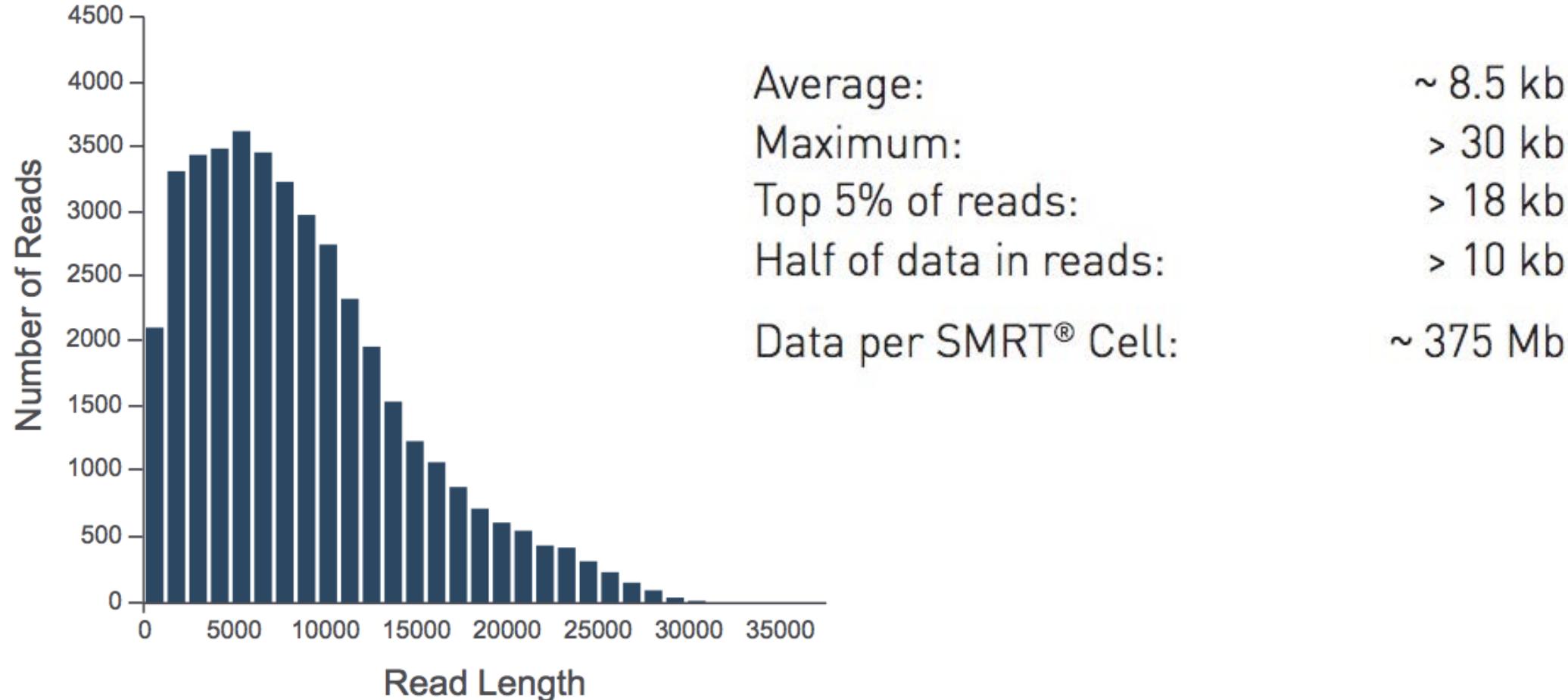


Sequel

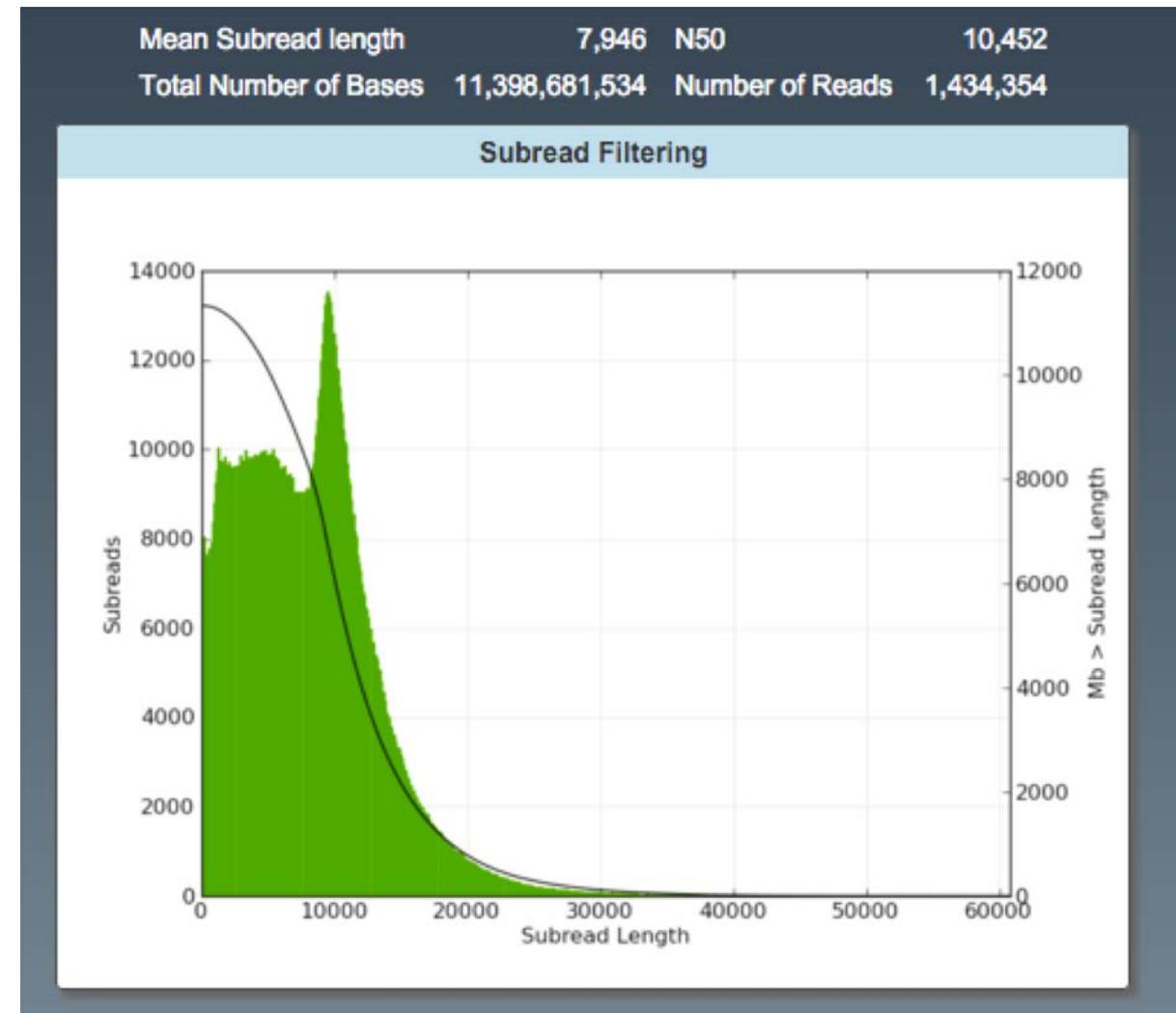
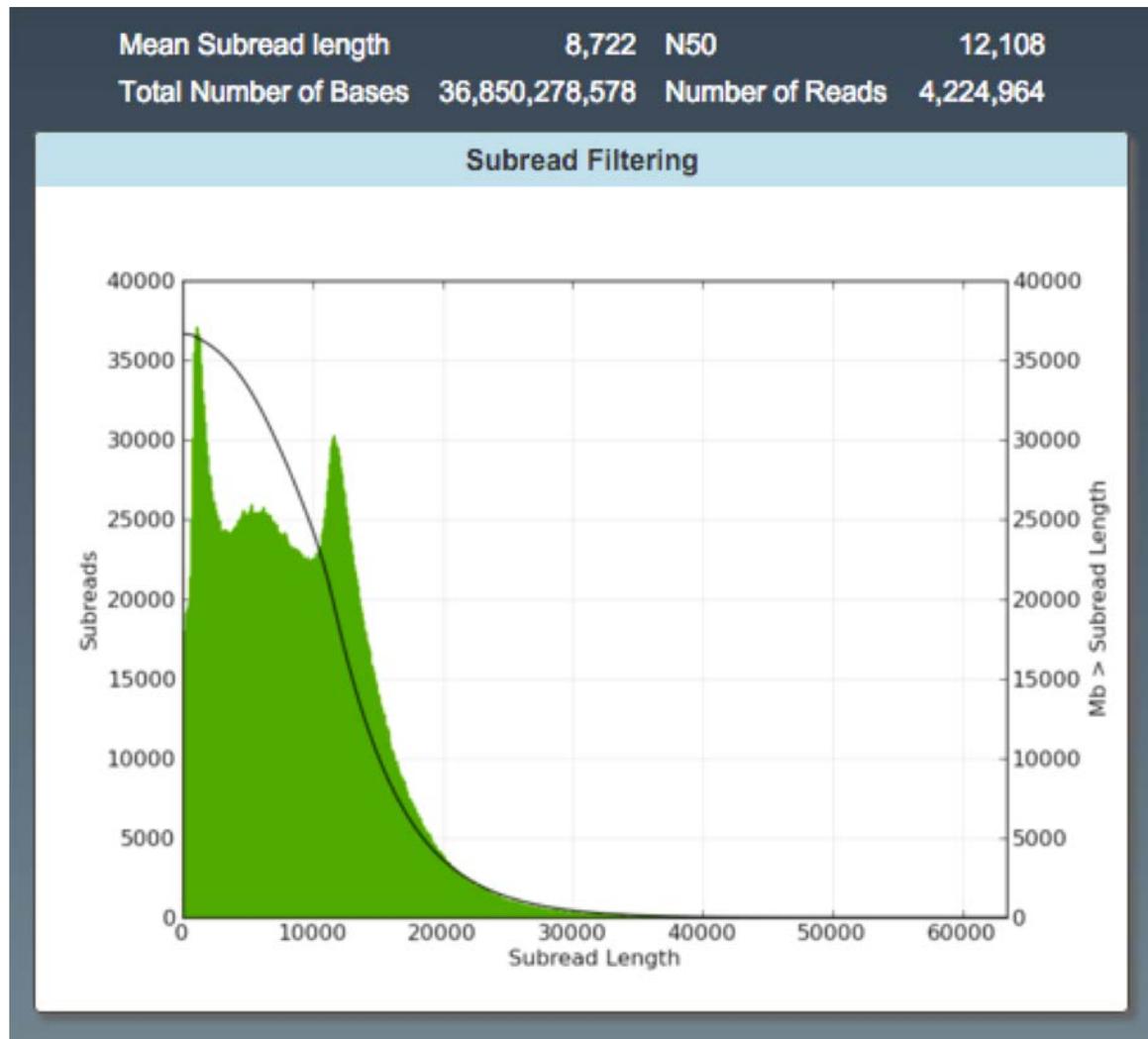
# Single molecule sequencing

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

# PacBio (Pacific Biosciences)



# PacBio (Pacific Biosciences)



# Oxford Nanopore

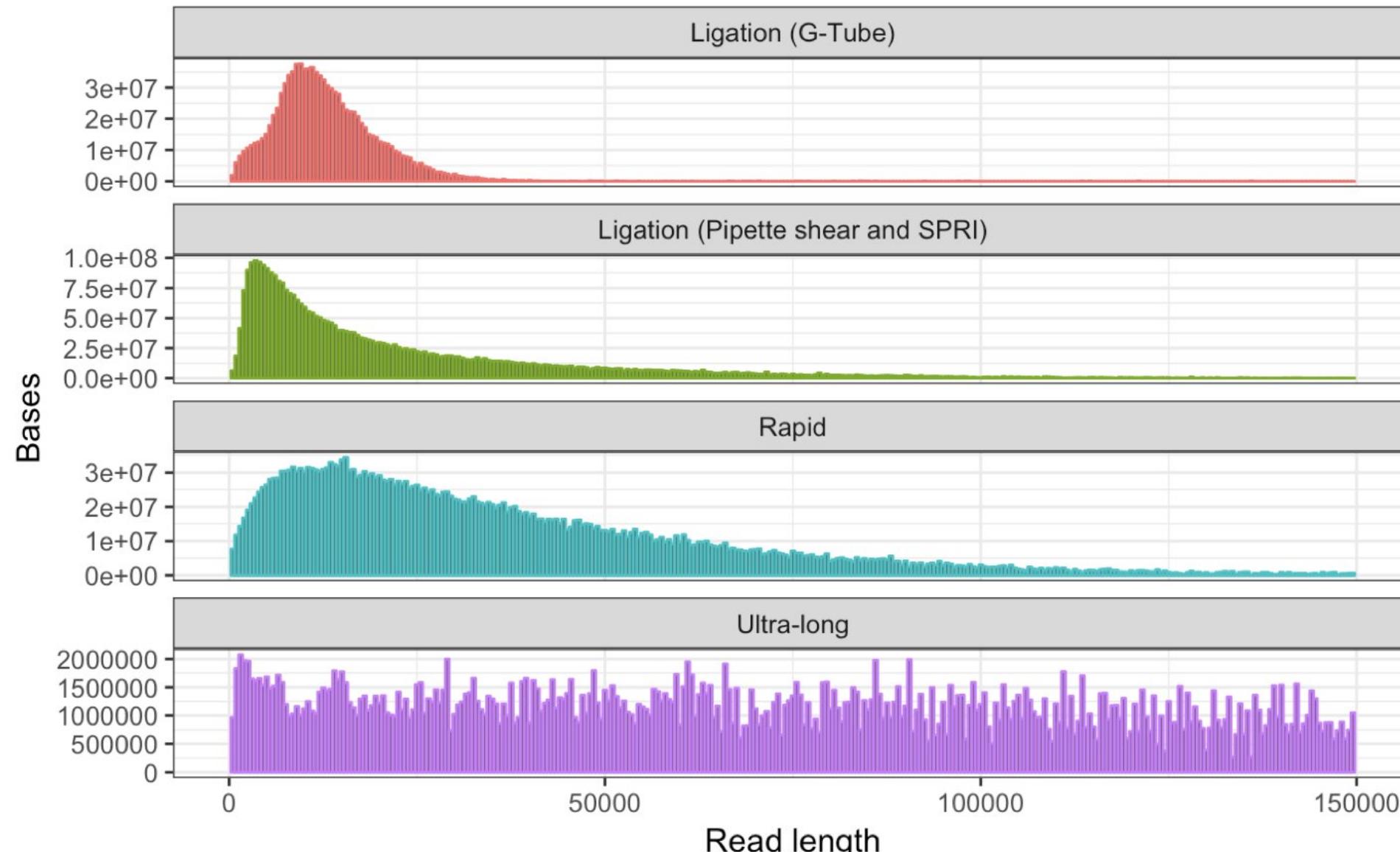


Key	SmidgION	Flongle	MinION	GridION	PromethION
System Price	TBC	Included in \$5K Starter Pack	Included in \$1K Starter Pack	Included in \$50K Starter Pack	Included in \$135K Starter Pack
Number of channels	200 channels	128 channels	512 channels	$5 \times 512 = 2,560^*$	$48 \times 3,000^* = 144,000$
Per 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	<b>TBC</b>	<b>\$90 - \$30</b>	<b>\$30 - \$12.5</b>	<b>\$17.5 - \$7.5</b>	<b>\$5 - \$2</b>

# Oxford Nanopore – how it works

<https://nanoporetech.com/how-it-works>

# Read length go beyond



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

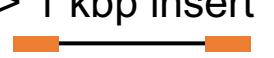
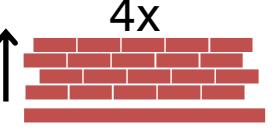
Genome reference is NOT available

- **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

 50-500 bp	Read	A sequenced piece of DNA
 300-600 bp insert	Paired-end read	Sequencing both ends of a short DNA fragment
 > 1 kbp insert	Mate-pair read	Sequencing both ends of a long DNA fragment
 length	Insert size	The length of the DNA fragment
	Contig	A set of overlapping DNA segments that represents a consensus region of DNA
 N	Scaffold	Contigs separated by gaps of known length
 4x	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--

**Scoring scheme is needed:**

1 for match

-1 for mismatch

-2 for gap

2 :

ATTGAAA-GCTA---

| | | | | |

---GAAATGAAAAGG

insertions / deletions (indels) mismatches

Which alignment is better?

# Assembly

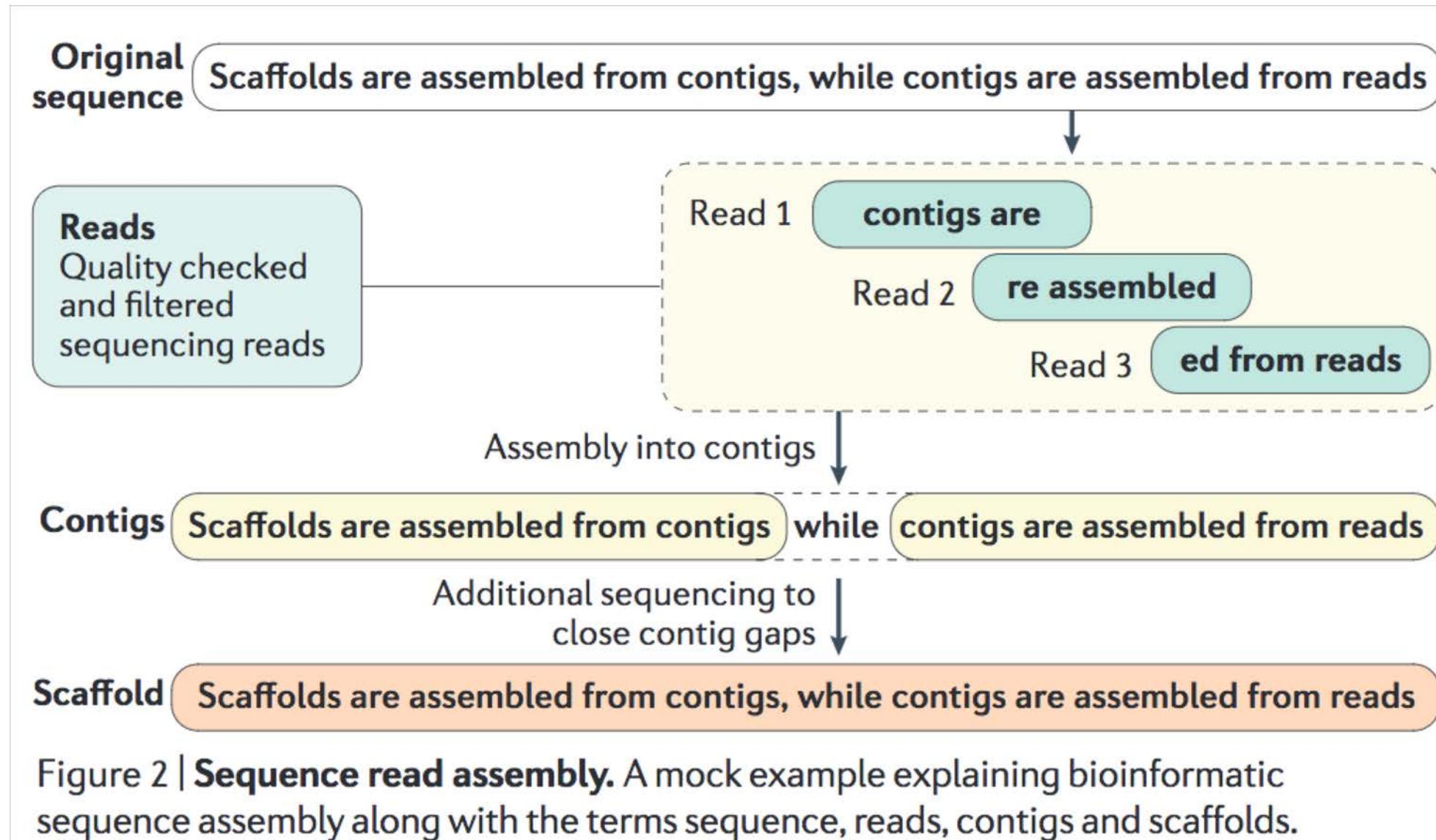
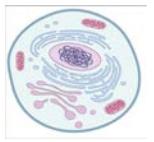


Figure 2 | **Sequence read assembly.** A mock example explaining bioinformatic sequence assembly along with the terms sequence, reads, contigs and scaffolds.

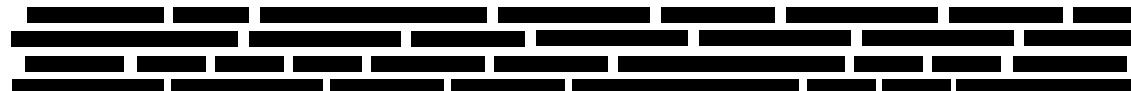
# Assembly



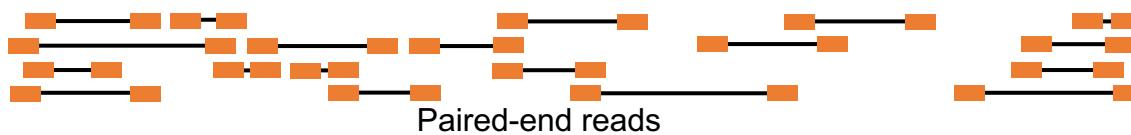
Genome



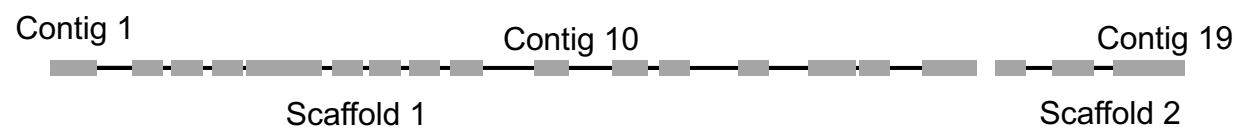
Fragment



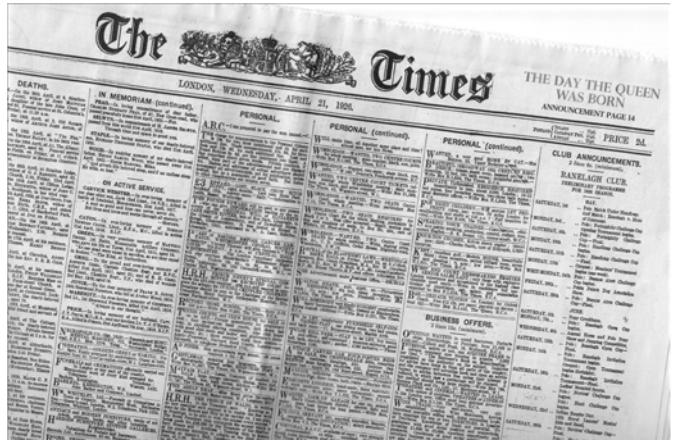
Sequence



Assemble



# Assembly



Genome  
(3.000.000 letters)

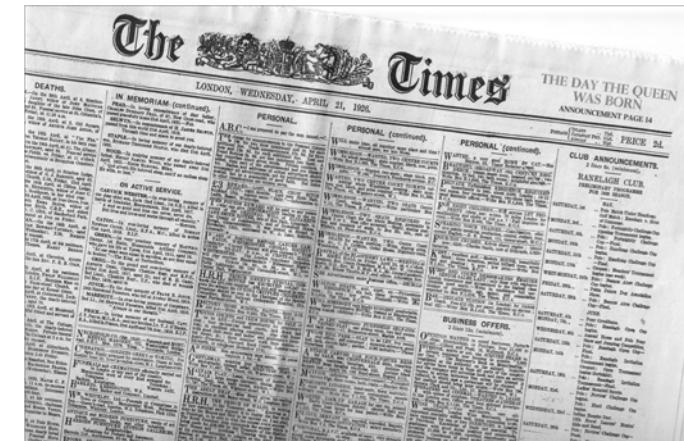
Sequencing



Reads

(50-500 letters each)

Assembly



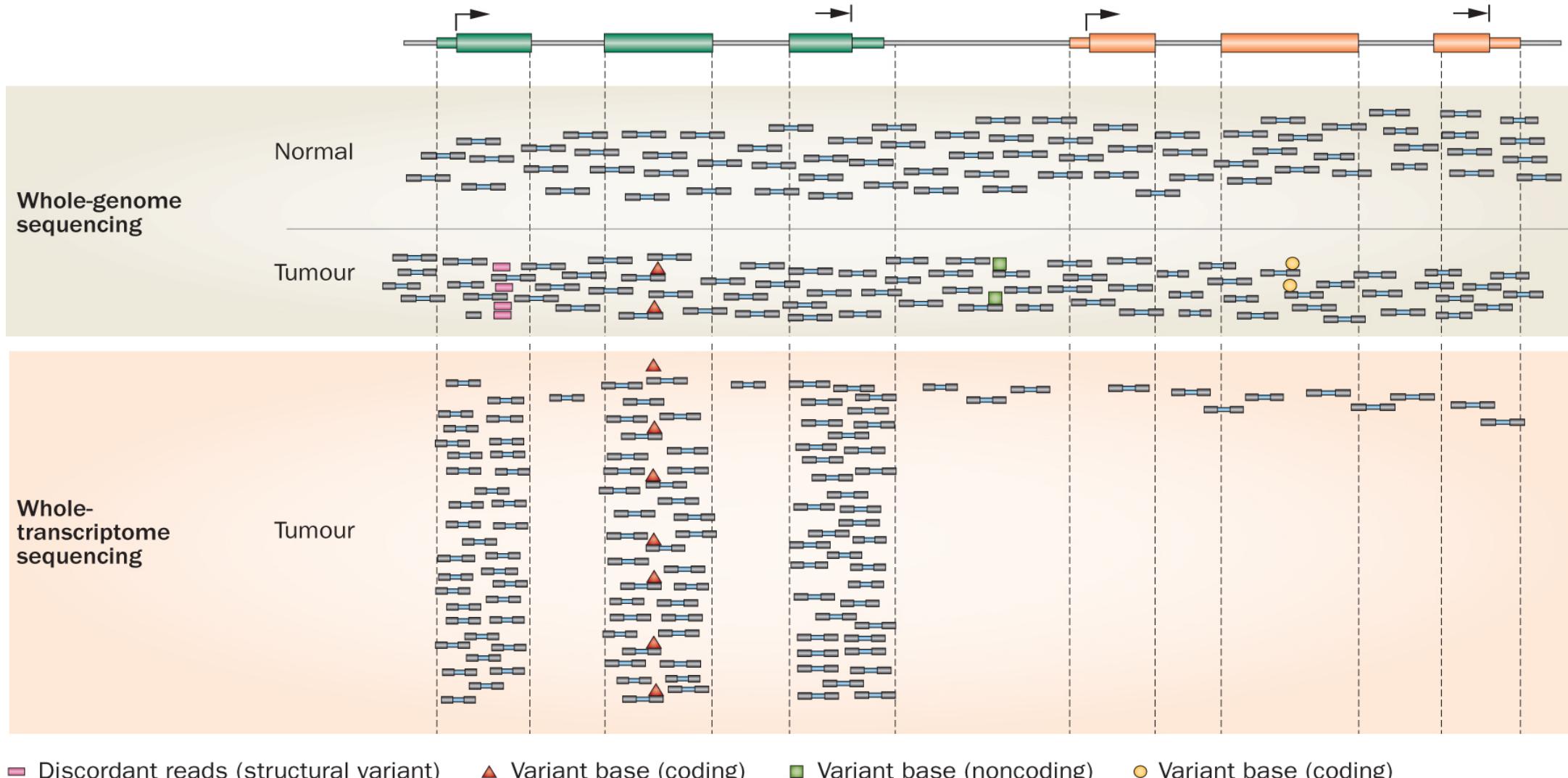
Genome  
(3.000.000 letters)

# After assembly

- Say you have an assembly with 200 contigs and 34 scaffolds.  
What do you do next?
- How accurate is it?
- Have you tried different assemblers?
- Can you improve with additional data or diminishing returns?
- Is there contamination?
- How does it compare to other species?

# Mapping

Reference genome depicting two example genes



■ Discordant reads (structural variant)

▲ Variant base (coding)

■ Variant base (noncoding)

○ Variant base (coding)

# Read length matters in sequencing

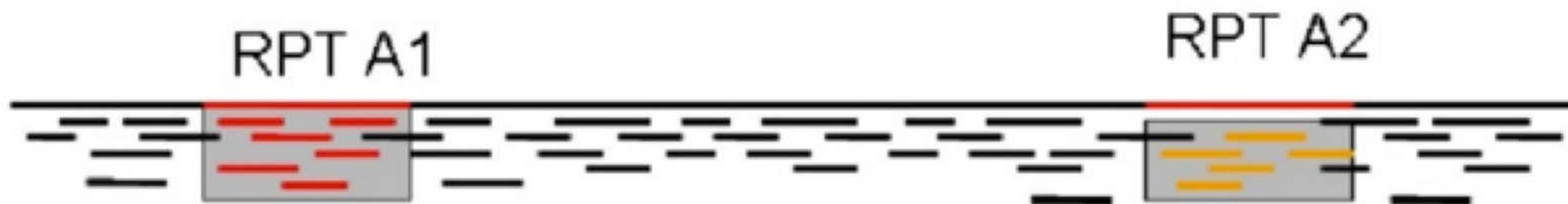


Figure 5. Two copies of a repeat along a genome. The reads colored in red and those colored in yellow appear identical to the assembly program.

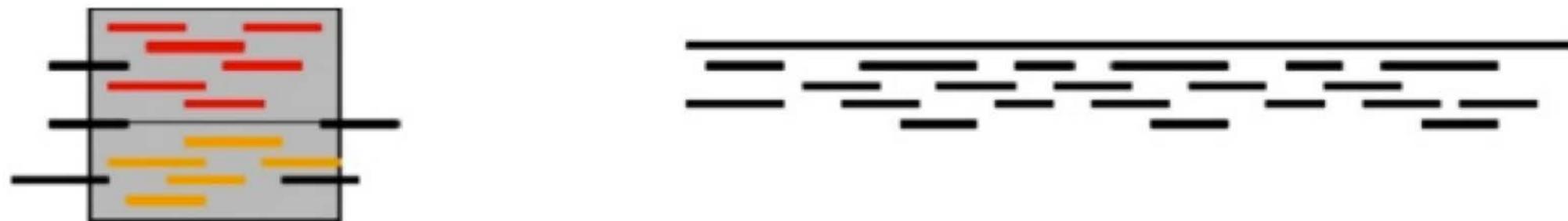
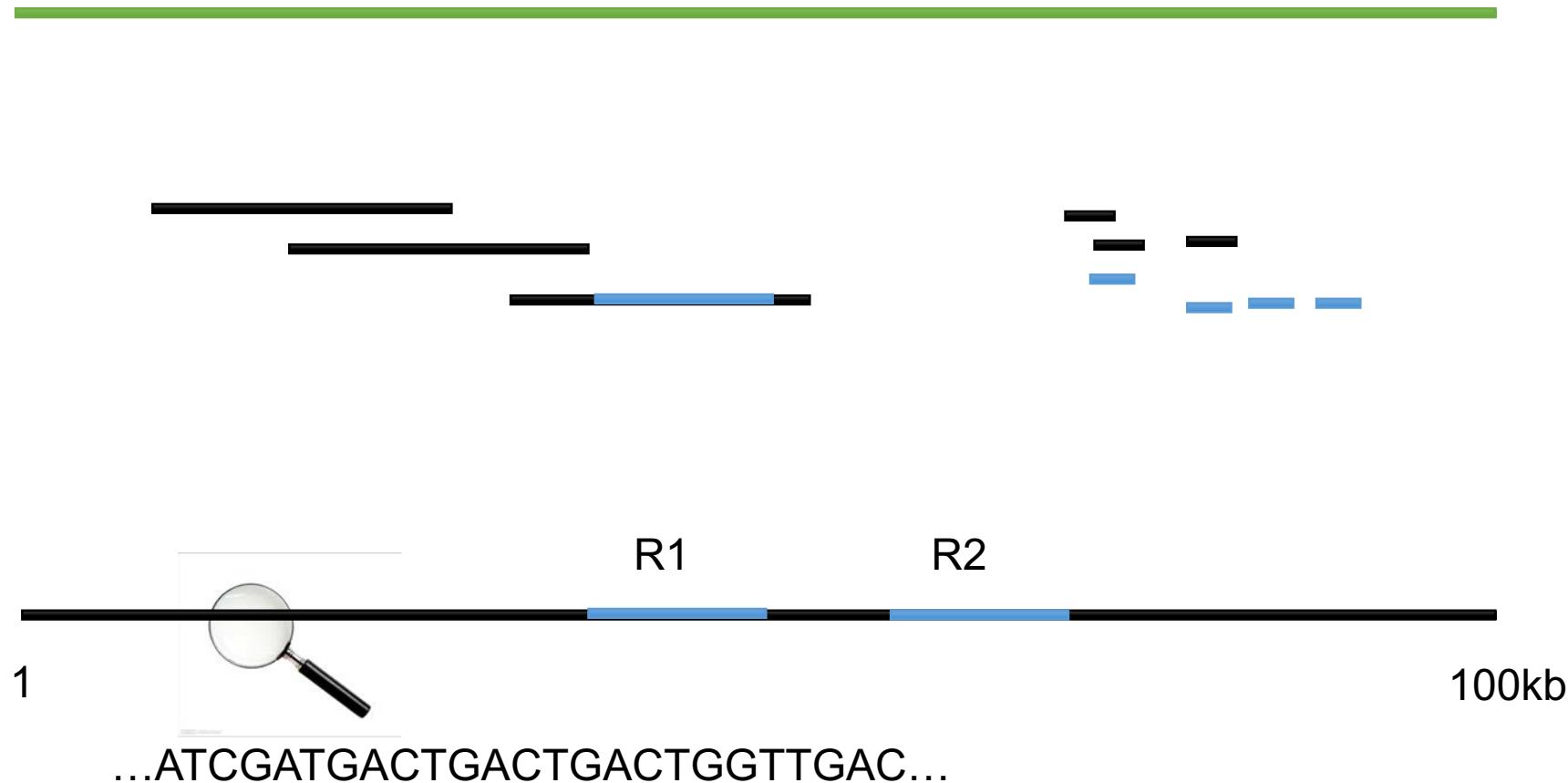
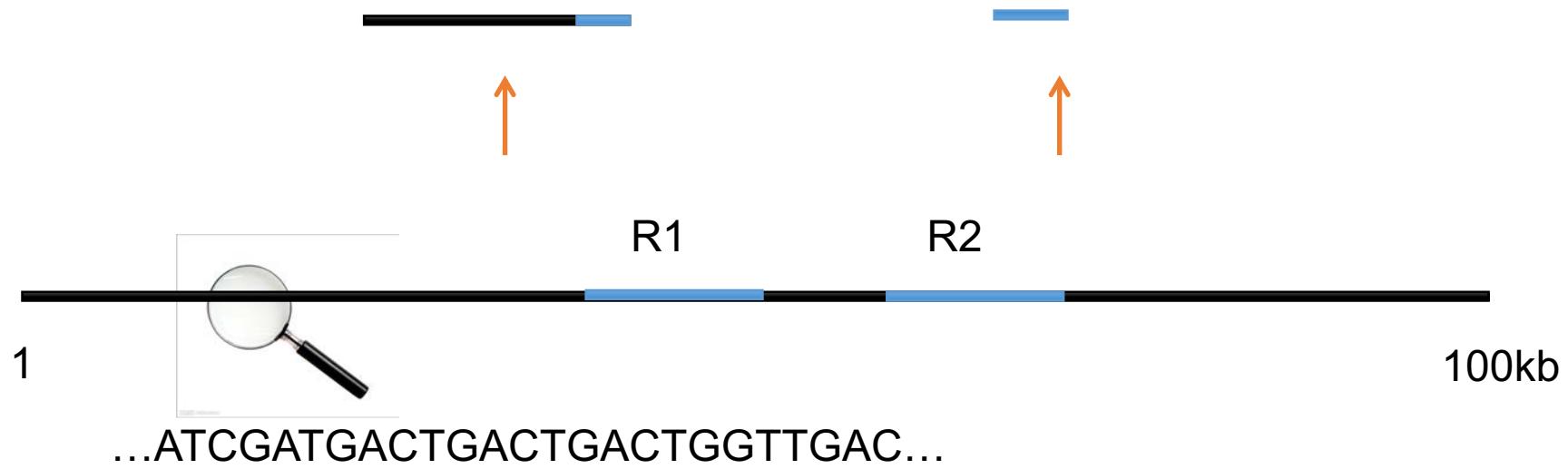


Figure 6. Genome mis-assembled due to a repeat. The assembly program incorrectly combined the reads from the two copies of the repeat leading to the creation of two separate contigs

# Read length matters in sequencing



# Paired end and insert size matter in sequencing



# Depth matters in sequencing

ATCGATGACTGACTGAATGGTTGAC  
ATCGATGACTGACTGAATGGTTGAC  
ATCCATGACTGACTGAATGGTTGAC  
ATCGATGACTGACTGAATGGTTGAC  
ATCGATGACTGACTGAATGGTTGAC  
ATCGATGACTGAGTGAATGGTTGAC  
ATCGATGACTGAGTGAATGGTTGAC  
ATCGATGACTGAGTGAATGGTTGAC  
ATCGATGACTGAGTGAATGGTTGAC  
10X ATCGATGACTGAGTGAATGGTTGAC

1X ATCGAT<sup>C</sup>ACTGACTGACTGGTTGAC

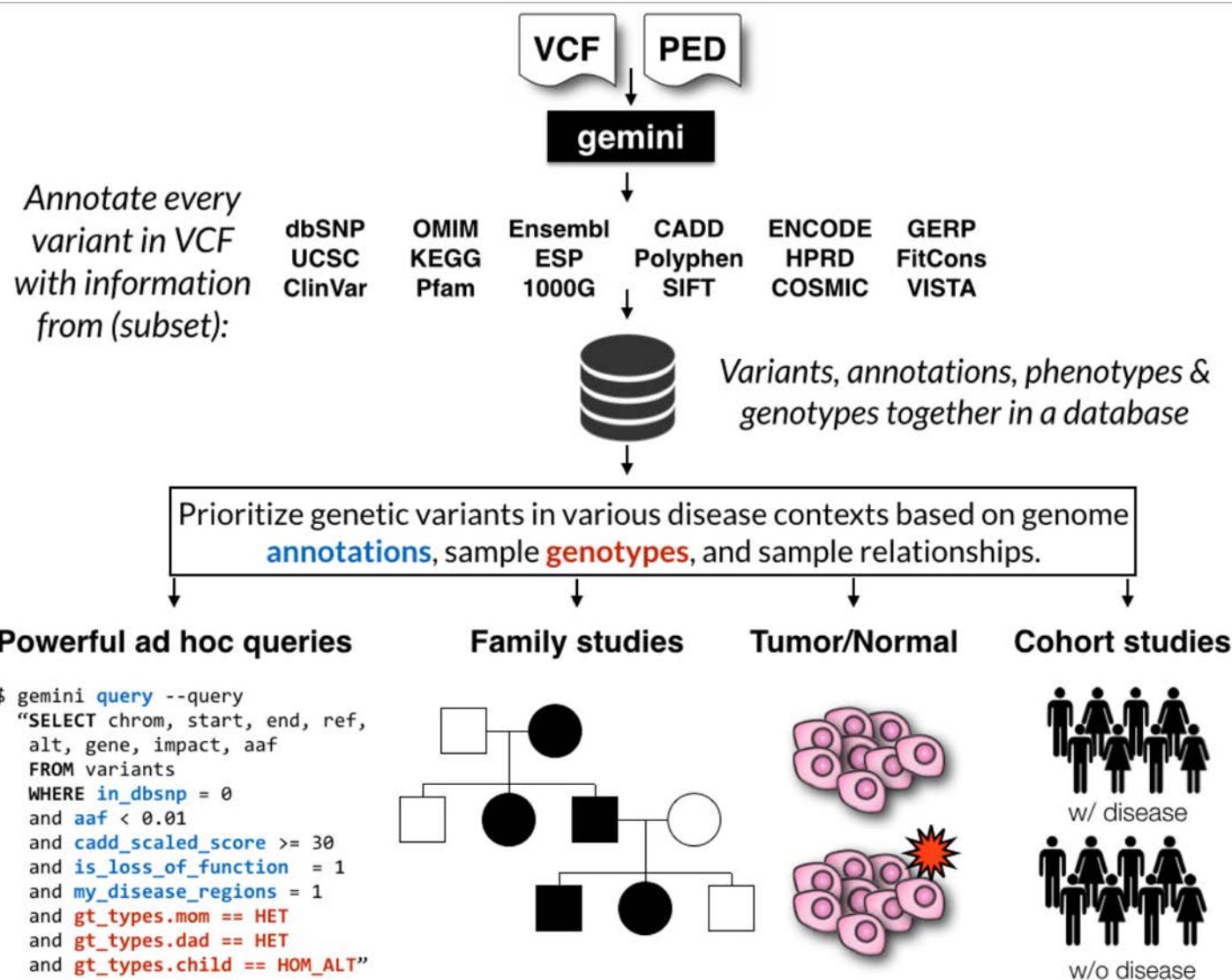
Homozygous? Heterozygous?

---

...ATCGATGACTGACTGACTGGTTGAC...

reference

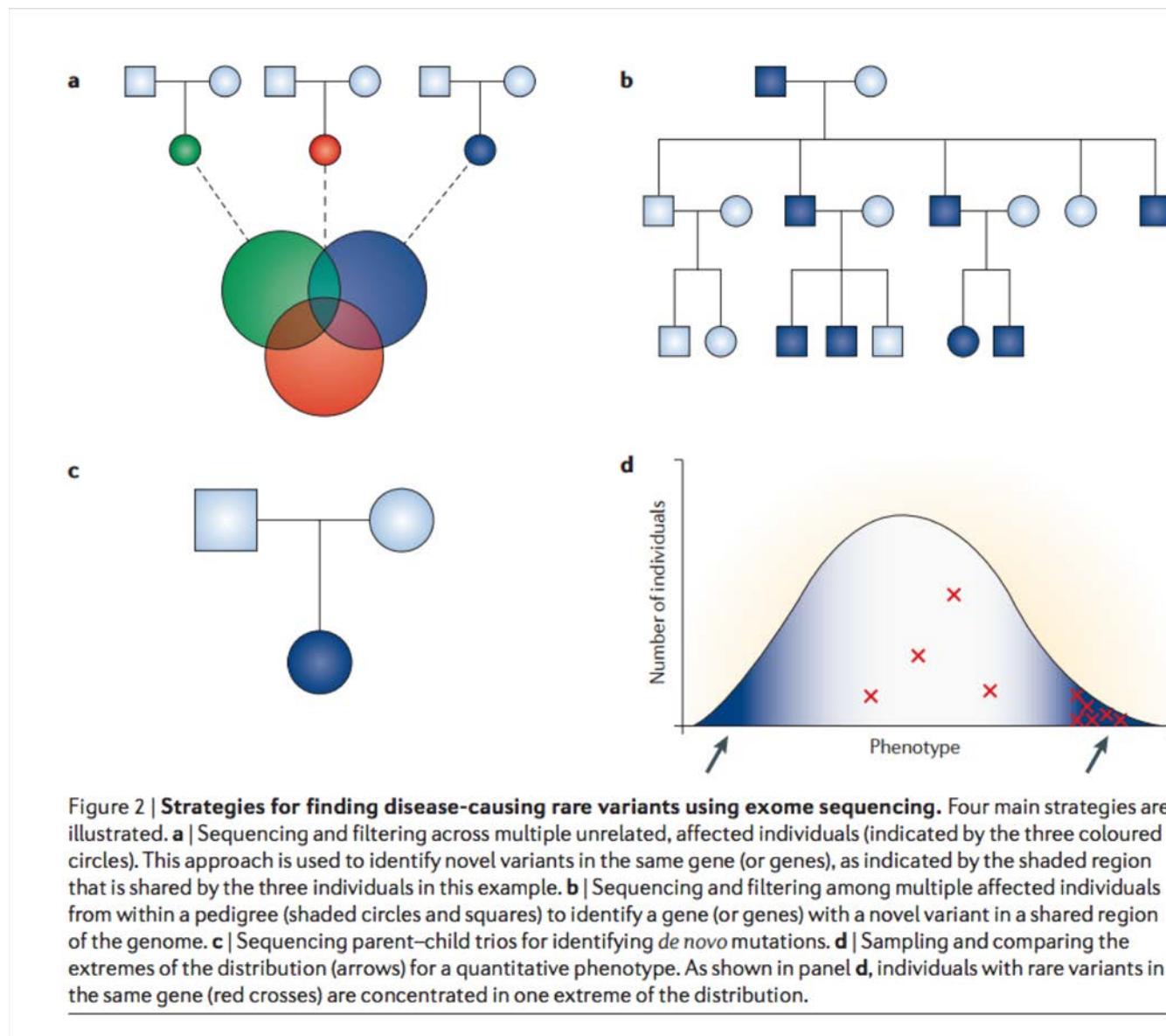
# Filtering and annotating variants



ClinVar  
SIFT  
Function  
MAF (Minor Allele frequency)

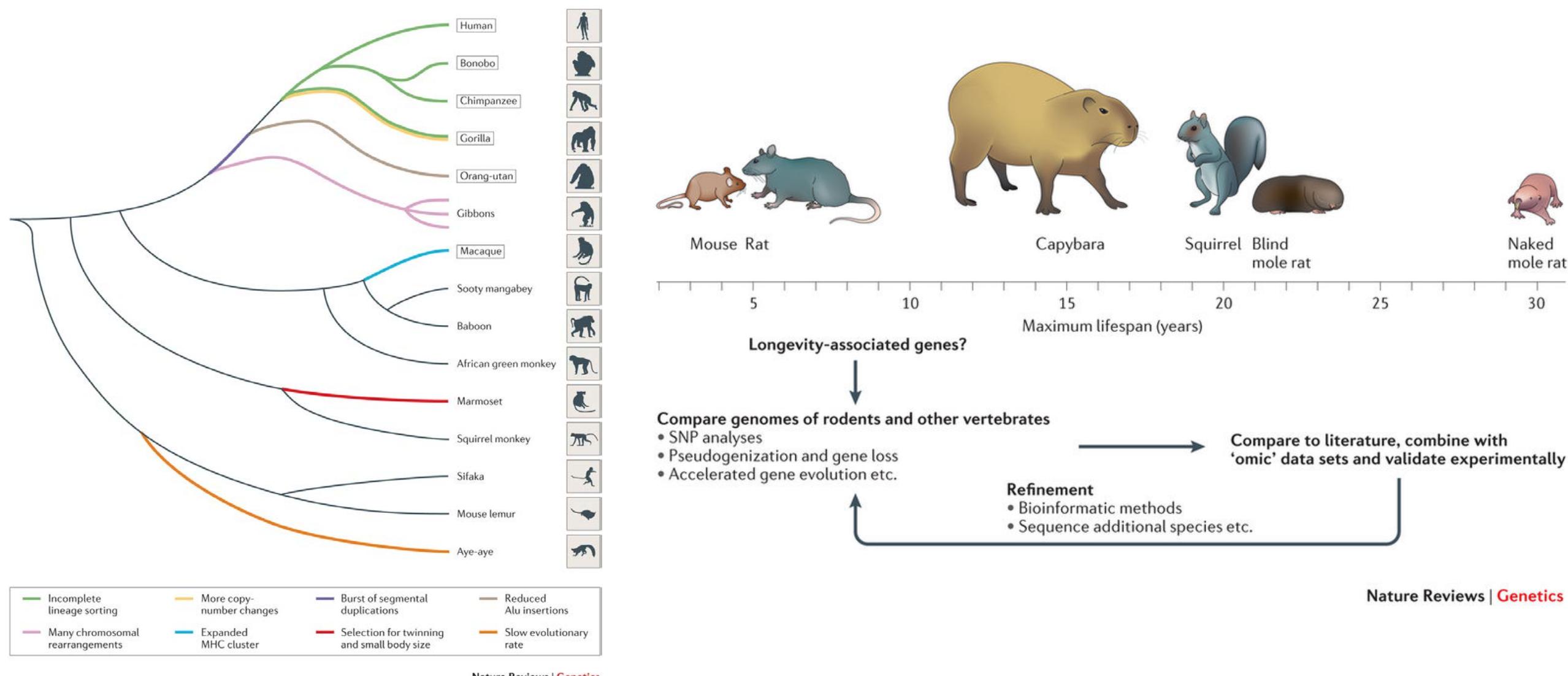
# Case studies

# Classical genetics

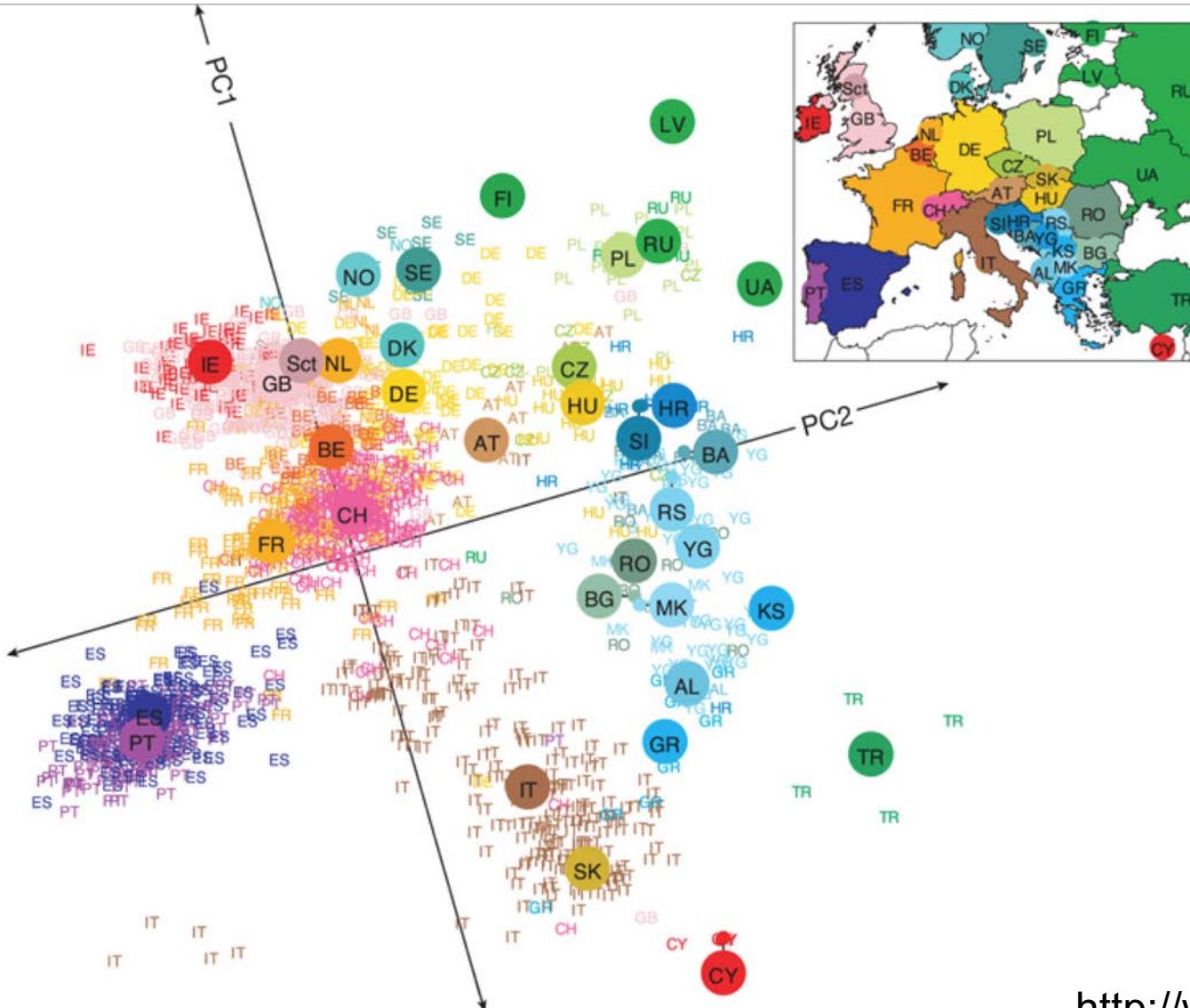


**Figure 2 | Strategies for finding disease-causing rare variants using exome sequencing.** Four main strategies are illustrated. **a** | Sequencing and filtering across multiple unrelated, affected individuals (indicated by the three coloured circles). This approach is used to identify novel variants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

# Comparative genomics



# Population genomics

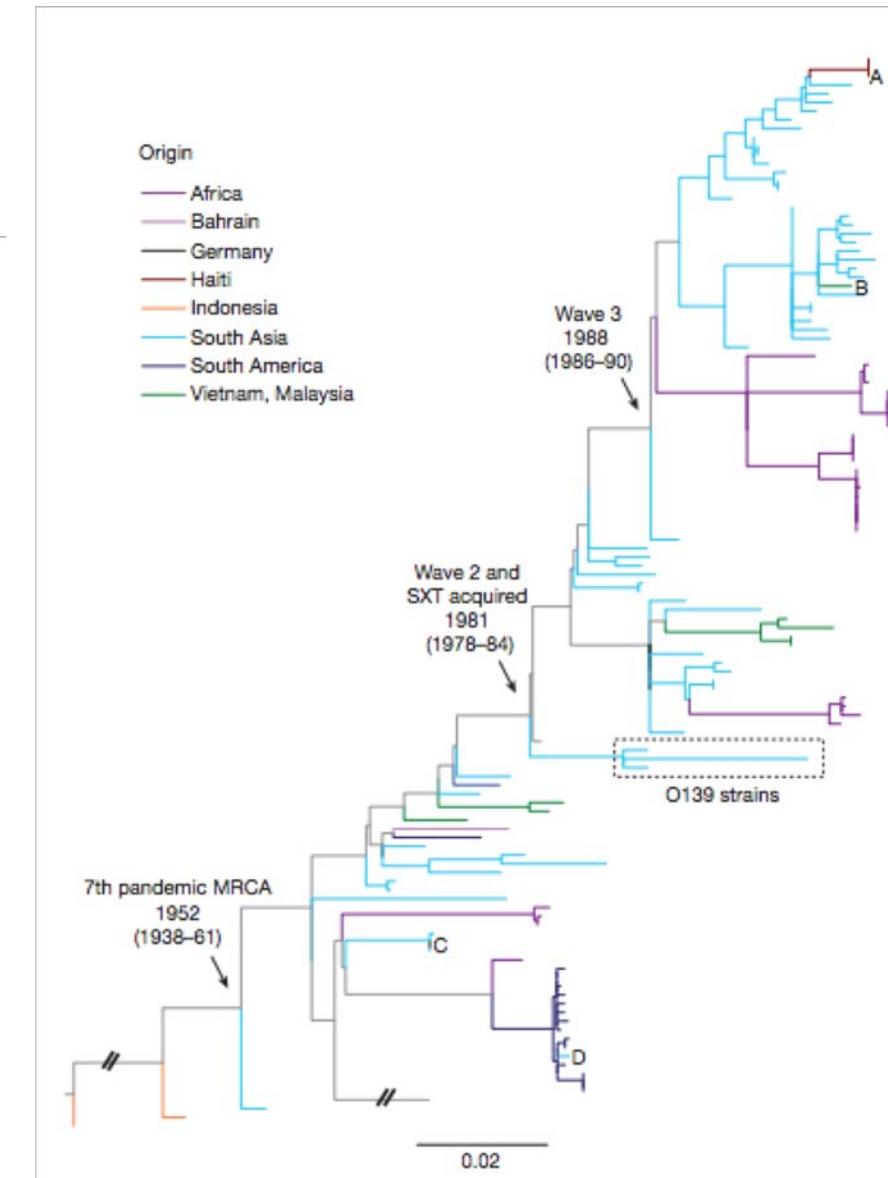
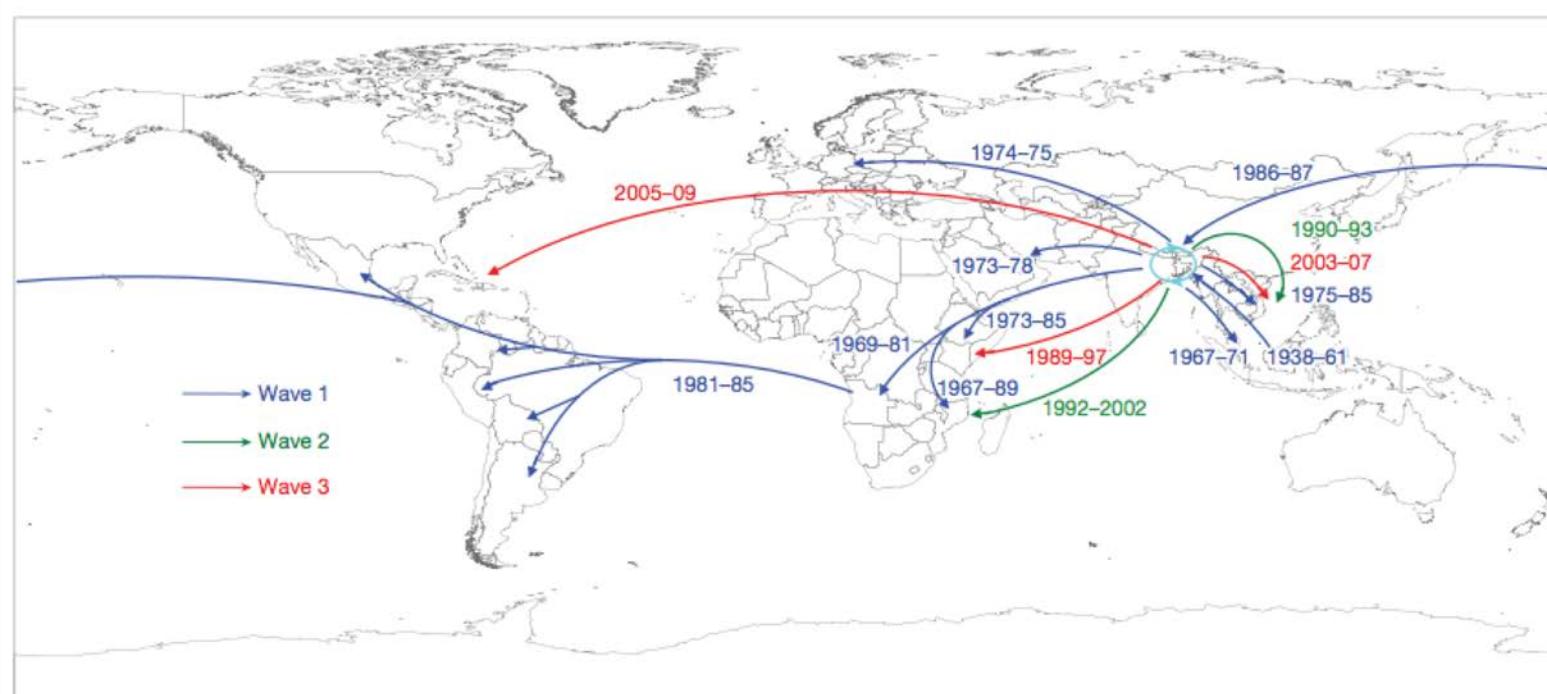


Novembre et al Nature (2008)



[http://www.genomenext.com/casestudies\\_post/population-scale-analysis-genomic-samples-analyzed-from-2504-individuals-in-1-week/](http://www.genomenext.com/casestudies_post/population-scale-analysis-genomic-samples-analyzed-from-2504-individuals-in-1-week/)

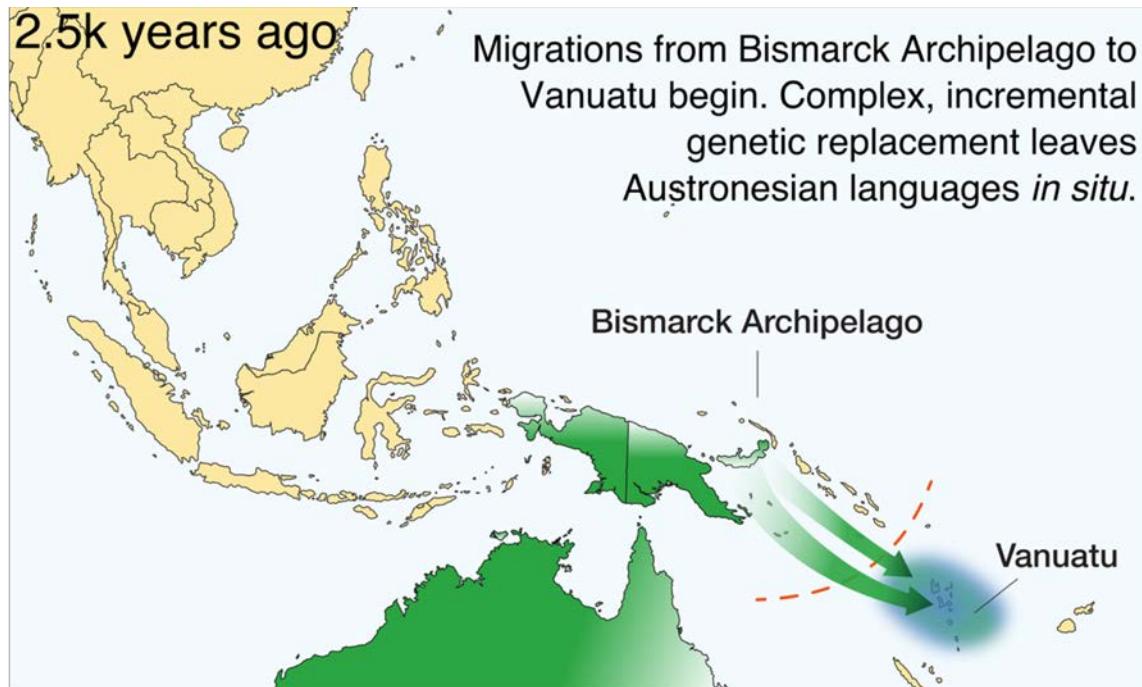
# Population genomics



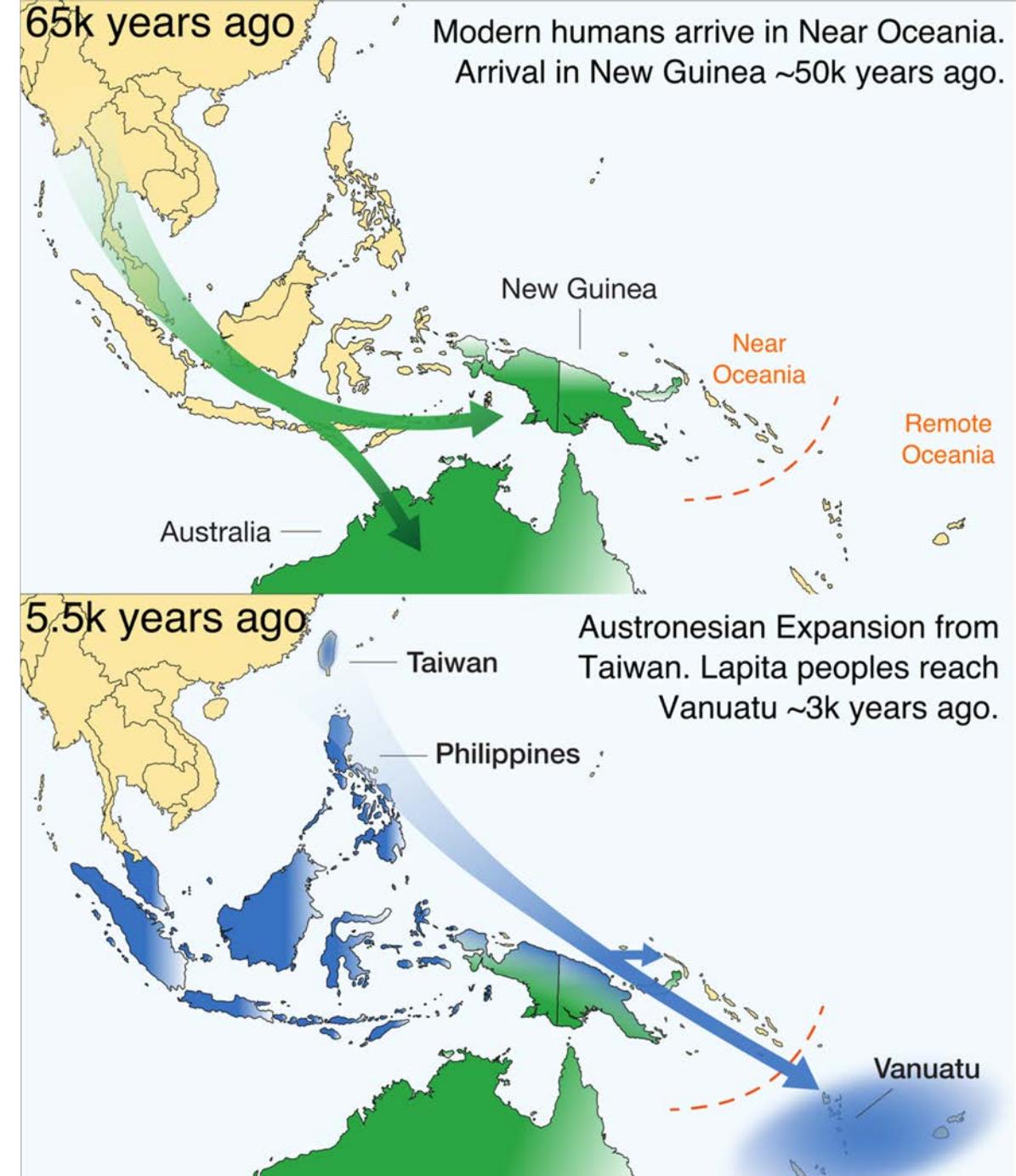
Mutreja *et al.*, Nature Genetics (2011)

# Population genomics

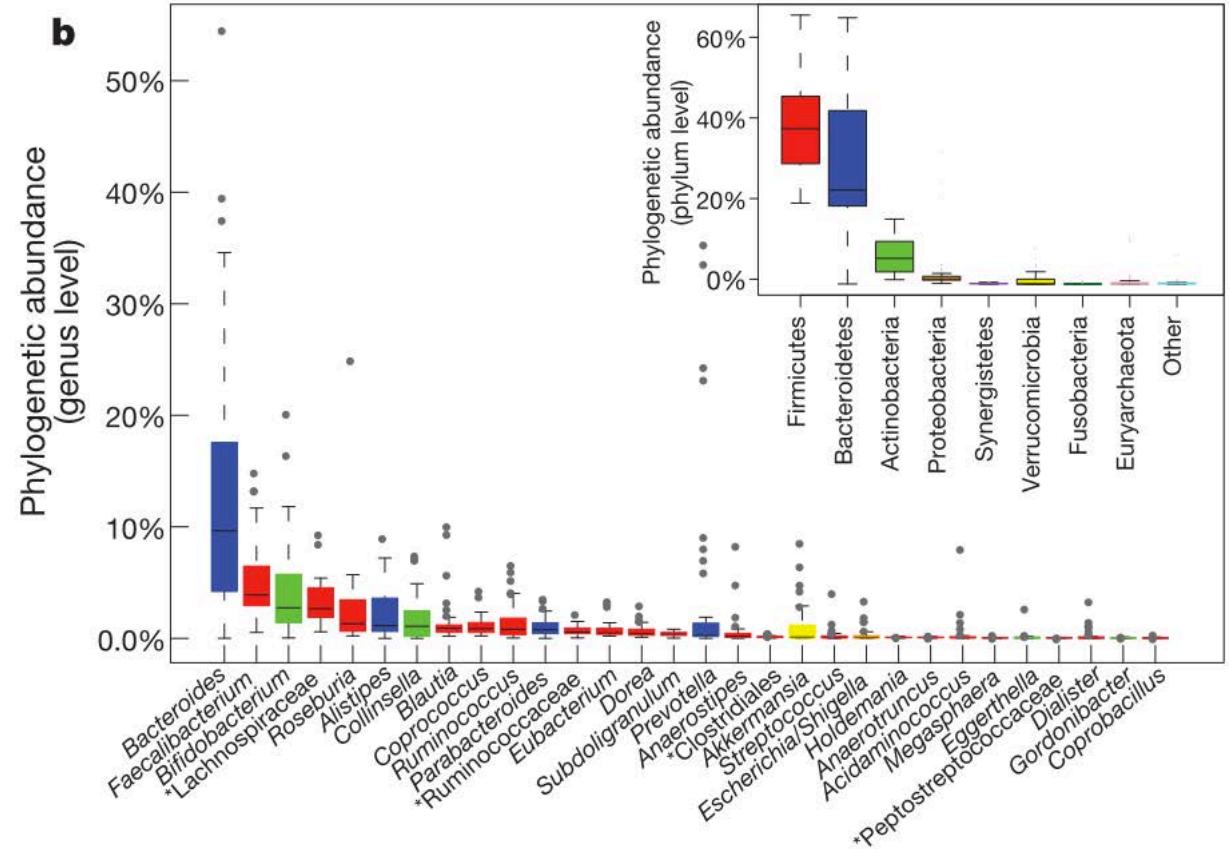
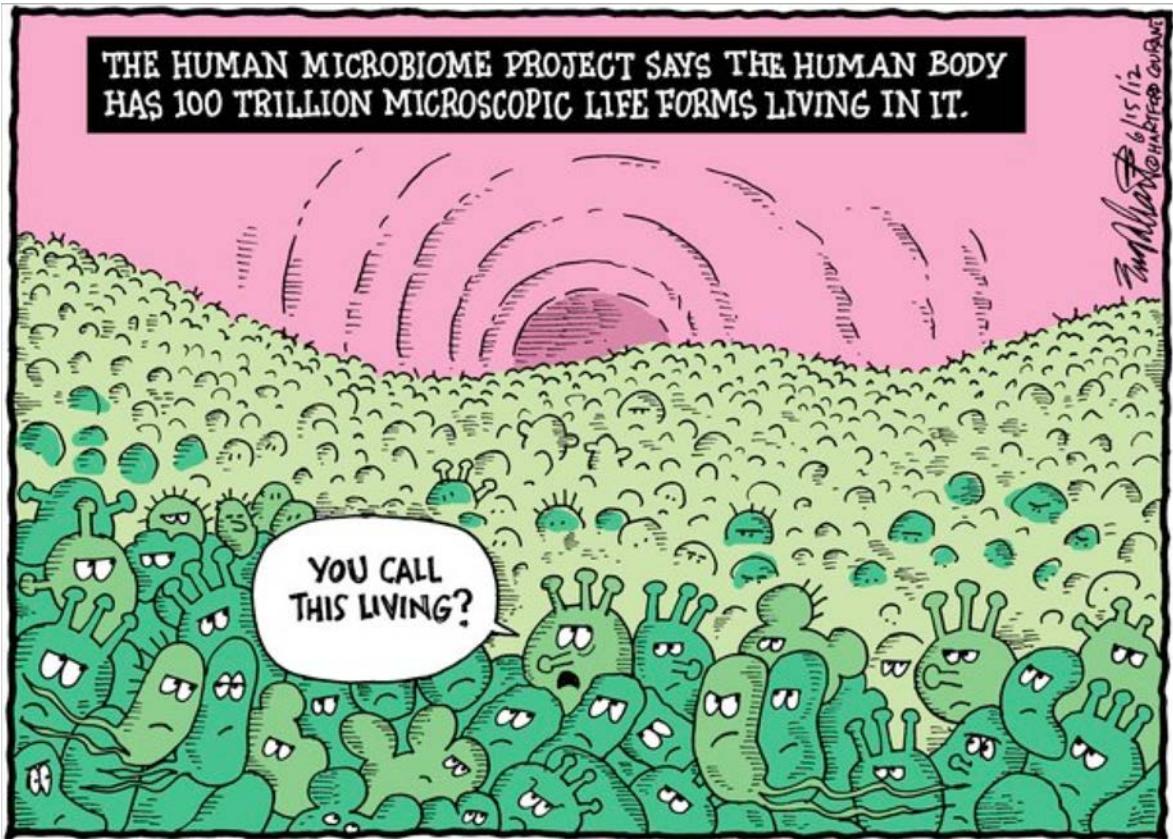
There is a puzzling mismatch in the population history of the Pacific: why do most people across Remote Oceania (the vast area stretching from Vanuatu to Easter Island-Rapa Nui) speak languages of the Austronesian language family that expanded into this region only 3,000 years ago, yet carry a component 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)?



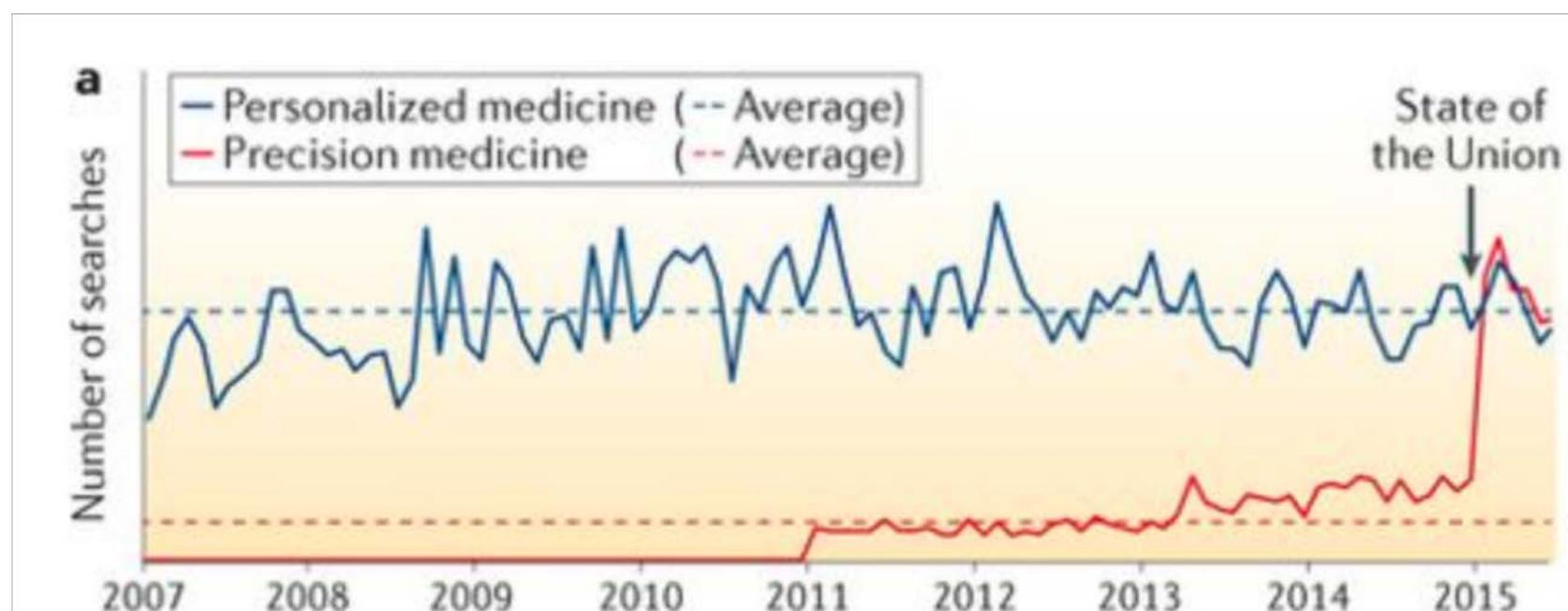
Credit: Hans Sell, MPI-SHH, adapted from Skoglund *et al.* 2016 *Nature*



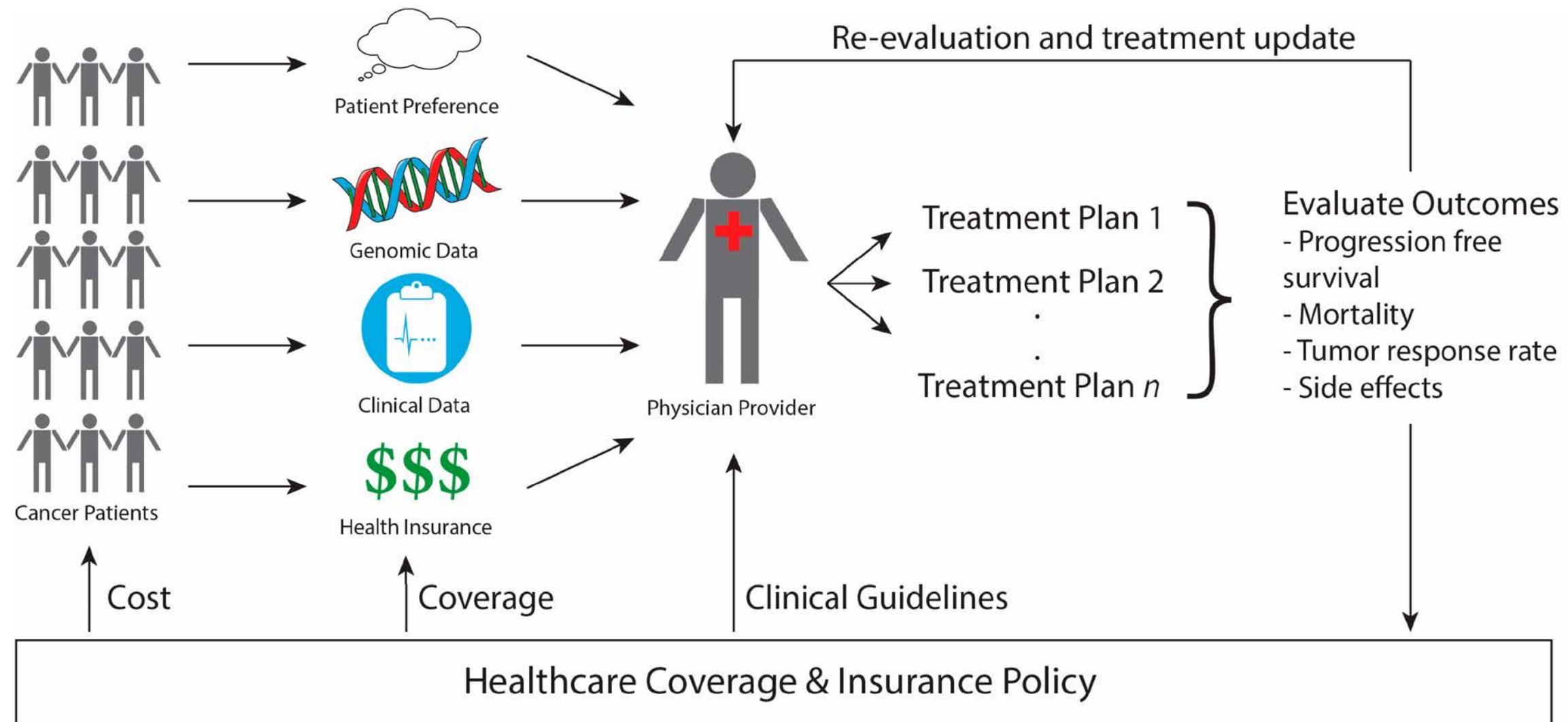
# Metagenomics



# Precision medicine



# Outline of precision medicine



# PM examples

Table 1 | Examples of precision medicine

Condition	Gene	Action
<b>Mendelian disease</b>		
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	<ul style="list-style-type: none"> <li>Heterozygous FH (HeFH): eligible for PCSK9 inhibitor drugs</li> <li>Homozygous FH (HoFH): eligible for PCSK9 inhibitor drugs in addition to lomitapide and mipomersen</li> </ul>
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
<b>Precision oncology</b>		
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
<b>Pharmacogenomics</b>		
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

# Summary of outcomes in Oncology PM Studies

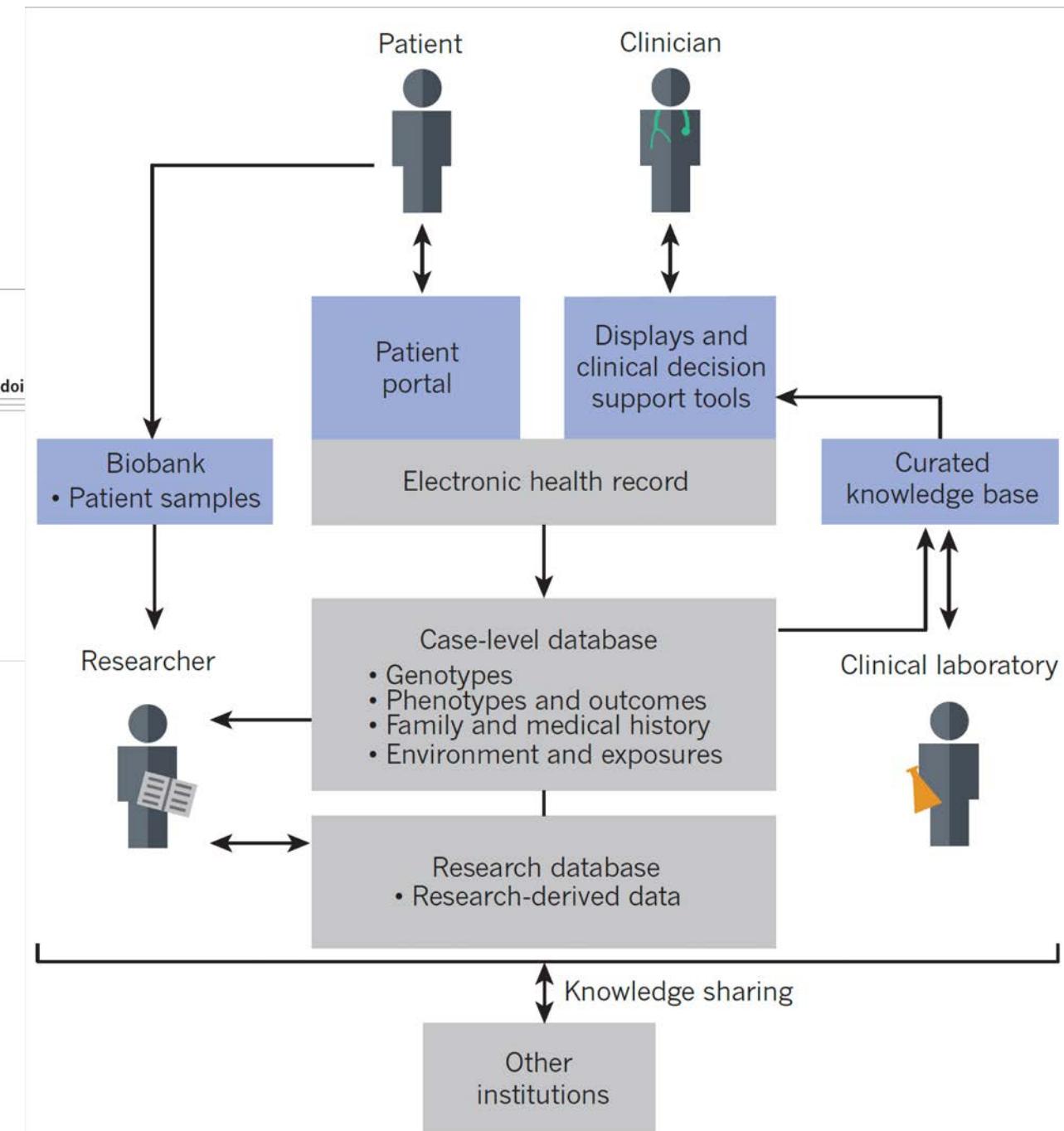
Study	Sample Size	Most Prevalent Tumor Types	Outcomes Reported
Tsimberidou et al. <i>Clin. Cancer Res.</i> 2012 [5]	291 patients with one molecular aberration (175 treated with matched therapy, 116 control)	Colorectal, melanoma, lung, ovarian	Matched group had improved ORR (27% vs. 5%), TTF (median 5.2 vs. 2.2 month), OS (median 13.4 vs. 9.0 month)
Radovich et al. <i>Oncotarget</i> 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. <i>Mol. Cancer Ther.</i> 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. <i>JAMA</i> 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. <i>J. Clin. Oncol.</i> 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. <i>Genome Med.</i> 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. <i>Lancet Oncol.</i> 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.

# REVIEW

## Building the foundation for genomics in precision medicine

Samuel J. Aronson<sup>1,2</sup> & Heidi L. Rehm<sup>1,3,4,5</sup>



# Building the foundation for genomics in precision medicine

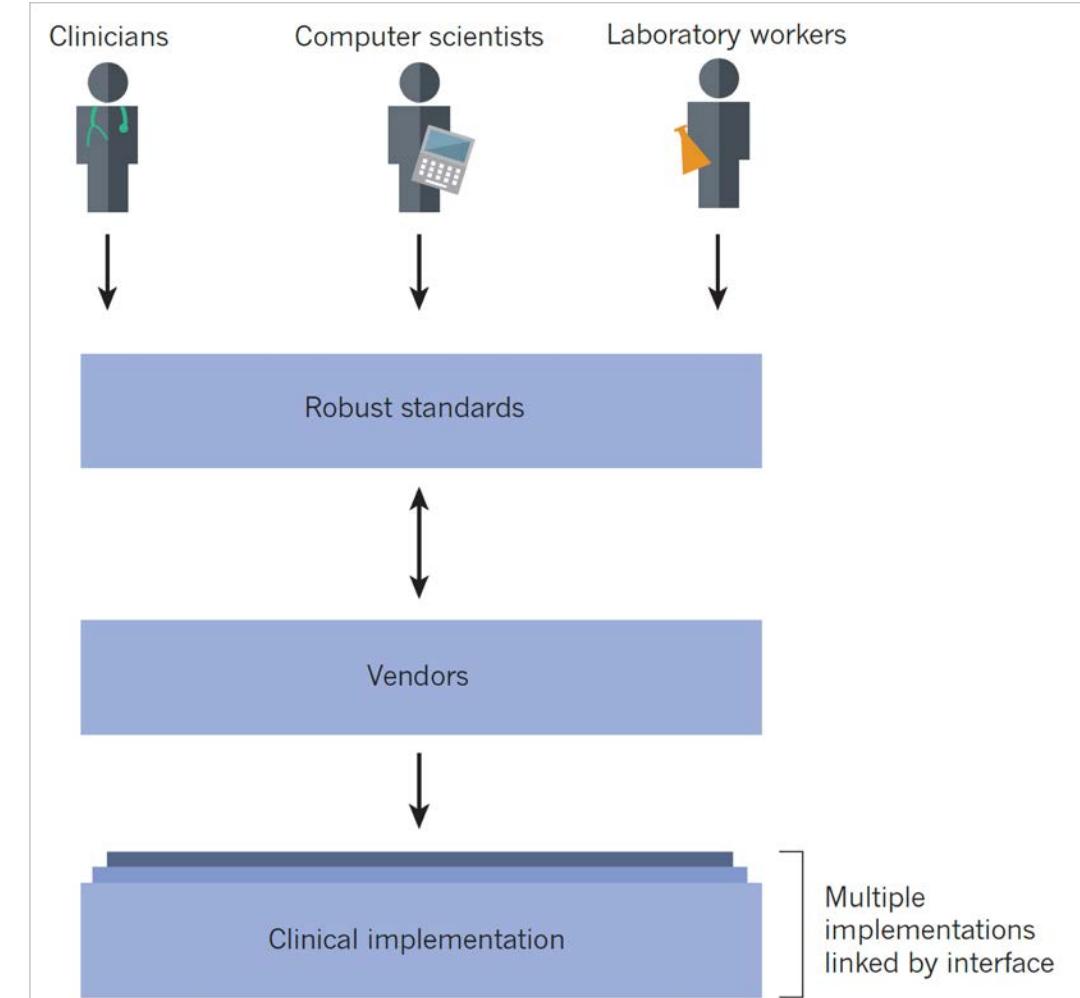
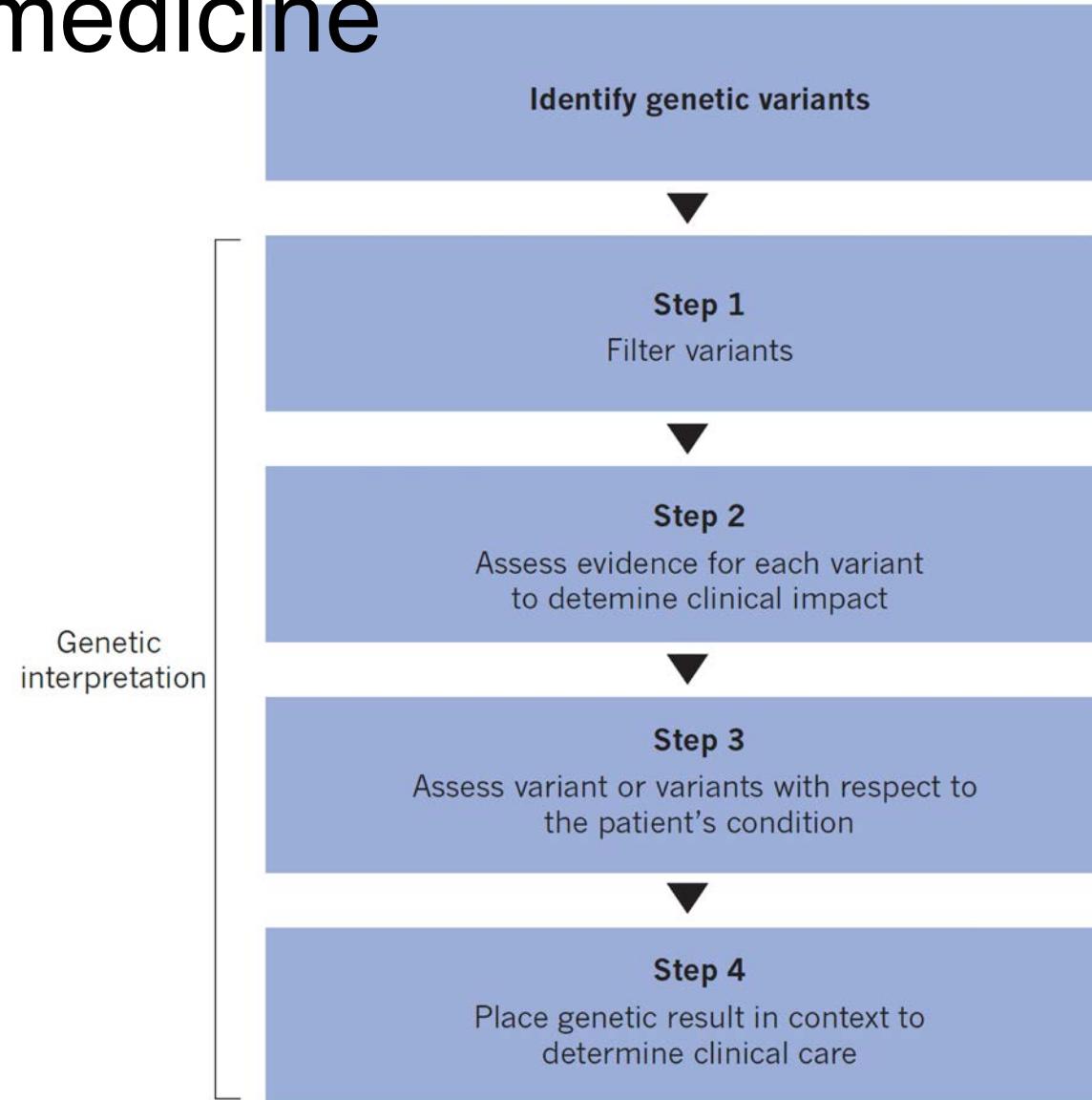
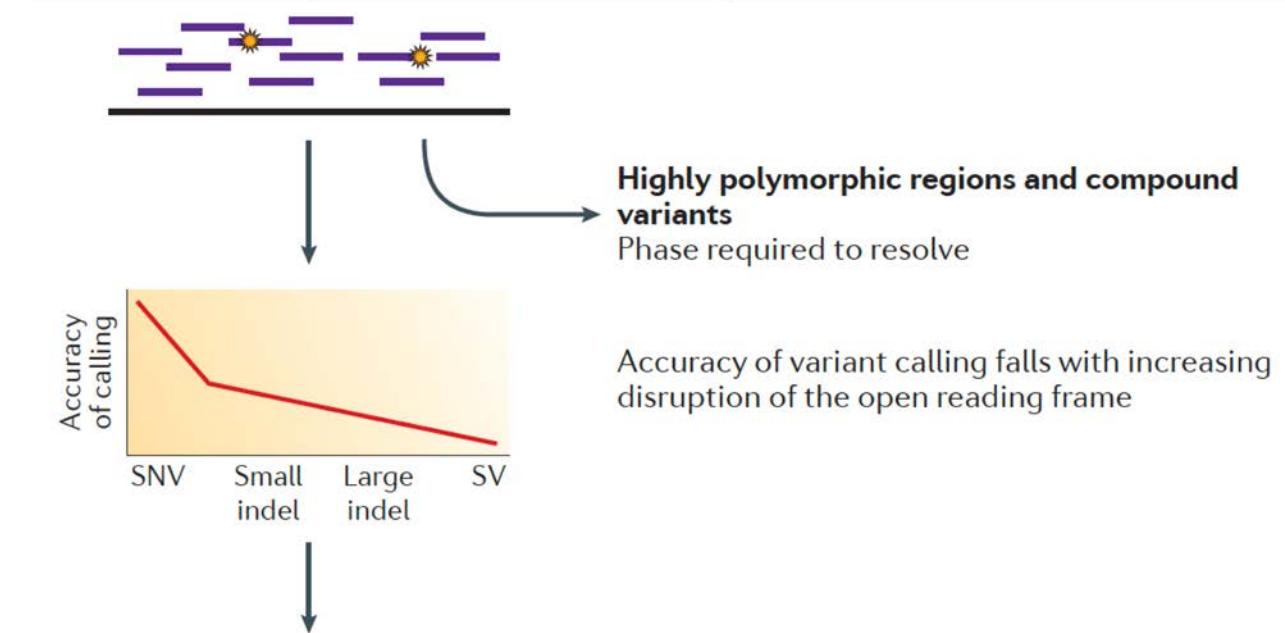
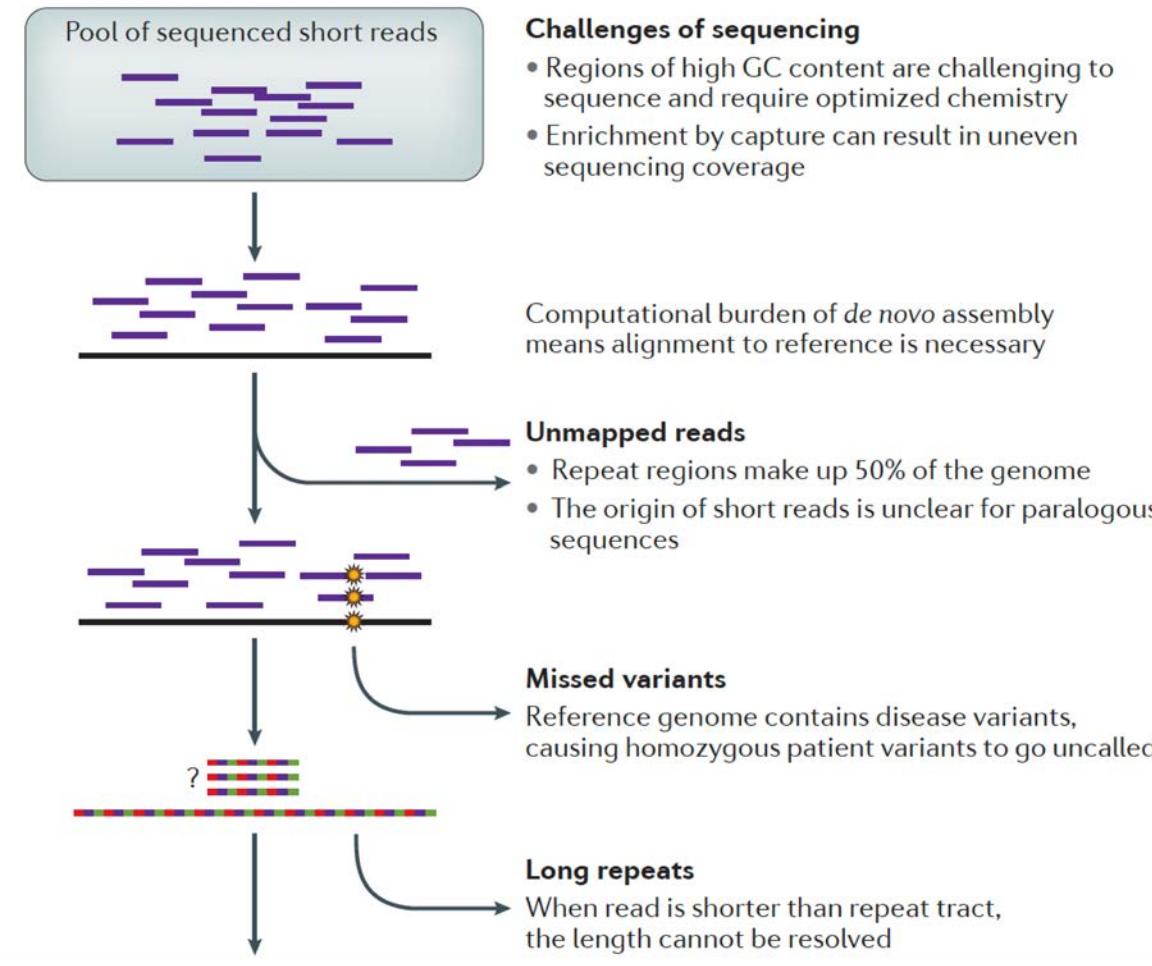
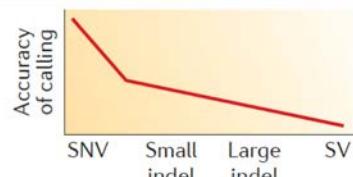


Figure 3 | Creating and implementing robust standards for the description and structuring of data in laboratory processing and patient-care systems.

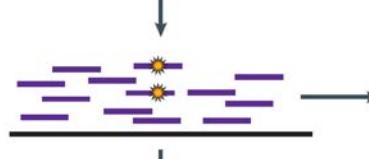
# Challenges and reduced accuracies



# Challenges and reduced accuracies



Accuracy of variant calling falls with increasing disruption of the open reading frame



Position	REF	ALT	Call
Chr14:23,456,332	T	A	0/1

#### Final VCF file

- File of appropriately called variants
- The VCF should contain a call at every position or patients homozygous for risk alleles present in the reference will be missed

Position	REF	ALT	Call
Chr14:23,456,332	T	A	0/1
Chr14:23,678,972	C	G	1/1
...			

Variants filtered based on standard metrics, such as population frequency and known disease-associated genes

Position	REF	ALT	Call
Chr14:23,456,332	T	A	0/1
Chr14:23,678,972	C	G	1/1
...			

Causality determined by magnitude and dependency of effect

Downstream treatment and disease management are influenced by knowledge of disease-causing gene and variant

a  
Repeat location

Disease

Gene

Repeat sequence

Repeat length

5' UTR

Exon

Intron

Exon

3' UTR

Fragile X syndrome  
Fragile X-associated tremor/ataxia syndrome  
Fragile XE mental retardation  
Spinocerebellar ataxia 12

FMR1  
FMR1  
FRAXE  
ATXN12

CGG  
CGG  
GCC  
CAG

0 50 100 200 1,000 2,000 11,000

Spinocerebellar ataxias 1, 2, 3, 6, 7 and 17  
Huntington disease  
Spinal and bulbar muscular atrophy  
Dentatorubral-pallidoluysian atrophy

ATXN1, 2, 3, 7,  
CACNA1A and TBP  
HTT  
AR  
ATN1

CAG

50 100 200 1,000 2,000 11,000

Friedreich's ataxia  
Myotonic dystrophy 2  
Spinocerebellar ataxia 10  
Spinocerebellar ataxia 31  
Spinocerebellar ataxia 36  
Amyotrophic lateral sclerosis

FXN  
CNBP  
ATXN10  
BEAN1  
NOP56  
C9orf72

GAA  
CCTG  
ATTCT  
TGGA  
GGCTG  
GGGGCC

50 100 200 1,000 2,000 11,000

Multiple skeletal dysplasias  
Synpolydactyly syndrome  
Hand–foot–genital syndrome  
Cleidocranial dysplasia  
Holoprosencephaly  
Oculopharyngeal muscular atrophy  
Congenital central hypoventilation syndrome  
Blepharophimosis, ptosis and epicanthus inversus syndrome  
ARX-related X-linked mental retardation

COMP  
HOXD13  
HOXA13  
RUNX2  
ZIC2  
PABPN1  
PHOX2B  
FOXL2  
ARX

GAC  
GCG  
GCG  
GCG  
GCG  
GCG  
GCG  
GCG  
GCG

50 100 200 1,000 2,000 11,000

Myotonic dystrophy 1  
Spinocerebellar ataxia 8  
Huntington's disease-like 2

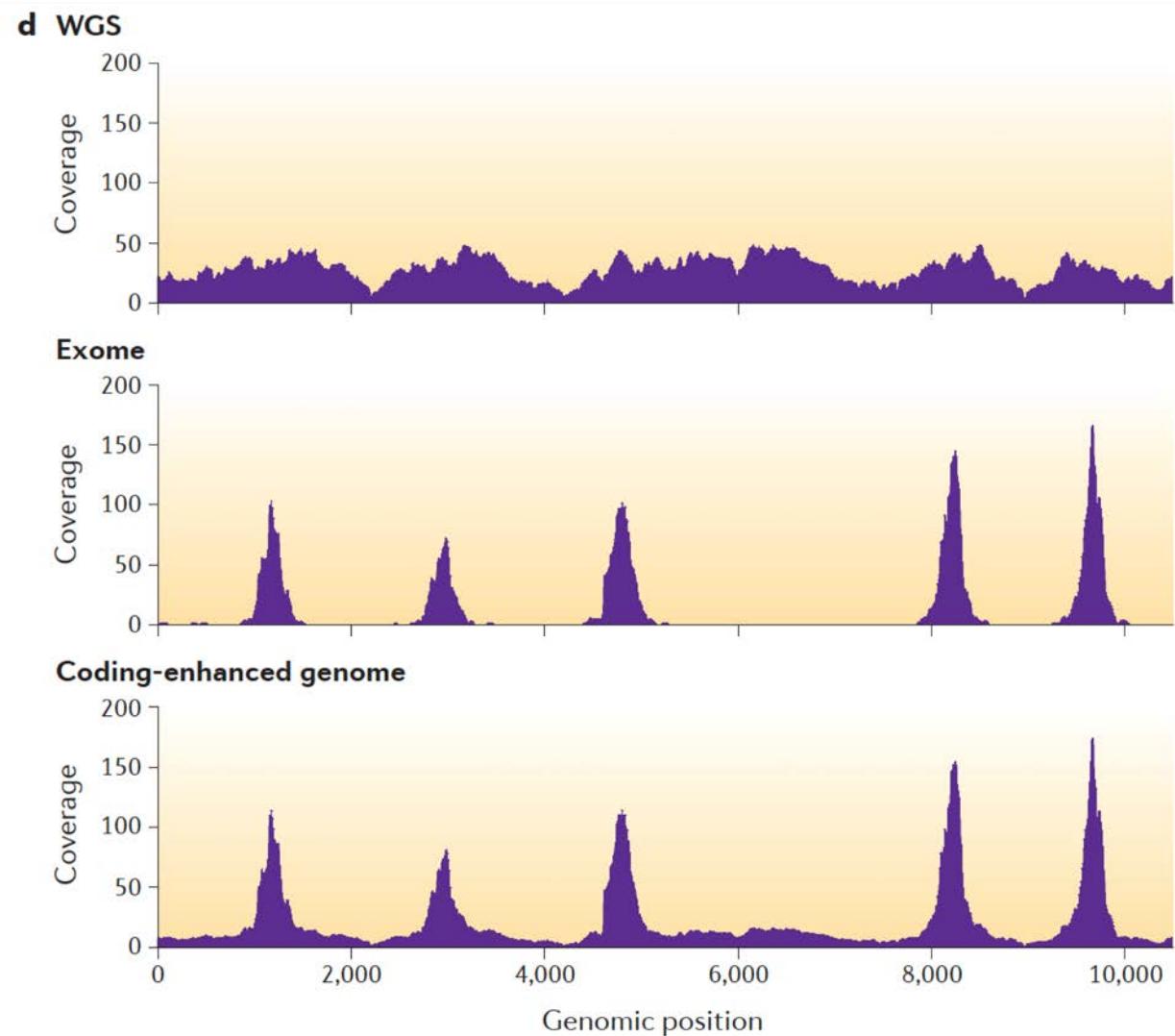
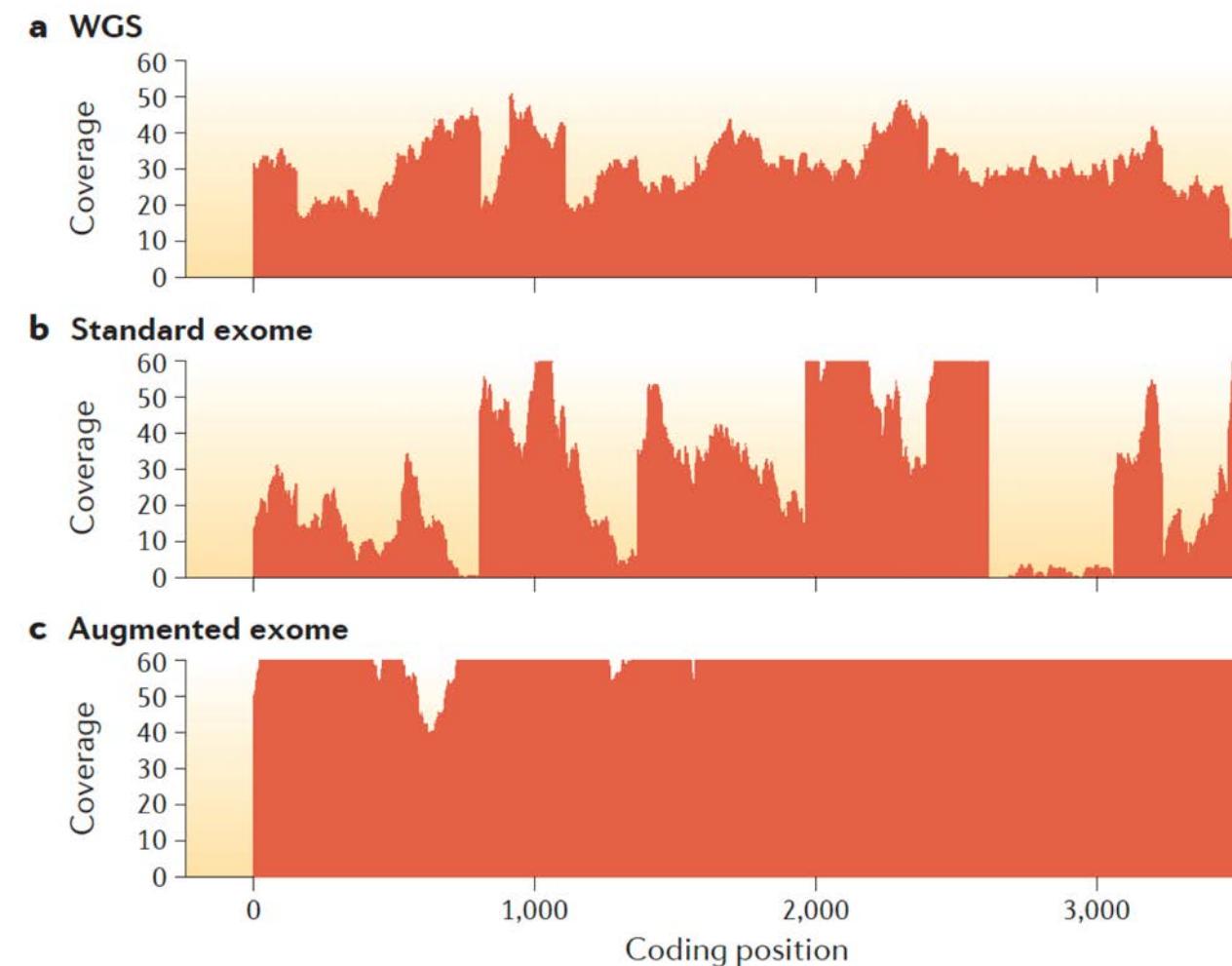
DMPK  
ATXN8/ATXN8OS  
JPH3

CTG  
CTG  
CTG

50 100 200 1,000 2,000 11,000

■ Unaffected  
■ Carrier  
■ Affected

# Resolution



# Transcriptomics / RNAseq

# Applications of RNAseq

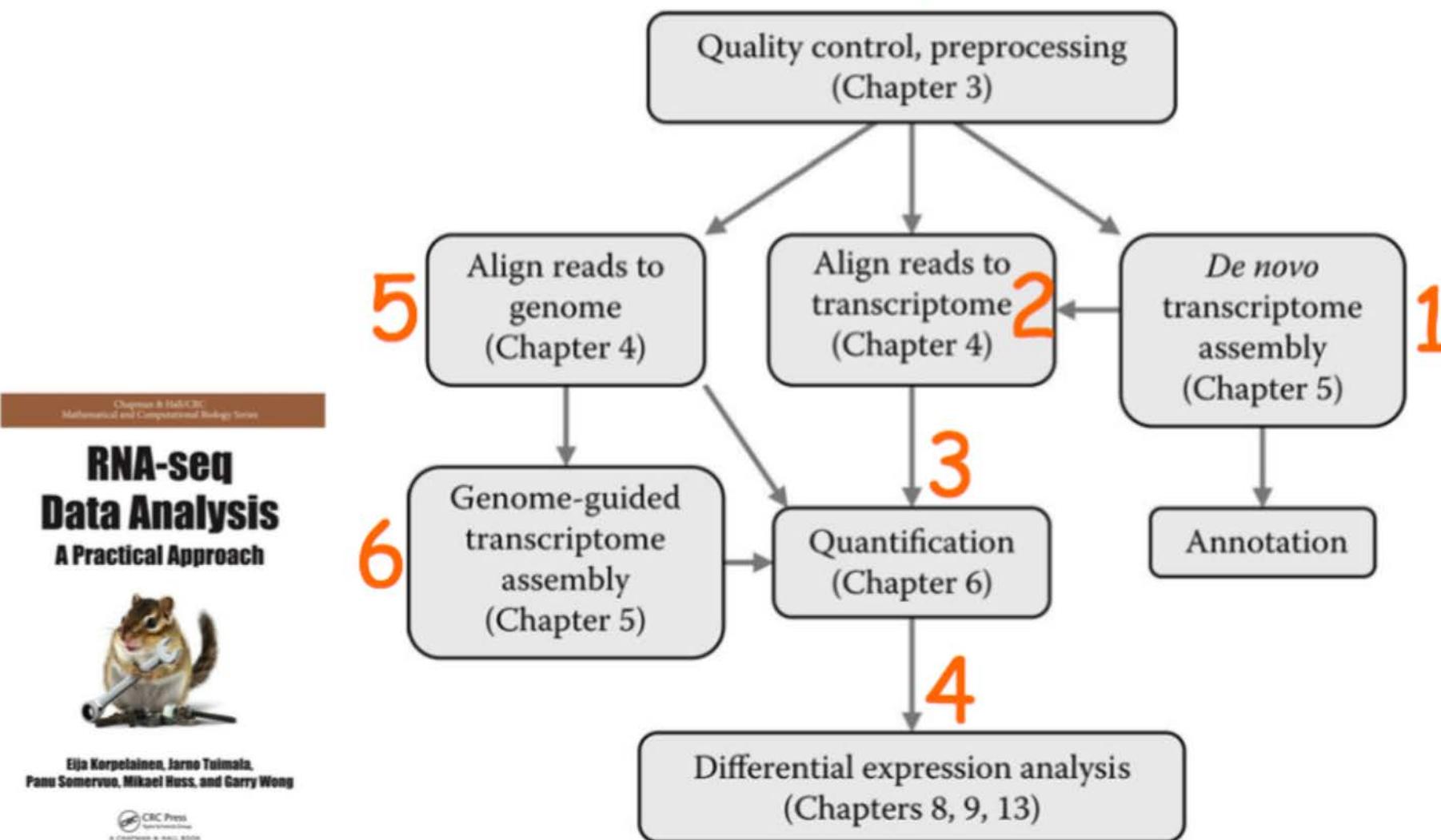
## **Discovery / Annotation**

- 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**

# Expression quantification and transcript assembly



September 9, 2014

E-Book - ISBN 9781482262353

FIGURE 2.1 Possible paths in RNA-seq data analysis. 3

---

# Introduction to differential gene expression analysis using RNA-seq

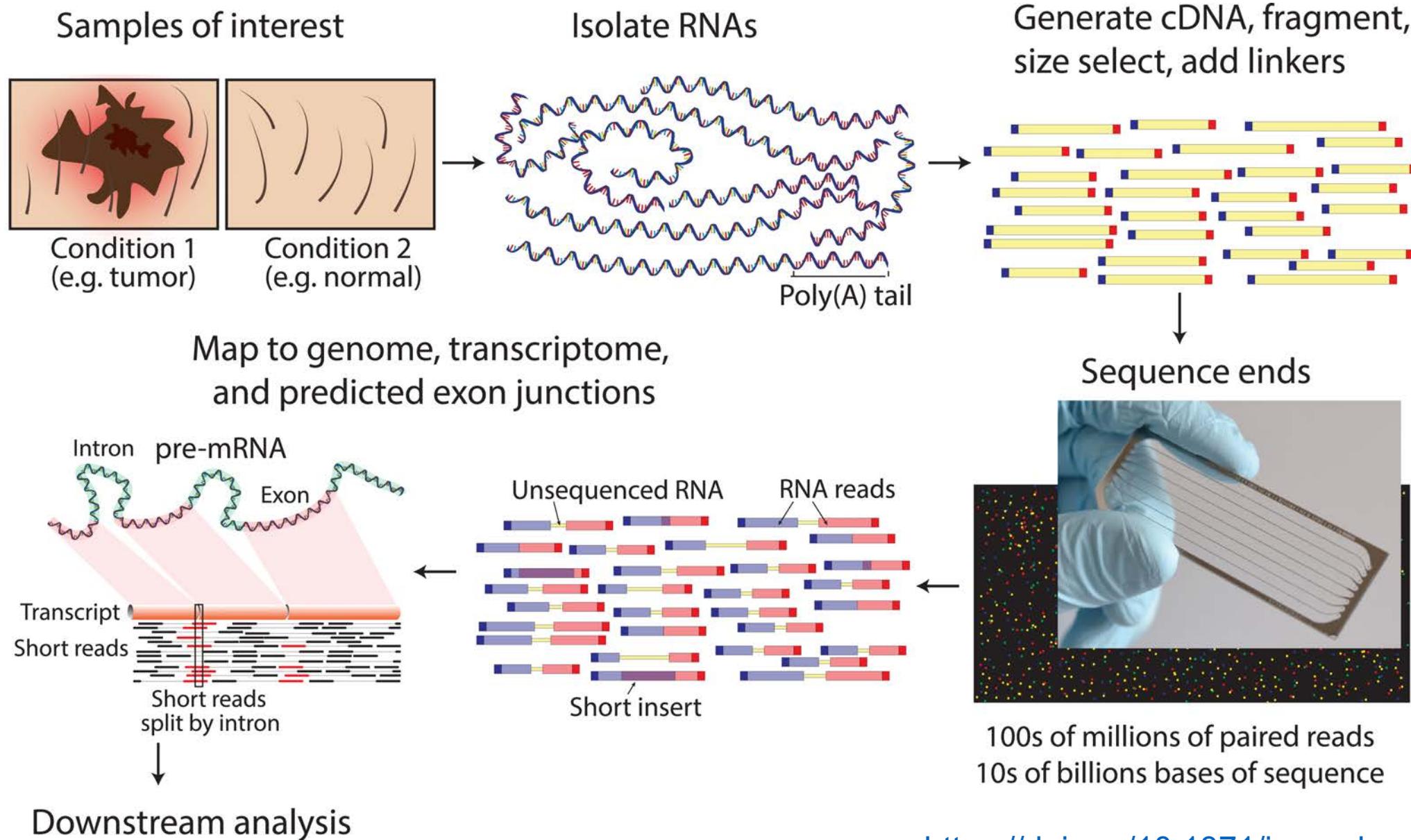
Written by Friederike Dündar, Luce Skrabaneck, Paul Zumbo

September 2015  
updated March 20, 2018

---

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

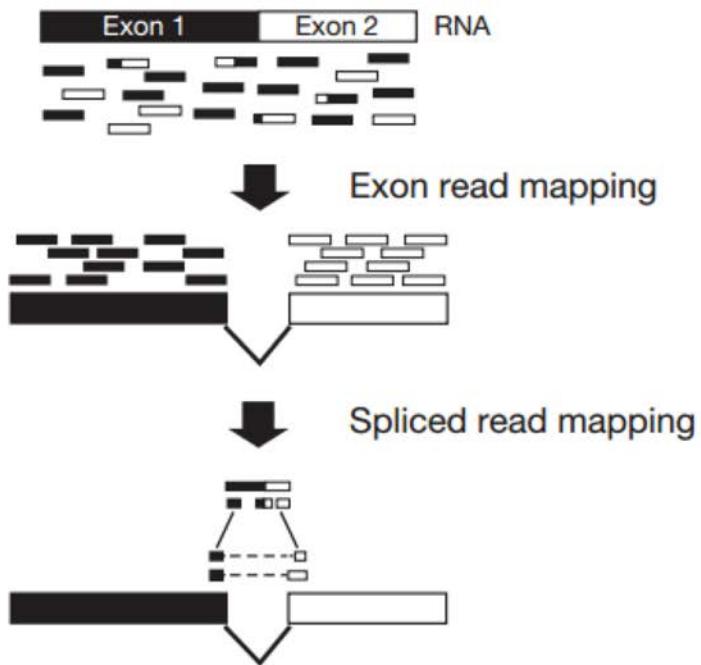
# RNA-seq data generation



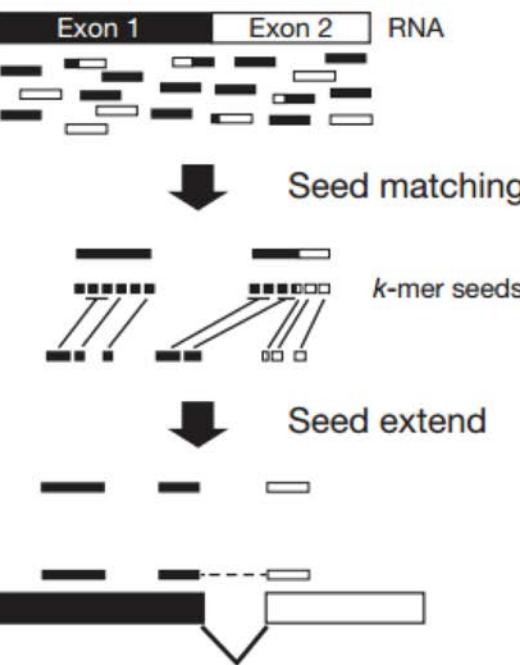
Once you have the raw data: **map** the RNAseq reads

# Strategies for gapped alignments of RNAseq reads

## a Exon-first approach

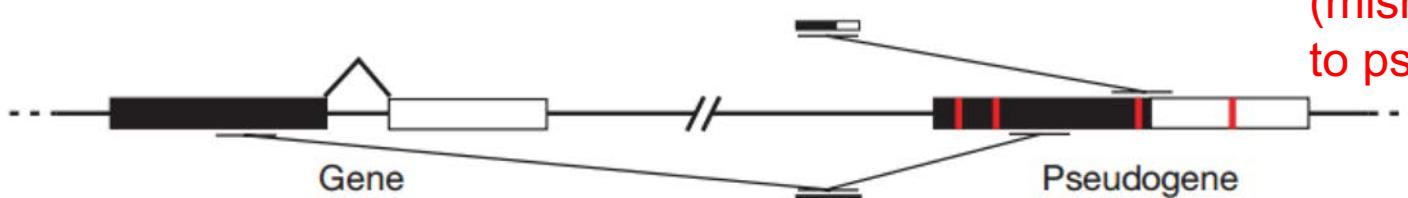


## b Seed-extend approach



## c Potential limitations of exon-first approaches

Preferential alignment  
(mismatch rather than split)  
to pseudogene

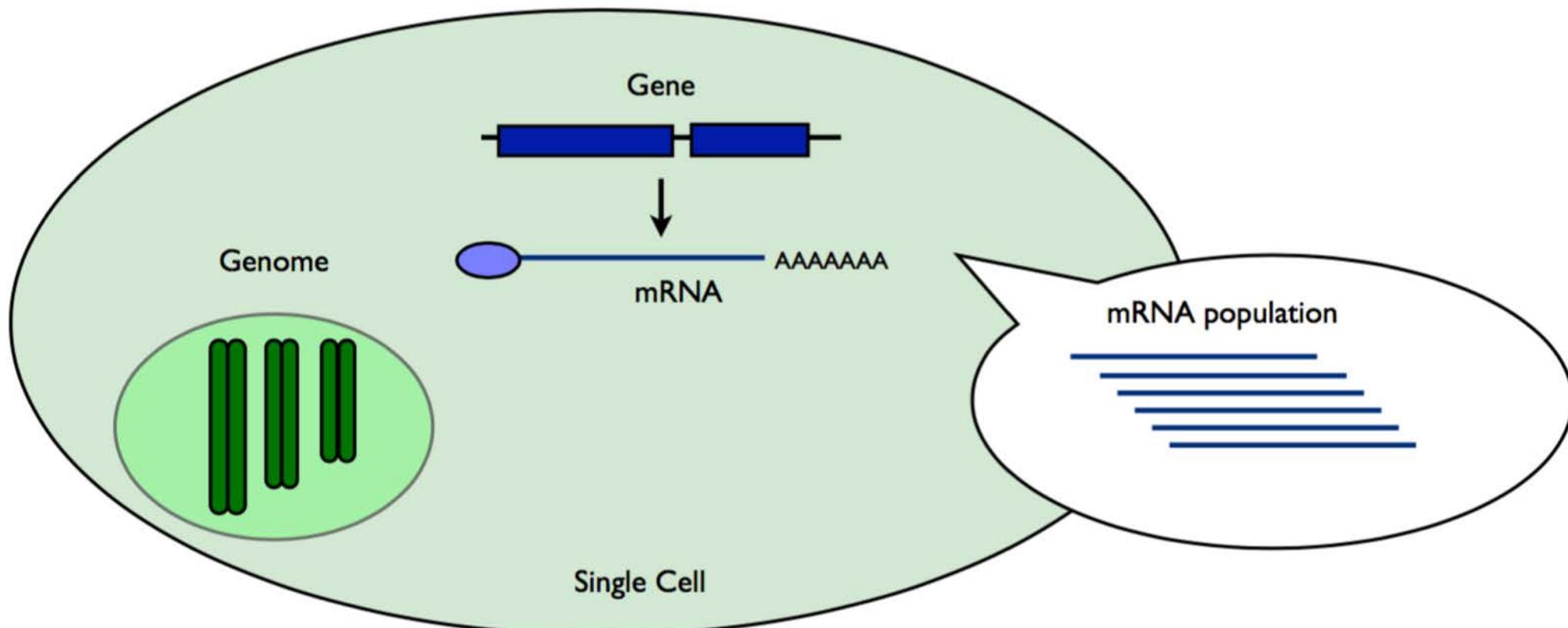


# Types of experiments

# Transcriptome Complexity:

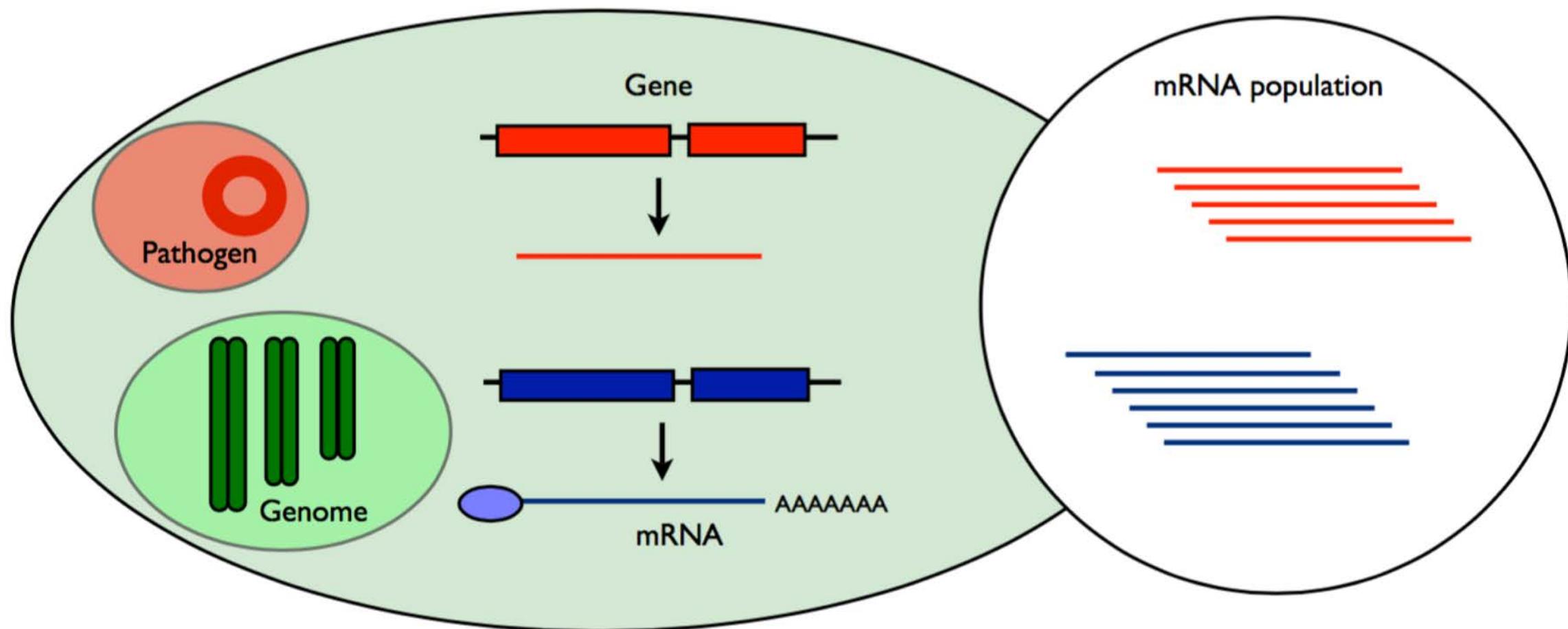
## Simple System:

**One Genome => Gene 1 copy => Single mRNA**



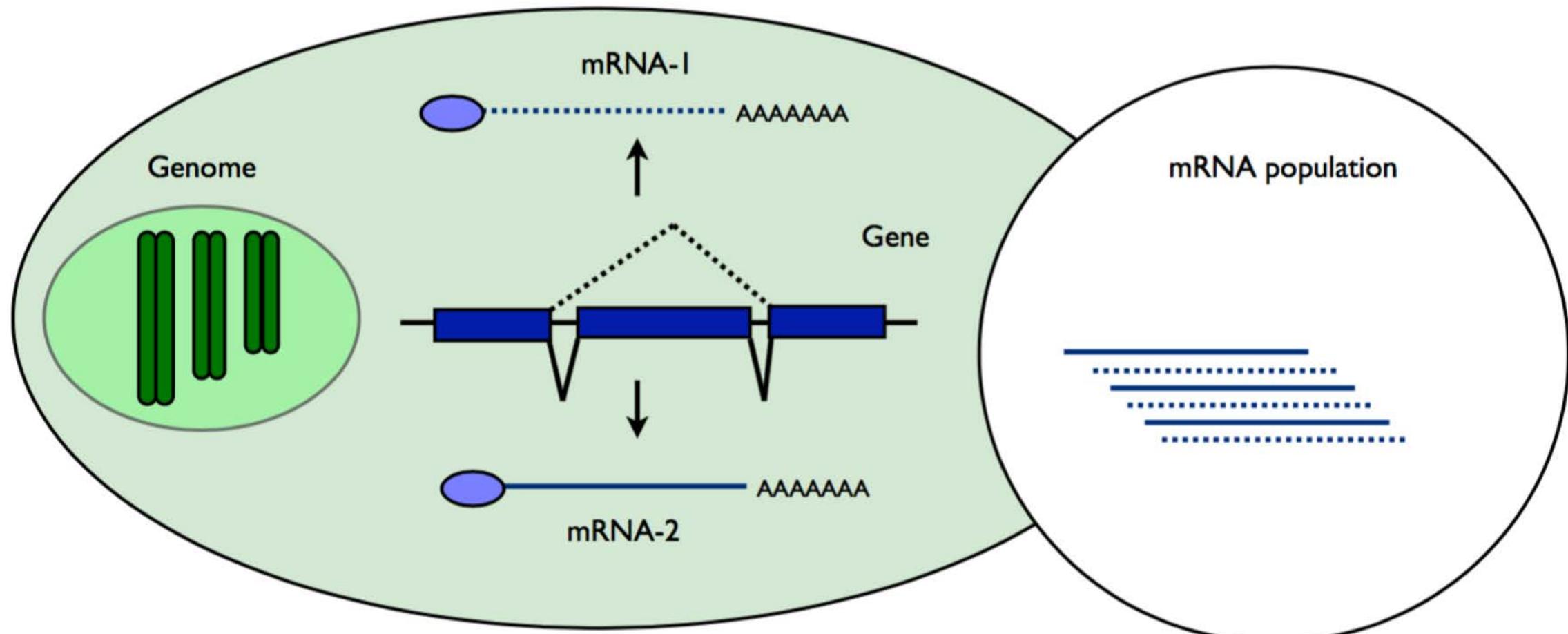
# How many species we are analyzing ?

- 1) Problems to isolate a single species (rhizosphere)
- 2) Species interaction study (plant-pathogen)



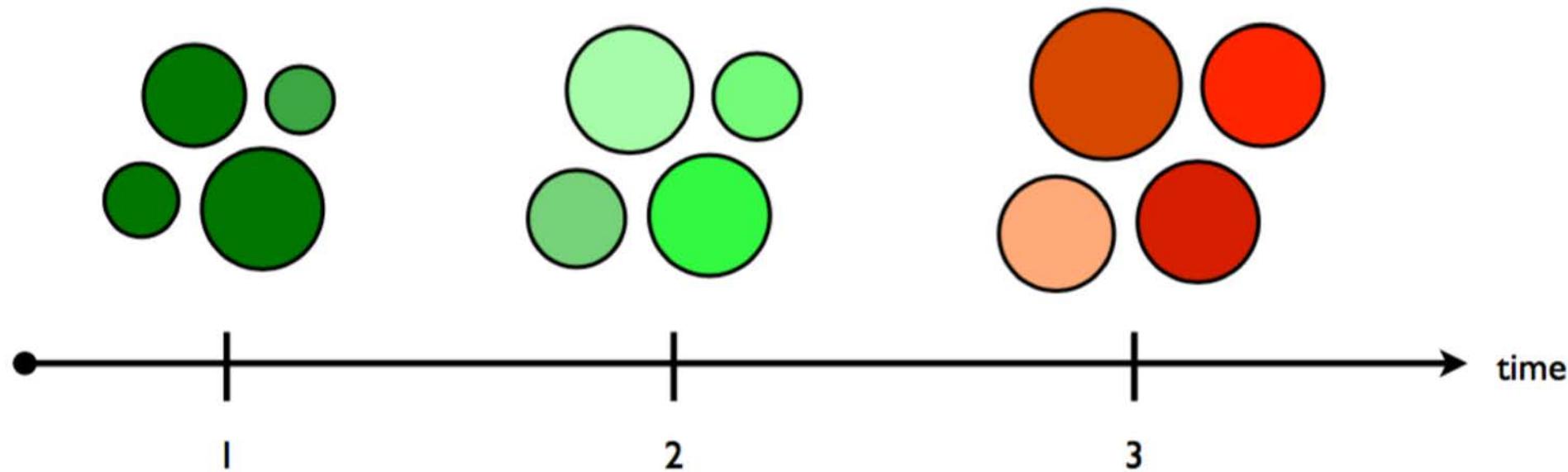
# How many isoforms we expect for each allele ?

## 1) Alternative splicings



**Is the study performed at different time points?**

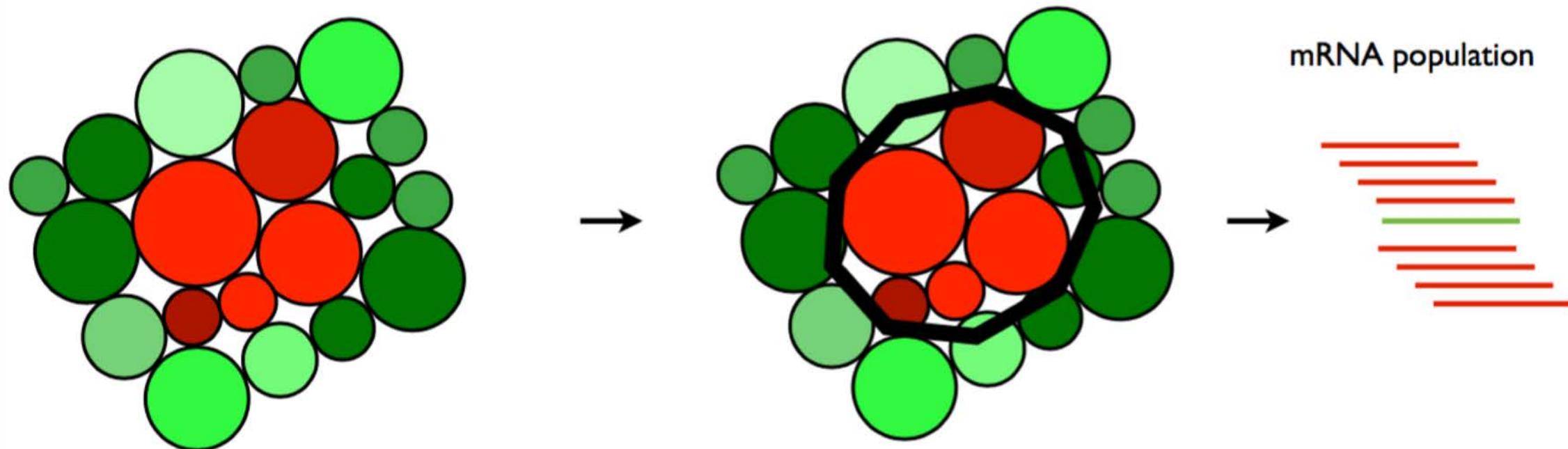
- 1) Developmental stages (difficult to select the same)**
- 2) Response to a treatment**



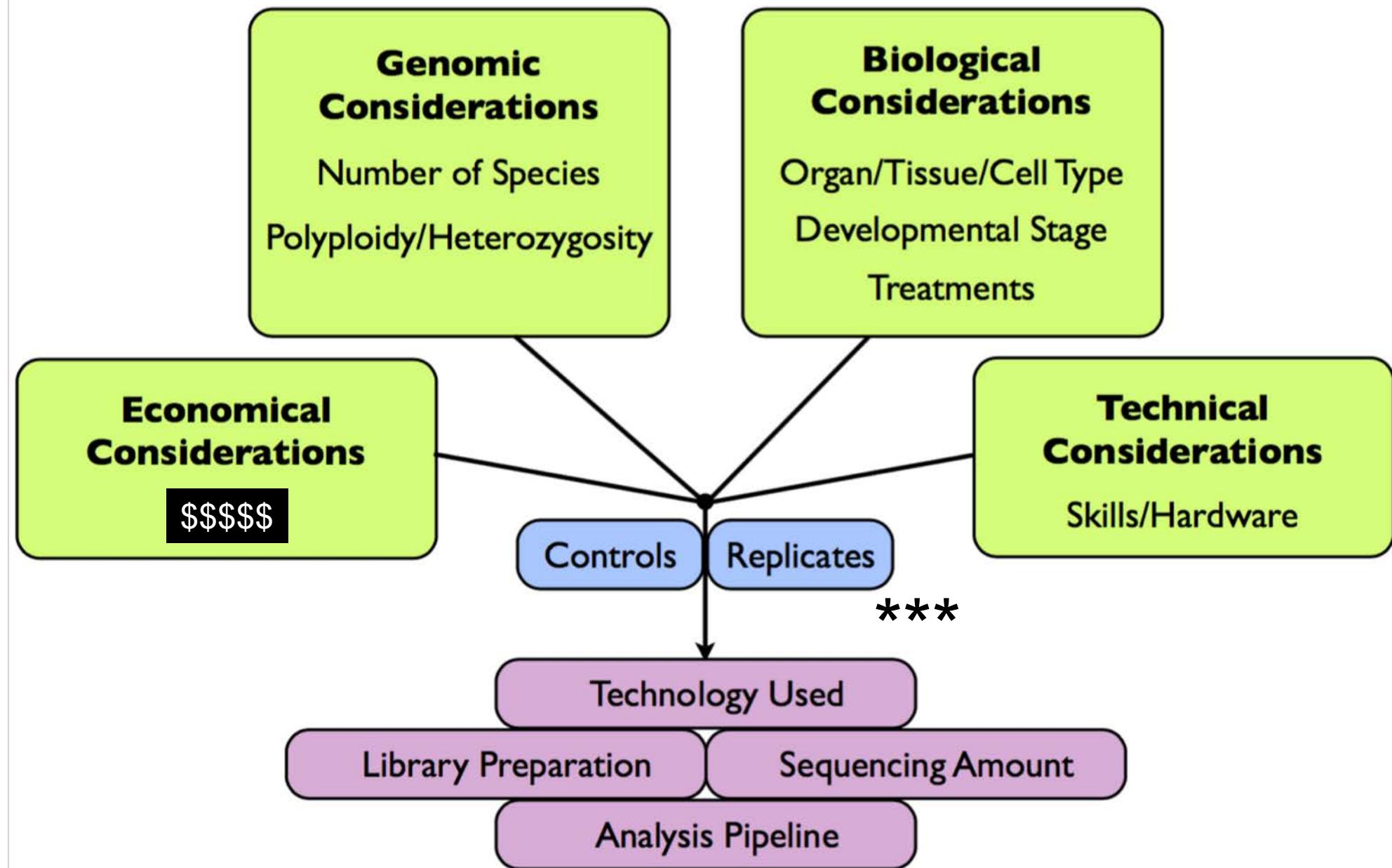
**Is the study performed with different parts?**

- 1) Organ specific**
- 2) Tissue/Cell type specific**

**(Laser Capture Microdissection, LCM)**



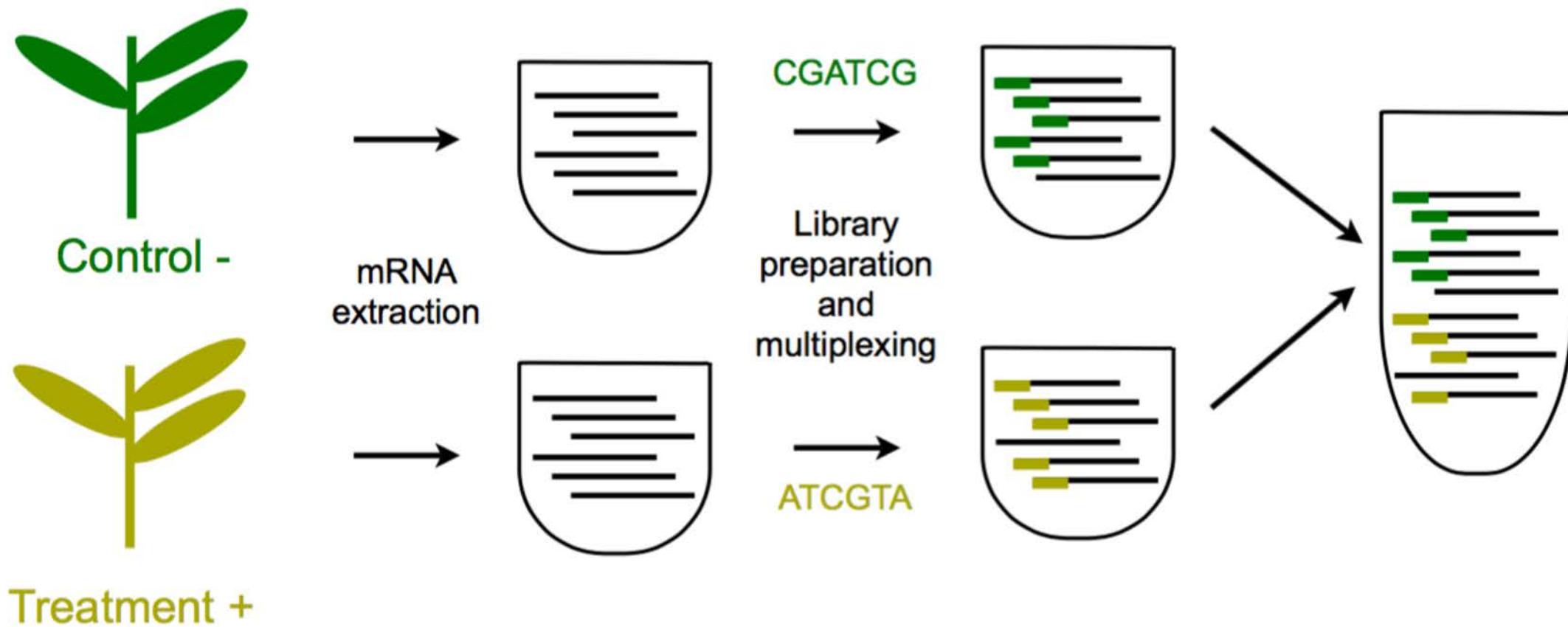
# Experimental design



# Prep and treatment

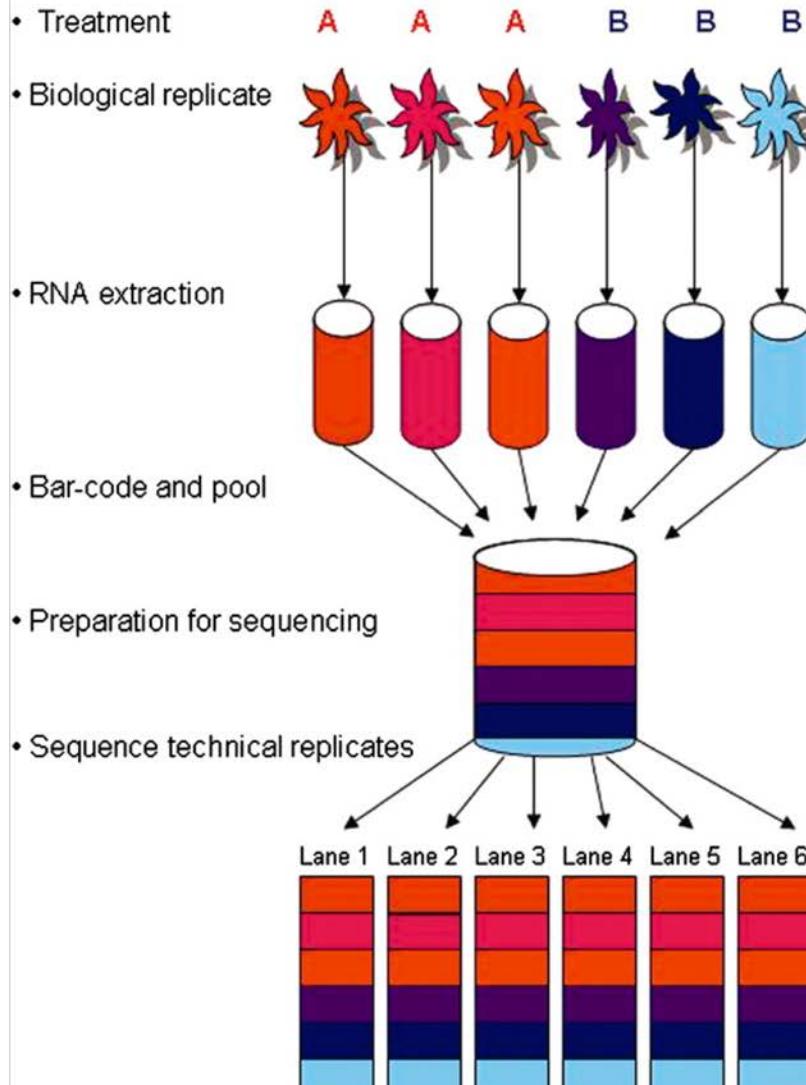
Sequencing of multiple samples can be performed using **multiplexing**.

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.

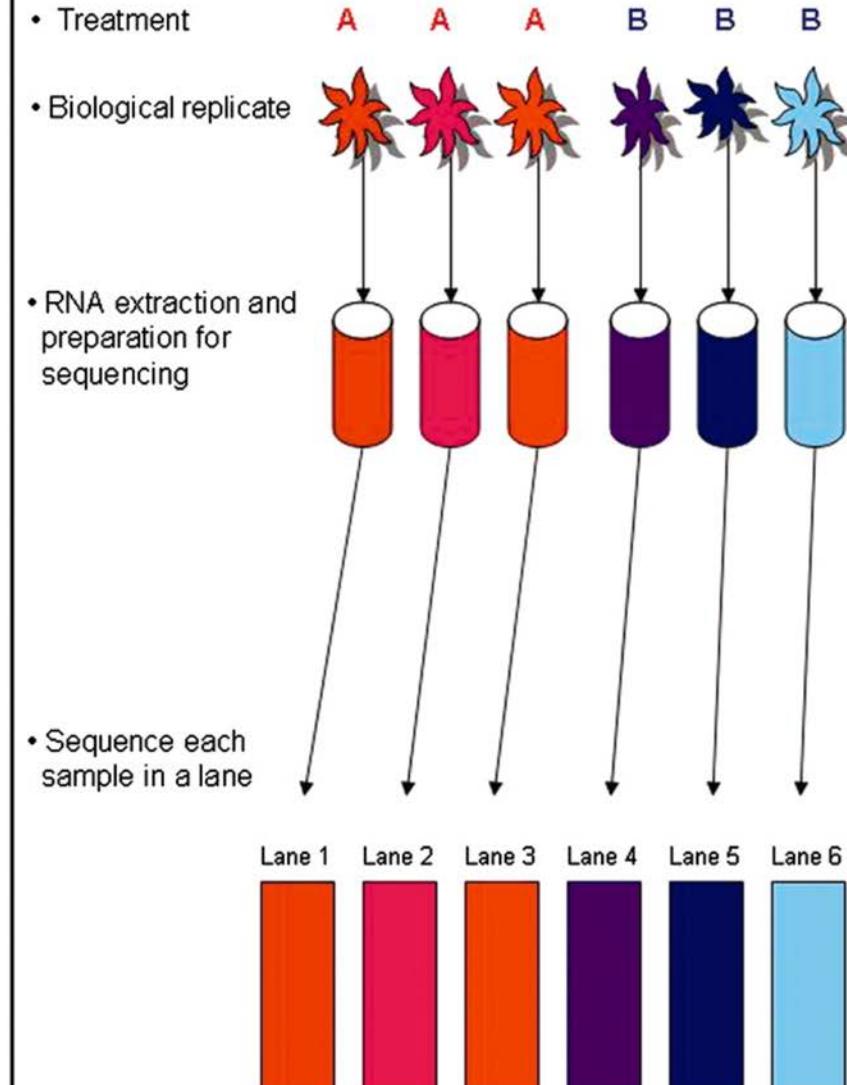


# You need to design experiment carefully

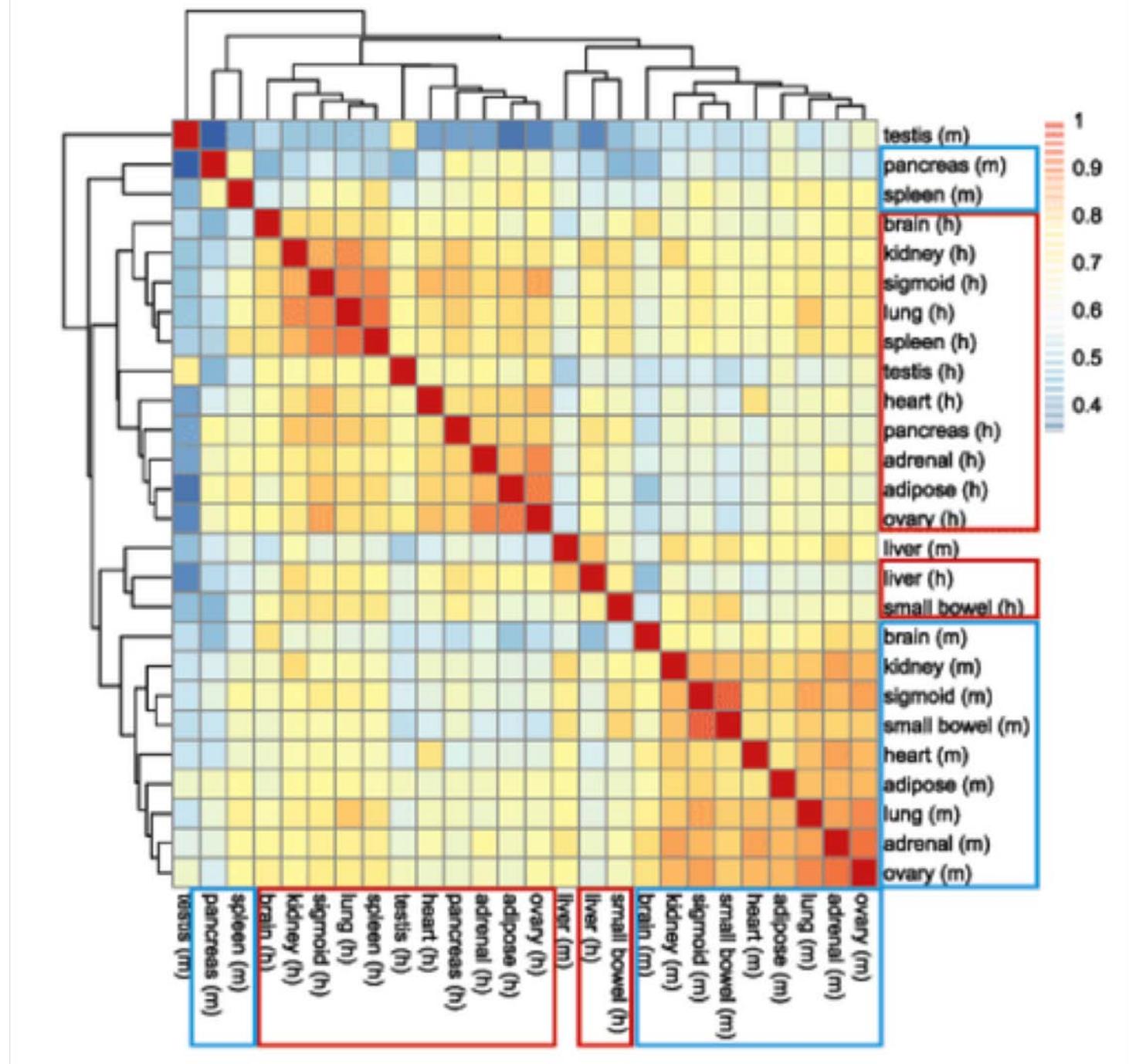
## Balanced Blocked Design



## Confounded Design



# Example of batch effect:



# Example of batch effect:

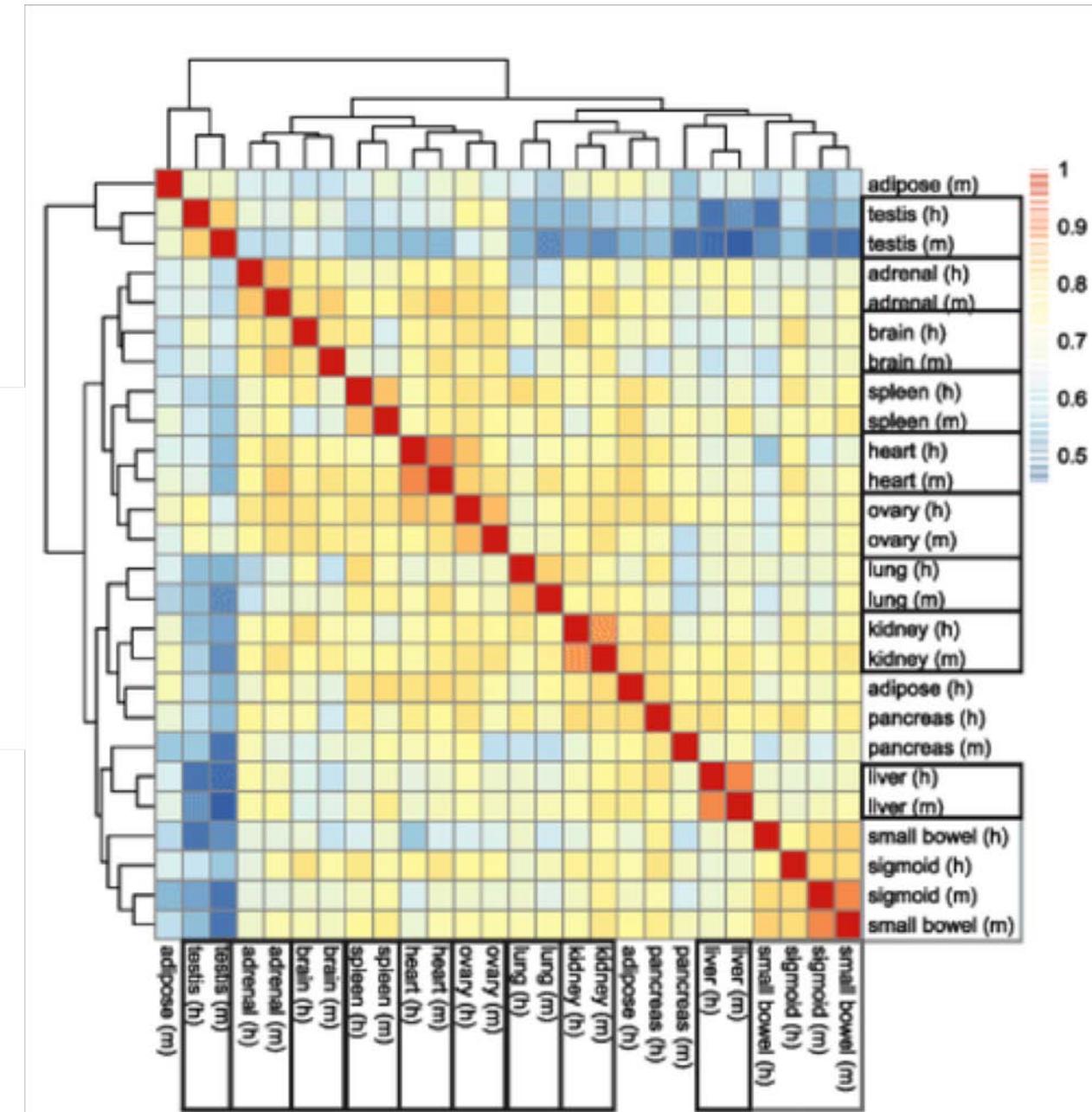


Yoav Gilad  
@Y\_Gilad



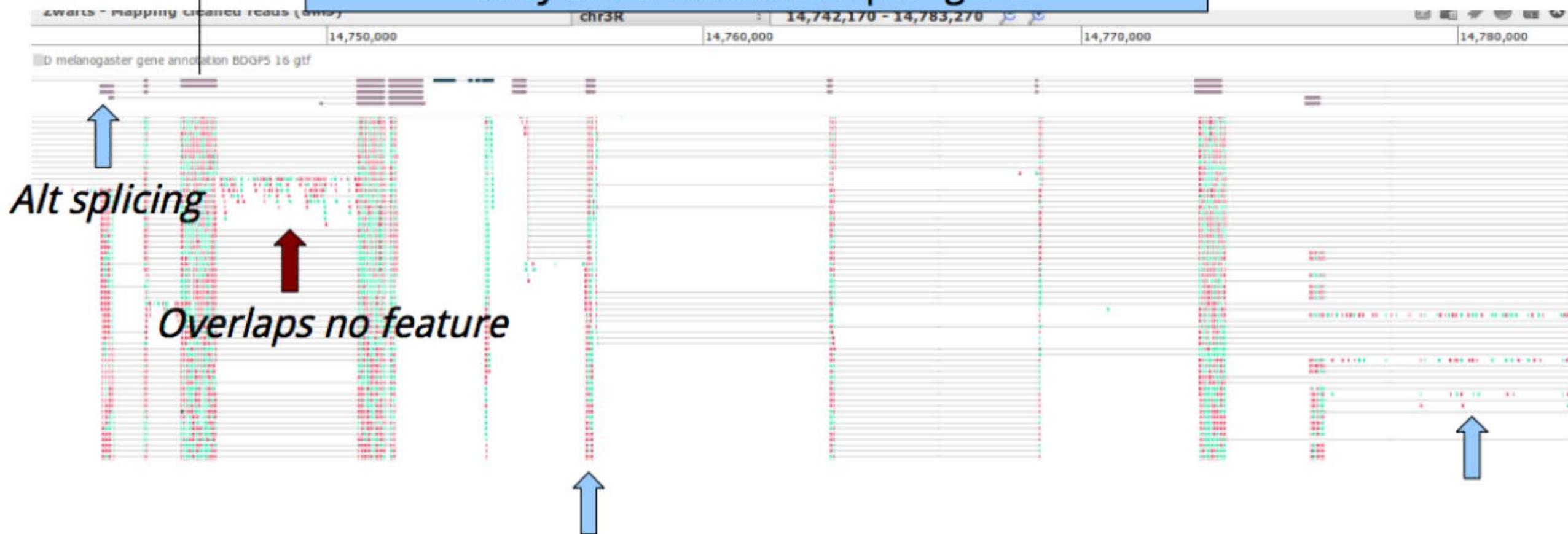
Following

We reanalyzed the data from  
[pnas.org/content/111/48...](http://pnas.org/content/111/48/) and found the  
following:



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

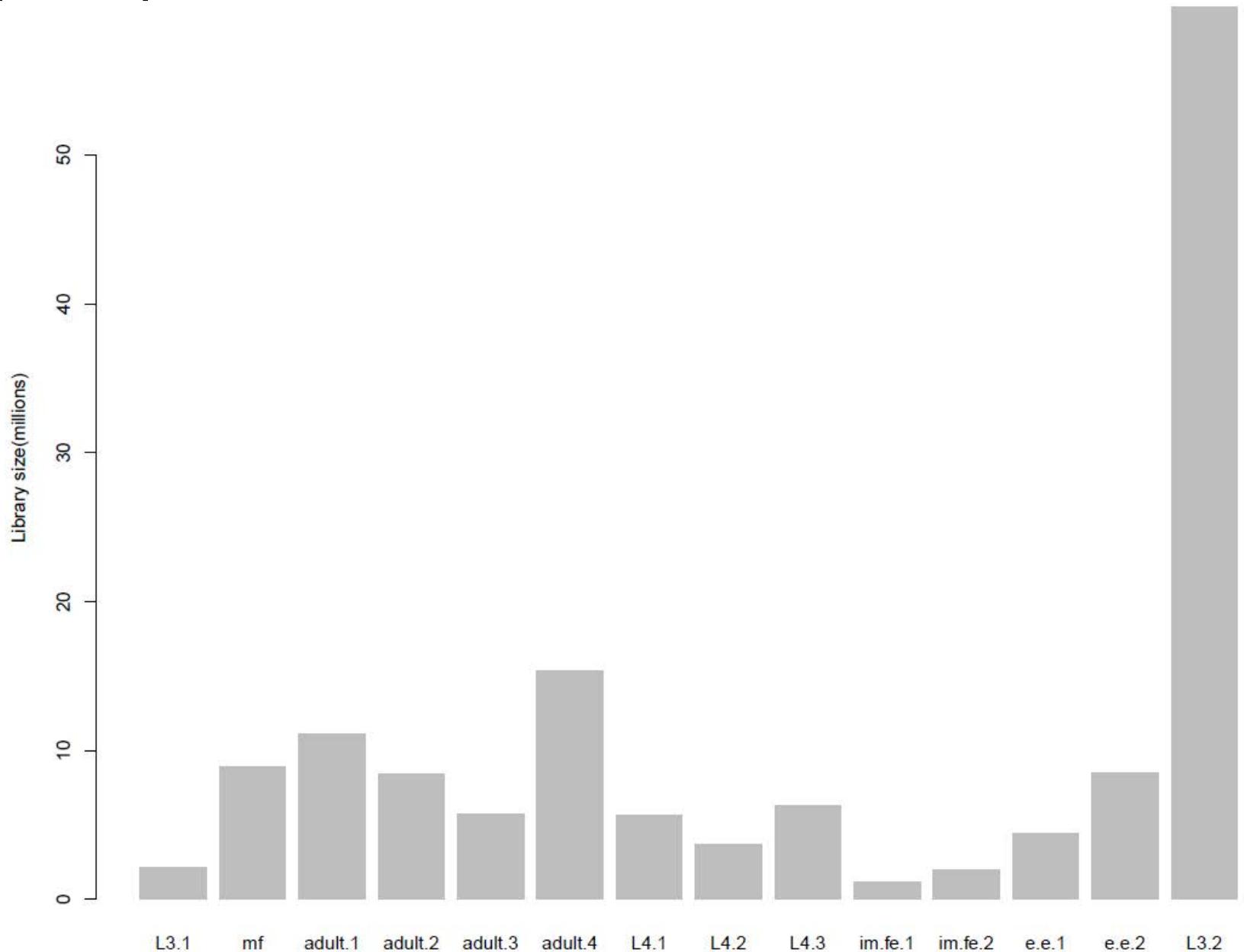
GeneB = exon 1 + exon 2 + exon 3 = 180 reads

This is the bit we care about!

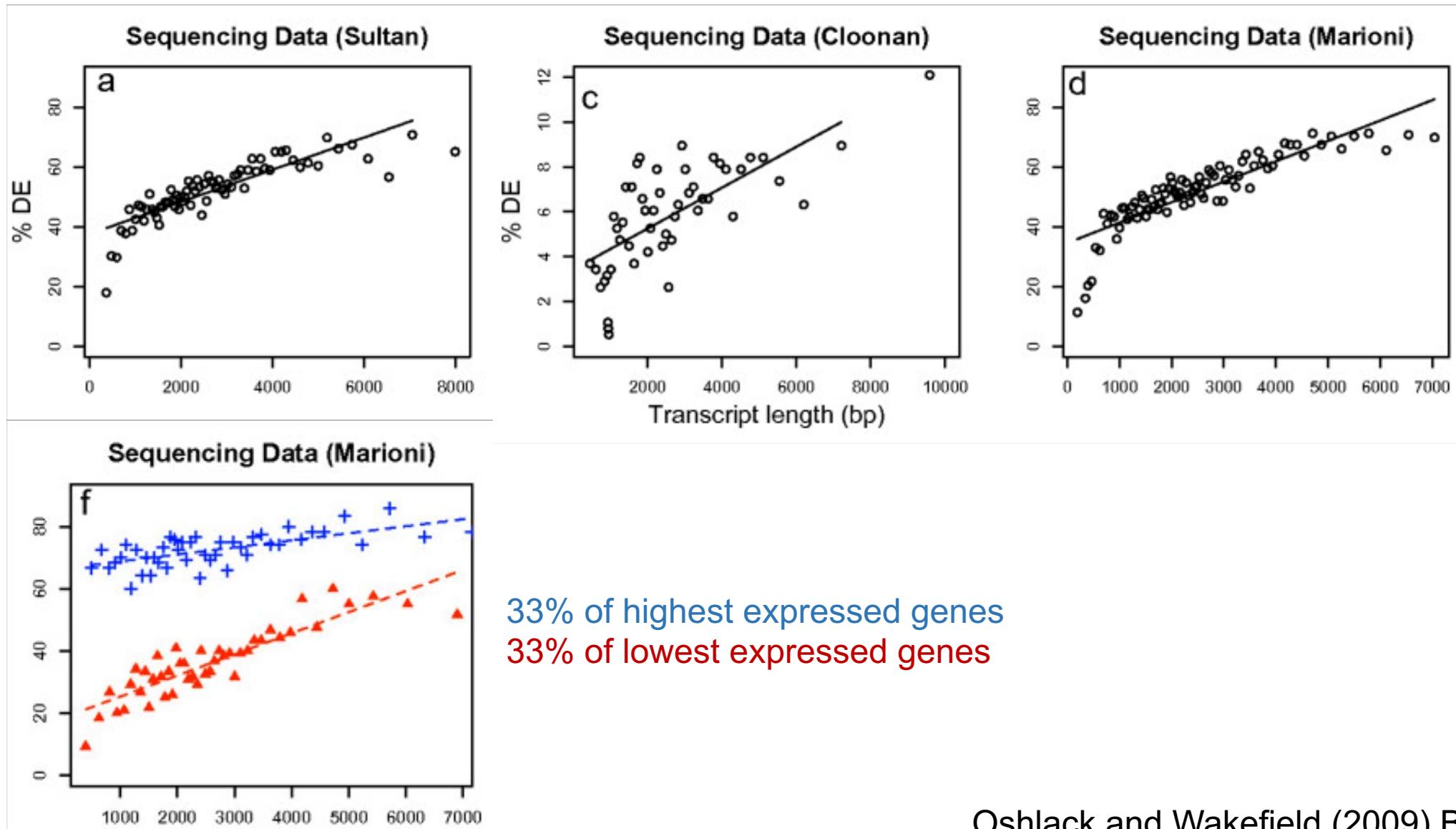


Counts of the gene depends on **expression** ,transcript length  
,sequencing depth and simply chance

# Higher sequencing depth equals more counts



# Counts are proportional to the transcript length x mRNA expression level



# 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):** (Law et 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

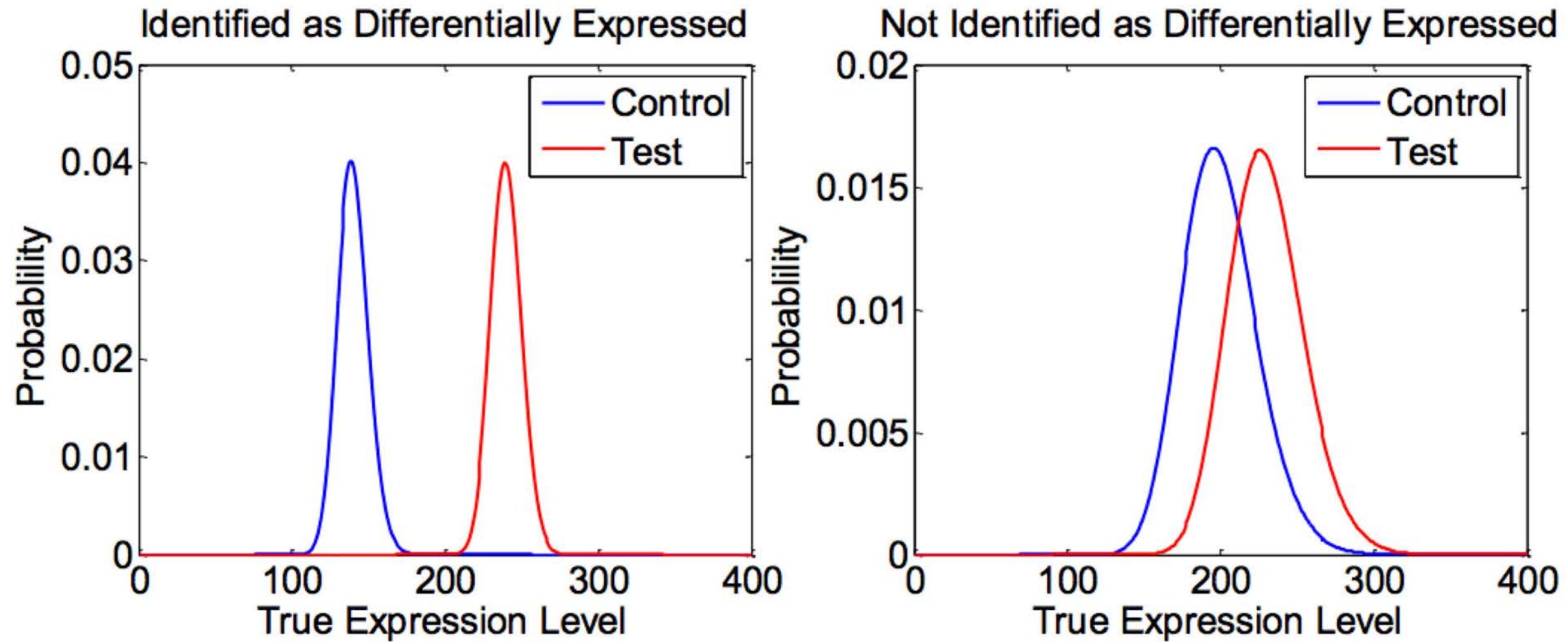
Published: Nov 06, 2013 • DOI: 10.1371/journal.pone.0077885

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

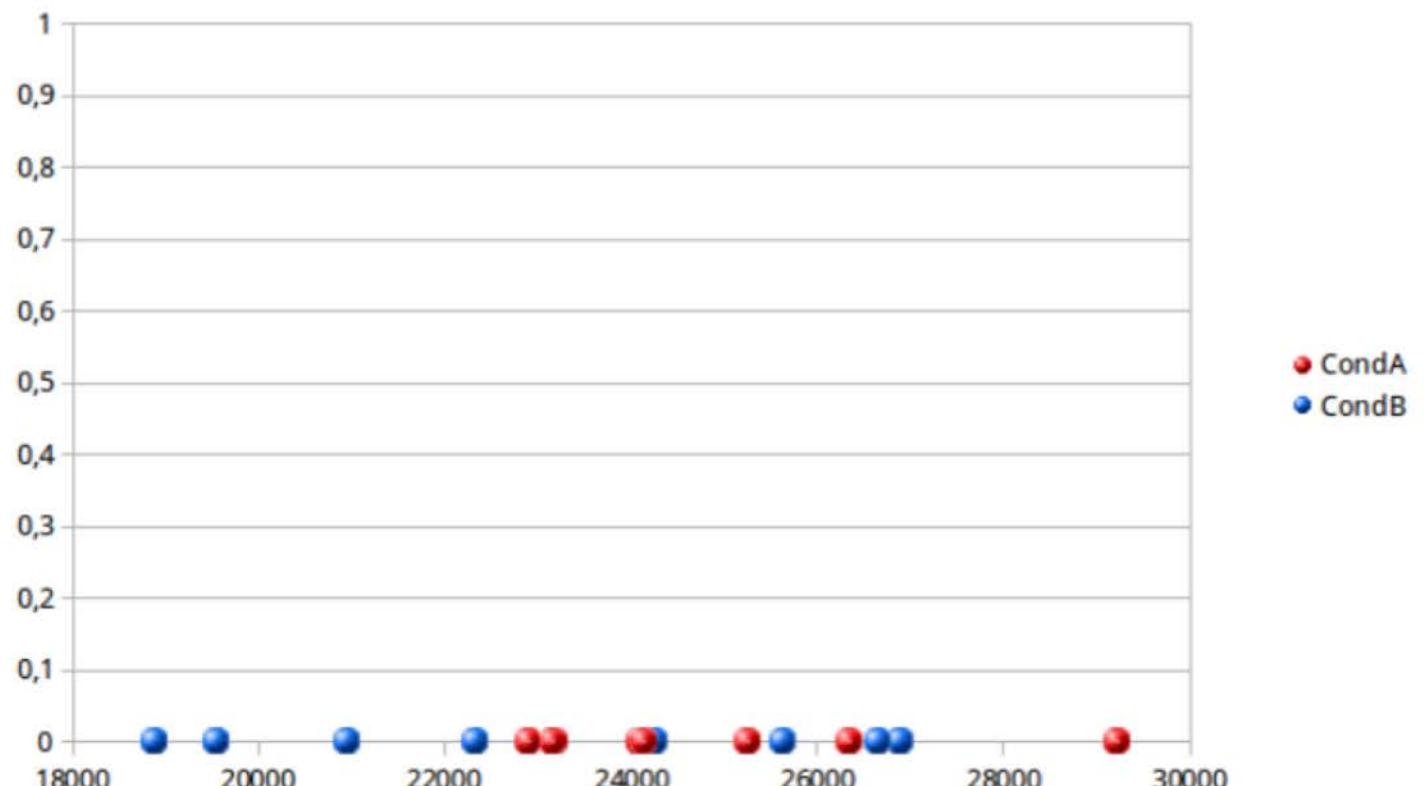
# Fitting a distribution for every gene for DE



# Scenario

gene\_id CAF0006876

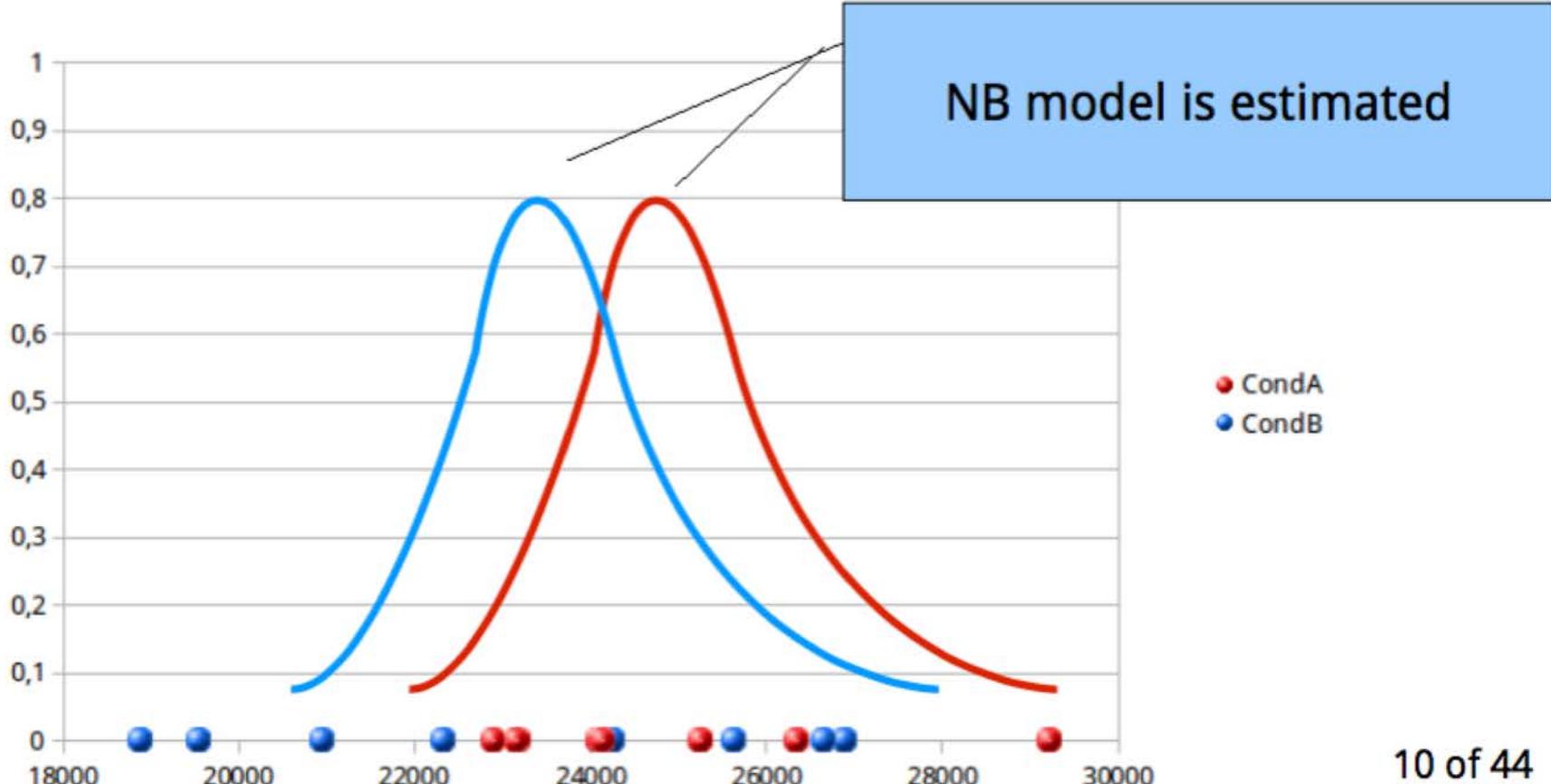
	sample1 23171	sample2 22903	sample3 29227	sample4 24072	sample5 23151	sample6 26336	sample7 25252	sample8 24122
Condition A								
Condition B	Sample9 19527	sample10 26898	sample11 18880	sample12 24237	sample13 26640	sample14 22315	sample15 20952	sample16 25629



# Scenario

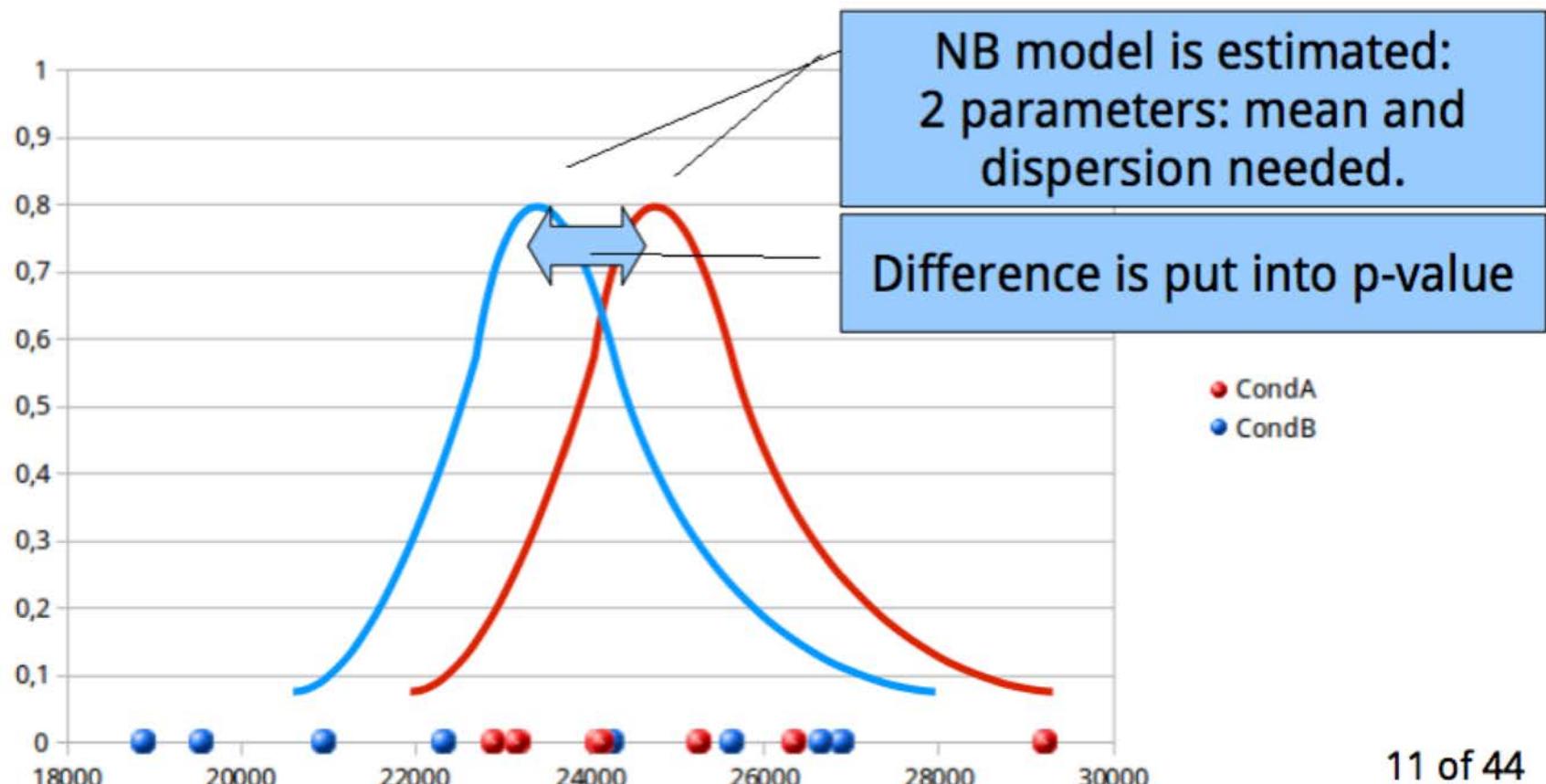
gene\_id CAF0006876

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

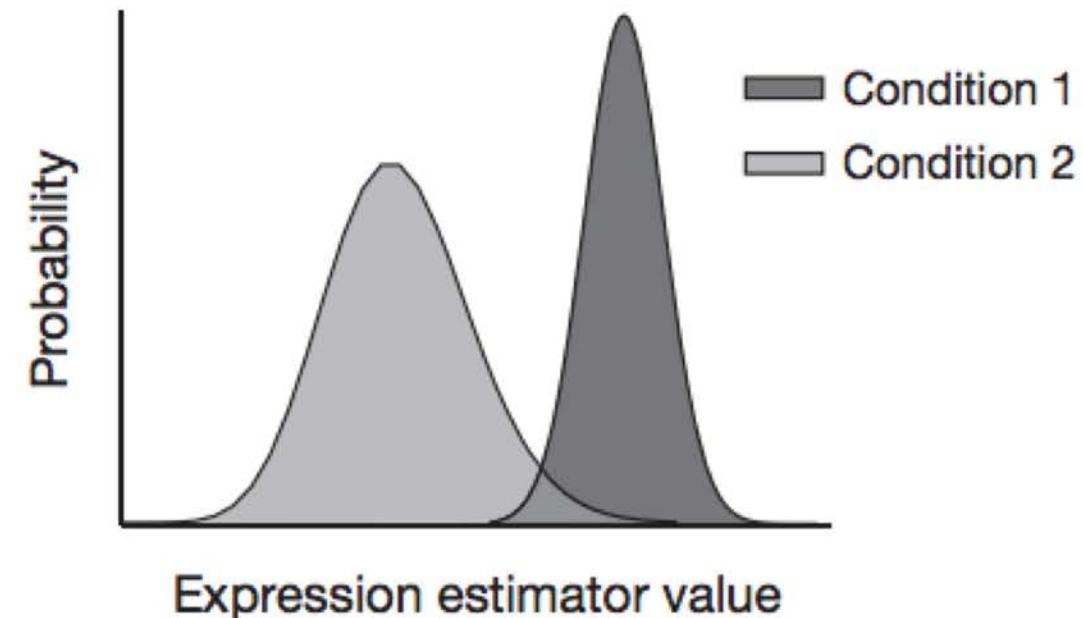
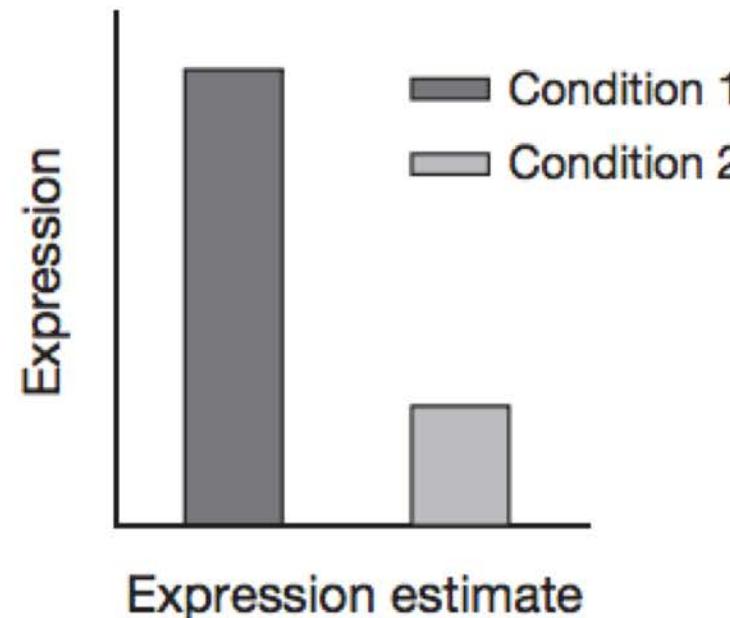
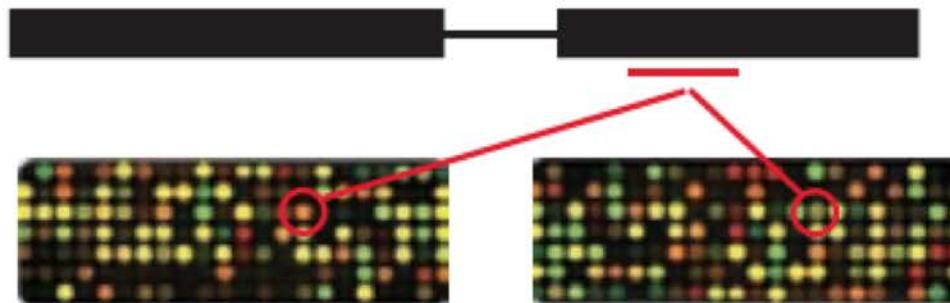


# Scenario

gene_id	CAF0006876	sample1	sample2	sample3	sample4	sample5	sample6	sample7	sample8
<b>Condition A</b>	23171	22903	29227	24072	23151	26336	25252	24122	
<b>Condition B</b>	19527	26898	18880	24237	26640	22315	20952	25629	

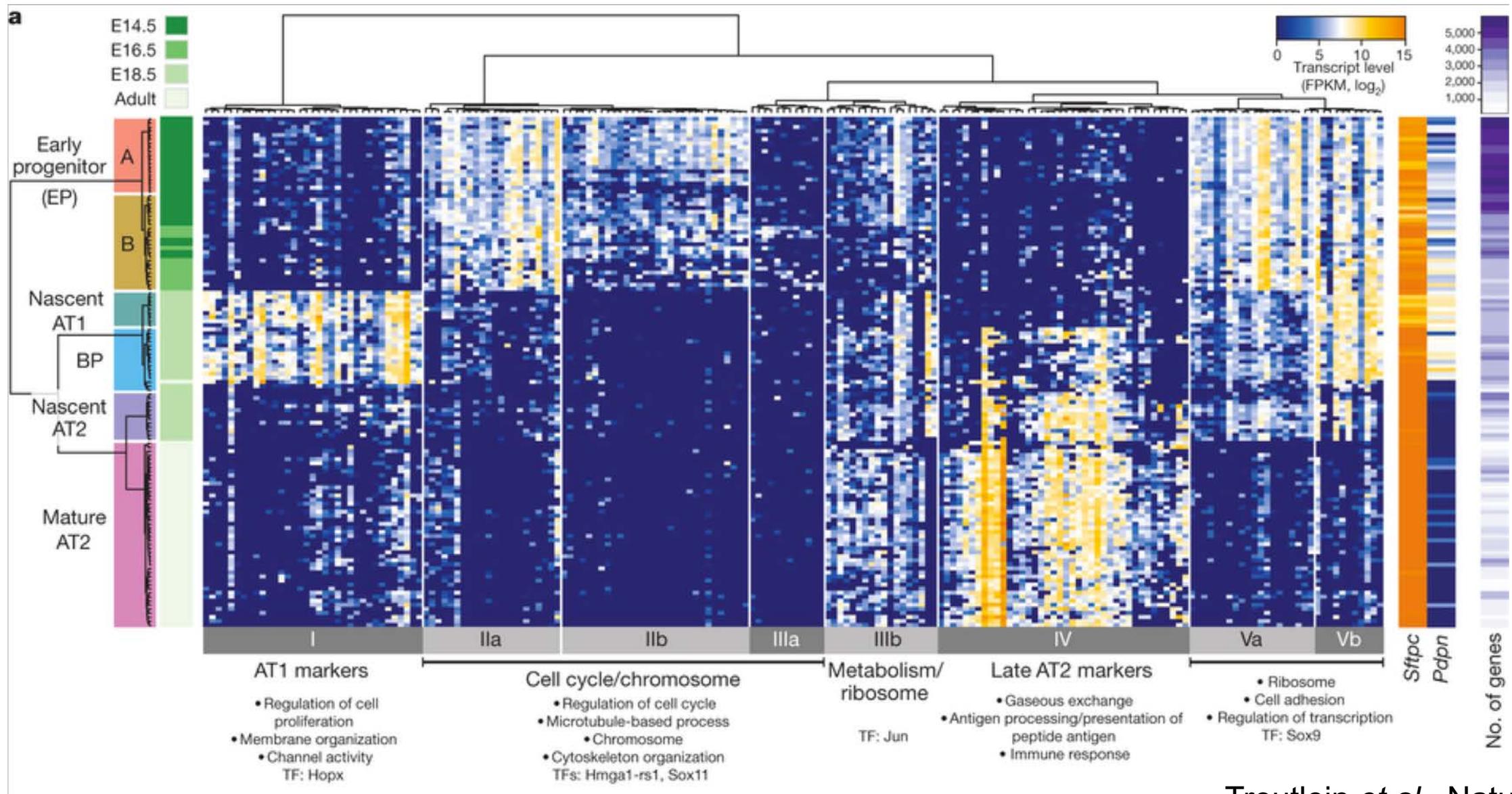


# RNAseq vs Microarray



Once you have set of differentially expressed genes

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



# Amplicon / Metagenomics: An Intro



# A clinician's guide to microbiome analysis

*Marcus J. Claesson<sup>1,2\*</sup>, Adam G. Clooney<sup>1-3\*</sup> and Paul W. O'Toole<sup>1,2</sup>*

**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.

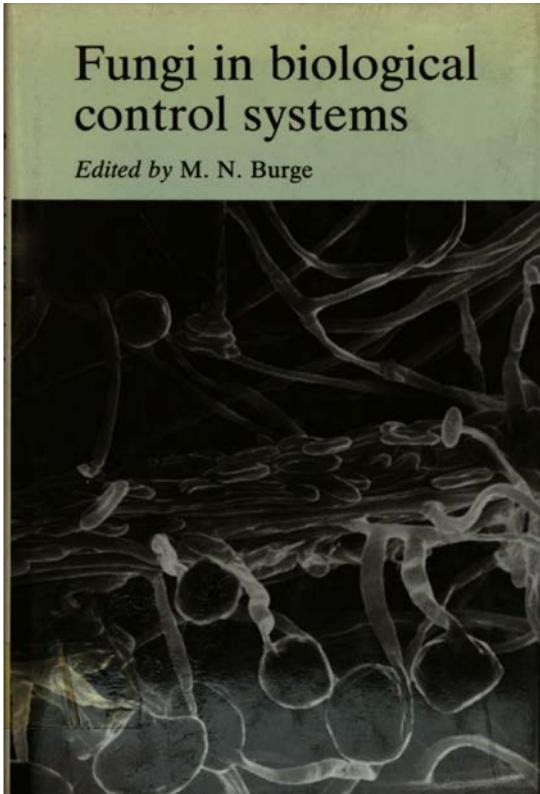
## **Key points**

---

- 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

# 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?

Crosstalk R245

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

Jo Handelsman<sup>1</sup>, Michelle R Rondon<sup>1</sup>, Sean F Brady<sup>2</sup>, Jon Clardy<sup>2</sup> and Robert M Goodman<sup>1</sup>

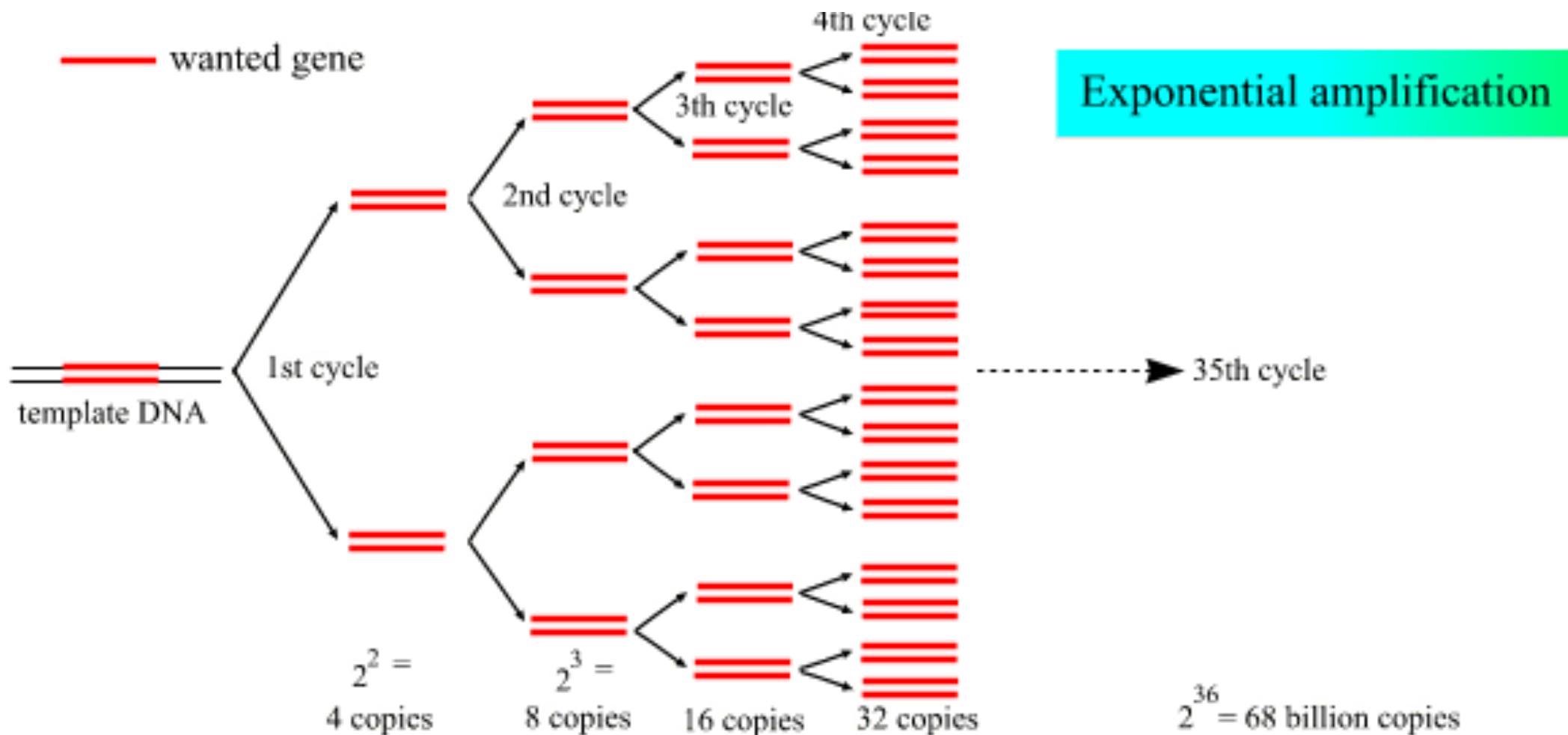


**Chemistry & Biology** October 1998, 5:R245–249  
<http://biomednet.com/elecref/10745521005R0245>

**... 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)



# And then what is the metagenome?

OPEN  ACCESS Freely available online

PLOS COMPUTATIONAL BIOLOGY

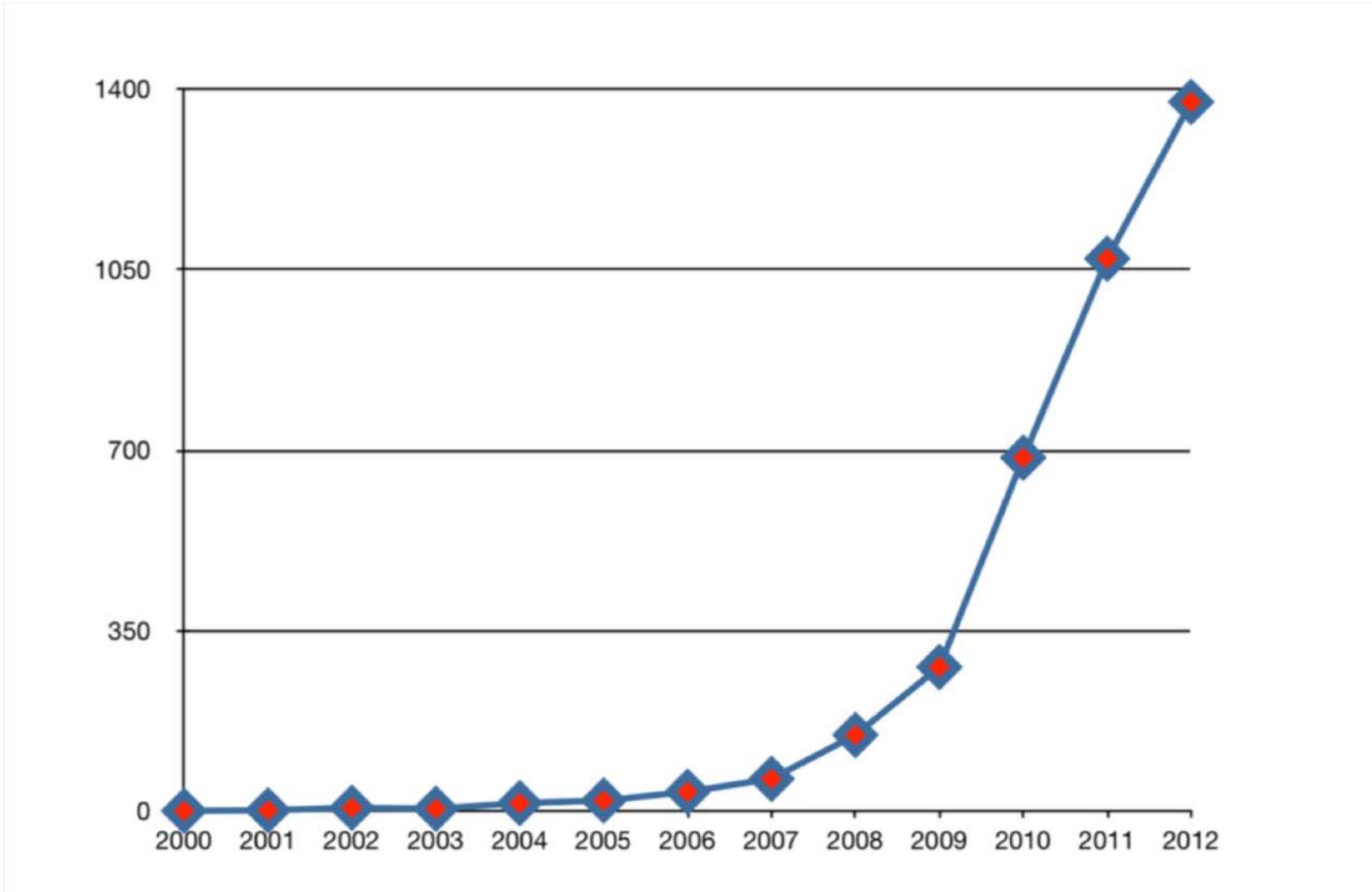
Review

## 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”



Metagenomics  $\neq$  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

# Basic Purpose

Characteristics of (microbial) community

**Who** are they?

**Where** do they come from?

Are their similarities (at what level)

between communities

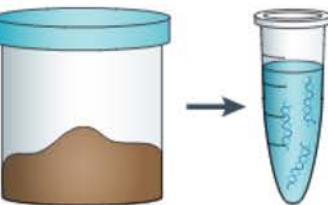
of different conditions

of similar conditions?

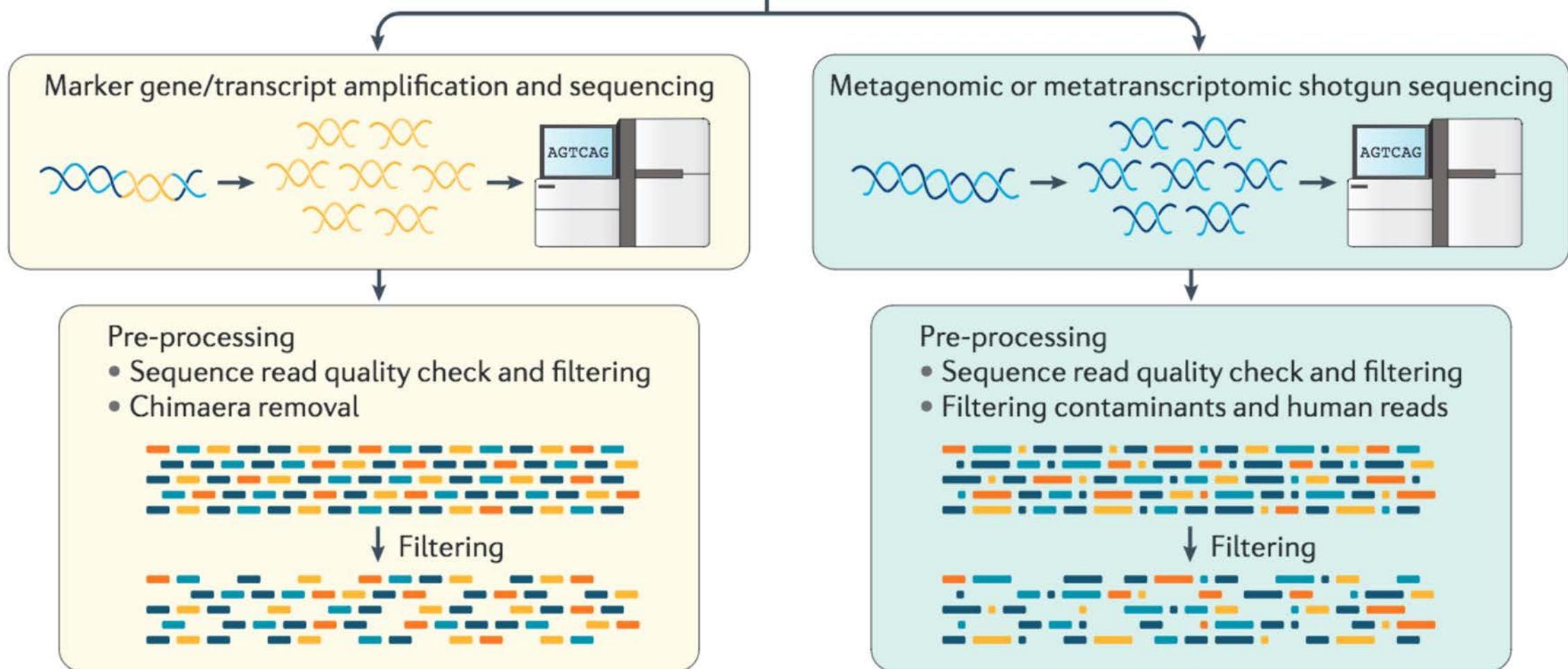
within a community?

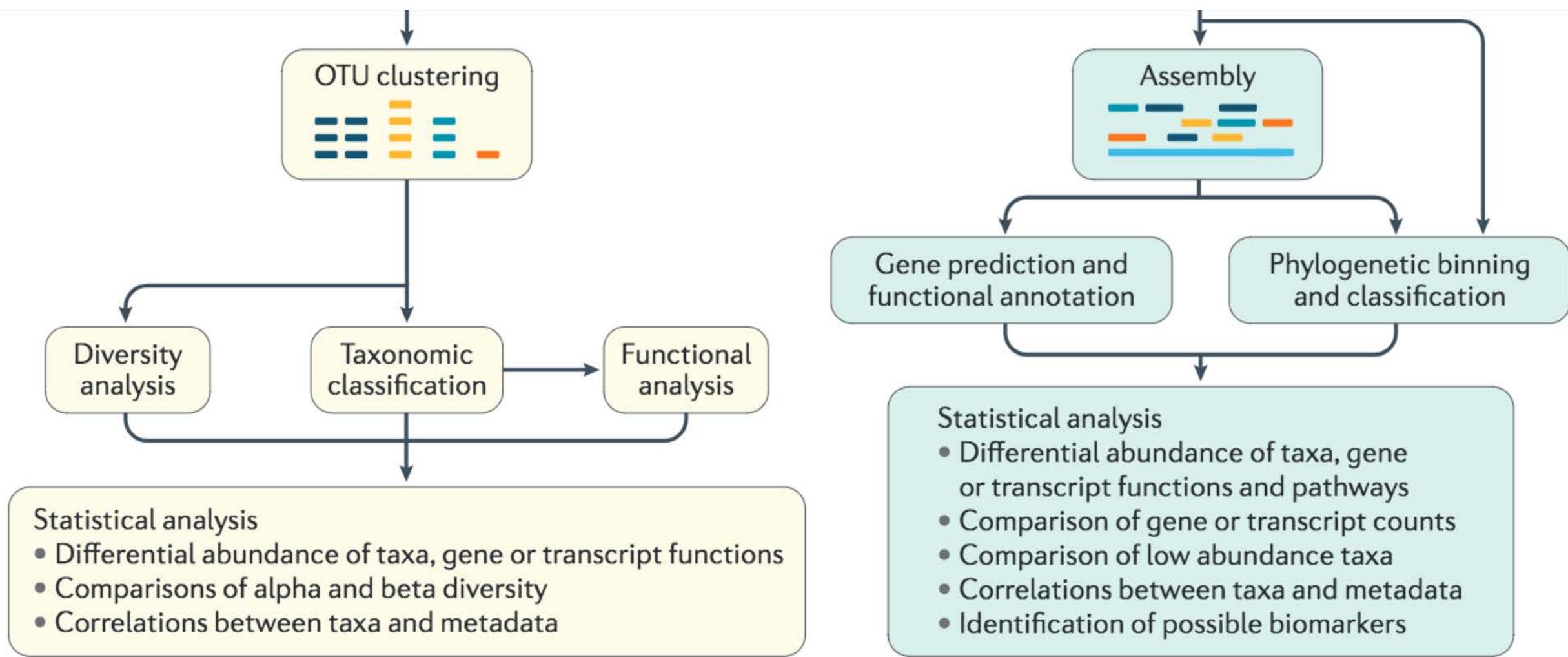
**What** are they doing?

**How** are they doing?



### Study design, sample collection, storage and DNA/cDNA/RNA extraction





**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

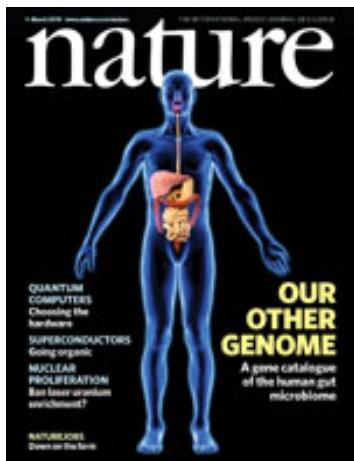
# What have metagenomics been used for?

## Exploration and categorisation



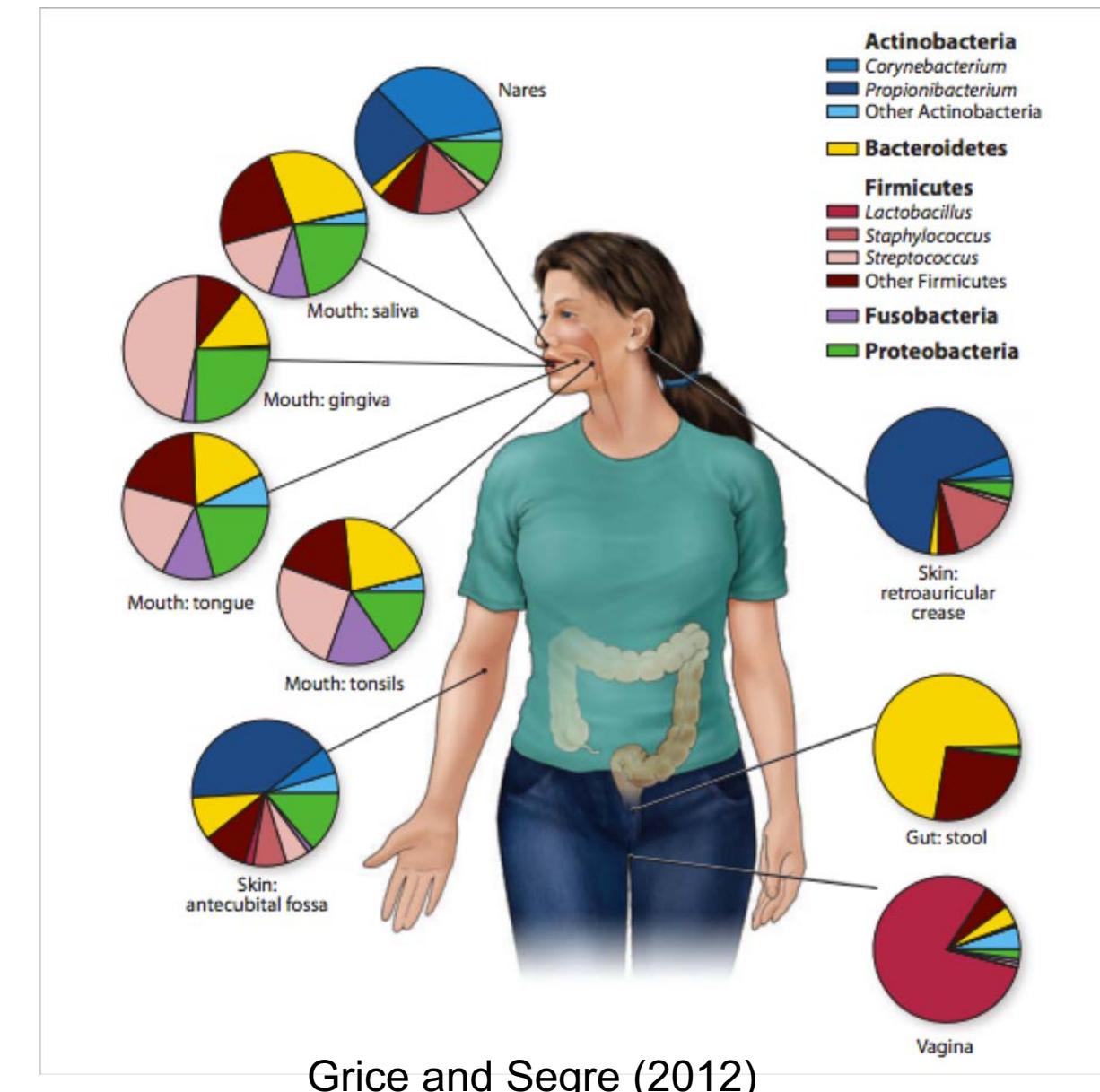
Rusch et al., 2007 Plos Biology

- 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

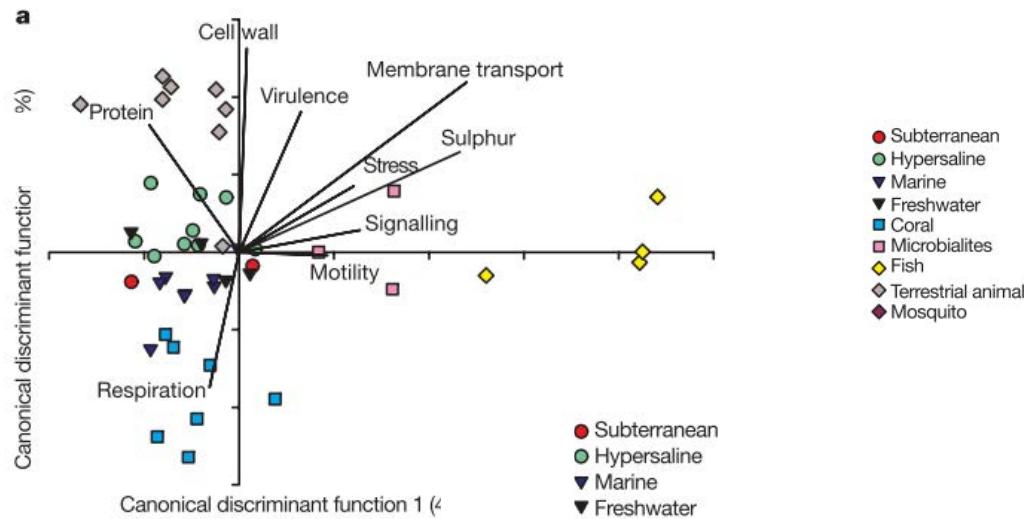
- 127 Human gut metagenomes
- 600 Gbp sequence (200 x Human genomes)
- 3.3 million genes identified
- Minimal gut metagenome defined



Grice and Segre (2012)

# What have metagenomics been used for?

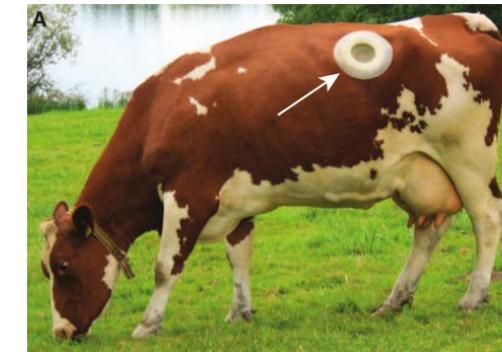
## Comparative



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

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

## Specific functions

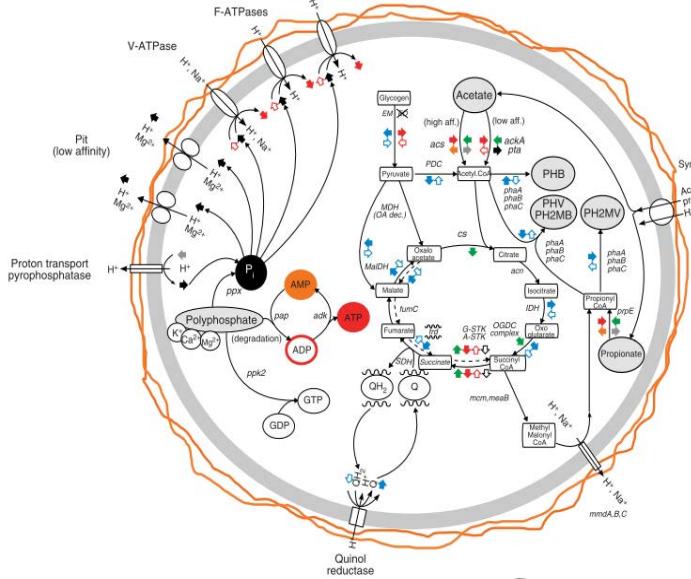


Hess *et al.*, 2011 **Science**

- Identified 27,755 putative carbohydrate-active genes from a cow rumen metagenome
- Expressed 90 candidates of which 57% had enzymatic activity against cellulosic substrates

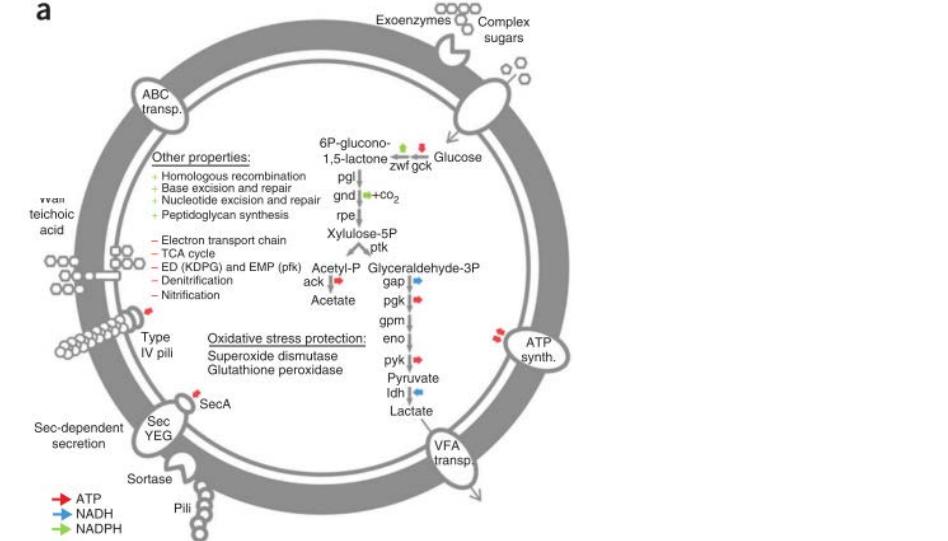
# What have metagenomics been used for?

## Extracting genomes



Garcia Martin et al., 2006 **Nat. Biotechnol.** Albertsen et al., 2013 **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



- 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

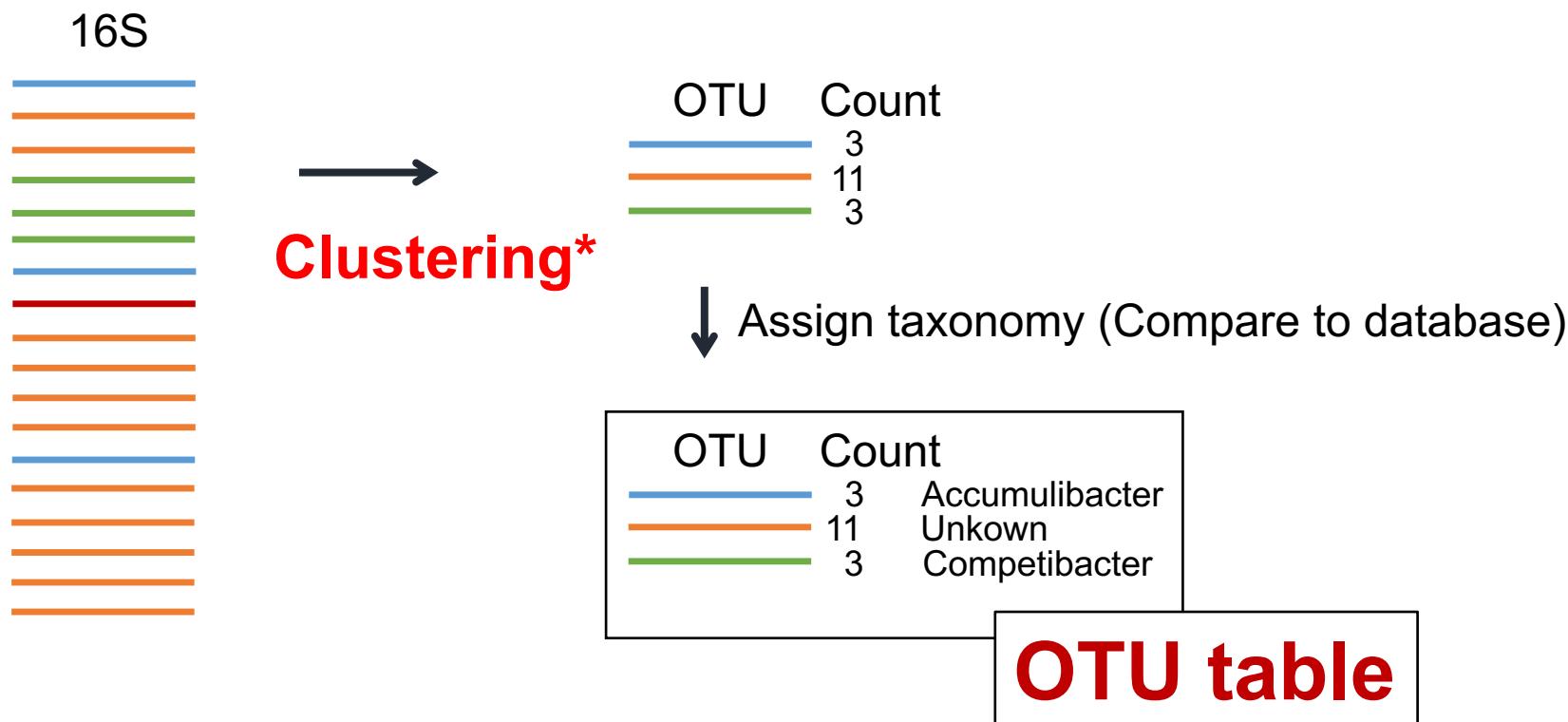
- Conceptually simple
- Mask effect of poor quality data
  - Sequencing error
  - in vitro recombination

## Weaknesses

- Limited resolution
- Logically inconsistent definition

# 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



# OTU “picking”

The process of bin sequences into clusters of OTUs.

## De Novo

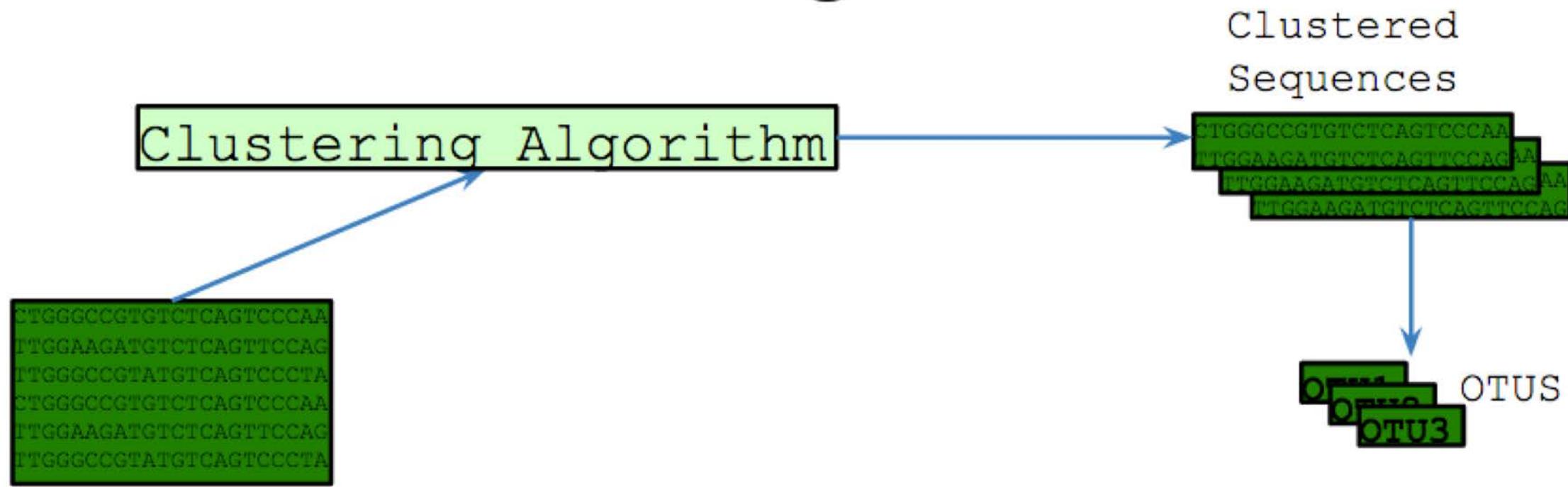
Reads are clustered based on similarity to one another.

## Reference-based

Closed reference: any reads which don't hit a reference sequence are discarded

Open reference: any reads which don't hit a reference sequence are clustered de novo

# De novo OTU picking



Experimental  
Sequences

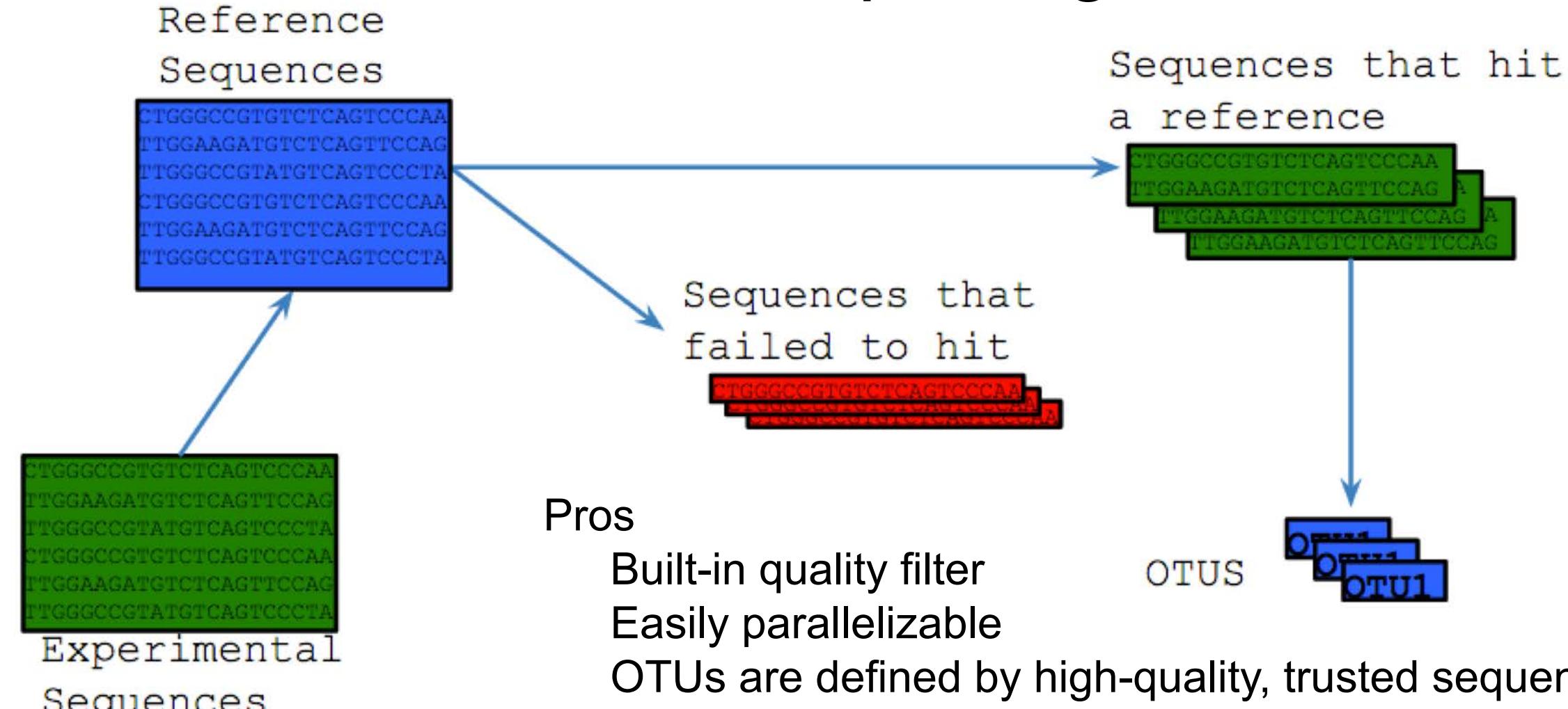
Pros  
All reads are clustered

Cons  
Not parallelizable  
OTUs may be defined by erroneous  
reads

# *De novo* OTU picking

- You **must** use if:
  - You do not have a reference sequence collection to cluster against, for example because you're working with an infrequently used marker gene.
- You **cannot** use if:
  - You are comparing non-overlapping amplicons, such as the V2 and the V4 regions of the 16S rRNA.
  - You are working with very large data sets, like a full HiSeq 2000 run.  
(Technically you can, but it will be *really* slow.)

# Closed-reference OTU picking



## Pros

Built-in quality filter

Easily parallelizable

OTUs are defined by high-quality, trusted sequences

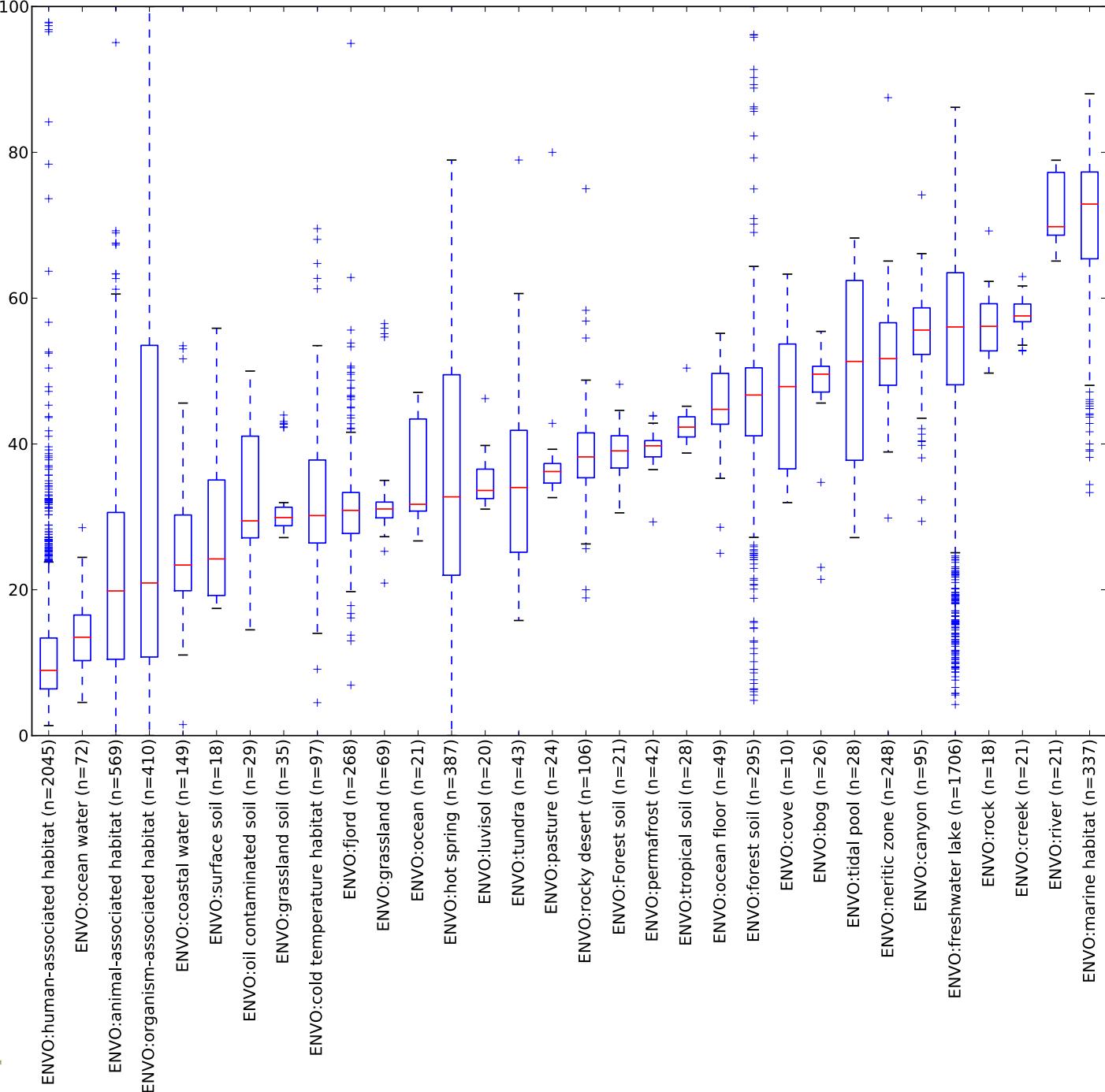
## Cons

Reads that don't hit reference dataset are excluded, so you can never observe new OTUs

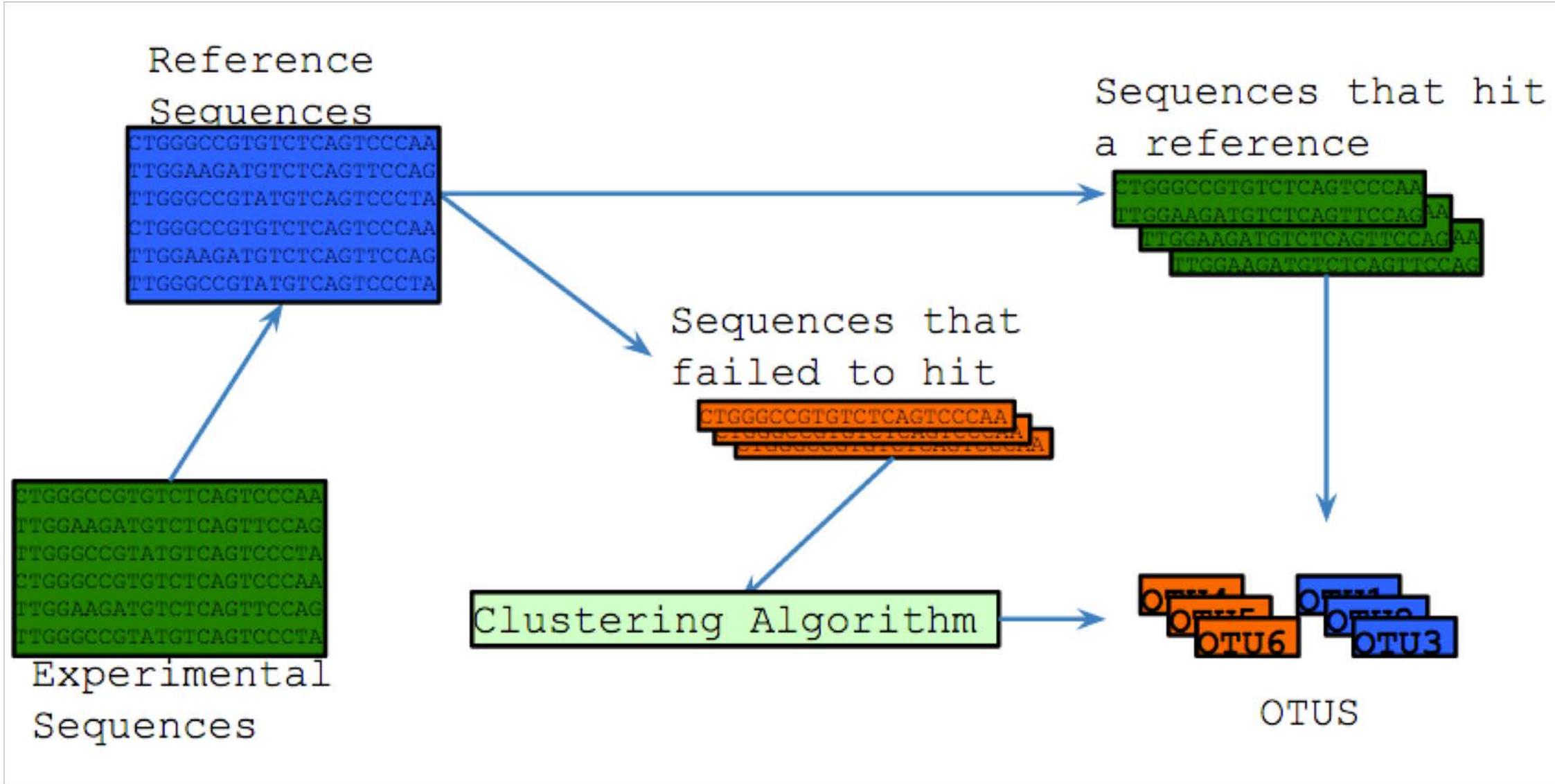
# Closed-reference OTU picking

- You **must** use if:
  - You are comparing non-overlapping amplicons, such as the V2 and the V4 regions of the 16S rRNA. Your reference sequences must span both of the regions being sequenced.
- You **cannot** use if:
  - You do not have a reference sequence collection to cluster against, for example because you're working with an infrequently used marker gene.

Percentage of  
reads that do not hit  
the reference  
collection, by  
environment type.



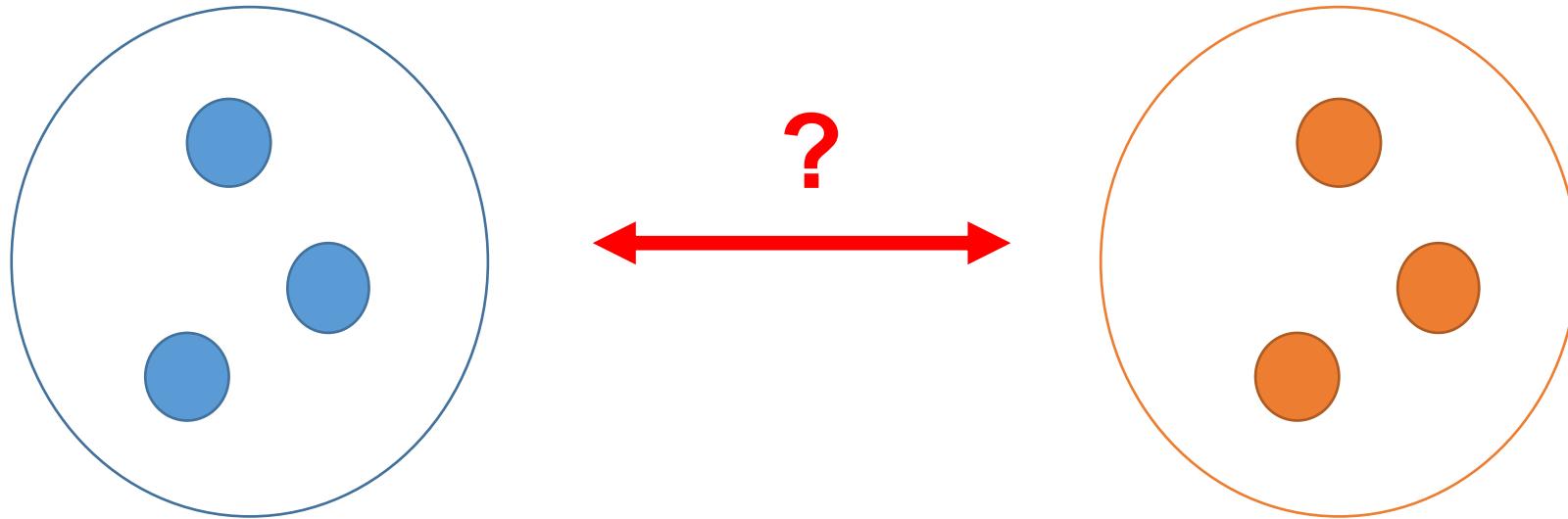
# Open-reference OTU picking



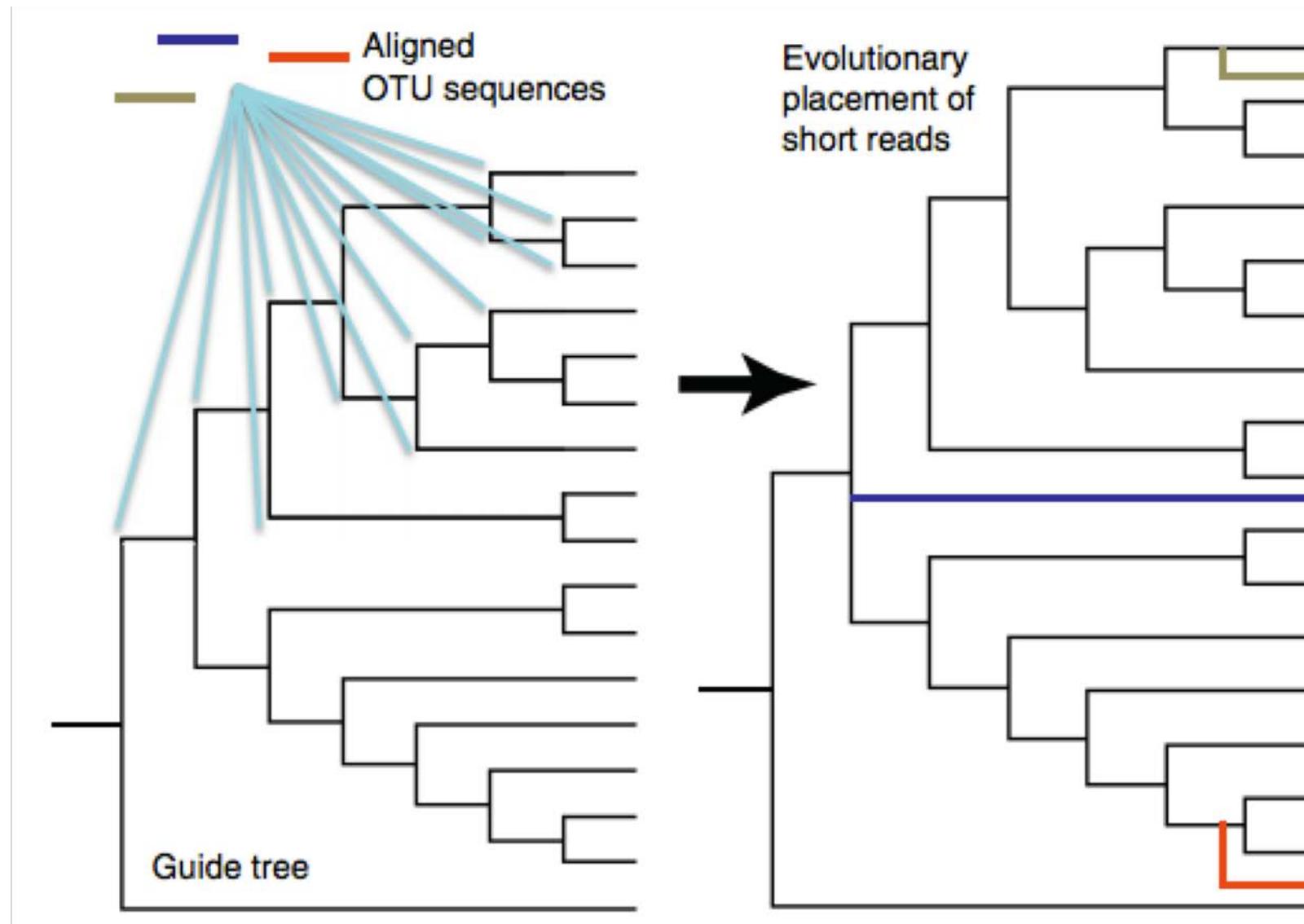
# Open-reference OTU picking

- Pros
  - All reads are clustered
  - Partially parallelizable
- Cons
  - Only *partially* parallelizable
  - Mix of high quality sequences defining OTUs (i.e., the database sequences) and possible low quality sequences defining OTUs (i.e., the sequencing reads)

# Assigned OTUs -> Loss of information



# OTU relationship using phylogenetics

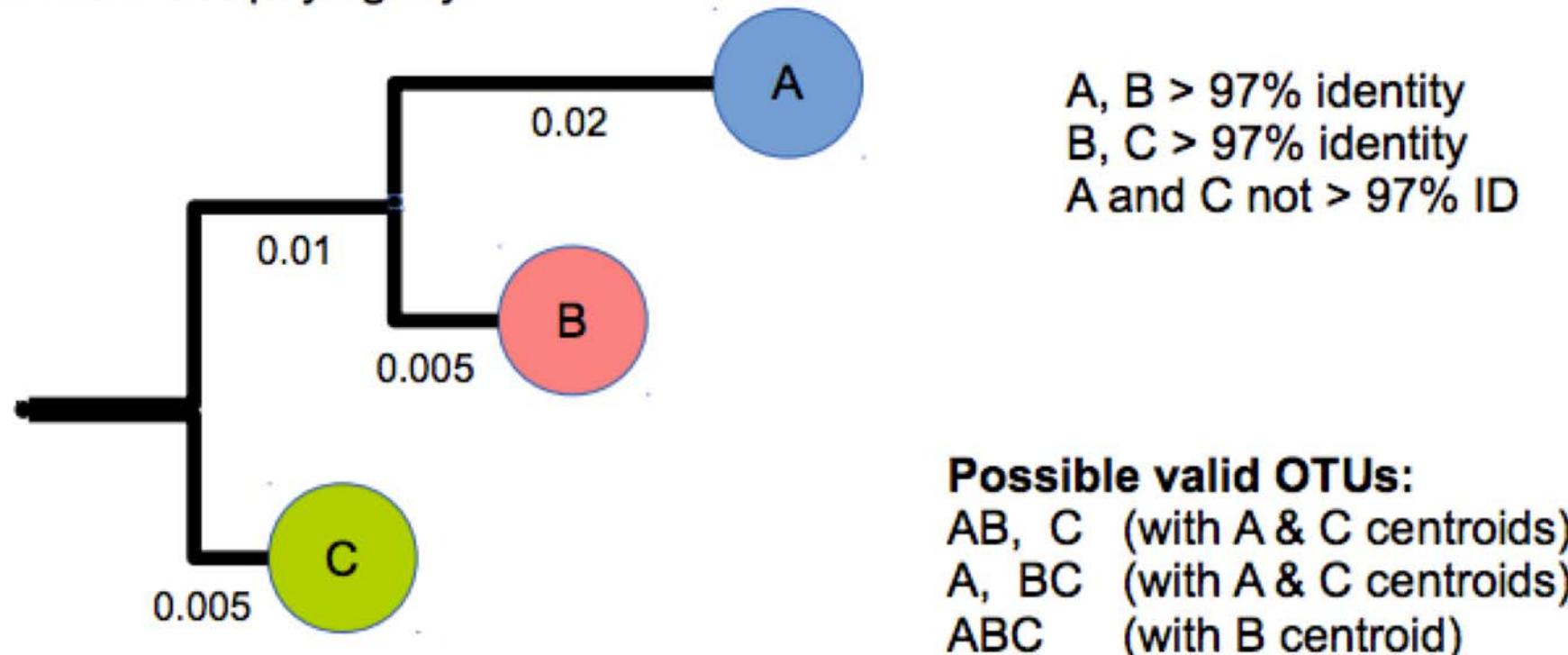


TRENDS in Ecology & Evolution

Bik et al (2011)

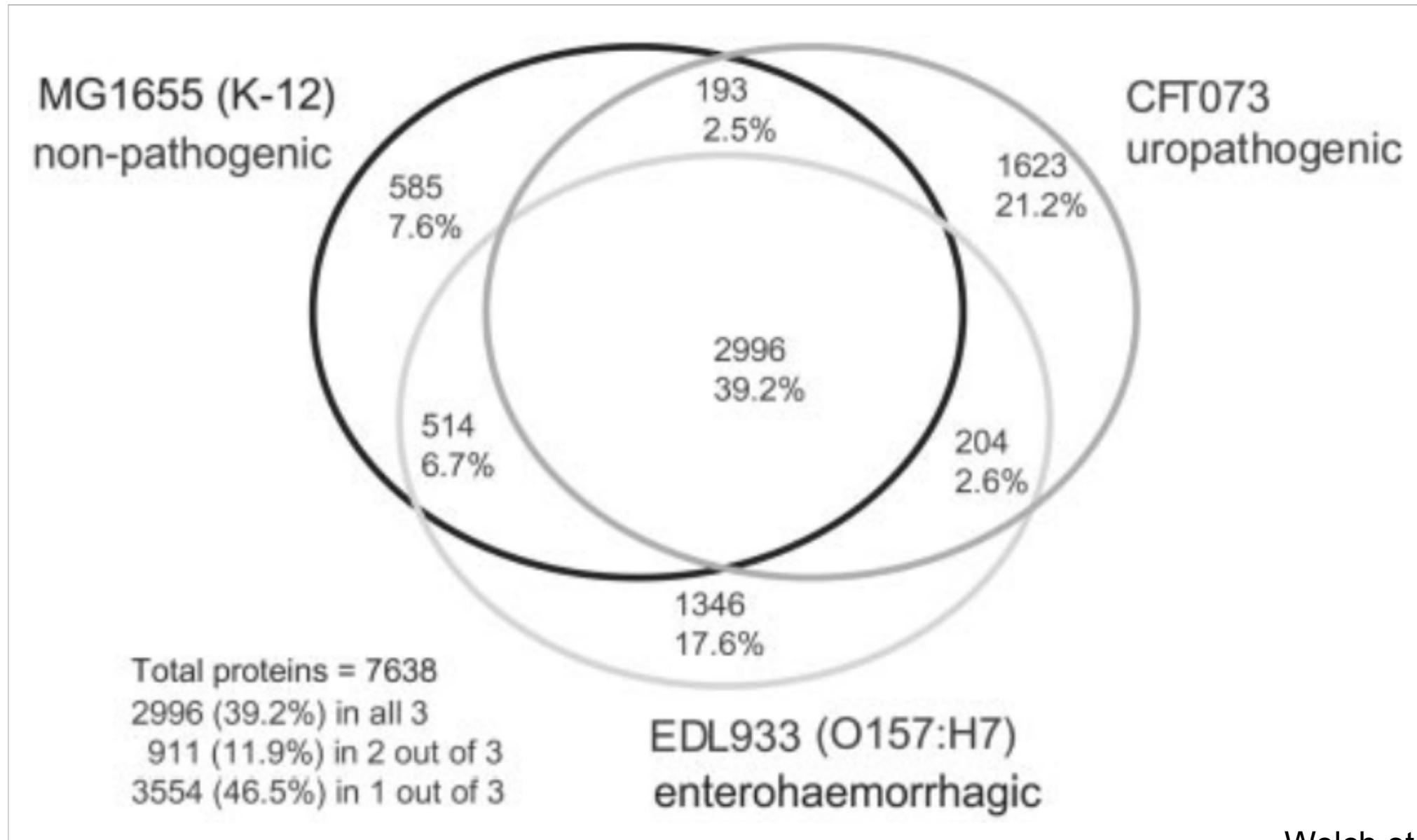
# Logical inconsistency: OTUs at 97% ID

Assume the true phylogeny:

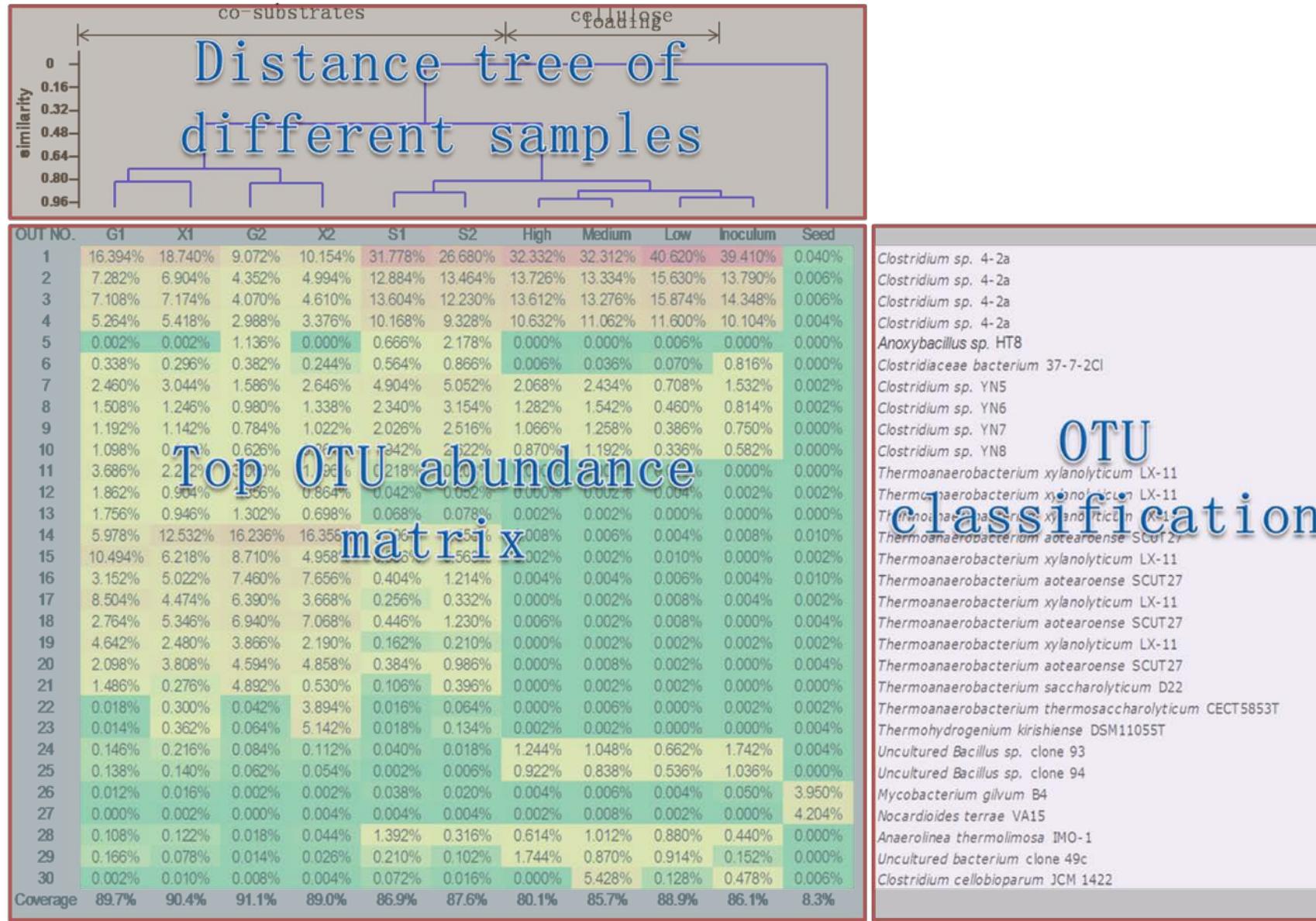


OTU pipelines will arbitrarily pick one of the three solutions.  
Is this actually a problem??

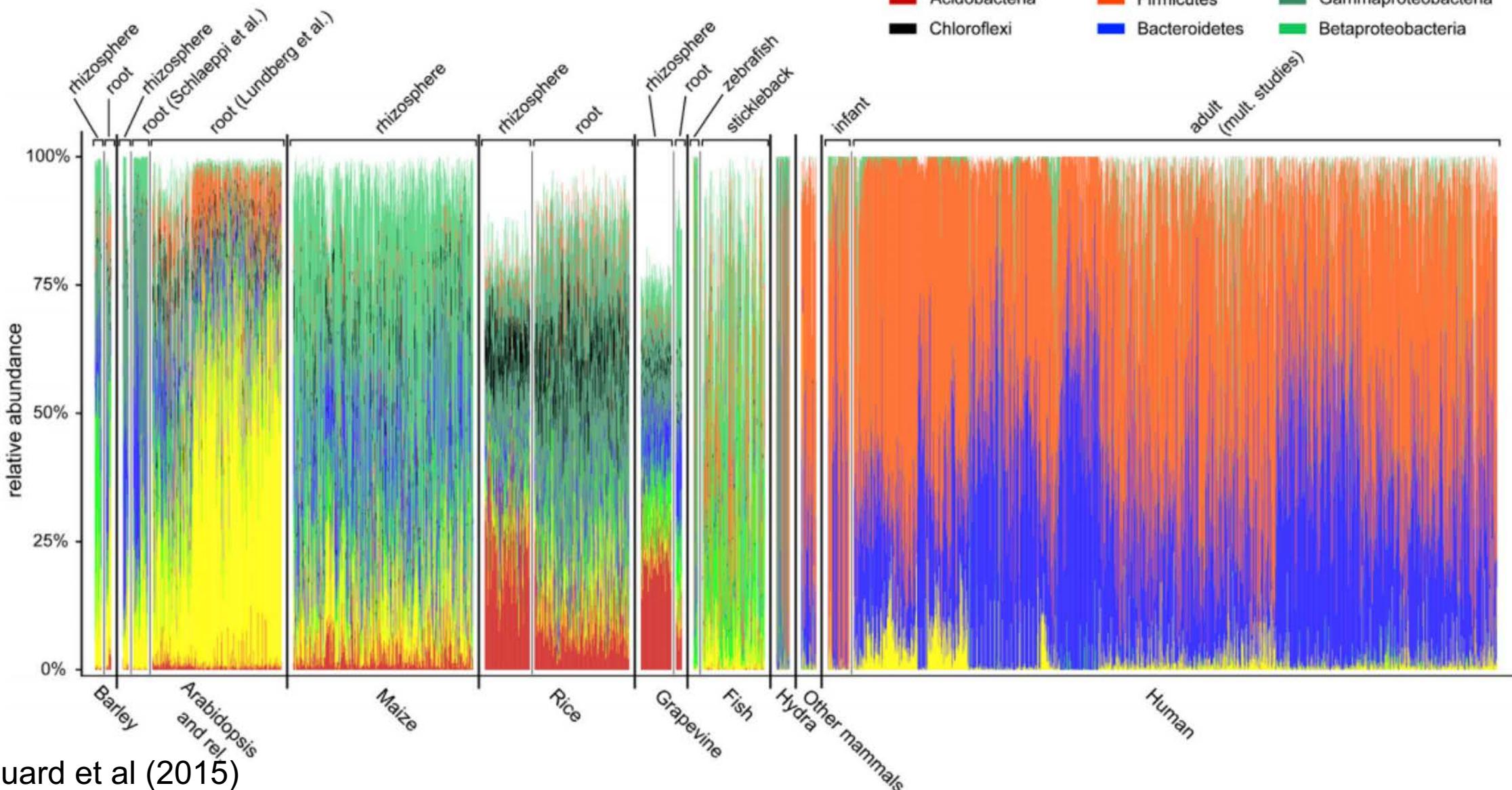
# Same species (16S): Different genomes



# Tree way plot with top OTUs abundance and classification

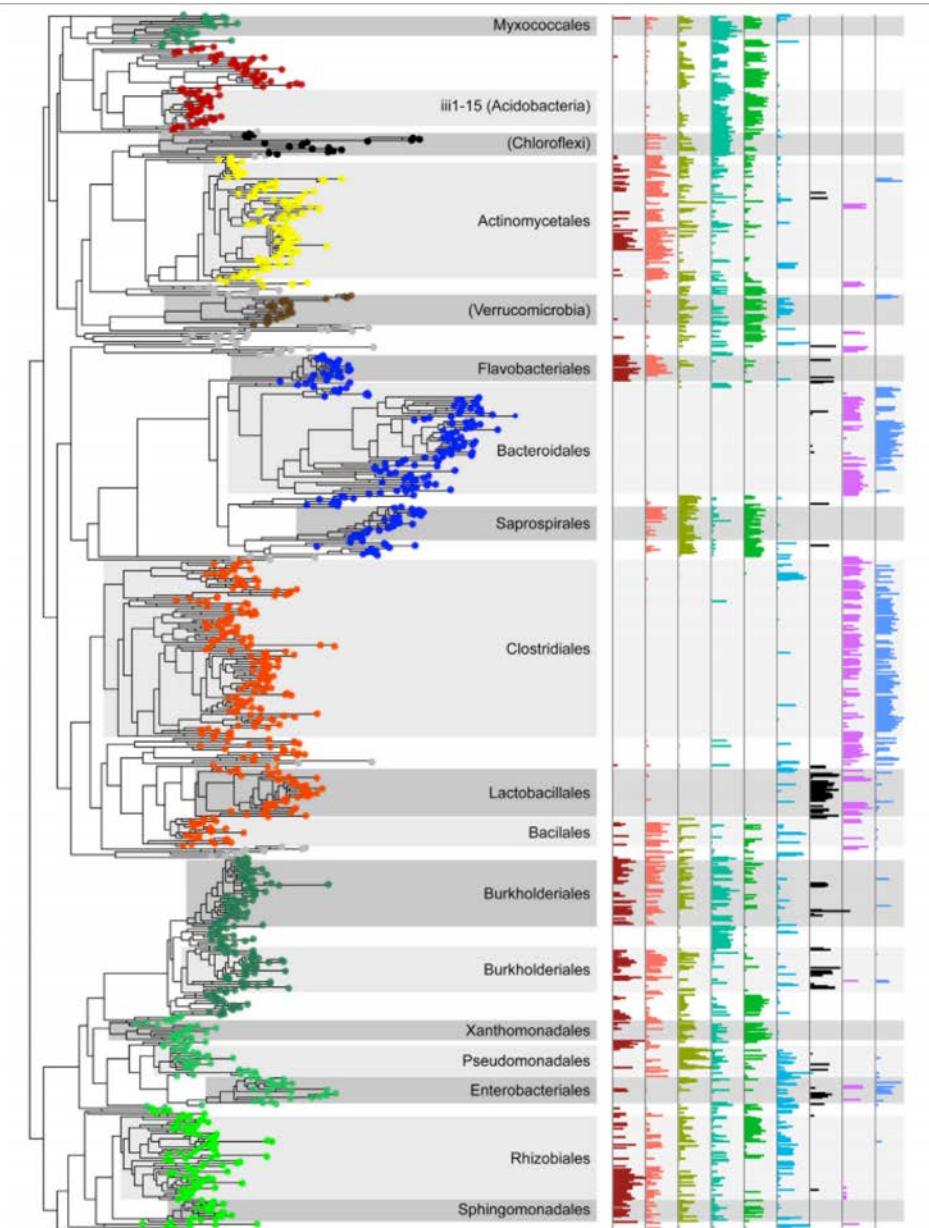


# Cumulative Abundance plots



# Phylogenetic Analysis of OTU abundances

Relationship between OTUs



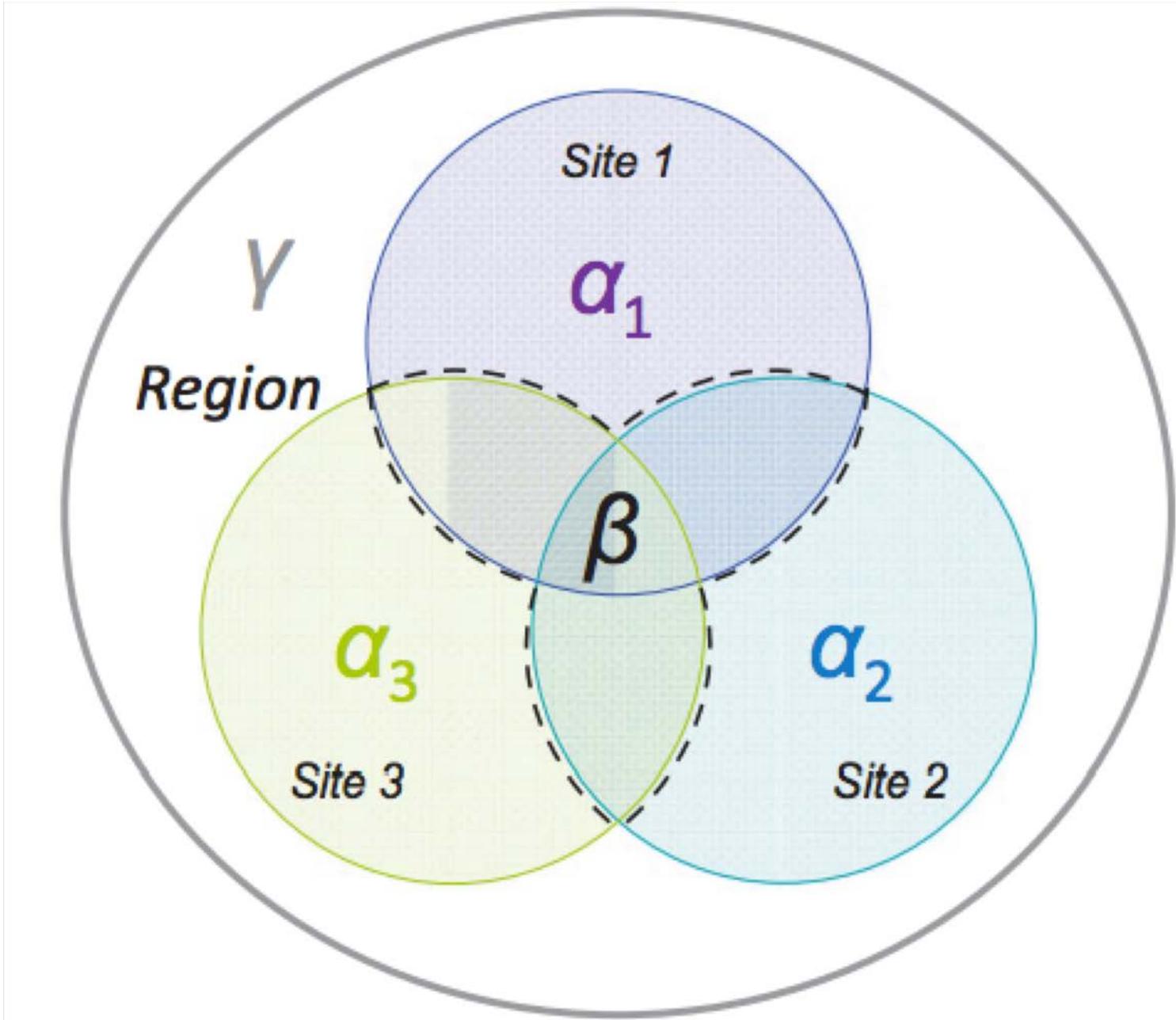
How do we compare  
between different  
samples?

# Concept: Diversity 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).

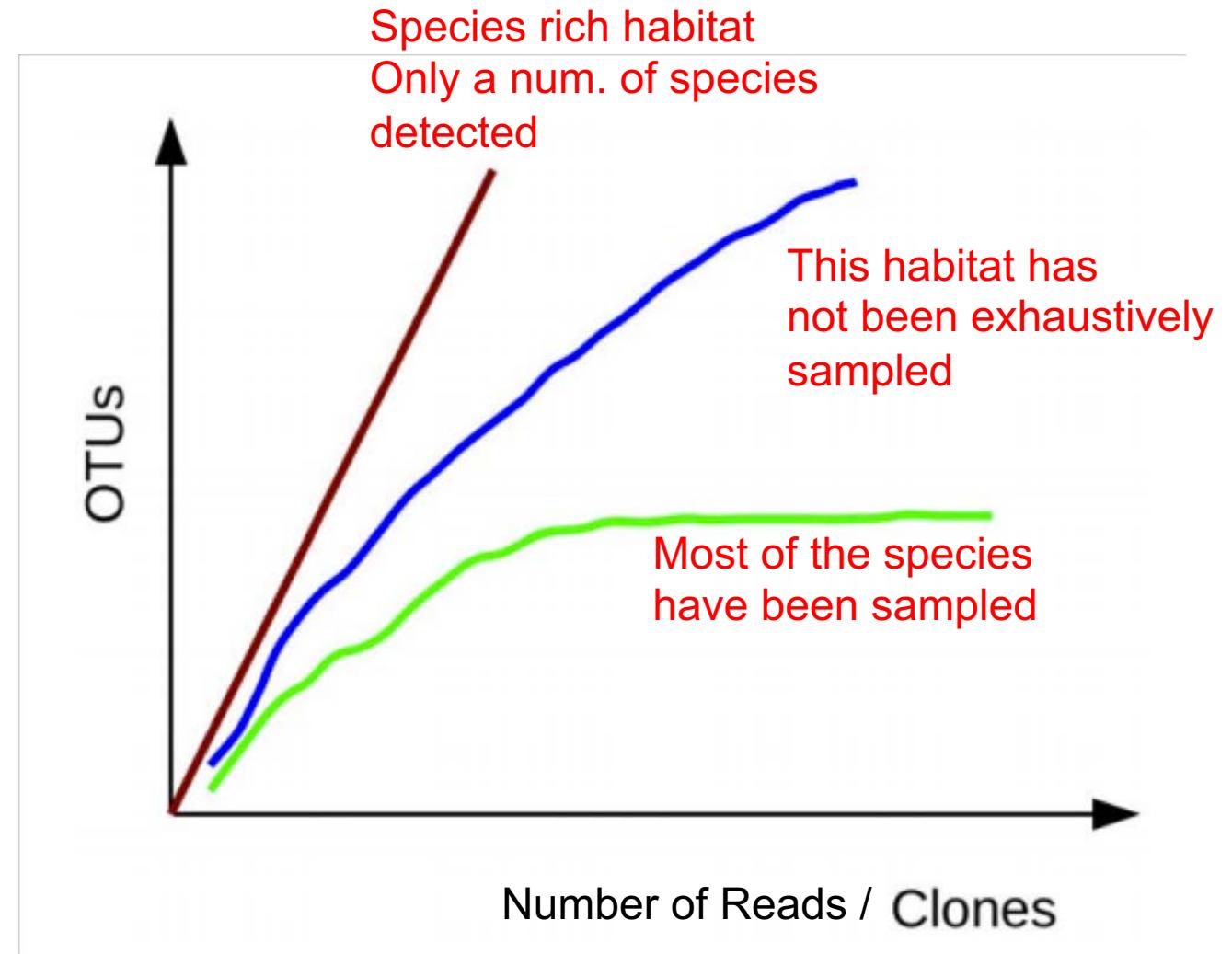
**Beta diversity** consists in determining the difference in diversity or community composition between two or more locations or samples (i) by considering species composition only, and use incidence data with associated metrics such as Jaccard or Sorensen similarity indices or (ii) by taking species 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



# 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

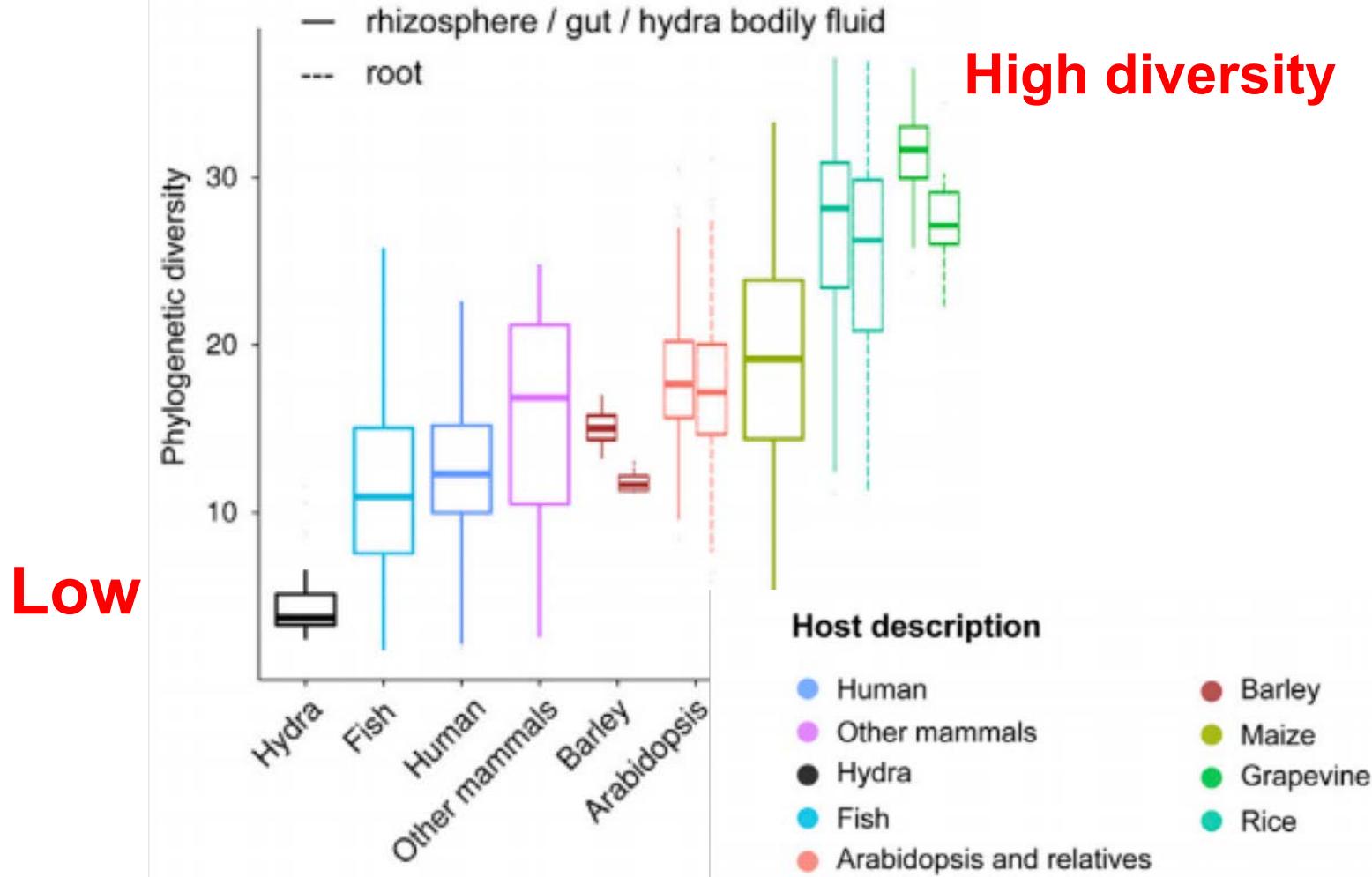
Phylogenetic diversity of genes = genetic PD

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

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

**B**

## Alpha-diversity (phylogenetic diversity)



# 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 occurrences across communities  
1 = present, 0 = below detection

List of observed OTUs	Occurrences in community A	Occurrences in community B	Shared occurrences A & B		
	OTU 1	OTU 2	OTU 3	OTU 4	OTU 5
	1	0			
	0	1			
	1	1	X		
	1	1	X		
	1	1	X		

## B. Composition:

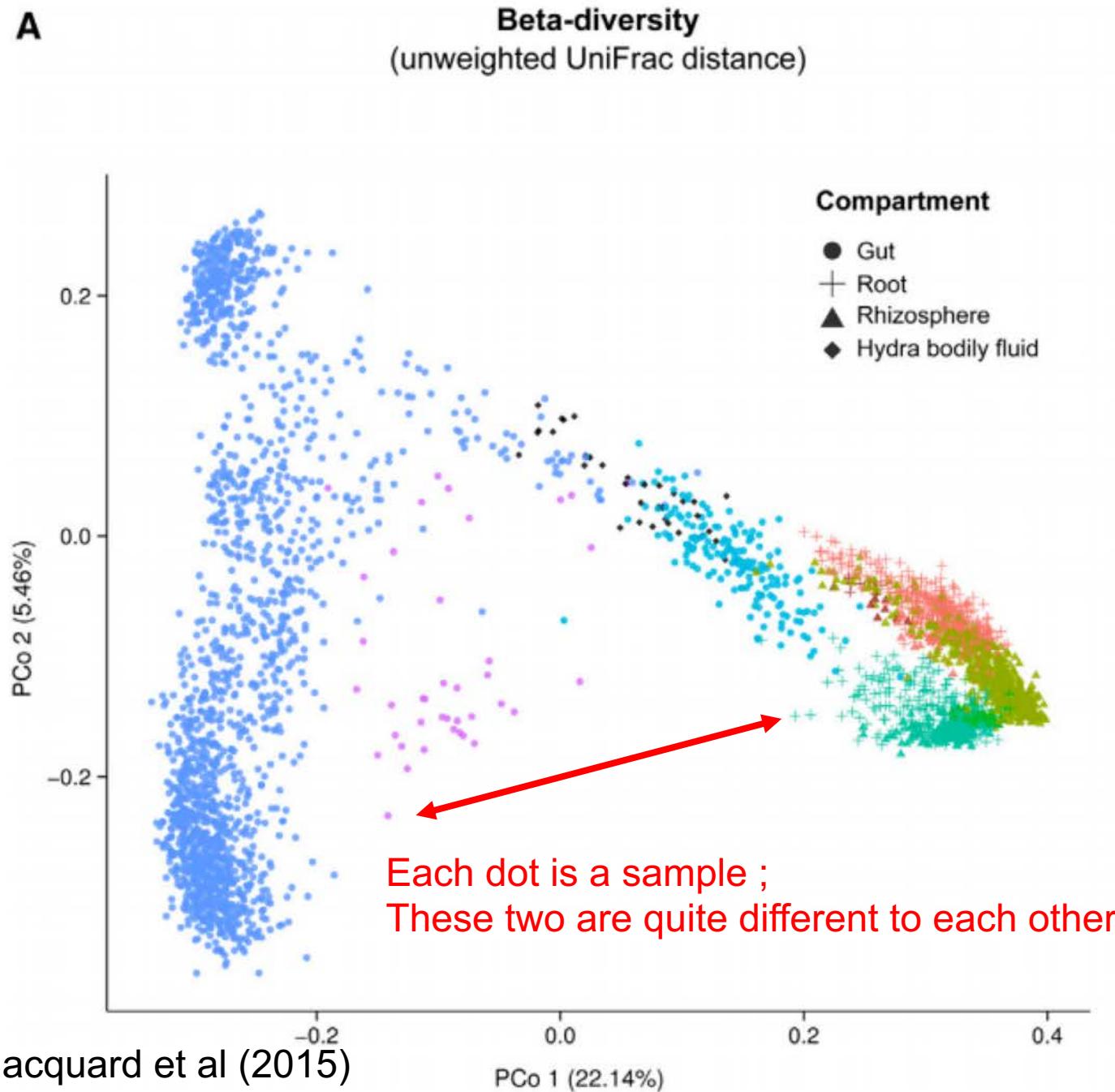
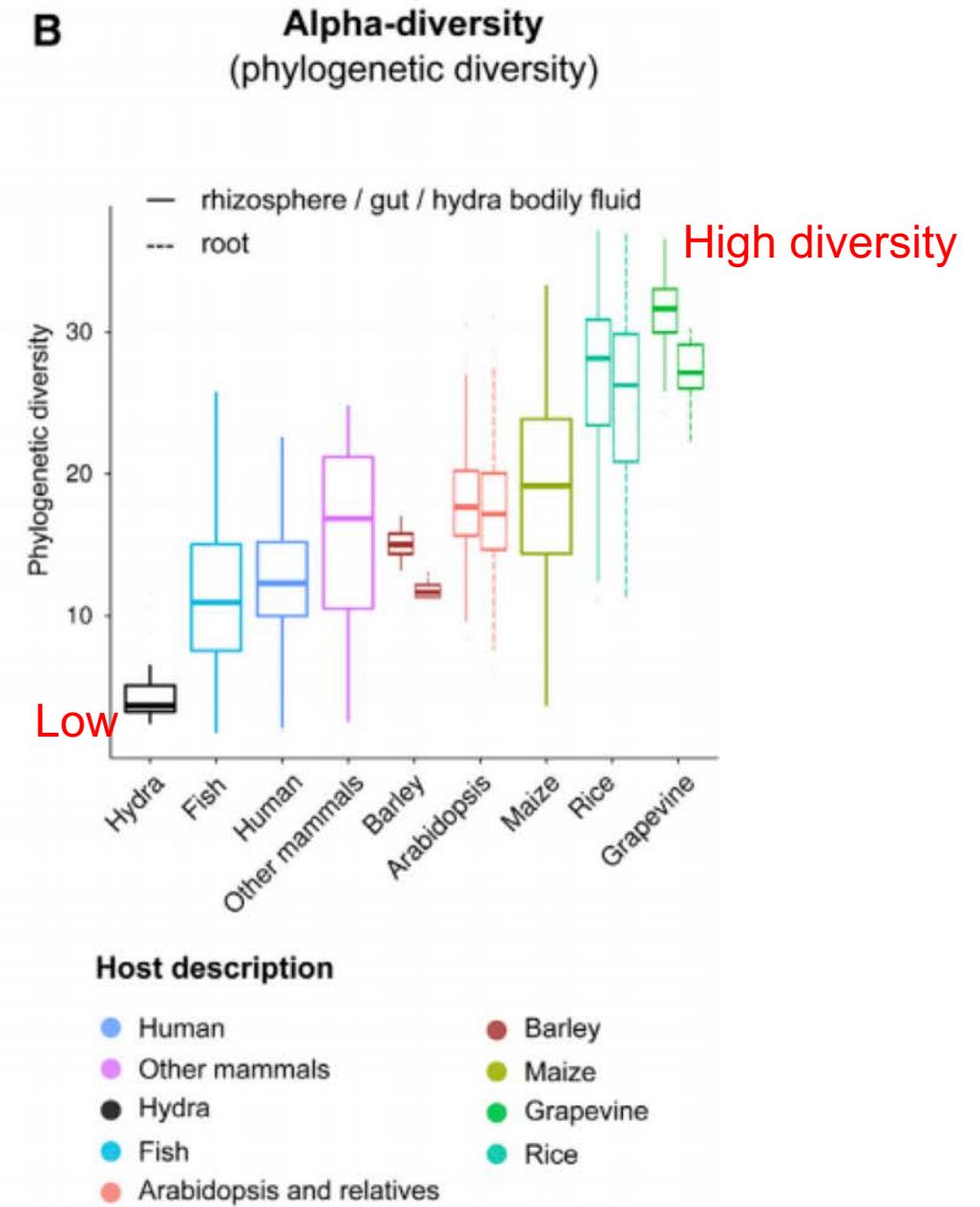
similar OTU abundances across communities

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

## Phylogeny:

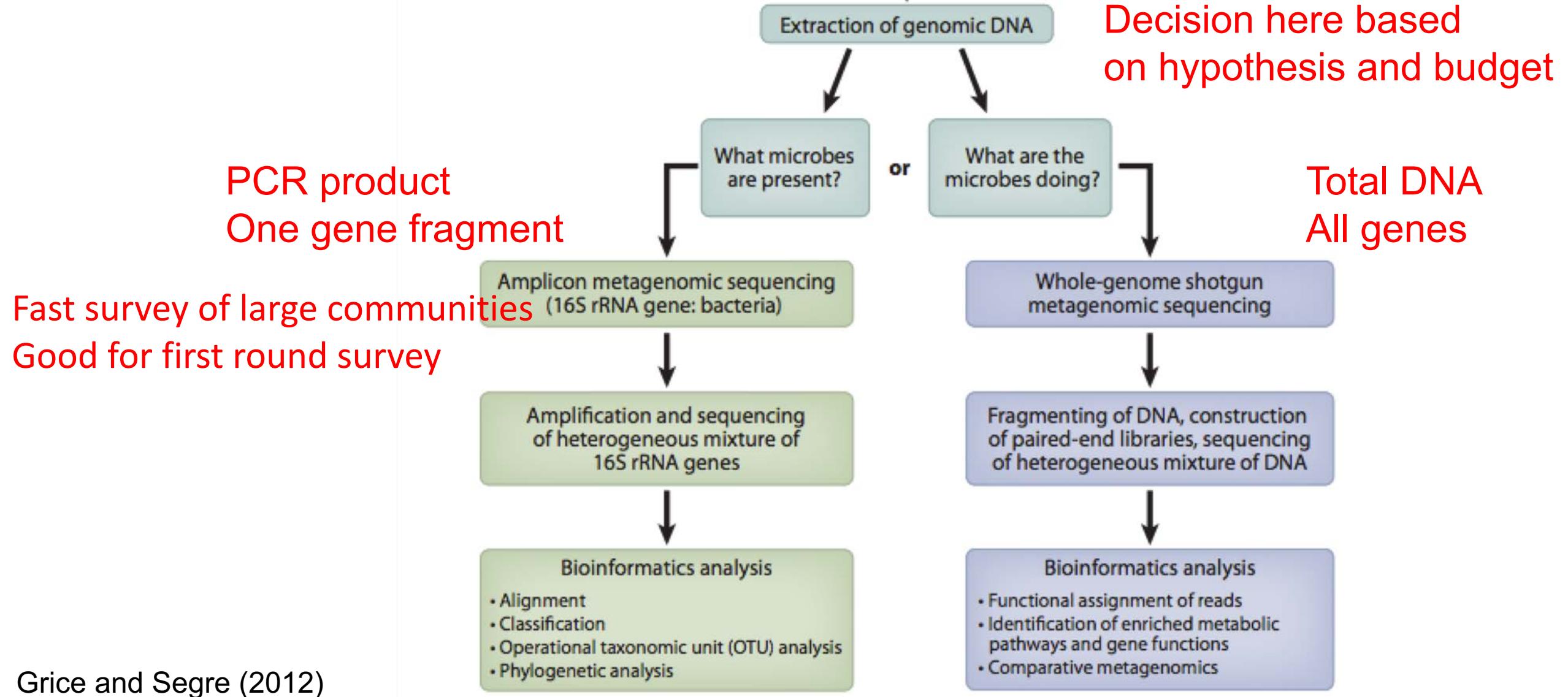
shared OTU lineages across communities

lineage	Abundances community A	Abundances community B	Similar abundances A & B		
	OTU 1	OTU 2	OTU 3	OTU 4	OTU 5
i.	0.4	0			
ii.	0	0.1			
iii.	0.1	0.1	X		
	0	0.8			
	0.5	0			

**A****B**

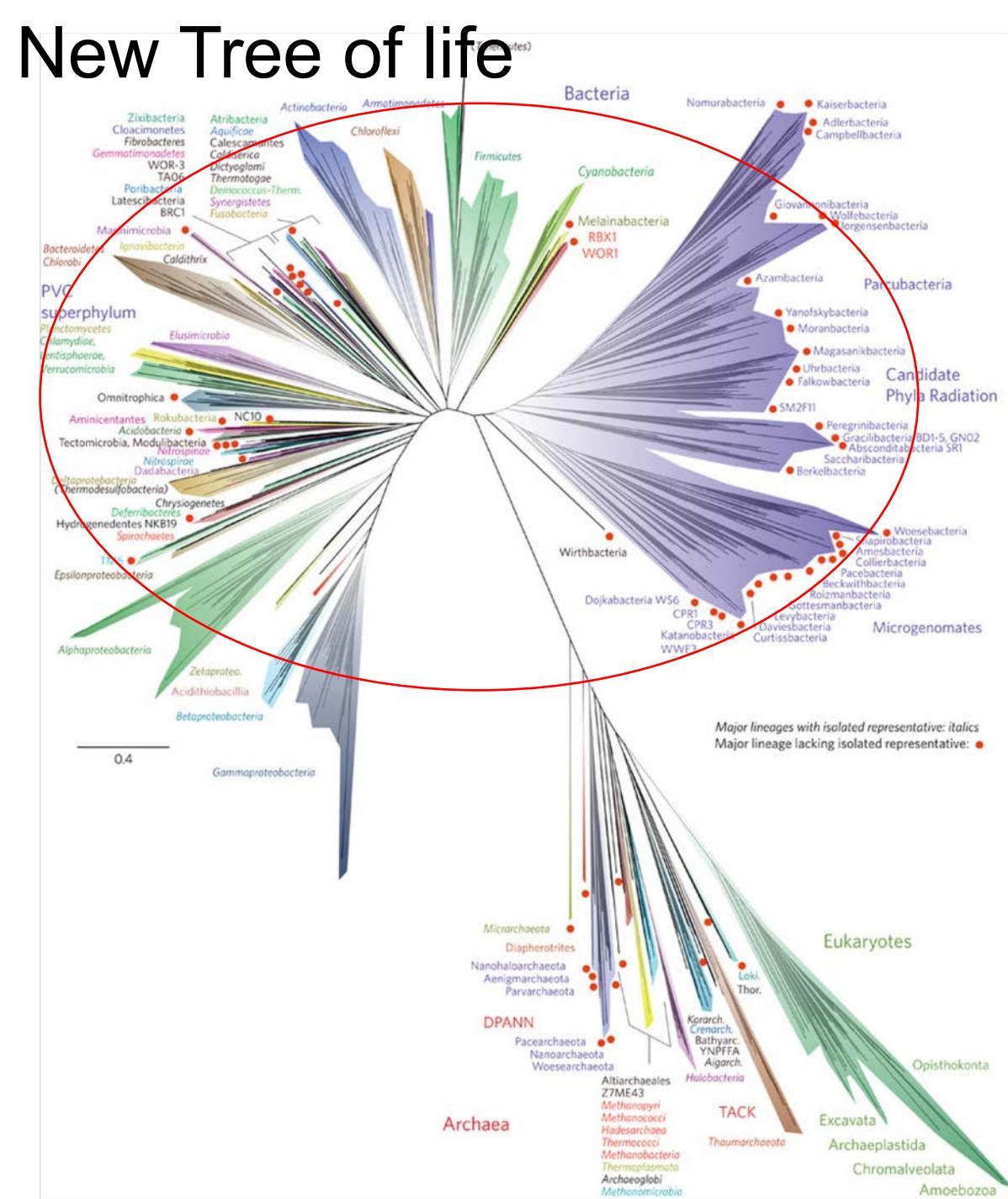
Amplicon sequencing or metagenomes?

# Workflow decision



# Amplicon sequencing

# New Tree of life

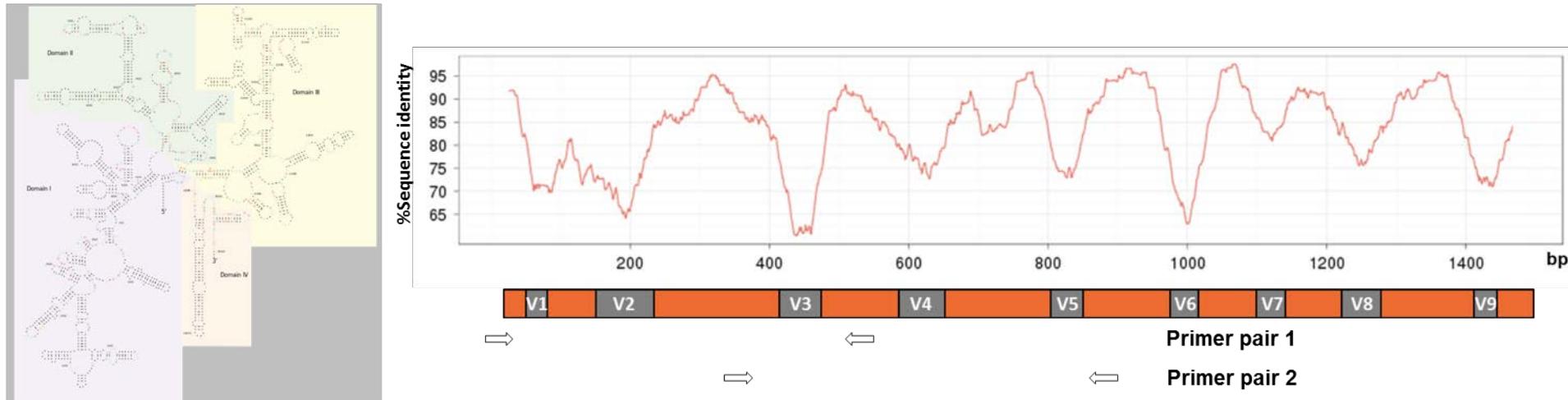


What do they have in common?

Hug et al (2016)

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

# 16S



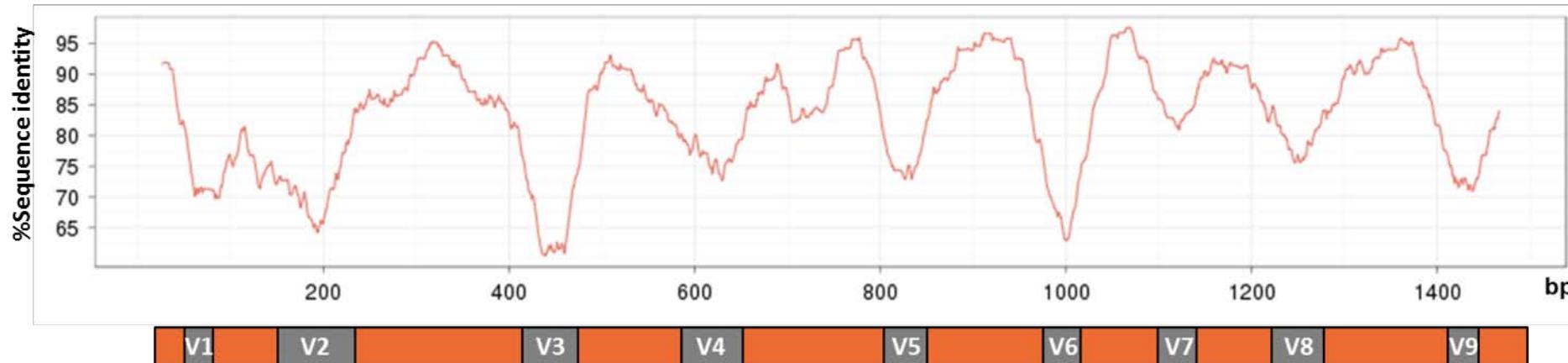
- **Advantages:**

- Universal: Every bacterial and archaea 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:**

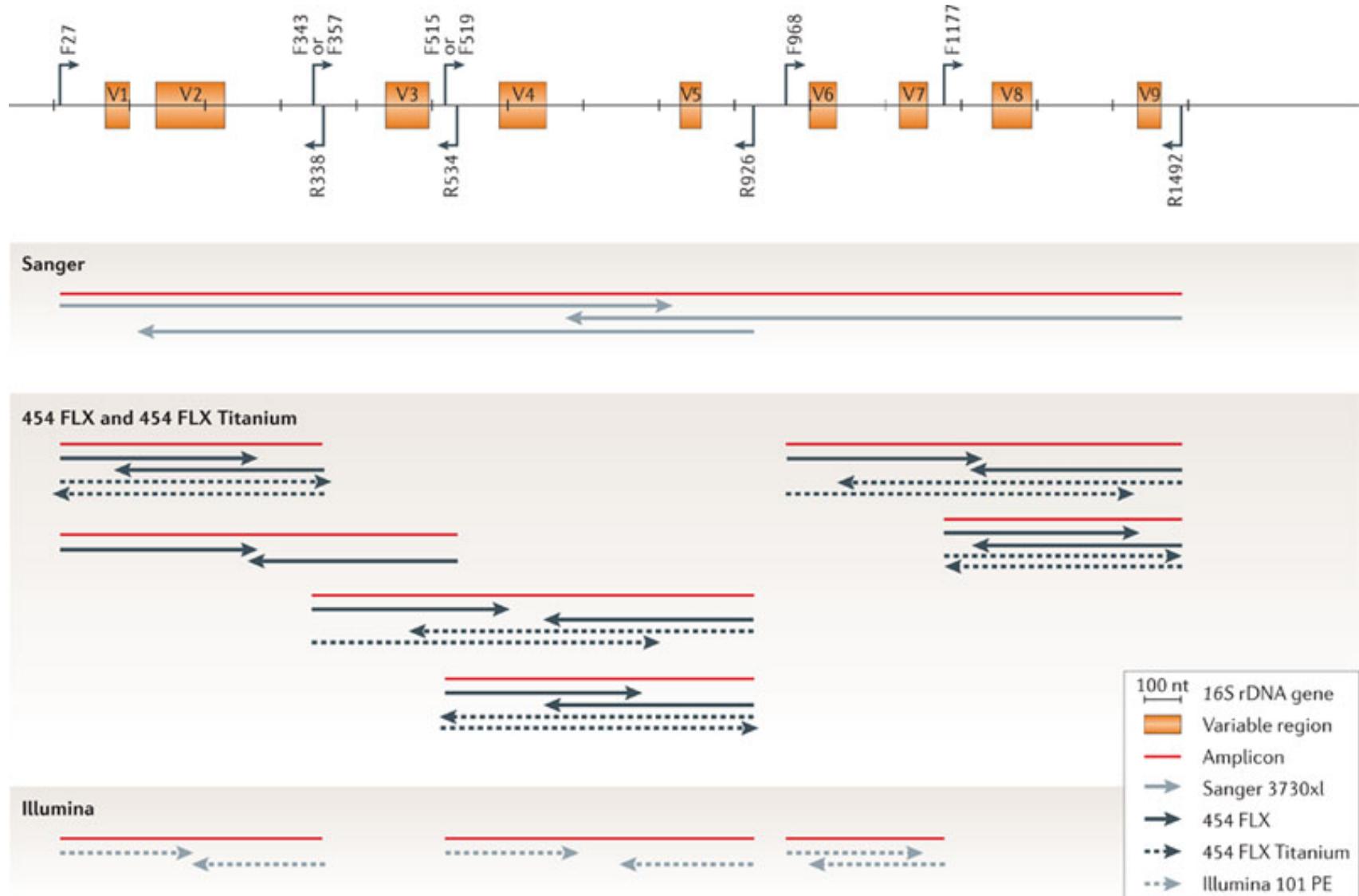
- Variable copy number in each species
- No universal (unbiased) primers
- (Not directly correlated with activity)
- (Lack of functional information)

# Typical workflow

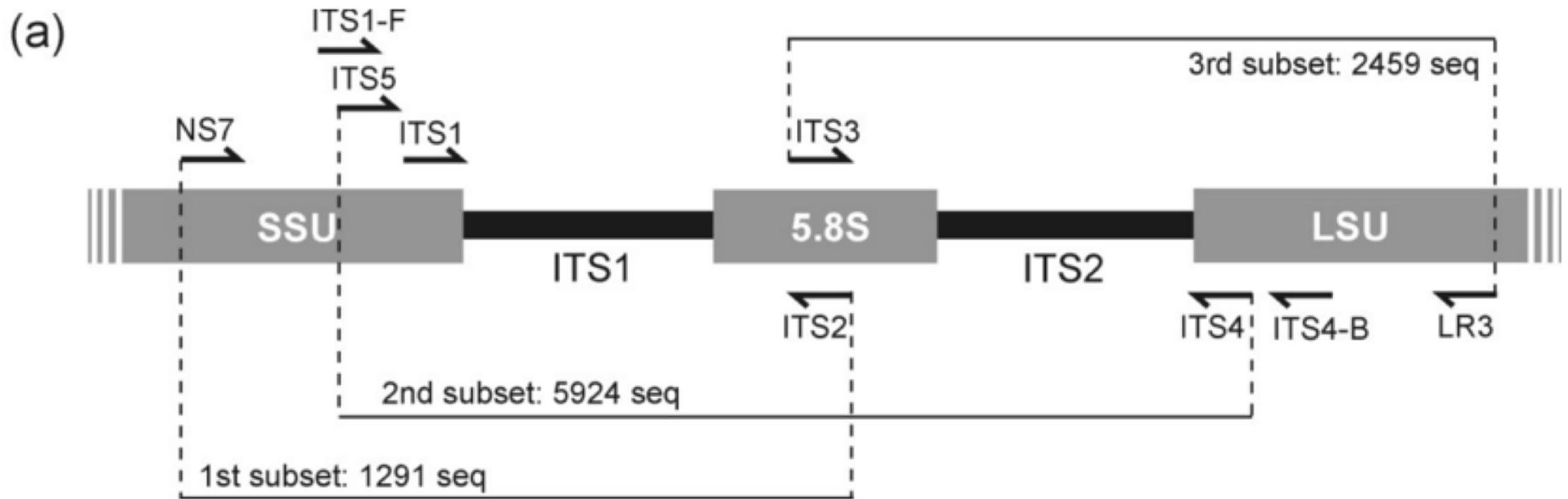


Which region to sequence?

# 16S amplified region



# ITS for characterization of fungi species



# Workflow

---

Filter for  
contaminants and  
low quality reads



Assemble  
overlapping reads



Reduce datasets  
(clustering)



Perform taxonomic  
classification and  
compute diversity  
metrics

- Quality plots and read trimming
  - FastQC  
<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
  - FASTX  
[http://hannonlab.cshl.edu/fastx\\_toolkit/](http://hannonlab.cshl.edu/fastx_toolkit/)
- Chimera removal
  - AmpliconNoise  
<http://code.google.com/p/ampliconnoise/>
  - UCHIME  
<http://www.drive5.com/uchime/>

# Workflow

Filter for  
contaminants and  
low quality reads



**Assemble  
overlapping reads**



Reduce datasets  
(clustering)



Perform taxonomic  
classification and  
compute diversity  
metrics

Because the read length (2x250 or 2x300) will be  
longer than the actual DNA fragment

- Merge overlapping paired end reads

- FLASH

- <http://www.genomics.jhu.edu/software/FLASH/index.shtml>

- FastqJoin

- <http://code.google.com/p/ea-utils/wiki/FastqJoin>

- CD-HIT read-linker

- <http://weizhong-lab.ucsd.edu/cd-hit/wiki/doku.php?id=cd-hit-auxtools-manual>

# Workflow

Filter for  
contaminants and  
low quality reads



Assemble  
overlapping reads



**Reduce datasets  
(clustering)**



Perform taxonomic  
classification and  
compute diversity  
metrics

- **Clustering with high stringency**

- UCLUST/USEARCH (16S only)

<http://www.drive5.com/usearch/>

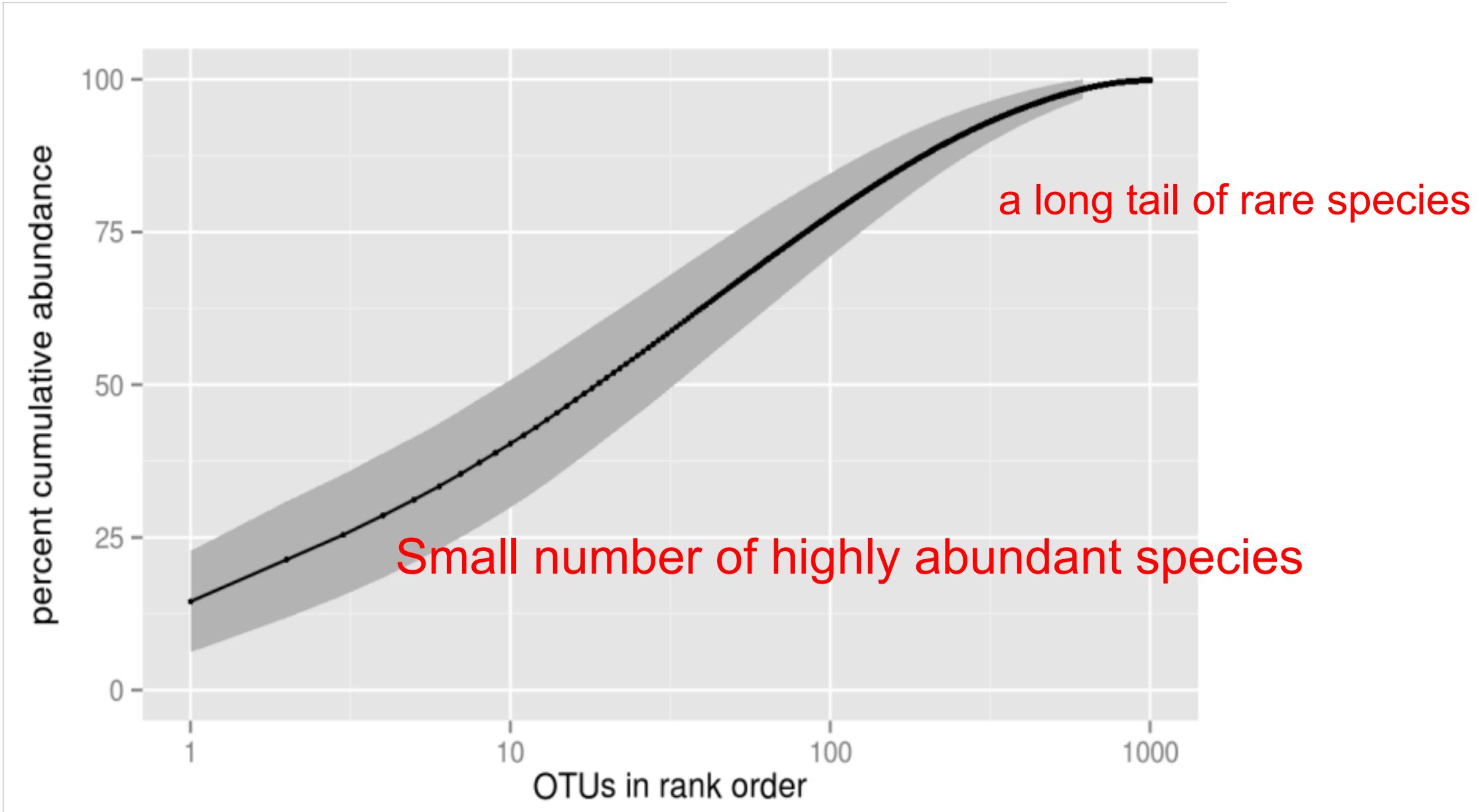
- CD-HIT-OTU (16S only)

<http://weizhong-lab.ucsd.edu/cd-hit-otu/>

- phylOTU (16S only)

<https://github.com/sharpton/PhyLOTU>

# Typical number of OTUs



# Workflow

Filter for  
contaminants and  
low quality reads



Assemble  
overlapping reads



Reduce datasets  
(clustering)



**Perform  
taxonomic  
classification and  
compute diversity  
metrics**

- **Composition based classifiers**
  - RDP database + classifier  
<http://rdp.cme.msu.edu/classifier/classifier.jsp>
- **Homology based classifiers**
  - ARB + Silva database (16S only)  
<http://www.arb-home.de/>
  - GreenGenes database (16S only)  
<http://greengenes.lbl.gov/cgi-bin/nph-index.cgi>
  - UNITE database (ITS only)  
<http://unite.ut.ee/>
  - FungalITS Pipeline (ITS only)  
<http://www.emerencia.org/fungalitspipeline.html>

# Using a “Classifier” to annotate OTUs

Uses an existing phylogeny

Find best unambiguous match to references



<http://greengenes.lbl.gov/cgi-bin/nph-index.cgi>



<http://www.arb-silva.de/>



<https://rdp.cme.msu.edu/>

# Analysis Packages



Qiime2 (<https://qiime2.org/>)



Mothur (mothur.org)

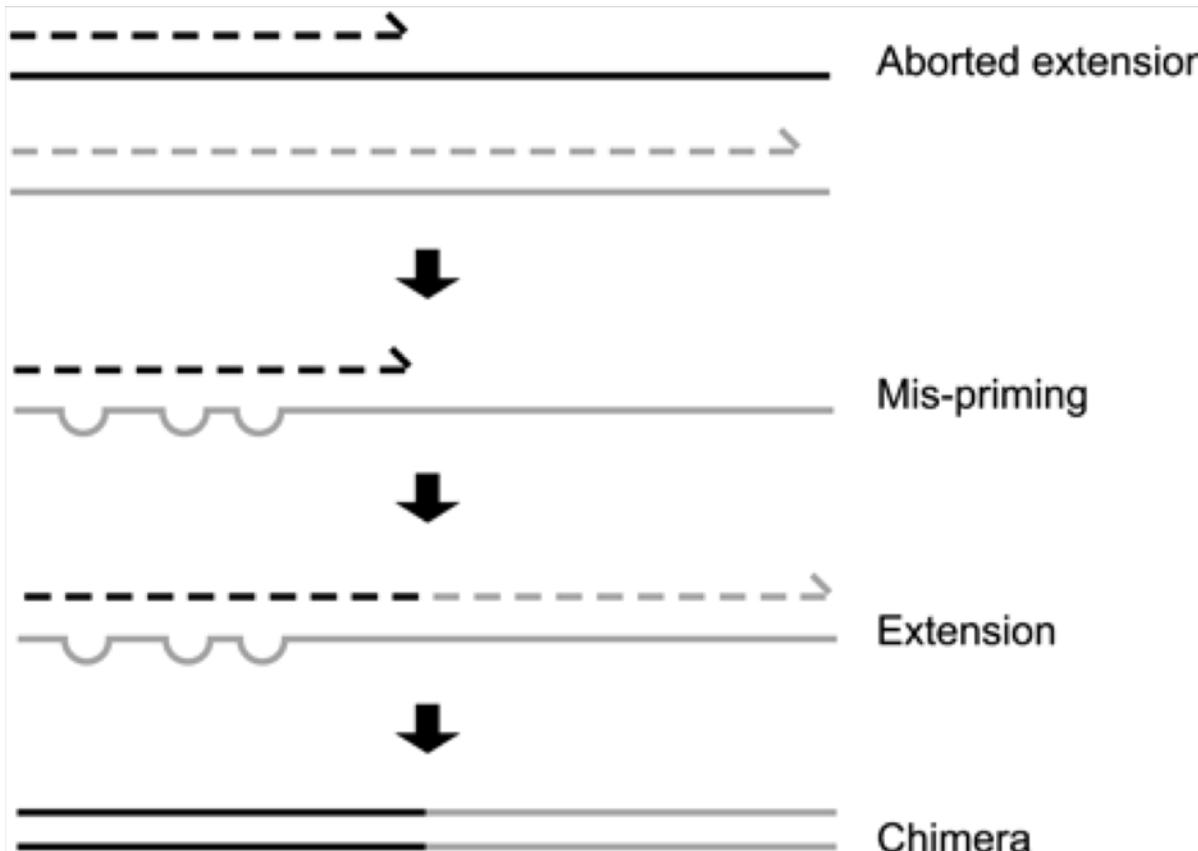
R

Phyloseq ( <https://joey711.github.io/phyloseq/> )

# 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

# 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%**”

# Metagenomics

# Advantage of metagenomics approach

**Better classification with Increasing number of complete genomes**

**Focus on whole genome based phylogeny (whole genome phlyotyping)**

- Advantages

No amplification bias like in 16S/ITS

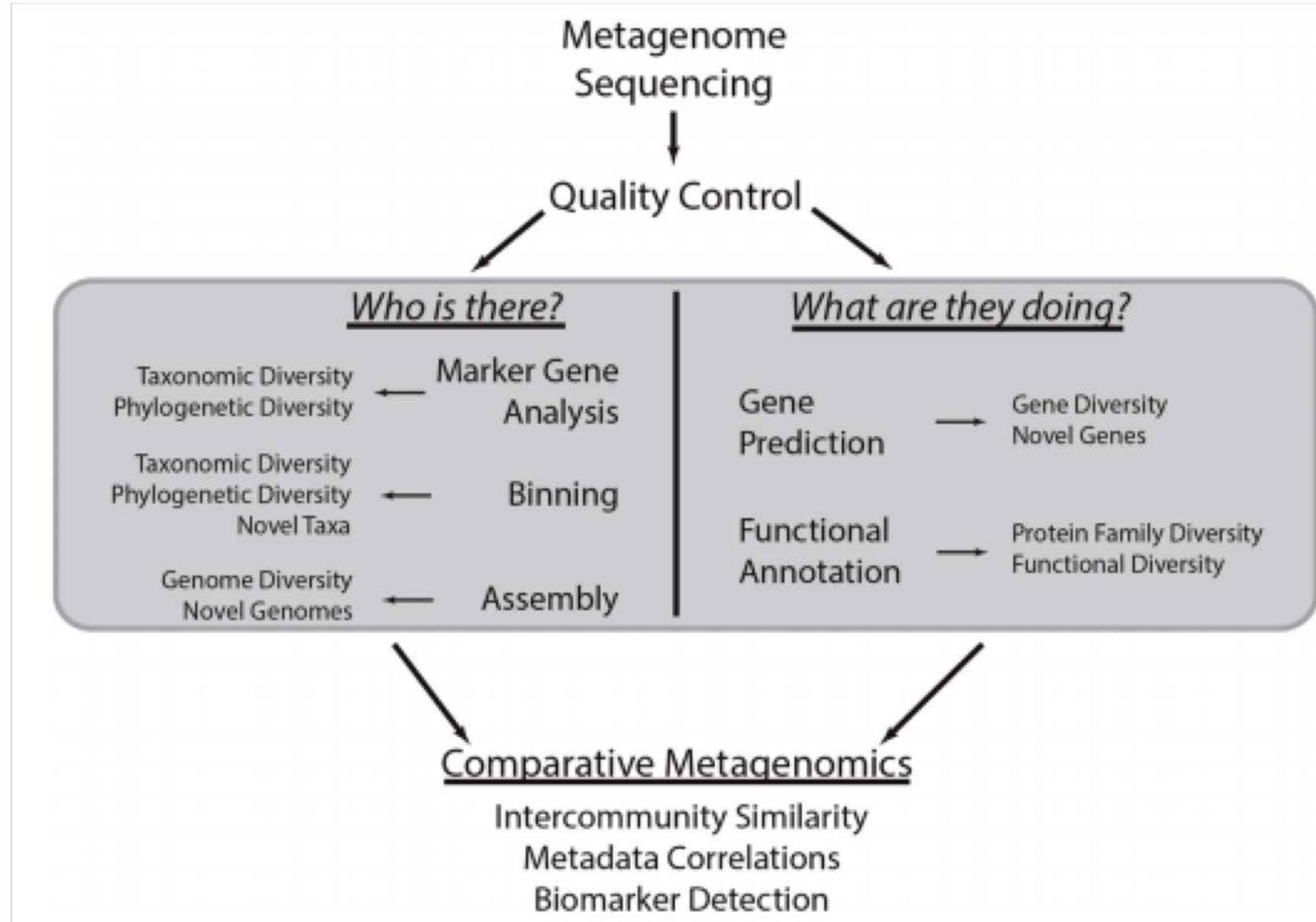
## Issues

**Poor sampling beyond eukaryotic diversity**

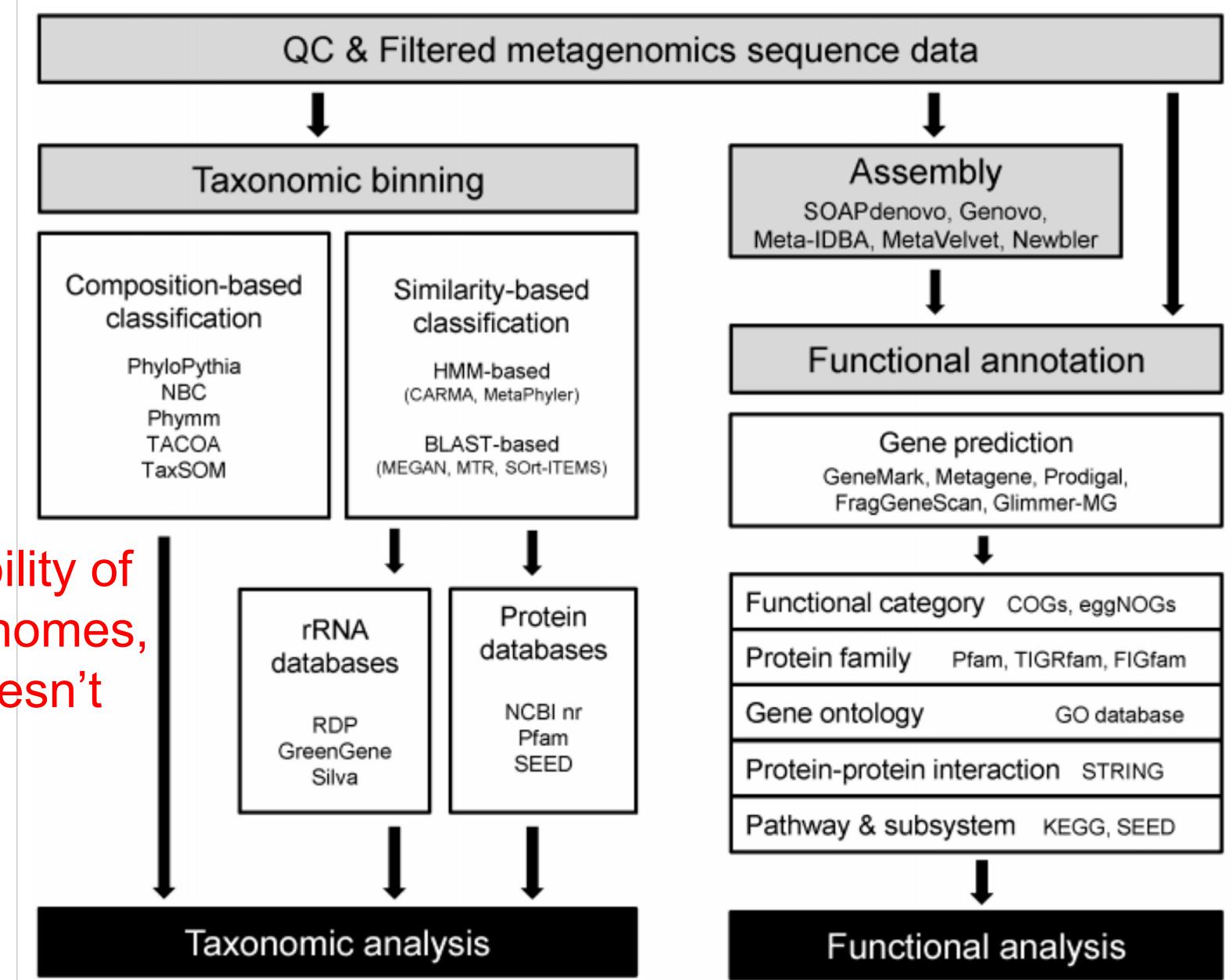
Assembly of metagenomes is **challenging** due to uneven coverage

Requires **high** depth of coverage

# Overall workflow

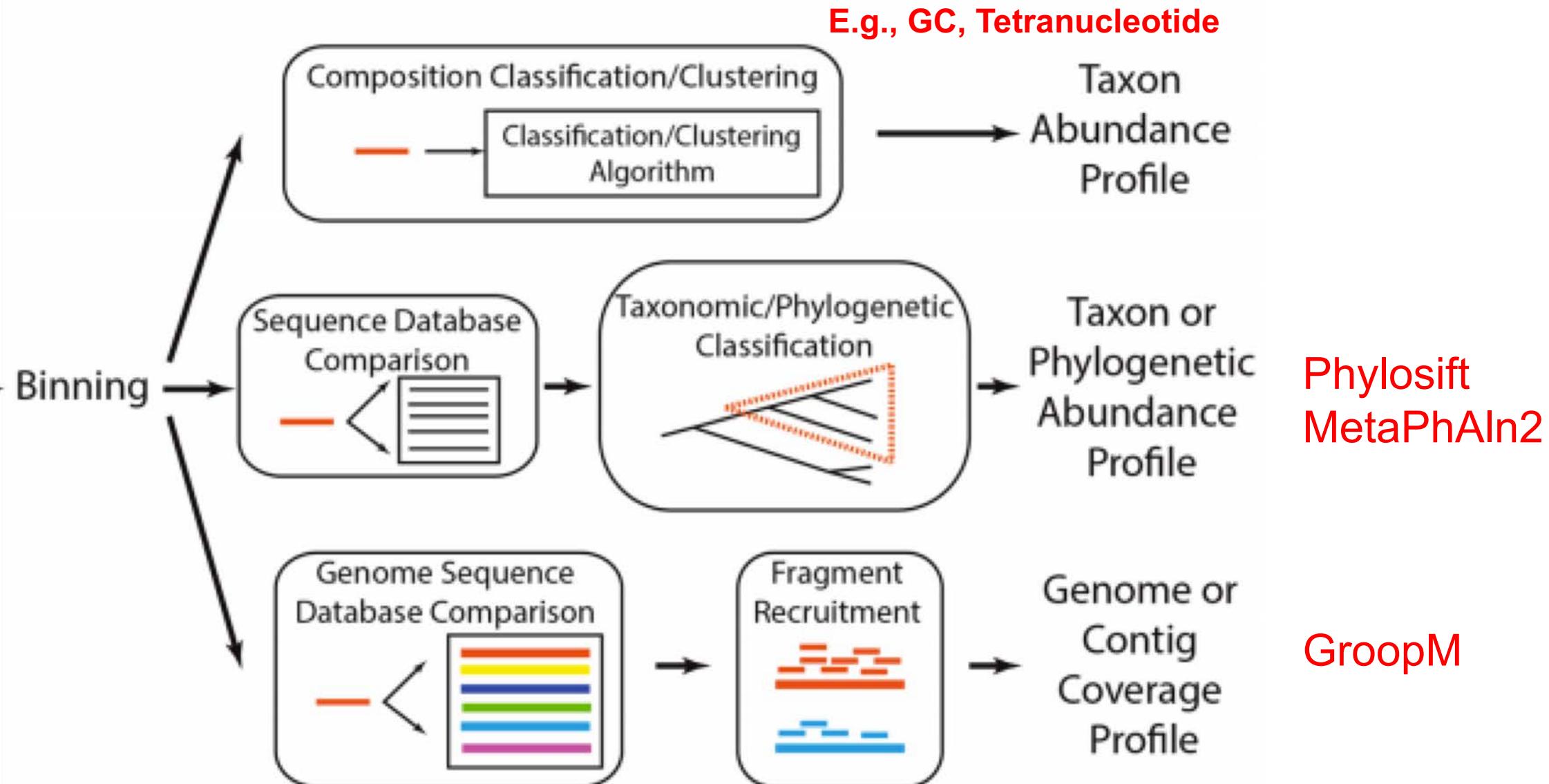


# Overall workflow



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

# Binning methods



# 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 .

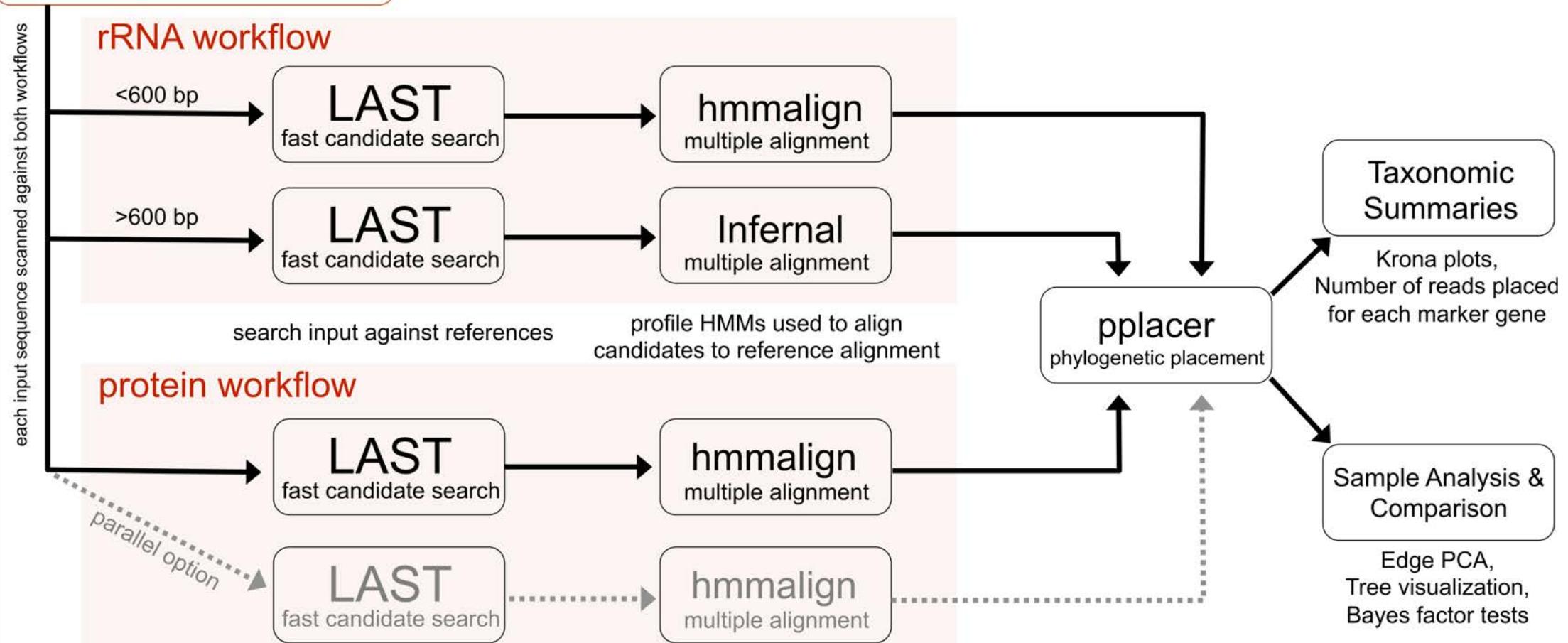
# PhyloSift

mining the global metagenome

<https://phylosift.wordpress.com/>

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

## Input Sequences



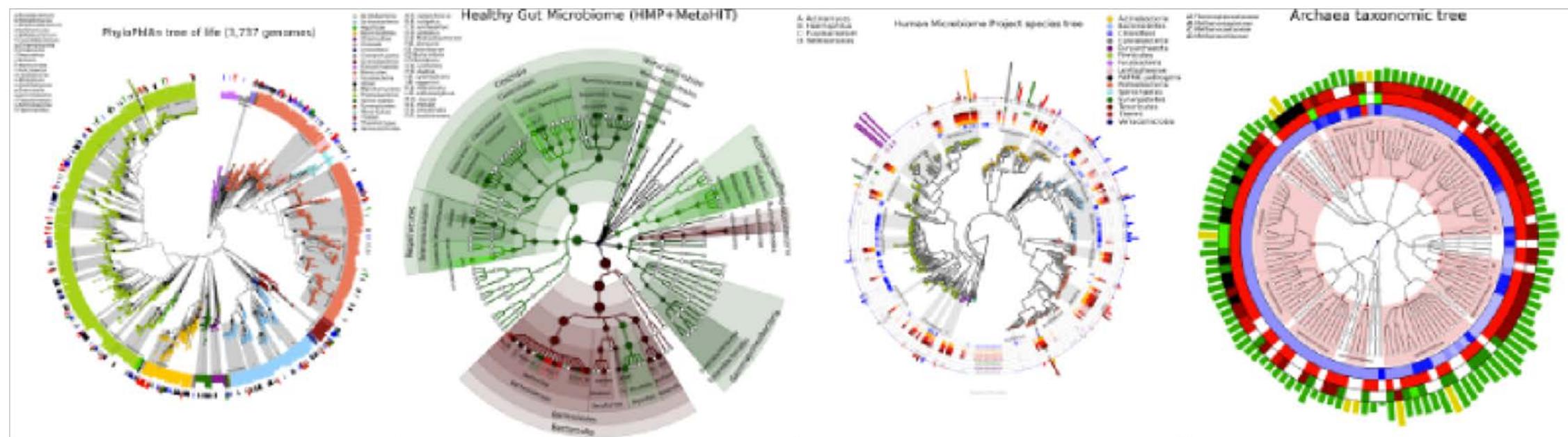
# MetaPhAIn2 – enhanced metagenomic taxonomic profiling

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:

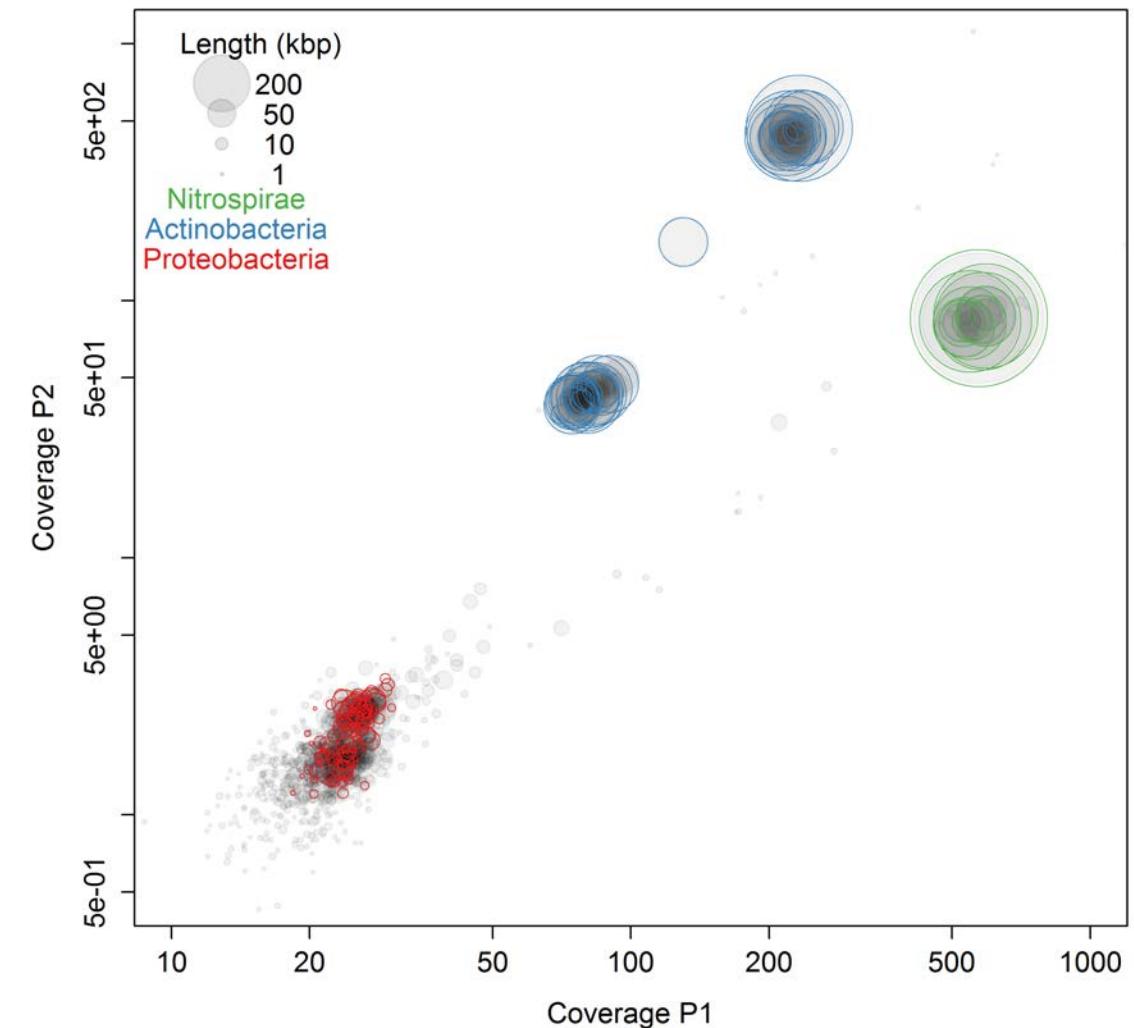
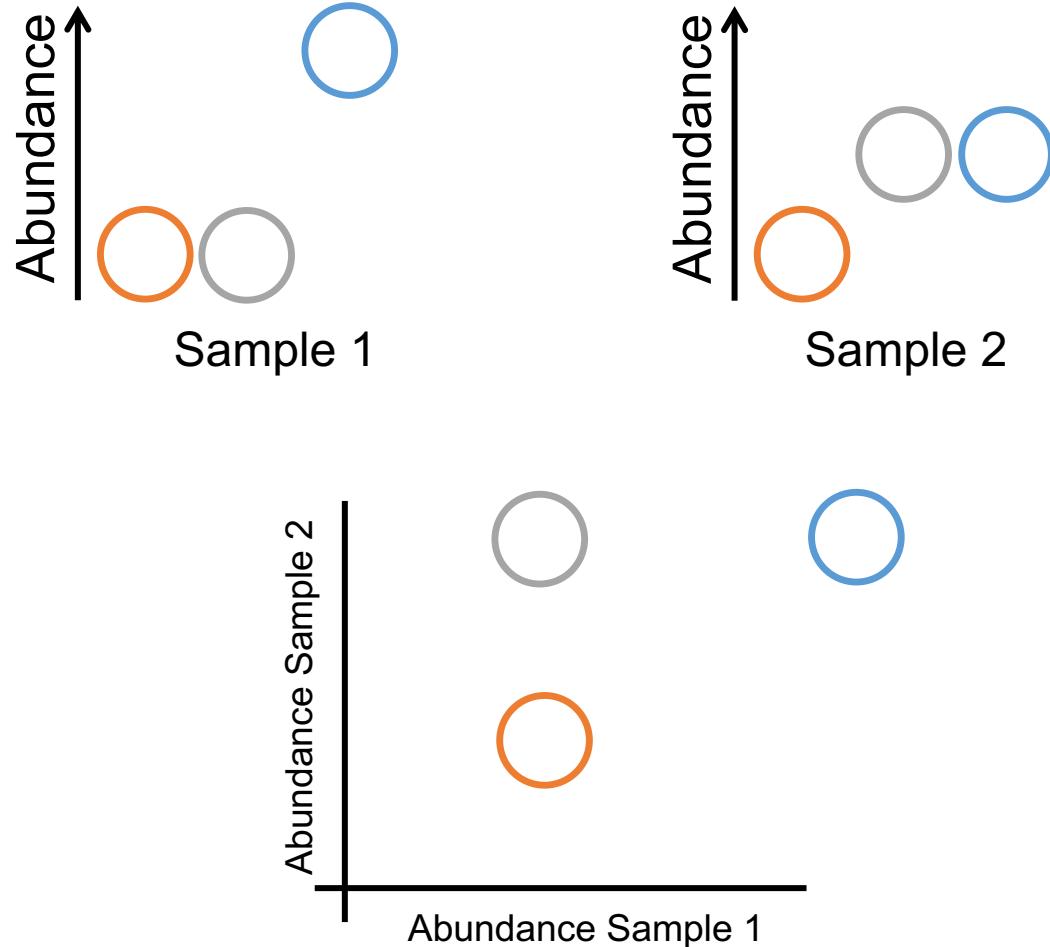
**Species level resolution**

Good visualisation with **GraphAIn**

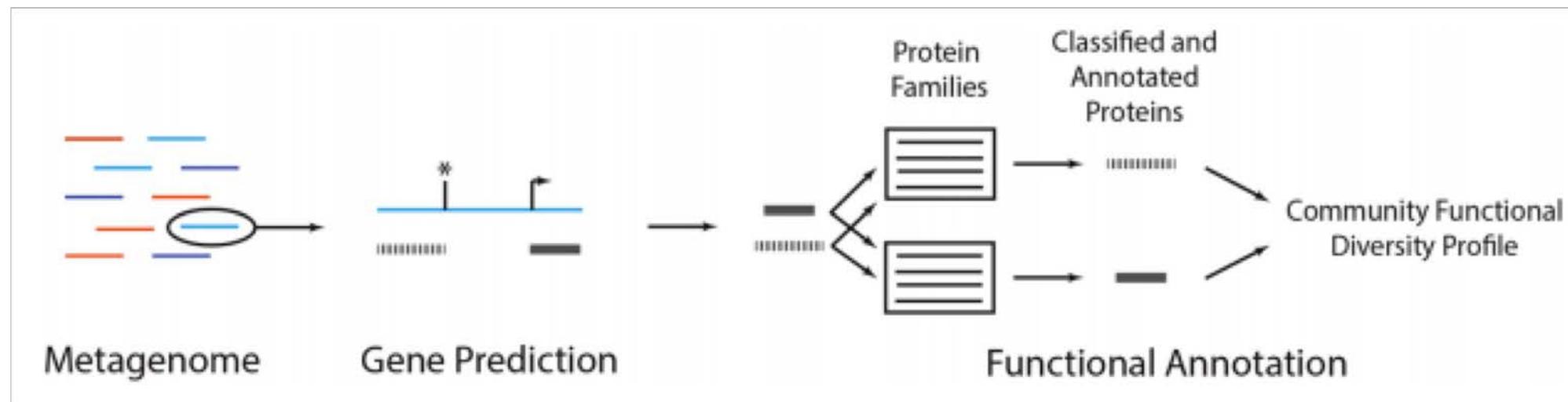
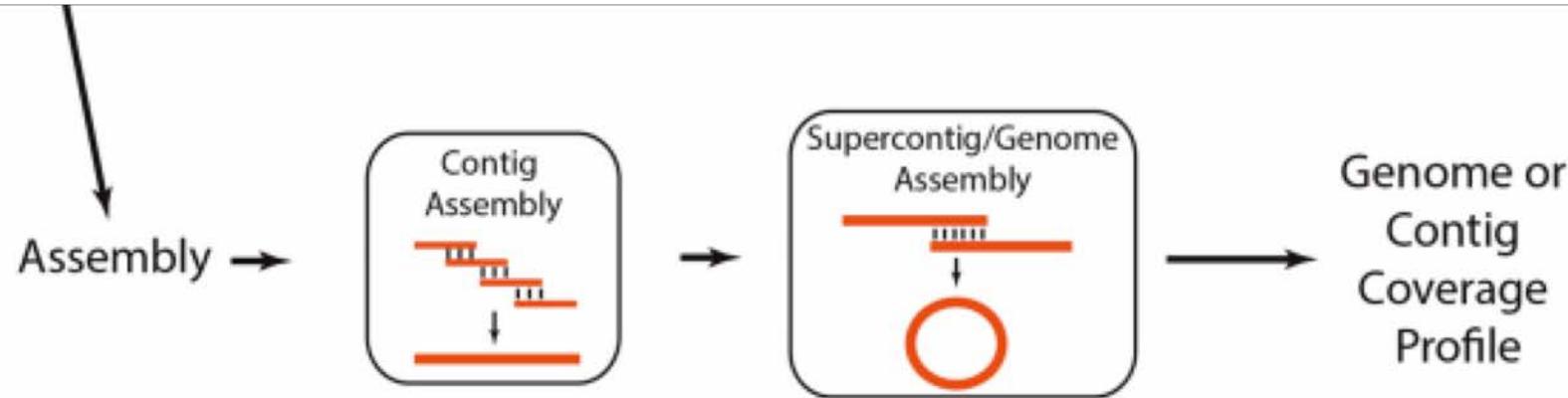
**(So it's useful with known ecosystems)**



# Example of binning based on differential coverage



# Actual assembly



# 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



Soil Biology & Biochemistry 36 (2004) 1607–1614

Soil Biology &  
Biochemistry

[www.elsevier.com/locate/soilbio](http://www.elsevier.com/locate/soilbio)

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

Julia R. de Liphay<sup>a,b</sup>, Christiane Enzinger<sup>b,1</sup>, Kaare Johnsen<sup>a,2</sup>, Jens Aamand<sup>a</sup>,  
Søren J. Sørensen<sup>b,\*</sup>

<sup>a</sup>Department of Geochemistry, Geological Survey of Denmark and Greenland, Øster Voldgade 10, DK-1350 Copenhagen K, Denmark

<sup>b</sup>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 group—and thus the indication of metabolic activity—may depend on the nature of the applied DNA extraction procedure.

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

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



**Microbiome**

Open Access

### RESEARCH

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

Agata Wesolowska-Andersen<sup>1</sup>, Martin Iain Bah<sup>2</sup>, Vera Carvalho<sup>2</sup>, Karsten Kristiansen<sup>3</sup>, Thomas Sicheritz-Pontén<sup>1</sup>, Ramneek Gupta<sup>1\*</sup> and Tine Rask Licht<sup>2\*</sup>

### 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 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.

Read filtering	Number of OTUs at percentage identity thresholds						% errorless reads	% reads used
	100	99	98	97	95	90		
<b>5' forward (V1 and V2)</b>								
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
<b>3' reverse (V8)</b>								
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<sup>1\*</sup>, Michael J Cox<sup>2</sup>, Elena M Turek<sup>2</sup>, Szymon T Calus<sup>3</sup>, William O Cookson<sup>2</sup>, Miriam F Moffatt<sup>2</sup>, Paul Turner<sup>4,5</sup>, Julian Parkhill<sup>1</sup>, Nicholas J Loman<sup>3</sup> and Alan W Walker<sup>1,6\*</sup>

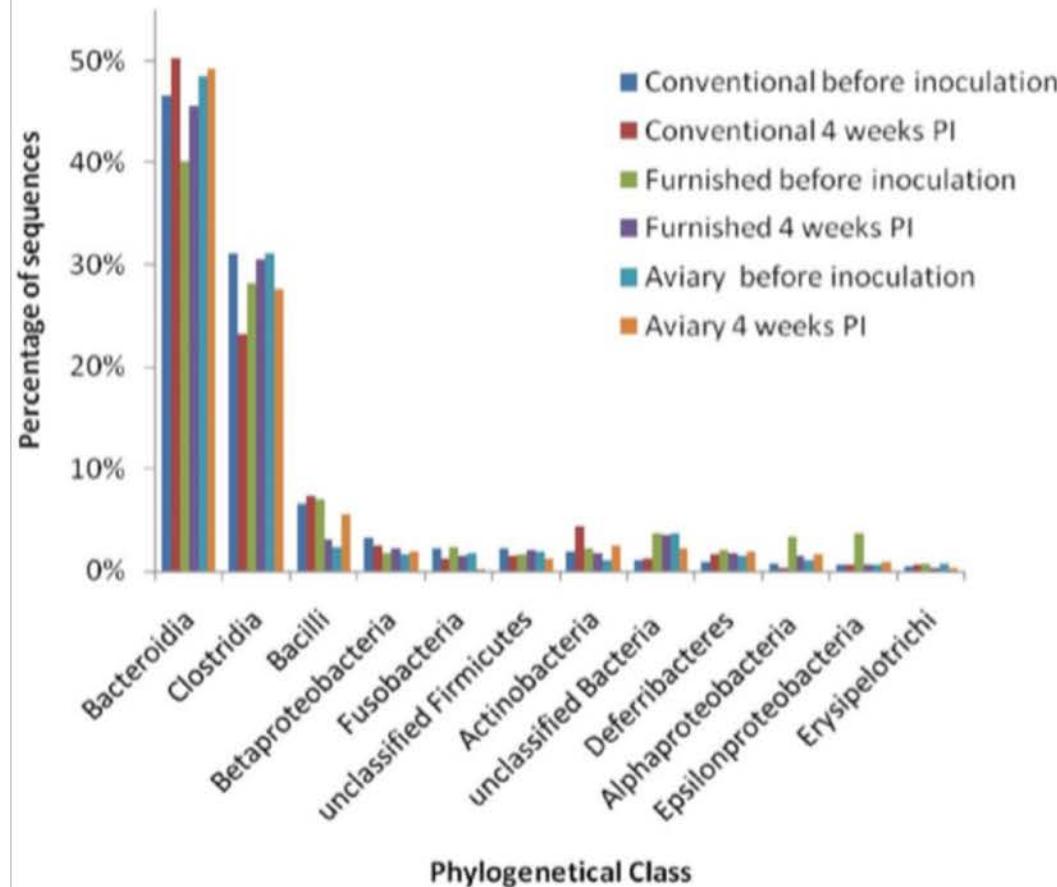
RESEARCH HIGHLIGHT

## Tracking down the sources of experimental contamination in microbiome studies

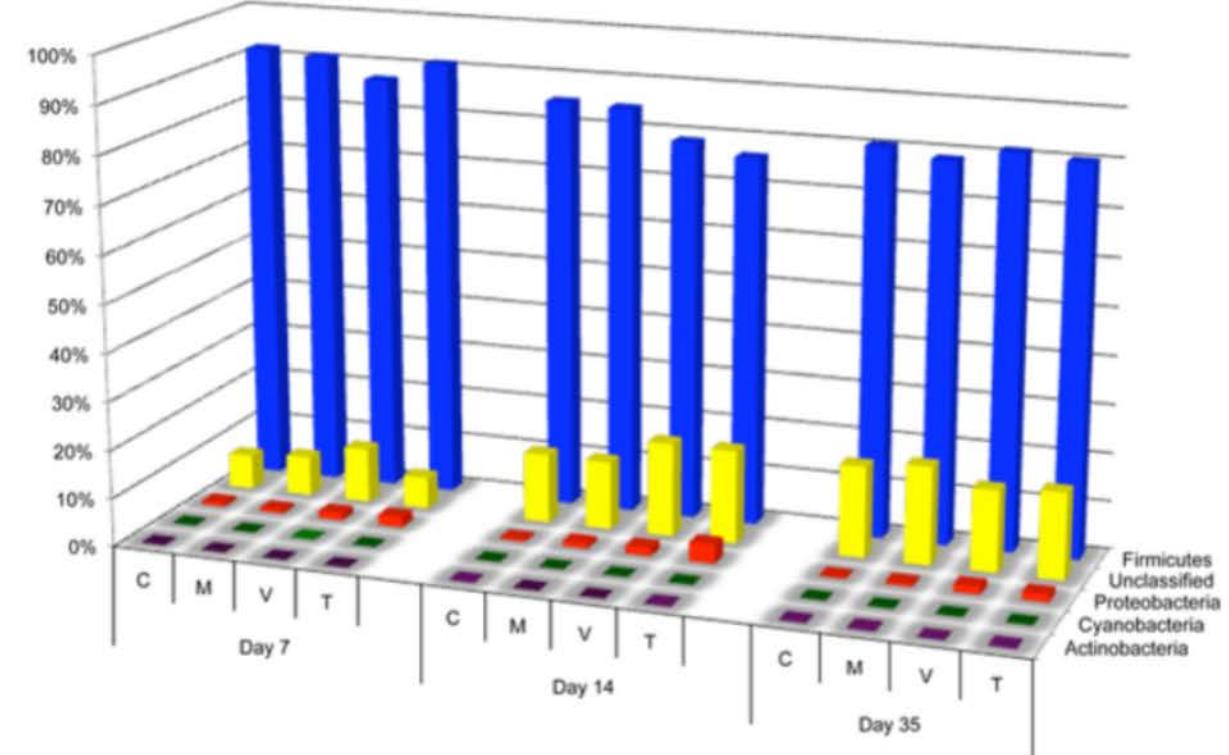
Sophie Weiss<sup>1</sup>, Amnon Amir<sup>2</sup>, Embriette R Hyde<sup>2</sup>, Jessica L Metcalf<sup>2</sup>, Se Jin Song<sup>2</sup> and Rob Knight<sup>2,3,4\*</sup>

# 2 papers with different results at the same year

Bacteroidetes >> rest

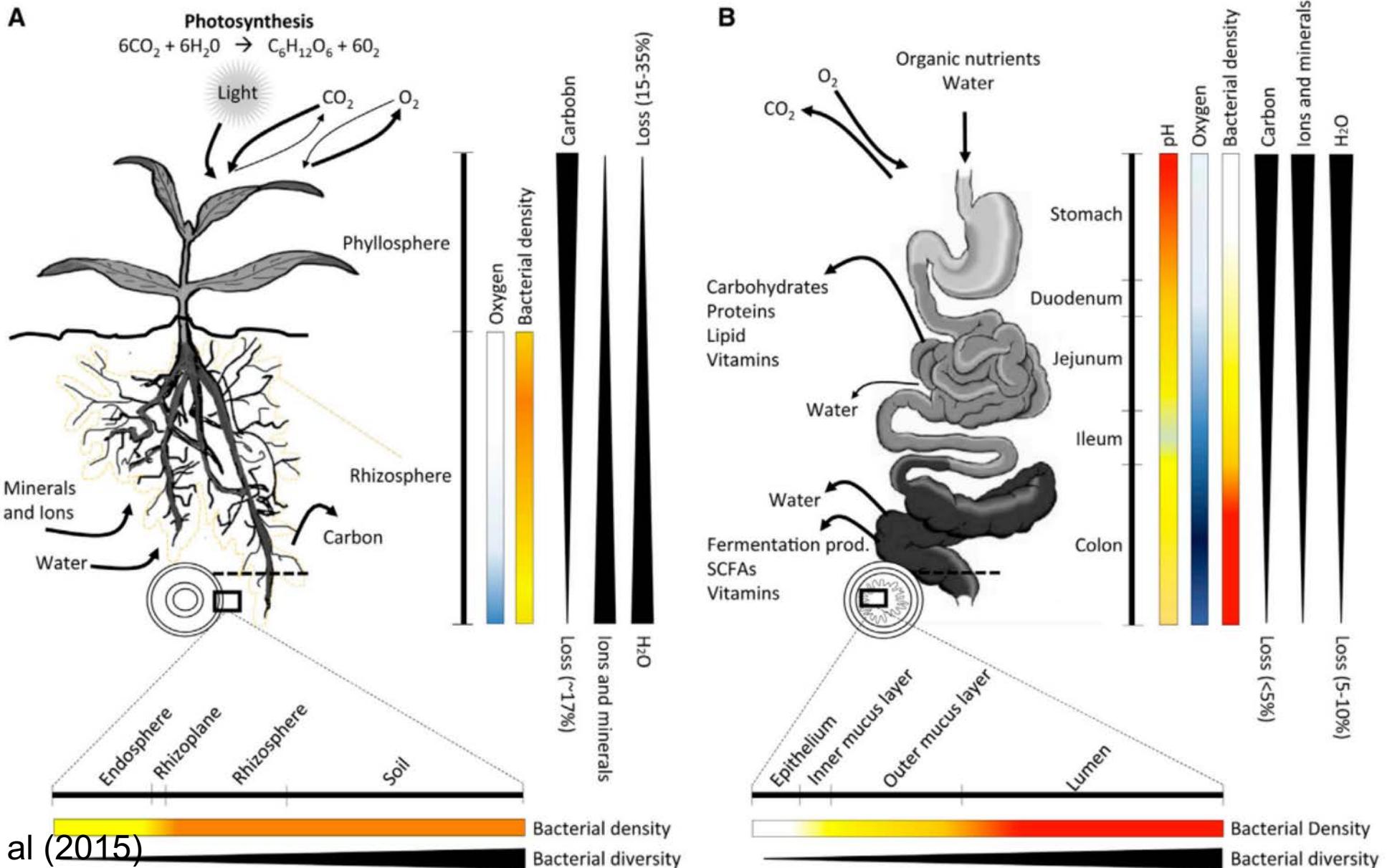


firmicutes >> rest > bacteroidetes

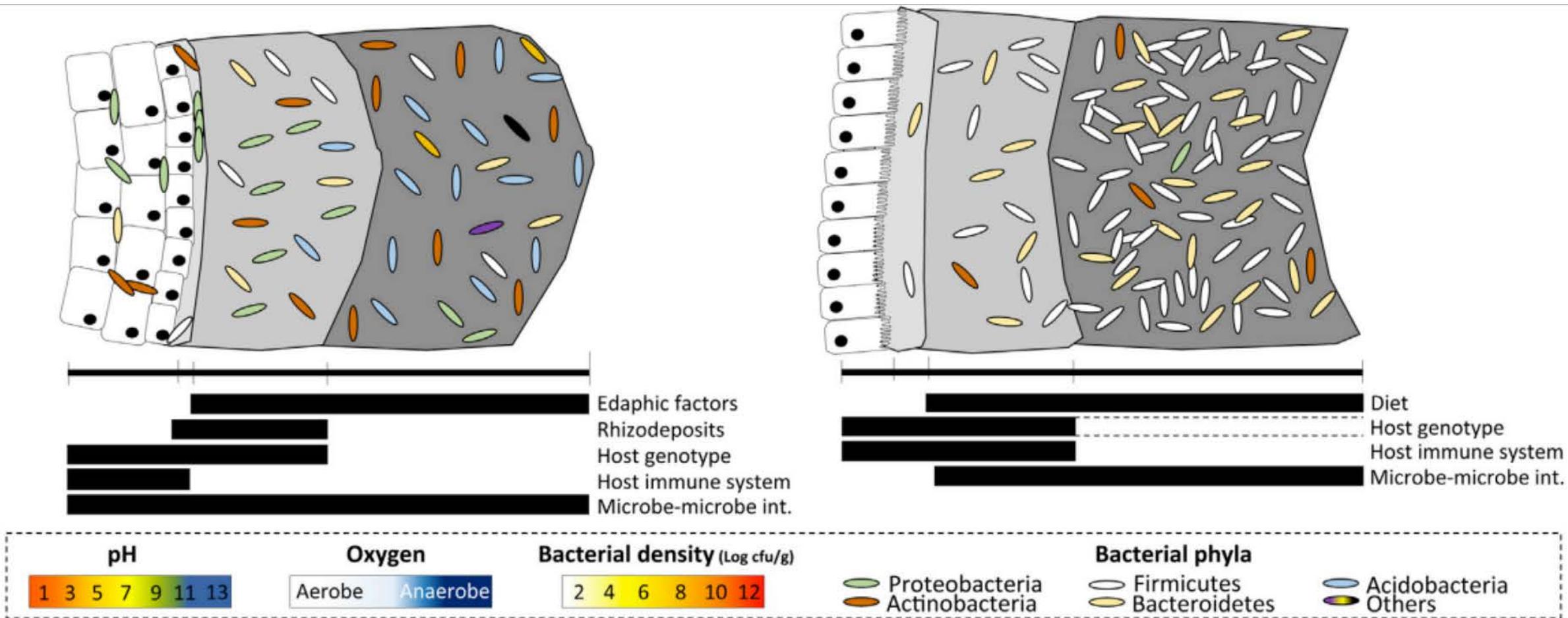


# Case studies

# Two most common systems



# Two most common systems



# Two most common systems

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

	Cucumber <sup>a</sup>	Wheat <sup>a</sup>	Soybean <sup>b</sup>	Wheat <sup>c</sup>	Oat <sup>c</sup>	Pea <sup>c</sup>	Barley <sup>d</sup>	Gut <sup>e</sup>
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

<sup>a</sup>Ofek-Lalzar et al. (2014) (metagenomics of rhizoplane samples).

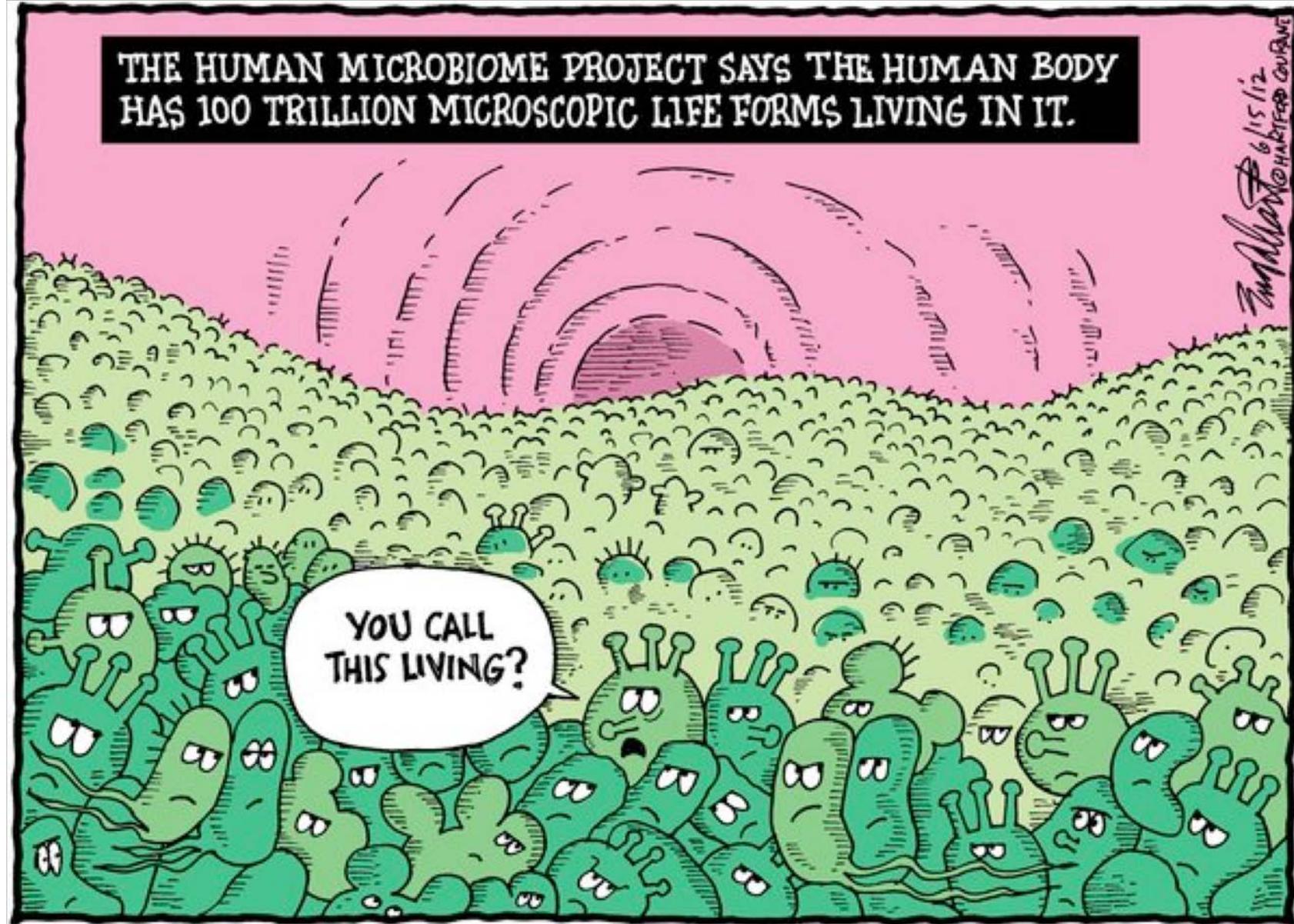
<sup>b</sup>Mendes et al. (2014) (metagenomics of rhizosphere samples).

<sup>c</sup>Turner et al. (2013) (metatranscriptomics of rhizosphere samples).

<sup>d</sup>Bulgarelli et al. (2015) (metagenomics of rhizosphere samples).

<sup>e</sup>Qin et al. (2010) (metagenomics of gut samples).

# Human gut microbiome

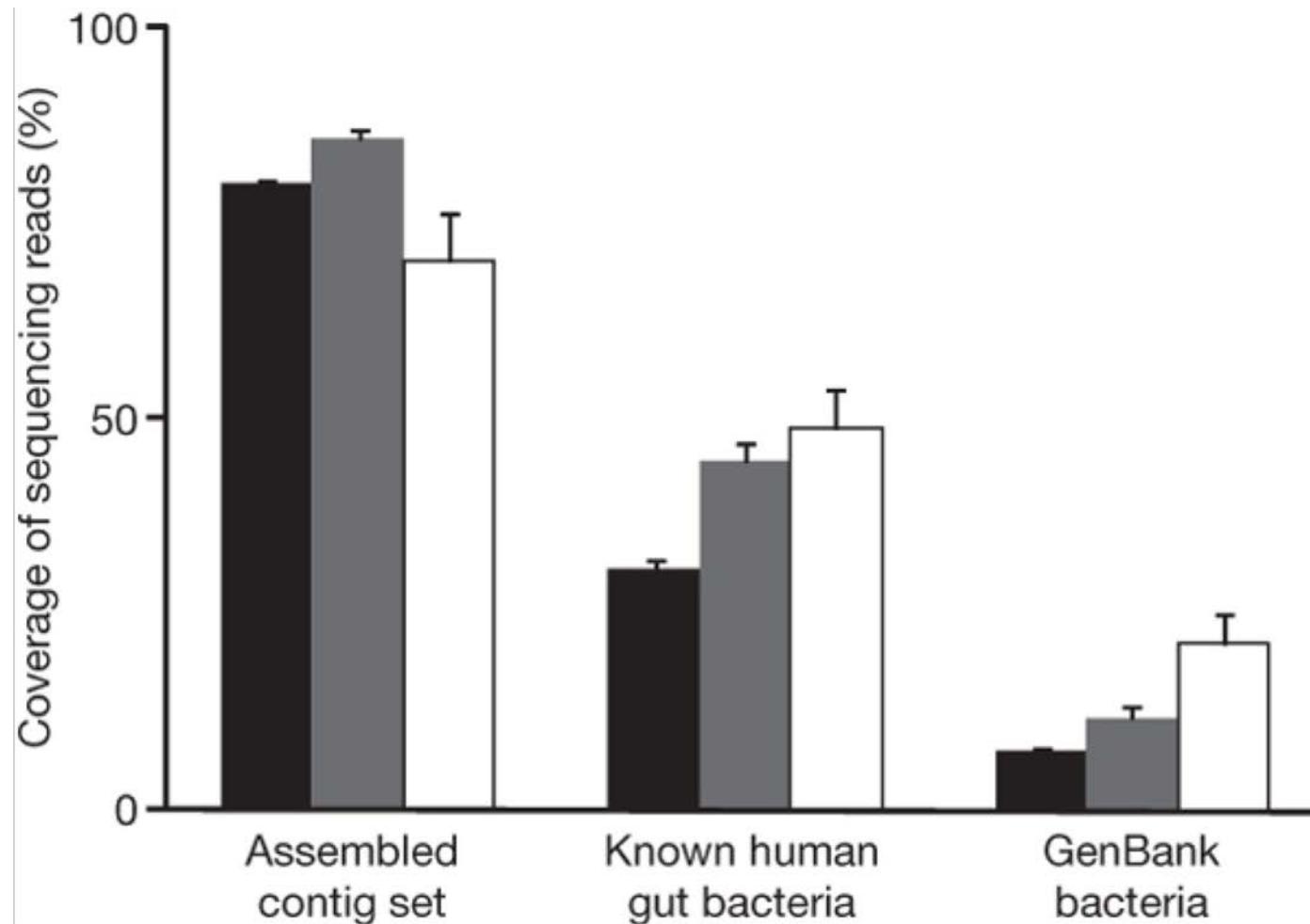


## ARTICLES

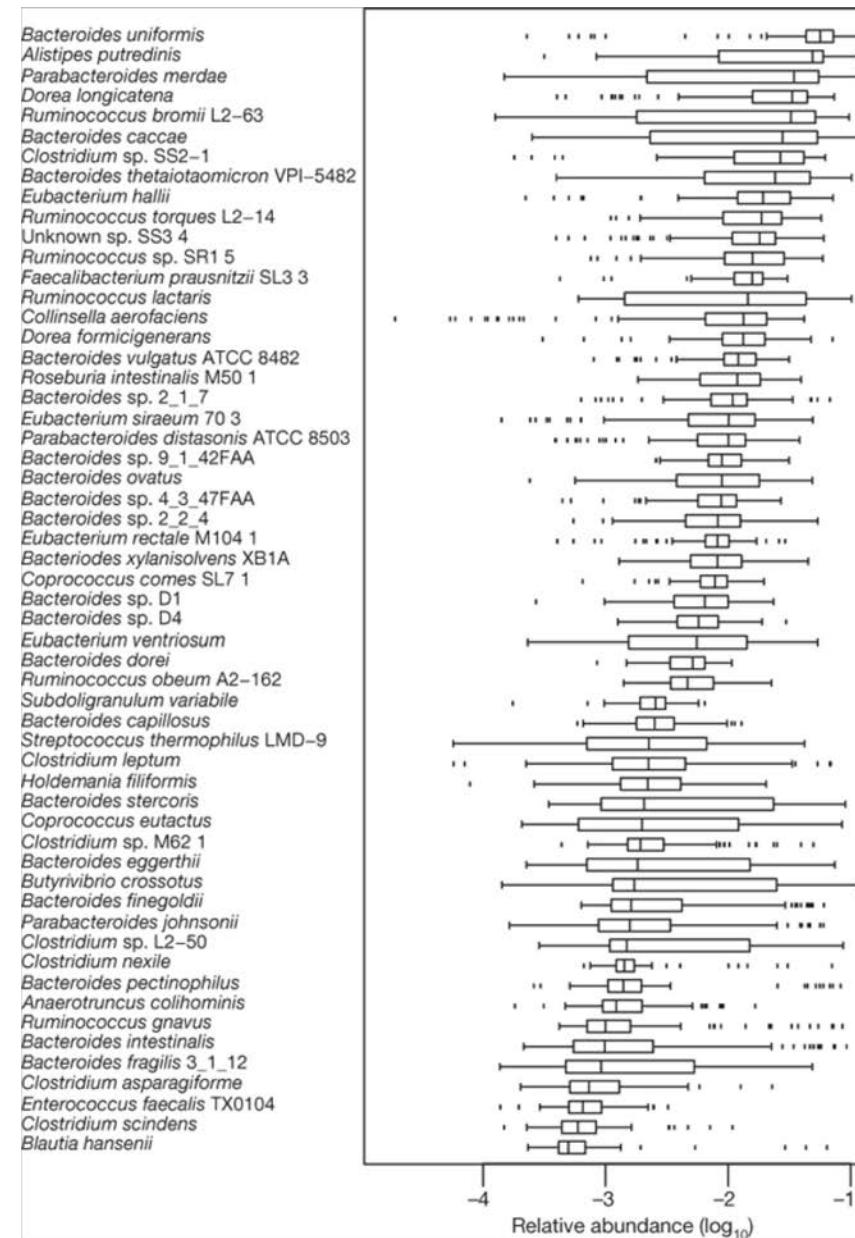
# A human gut microbial gene catalogue established by metagenomic sequencing

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

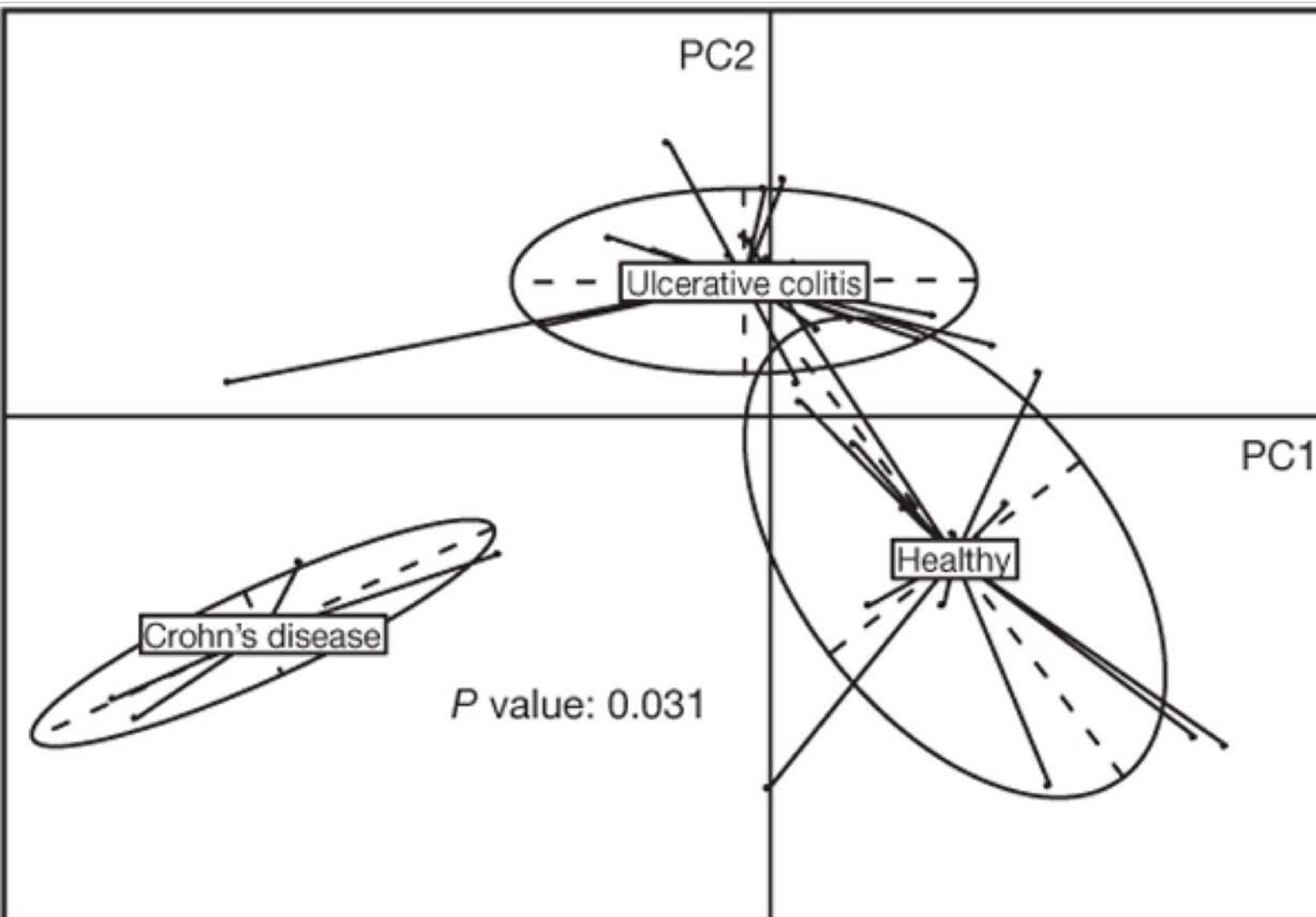
# Human gut microbiome



# Human gut microbiome



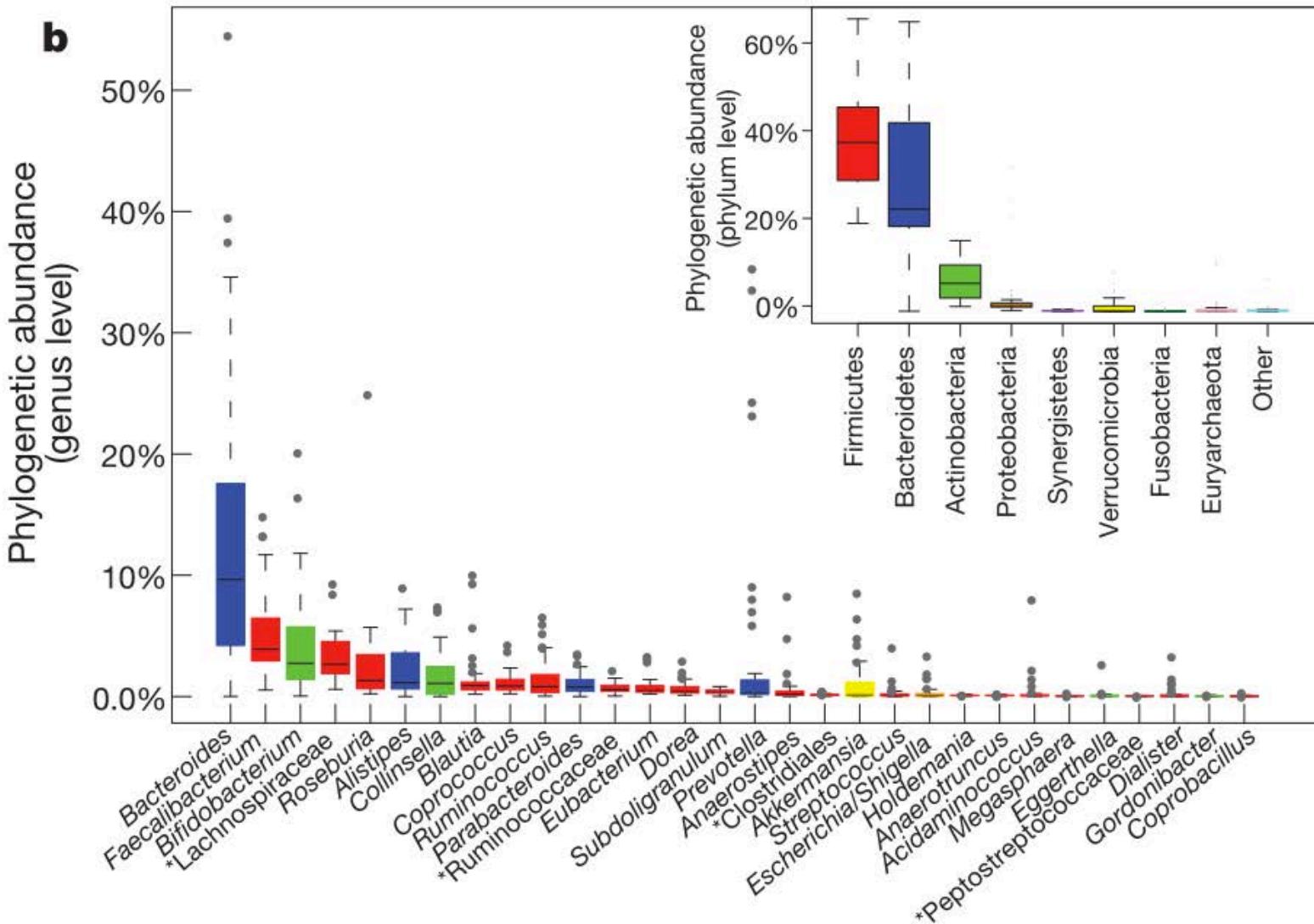
# Human gut microbiome

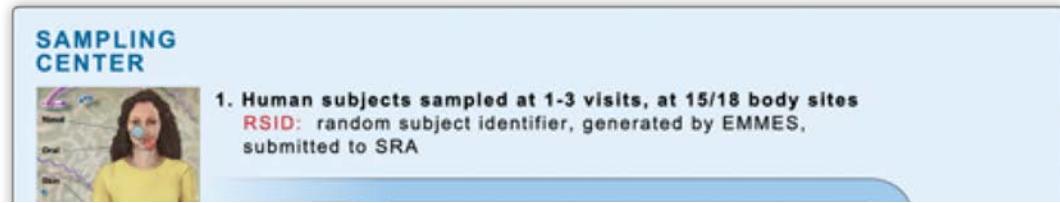


We can check which OTUs constitute the clustering (and separation) patterns

- > Biology
- > Biomarkers

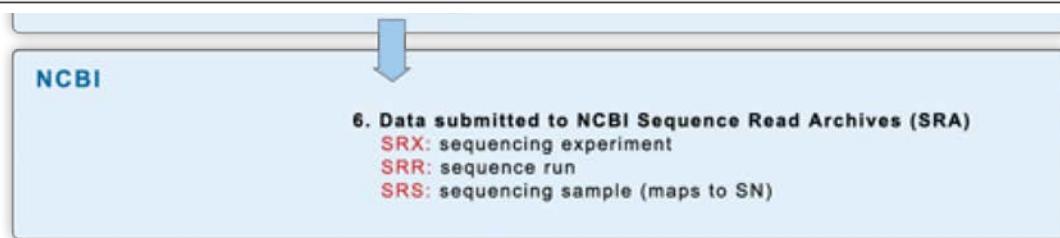
# Human gut microbiome





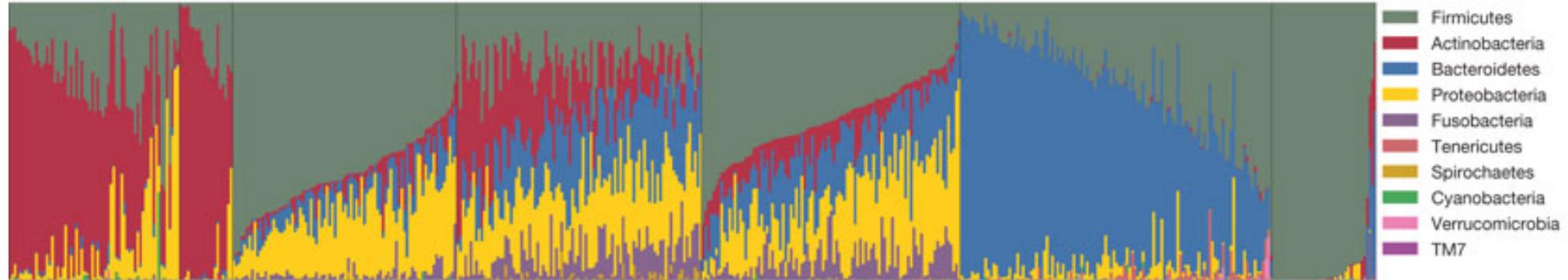
**Table 1 | HMP donor samples examined by 16S and WGS**

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
Airway	Tongue dorsum	348	332	193	1.3	320	2.0	181	122	1,620.1	15	19	1,084.3	106
	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
Vagina	Right retroauricular crease	319	316	190	1.4	304	1.6	178	15	181.9	18	59	42.4	12
	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

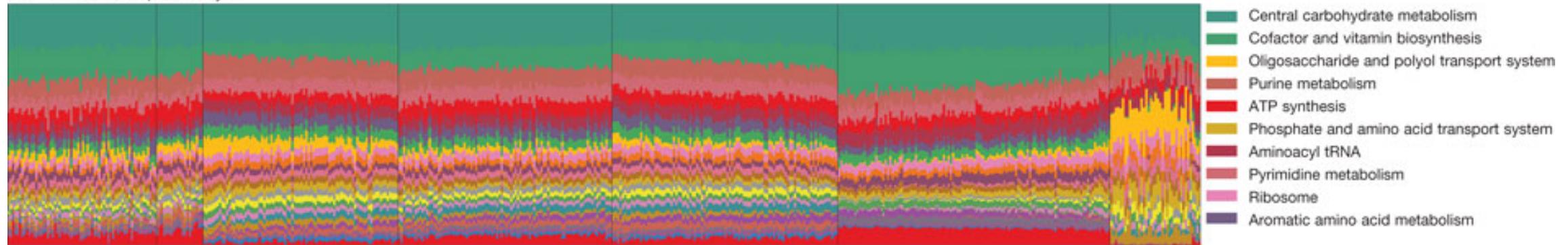


# Human microbiome

**a** Phyla



**b** Metabolic pathways



Anterior nares

RC

Buccal mucosa

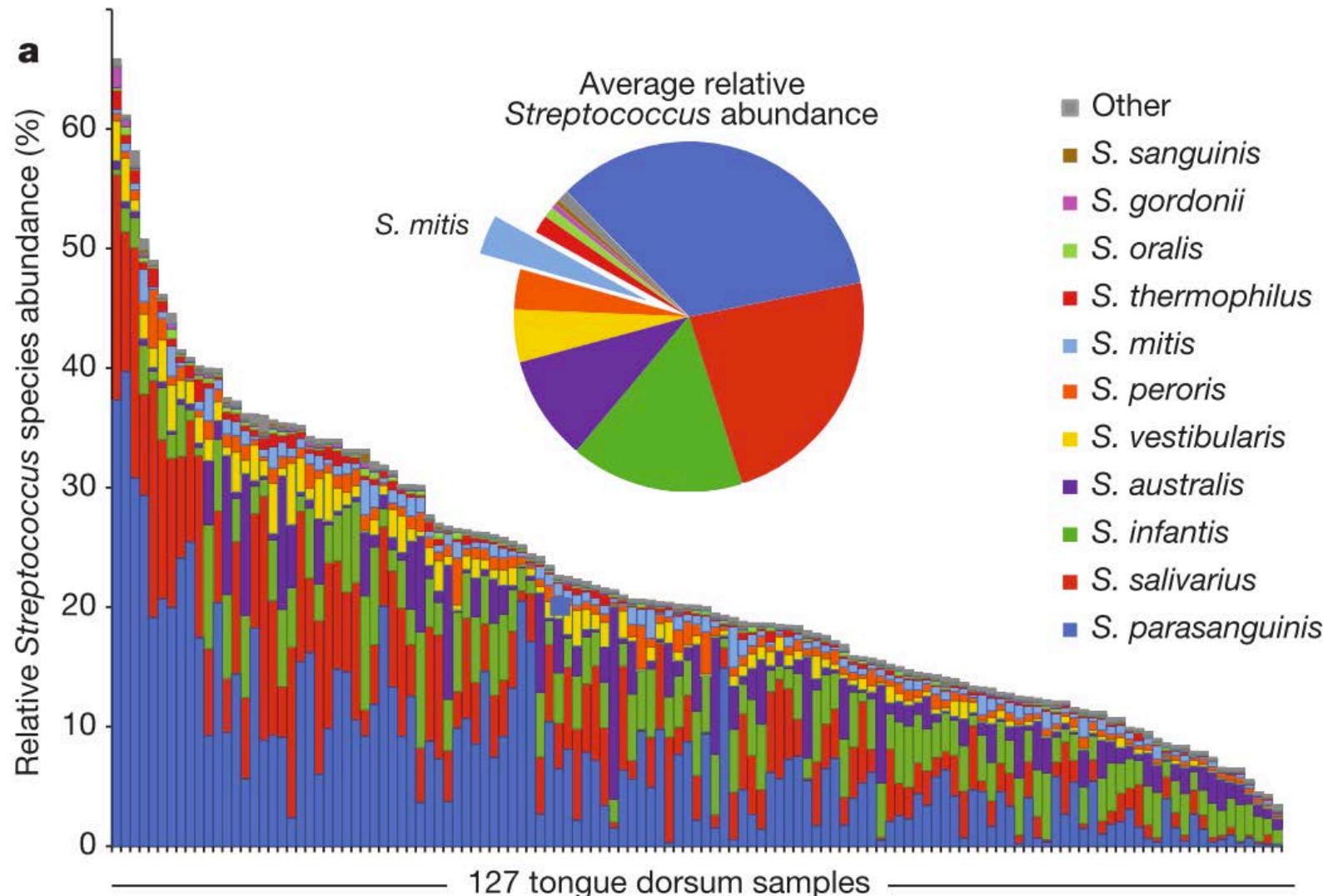
Supragingival plaque

Tongue dorsum

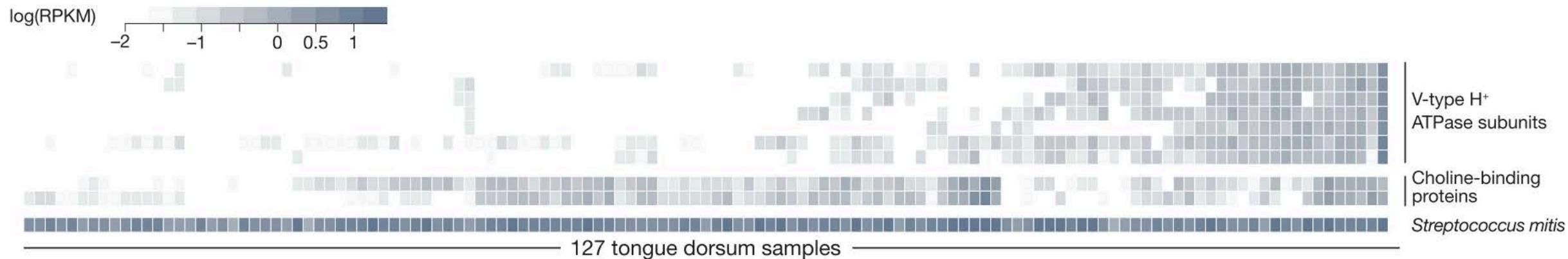
Stool

Posterior fornix

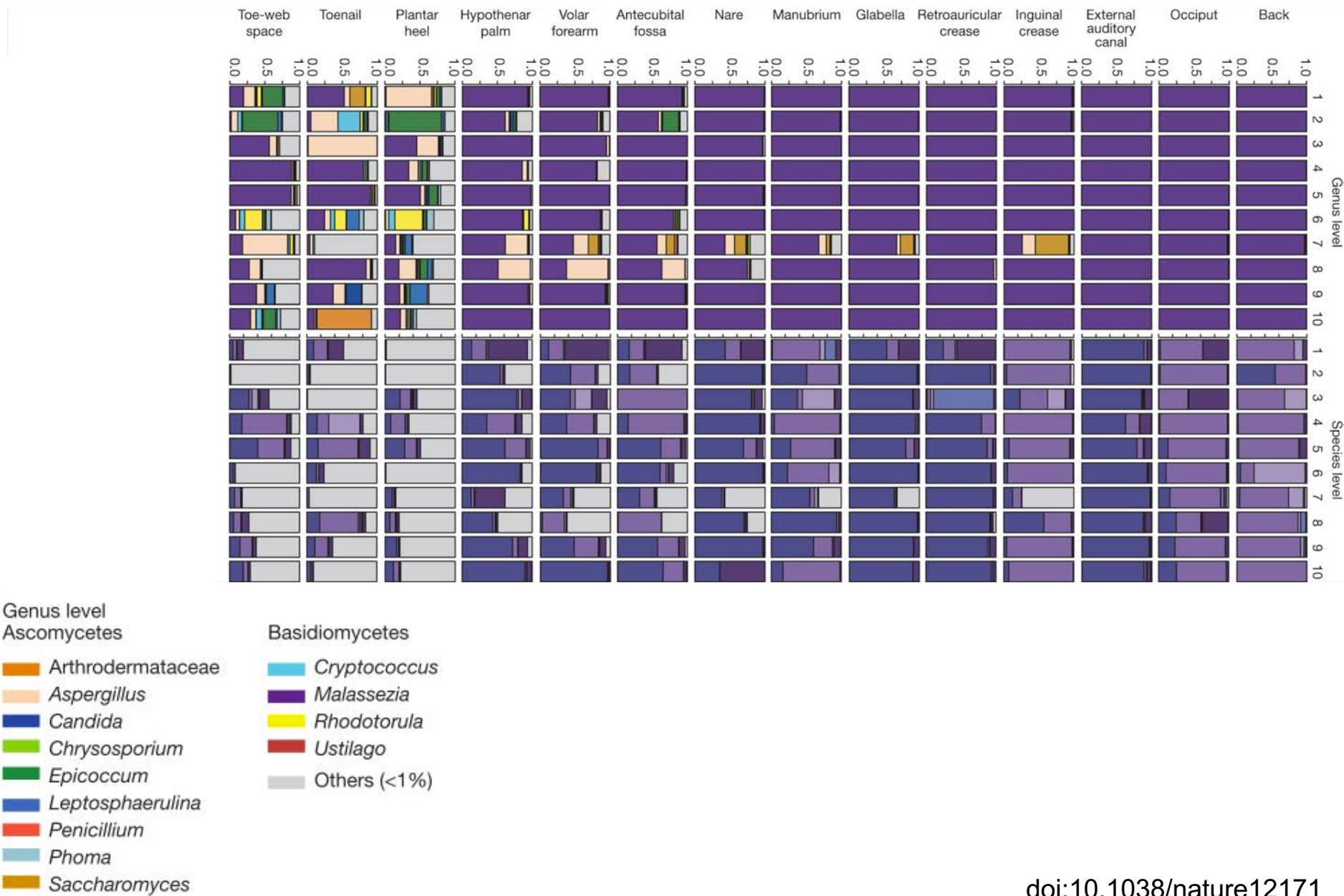
# Inter-individual variation in the microbiome proved to be specific, functionally relevant and personalized



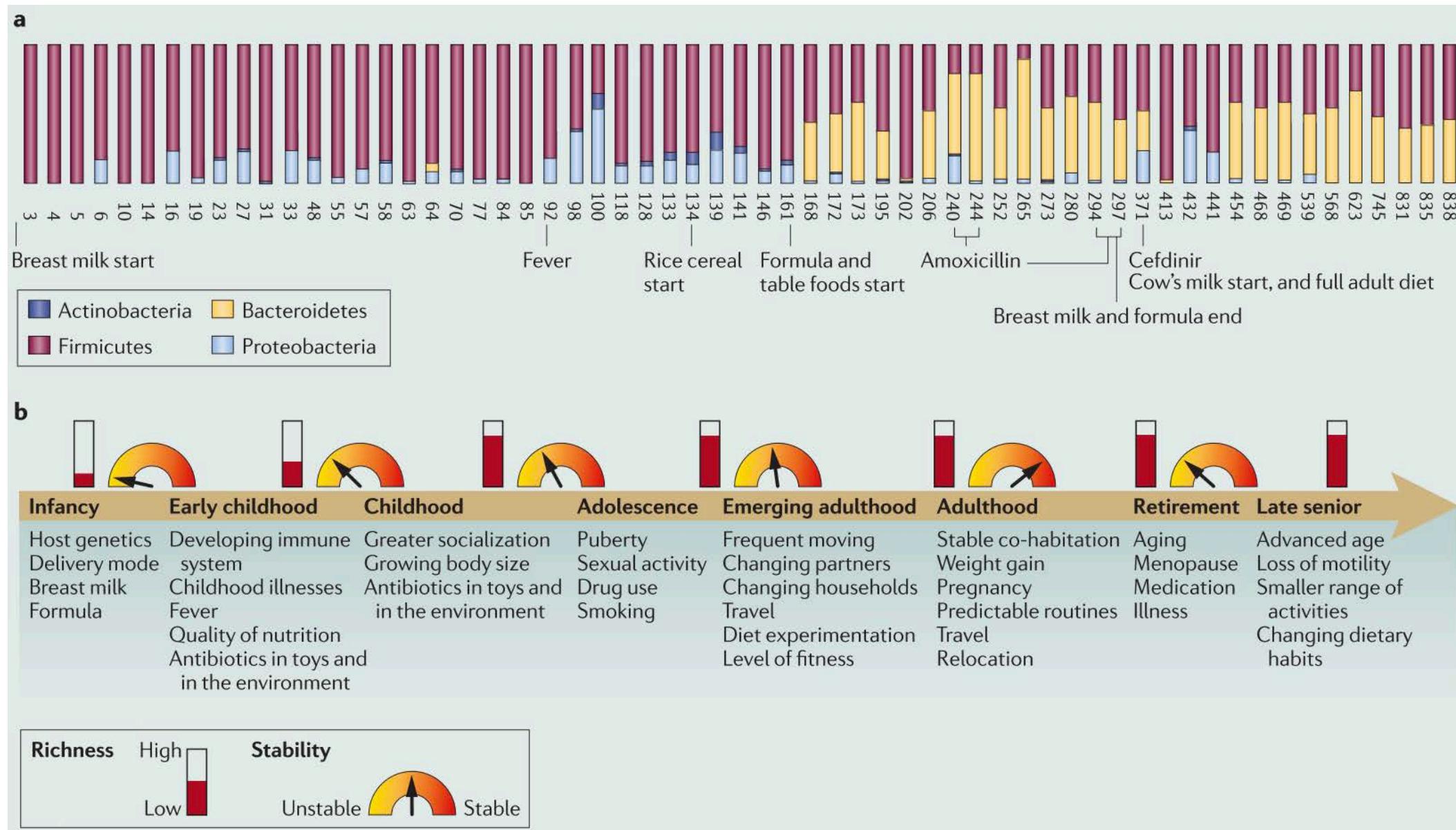
# Gene loss & Structural variants are common



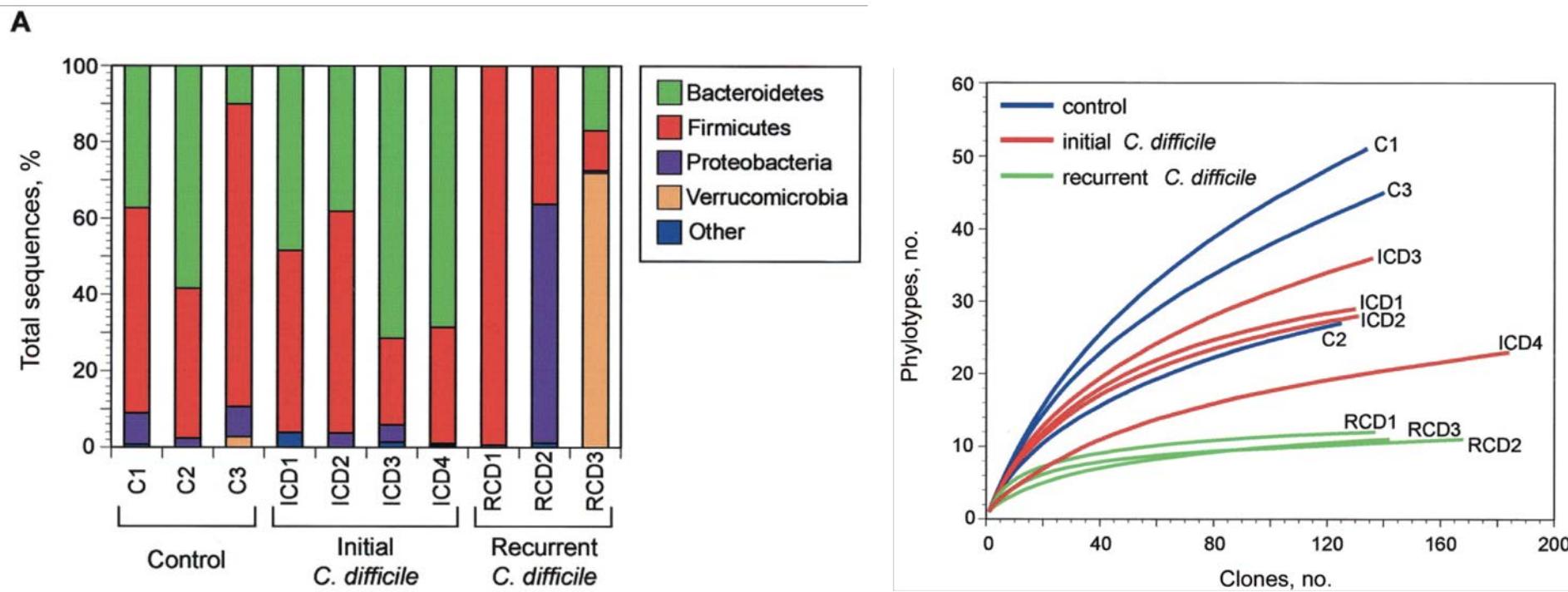
# Skins



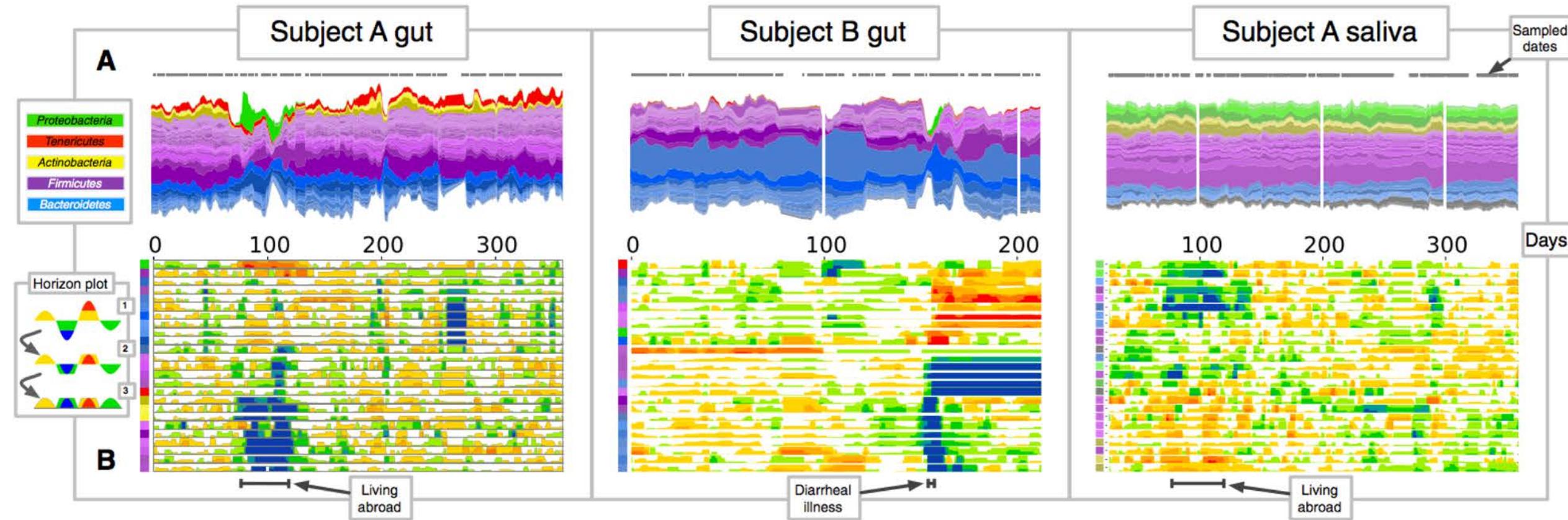
# The gut microbiome during life



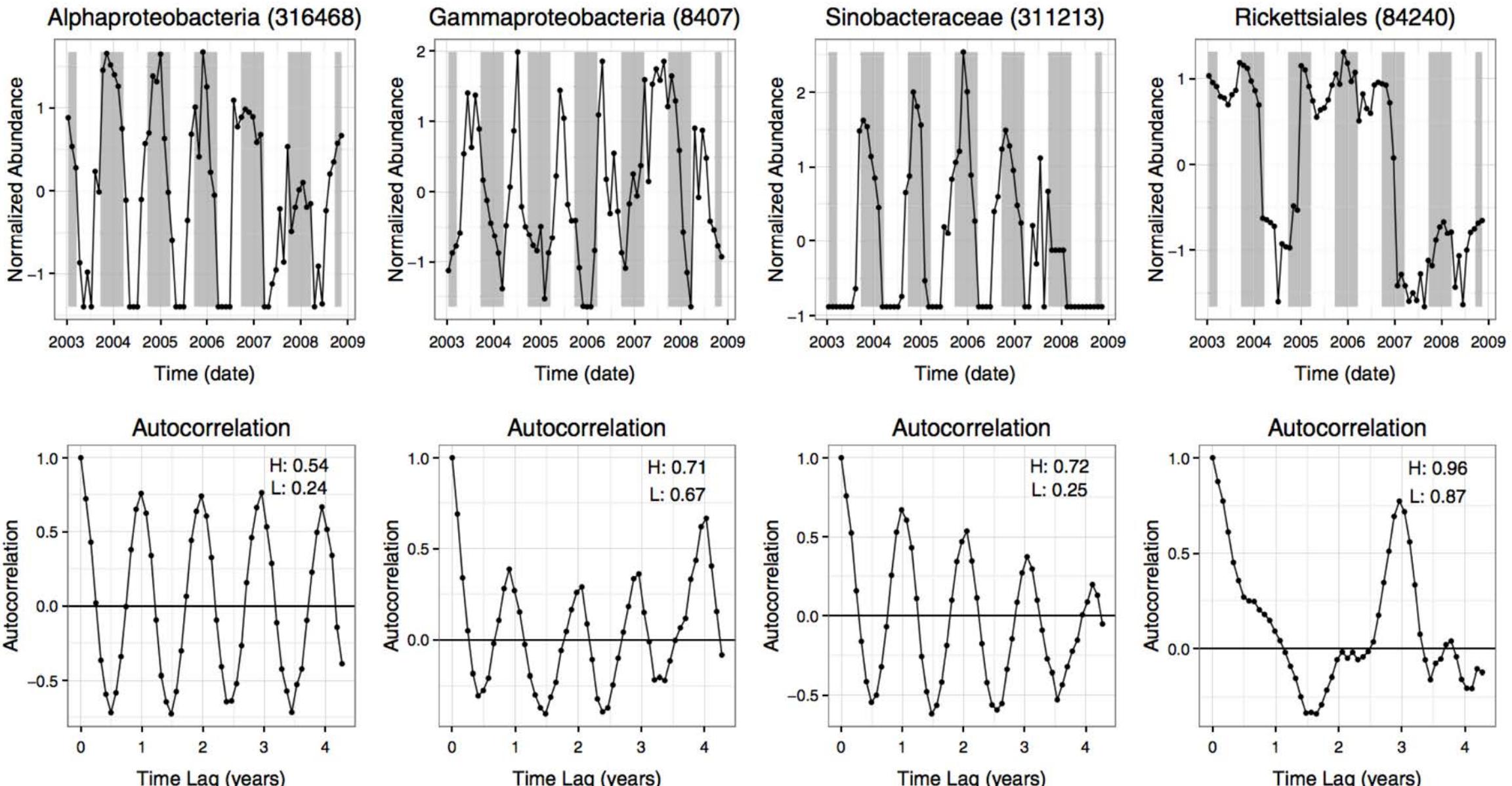
# Decreased diversity with *Clostridium difficile* – associated diarrhea



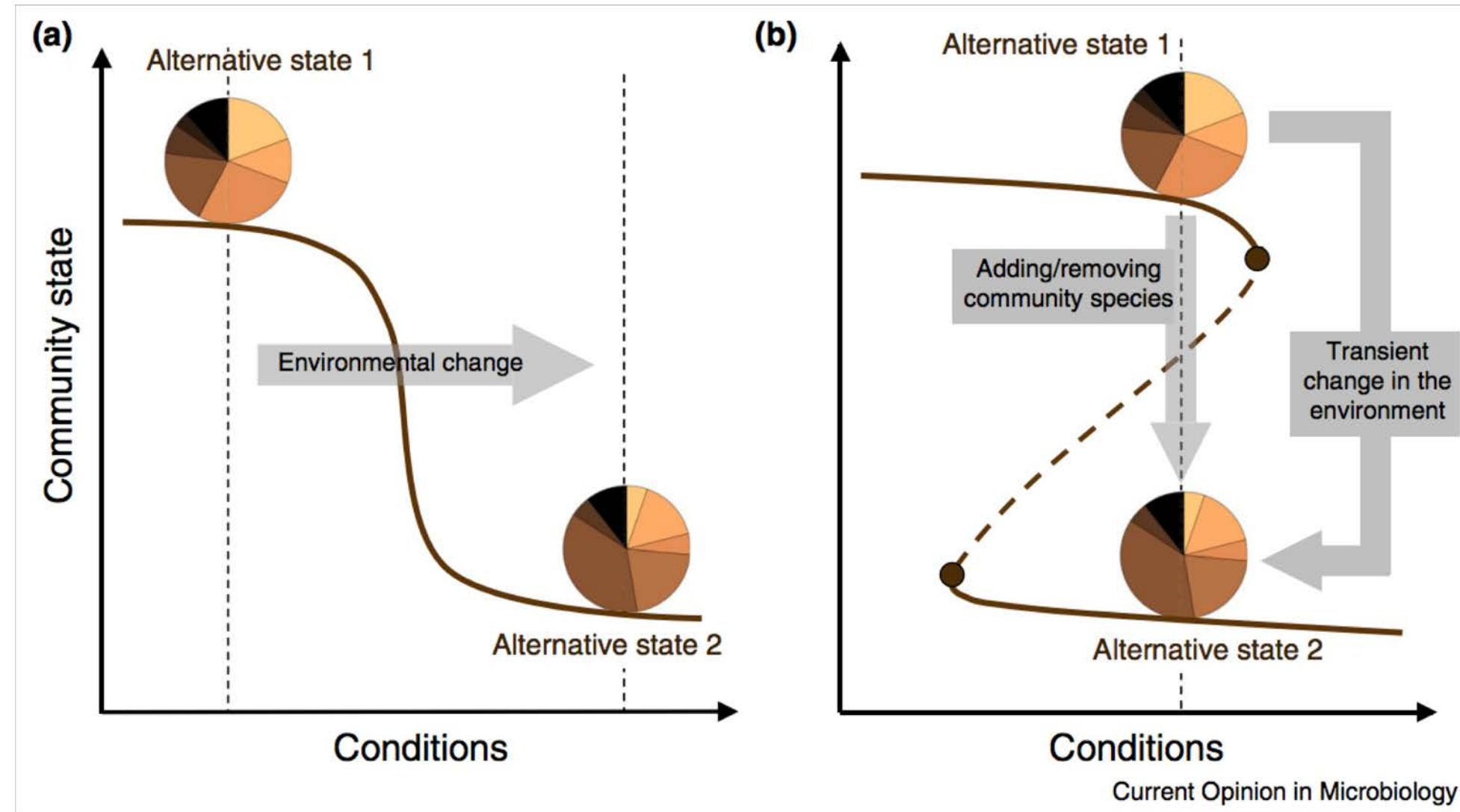
# Tracking microbiome on a daily scale



# Tracking microbiome spanning 6 years



# Tracking microbiome on a daily scale



Current Opinion in Microbiology

# 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 **Yes**  
ISME J, 6 (2012), pp. 2153–2167

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

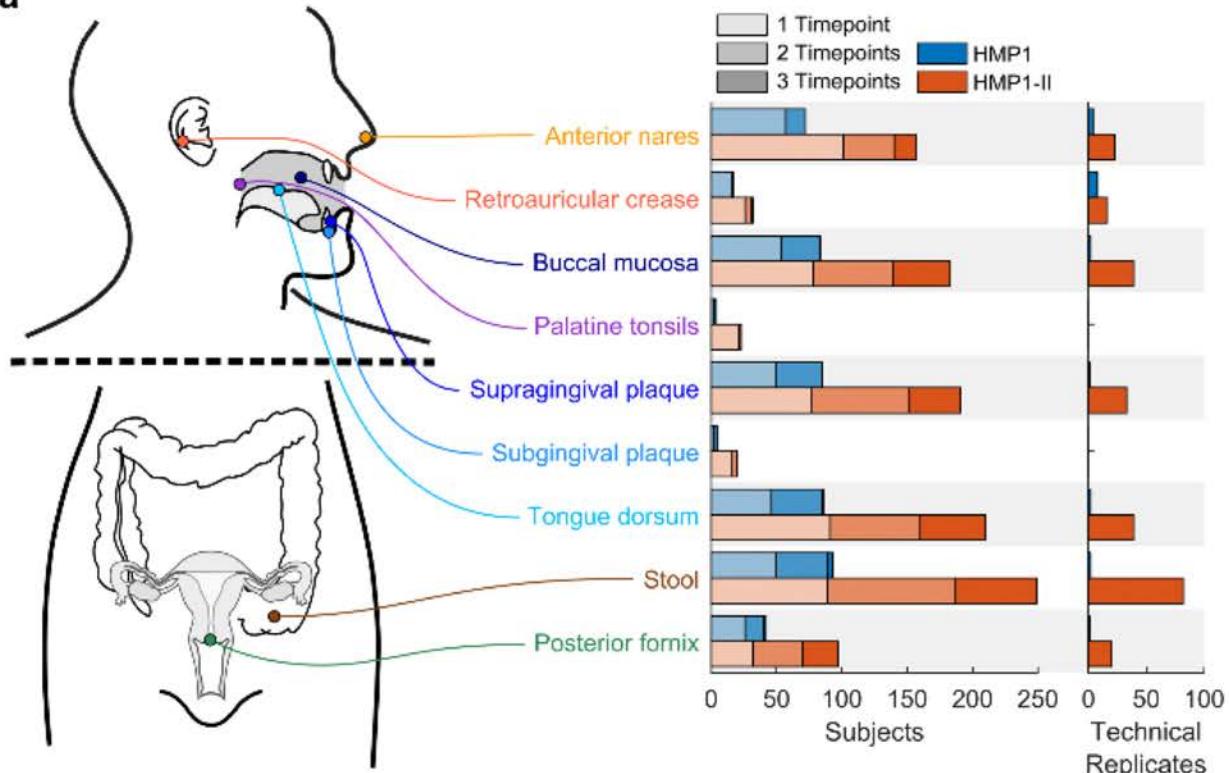
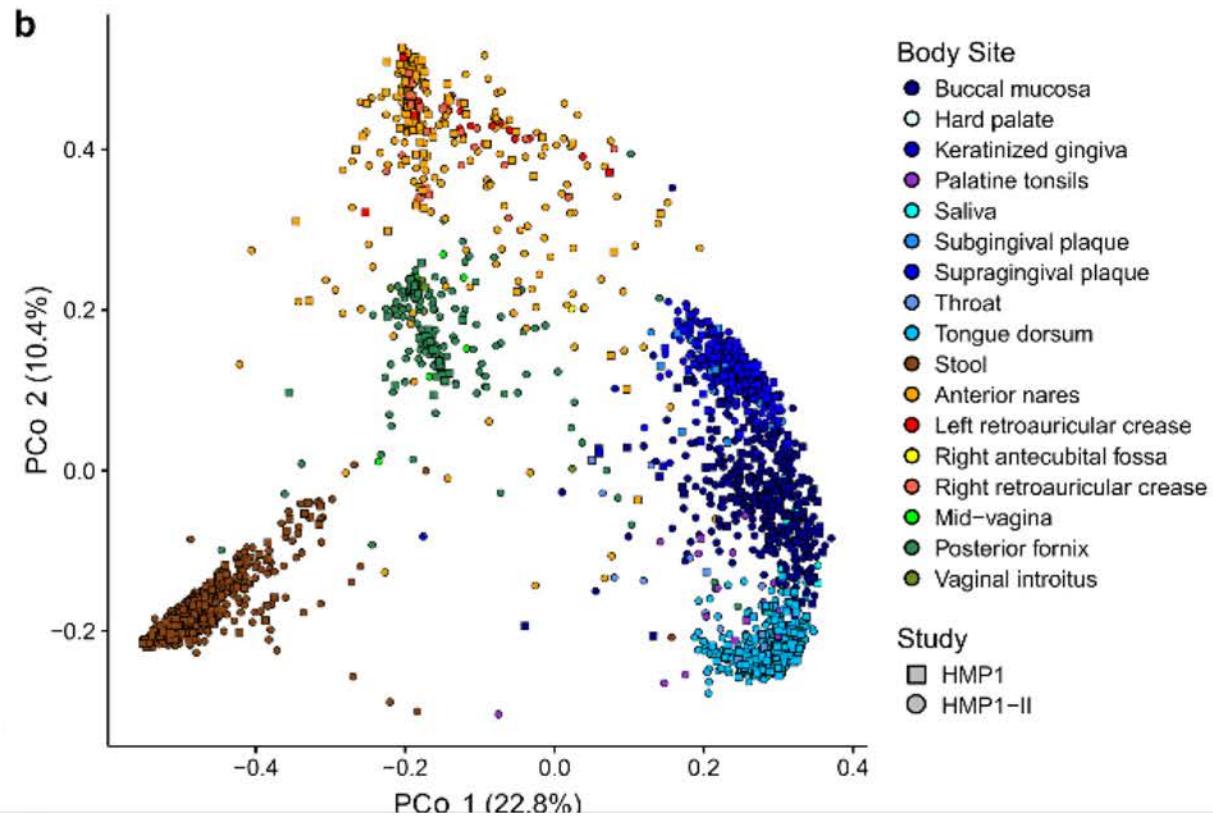
Host lifestyle affects human microbiota on daily timescales  
Genome Biol, 15 (2014), p. R89

**Yes and No**

# Strains, functions and dynamics in the expanded Human Microbiome Project

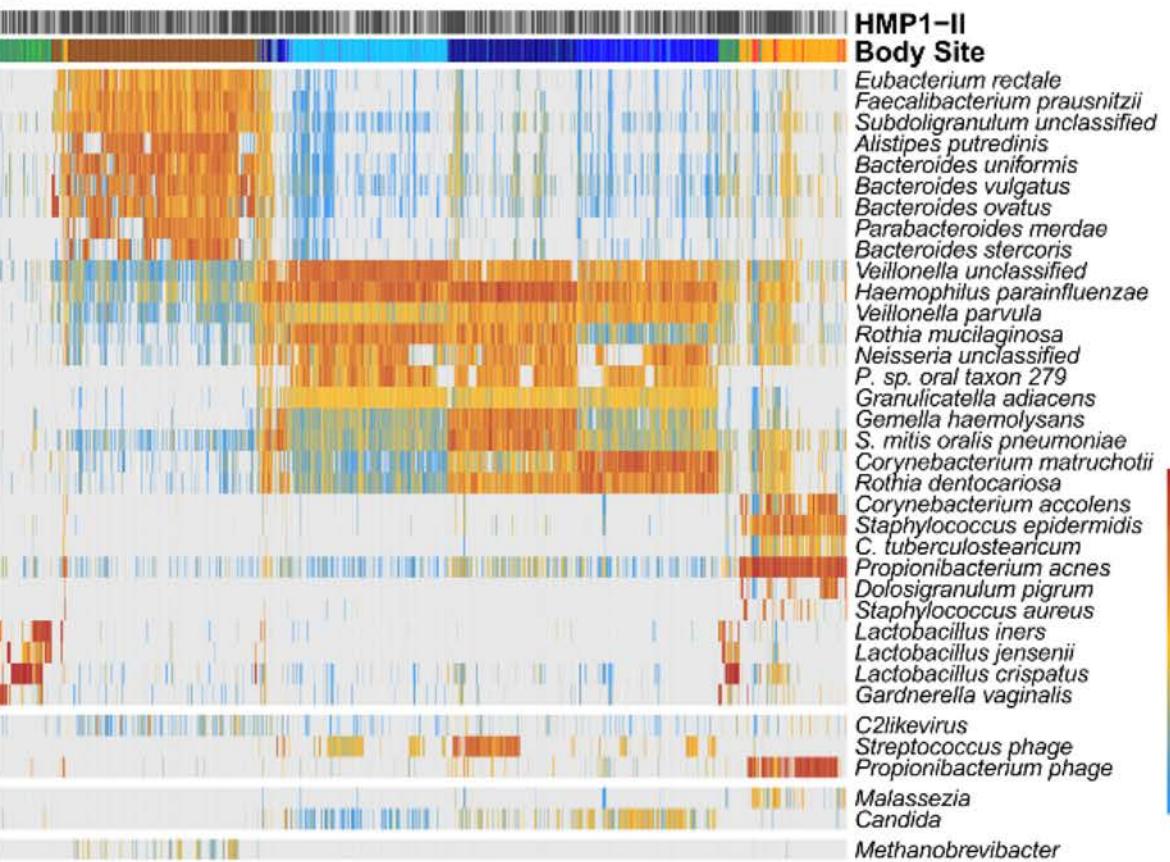
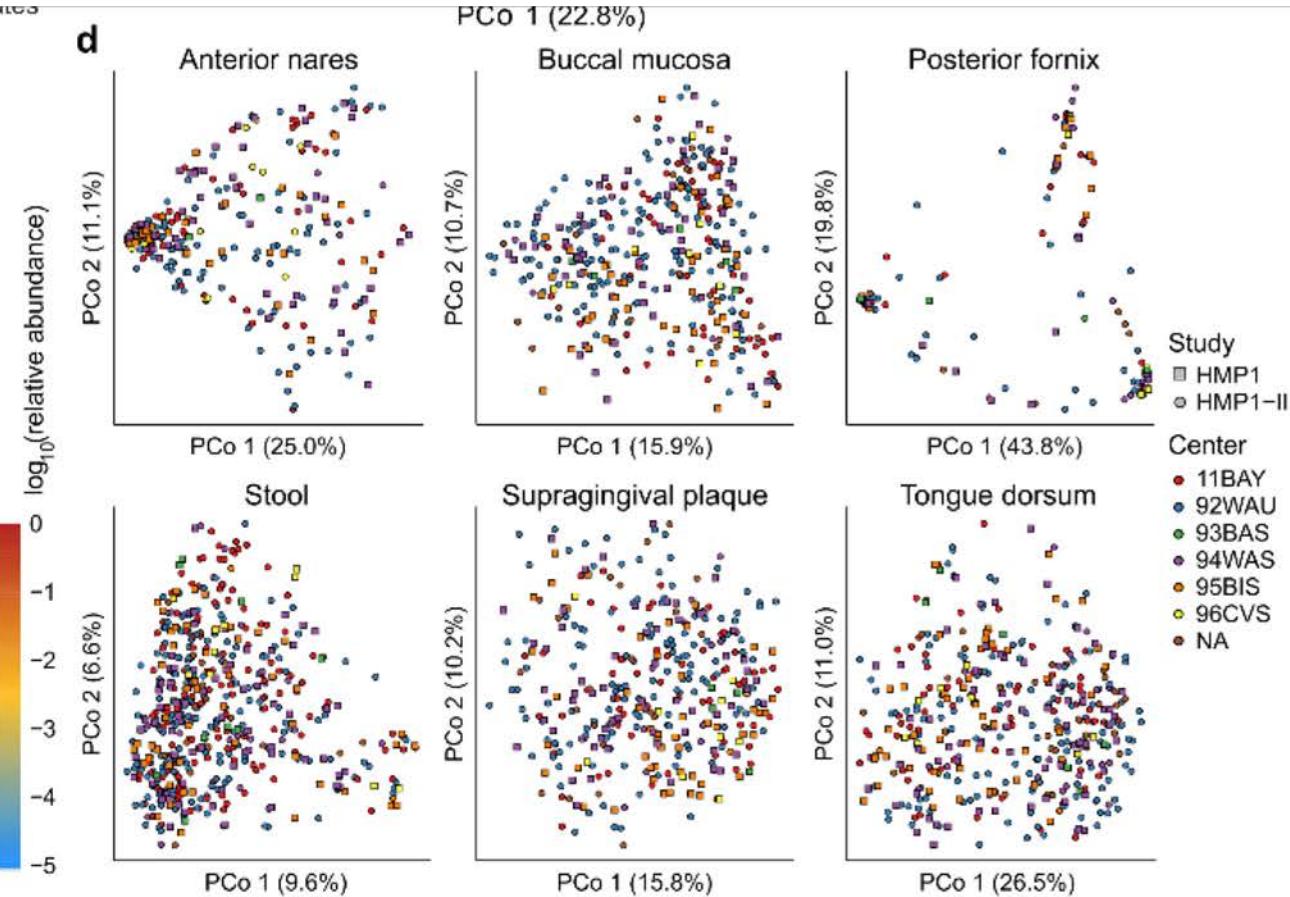
Jason Lloyd-Price<sup>1,2\*</sup>, Anup Mahurkar<sup>3\*</sup>, Gholamali Rahnavard<sup>1,2</sup>, Jonathan Crabtree<sup>3</sup>, Joshua Orvis<sup>3</sup>, A. Brantley Hall<sup>2</sup>, Arthur Brady<sup>3</sup>, Heather H. Creasy<sup>3</sup>, Carrie McCracken<sup>3</sup>, Michelle G. Giglio<sup>3</sup>,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

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.

**a****b**

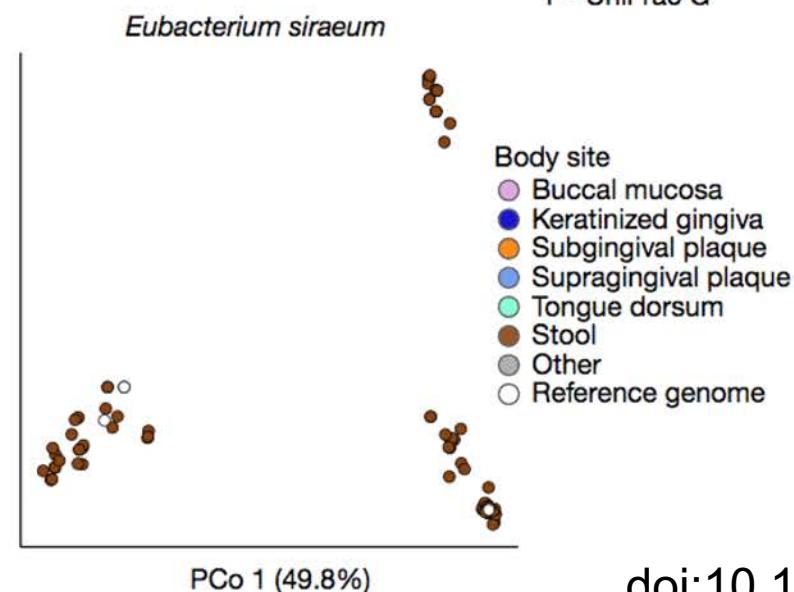
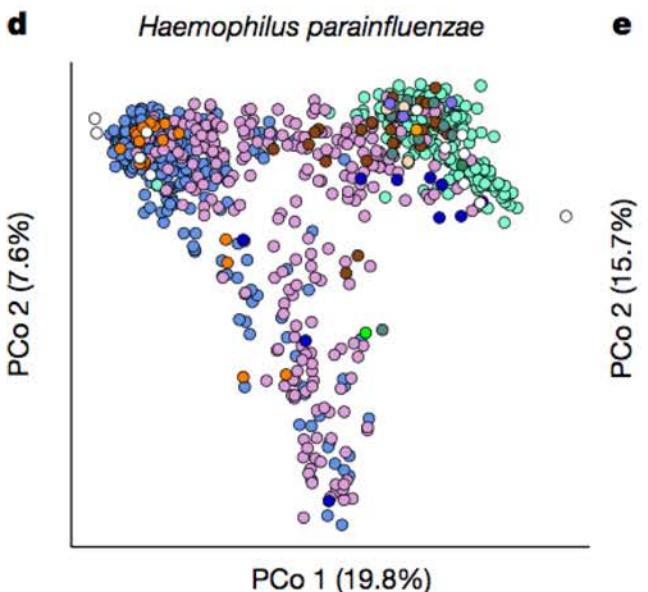
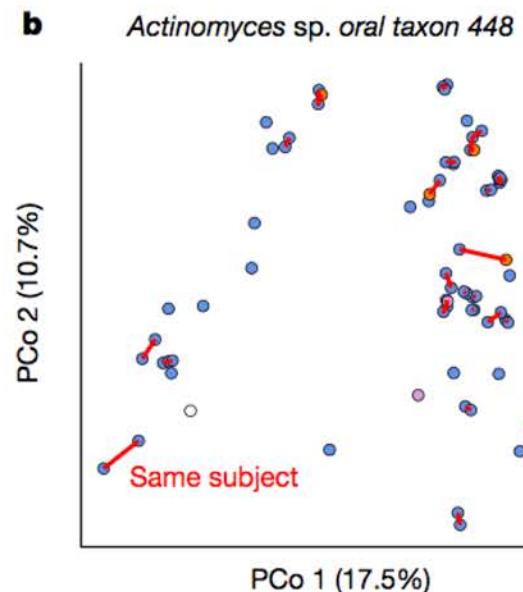
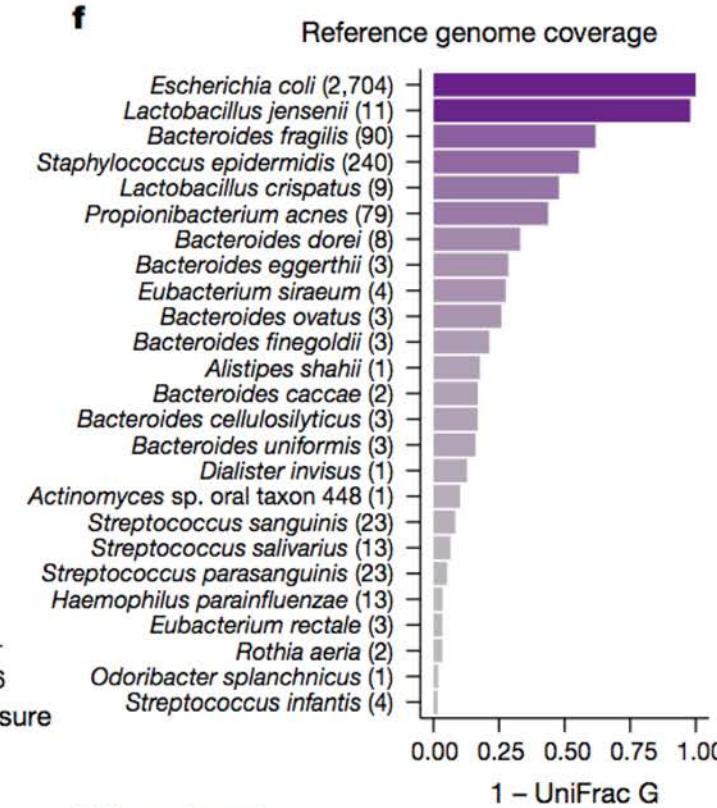
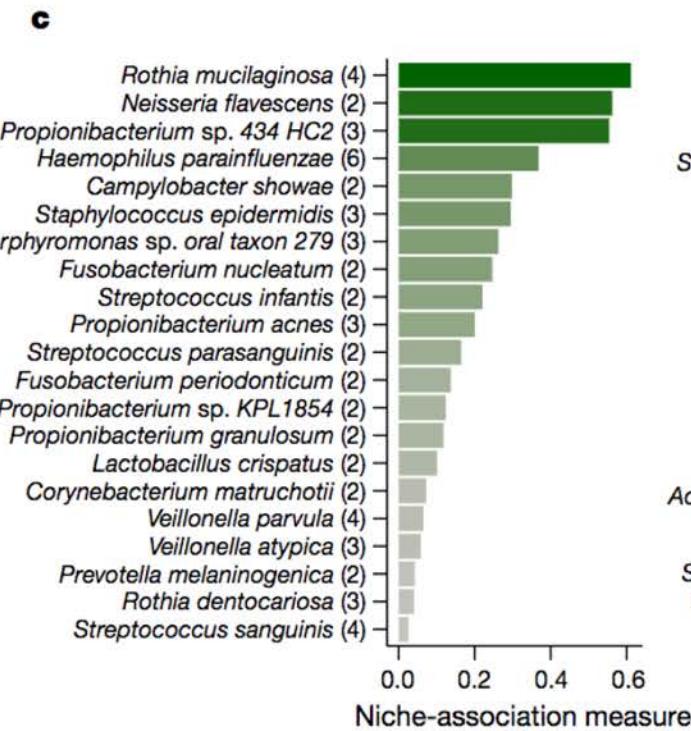
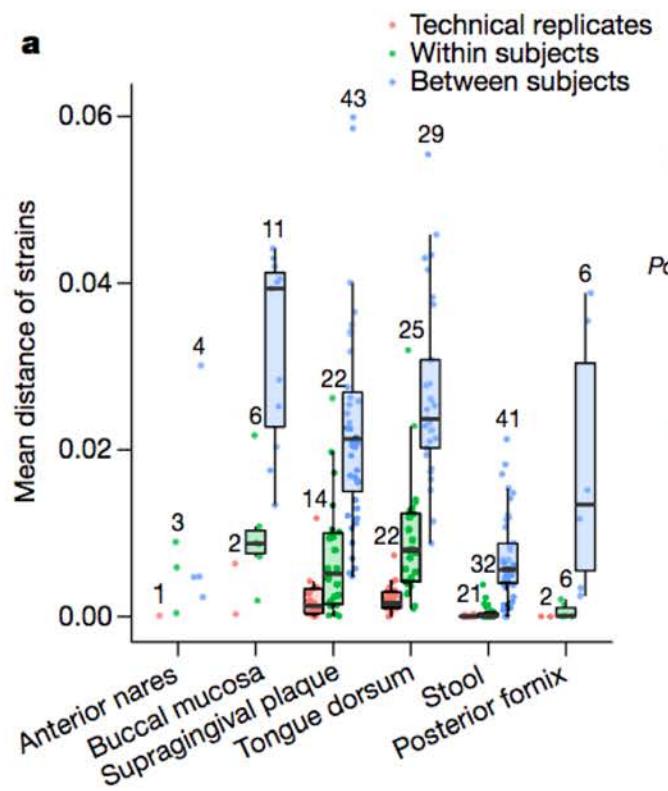
**Extended Data Figure 1 | Extended body-wide metagenomic taxonomic profiles in HMP1-II.** **a**, The combined HMP1-II datasets include a total of 2,355 metagenomes (724 previously published and 1,631 new, including 252 technical replicates). These span the project's six targeted body sites (anterior 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

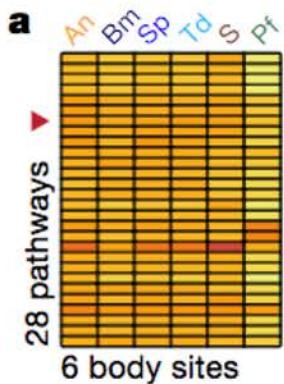
distances among all microbes at the species level. **c**, Relative abundances of the most prevalent and abundant microbes (bacterial, viral, eukaryotic, and archaeal) among all body sites, as profiled by MetaPhlAn2<sup>20</sup>. Prevalent eukaryotic microbes are shown at the genus level. **d**, Taxonomic 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<sup>1</sup>, with no divergence associated with technical variables along the first two ordination axes.

**C****d**

**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<sup>20</sup>. 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<sup>1</sup>, with no divergence associated with technical variables along the first two ordination axes.

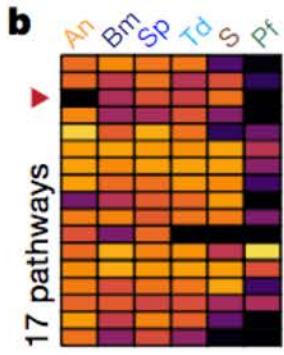




A broadly distributed 'supercore' pathway: coenzyme A biosynthesis (COA-PWY)

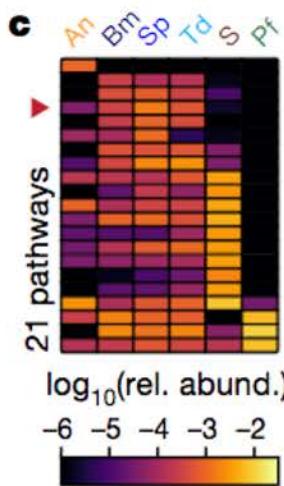
Major contributors

- Streptococcus
- Lactobacillus
- Bacteroides
- Haemophilus
- Corynebacterium
- Propionibacterium
- Prevotella
- Other
- Unclassified



A human microbiome-enriched pathway: adenosylcobalamin salvage from cobinamide (PWY-6269)

Bacteroides  
Porphyromonas  
Parabacteroides  
Streptococcus  
Prevotella  
Escherichia  
Klebsiella  
Other  
Unclassified



A body area (oral)-enriched pathway: nitrate reduction (DENITRIFICATION-PWY)

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

# Environment dominates over host genetics in shaping human gut microbiota

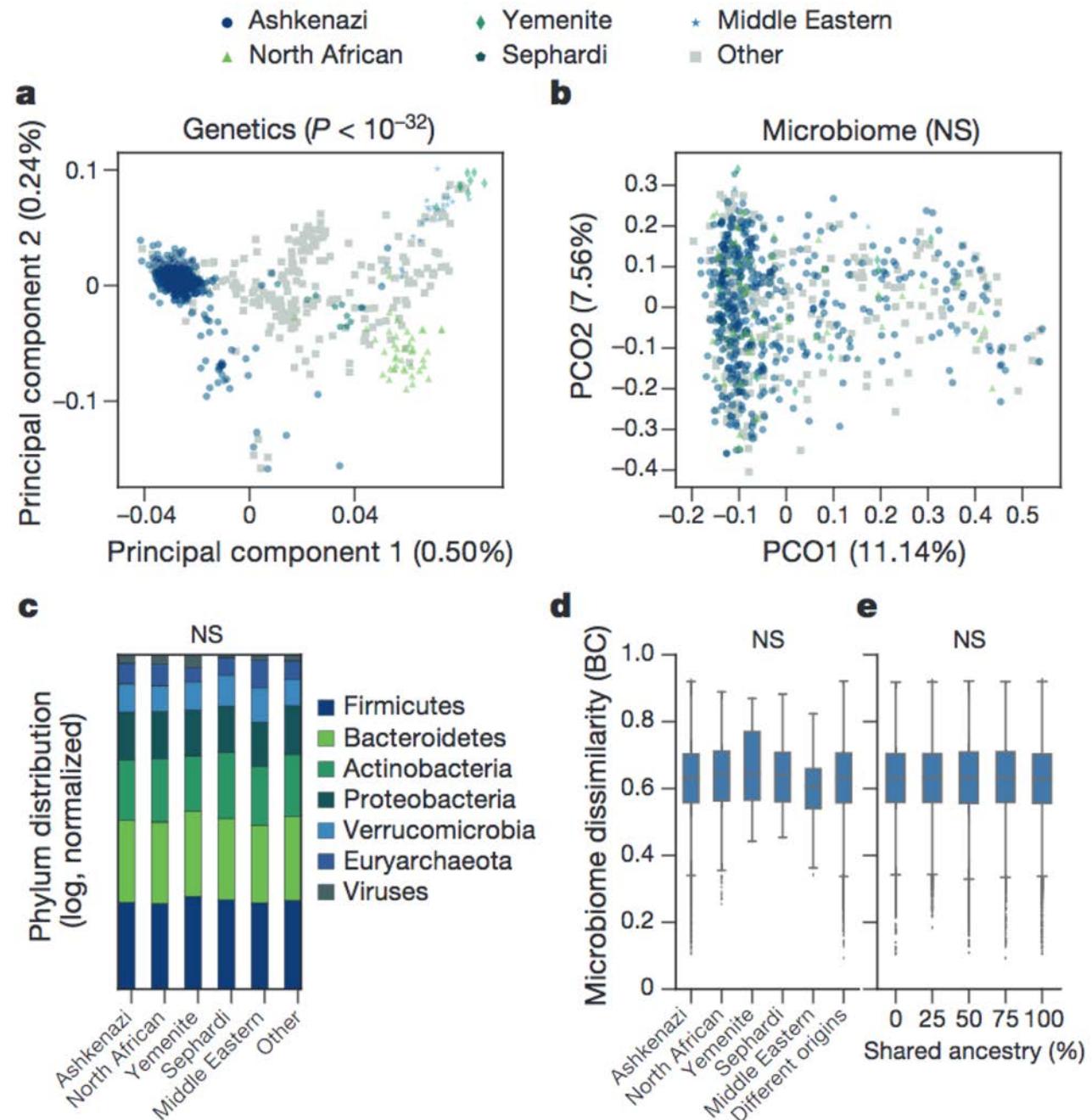
Daphna Rothschild<sup>1,2\*</sup>, Omer Weissbrod<sup>1,2\*</sup>, Elad Barkan<sup>1,2\*</sup>, Alexander Kurilshikov<sup>3</sup>, Tal Korem<sup>1,2</sup>, David Zeevi<sup>1,2</sup>, Paul I. Costea<sup>1,2</sup>, Anastasia Godneva<sup>1,2</sup>, Iris N. Kalka<sup>1,2</sup>, Noam Bar<sup>1,2</sup>, Smadar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

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.

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

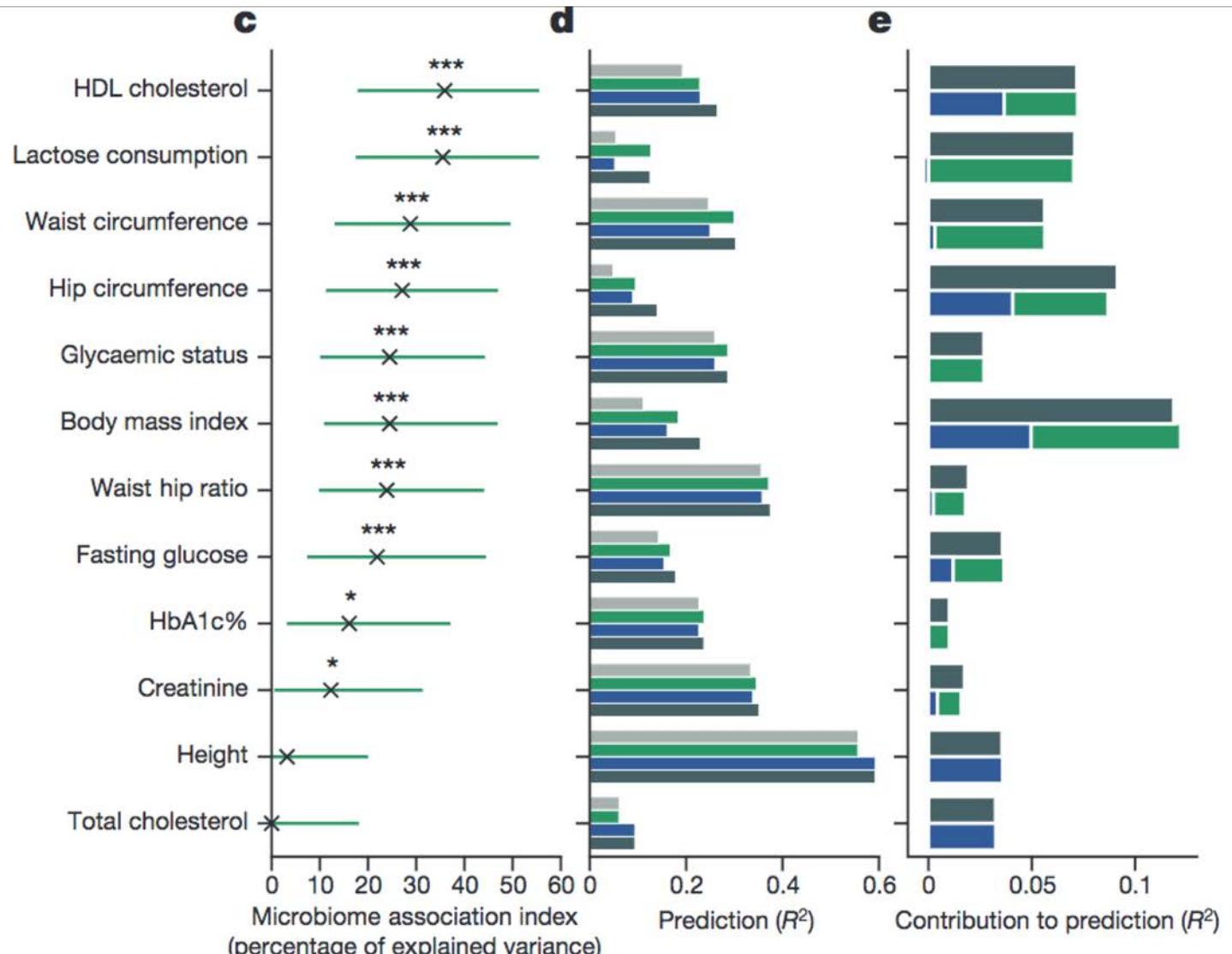
**Figure 1 | Genetic ancestry is not significantly associated with microbiome composition.** **a**, Genetic principal components are strongly associated with self-reported ancestry, with Ashkenazi ( $n = 345$ ), North African ( $n = 42$ ), Middle Eastern ( $n = 24$ ), Sephardi ( $n = 10$ ), Yemenite ( $n = 8$ ) 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).



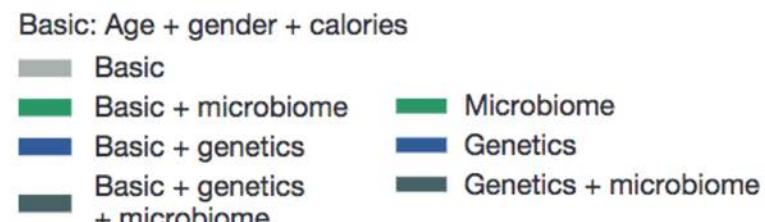
**b**

Phenotype	Microbiome association index		Genetic heritability (literature)
	Israeli cohort	LLD cohort	
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<sup>27–34</sup> (right column). \*FDR < 0.05, \*\*FDR < 0.01 and \*\*\*FDR < 0.001. Cons., consumption, circ., circumference. **c**,  $b^2$  estimates



\*\*\*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.



## 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.

# ARTICLE

OPEN

doi:10.1038/nature24621

## A communal catalogue reveals Earth's multiscale microbial diversity

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



## BY THE NUMBERS

27,751

samples

7 continents

43 countries

2,212,796,183

total DNA sequences

307,572

unique DNA sequences  
(approx. species)

50+

peer-reviewed  
publications

500+

scientists



92 environmental  
features

2 – 12 pH range

66 animal host  
species

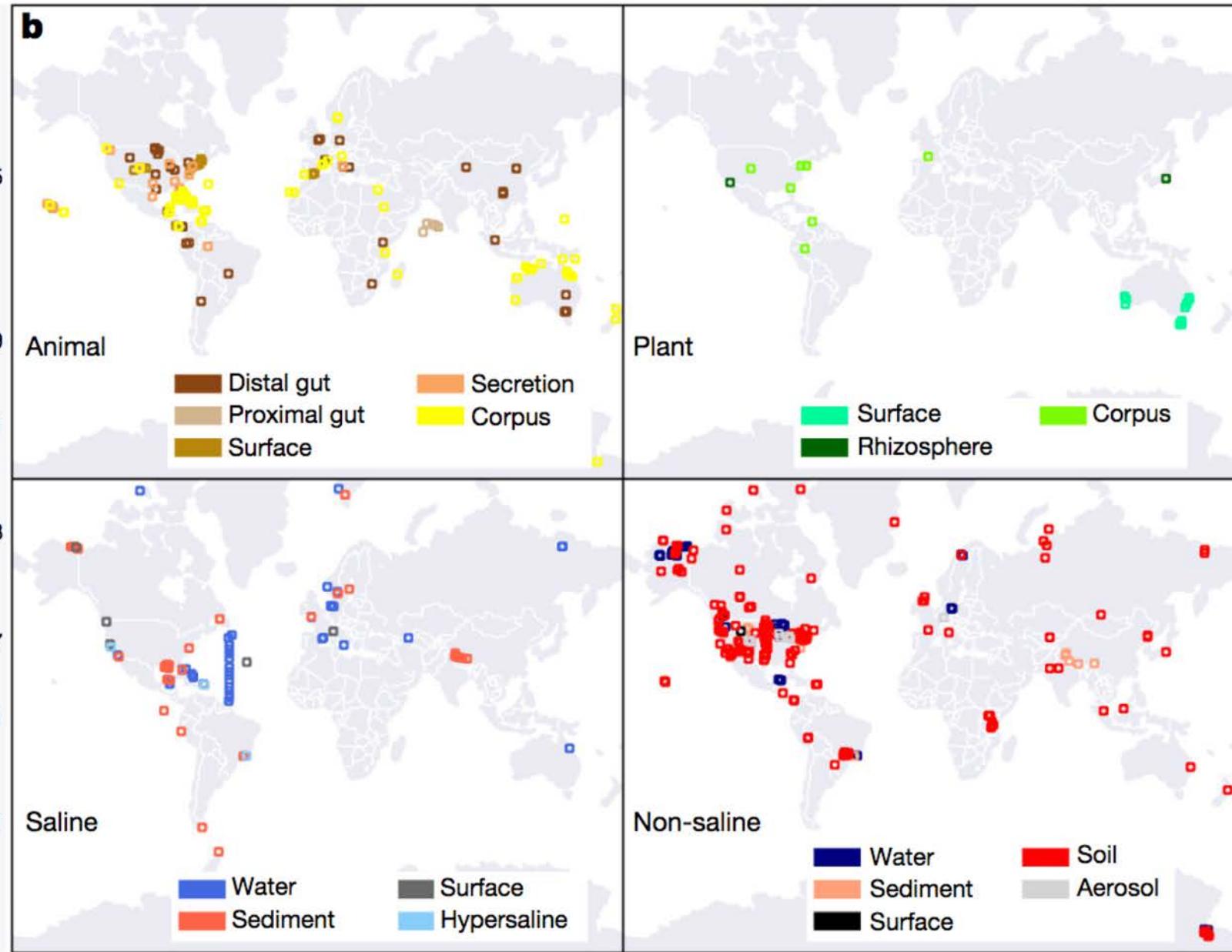
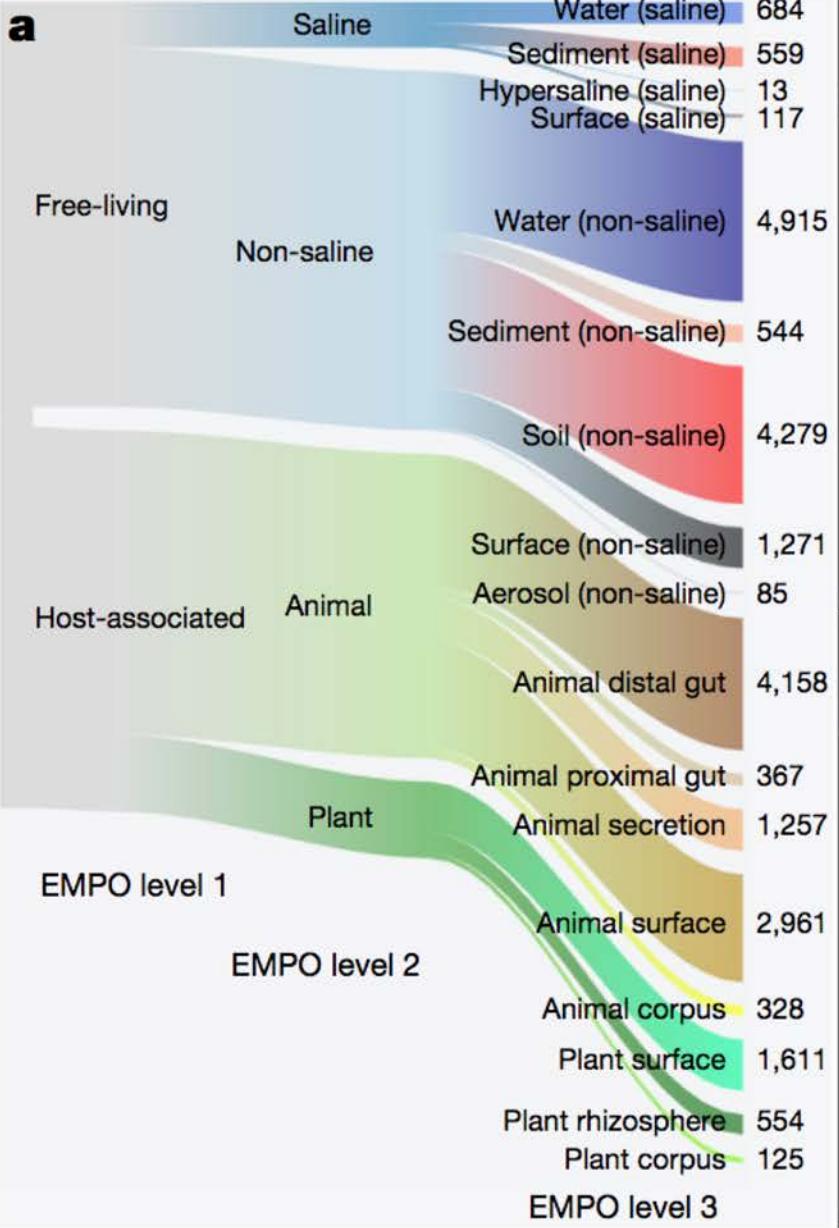
(stomach acid to household  
ammonia)

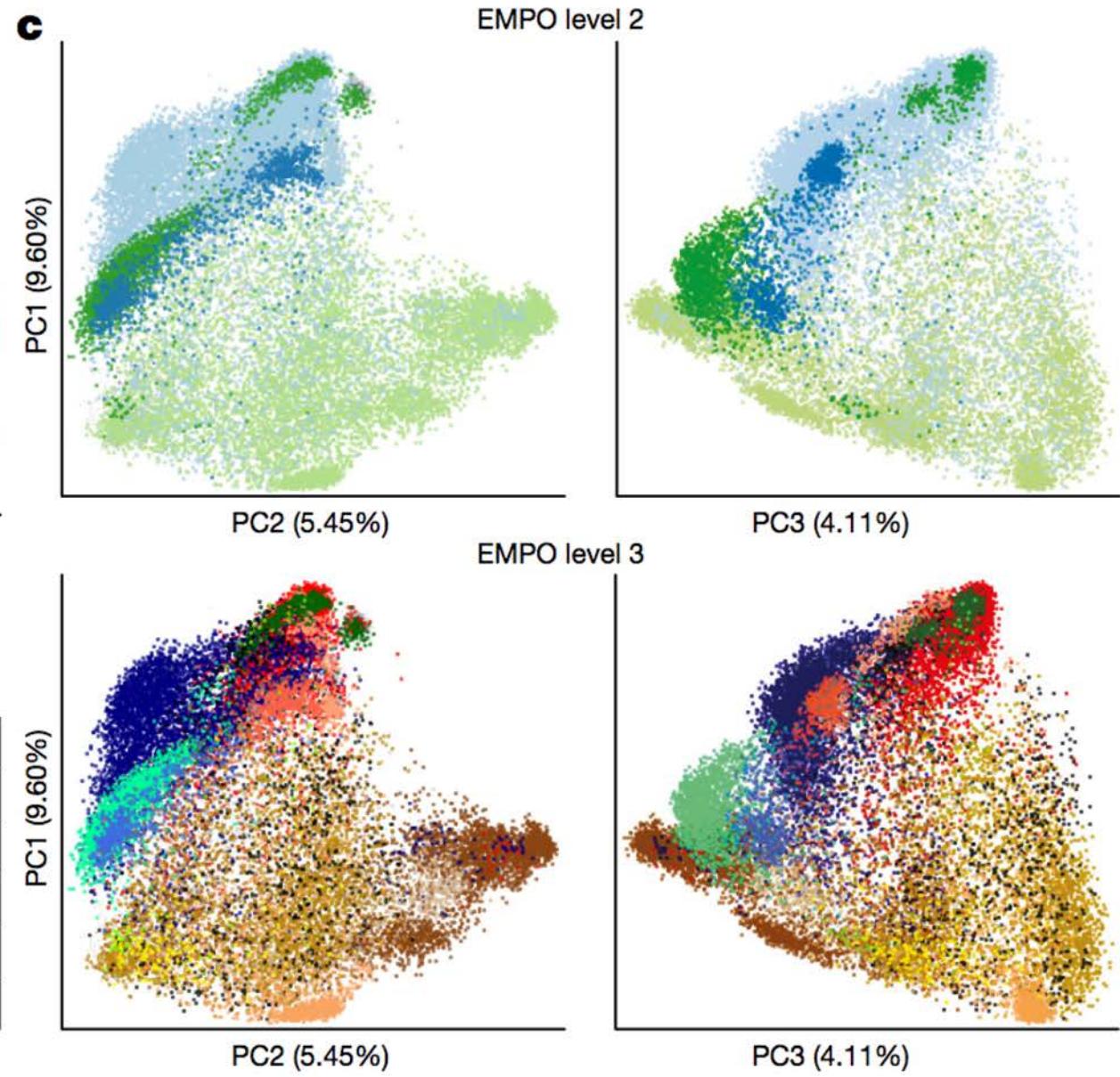
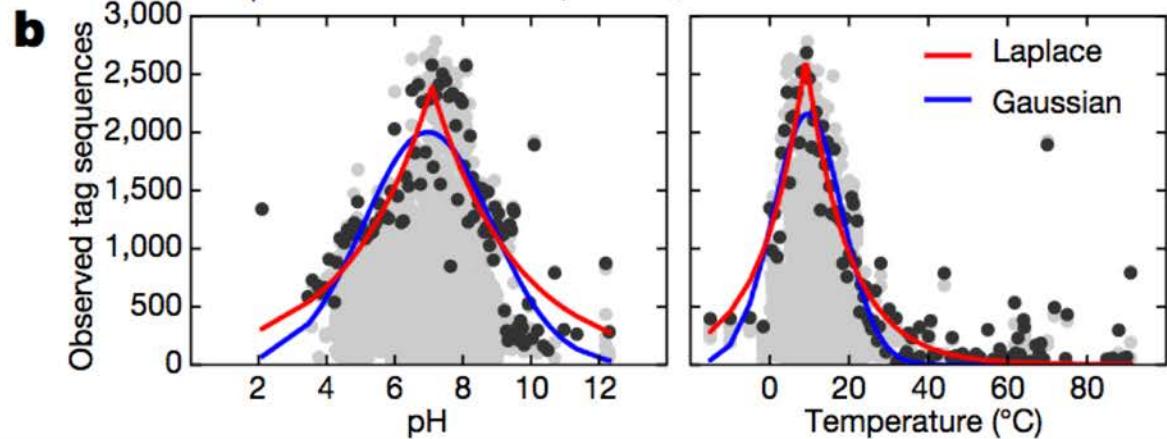
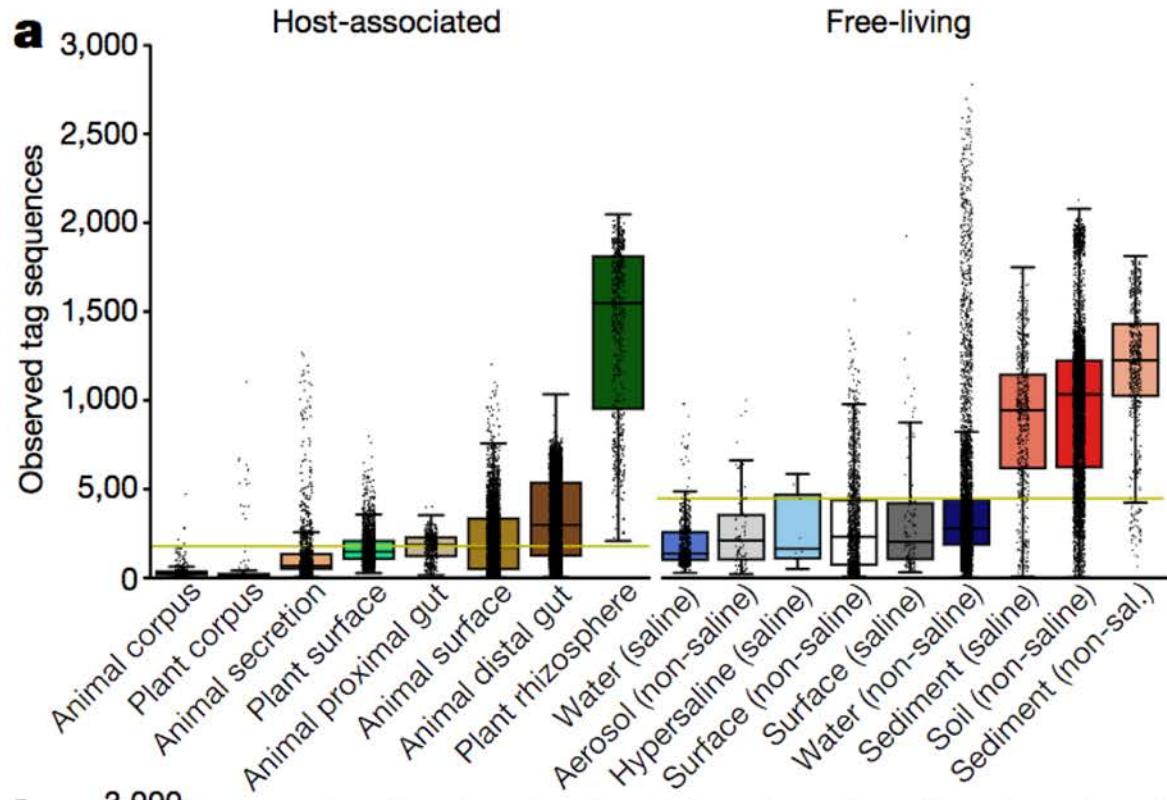
78.9 °N –  
78.2 °S

latitude range (Arctic  
Circle to Antarctica)

1 reference database of  
bacteria that reside on  
Planet Earth

Credit: UC San Diego Center for Microbiome Innovation  
Reference: Thompson et al. *Nature*. 2017. doi:10.1038/nature24621





# Additional references

# A good introductory popular science video about microbiome

<http://www.youtube.com/watch?v=5DTrENdWvvM>

# A clinician's guide to microbiome analysis

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

# TED talk

[https://www.ted.com/talks/rob knight how our microbes make us who we are](https://www.ted.com/talks/rob_knight_how_our_microbes_make_us_who_we_are)

#Coursera

<https://zh-tw.coursera.org/learn/microbiome>