

NGS-Analytics

Isheng Jason Tsai

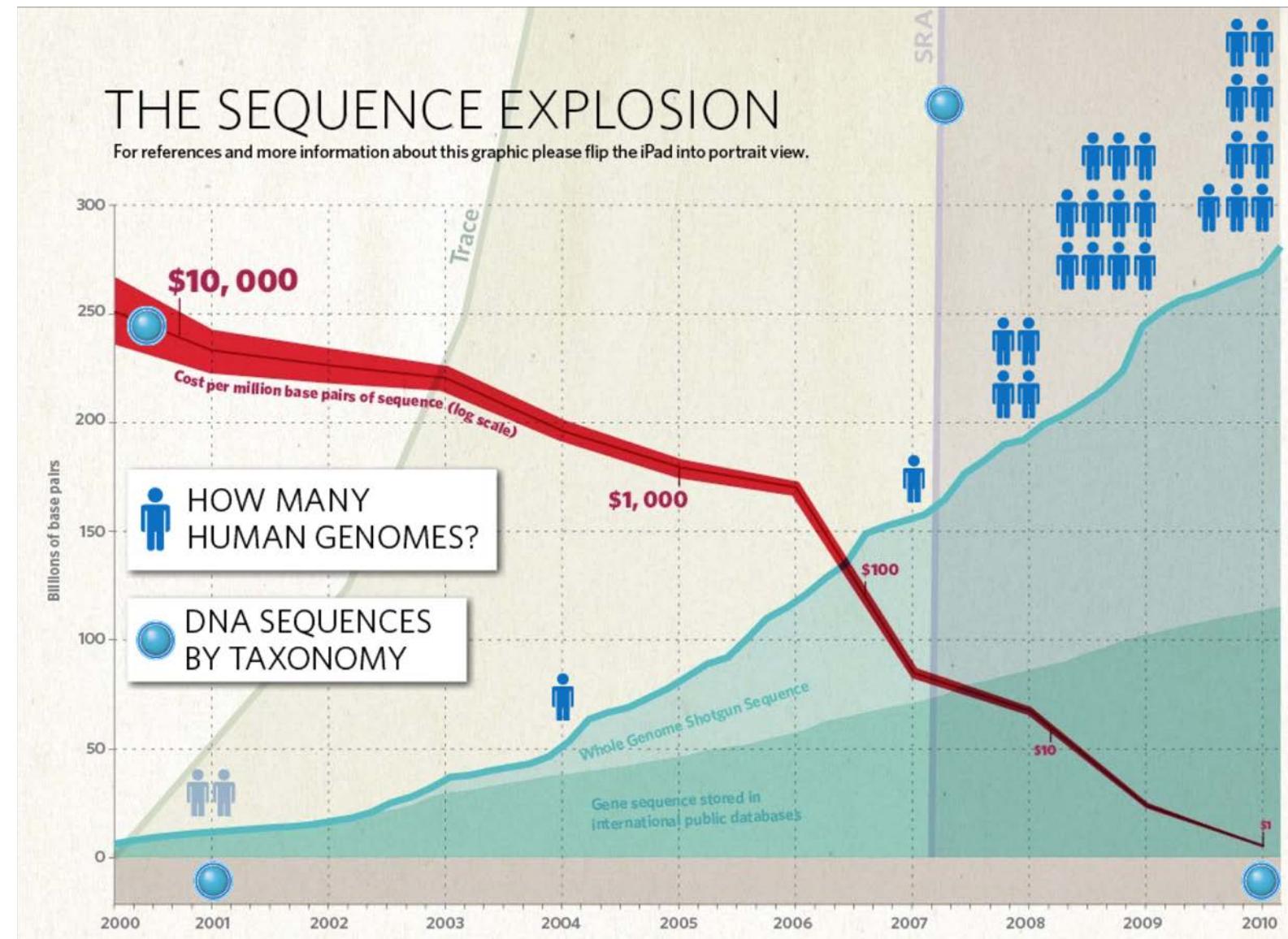
Lecture modified for B2



Welcome!

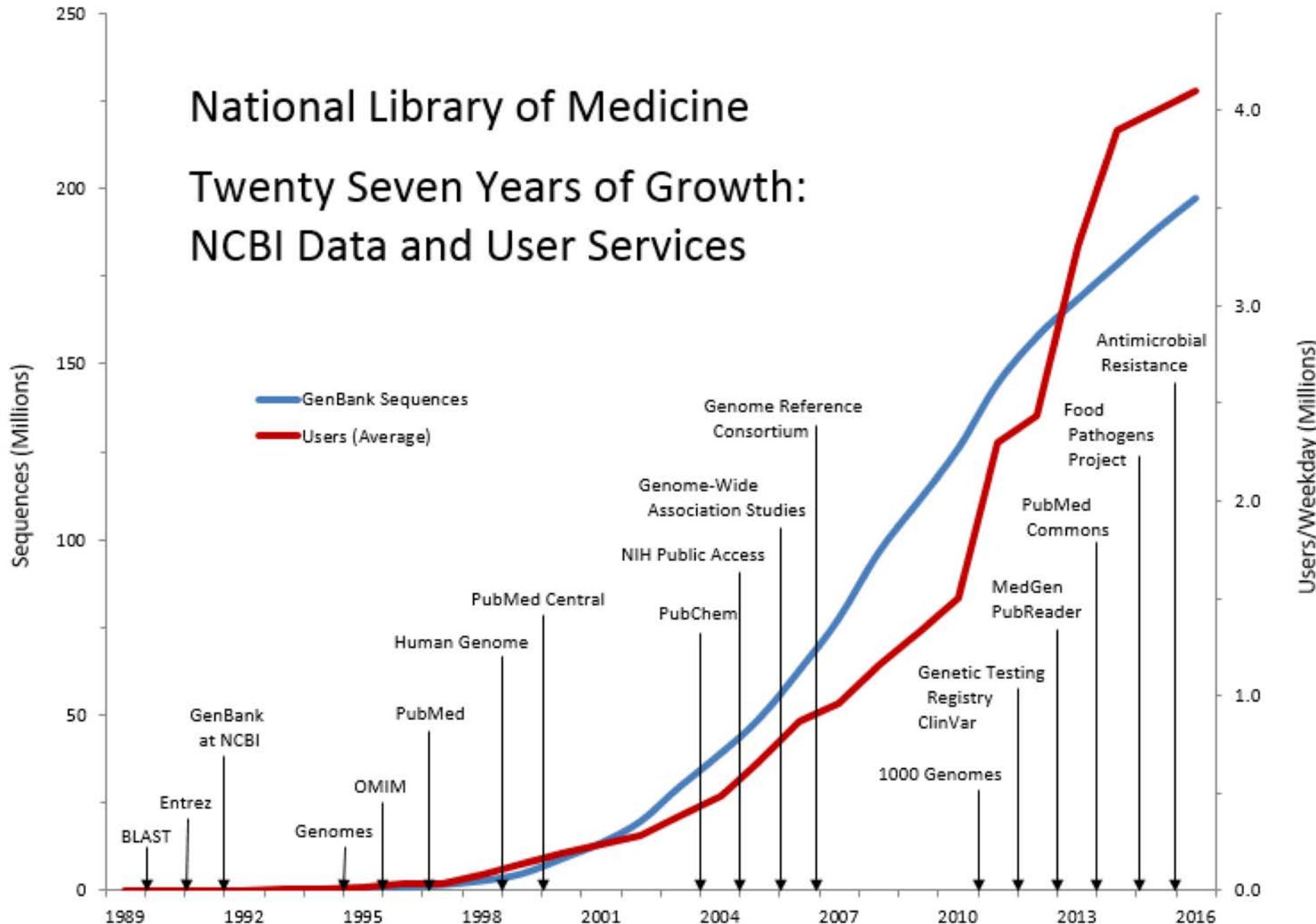
What is NGS?

- = 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



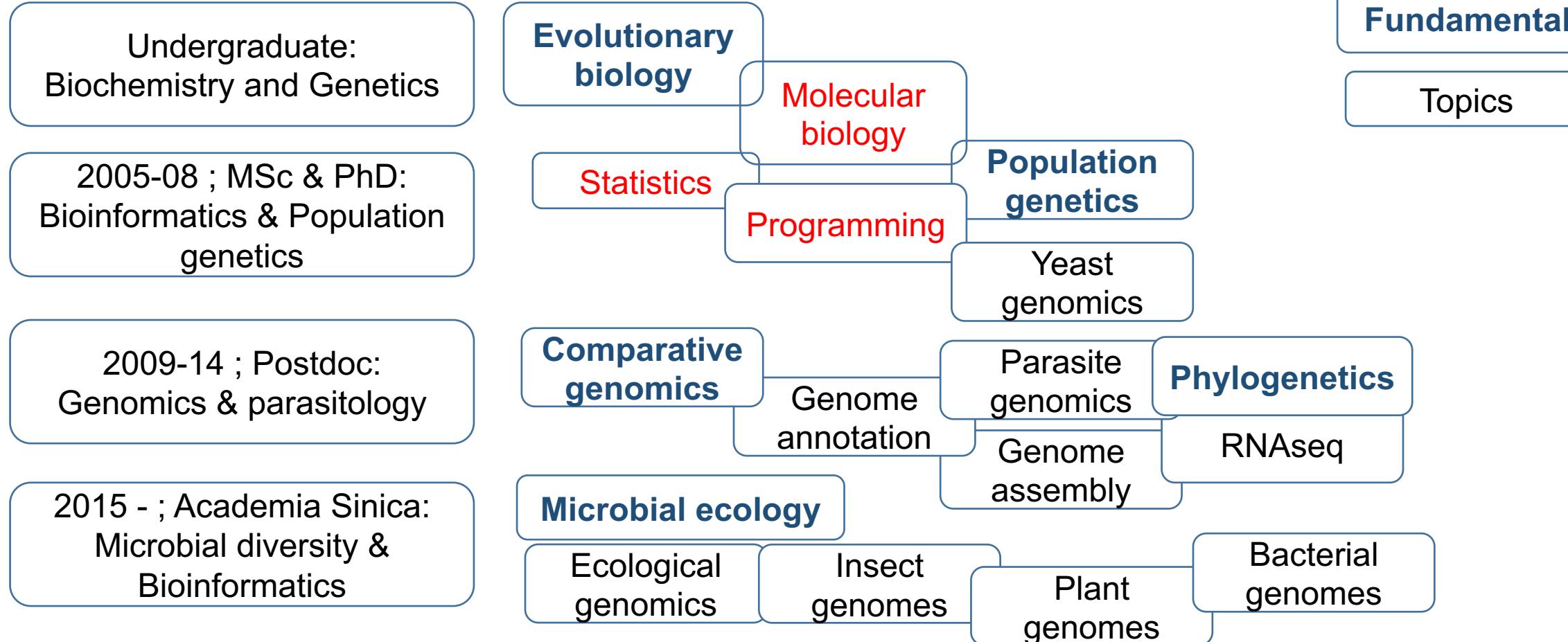
NGS = sequencing made cheaper, faster and higher throughput

This lecture is called “**NGS-Analytics**”

Actually

- NGS= 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?)

My background



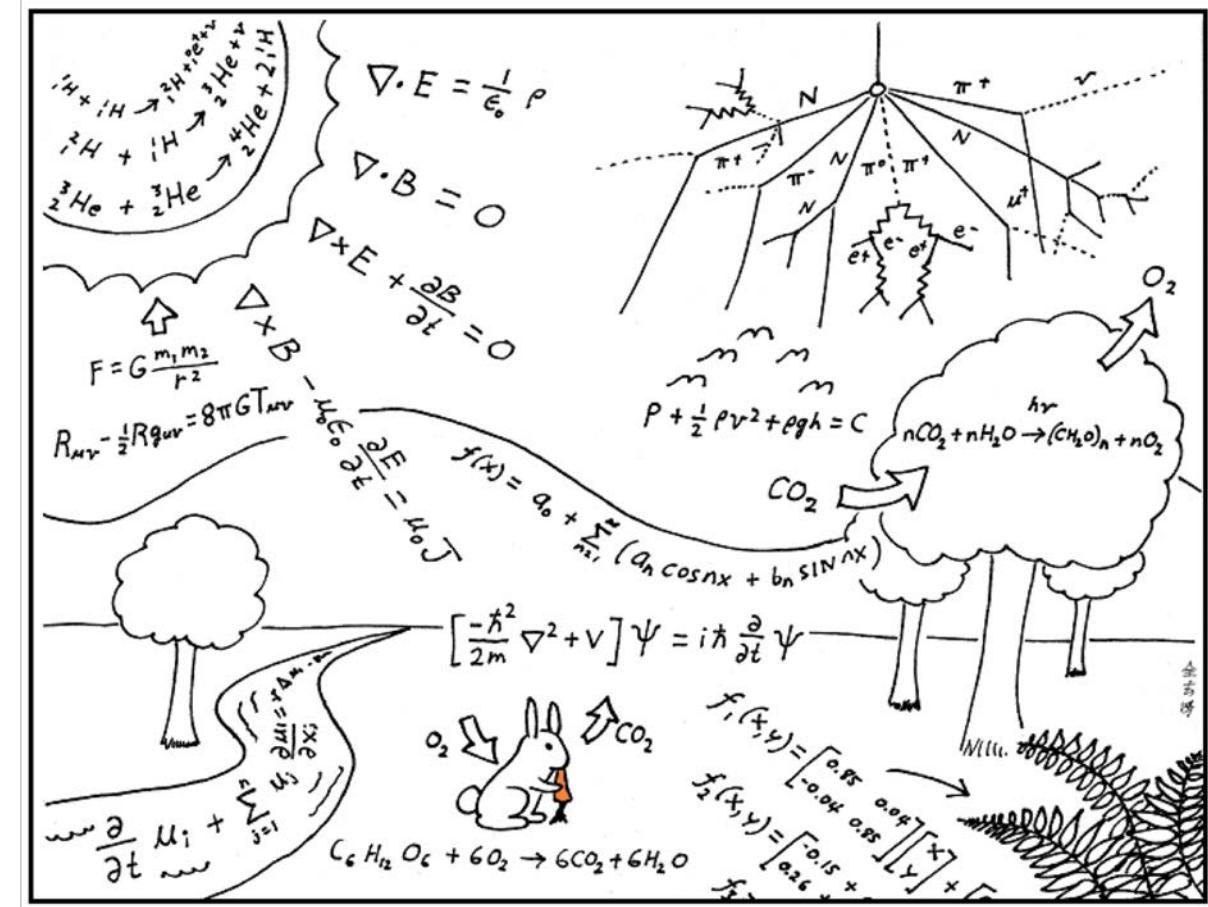
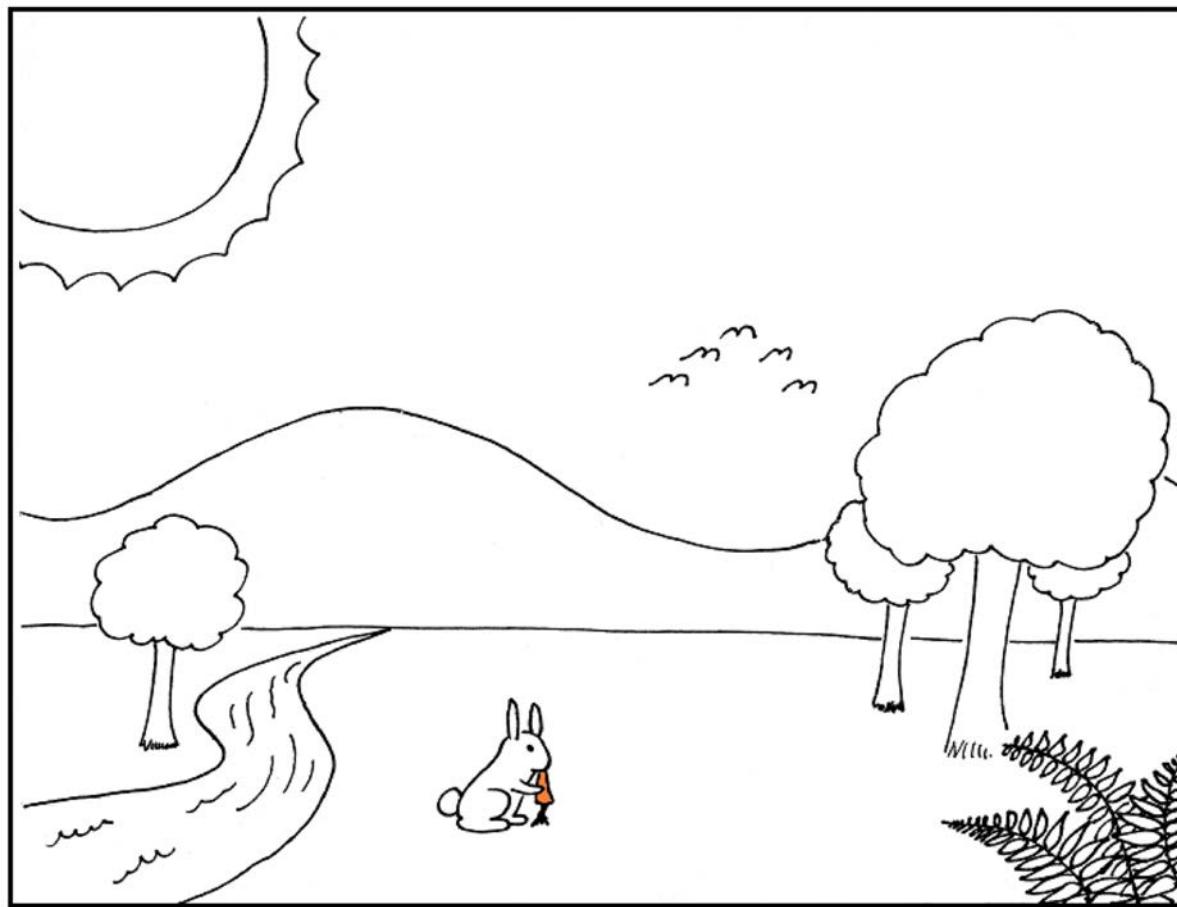
- **39 publications**
(2 Nature, 1 Science, 2 Nature Genetics, 2 PNAS,
3 Genome Biology, 1 Molecular Ecology)

TIGP Introduction to NGS

1. Introductory lecture [!\[\]\(cd3e54d951a9fb854f48e4697cf550f9_img.jpg\) Download](#)
2. Linux and R ; basic usage [!\[\]\(cc729e263f29c0a76fbdc4cfe67fceb0_img.jpg\) Download](#)
3. Genome Assembly and case studies [!\[\]\(90d36d418f8f7ab67431ba2525e00a5e_img.jpg\) Download](#)
4. Mapping and Case studies [!\[\]\(f70e40faeec369ff477dbaef549ee05b_img.jpg\) Download](#)
5. From Alignment to phylogeny (Jiang Ming Chang) [!\[\]\(ca68c0c79a5dc0026aa1d011fda2b676_img.jpg\) Download](#)
6. DNA/RNA preparation and different sequencing technologies (Meiyeh Lu) [!\[\]\(caba7331972dceb944f99aa56fee2f81_img.jpg\) Download](#)
7. RNAseq and Genome annotation [!\[\]\(c8a030d79816aa5f757cd6099c7d9a8e_img.jpg\) Download](#)
8. Comparative Genomics [!\[\]\(4a09161e9d0b0aaec8ec1149656d0974_img.jpg\) Download](#)
9. Population Genomics (John Wang)
10. Amplicon / Metagenomics [!\[\]\(9223deec401f24808aa5e7273d7c177a_img.jpg\) Download](#)
11. Practical one: Linux and R
12. Practical two: RNAseq mapping and EdgeR
13. Discussion
14. Final Report

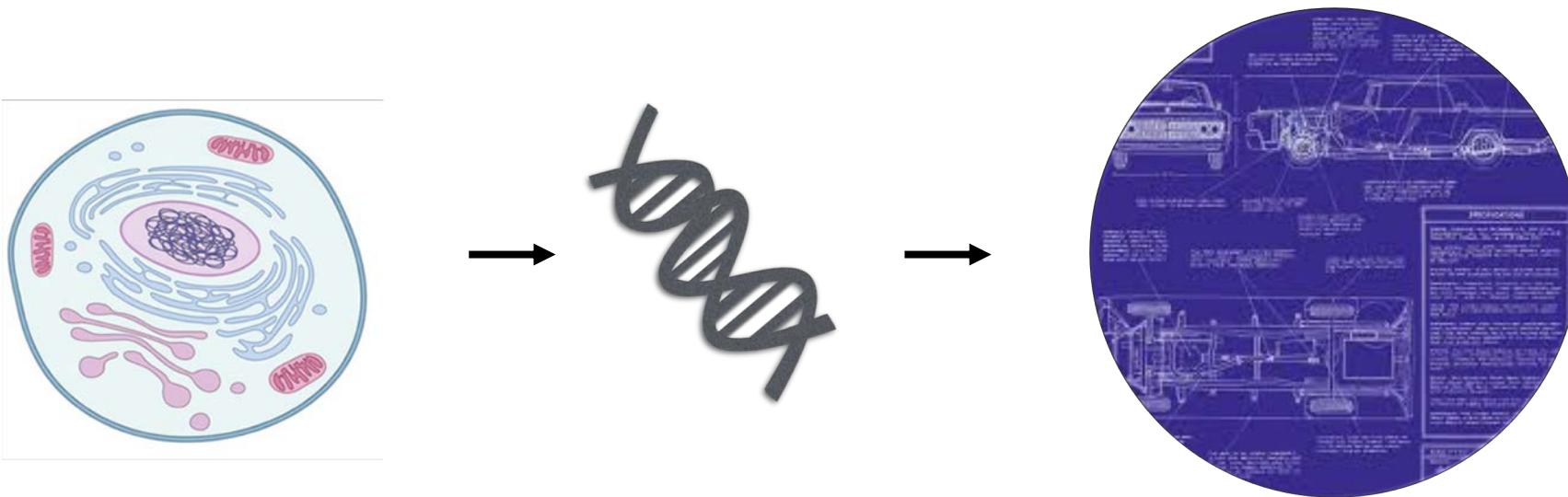
Always start with a question

This is how scientists see the world



How? Who? Where? What?

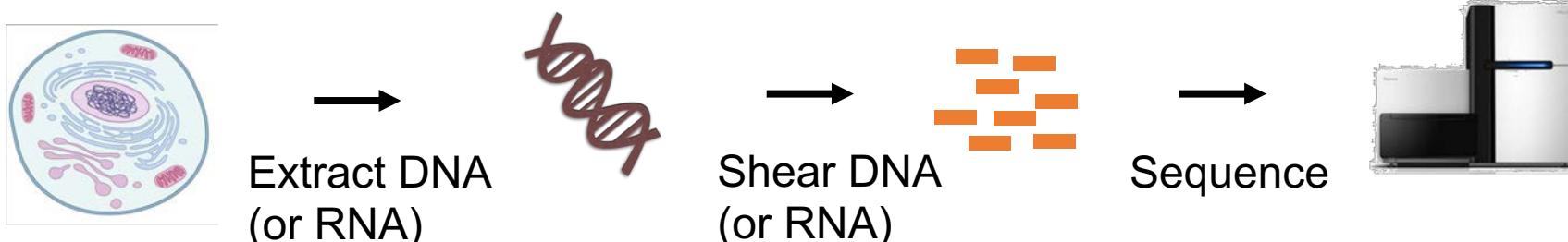
Genome



Genome = Parts list of a single genome

A genome project

Wet lab work



Bioinformatics

Data QC

Variant calling

ATCG
AT~~G~~G
ATCG

Annotation



DNA or RNA Reads
50-500 bp

Mapping
RNaseq

Reads
50-500 bp

Assembly

Contigs
1kb – 100 kbp

Scaffolding

Scaffolds
Hopefully Mbp

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



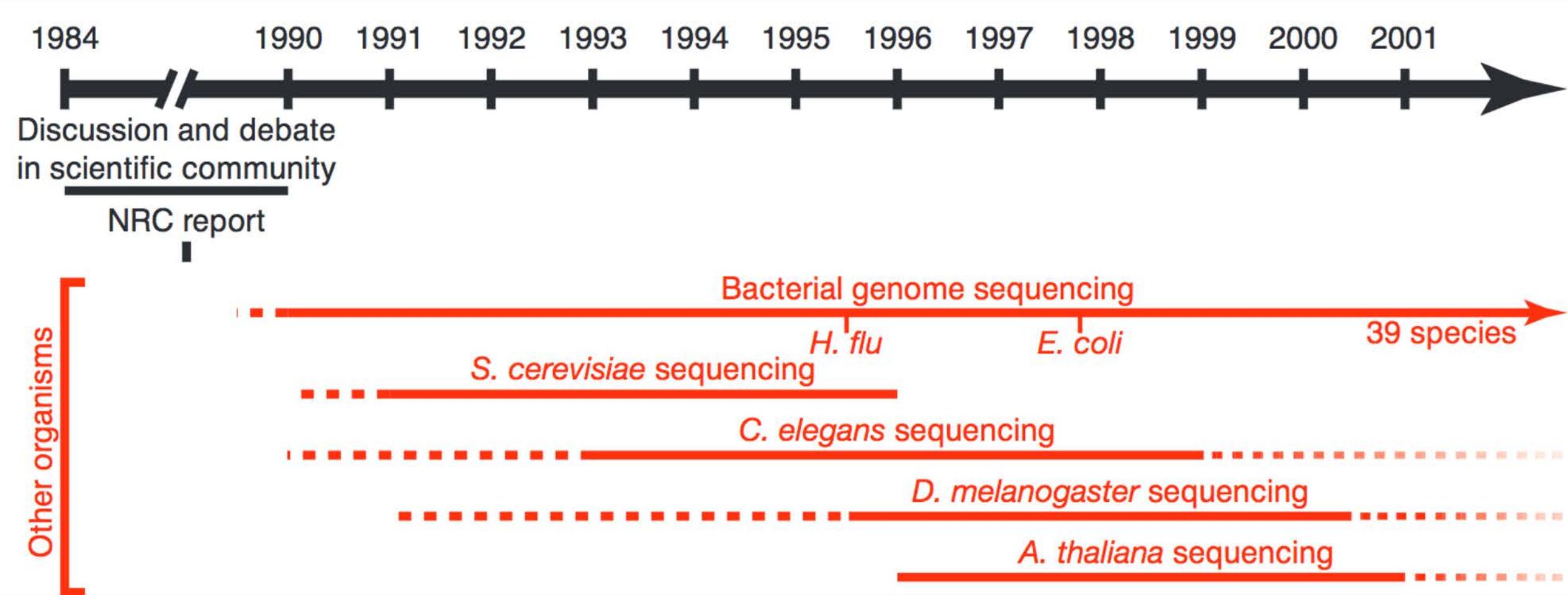
What we
think we do

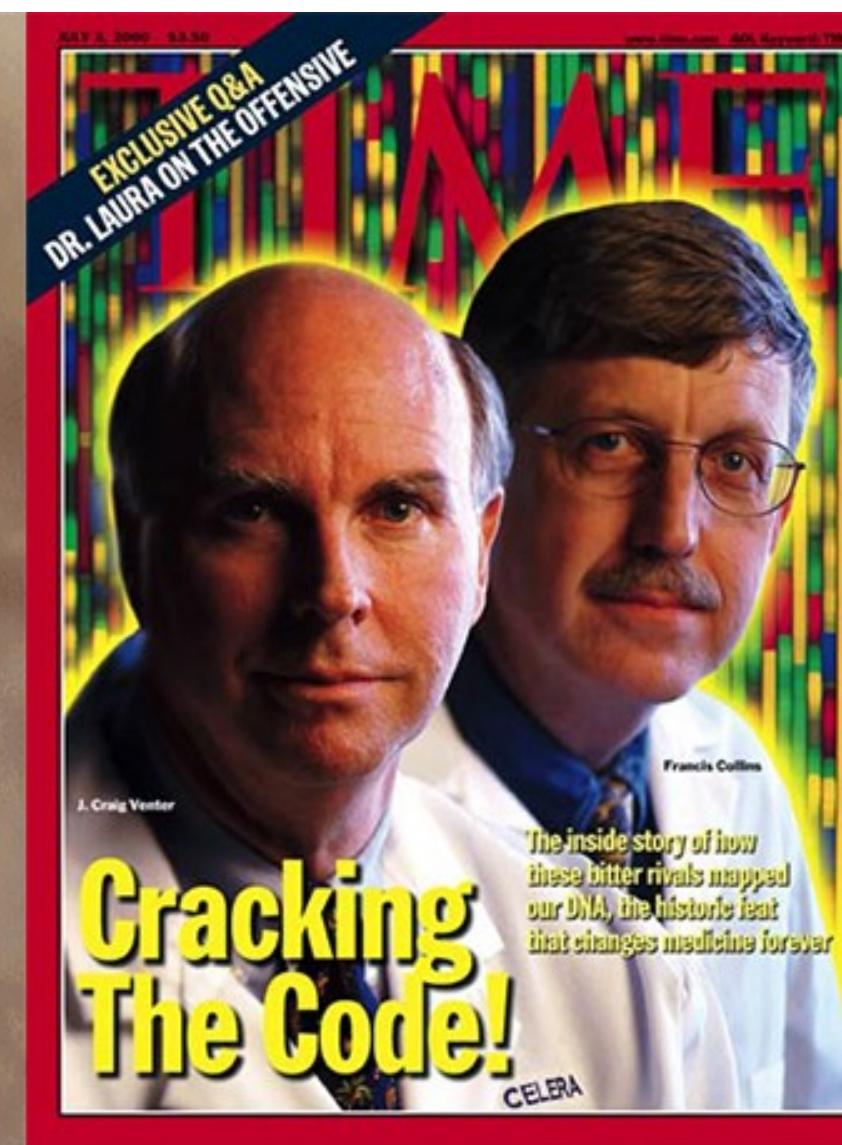


What we
actually do

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**
 - ~~Some lab just had the money ; don't do it~~





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.

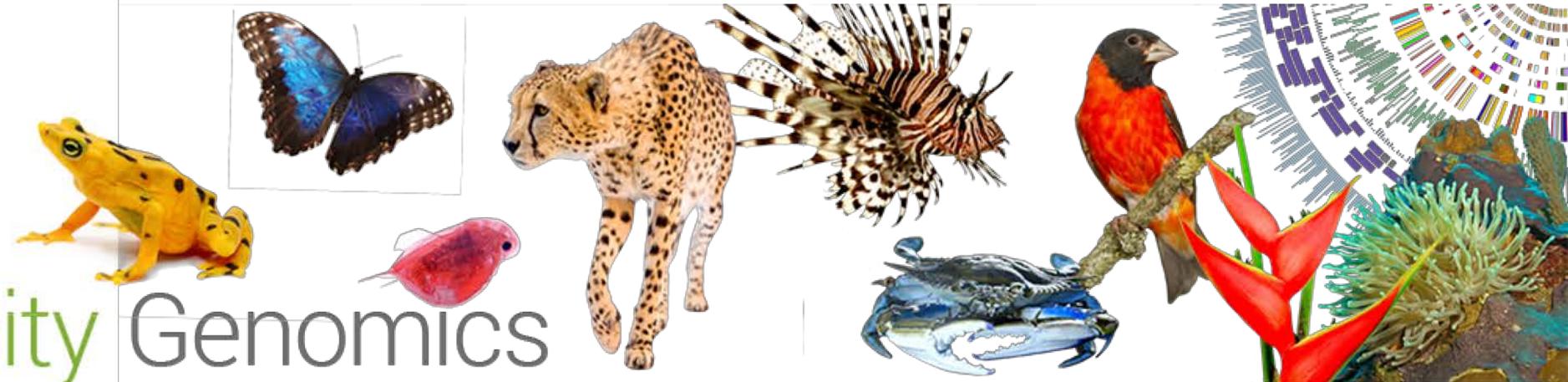
The project received a £10.5 million funding award from Wellcome in March 2010 and sequencing started in late 2010. For more information, please use the links on the right hand side.



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.

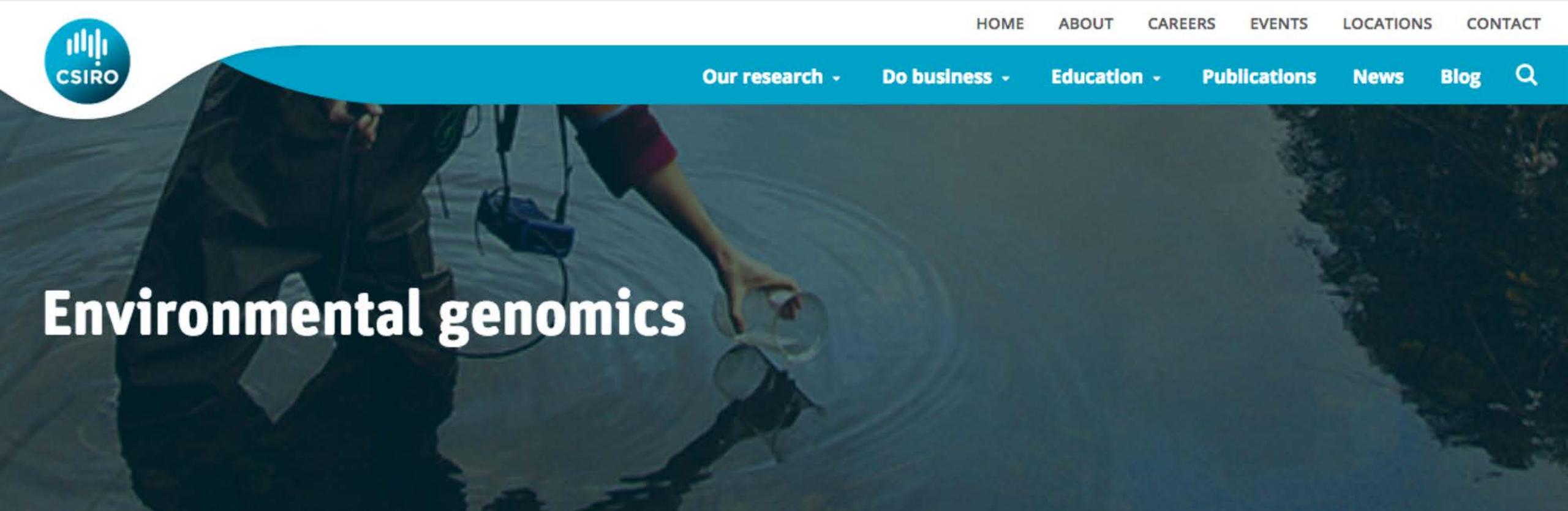


CENTER FOR CONSERVATION GENOMICS

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 up, wearing a dark wetsuit and a red shirt underneath. They are holding a clear plastic container with both hands, positioned over a body of water. The background is slightly blurred, showing more of the water and some foliage on the right side.

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)

Problem

Most people doing genomics not actually doing genomics

Posted on July 27, 2015 by jovialscientist

CAMBRIDGE. Most people who claim to be genomics researchers are not actually doing genomics at all, and instead are just sequencing things and calling it genomics, it has been found.

“Genomics is the study of genomes” said Barney Ewingsworth III from the Excellent Biology Institute (EBI) “and genomes are incredibly complex, with repeat regions, duplications, deletions, selective sweeps, gene deserts, 3D structure, mobile elements etc etc. ... and it turns out that many people who say they are genomic researchers are actually just people with a few quid who paid to sequence a stupid genome, like the lesser spotted tree trout. Then they assemble it (badly), submit it to GenBank still full of adapters, and bloody PhiX, and get a paper in *BMC I couldn’t get this into Genome Research*. It’s a scandal – they give genomics a bad name!” he finished, and then went back to his day job as Mayor of London.

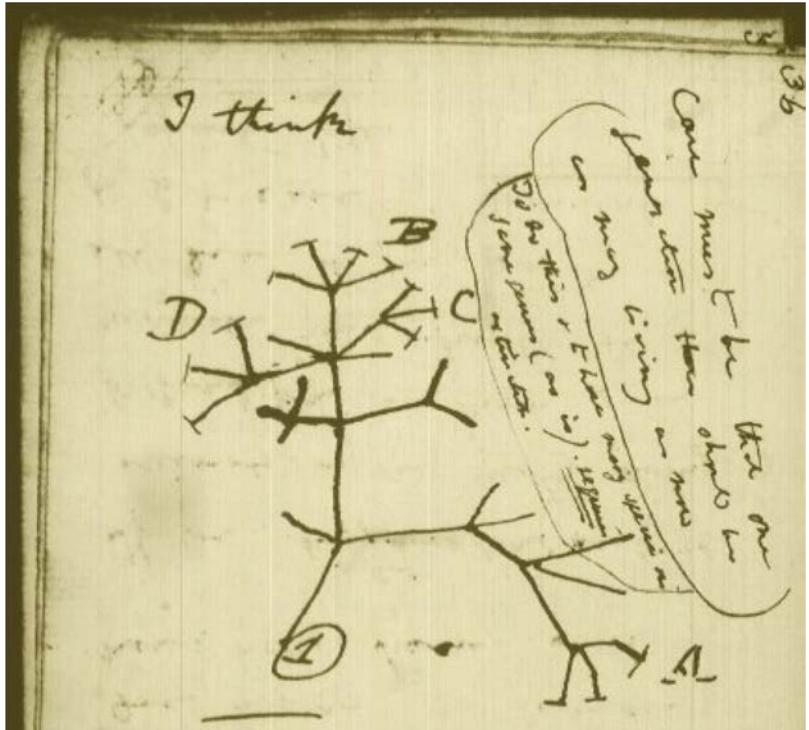
In an earlier survey, it was found that many scientists are sequencing things because they can’t think of anything else to do. Now it would appear that those very same scientists have no idea how to handle the data, and are poisoning the well with hundreds of crappy genomes.

NGS dos and don't

- Embrace it
- Don't just do it without a question
- Don't do it because you can (lots of \$\$, want to jump in)
- Don't just hate it because you don't know how to do it
 - Typical scenario: “We should focus on more traditional methods because NGS is expensive”
 - Typical scenario 2: “These people who do mathematics (?) don't know what ecology/biology/conservation are”

Nothing makes sense in the light of evolution

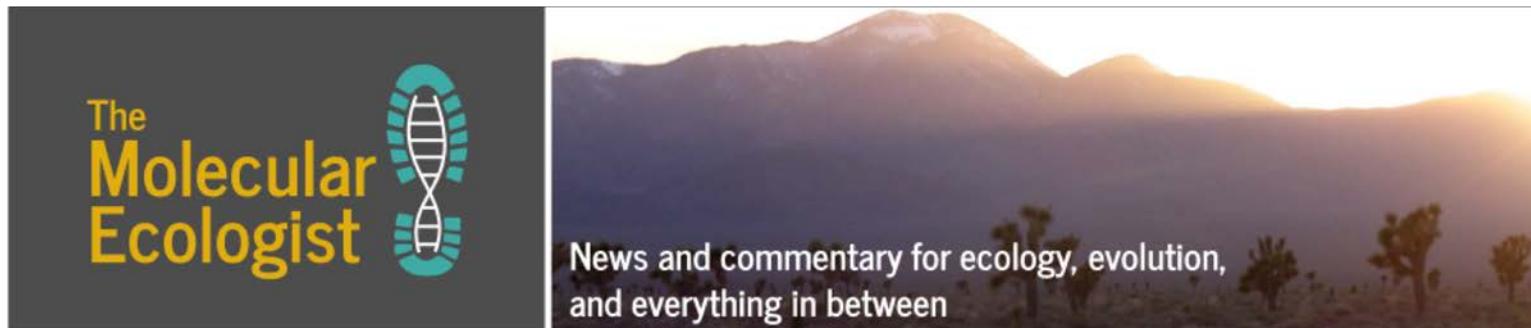
Theodosius Dobzhansky 1973



Survey 2016

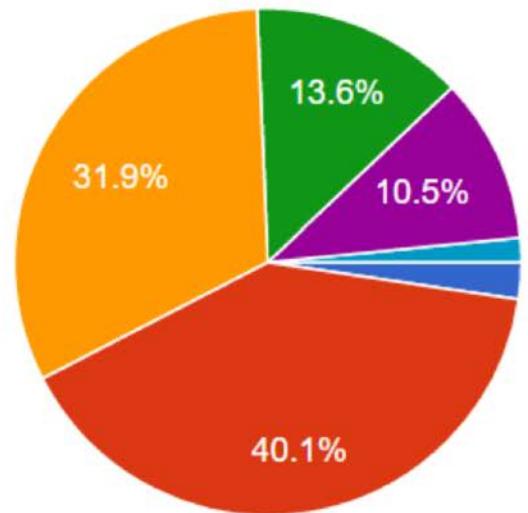
- Total number of students: **39**
- Anyone already have a dataset? **10**
- Anyone about to design their own experiment, produce sequences and analyse themselves? **3**
- Assembly? **4**
- Resequencing? **5**
- RNAseq? **9**
- Familiarity with Linux environment? **5**
- Programming experiences? **6**

Who's doing NGS? (Ecological context)

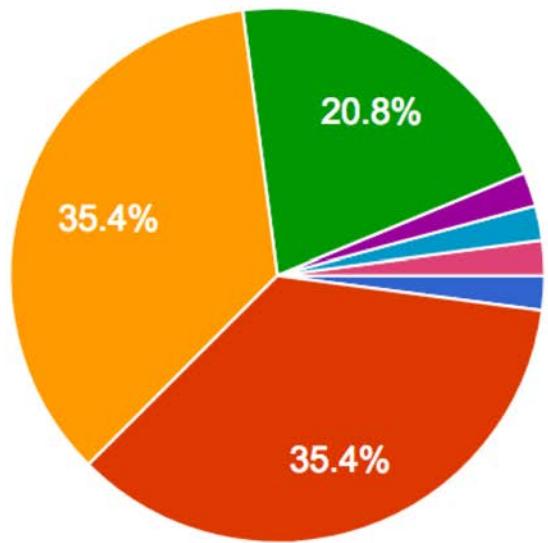


<http://www.molecularecologist.com/2016/04/results-of-the-molecular-ecologists-survey-on-high-throughput-sequencing/>

Which level are you at? (257 responses)

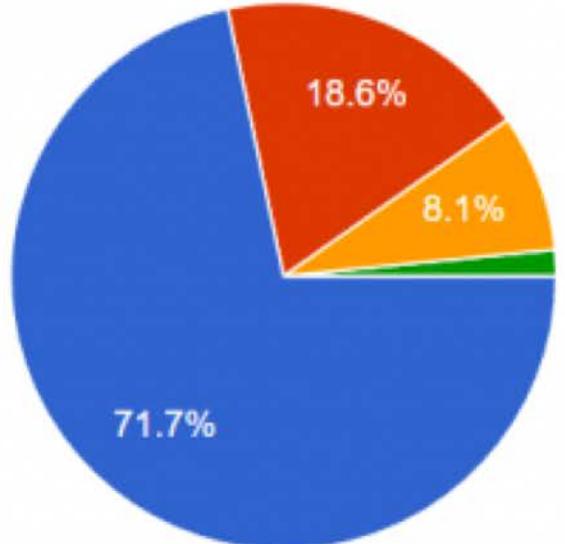


Rest of the world

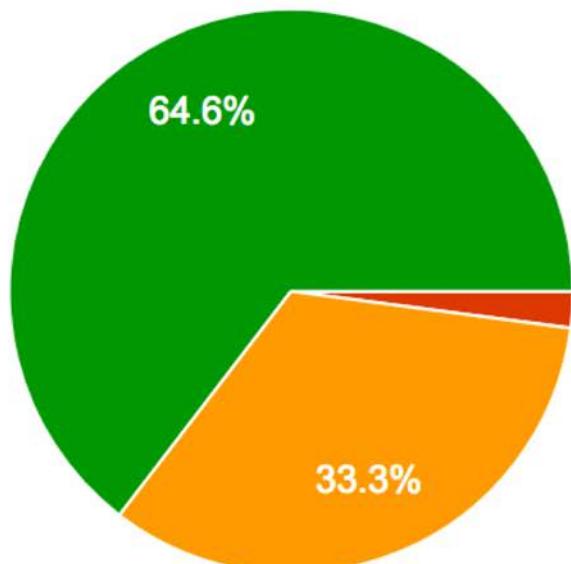


Taiwan

Do you use NGS techniques in your research?

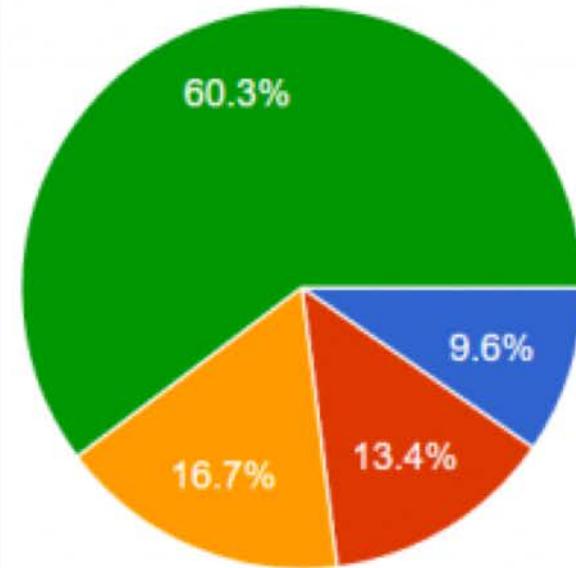


- Yes, extensively
- Yes, a bit
- No, but I plan to use it
- No

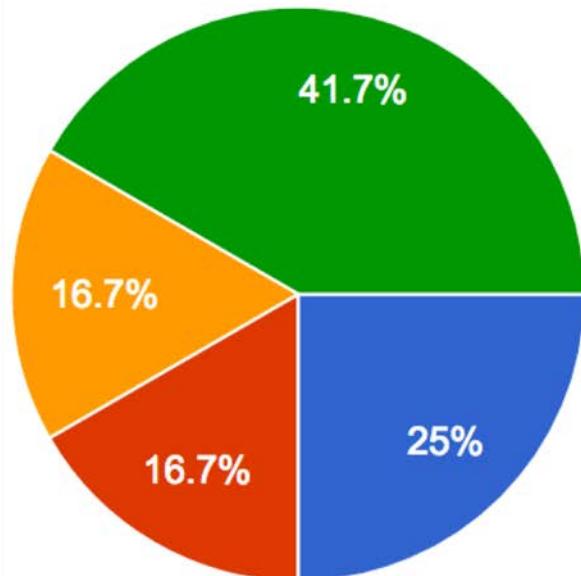


- No, and I don't know why I am here
- No, but I plan to use it
- Yes, a bit
- Yes, extensively

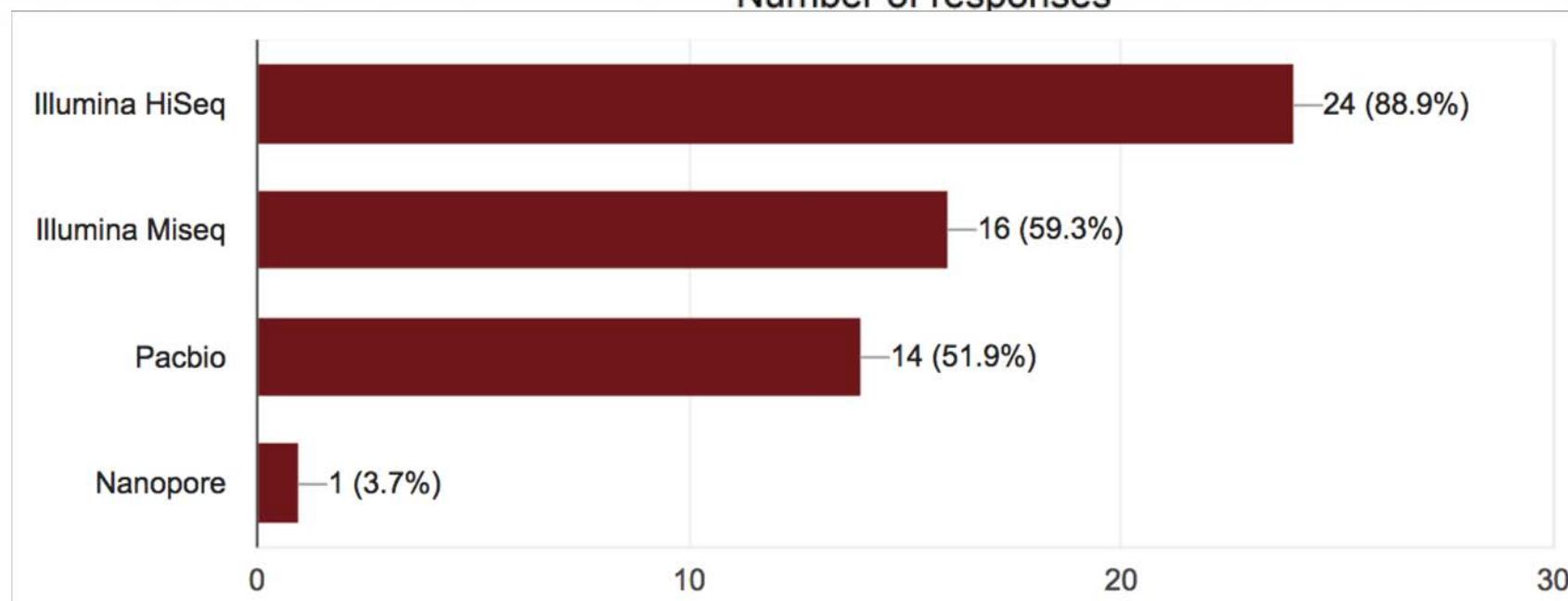
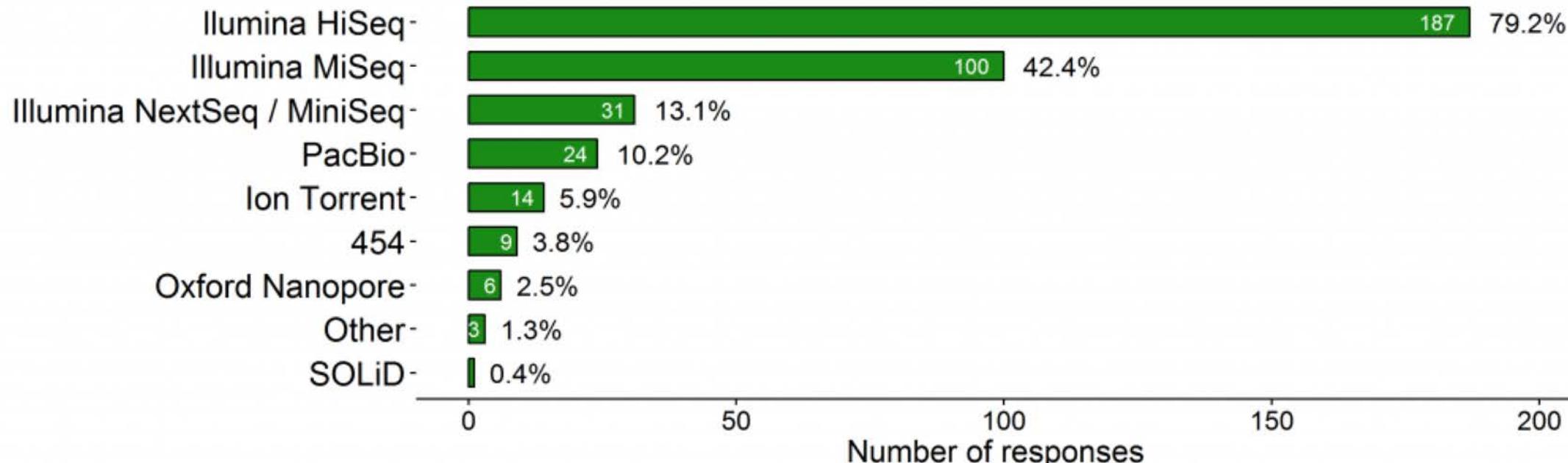
Does your lab have their own sequencing machines?

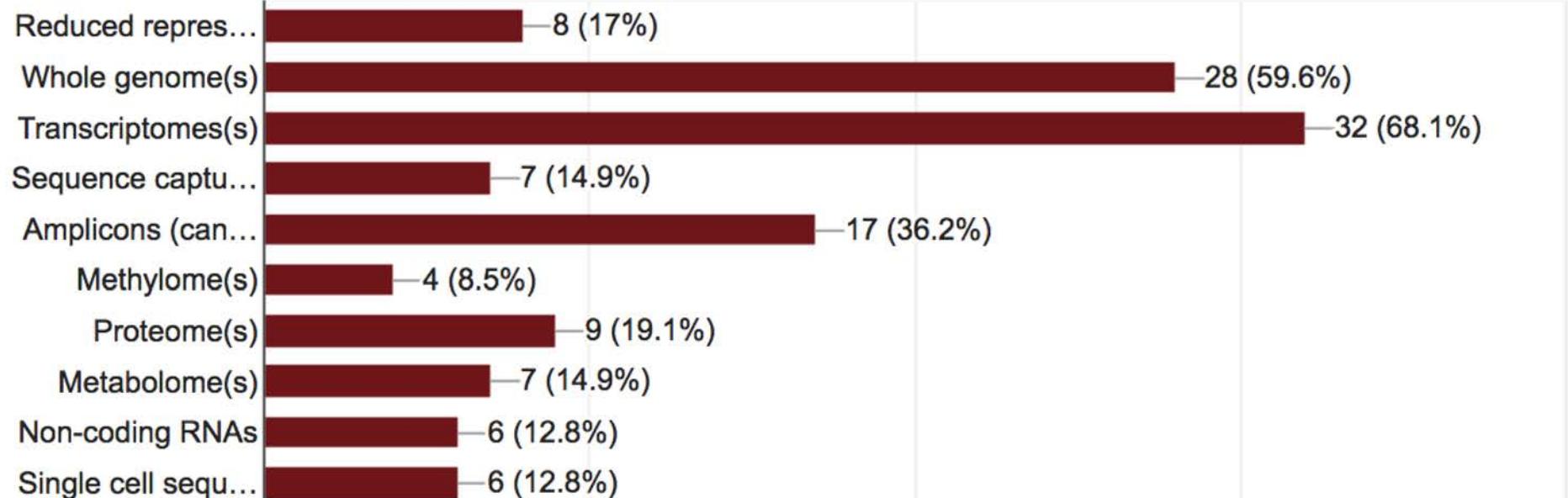
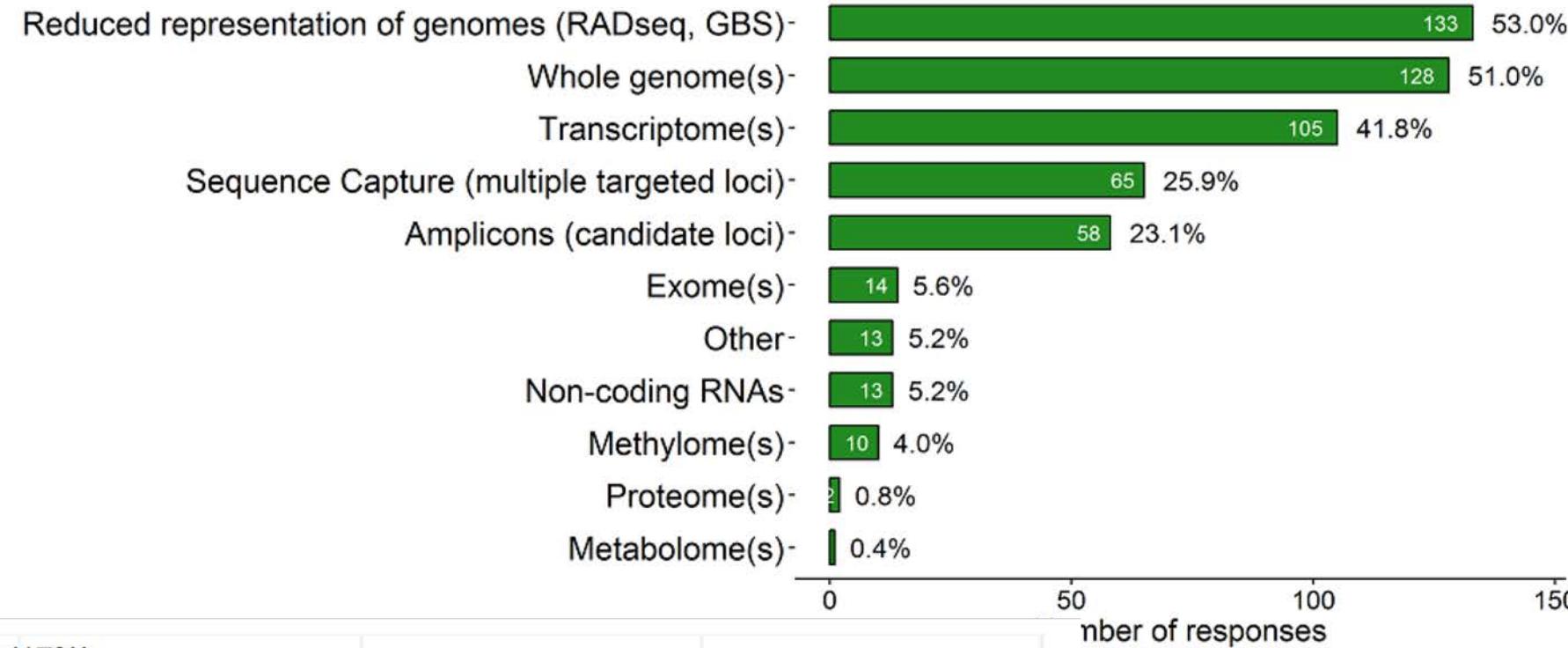


- Yes, several machines
- Yes, two or three
- Yes, one
- No

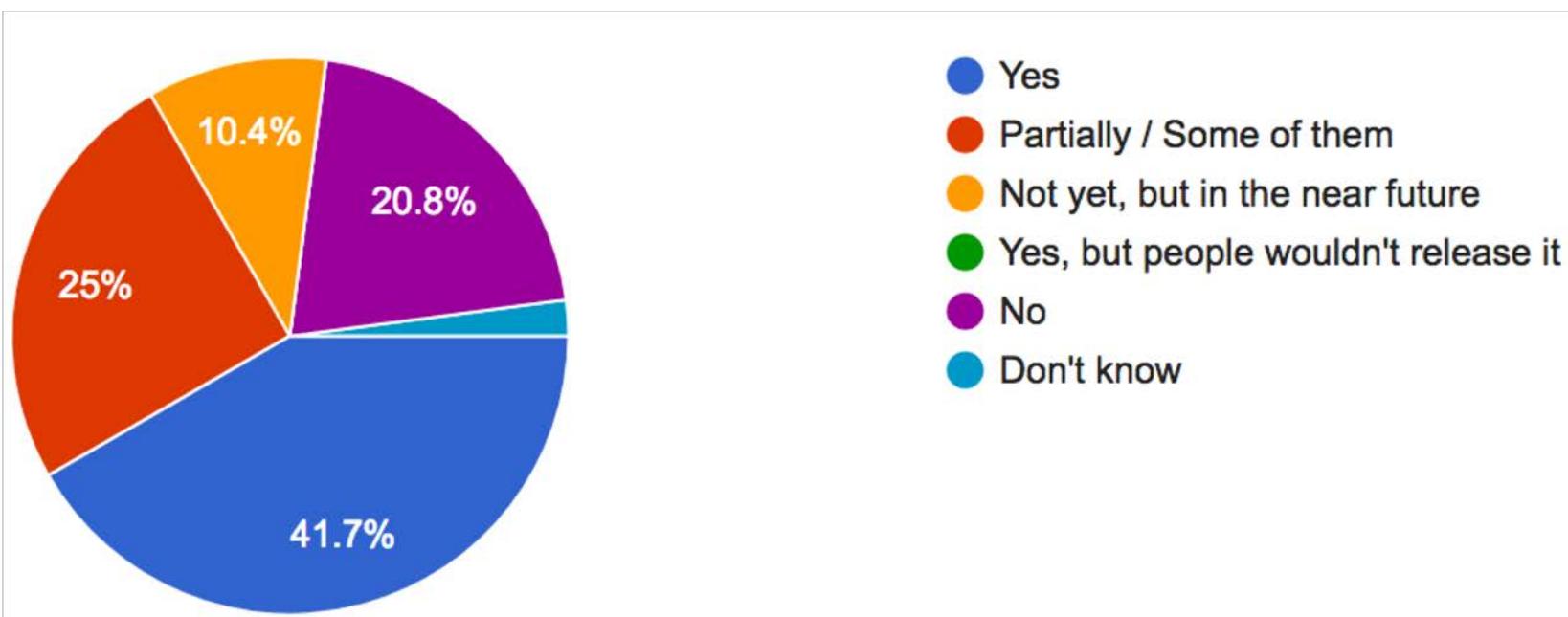


- Yes, several
- Yes, one to three
- Yes, but I can't use them
- No

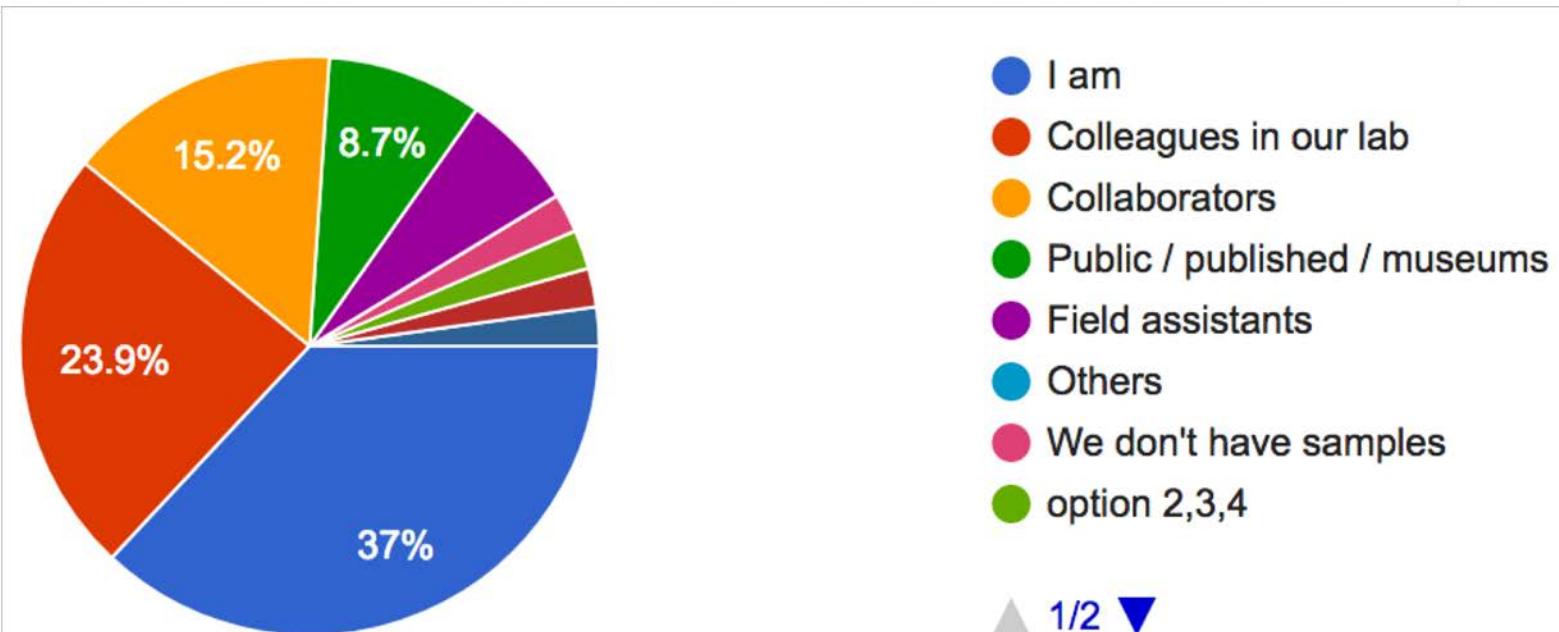
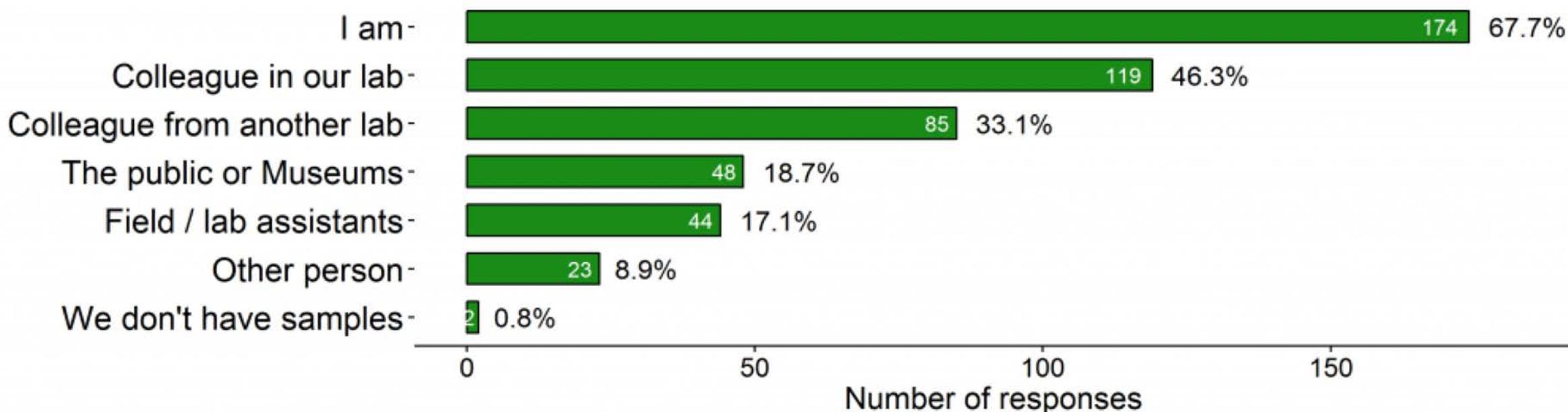




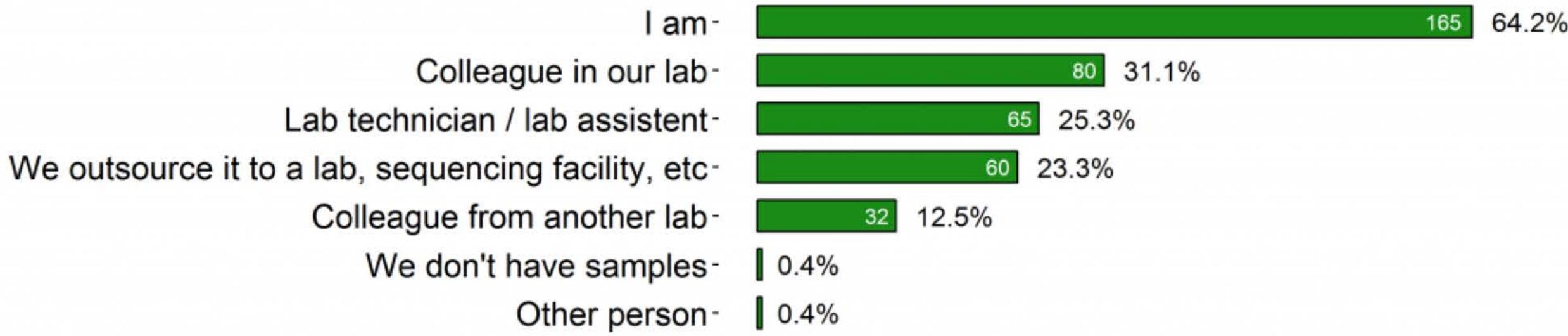
Does your study organism have a reference genome?



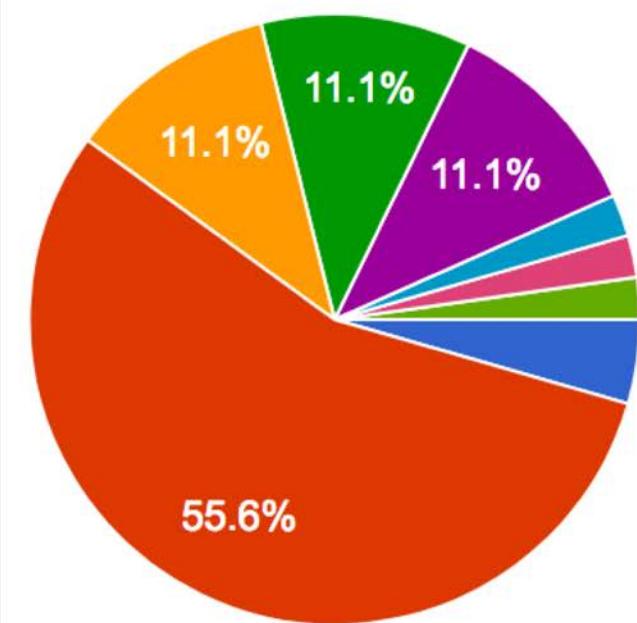
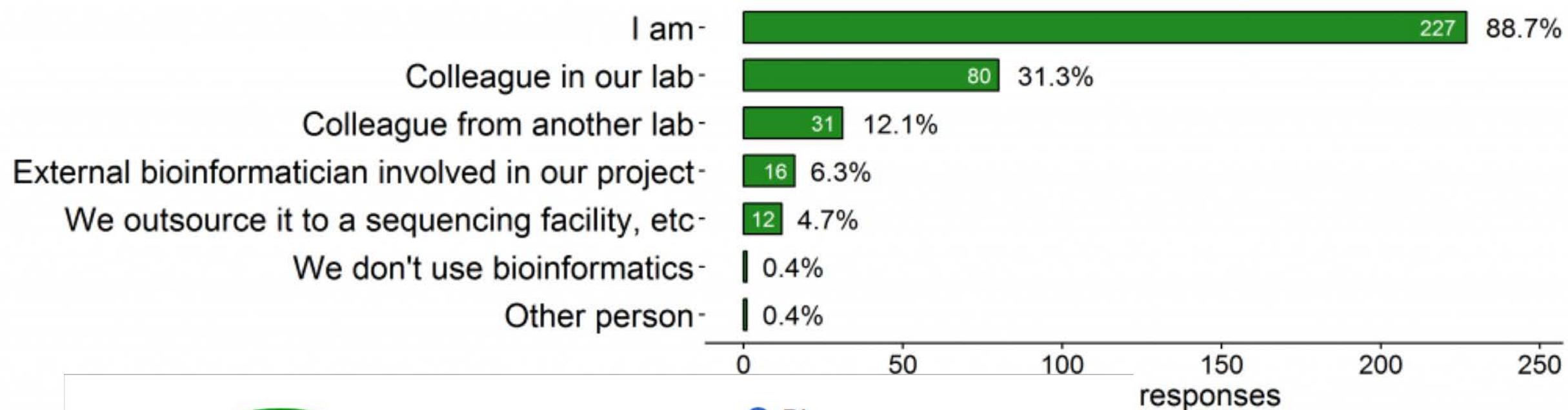
Who is collecting the samples in your main project? (field work or lab work)



Who is doing the wet lab processing of samples in your main project? (DNA extraction and library preparation)



Who is doing the bioinformatic analyses in your main project?



- PI
- I am
- Colleagues in our lab
- Collaborators
- We outsource to a company
- We don't use bioinformatics
- options 2,3,4
- I am and Collaborators

Some food for thought

- A good portion of research scientists are now expected to carry out everything from field to lab to sequencing to analysis
- Research still mainly **human** and/or **reference genome is available** in Taiwan
- Analysis becomes integral part of research (88% vs. 55.6%)
- Scientists in Taiwan seems to be lacking in either collection (67% vs. 37%), sample processing (64% vs 30%) or analysis component (88% vs. 55.6%)
- ... who's doing all the work?

Different sequencing platforms /
History of sequencing

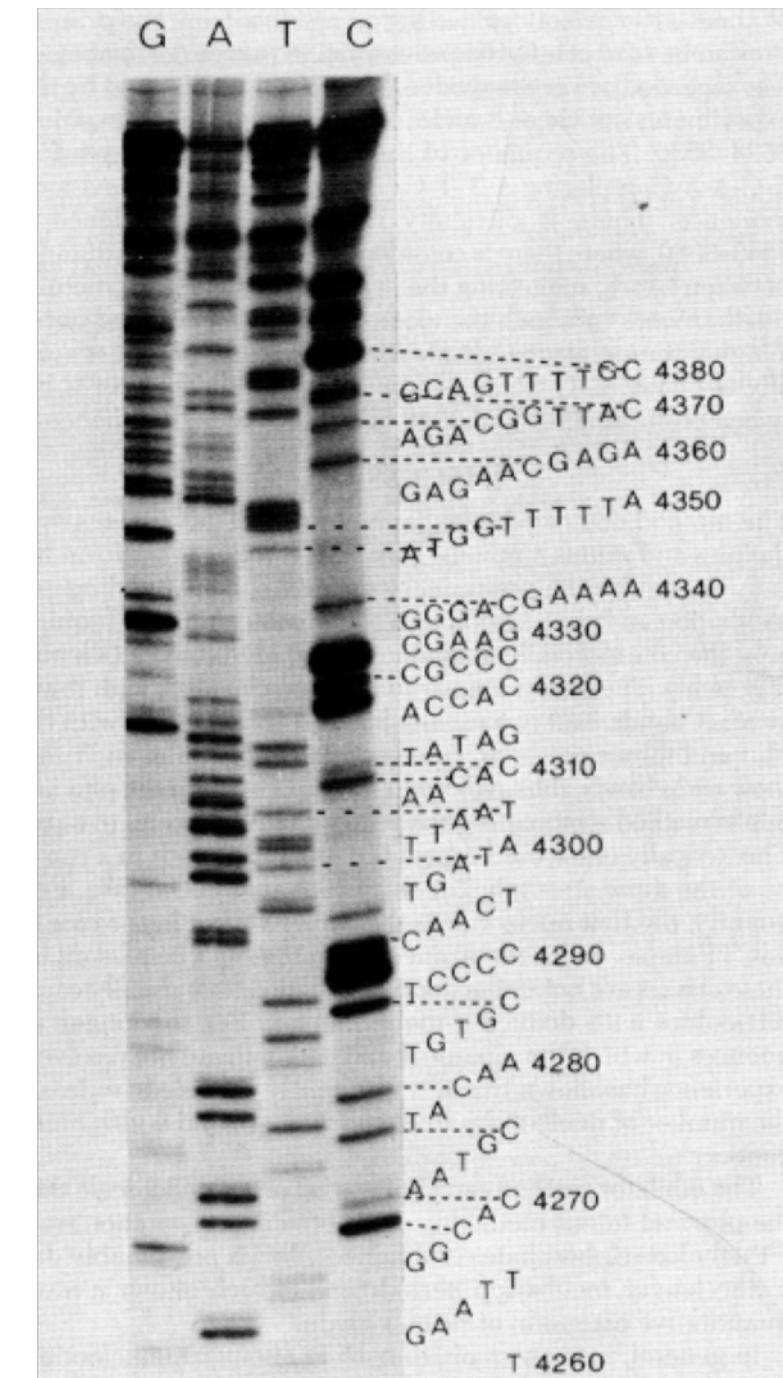
DNA sequencing with chain-terminating inhibitors

(DNA polymerase/nucleotide sequences/bacteriophage ϕ X174)

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

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

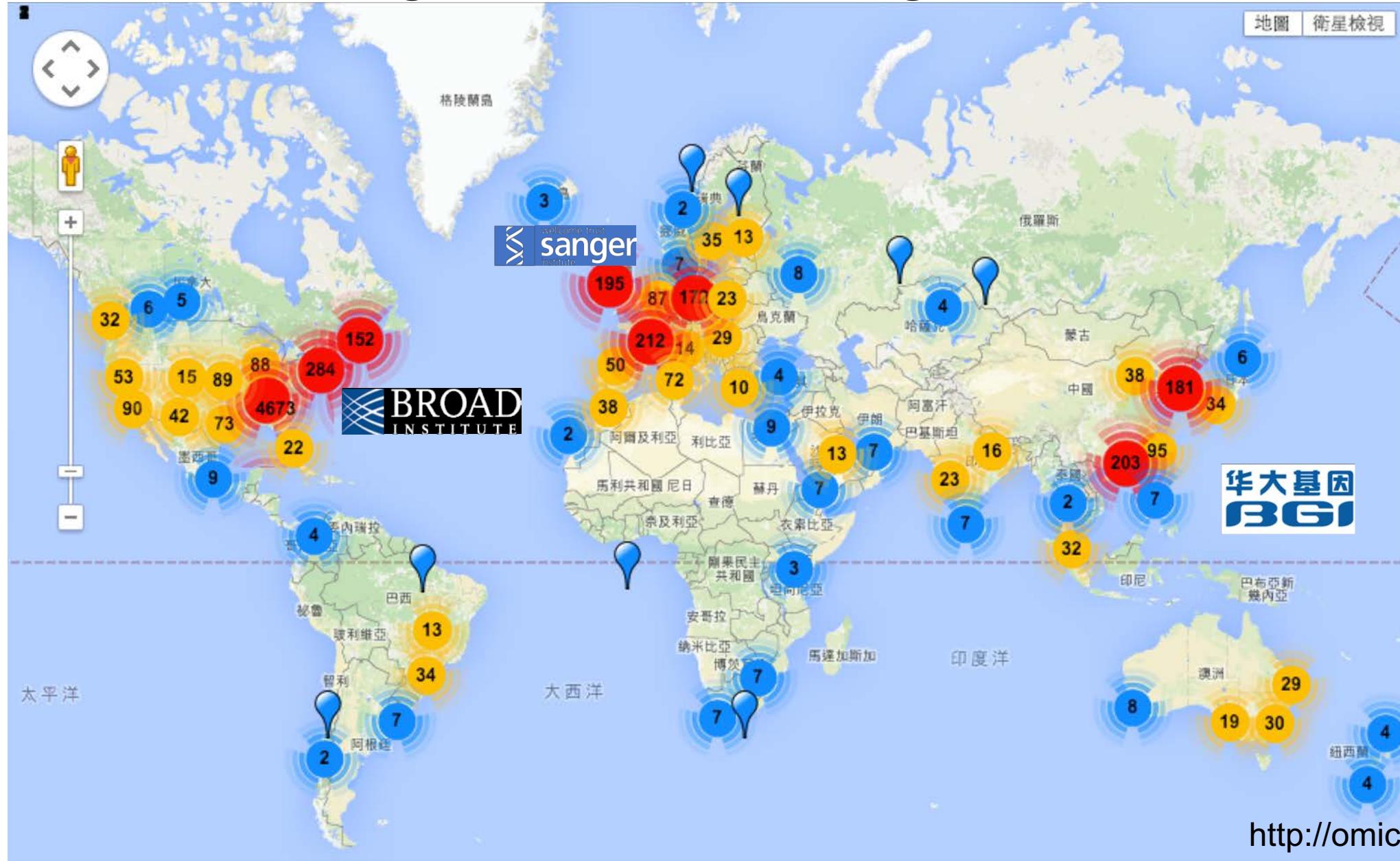
Contributed by F. Sanger, October 3, 1977



ABI 3730xi at TIGR

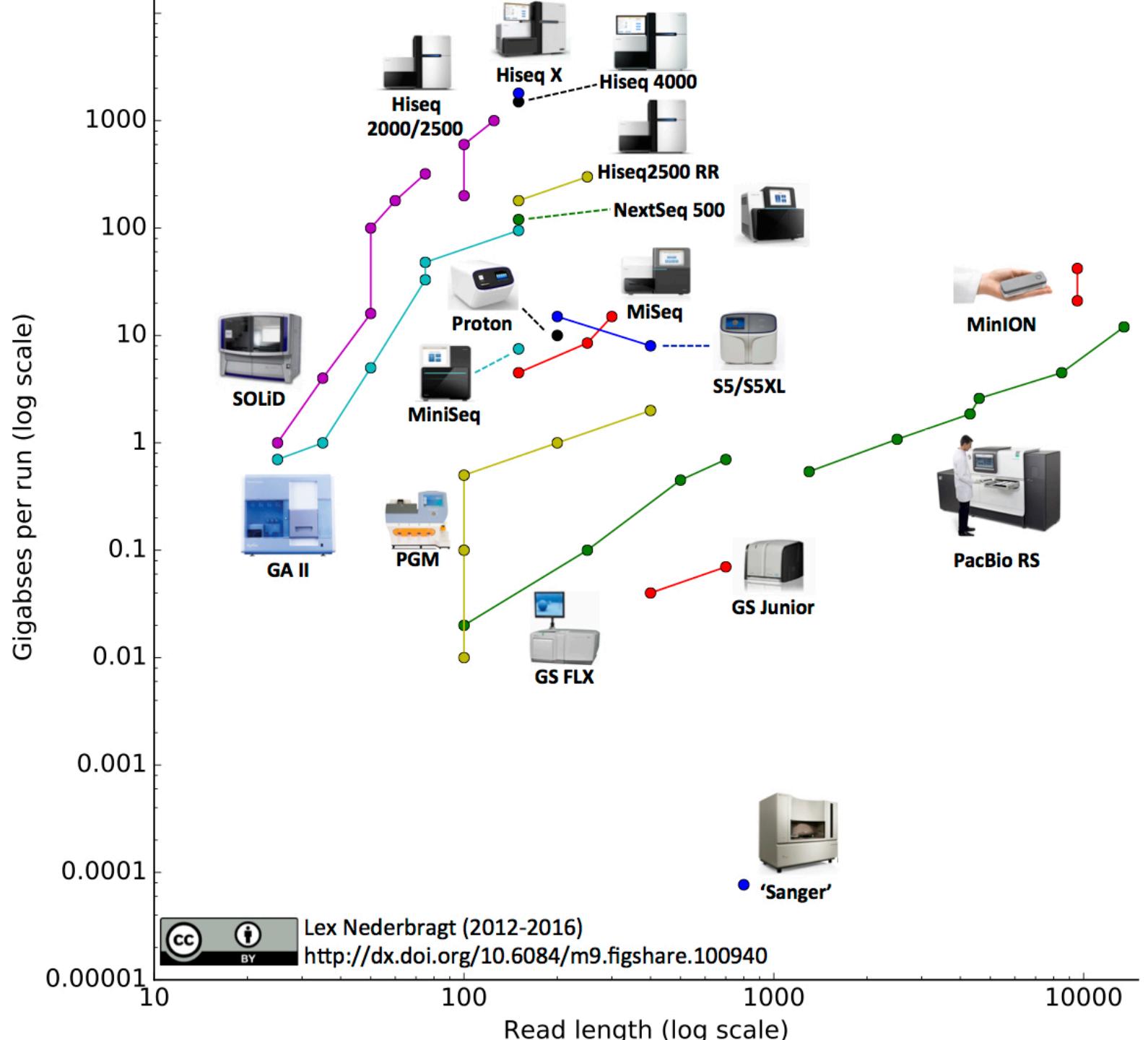


World competing for sequencing power



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



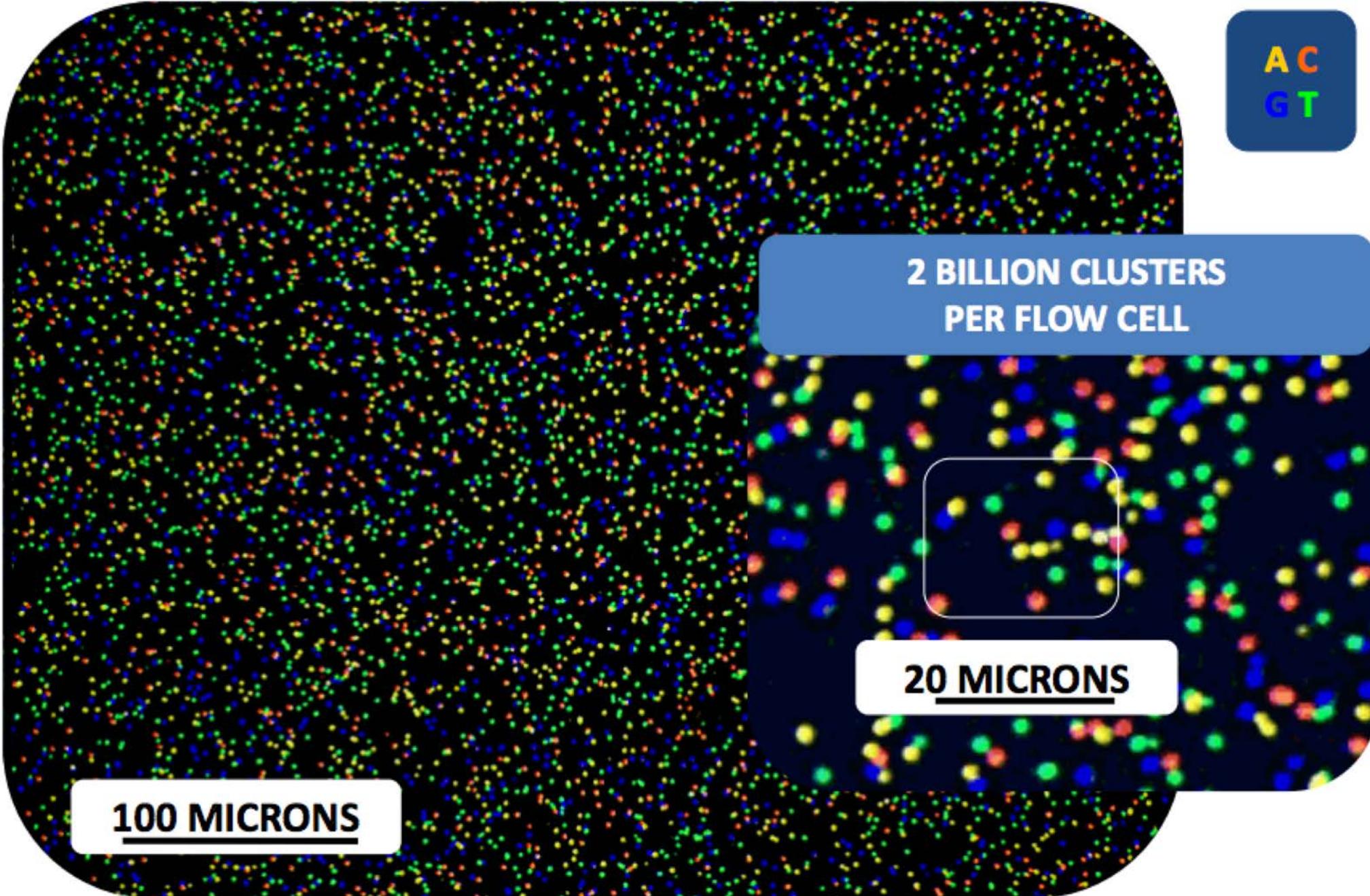
		Read length: (paired-end*, Half of data in reads**)								
Platform	Reads x run: (M)		Run time: (d)	Yield: (Gb)	Rate: (Gb/d)	Reagents: (\$K)	per-Gb: (\$)	hg-30x: (\$)	Machine: (\$)	
iSeq 100 1fcell	4	150*	0.77	1.2	1.56	0.625	521	62500	19.9K	
MiniSeq 1fcell	25	150*	1	7.5	7.5	1.75	233	28000	49.5K	
MiSeq 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>

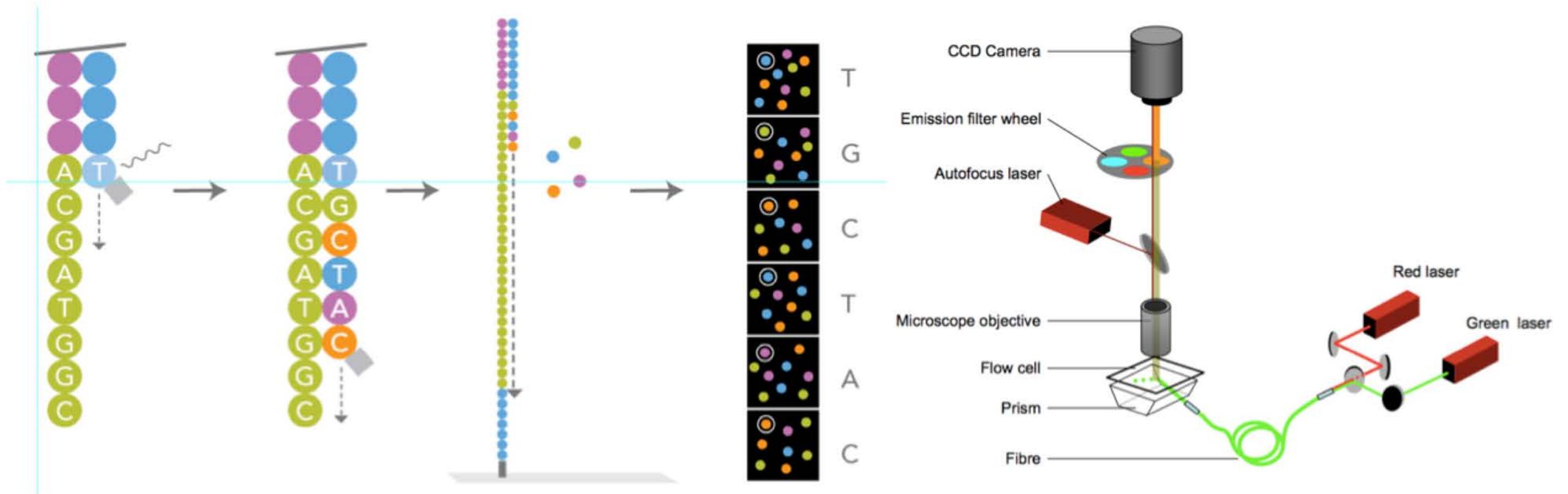


AC
G
T

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

Of Course Size Matters
No one wants
A Small
Glass of Wine



someecards
user card

PacBio (Pacific Biosciences)



RSII

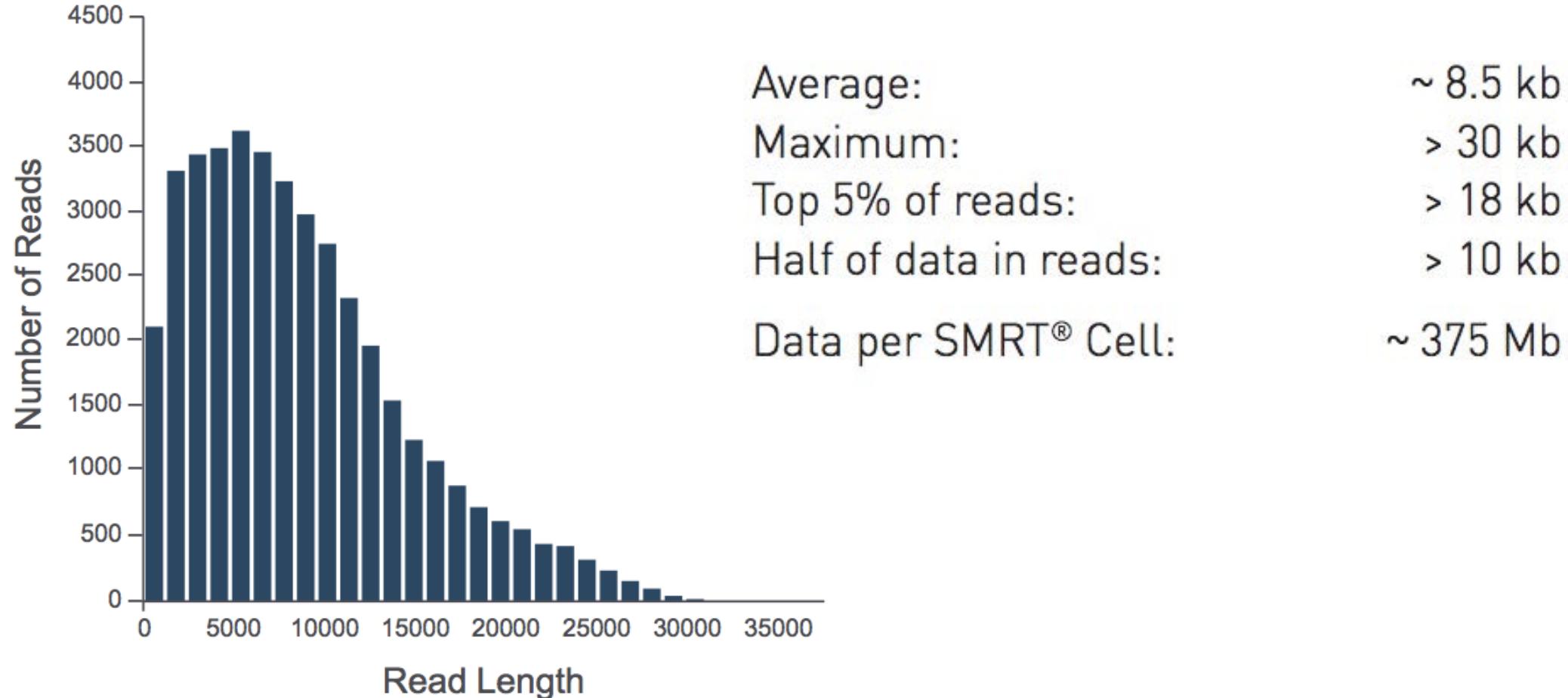


Sequel

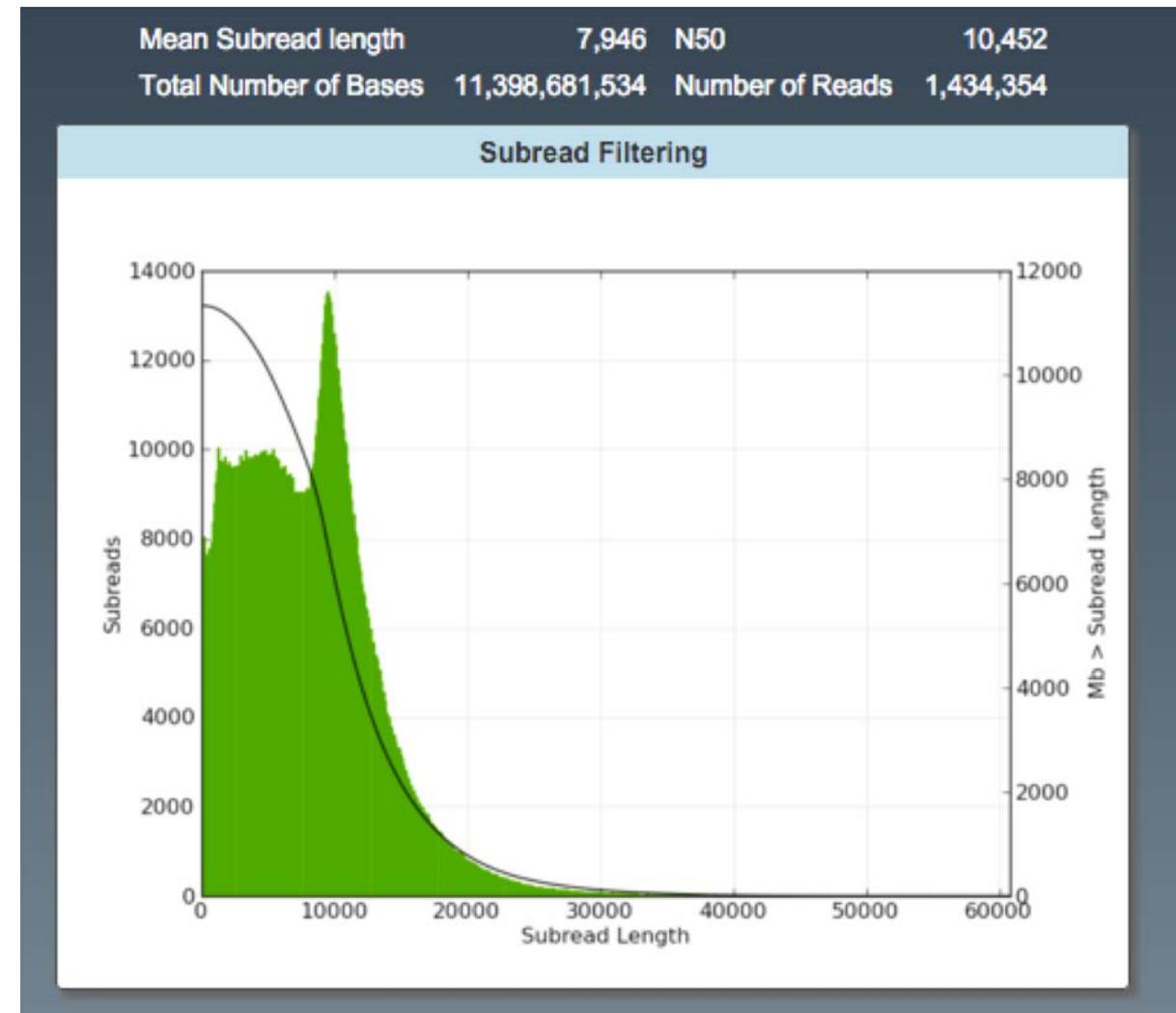
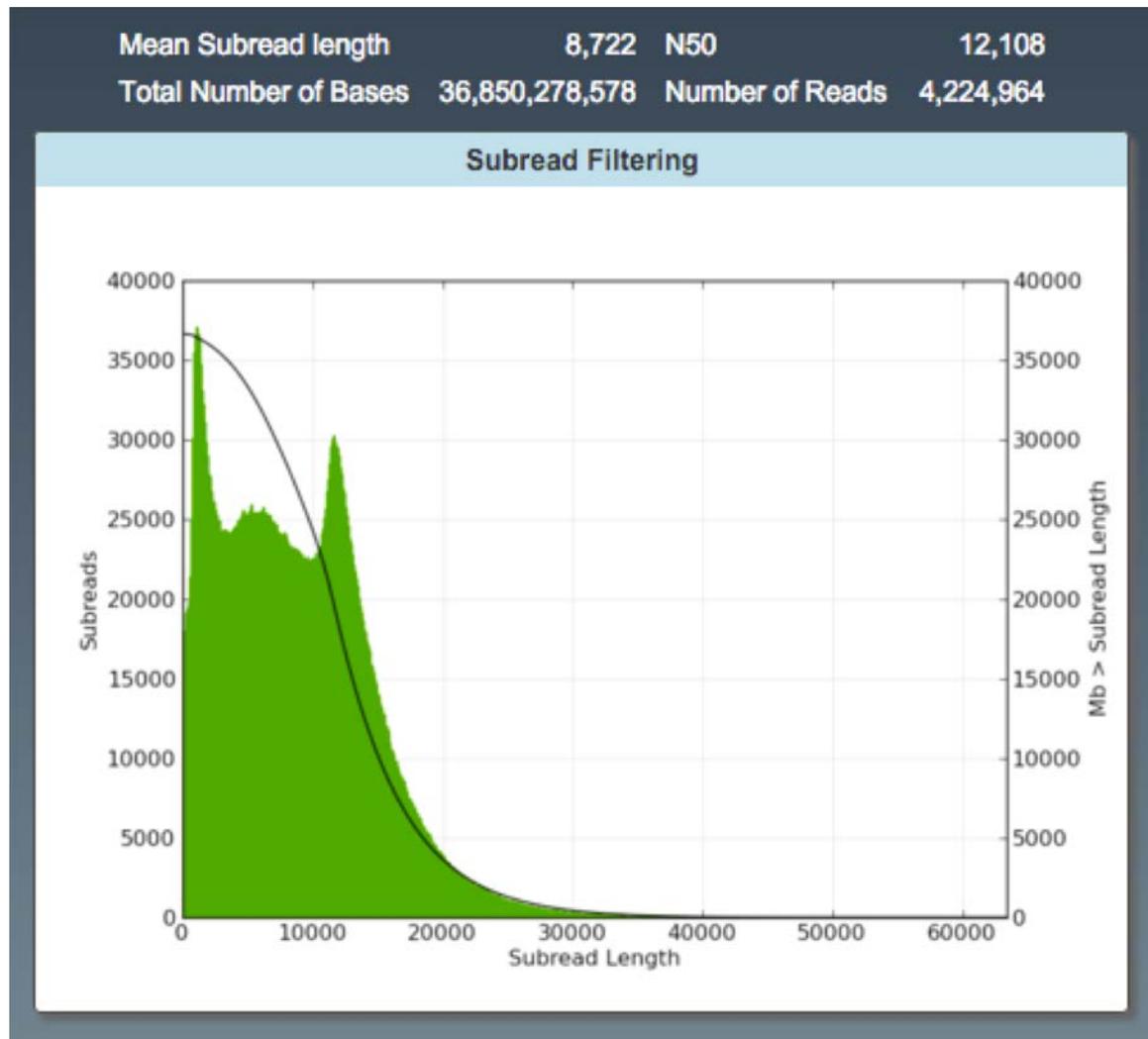
Single molecule sequencing

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

PacBio (Pacific Biosciences)



PacBio (Pacific Biosciences)



Oxford Nanopore

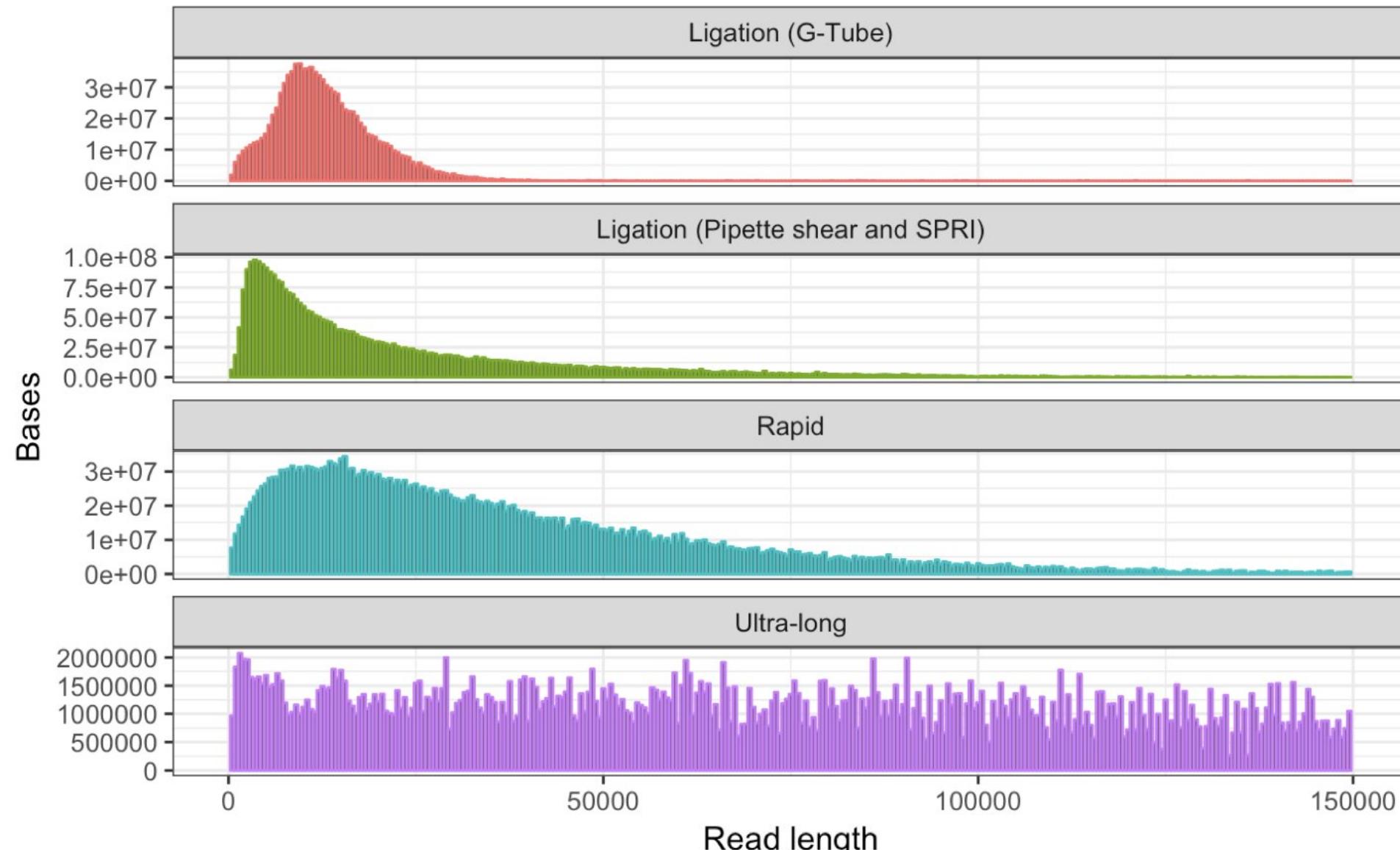


Key	SmidgION	Flongle	MinION	GridION	PromethION
System Price	TBC	Included in \$5K Starter Pack	Included in \$1K Starter Pack	Included in \$50K Starter Pack	Included in \$135K Starter Pack
Number of channels	200 channels	128 channels	512 channels	$5 \times 512 = 2,560^*$	$48 \times 3,000^* = 144,000$
Per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

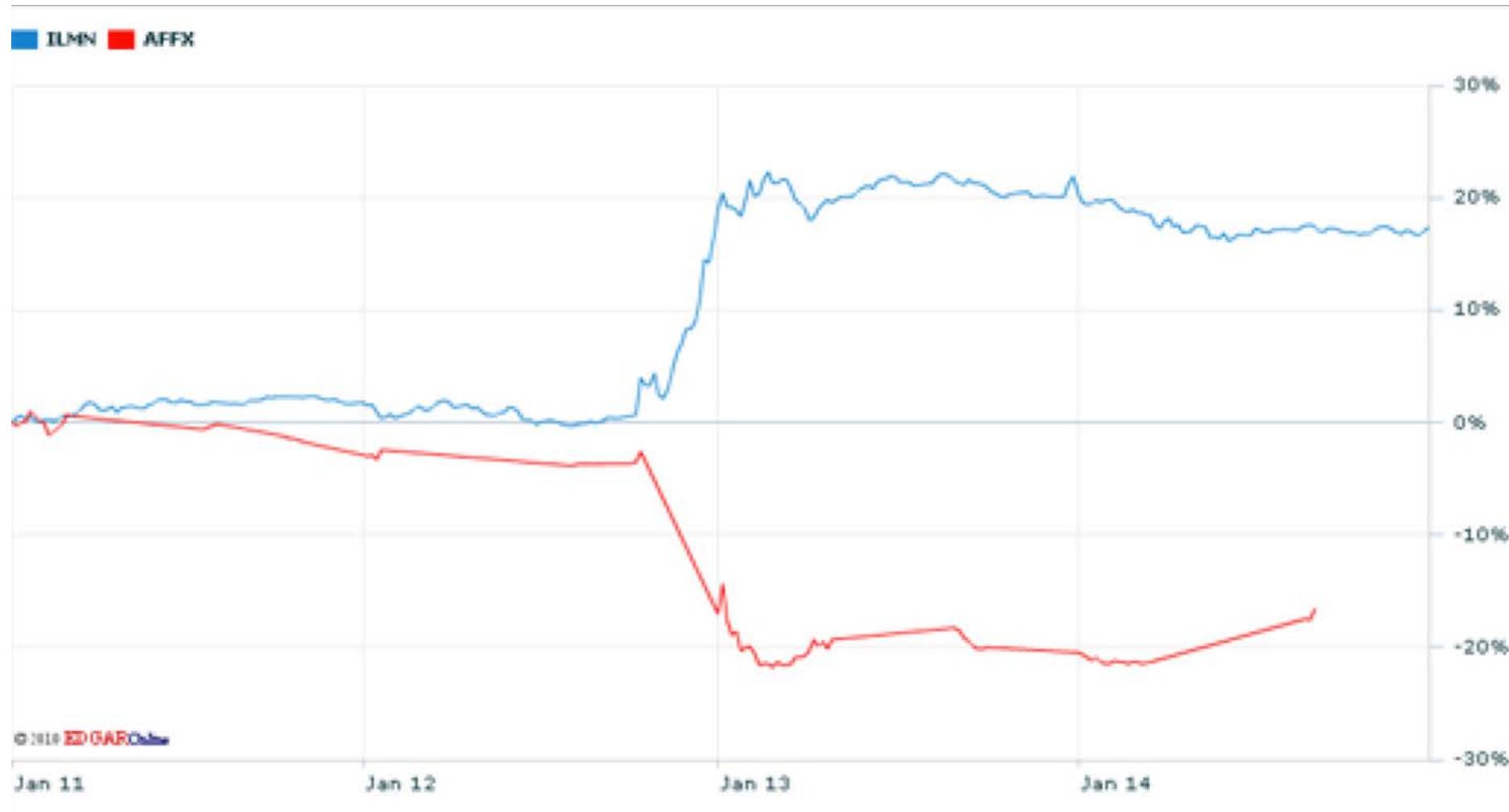
Oxford Nanopore – how it works

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

Read length go beyond



Come and go of technologies



Break here

A lot of data

- We biologists generate a lot of data
 - Experiments, sequencing
 - Everything is more high throughput, but not necessarily less noisy
- Different data types
 - Images, Sequences, Signals, Locations, Linkage, Frequencies...
- How do we
 - analyse them?
 - store them?
 - publish them?
 - reuse them?

A small project's typical output

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

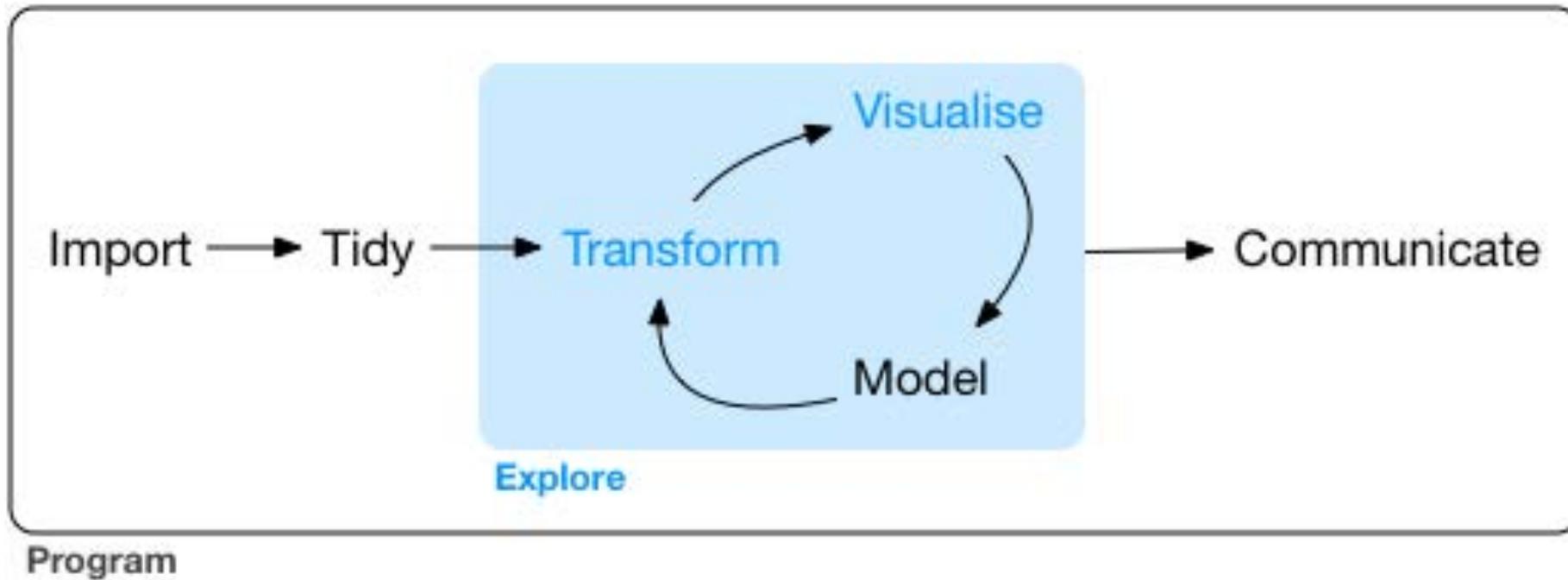
8 exome samples ;

2 Illumina Hiseq lanes with 184GB of data

~100X of human exome to detect disease causing SNP

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

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



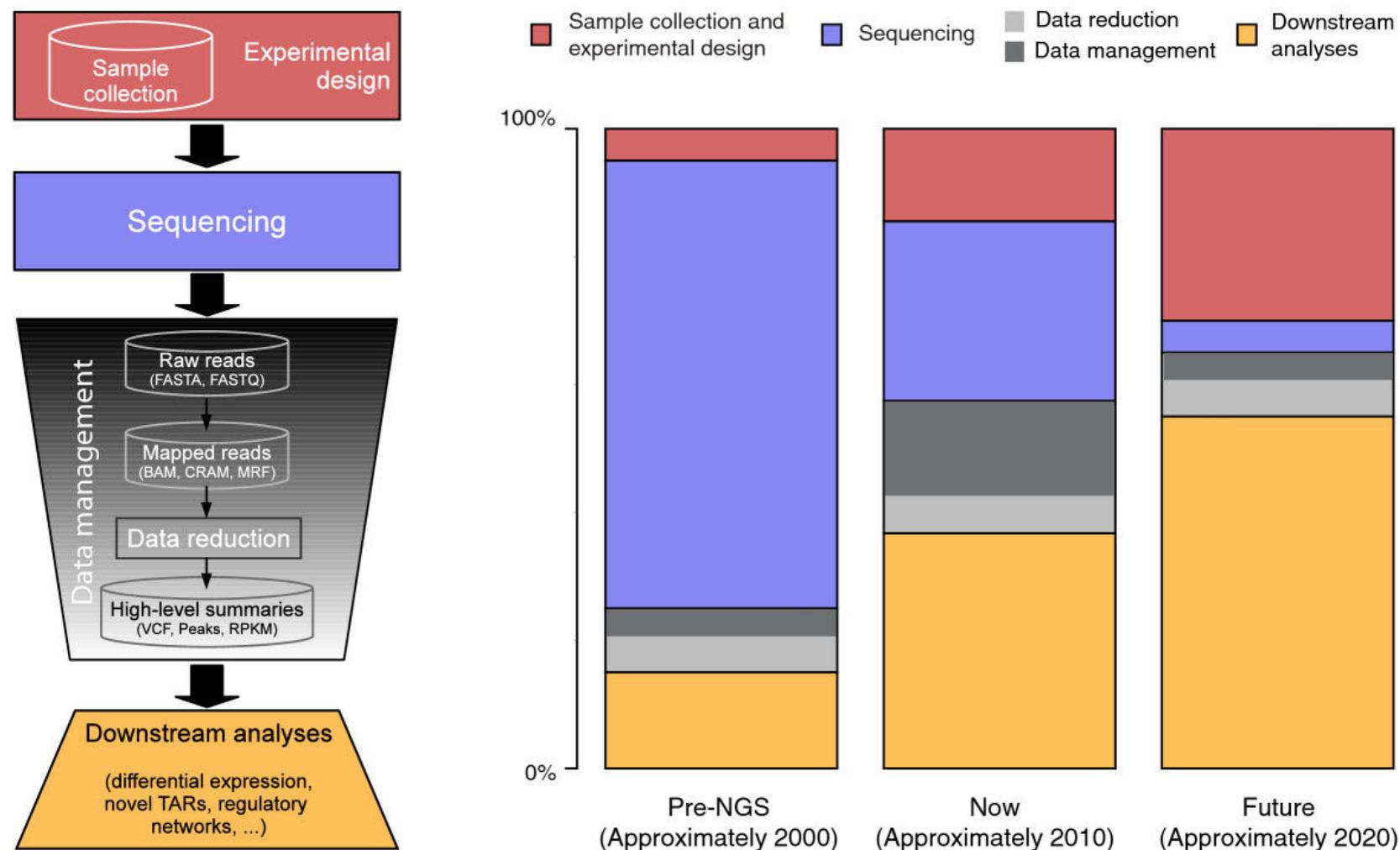
More data but less people with informatics skills

- 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**

OPINION

The real cost of sequencing: higher than you think!

Andrea Sboner^{1,2}, Xinmeng Jasmine Mu¹, Dov Greenbaum^{1,2,3,4,5}, Raymond K Auerbach¹ and Mark B Gerstein^{*1,2,6}



You will end up with an analysis pipeline

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

Important problems:

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

No 'perfect' pipeline – learn through experience



Always understand your data / programs

- Understand:
 - Data format
 - The nature of your data
- Please don't
 - assume data you are given is 'correct'
 - Scenario 1: We got the assemblies and analysis from company XXXX, and we don't know what to do with it
 - assume everything's correct online
 - Run everything in 'default' mode

If unsure – always check **benchmark** studies

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

Resource

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

(GIGA)ⁿ
SCIENCE

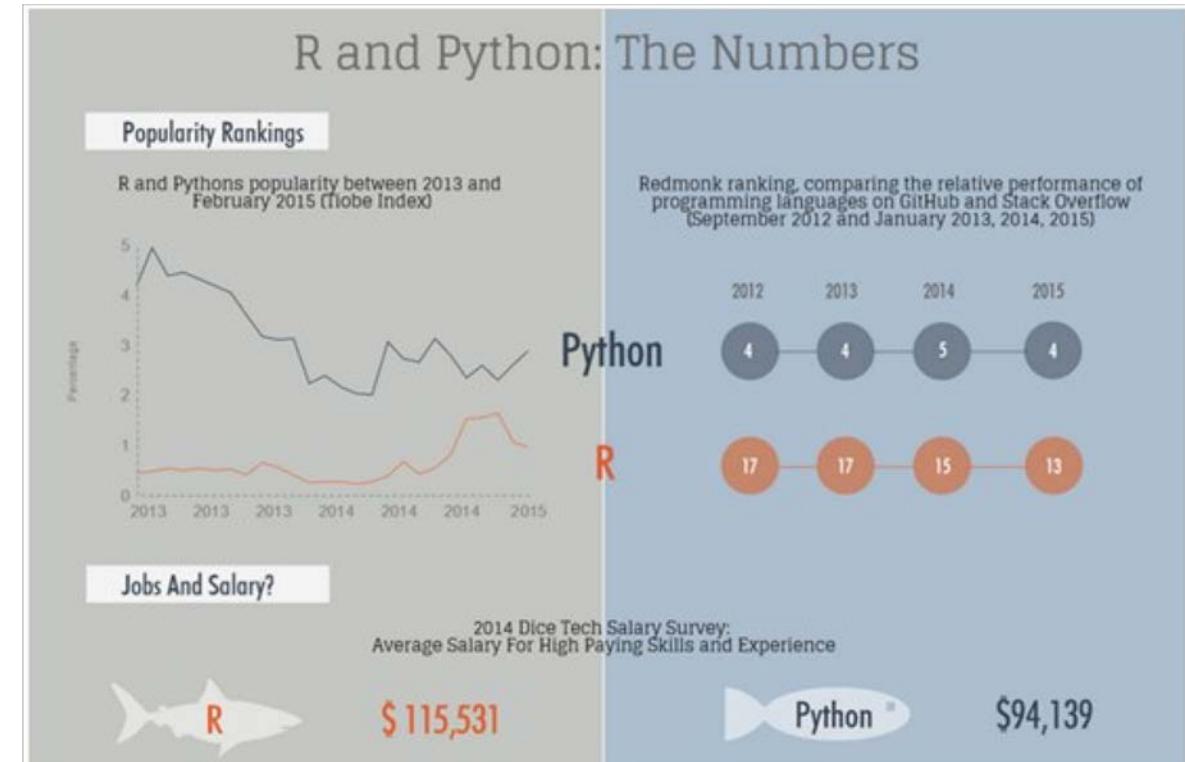
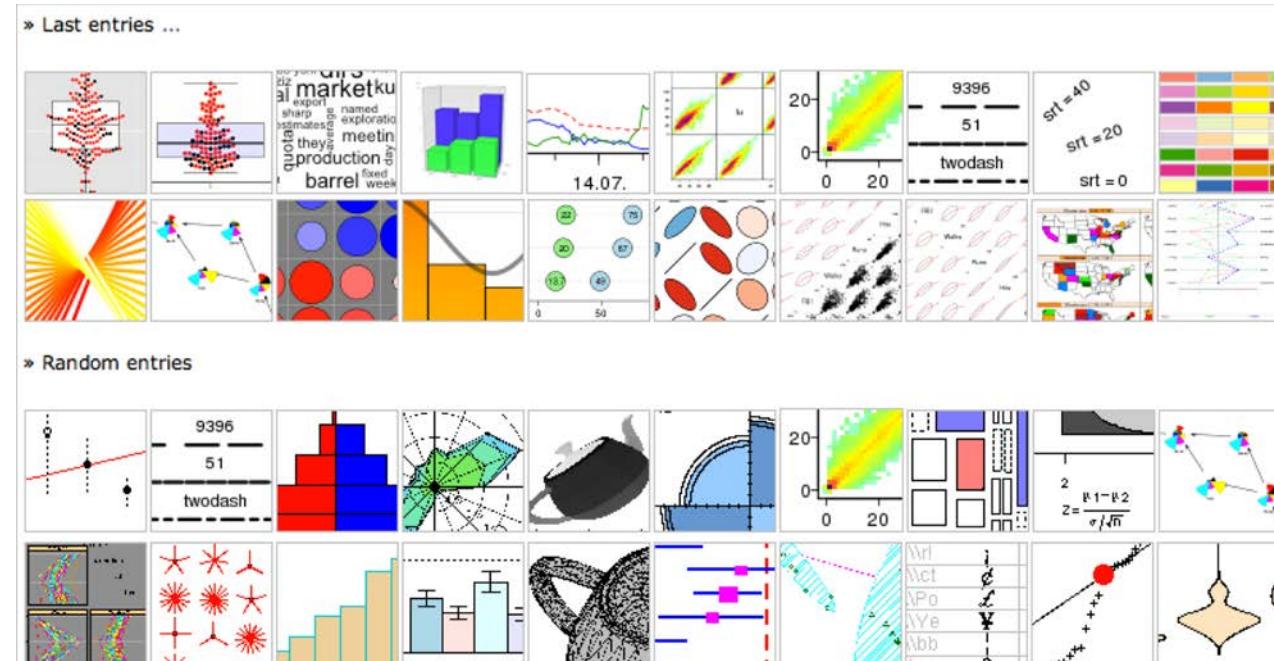
RESEARCH **Open Access**

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

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

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

Python and R



FASTA format

>Name_of_sequence

GCAGGGCATCCGCTGCGTGCTGGCAAAGTCTGTGGCGA
CCGTATGAAACCCCTGAACCCGGCAATACGGTTGCAGATG
CTGGAGGCAGGGCAACGGCAATCAGCTTATTGAGGT
GAGGTACATAAGCCCCGCCAGGCGAGCAGGCAAAGCTGT
TCTGGTCCGAGCCCCAACACAGGGTTCACCAAGTGGCCTG
CCGACGAAAGCGCCGAAGCCCCAACCCATCAATCACTG
GCAGCGTGCAGTCCAGGCCATCGACGAGGCCATCATTGA
AGCGCGGTACGACCCCGAAACGGCACGCTCATTGTTGC
GTTGGCTTCCTATGGTCGGCGCGACCCAGCTTCCCTGGA
ACAGTTGCGCGCCACCTCGCGAAGGAAGGCATTCCCC
CGGAATTCTGTCACATTATGAGCCTGACGGACCCTTGC

Alignment format

- Some programs need slightly modified format

```
>Name_of_sequence_1
GCAGGGCATCCGCTGCGTGCTGGGCAAAGTCTGTGGCGA
CCGTATGAAACCCCTGAACCCGGCAATACGGTTGCAGATG
CTGGAGGCAGGGCAACGGCAATCAGCTTATTGAGGT
GAGGTACATAAGCCCCGCCAGGCGAGCAGGCAAAGCTGT
TCTGGTCCGAGCCCCAACAGGGTTCACCAAGTGGCCTG
CCGACGAAAGCGCCGAAGCCCCG

>Name_of_sequence_2
GCAGGGCATCCGCTGCGTGCTGGGCAAAGTCTGTGGCGA
CCGTATGAAACCCCTGAACCCGGCAATACGGTTGCAGATG
CTGGAGGCAGGGCAACGGCAATCAGCTTATTGAGGTG
AGGTACATAAGCCCCGCCAGGCGAGCAGGCAAAGCTGTTC
TGGTCCGAGCCCCAACAGGGTTCACCAAGTGGCCTGCC
GACGAAAGCGCCGAAGCCCCG
```

Data type keep evolving

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

>Name_of_sequence_1

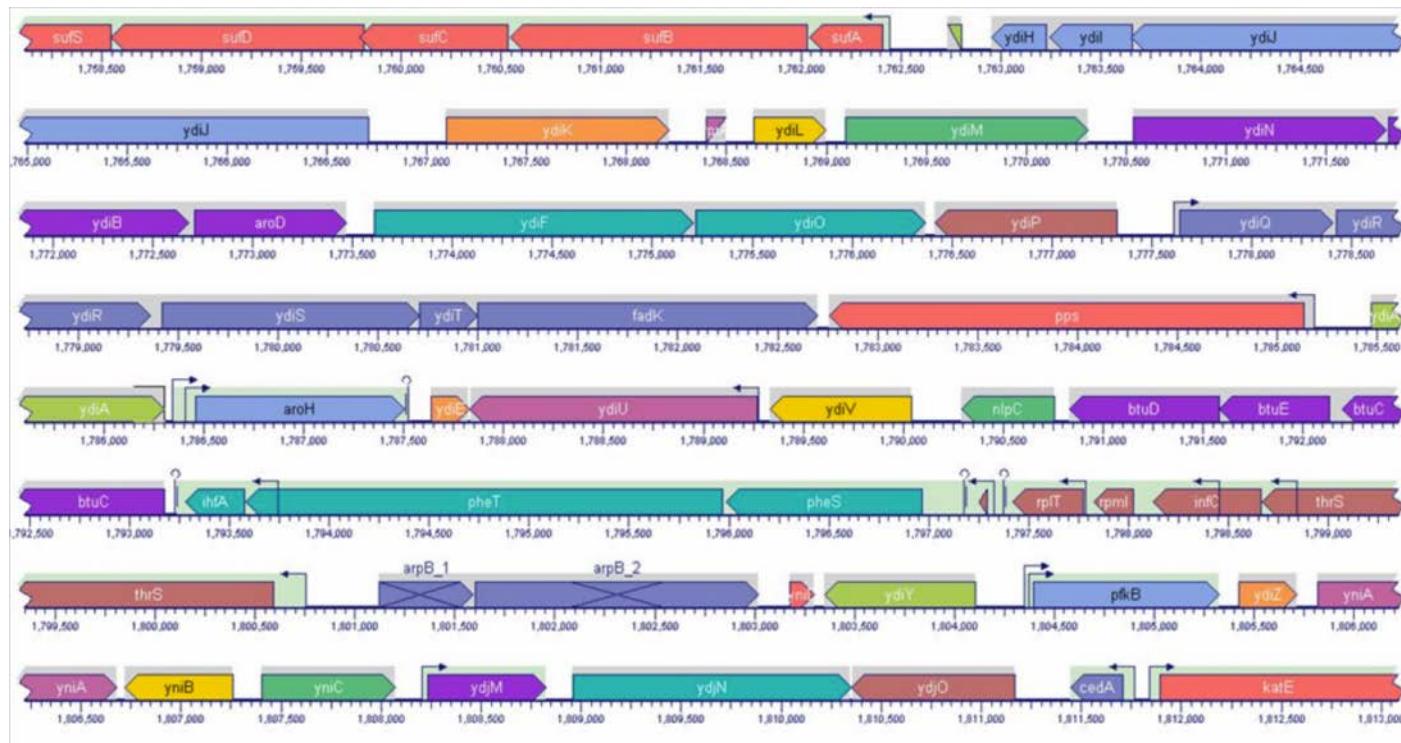
GCGGGTA

>Name_of_sequence_1

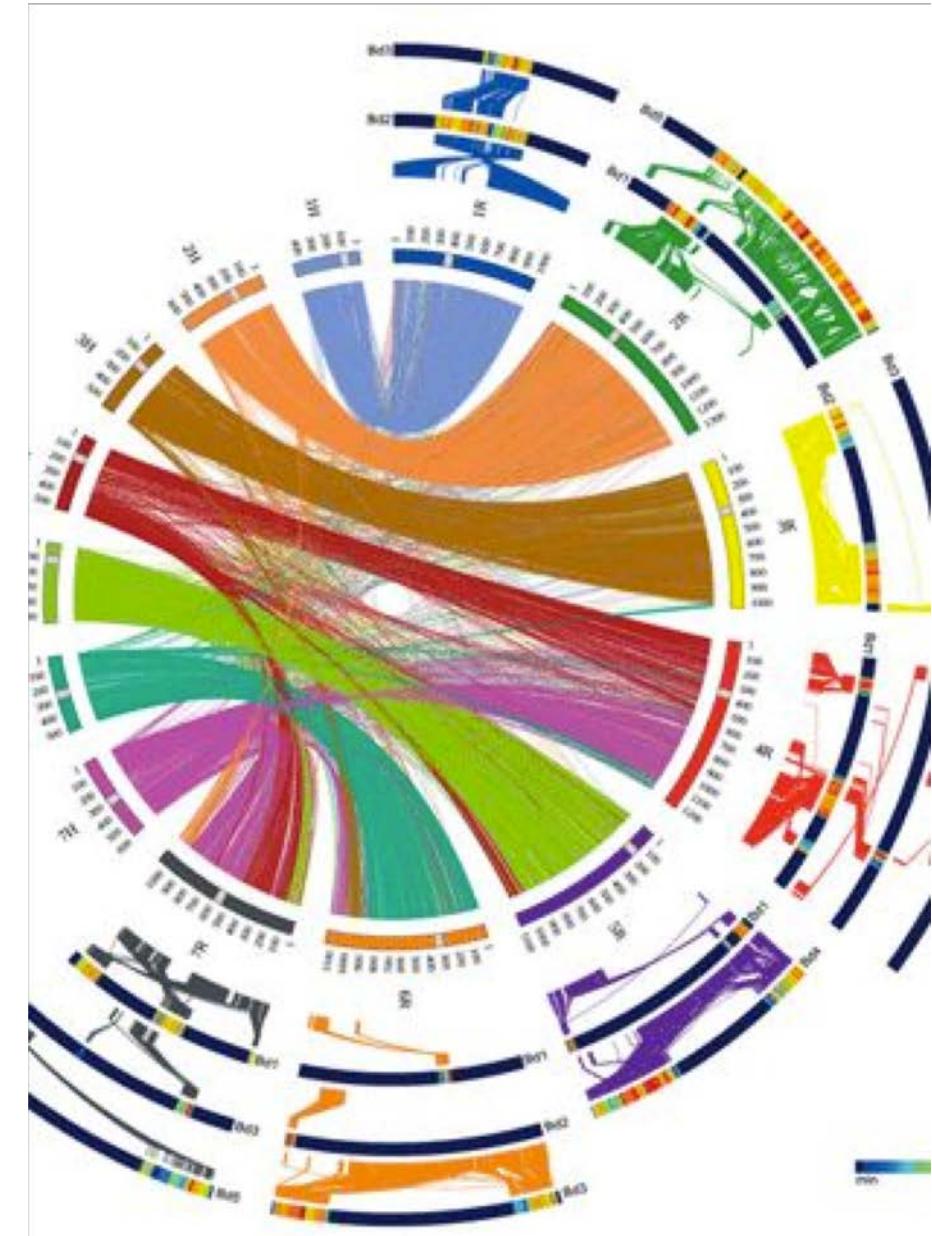
20 30 33 30 20 33 19

Locations / maps

- How do we represent/visualise them?



Gene locations / strand



Circos

BED/gff format

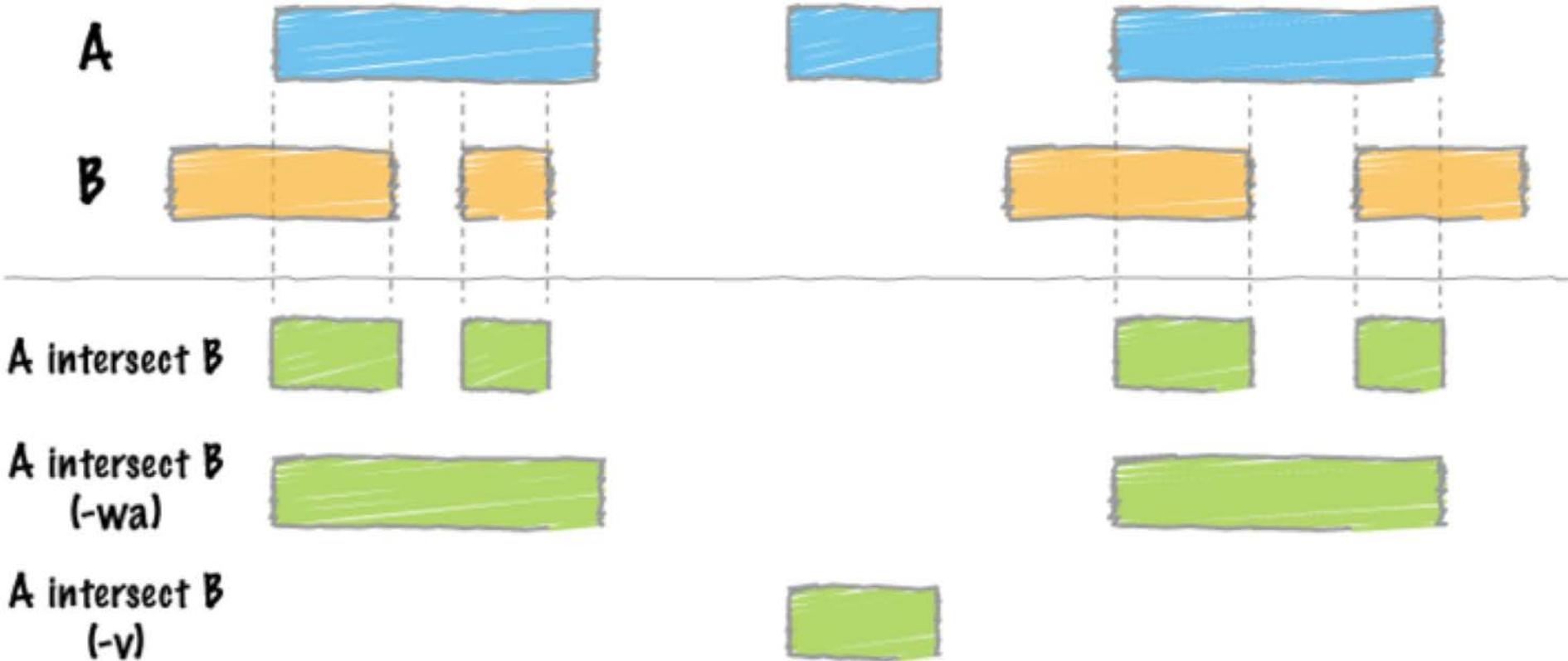
- Features on genome use bed / gff files to represent their locations
- “Optional field” can be added for additional information

chr7	127471196	127472363
chr7	127472363	127473530
chr7	127473530	127474697
chr7	127474697	127475864
chr7	127475864	127477031
chr7	127477031	127478198
chr7	127478198	127479365
chr7	127479365	127480532
chr7	127480532	127481699

IV	curated exon	5506900	5506996	.	+	.	Transcript	B0273.1
IV	curated exon	5506026	5506382	.	+	.	Transcript	B0273.1
IV	curated exon	5506558	5506660	.	+	.	Transcript	B0273.1
IV	curated exon	5506738	5506852	.	+	.	Transcript	B0273.1

Bedtools – extremely useful

Intersect w/
1 database



SAM format

- 1 DNA is extracted from a sample.
 - 2 DNA is sequenced.
 - 3 Raw sequencing reads are aligned to a reference genome.
 - 4 Aligned reads are evaluated and visualized.
 - 5 Genomic variants, including single nucleotide polymorphisms (SNPs), small insertions and deletions are identified.
-]
- samtools

SAM format

- Everything in one line

```
HISEQ:134:C6H9FANXX:8:1110:10236:94013 99    chr1  11844  0    150M  =    12057  363
GGTATCATTACCCATTTCCTTCTGTTAACCTGCCGTCAGCCTTTCTTGACCTCTTCTGTTC
ATGTGTATTGCTGTCTCTAGCCCAGACTTCCCGTATCCTTCCACCGGGCCTTGAGAGGTACAC
GGGTCTTGATGCTG
>A=>AFDEEGEGGEFFFFFFFGCDFBEGFFHFGCDGEHGGFFFFGFFEDGGFGFFGFFFDFEDF
GCFHCHDBEFFHFEGCFEFED@CEEEBEADCBBCB>?,?AA@@@?@?>@?;?@??==?=?:<=?@GEH
GFDGFFHAC=?=@ MC:Z:150M
BD:Z:NNOOPSQQNOPMNOOMGGGNMMGGNONNLNONOPOMNPOPQOONHGONNHOPOOOO
NNONNHONPMNOPPPNMONNHOPPPMQNONNM
```

SAM format

- Bitwise flag

1	QNAME	Query template/pair NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition/coordinate of clipped sequence
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	extended CIGAR string
7	MRNM	Mate Reference sequence NaMe ('=' if same as RNAME)
8	MPOS	1-based Mate POSition
9	TLEN	inferred Template LENGTH (insert size)
10	SEQ	query SEQuence on the same strand as the reference
11	QUAL	query QUALity (ASCII-33 gives the Phred base quality)
12+	OPT	variable OPTIONAL fields in the format TAG:VTYPE:VALUE

mapped in correct orientation and within insert size

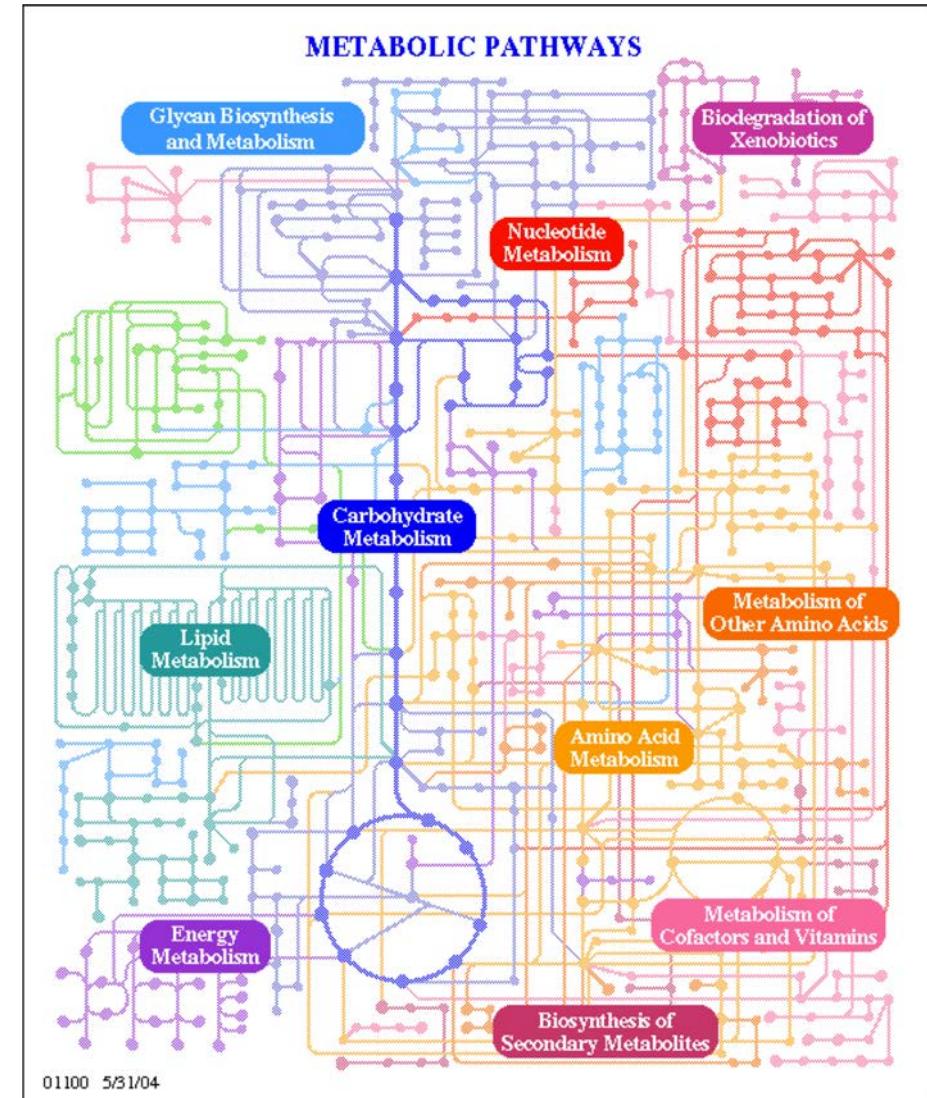
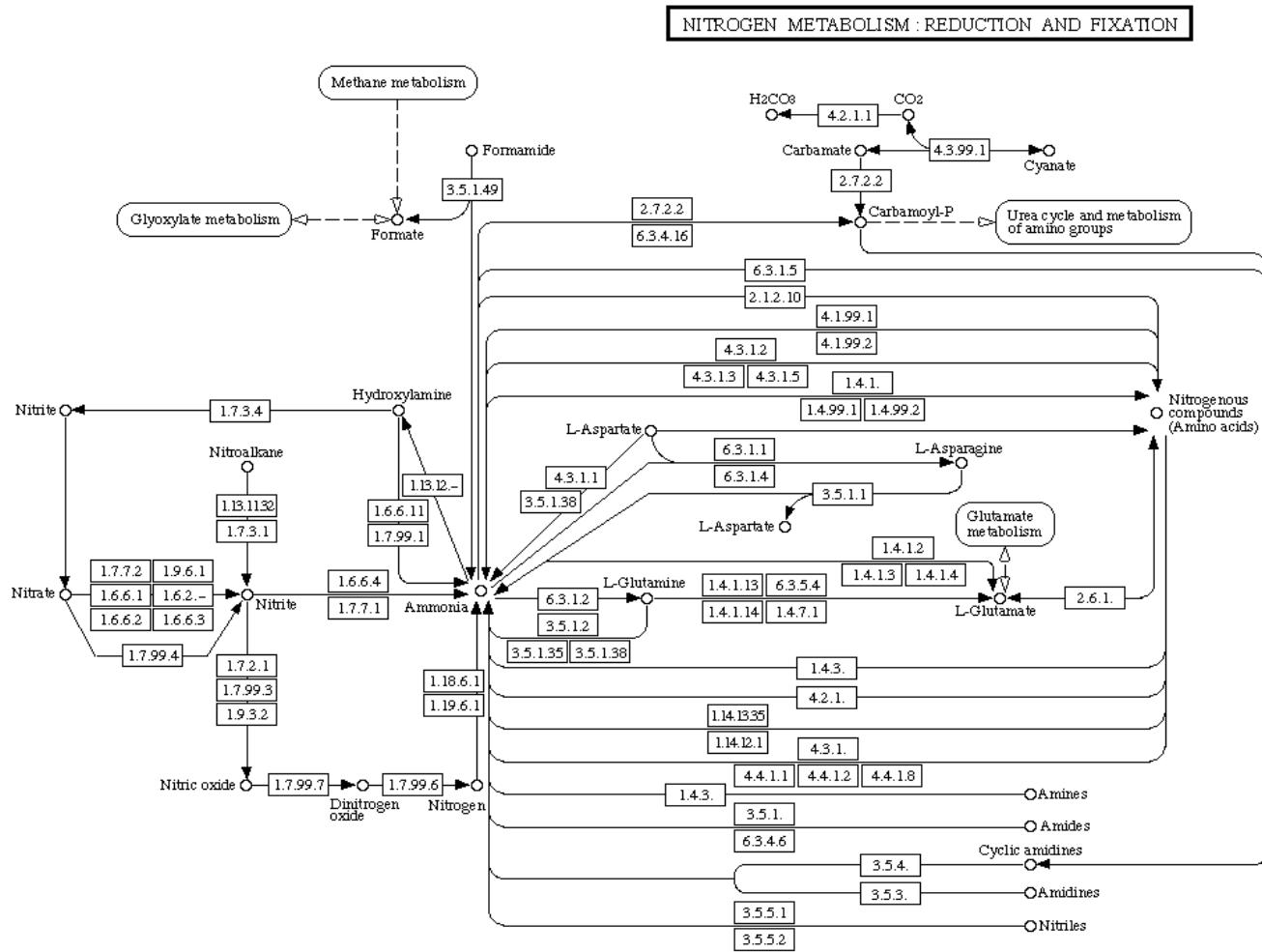
99	1+2+32+64 99	1	map	+	map	-	y
147	0+1+2+16+128 147	2	map	-	map	+	y

83	1+2+16+64 83	1	map	-	map	+	y
163	1+2+32+128 163	2	map	+	map	-	y

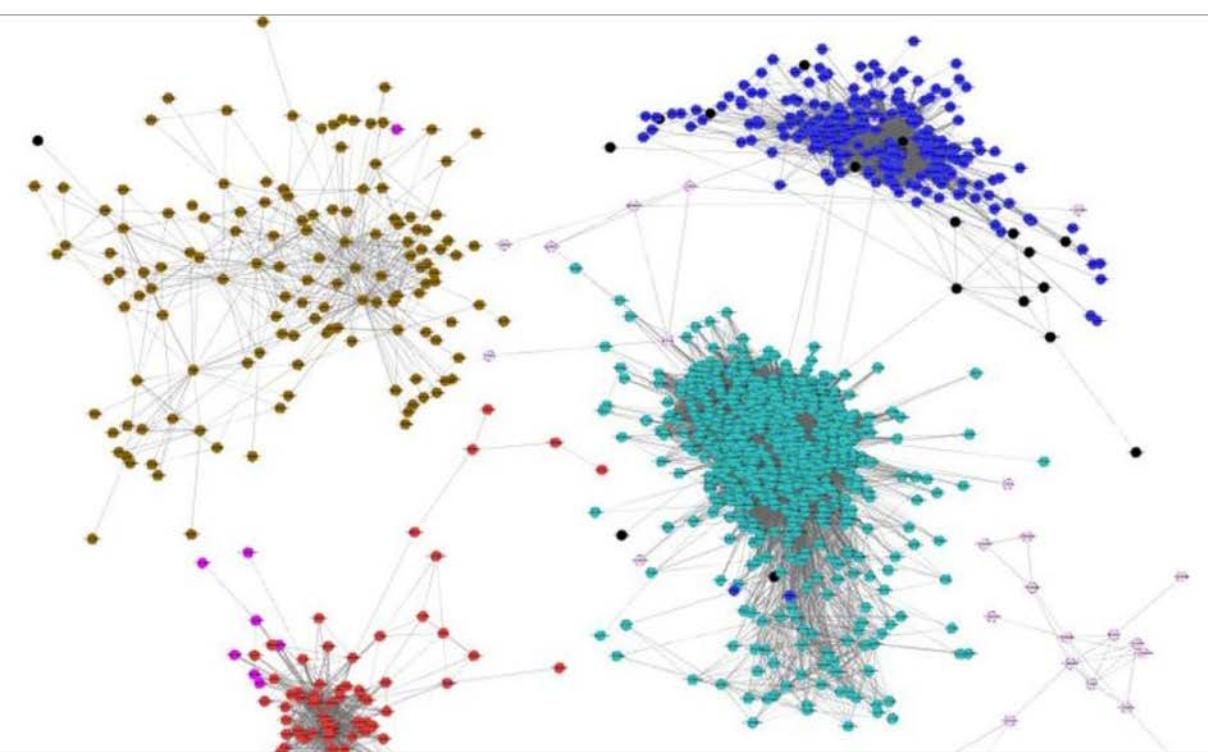
Public databases

- Access public data is key to bioinformatics analysis
 - We can't survive without pubmed
 - Any closely related species to your working species?
 - Any additional experimental data?
 - Any functional annotation to your SNP?
- Remember to deposit your own data to contribute

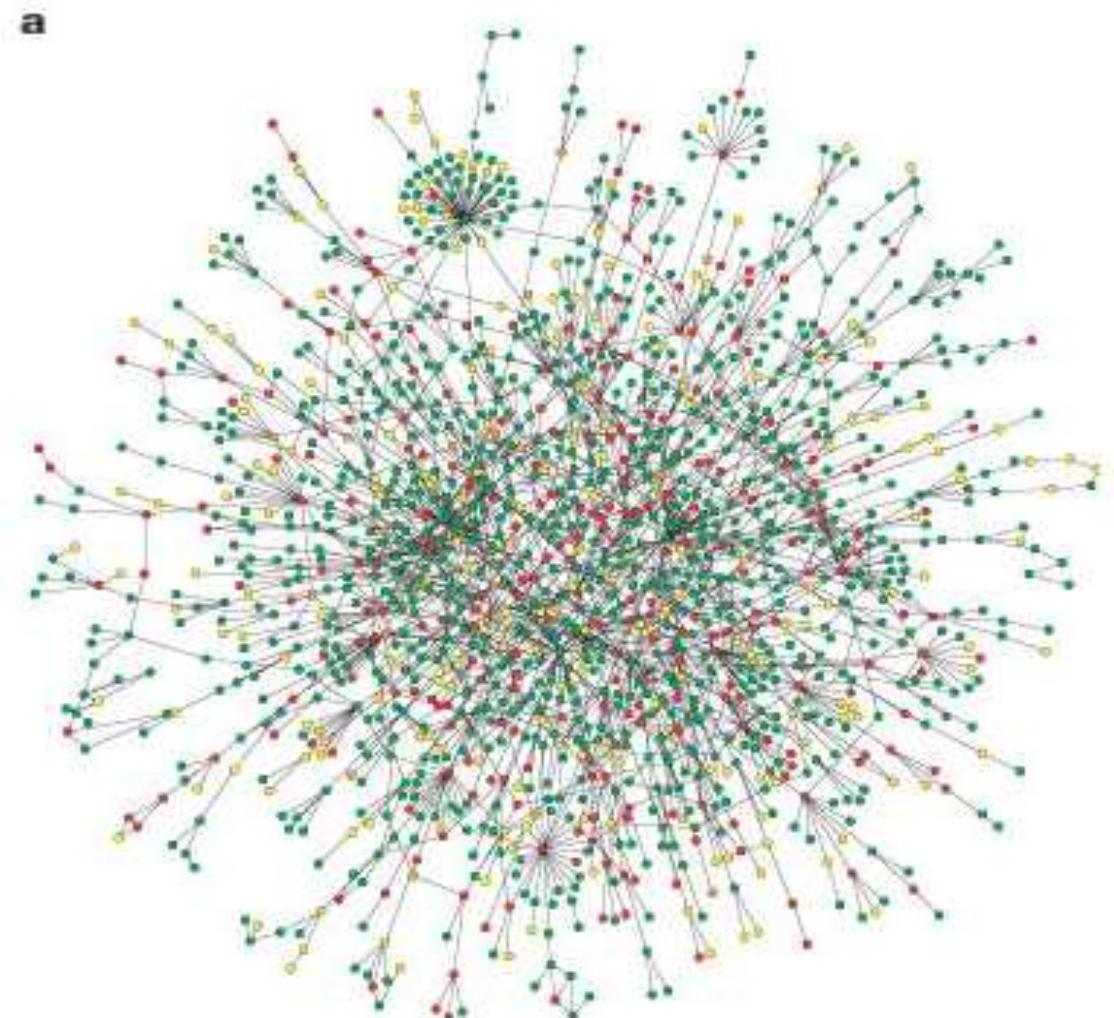
KEGG: Kyoto Encyclopedia of Genes and Genomes



Importance of networks in biology



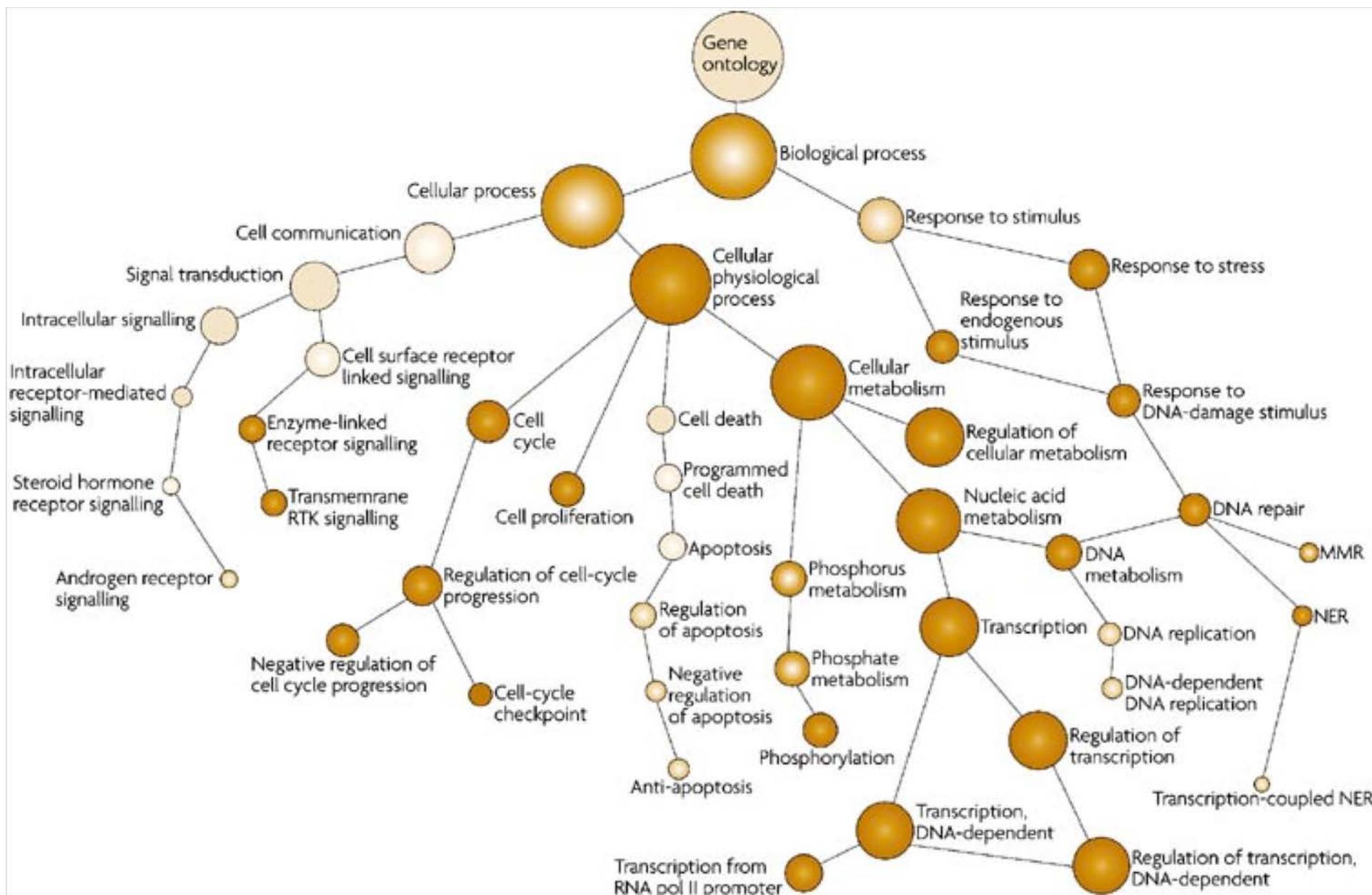
Gene interaction networks



Protein interaction network

Genome Ontology

- Biologists are fun:
 - “sonic hedgehog”
 - RING domain = Really Interesting New Gene
 - The ken and barbie gene
- An attempt to unify the names and functions across all species
- Genome Ontology uses a single 3 part system
 - Molecular function (specific tasks)
 - Biological process (broad biological goals - e.g cell division)
 - Cellular component (location)



Analysis and interpretation

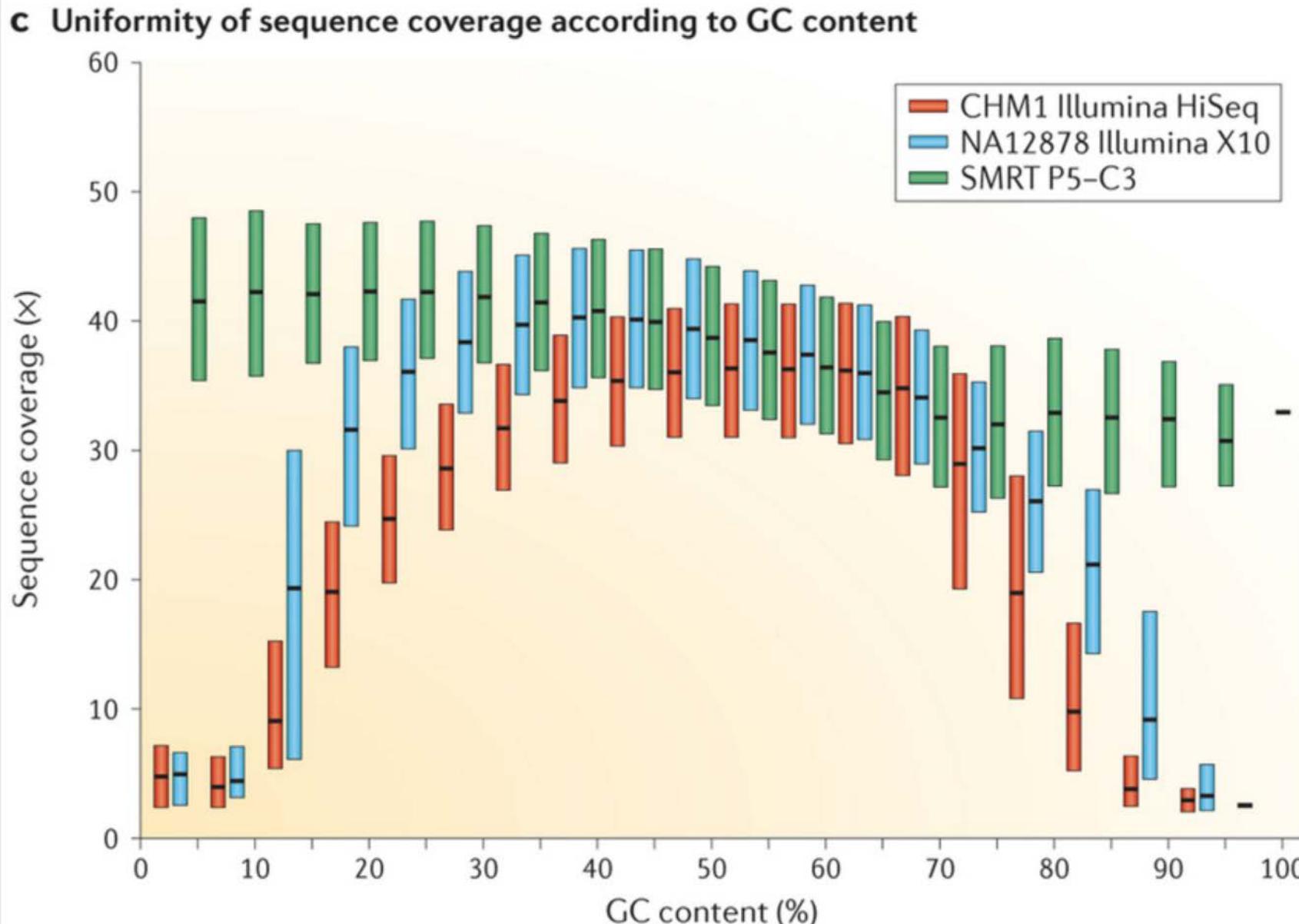


Is your data good enough?



<https://sequencing.qcfail.com>

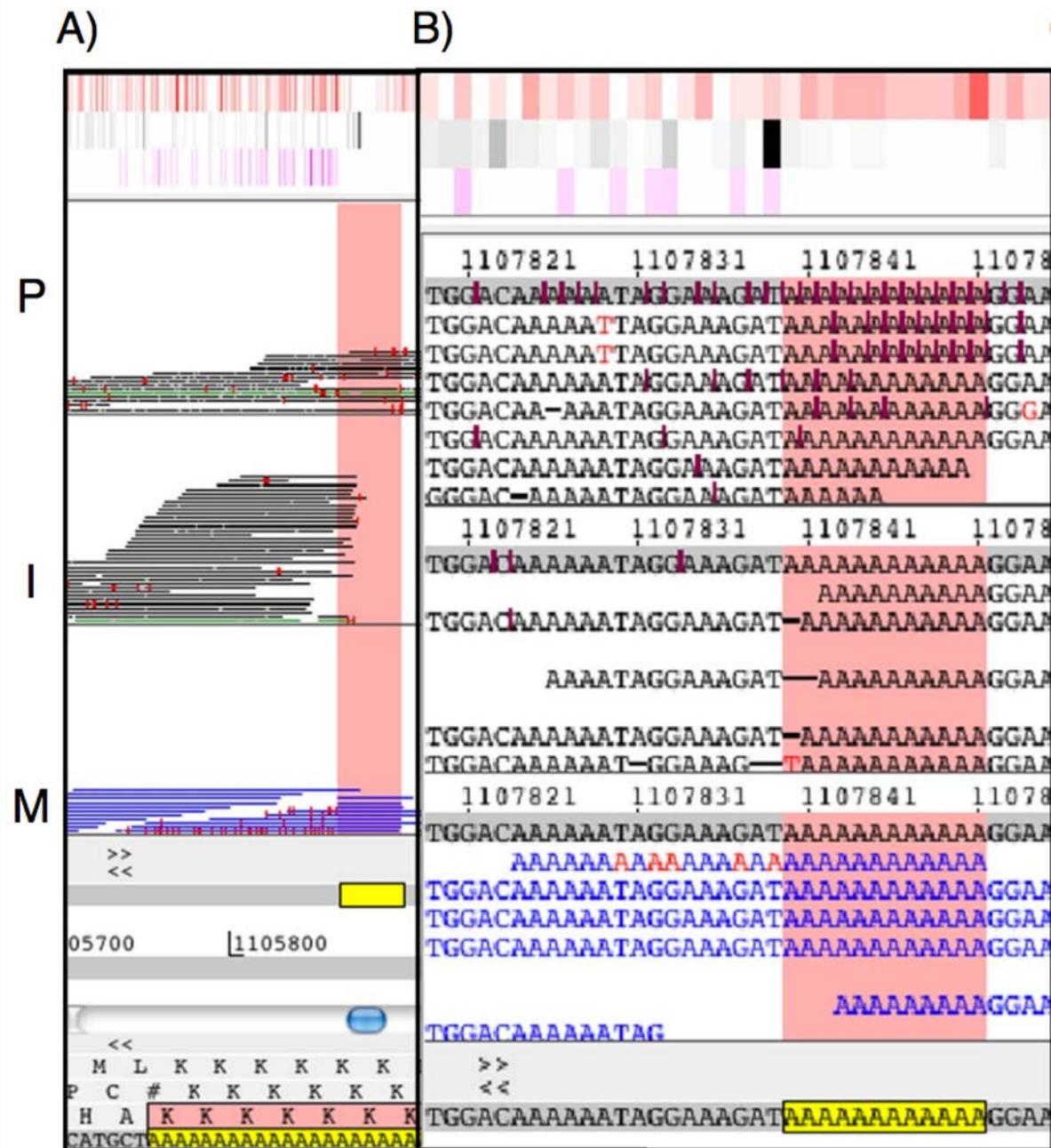
Sequencing Biases



Sequencing Errors

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

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



Three situations you are most likely to encounter

Genome reference is available (for example, humans):

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

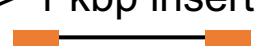
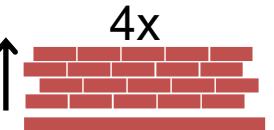
Genome reference is NOT available

- **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?

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---

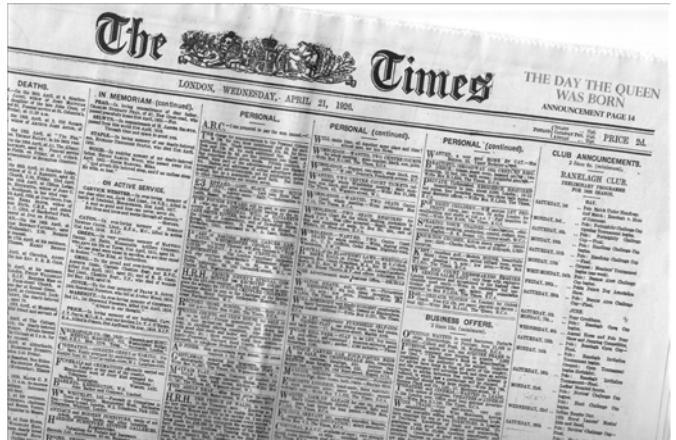
| | | | | |

---GAAATGAAAAGG

insertions / deletions (indels) mismatches

Which alignment is better?

Assembly



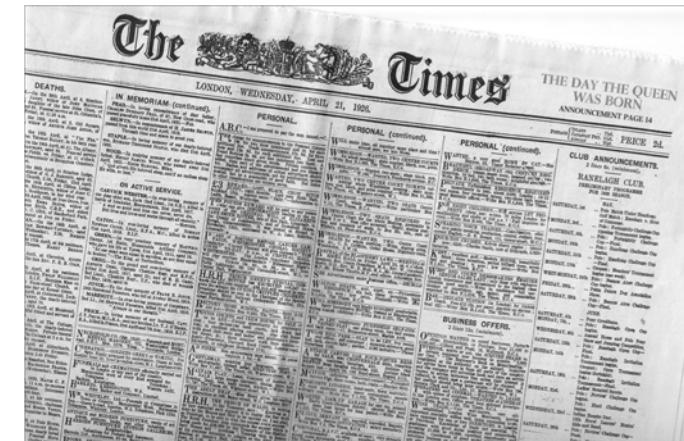
Genome
(3.000.000 letters)

Sequencing



Reads
(50-500 letters each)

Assembly

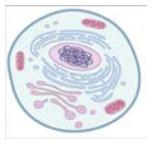


Genome
(3.000.000 letters)

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



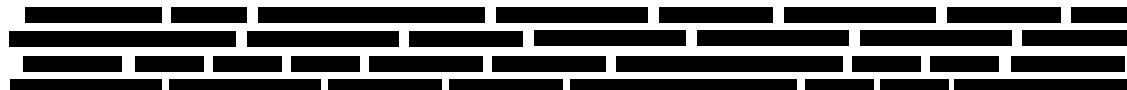
Assembly



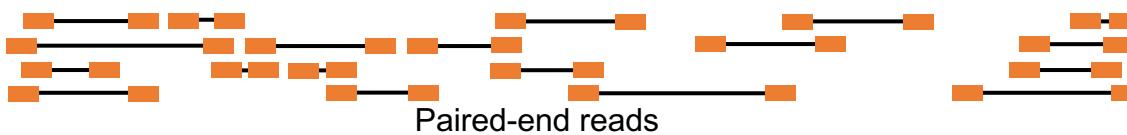
Genome



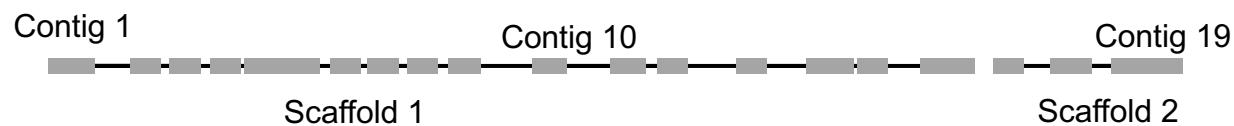
Fragment



Sequence



Assemble



After assembly

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

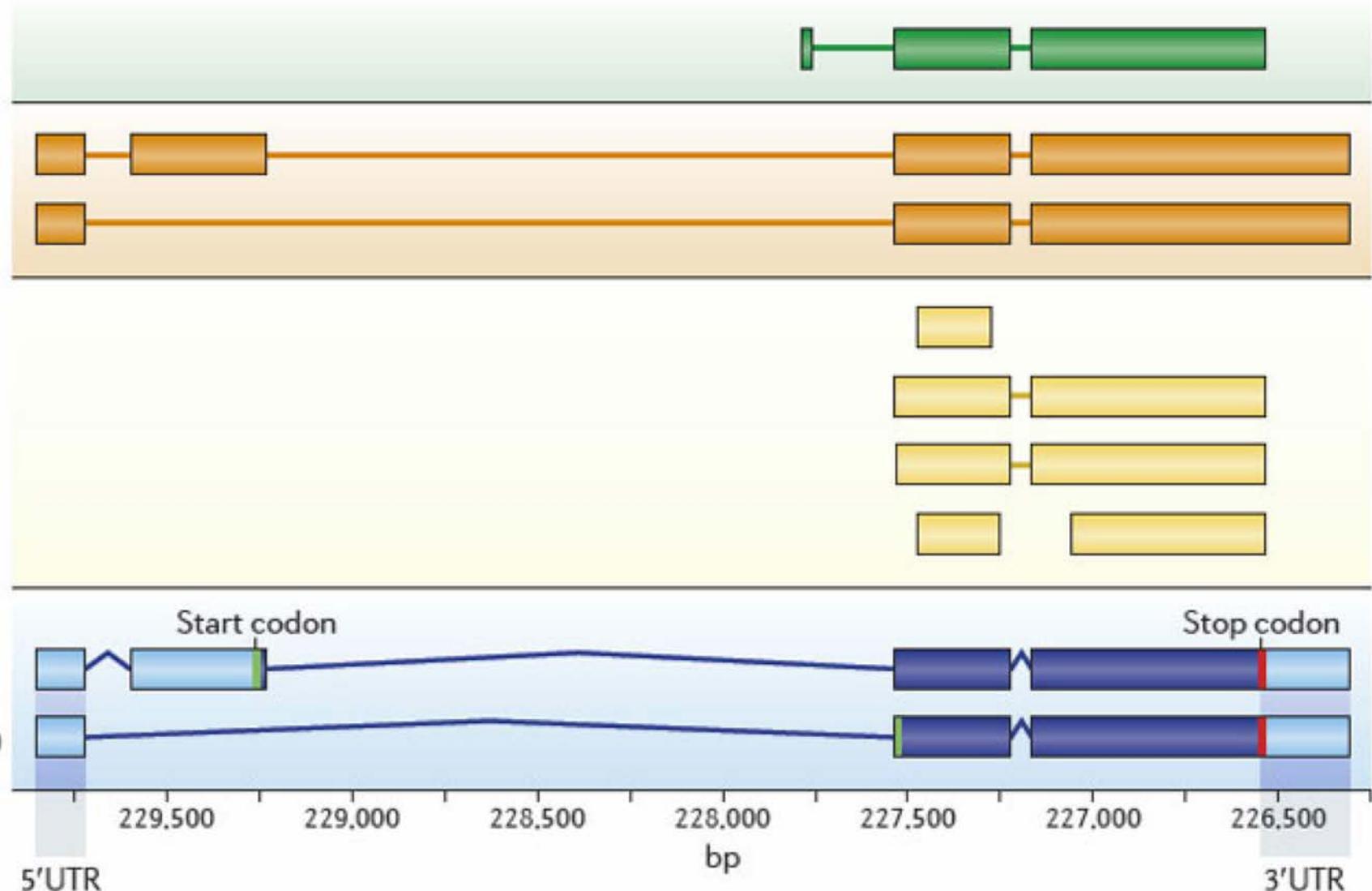
Annotation

Gene prediction
(SNAP)

de novo

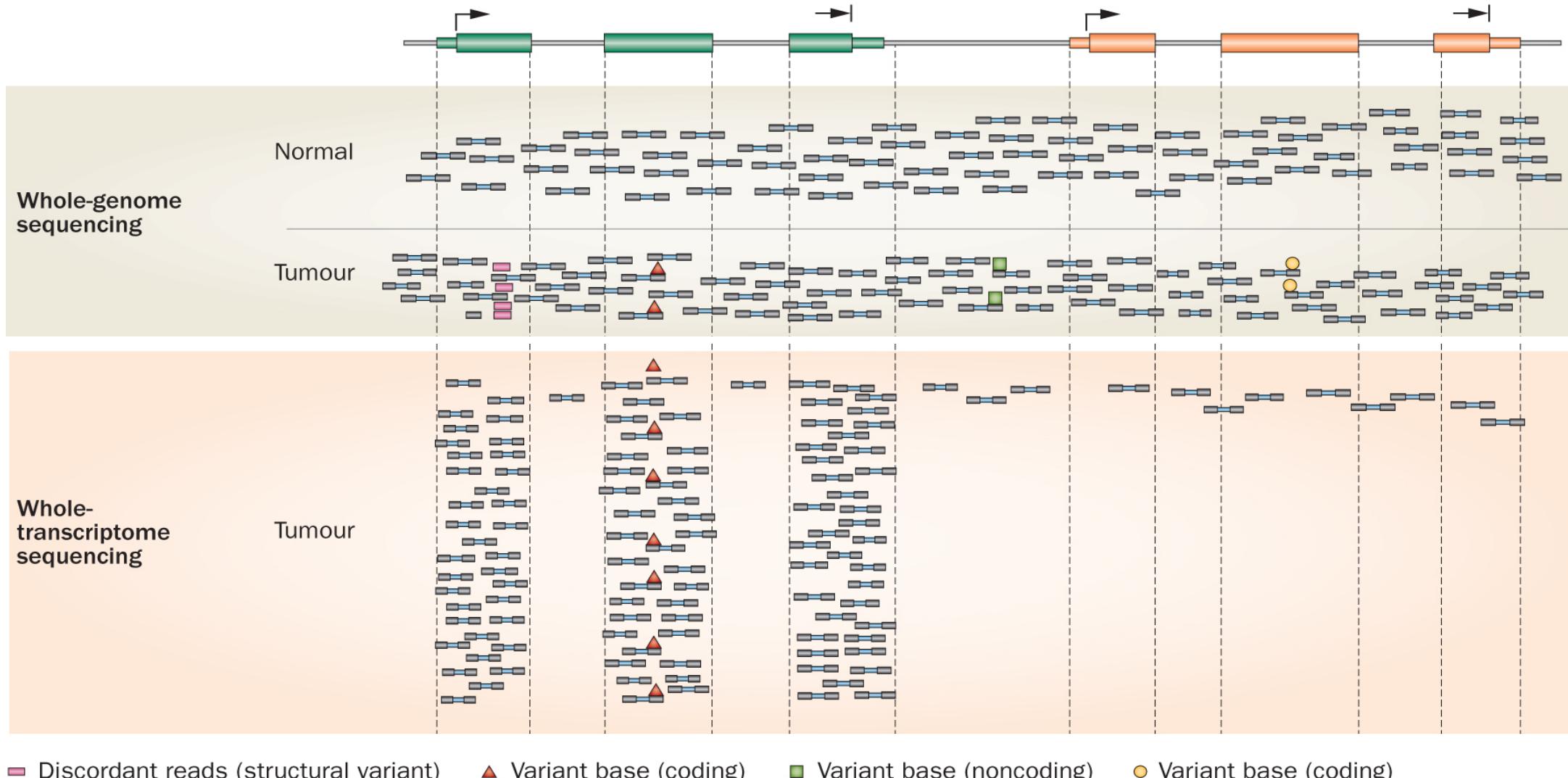
mRNA or EST evidence
(Exonerate)

Now with RNAseq



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

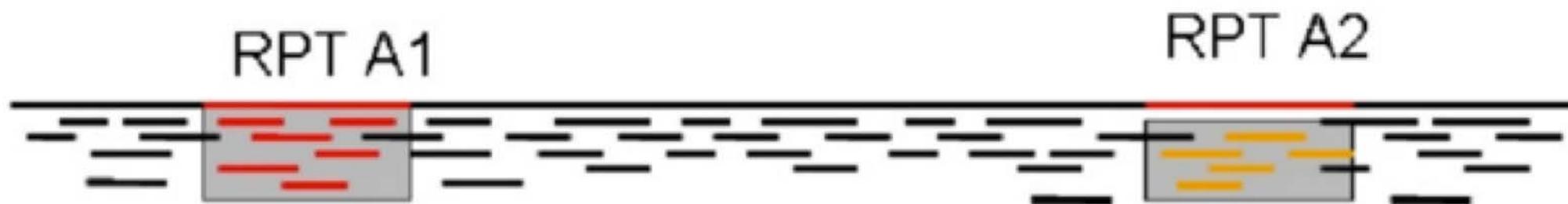


Figure 5. Two copies of a repeat along a genome. The reads colored in red and those colored in yellow appear identical to the assembly program.

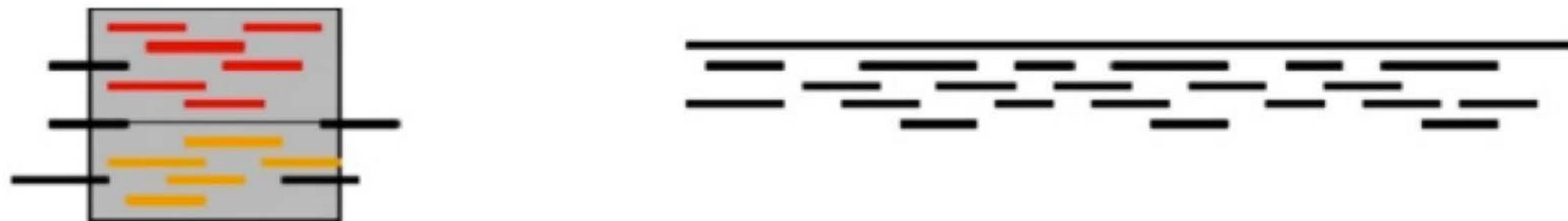
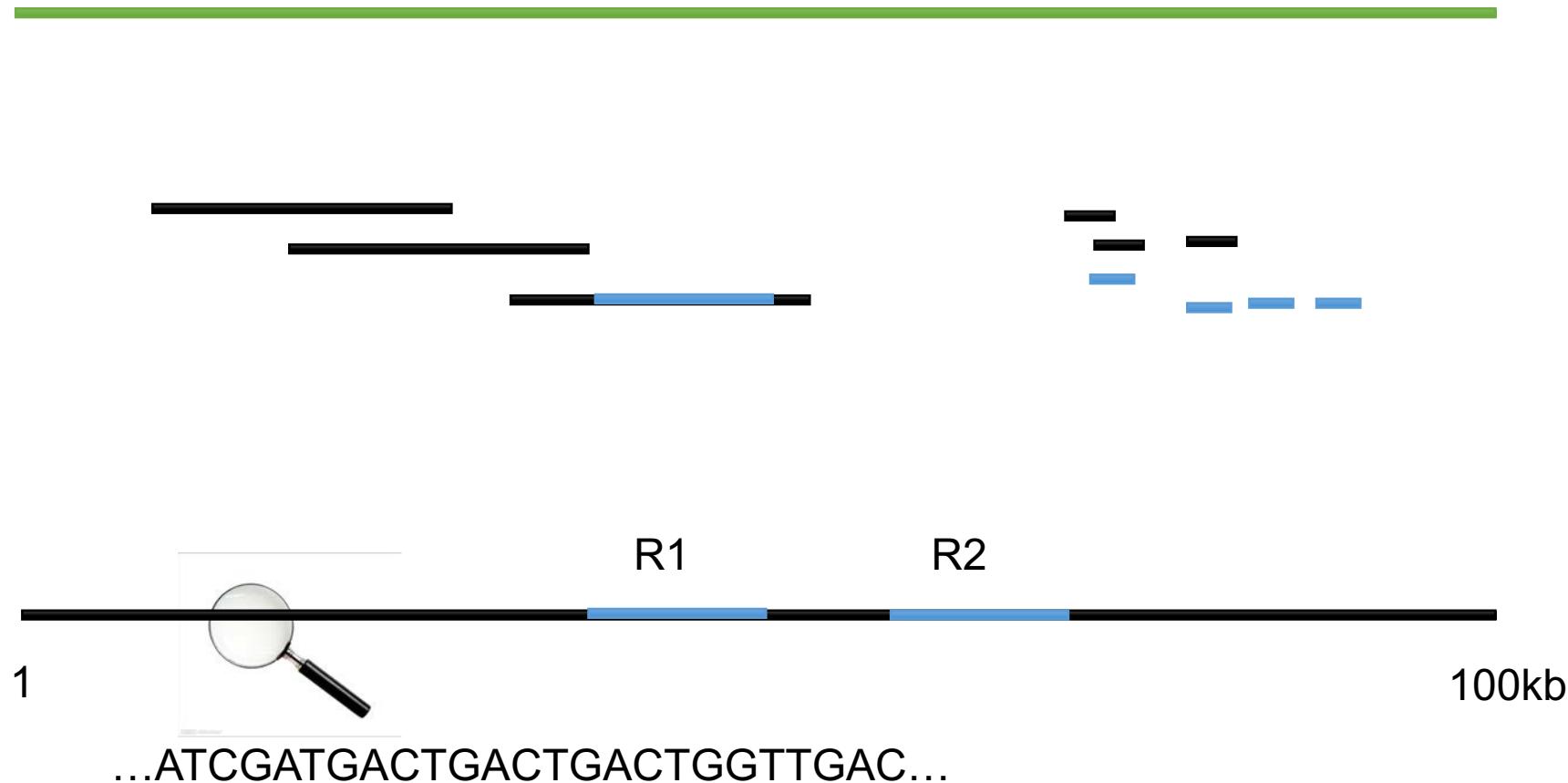
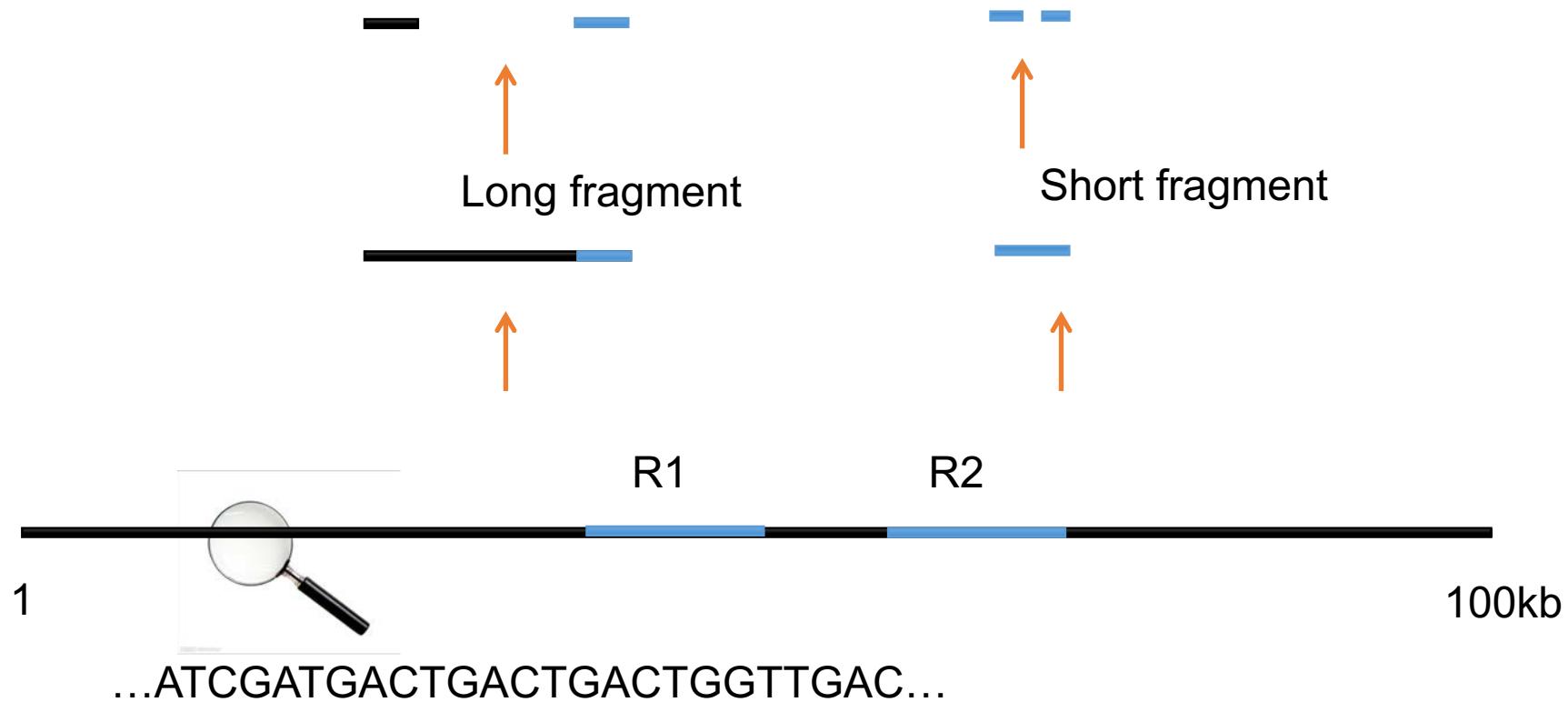


Figure 6. Genome mis-assembled due to a repeat. The assembly program incorrectly combined the reads from the two copies of the repeat leading to the creation of two separate contigs

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^CACTGACTGACTGGTTGAC

Homozygous? Heterozygous?

...ATCGATGACTGACTGACTGGTTGAC...

reference

Case studies

Classical genetics

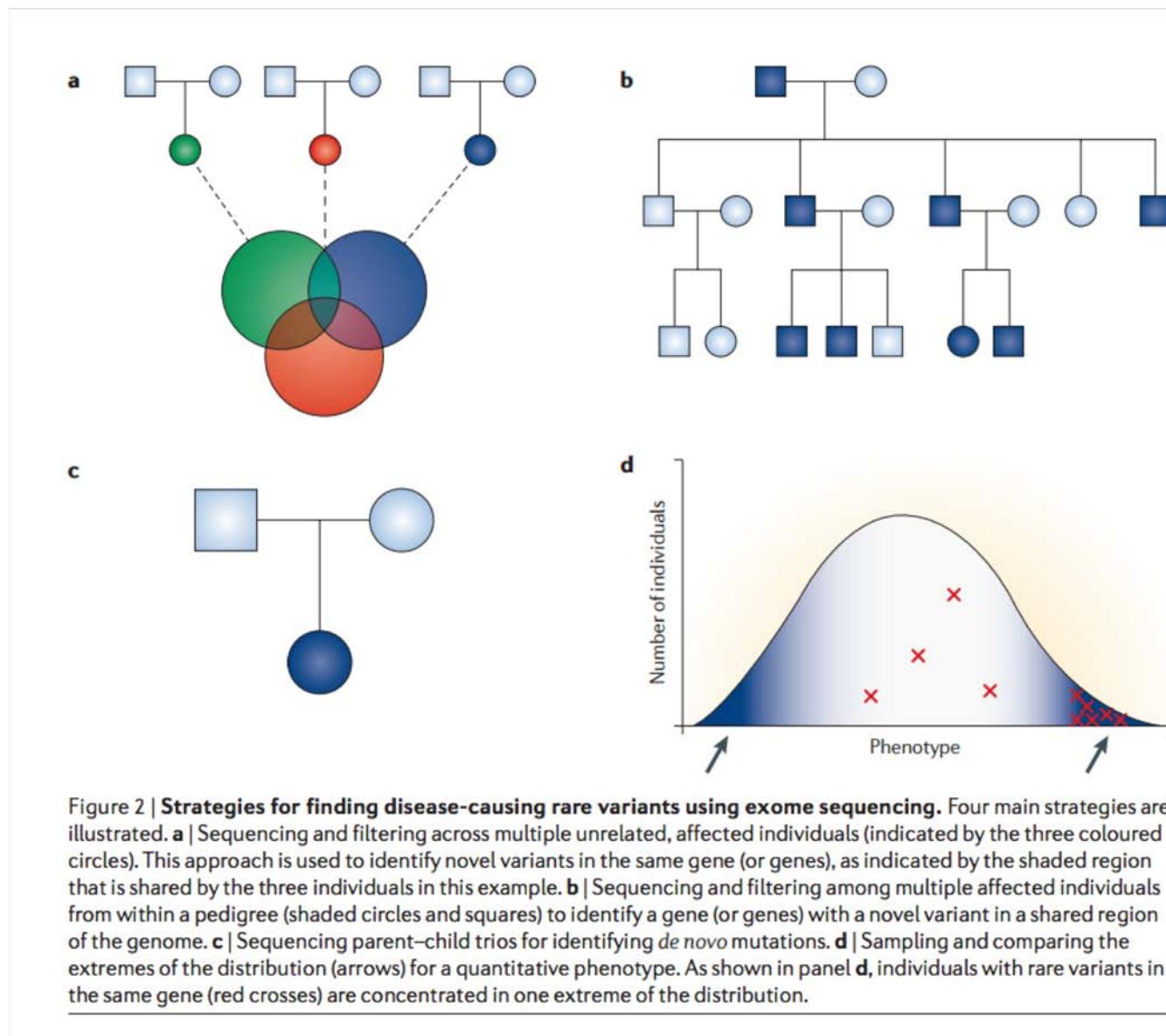
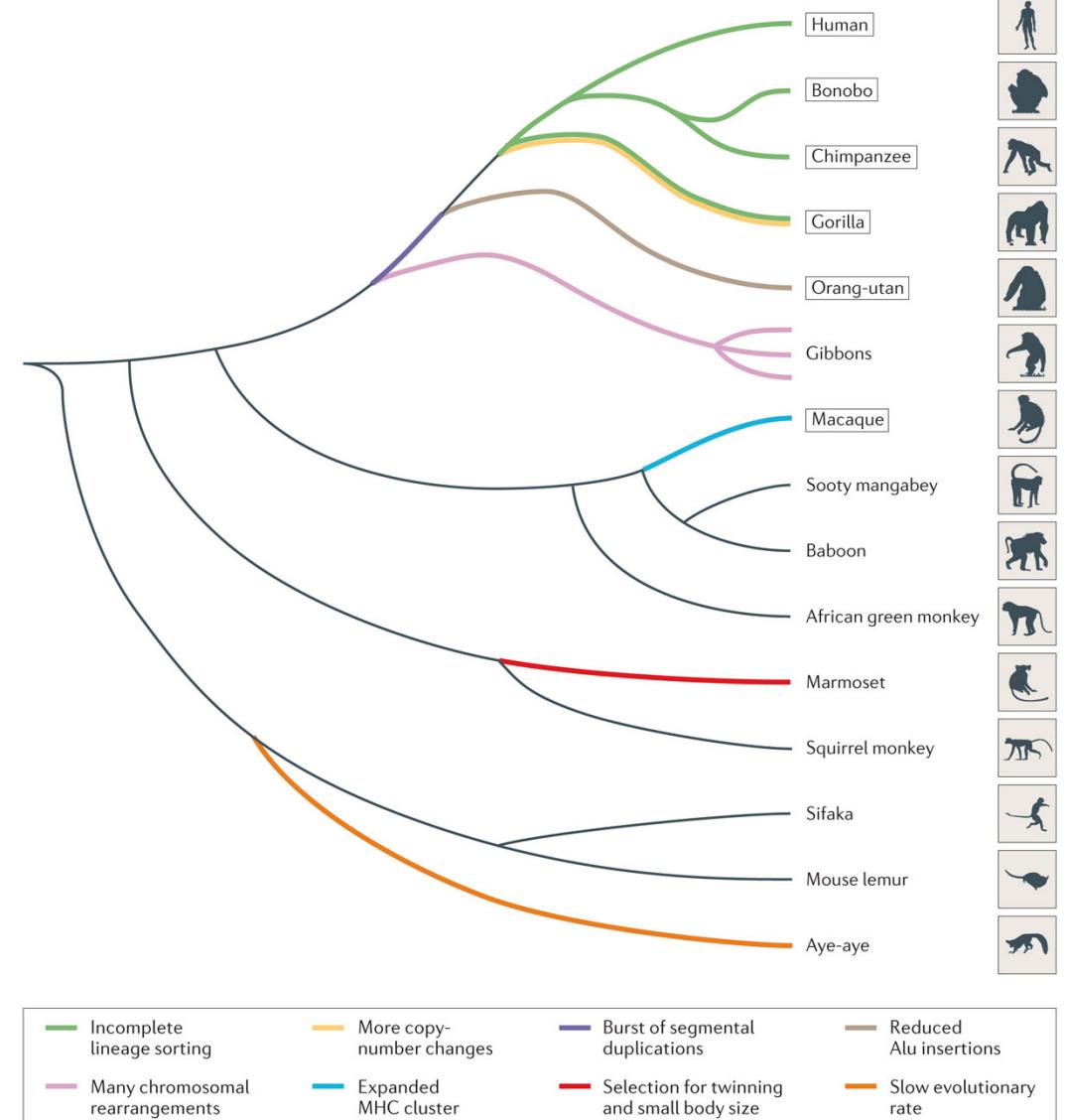
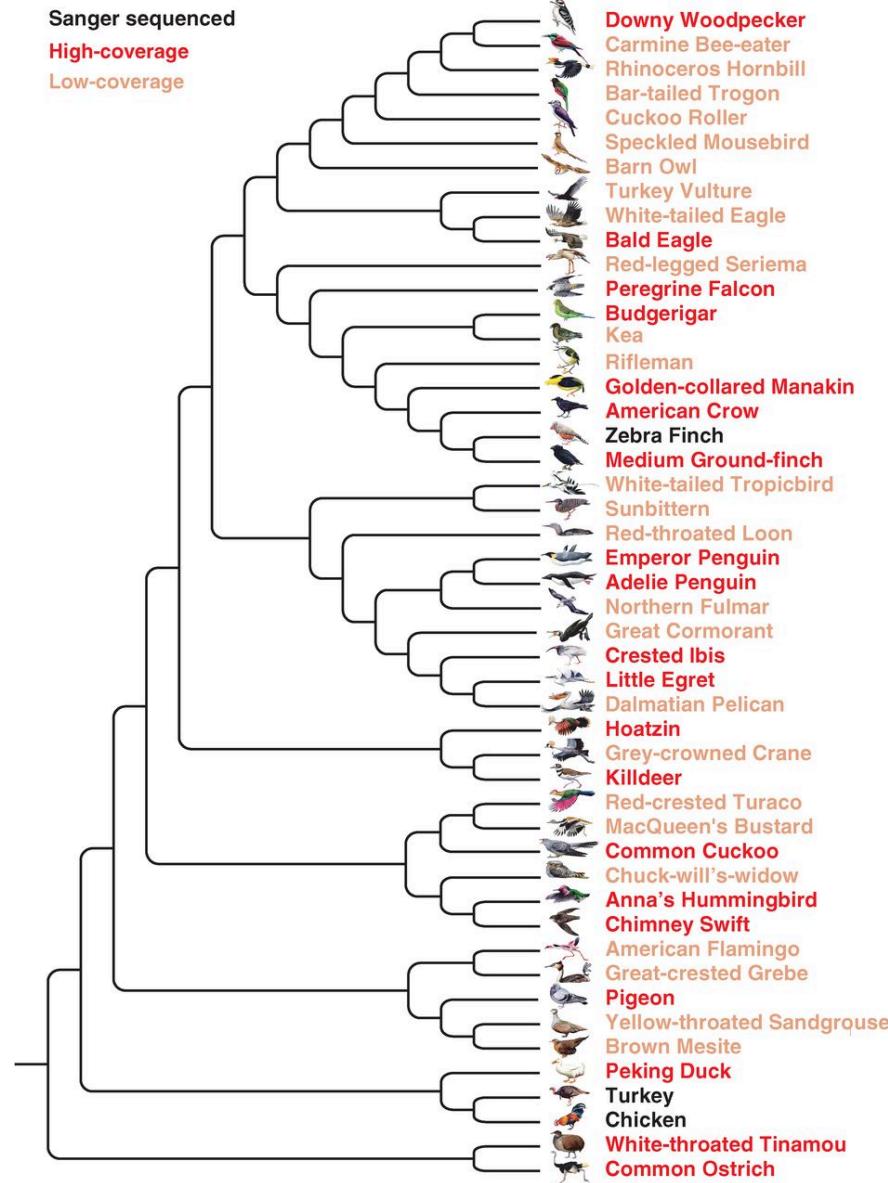


Figure 2 | Strategies for finding disease-causing rare variants using exome sequencing. Four main strategies are illustrated. **a** | Sequencing and filtering across multiple unrelated, affected individuals (indicated by the three coloured circles). This approach is used to identify novel variants in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



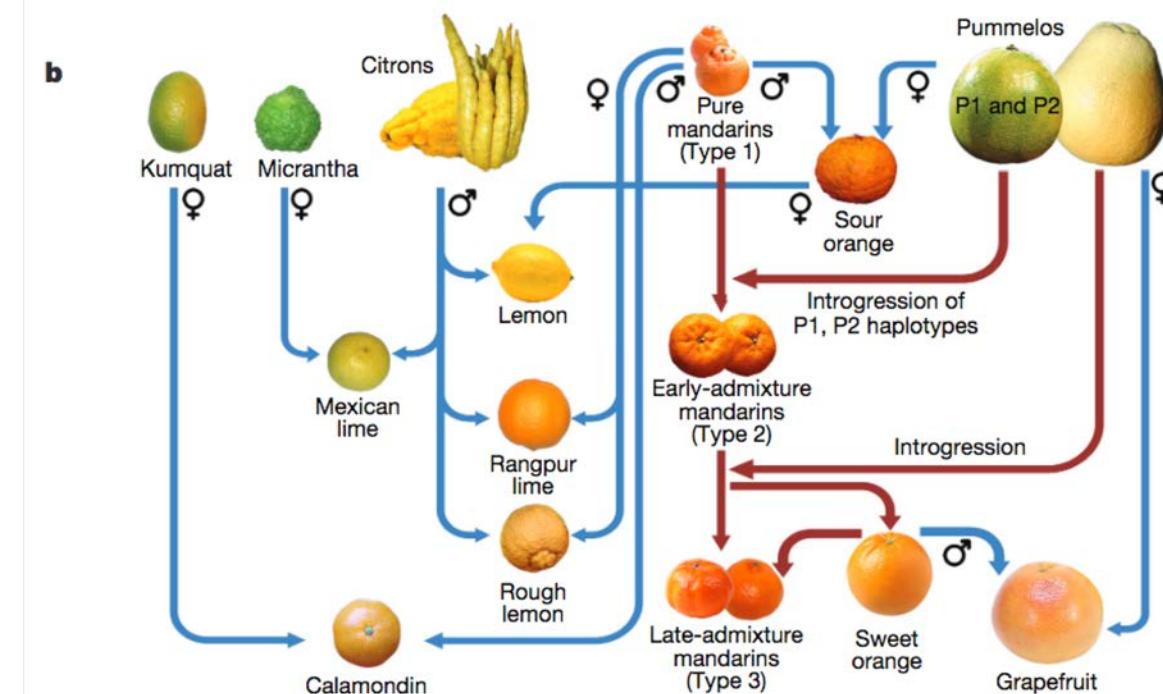
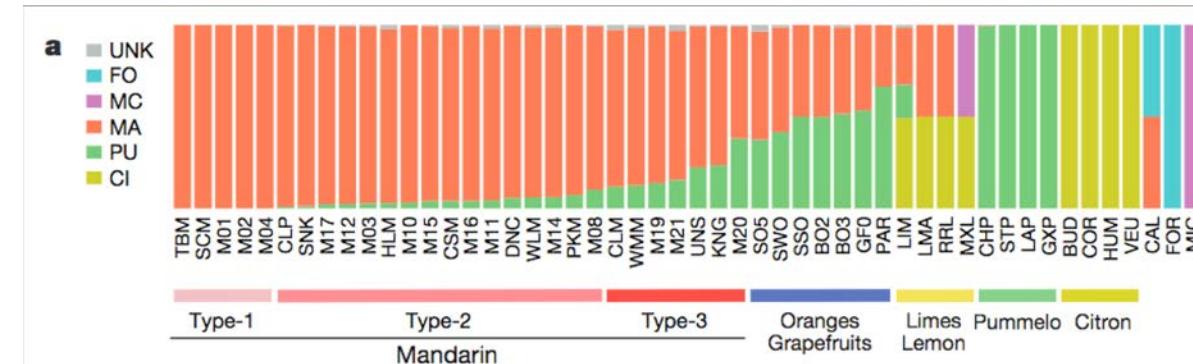
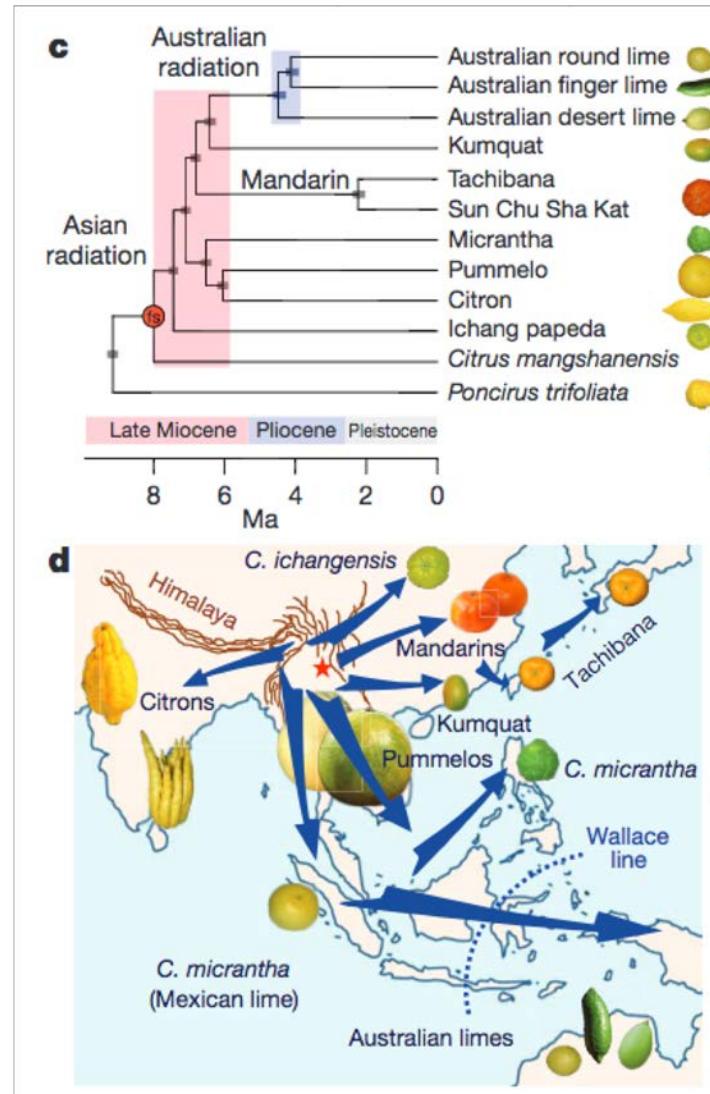
Nature Reviews | Genetics

Guojie Zhang et al. Science (2014)

Roger & Gibbs Nature Reviews Genetics (2014)

Comparative genomics

Genomics of the origin and evolution of Citrus

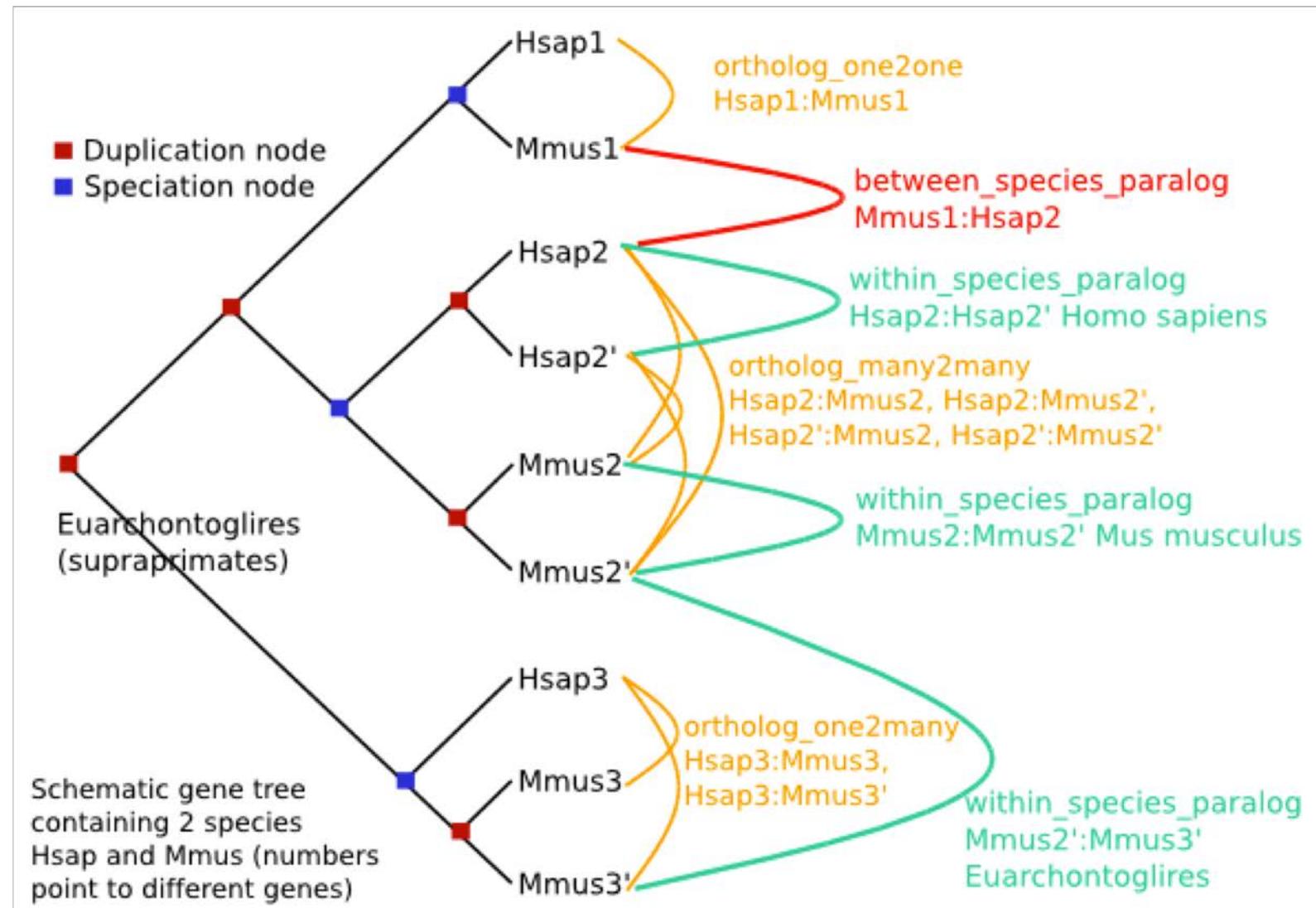


Homologs: Orthologs and paralogs

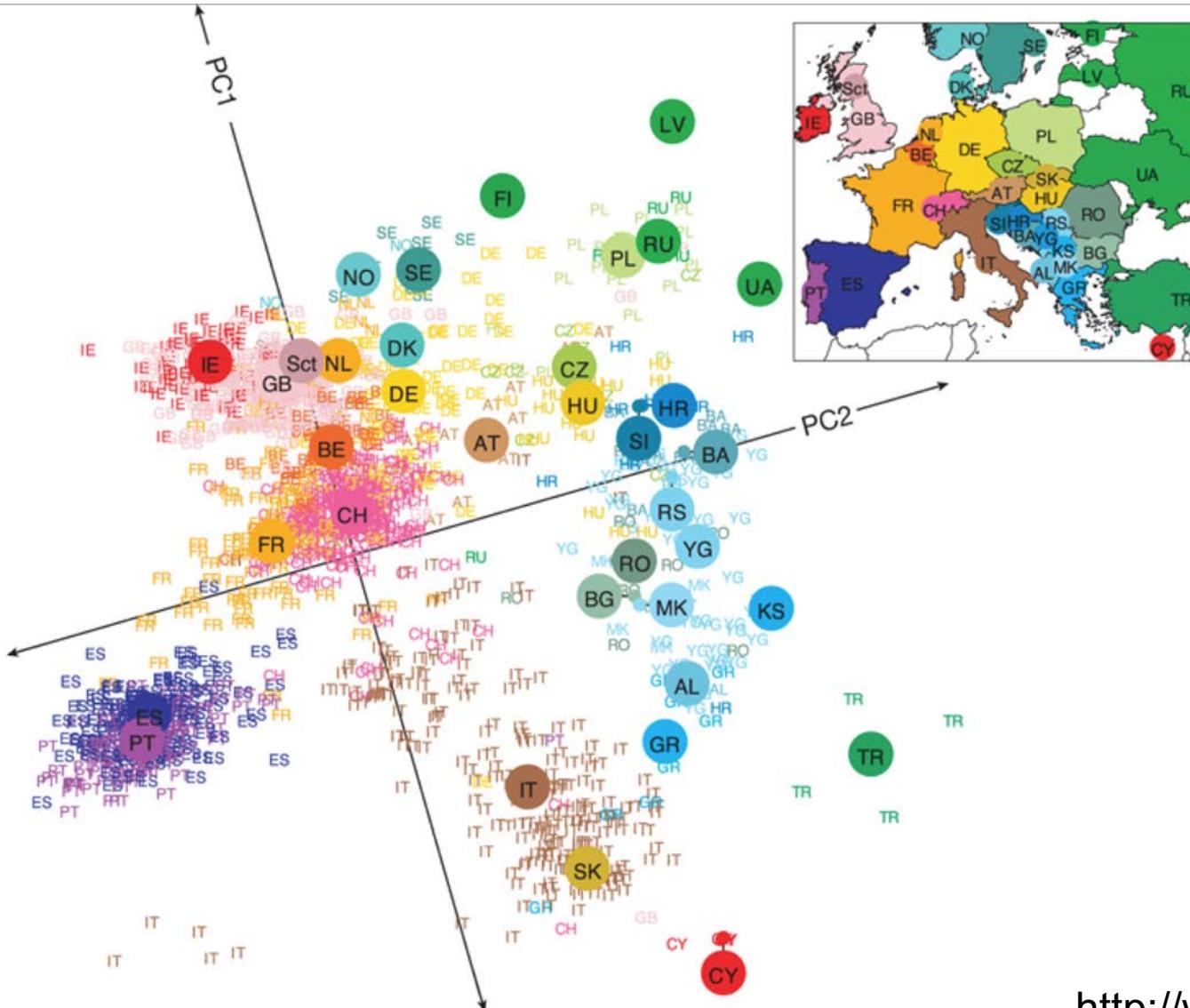
Genes in different species and related by a speciation event are defined as **orthologs**.

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

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



Population genomics

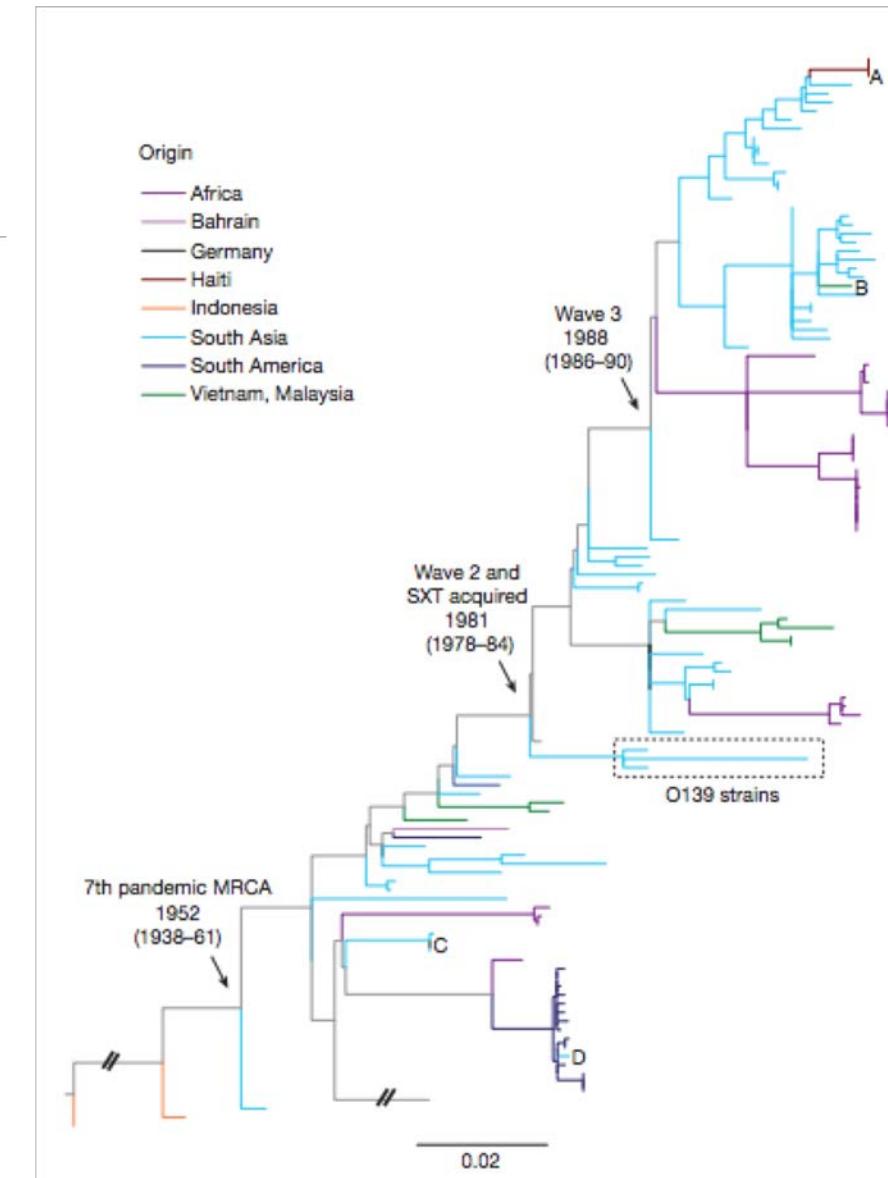
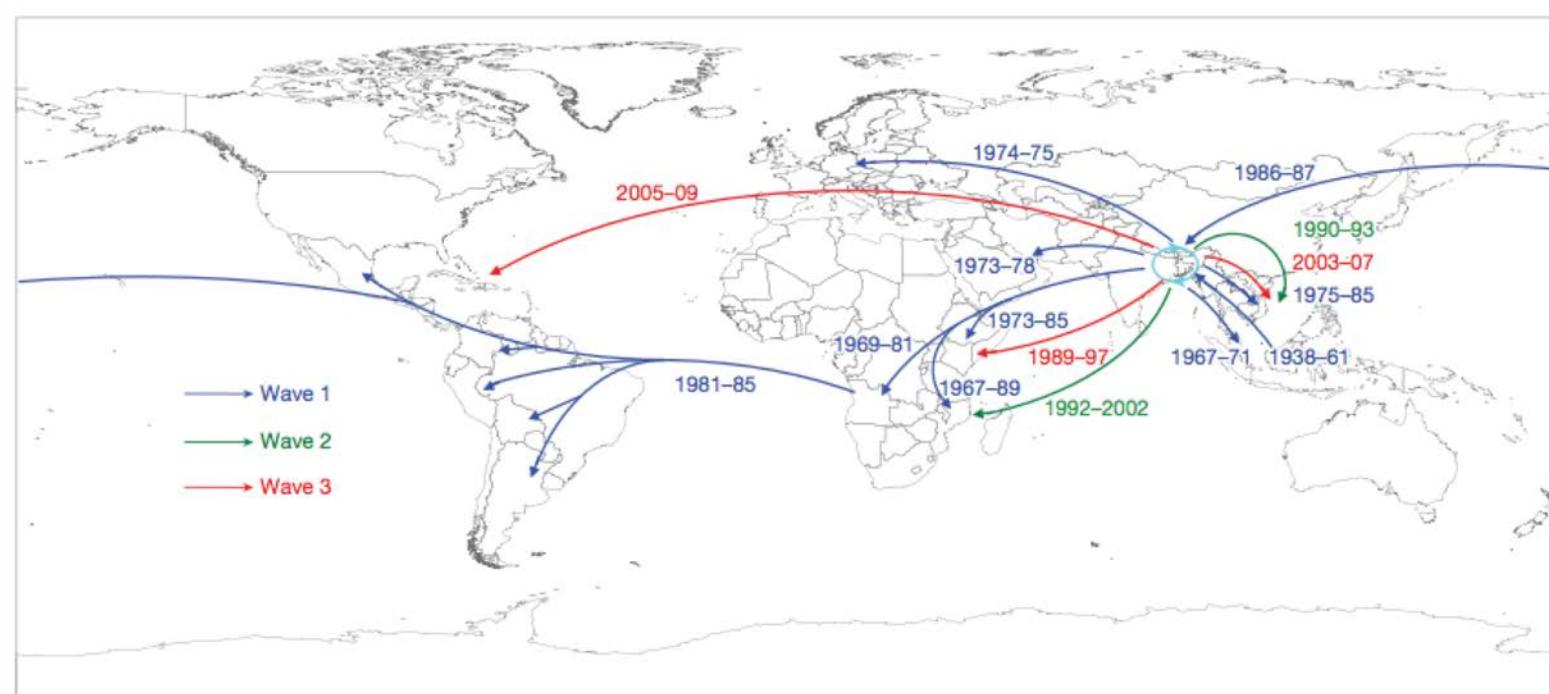


Novembre et al Nature (2008)

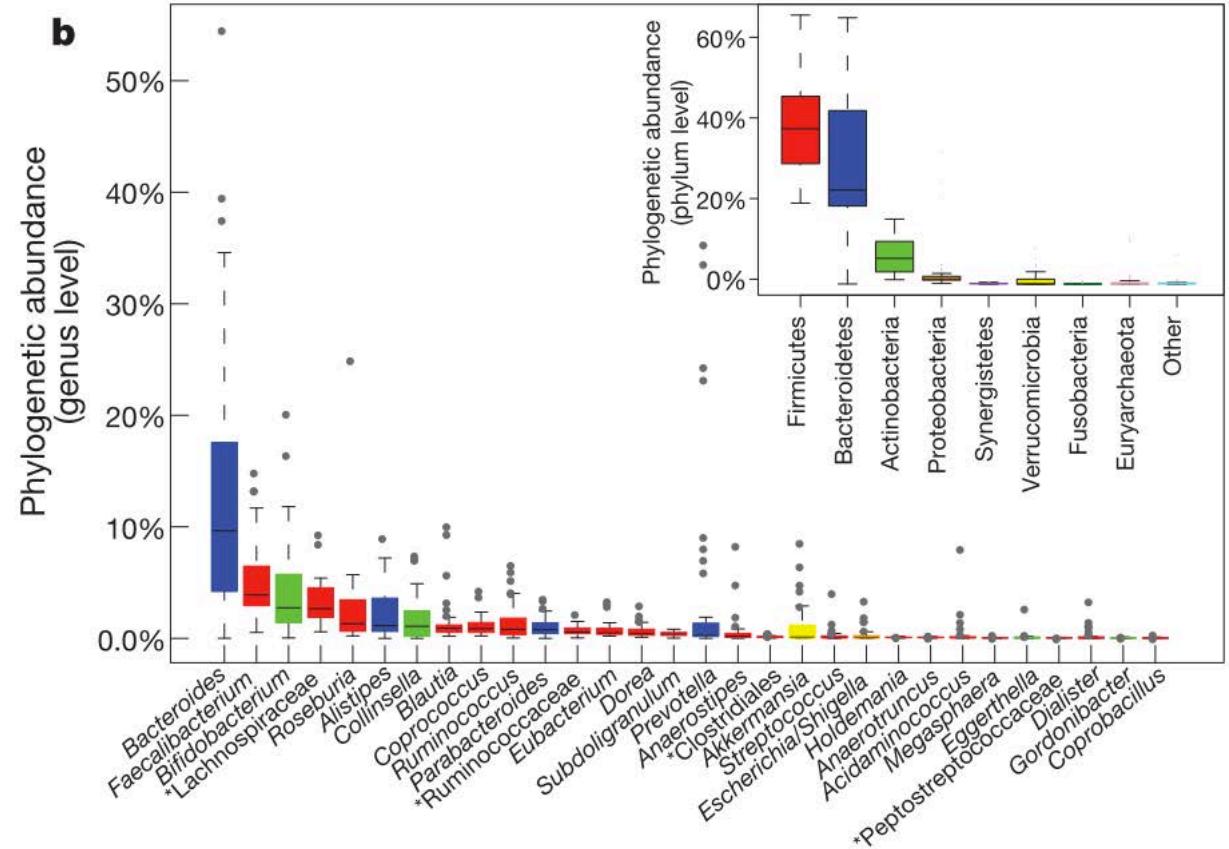
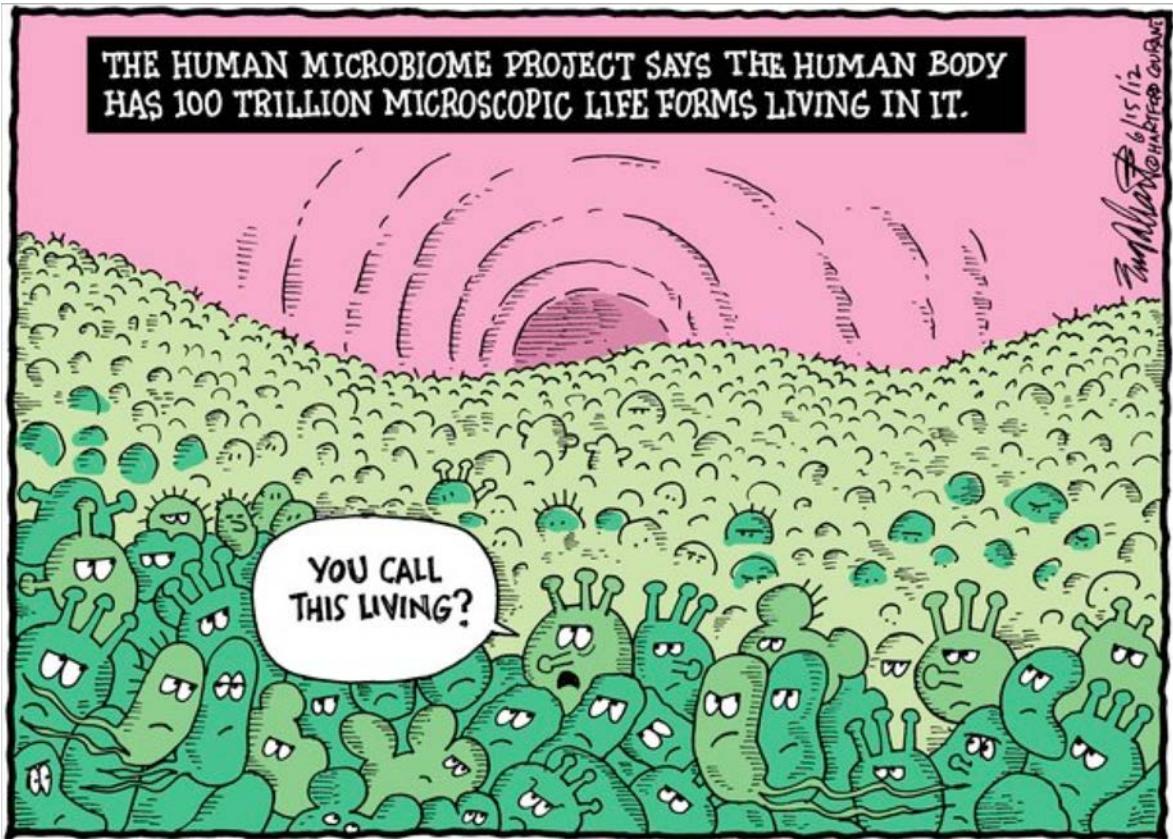


http://www.genomenext.com/casestudies_post/population-scale-analysis-genomic-samples-analyzed-from-2504-individuals-in-1-week/

Population genomics

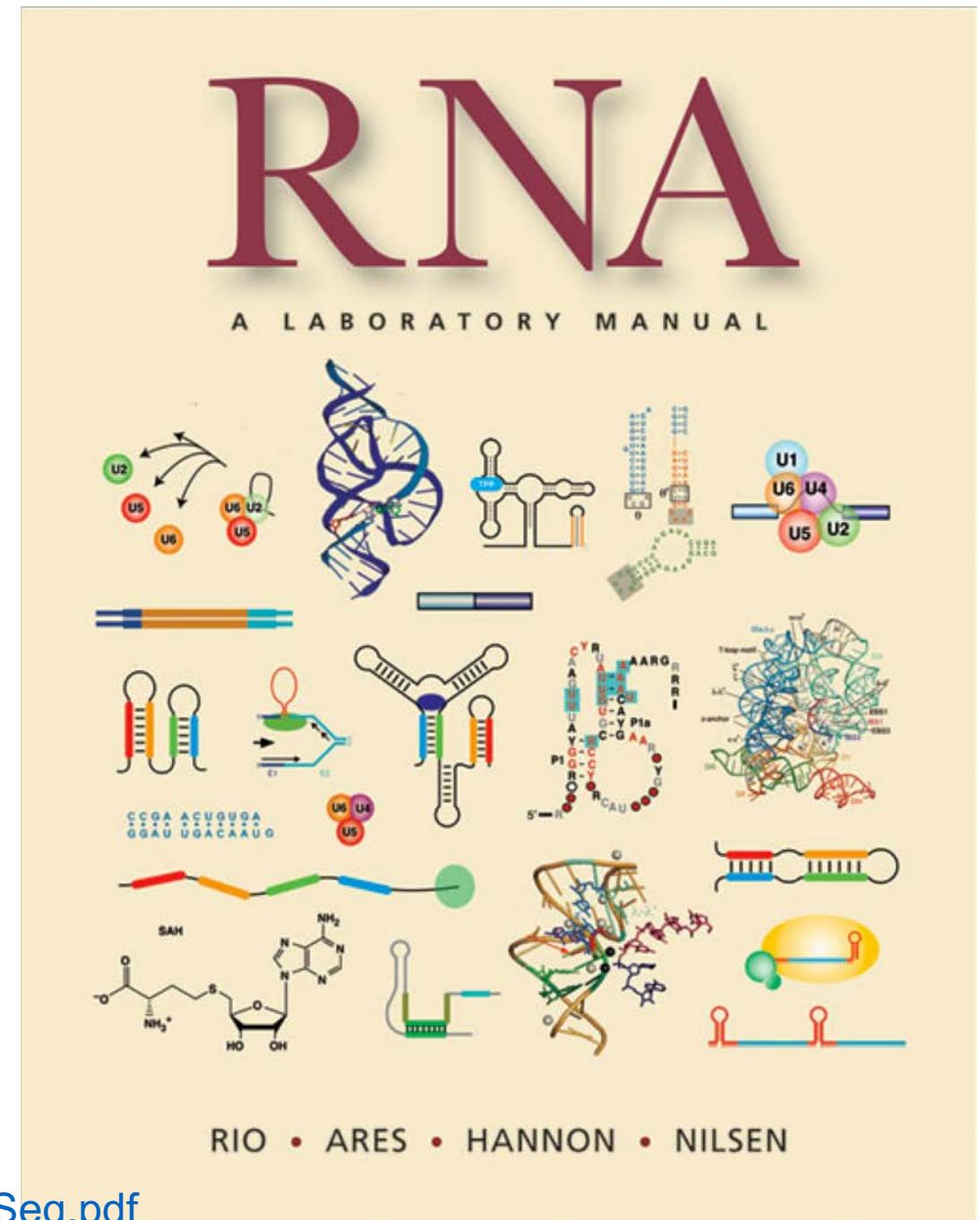
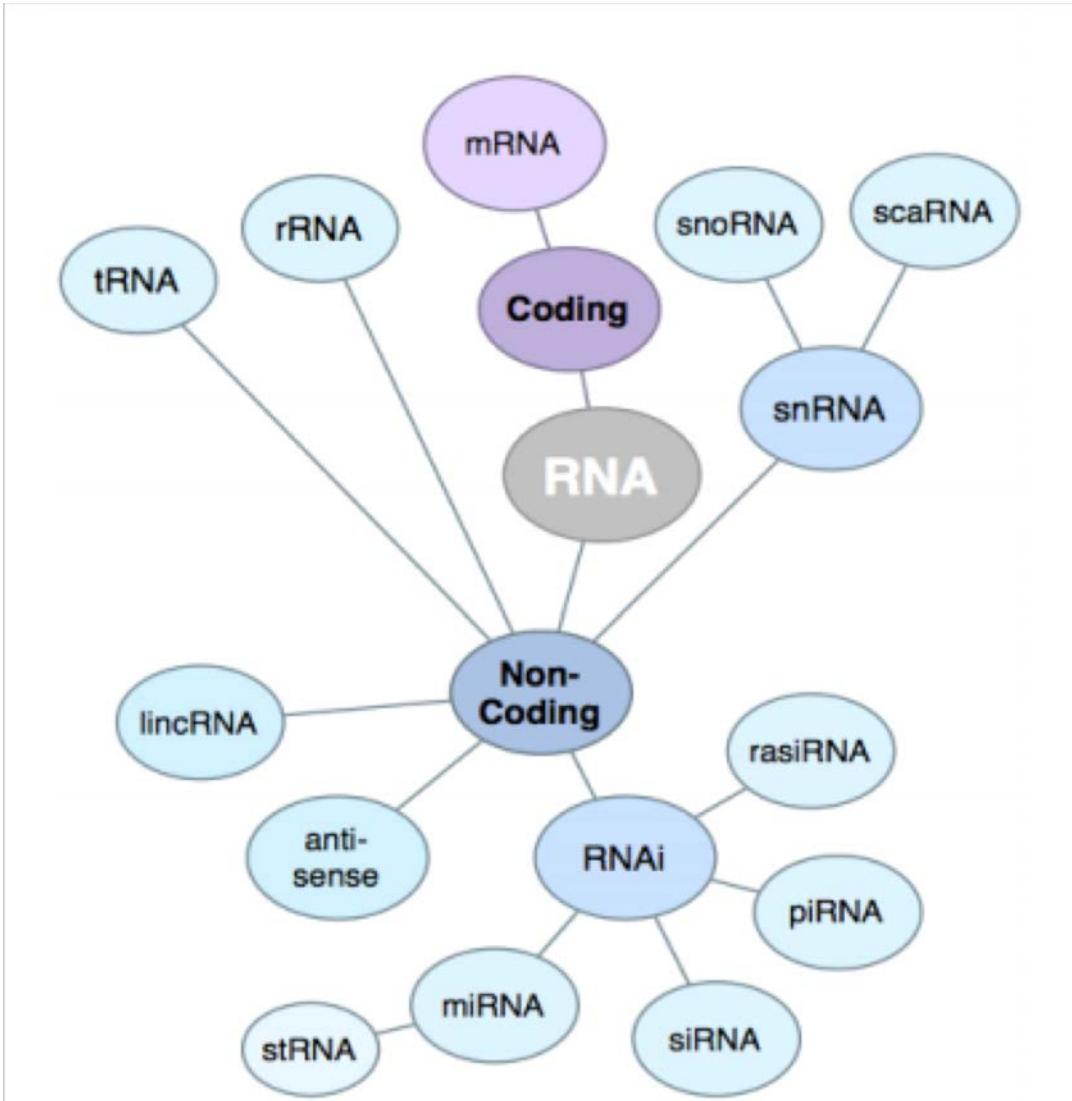


Metagenomics

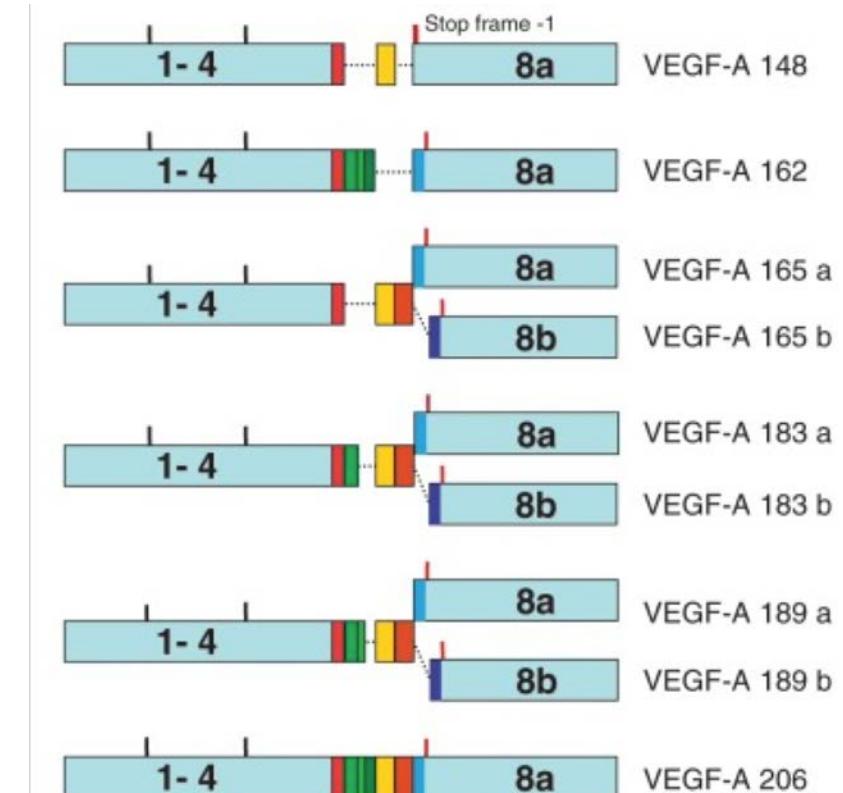
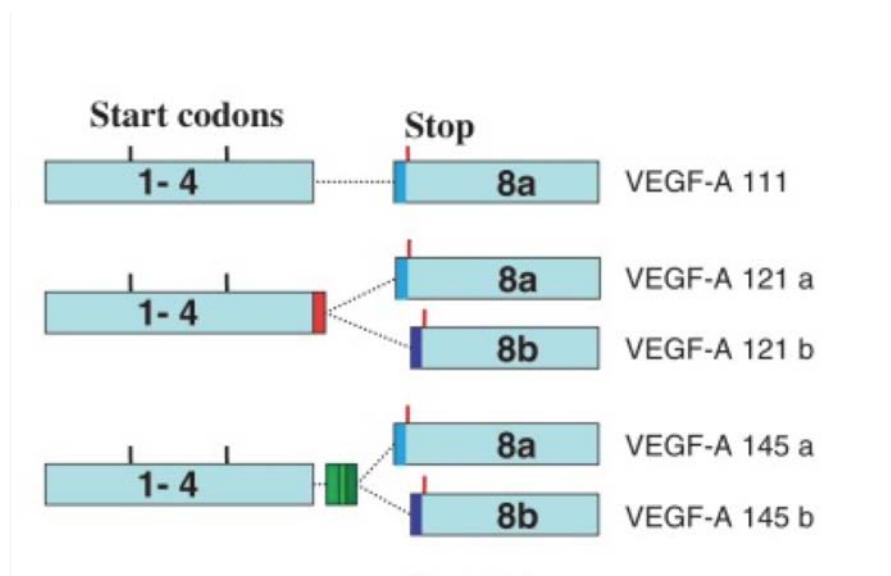
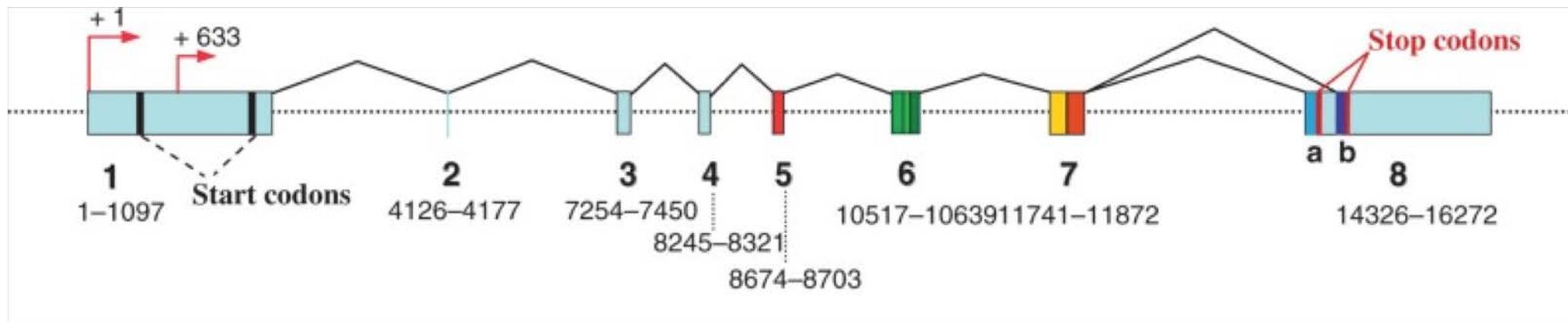


Transcriptomics / RNAseq

Types of RNA



Gene and isoforms



REVIEW

Open Access

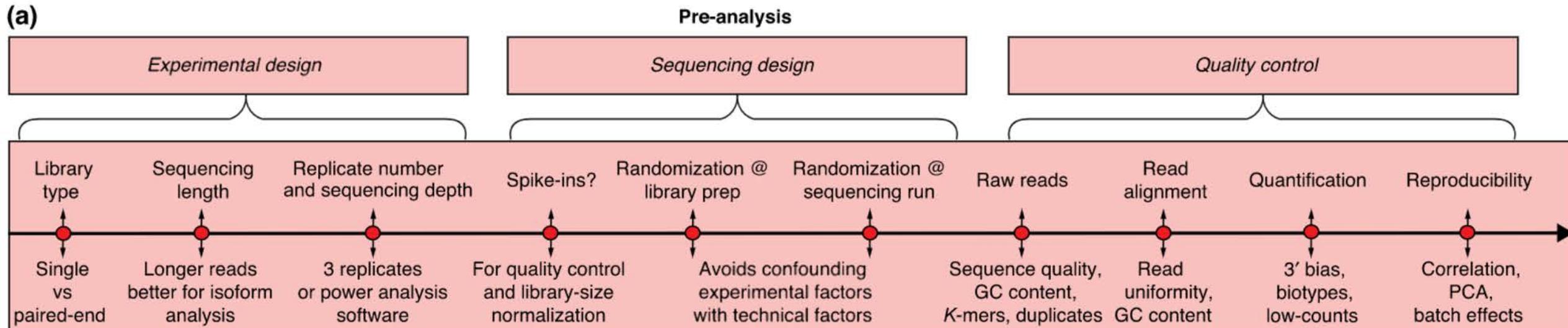


A survey of best practices for RNA-seq data analysis

Ana Conesa^{1,2*}, Pedro Madrigal^{3,4*}, Sonia Tarazona^{2,5}, David Gomez-Cabrero^{6,7,8,9}, Alejandra Cervera¹⁰, Andrew McPherson¹¹, Michał Wojciech Szcześniak¹², Daniel J. Gaffney³, Laura L. Elo¹³, Xuegong Zhang^{14,15} and Ali Mortazavi^{16,17*}

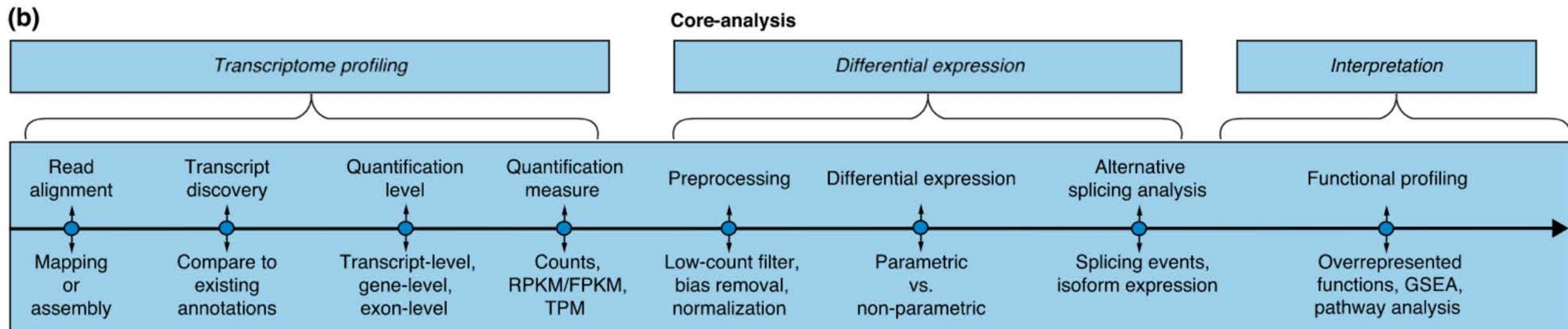
Pre-analysis

(a)



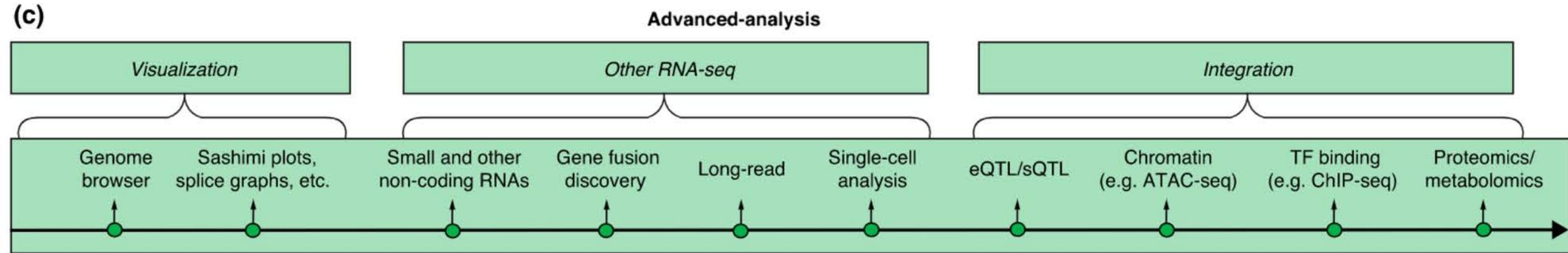
Core-analysis

(b)

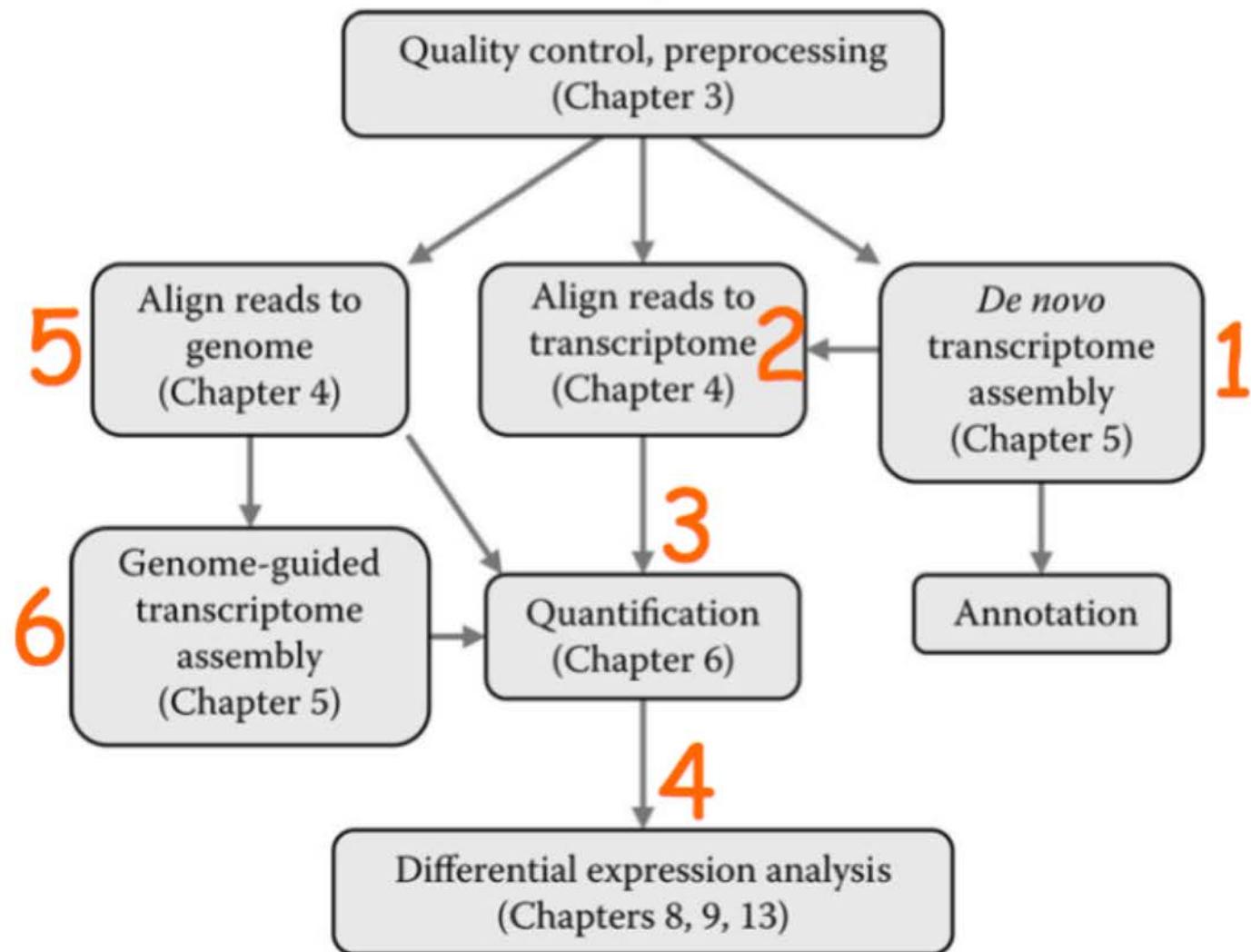


Advanced-analysis (not covered in this lecture but should be mentioned)

(c)



Expression quantification and transcript assembly

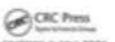


Chapman & Hall/CRC
Mathematical and Computational Biology Series

RNA-seq Data Analysis A Practical Approach



Eija Korpelainen, Jarno Tuimala,
Pauu Somervuo, Mikael Huss, and Garry Wong



September 9, 2014

E-Book - ISBN 9781482262353

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

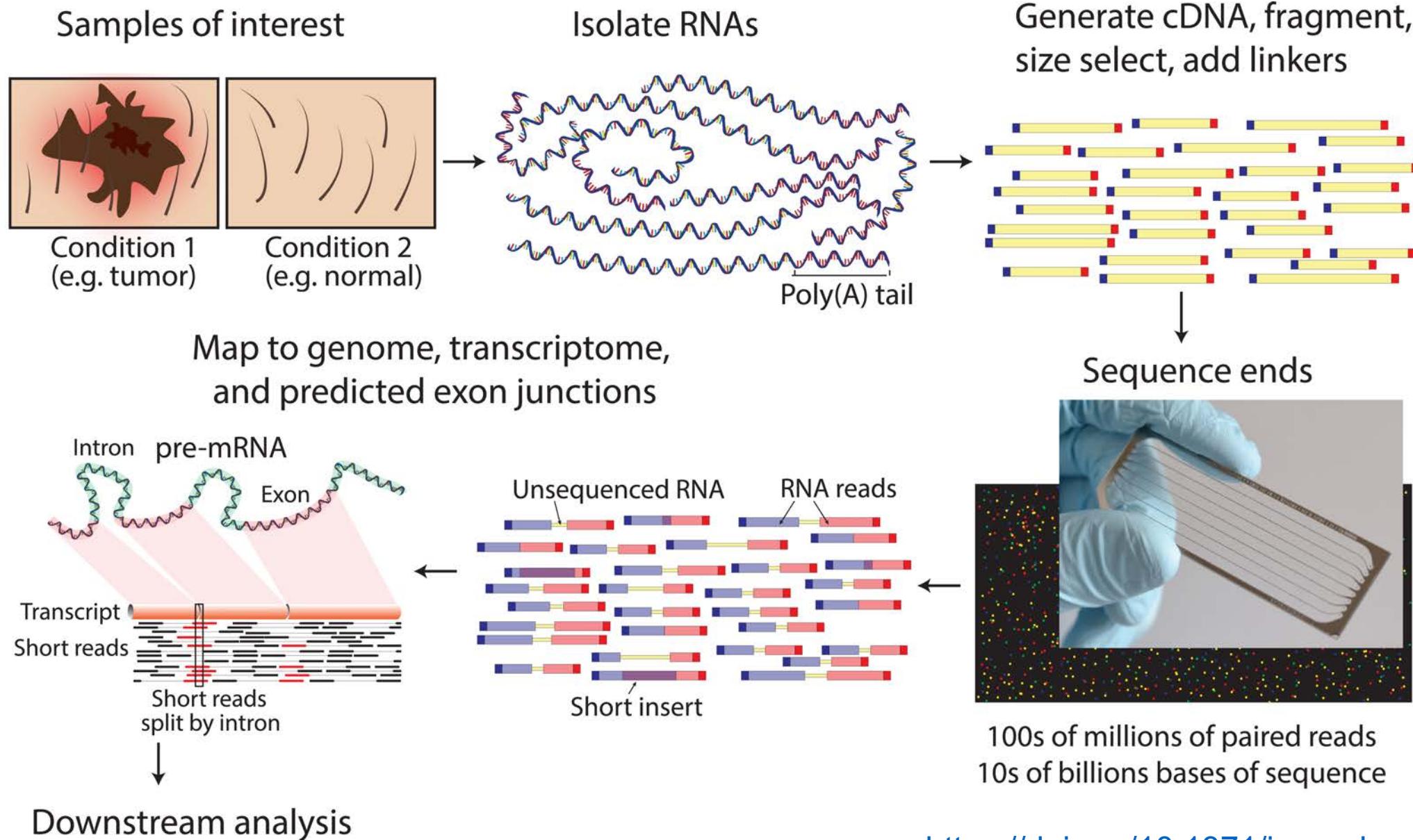
Introduction to differential gene expression analysis using RNA-seq

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



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

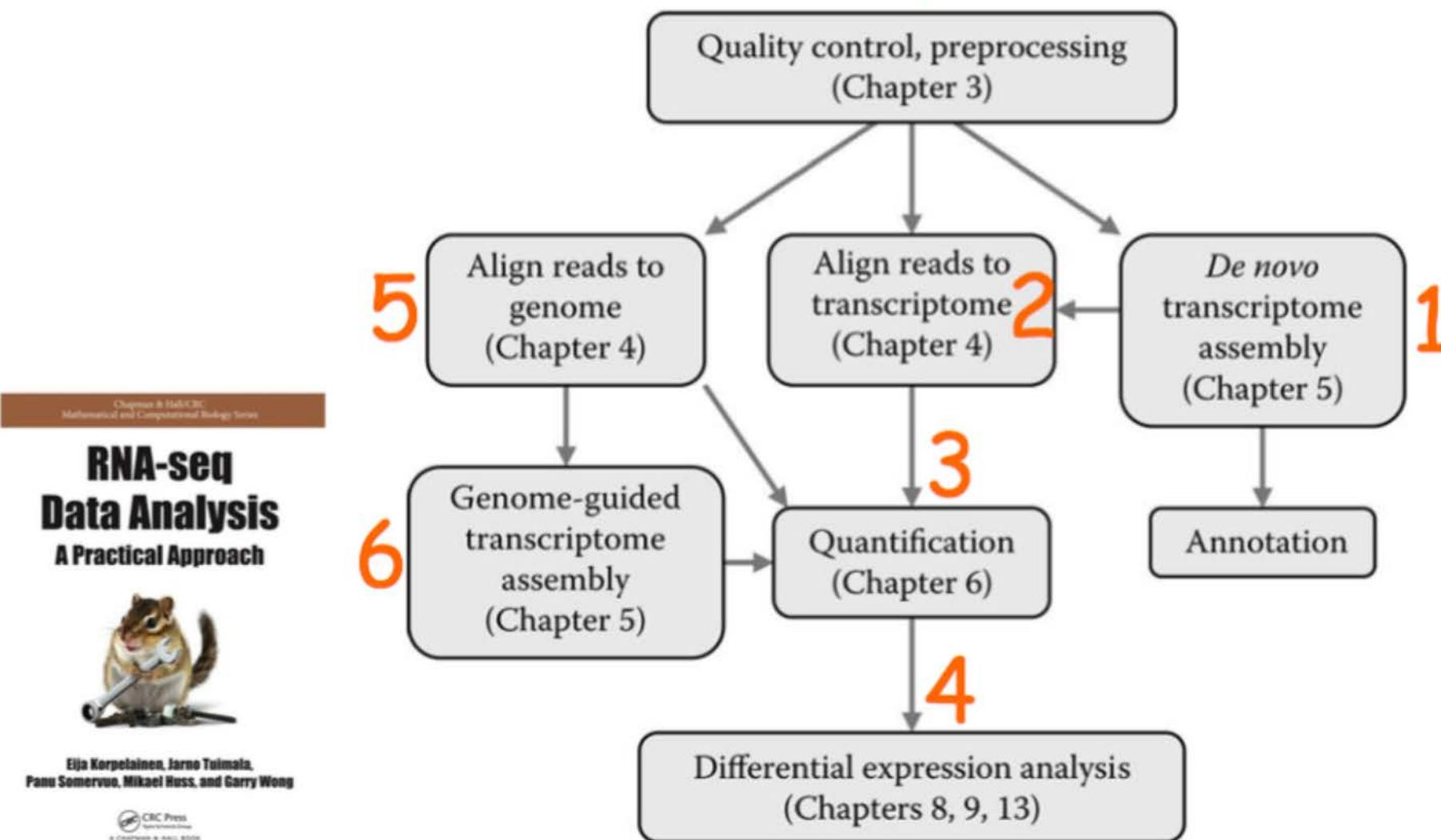
- expression
- Isoform abundance, splice patterns, transcript boundaries

New field:

RNA:RNA (CLASH) (Travis *et al.*, 2014)

RNA:protein (RIP-seq) (Cook *et al.*, 2015)

Expression quantification and transcript assembly



September 9, 2014

E-Book - ISBN 9781482262353

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

Basic rule of thumb

Just genome with no closely related species

Different *de novo* predictors, and combine them with combiners

Genome + closely related species + RNAseq

de novo predictors + evidence + combiners

Genome + closely related species available + RNAseq

de novo predictors + evidence + RNAseq evidence + combiners

Genome + closely related species available + RNAseq + manual efforts

manual curation to train *de novo* predictors

Trained predictors + protein evidence + RNAseq evidence + combiners

Genome + initial annotations + RNAseq

protein evidence -> Trying to improve existing annotations

RNAseq Raw data type

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

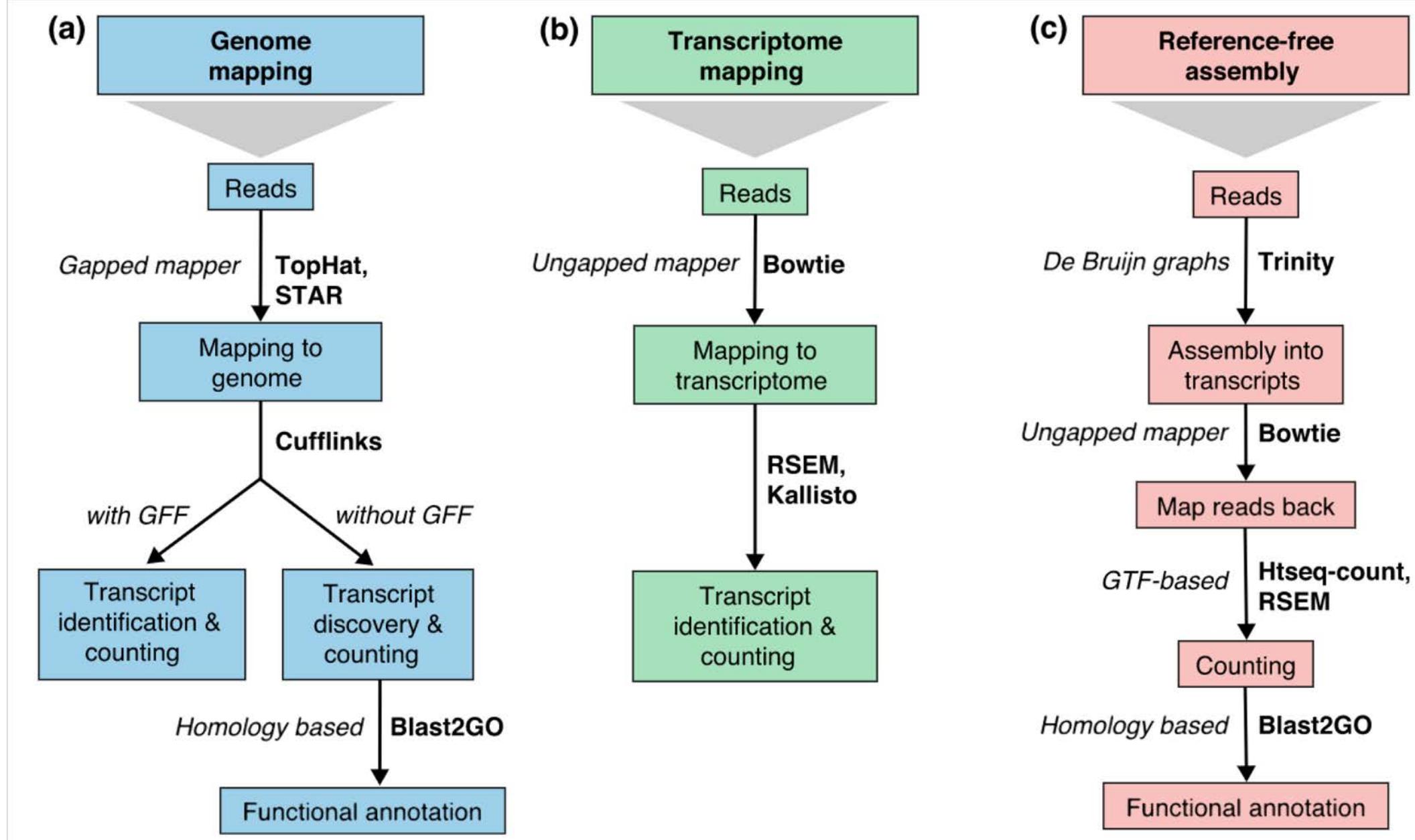
Read mapping

1. **Fasta** file with genome sequence
2. **Gff** file with gene model annotations

```
##gff-version 3
##sequence-region ctg123 1 1497228
ctg123 . gene          1000  9000  .  +  .  ID=gene00001;Name=EDEN
ctg123 . exon          1300  1500  .  +  .  ID=exon00001;Parent=mRNA00003
```

<http://www.sequenceontology.org/resources/gff3.html>

Read mapping and transcript identification strategies



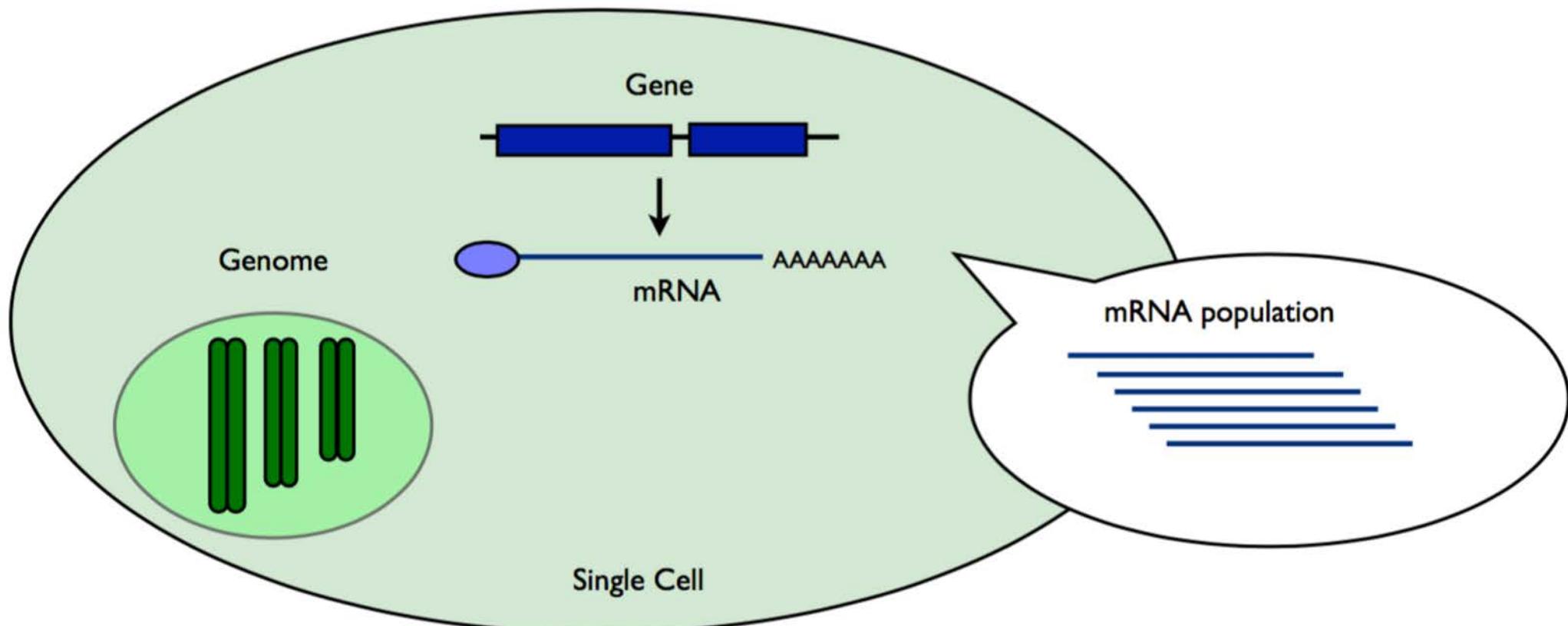
Differential expression

Types of experiments

Transcriptome Complexity:

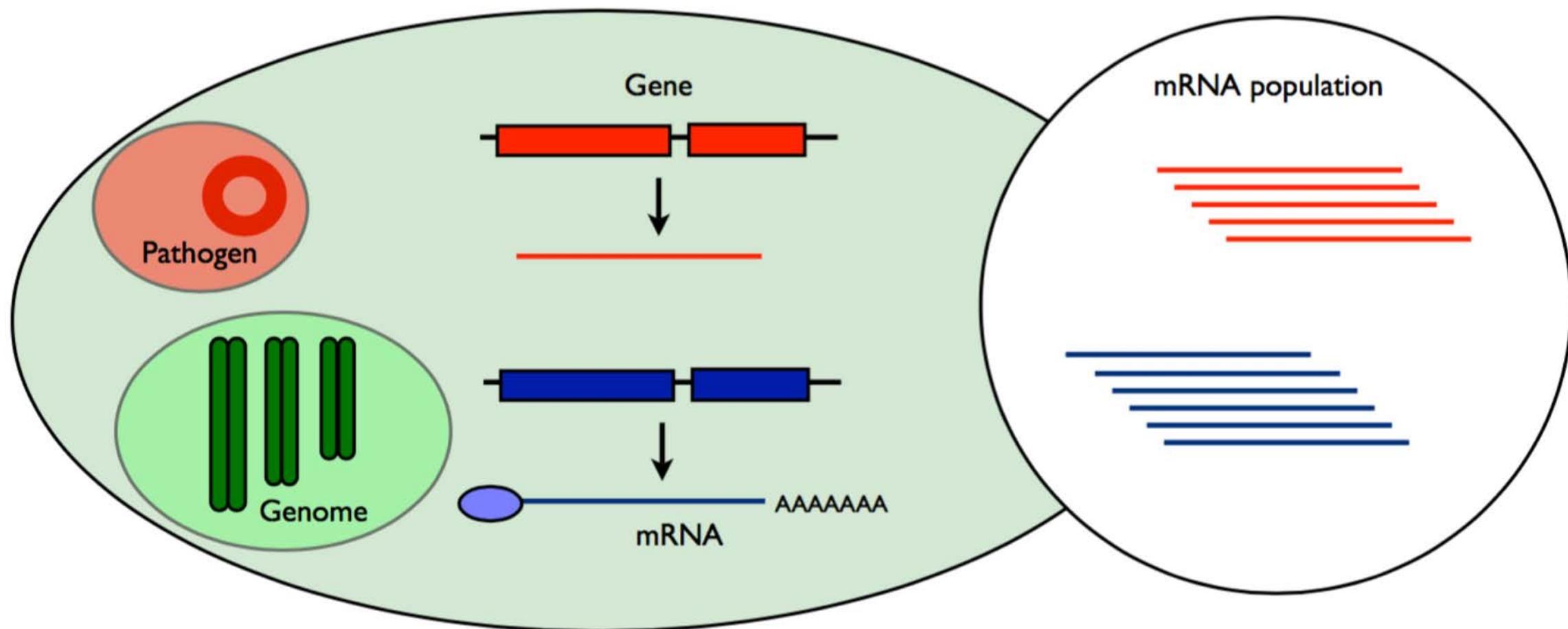
Simple System:

One Genome => Gene 1 copy => Single mRNA



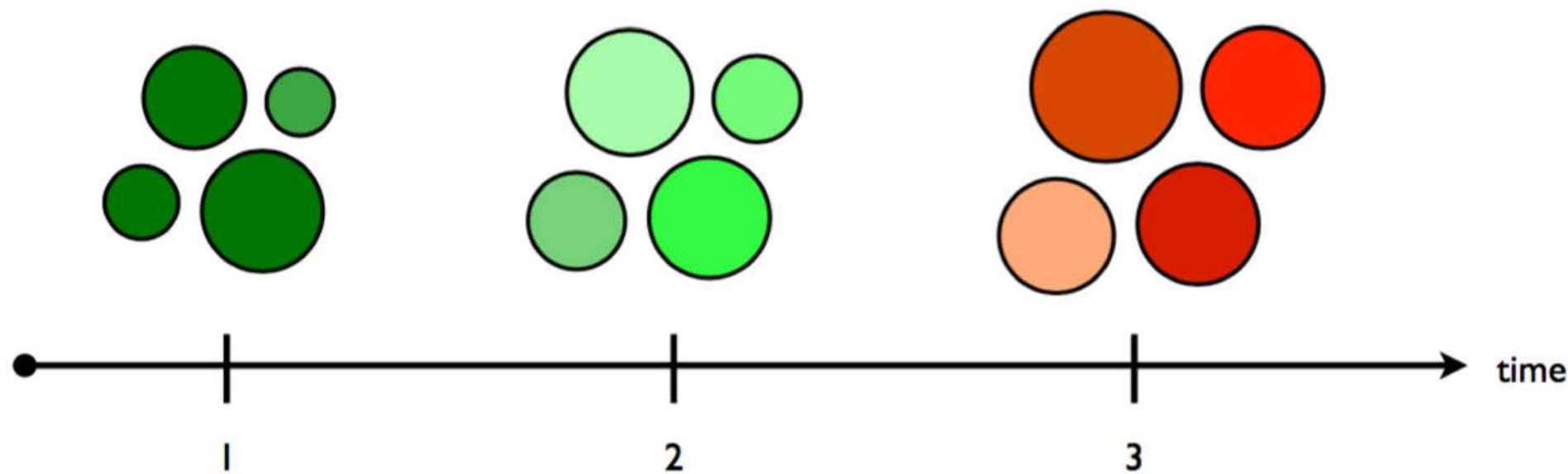
How many species we are analyzing ?

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



Is the study performed at different time points?

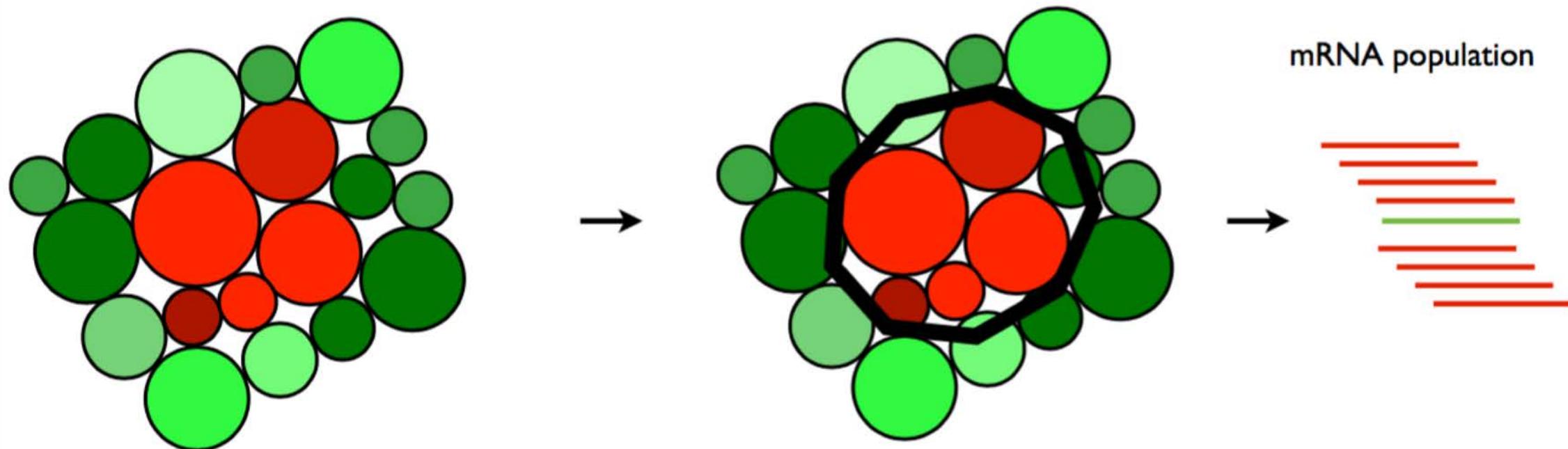
- 1) Developmental stages (difficult to select the same)**
- 2) Response to a treatment**



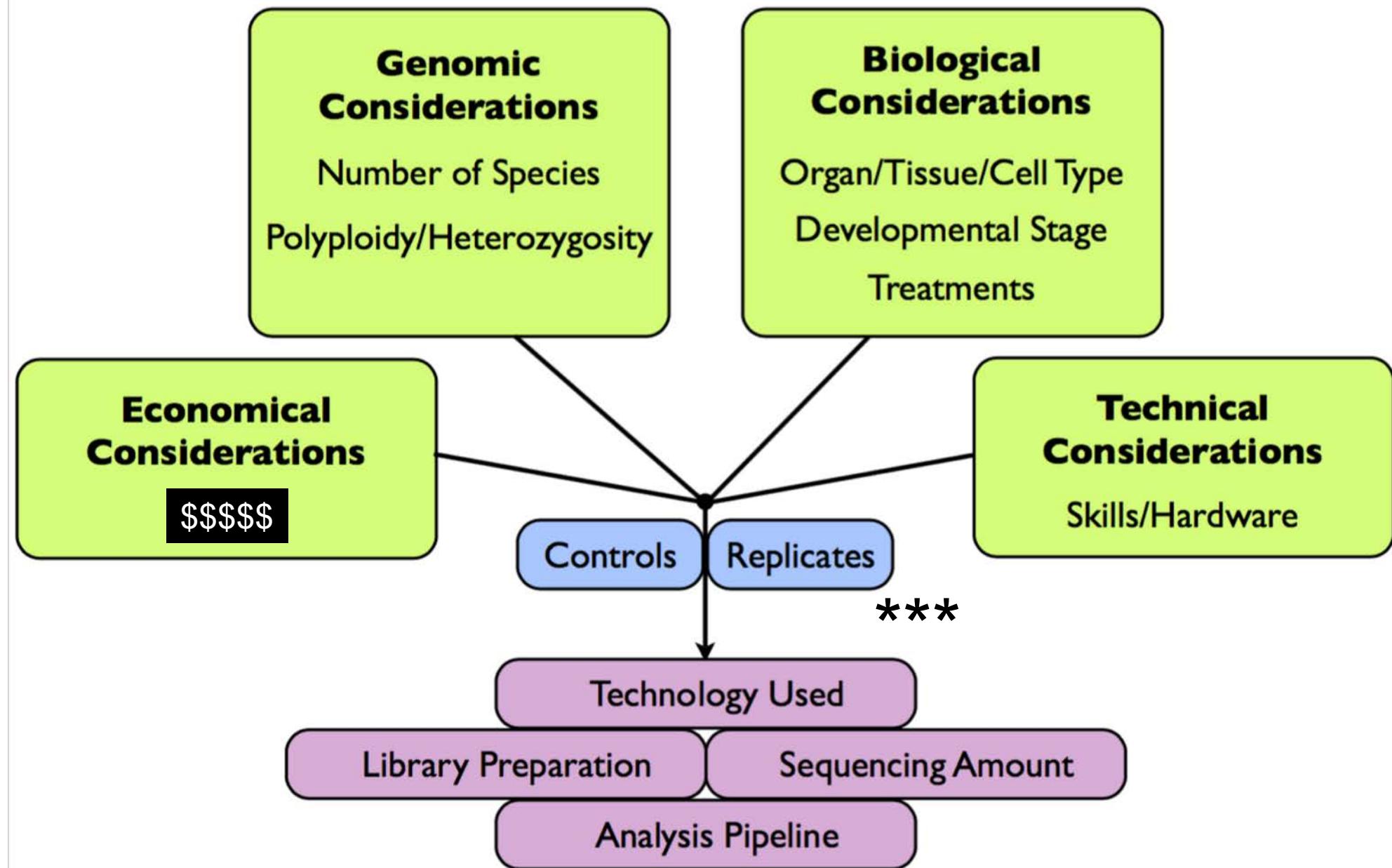
Is the study performed with different parts?

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

(Laser Capture Microdissection, LCM)



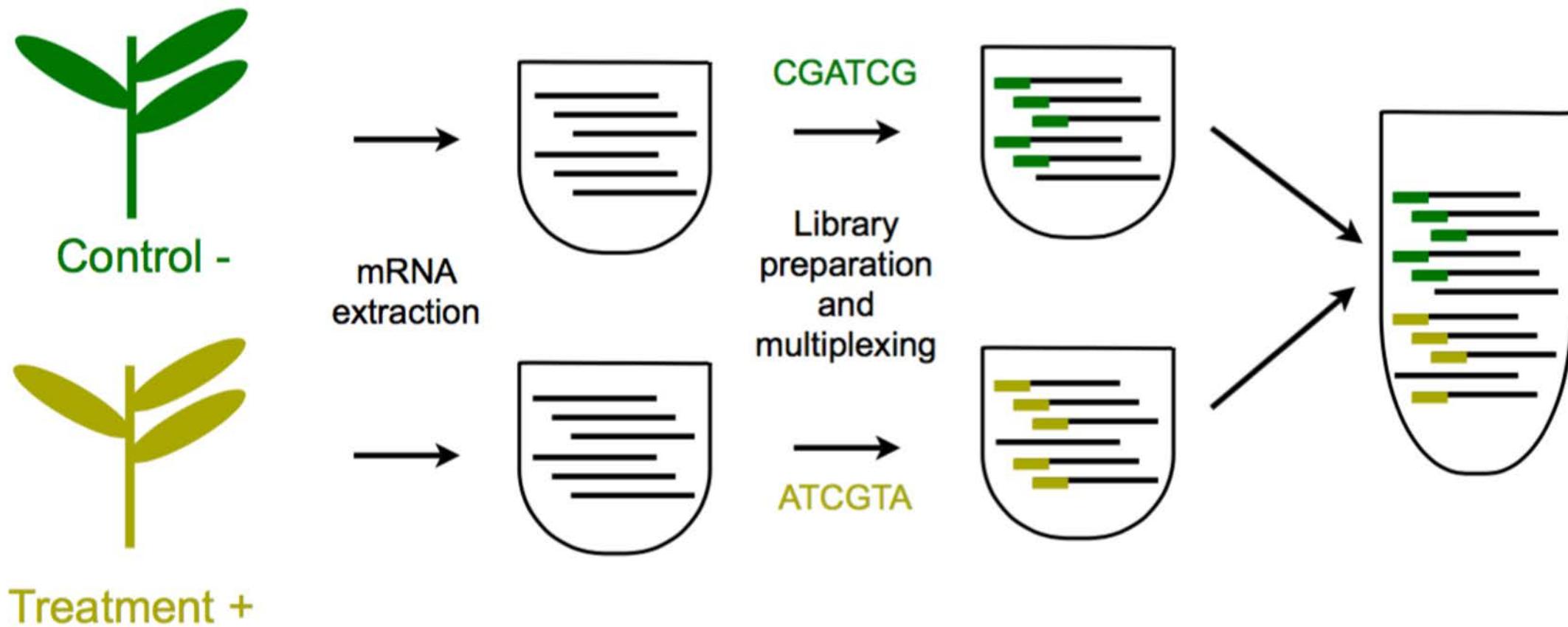
Experimental design



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

- Treatment

A A A B B B

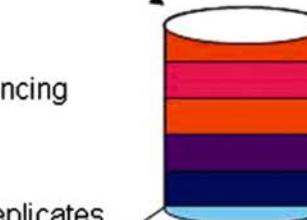
- Biological replicate



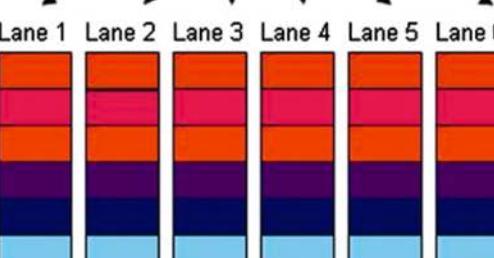
- RNA extraction



- Bar-code and pool



- Preparation for sequencing



- Sequence technical replicates

Confounded Design

- Treatment

A A A B B B

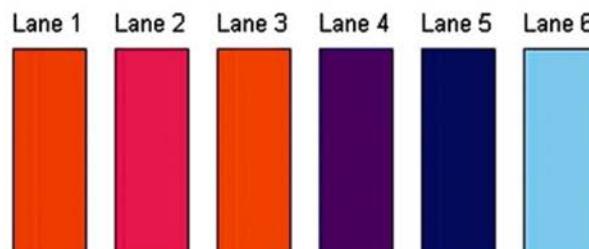
- Biological replicate



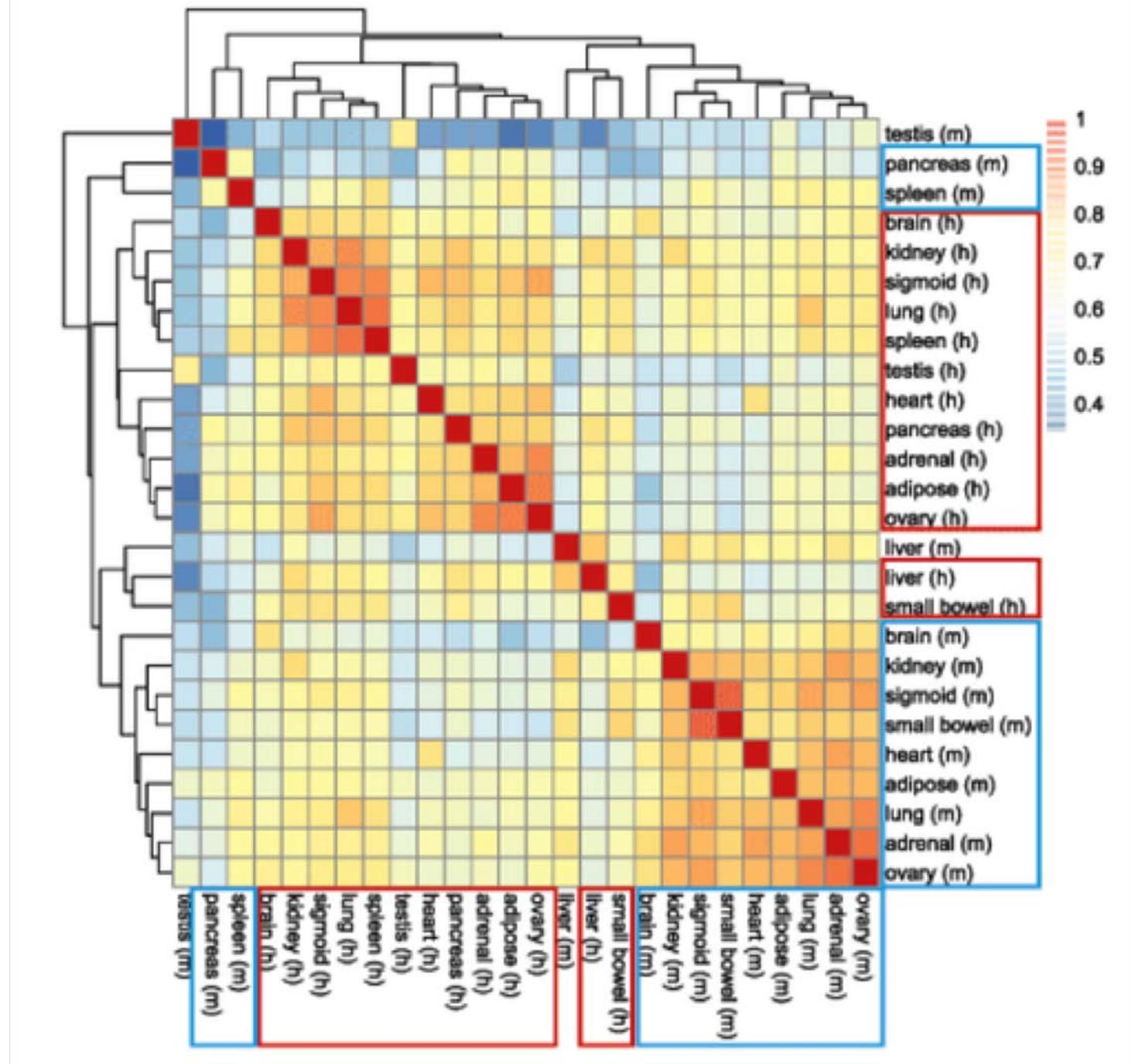
- RNA extraction and preparation for sequencing



- Sequence each sample in a lane



Example of batch effect:



Example of batch effect:

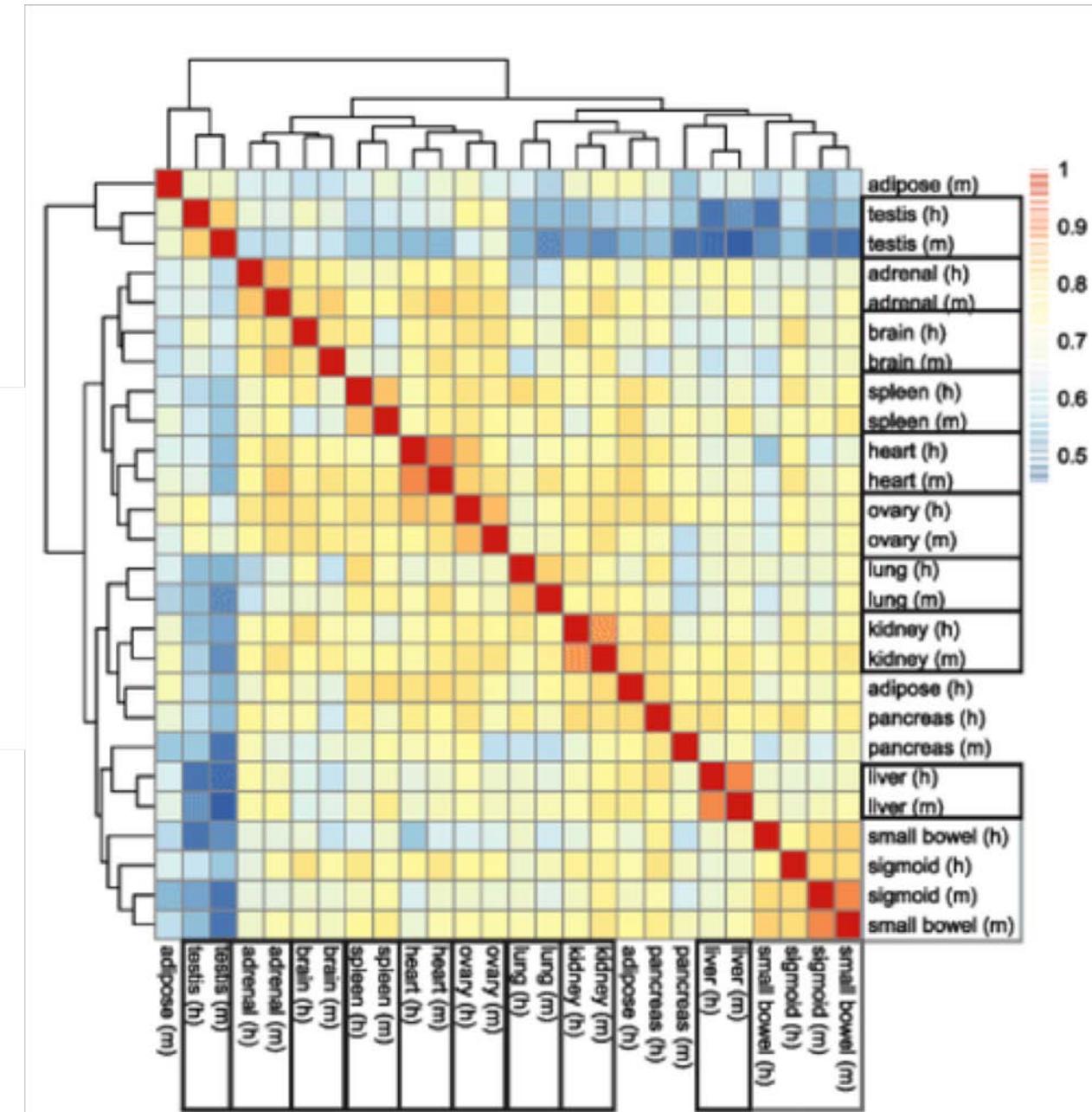


Yoav Gilad
@Y_Gilad



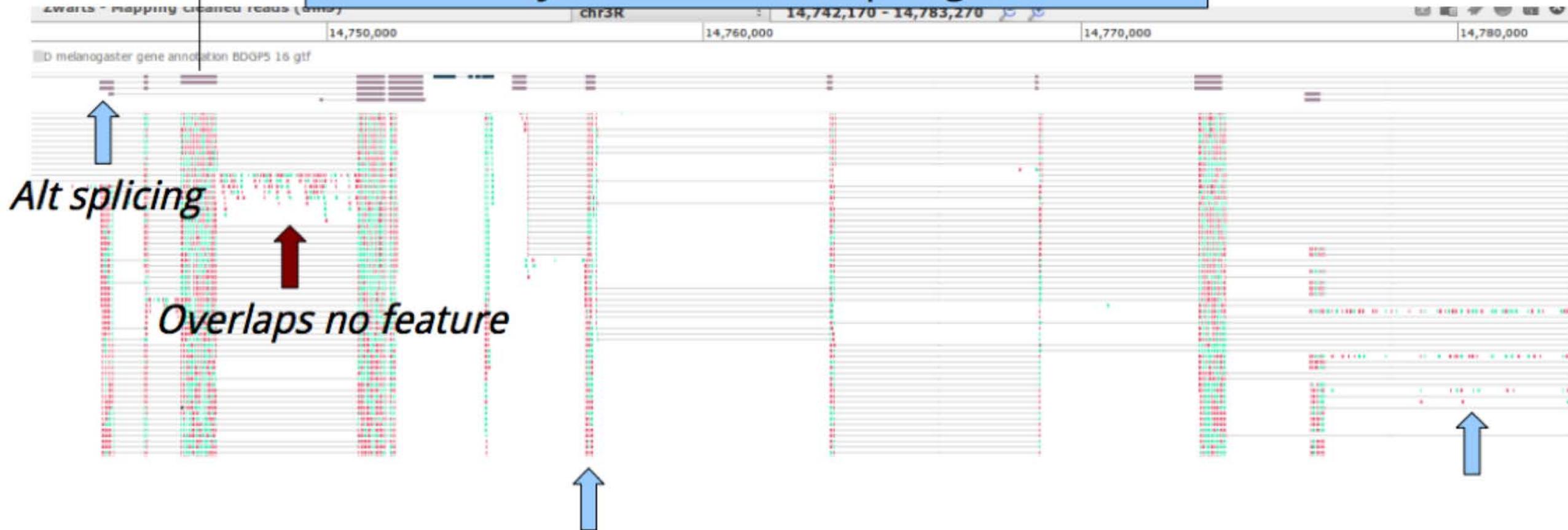
Following

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

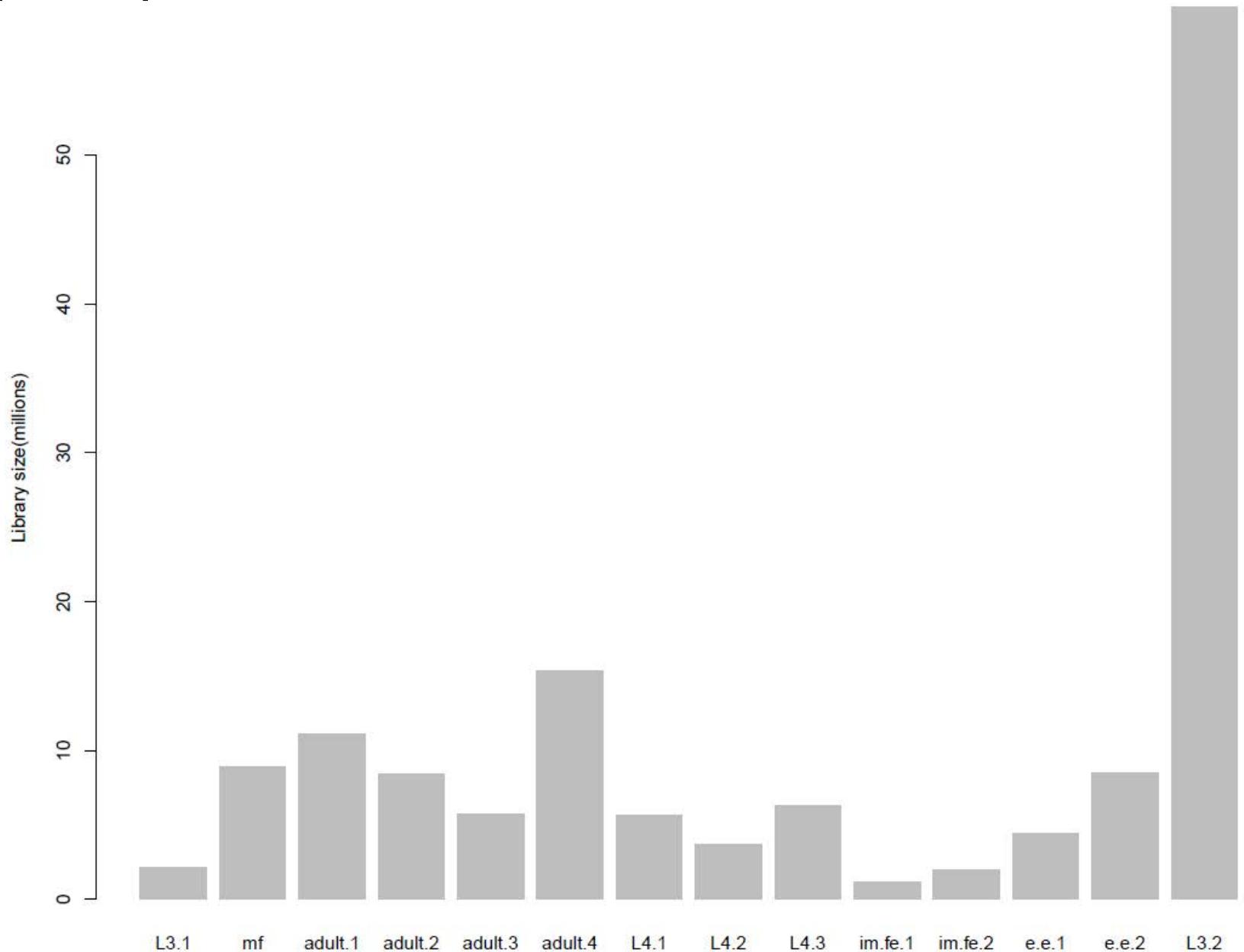
GeneB = exon 1 + exon 2 + exon 3 = 180 reads

This is the bit we care about!

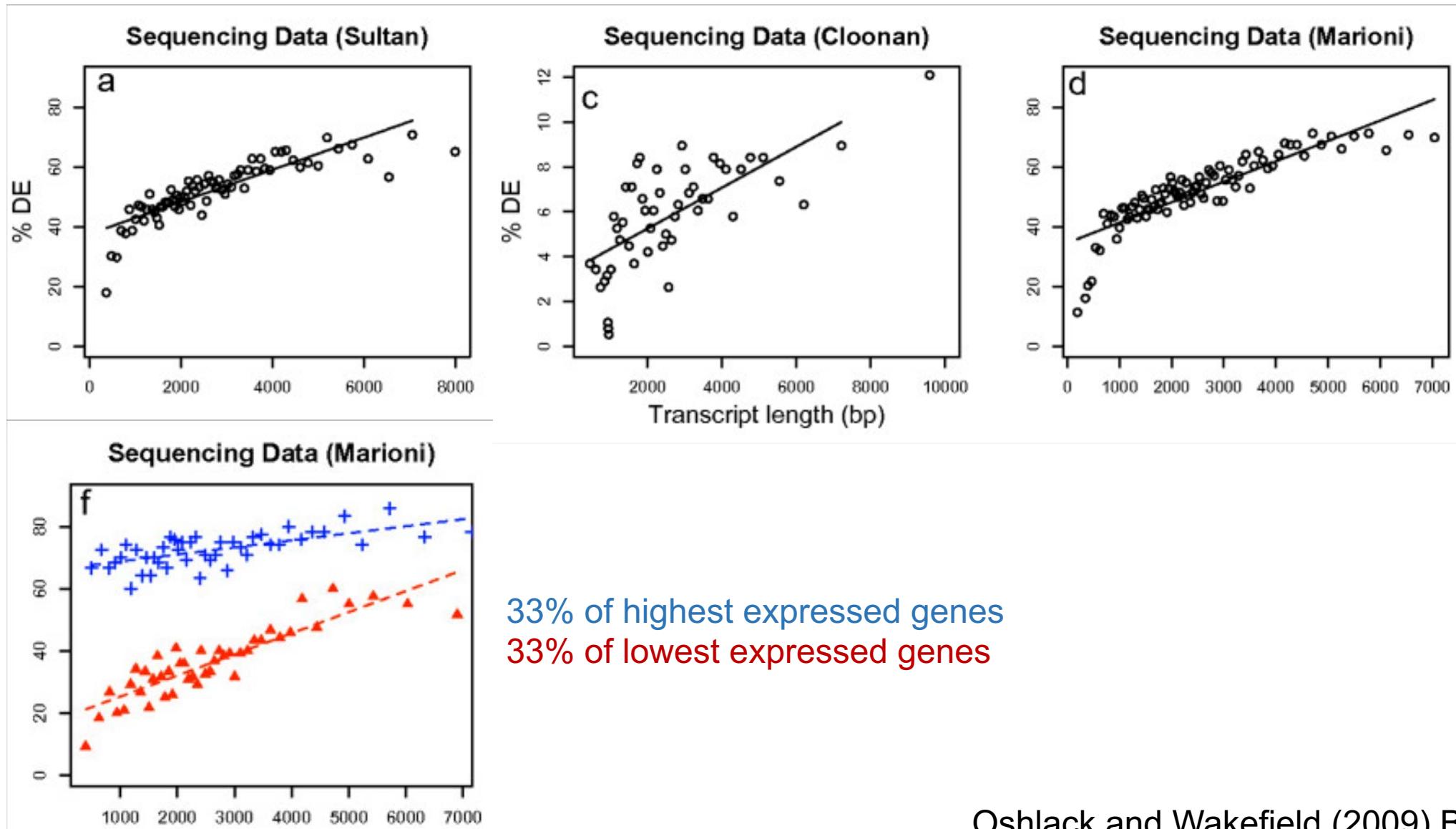


Counts of the gene depends on **expression** ,transcript length
,sequencing depth and simply chance

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

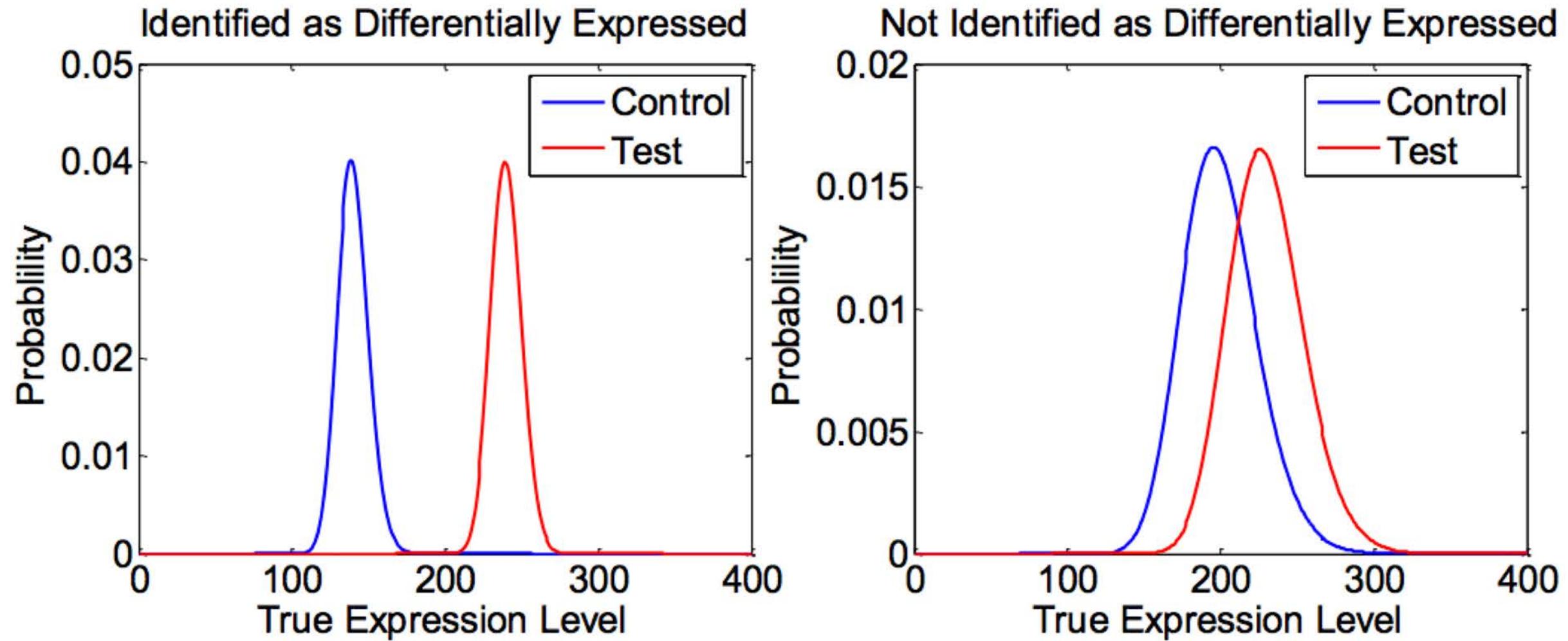
Gustavo Glusman , Juan Caballero, Max Robinson, Burak Kutlu, Leroy Hood

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

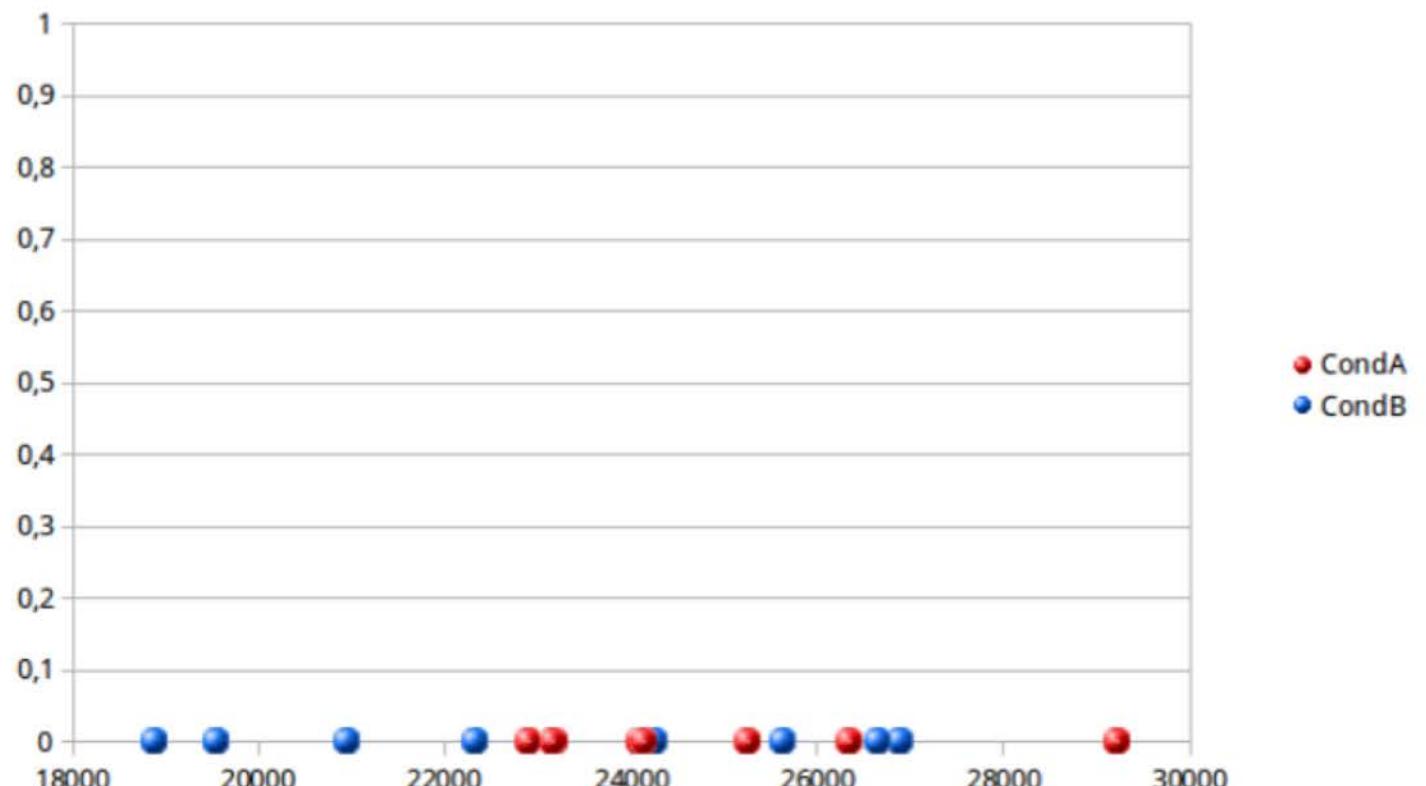
Fitting a distribution for every gene for DE



Scenario

gene_id CAF0006876

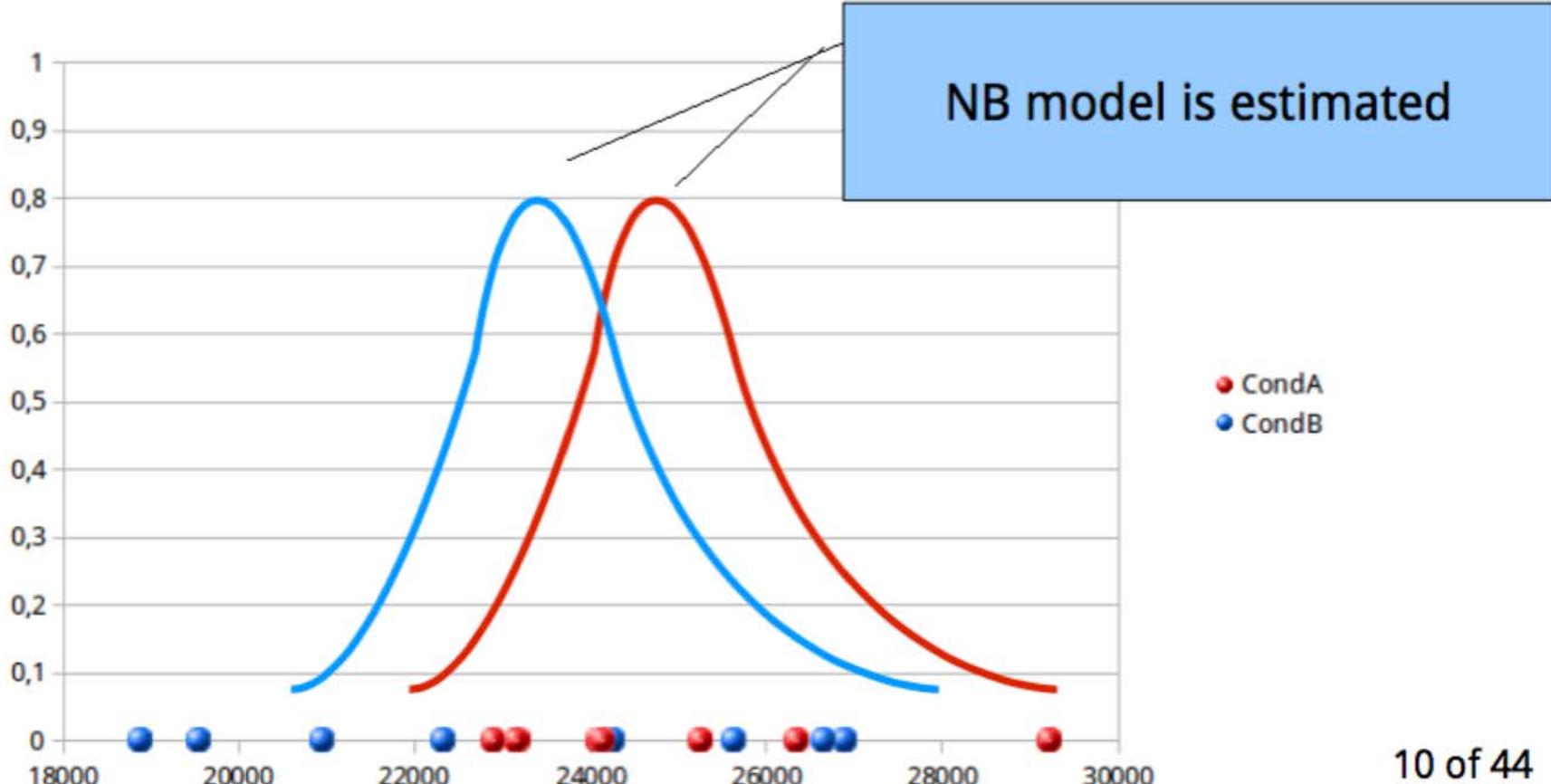
	sample1 23171	sample2 22903	sample3 29227	sample4 24072	sample5 23151	sample6 26336	sample7 25252	sample8 24122
Condition A								
Condition B	Sample9 19527	sample10 26898	sample11 18880	sample12 24237	sample13 26640	sample14 22315	sample15 20952	sample16 25629



Scenario

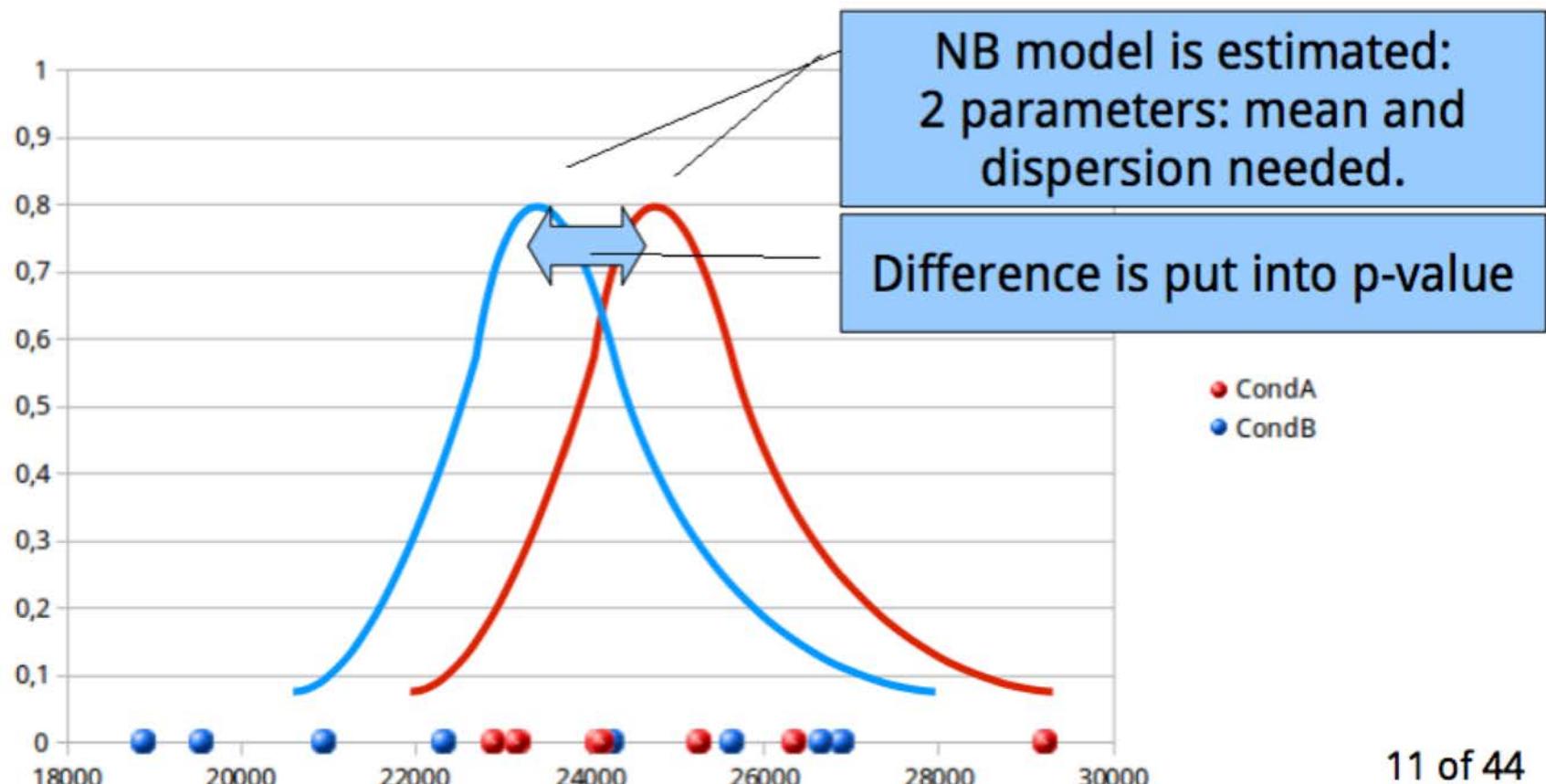
gene_id CAF0006876

	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

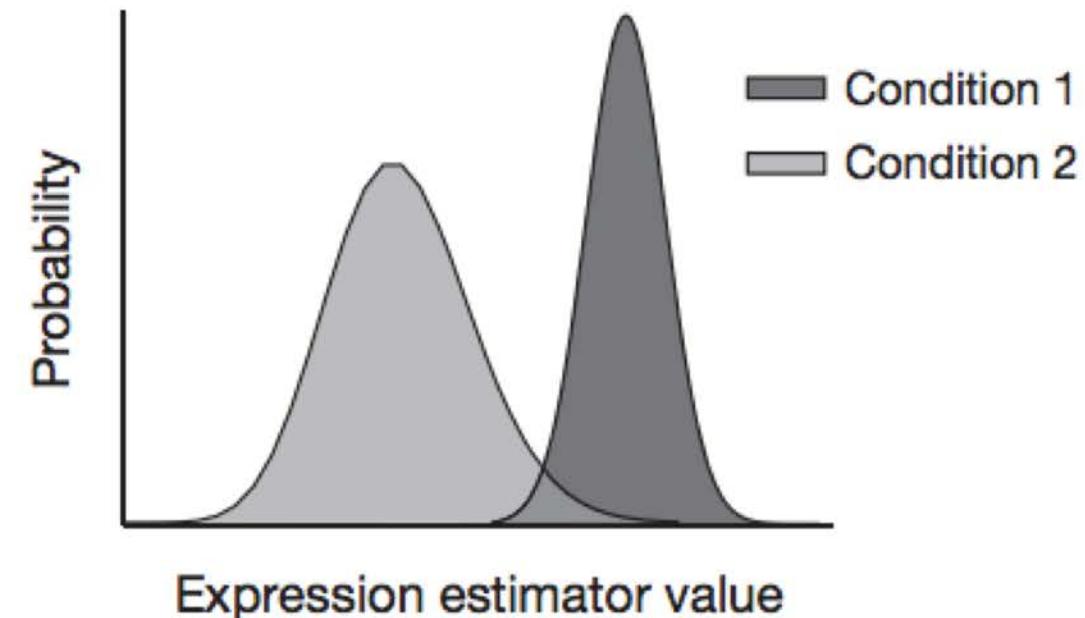
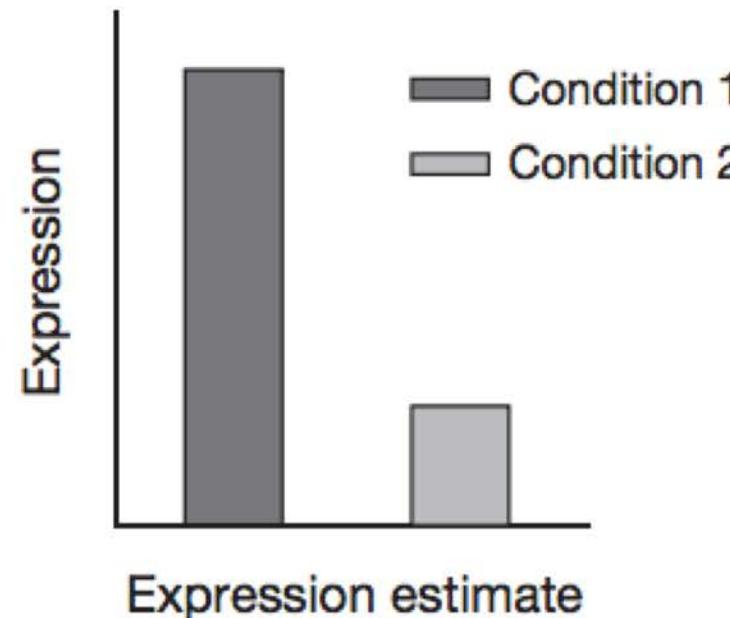
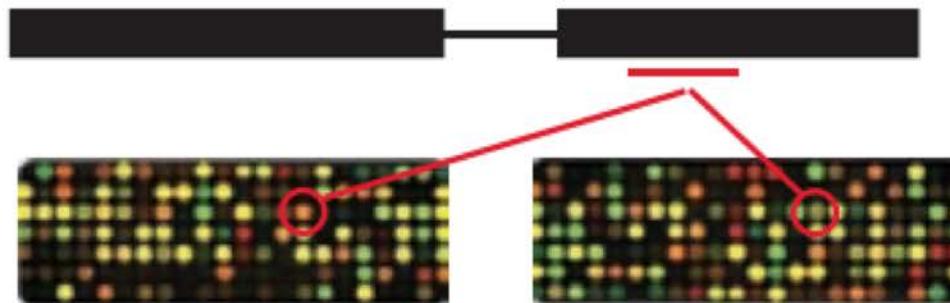


Scenario

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

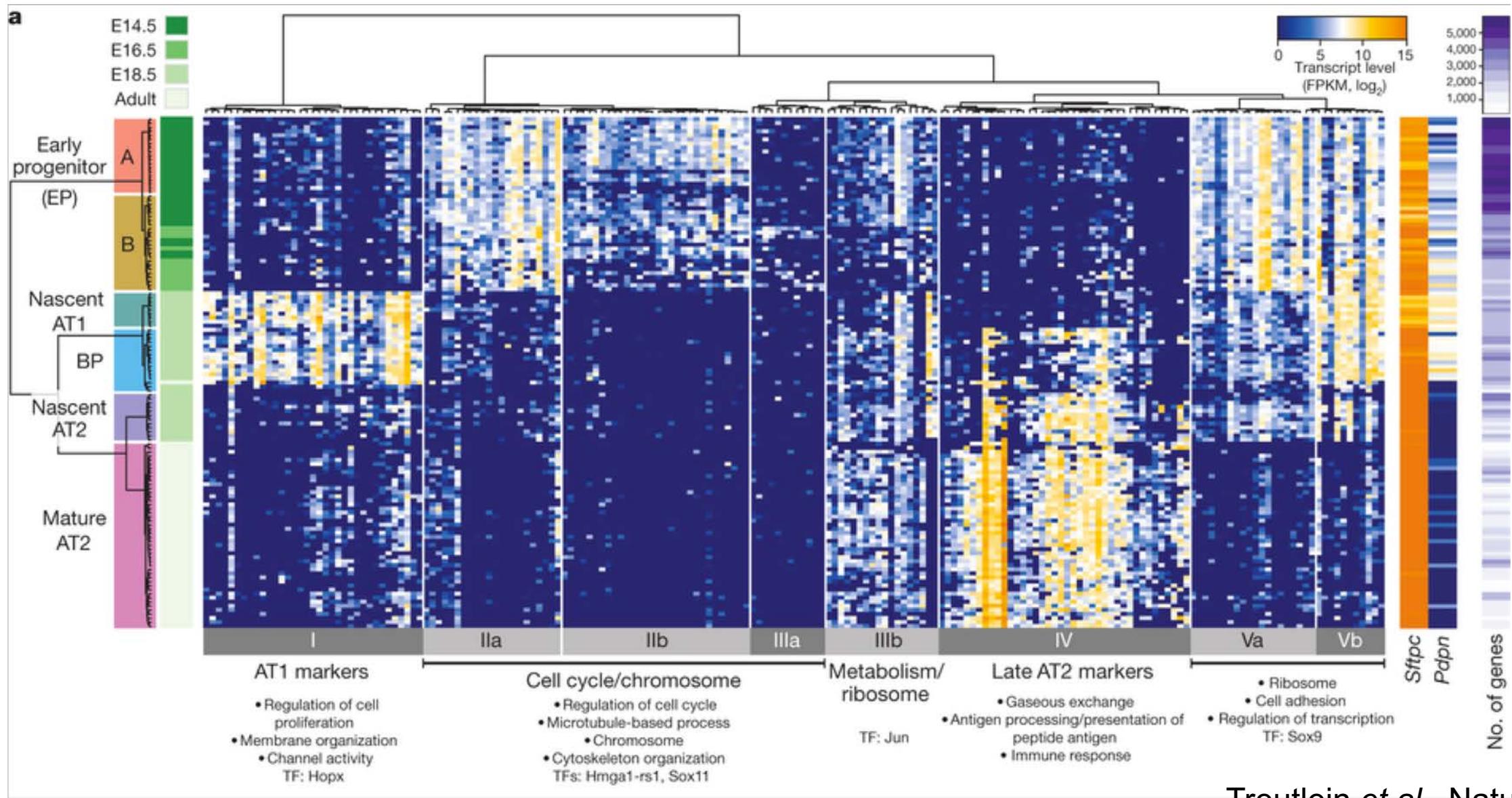


RNAseq vs Microarray



Once you have set of differentially expressed genes

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



Summary

- We have briefly covered “NGS”
- It’s really just sequencing but in a much higher throughput scale
- Only covered very basics of:
 - Genomics
 - Transcriptomics
- A lot of omics not covered..
 - Single cell sequencing
 - New technologies
 - And much more..

Personal journey

2005 – *Saccharomyces paradoxus*

- Capillary read sequenced full Chromosome III (~315kb) of 20 isolates
 - Costed £750k
 - One of the first scale re-sequencing projects
-
- Took me 3 years to sequence, align, annotate and analyse (= PhD)

Population genomics of the wild yeast *Saccharomyces paradoxus*: Quantifying the life cycle

Isheng J. Tsai, Douda Bensasson*, Austin Burt, and Vassiliki Koufopanou†

Division of Biology, Imperial College London, Silwood Park, Ascot, Berks SL5 7PY, United Kingdom

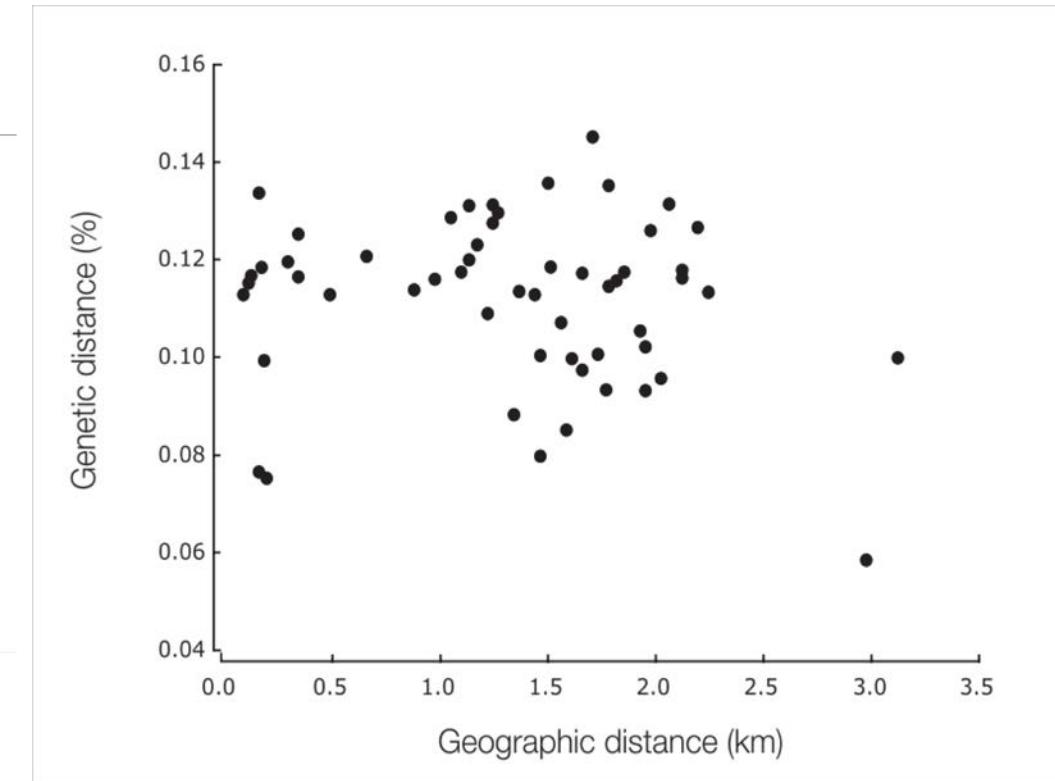
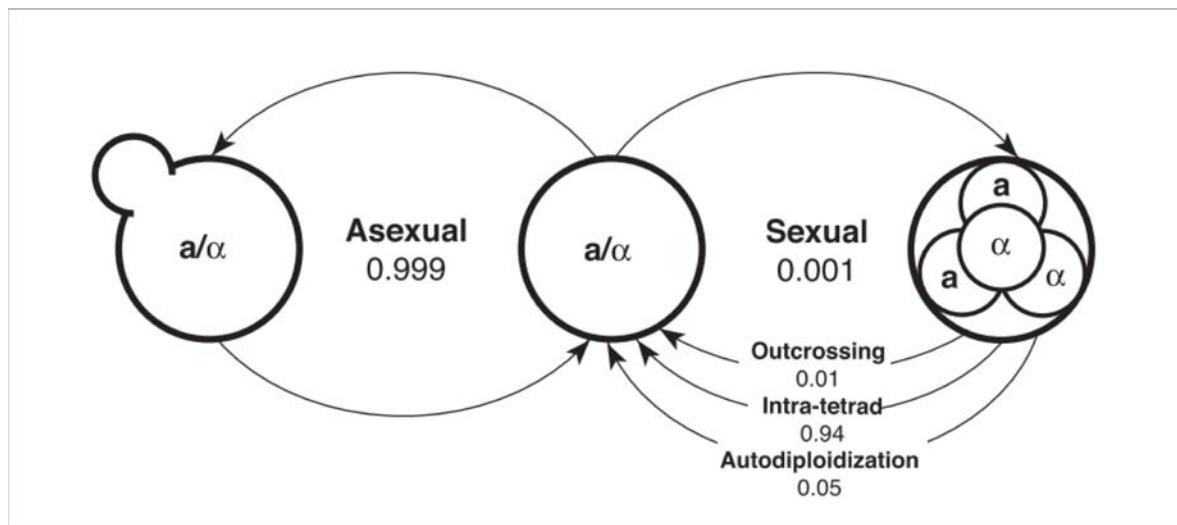
Edited by Mark Johnston, Washington University, St. Louis, MO, and accepted by the Editorial Board January 30, 2008 (received for review August 3, 2007)

Most microbes have complex life cycles with multiple modes of reproduction that differ in their effects on DNA sequence variation. Population genomic analyses can therefore be used to estimate the

are able to undergo mitoses, during which they repeatedly switch mating types, thus enabling matings between haploid clonemates (haplo-selfing or autodiploidization). This switch is possible be-

2005 – *Saccharomyces paradoxus*

- From population variation data we can infer frequencies of sex in yeast



2009 – *Saccharomyces* resequencing genome project

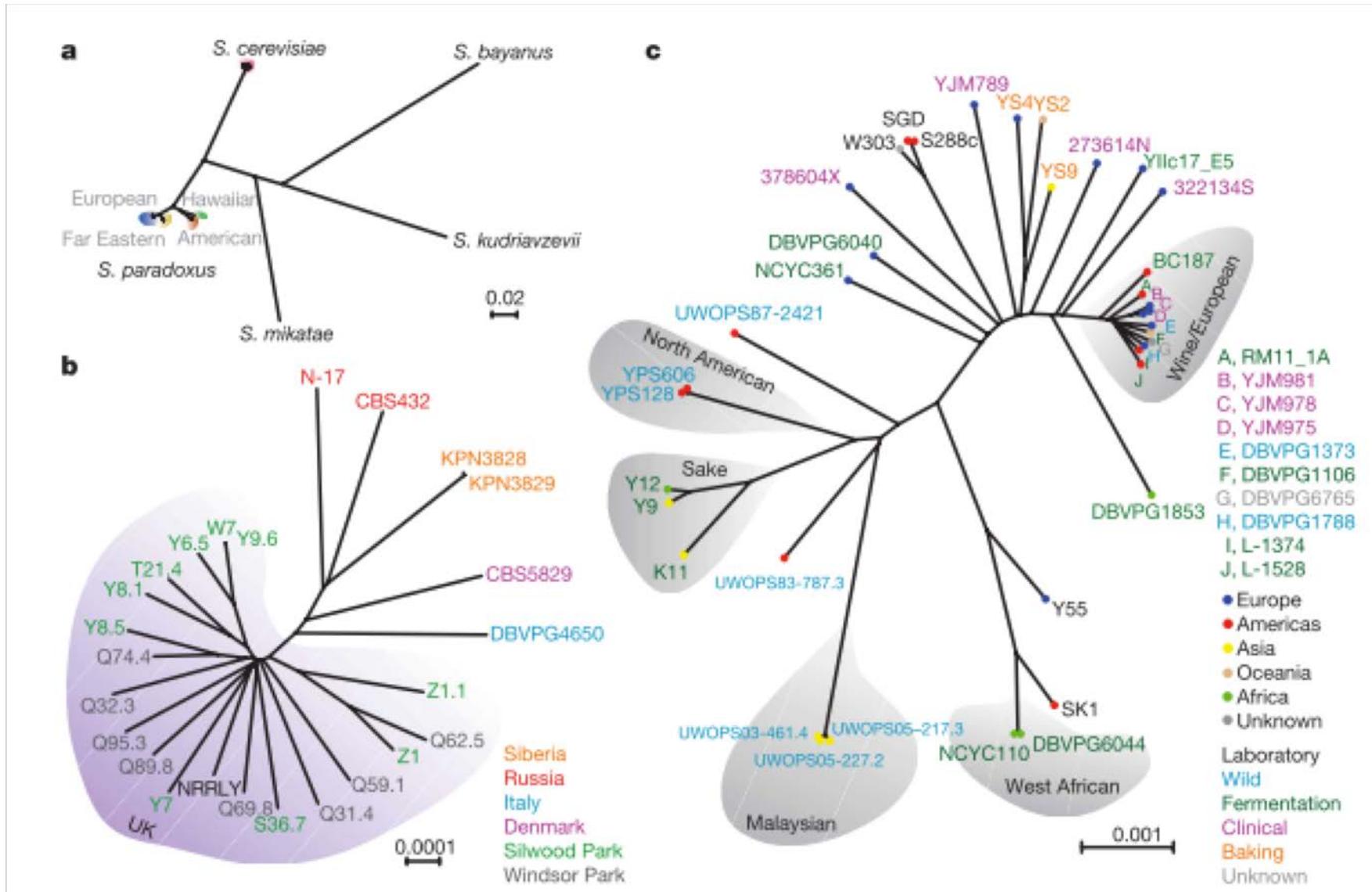
- 70 isolates at 1X-10X coverage
- ~2 years project with 26 authors
- At the start of NGS period (36bp Solexa reads)
- **Now= We are collecting and sequencing hundreds of isolates in Taiwan**

Population genomics of domestic and wild yeasts

Gianni Liti^{1*}, David M. Carter^{2*}, Alan M. Moses^{2,3}, Jonas Warringer⁴, Leopold Parts², Stephen A. James⁵, Robert P. Davey⁵, Ian N. Roberts⁵, Austin Burt⁶, Vassiliki Koufopanou⁶, Isheng J. Tsai⁶, Casey M. Bergman⁷, Douda Bensasson⁷, Michael J. T. O'Kelly⁸, Alexander van Oudenaarden⁸, David B. H. Barton¹, Elizabeth Bailes¹, Alex N. Nguyen Ba³, Matthew Jones², Michael A. Quail², Ian Goodhead^{2†}, Sarah Sims², Frances Smith², Anders Blomberg⁴, Richard Durbin^{2*} & Edward J. Louis^{1*}

2009 – *Saccharomyces* resequencing genome project

Phylogeny of ~70 isolates



2013 – Tapeworm genome project

- 4 tapeworm genomes (~100Mb) of different sequencing technologies (Illumina, 454, capillary)
- RNAseq of host infecting cycle ; sequencing of 7 isolates
- 2 years of work with 56 authors

ARTICLE

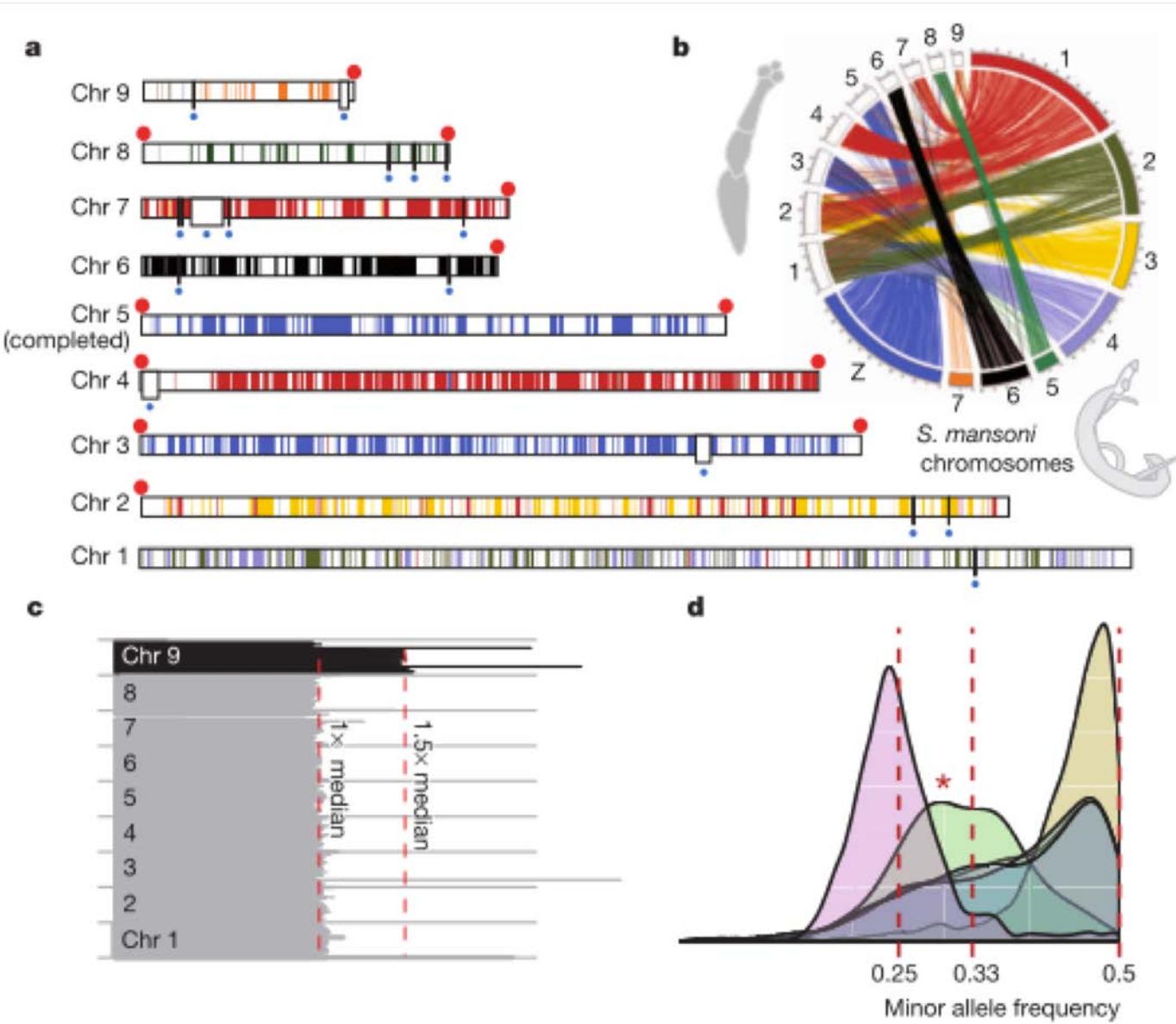
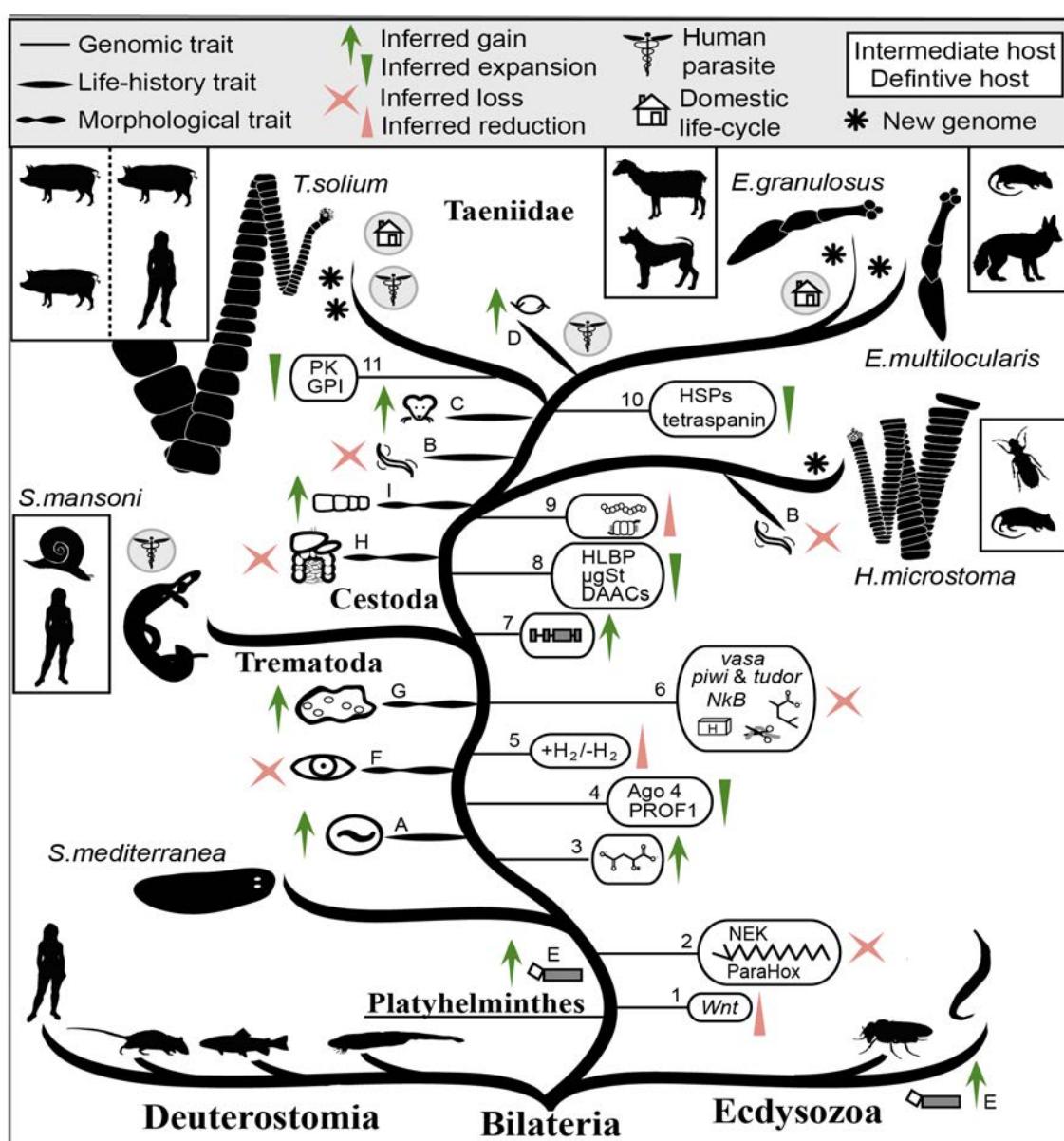
OPEN

doi:10.1038/nature12031

The genomes of four tapeworm species reveal adaptations to parasitism

Isheng J. Tsai^{1,2*}, Magdalena Zarowiecki^{1*}, Nancy Holroyd^{1*}, Alejandro Garciarrubio^{3*}, Alejandro Sanchez-Flores^{1,3}, Karen L. Brooks¹, Alan Tracey¹, Raúl J. Bobes⁴, Gladis Fragoso⁴, Edda Sciutto⁴, Martin Aslett¹, Helen Beasley¹, Hayley M. Bennett¹, Jianping Cai⁵, Federico Camicia⁶, Richard Clark¹, Marcela Cucher⁶, Nishadi De Silva¹, Tim A. Day⁷, Peter Deplazes⁸, Karel Estrada³, Cecilia Fernández⁹, Peter W. H. Holland¹⁰, Junling Hou⁵, Songnian Hu¹¹, Thomas Huckvale¹, Stacy S. Hung¹², Laura Kamenetzky⁶, Jacqueline A. Keane¹, Ferenc Kiss¹³, Uriel Koziol¹³, Olivia Lambert¹, Kan Liu¹¹, Xuenong Luo⁵, Yingfeng Luo¹¹, Natalia Macchiaroli⁶, Sarah Nichol¹, Jordi Paps¹⁰, John Parkinson¹², Natasha Pouchkina-Stantcheva¹⁴, Nick Riddiford^{14,15}, Mara Rosenzvit⁶, Gustavo Salinas⁹, James D. Wasmuth¹⁶, Mostafa Zamanian¹⁷, Yadong Zheng⁵, The *Taenia solium* Genome Consortium†, Xuepeng Cai⁵, Xavier Soberón^{3,18}, Peter D. Olson¹⁴, Juan P. Laclette⁴, Klaus Brehm¹³ & Matthew Berriman¹

2013 – Tapeworm genome project



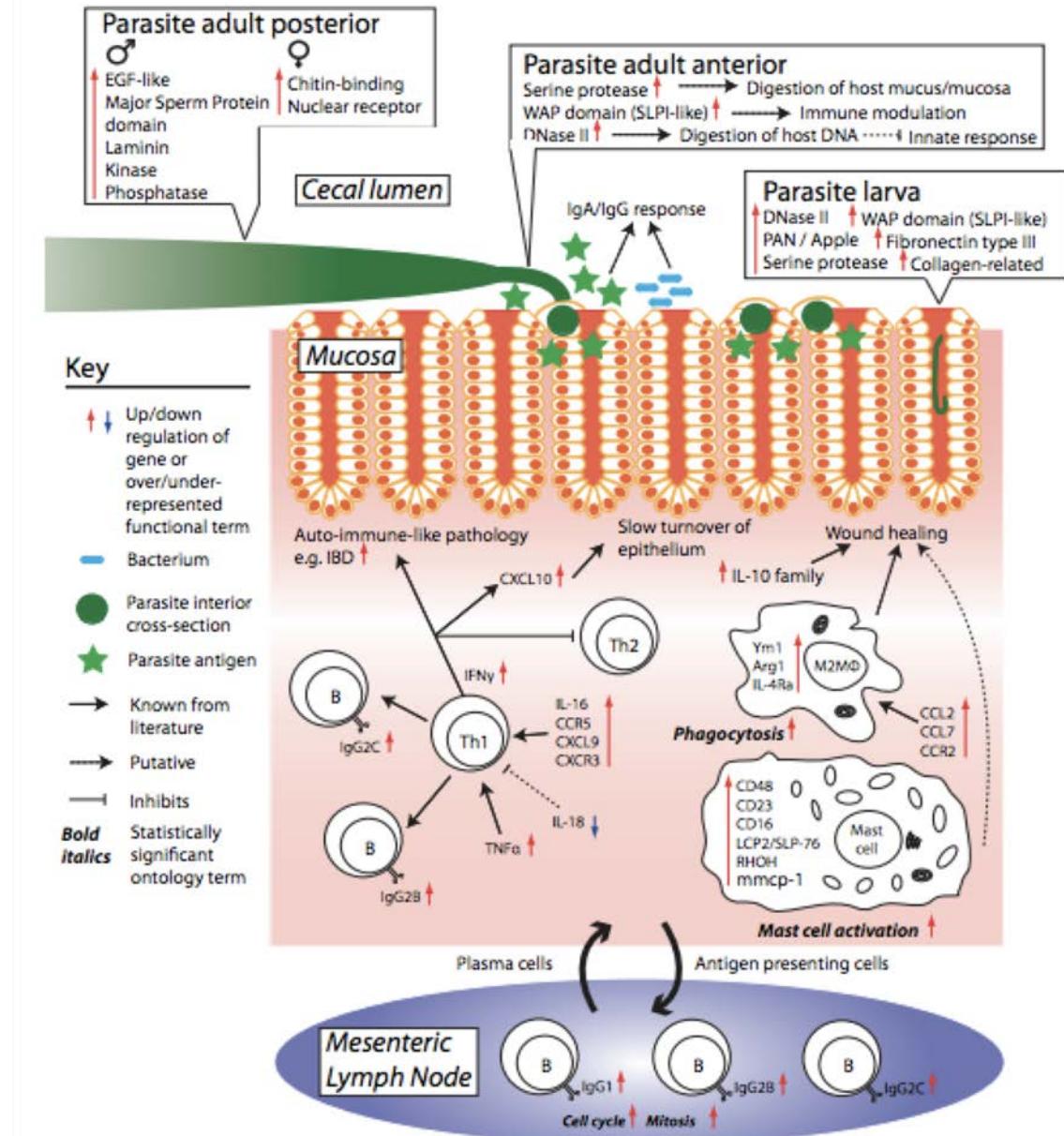
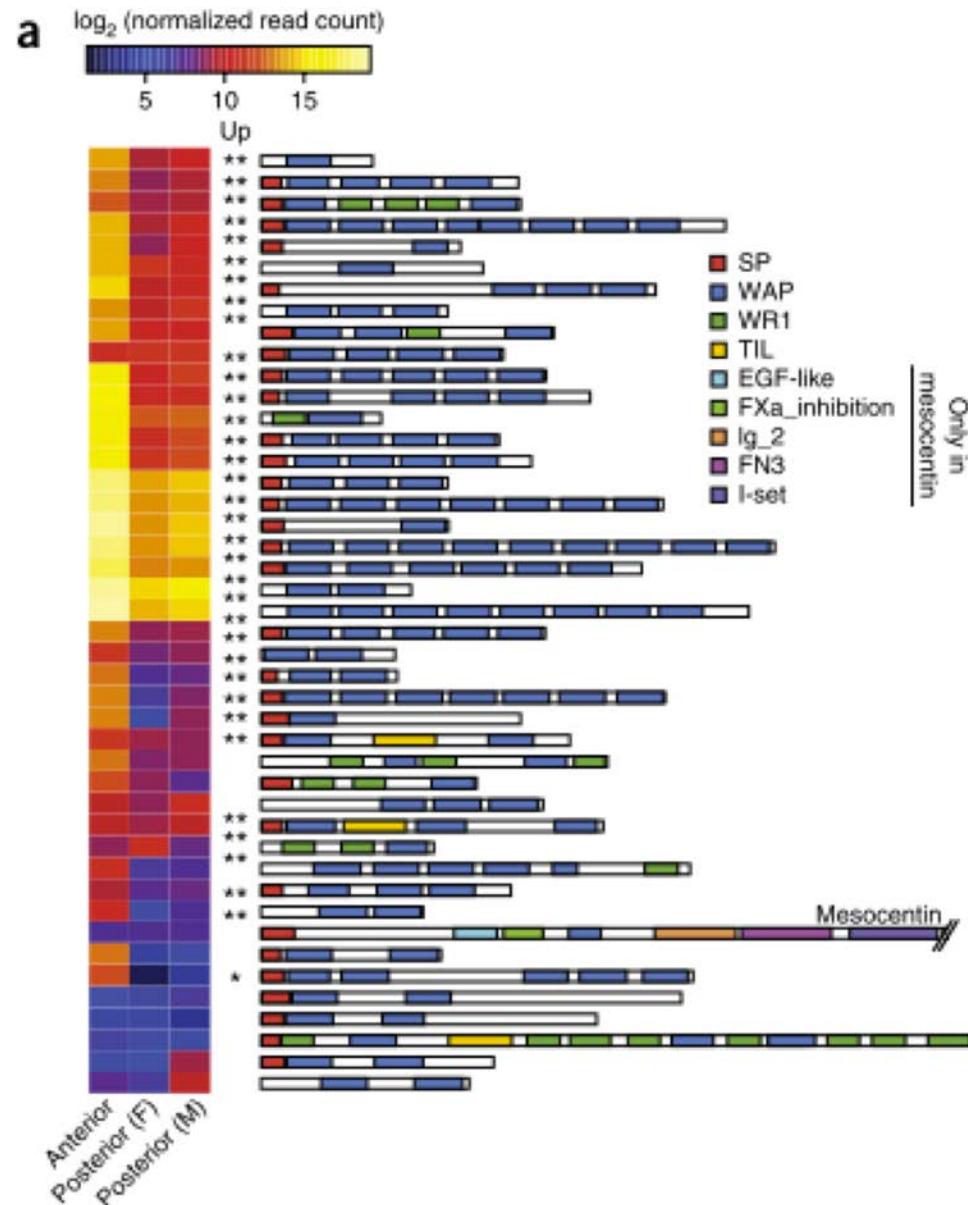
2014 – *Trichuris* genome project

- 2 genomes probably costs less than £10,000k
- About **40 RNAseq** libraries of different life cycle stages, host infecting stages
- Paradigm shifts to RNAseq

Whipworm genome and dual-species transcriptome analyses provide molecular insights into an intimate host-parasite interaction

Bernardo J Foth^{1,7}, Isheng J Tsai^{1,2,7}, Adam J Reid^{1,7}, Allison J Bancroft^{3,7}, Sarah Nichol¹, Alan Tracey¹, Nancy Holroyd¹, James A Cotton¹, Eleanor J Stanley¹, Magdalena Zarowiecki¹, Jimmy Z Liu⁴, Thomas Huckvale¹, Philip J Cooper^{5,6}, Richard K Grencis³ & Matthew Berriman¹

2014 – *Trichuris* genome project



2014 – *Taphrina* genome project

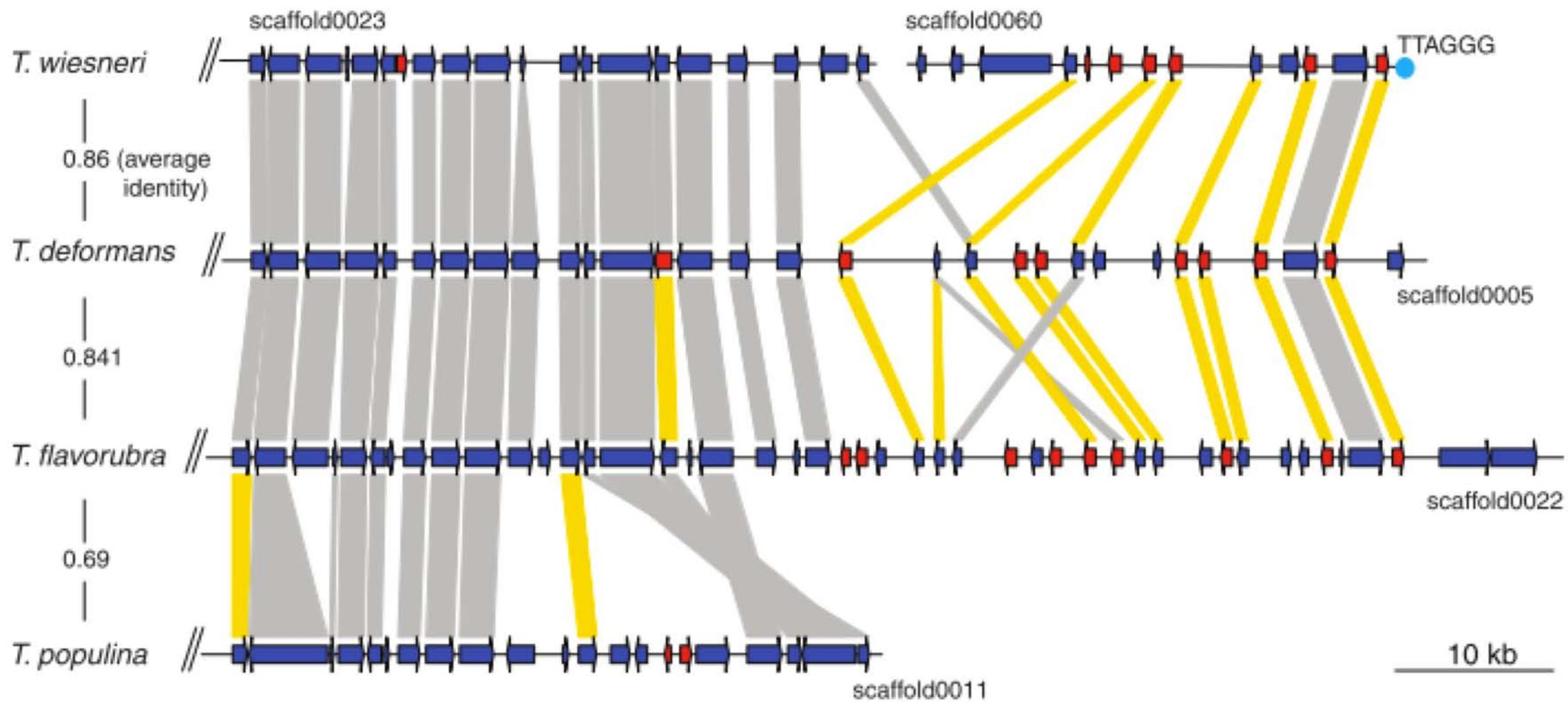
- 3 fungi genomes (~18Mb) of Illumina PE
- RNAseq for annotation purpose
- Costs probably less than 200,000 NT
- 2 months to analyse

GBE

Comparative Genomics of *Taphrina* Fungi Causing Varying Degrees of Tumorous Deformity in Plants

Isheng J. Tsai^{1,2}, Eiji Tanaka³, Hayato Masuya⁴, Ryusei Tanaka¹, Yuuri Hirooka⁵, Rikiya Endoh⁶, Norio Sahashi⁴, and Taisei Kikuchi^{1,4,*}

2014 – *Taphrina* genome project



2017 – *Phellinus* genome project

- 4 fungi genomes (30~50 Mb) of Pacbio reads (**comparative genomics**)
- RNAseq for annotation and DEG purpose (**RNAseq**)
- Resequencing of 60 isolates at ~60X (**population genomics**)

Received: 3 July 2017

Revised: 8 September 2017

Accepted: 11 September 2017

DOI: 10.1111/mec.14359

ORIGINAL ARTICLE

WILEY MOLECULAR ECOLOGY

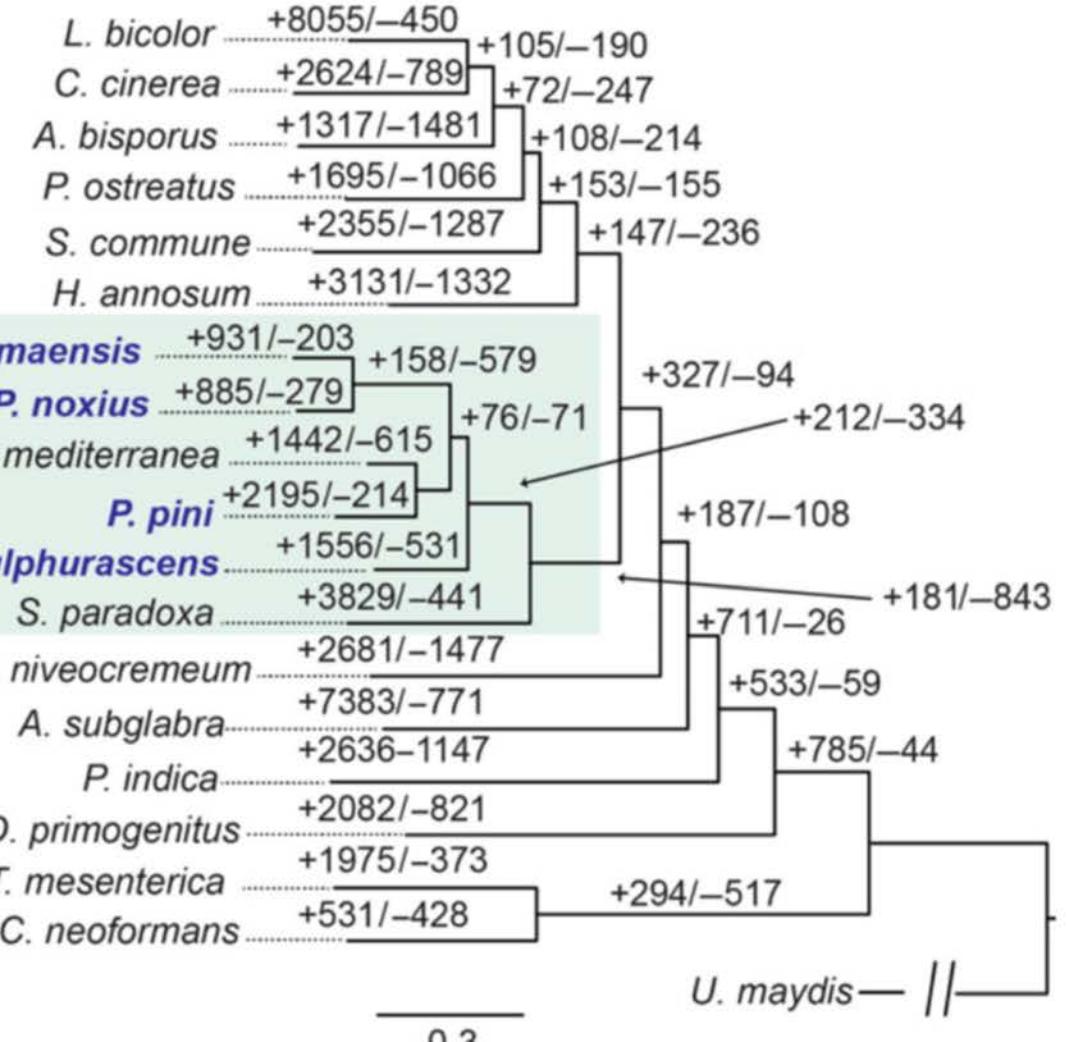
Comparative and population genomic landscape of *Phellinus noxius*: A hypervariable fungus causing root rot in trees

Chia-Lin Chung^{1,2*} | Tracy J. Lee^{3,4,5}  | Mitsuteru Akiba⁶ | Hsin-Han Lee¹ |
Tzu-Hao Kuo³  | Dang Liu^{3,7}  | Huei-Mien Ke³  | Toshiro Yokoi⁶ |
Marylette B. Roa^{3,8} | Mei-Yeh J. Lu³ | Ya-Yun Chang¹ | Pao-Jen Ann⁹ |
Jyh-Nong Tsai⁹ | Chien-Yu Chen¹⁰ | Shean-Shong Tzean¹ | Yuko Ota^{6,11} |
Tsutomu Hattori⁶ | Norio Sahashi⁶ | Ruey-Fen Liou^{1,2} | Taisei Kikuchi¹² |
Isheng J. Tsai^{3,4,5,7*} 

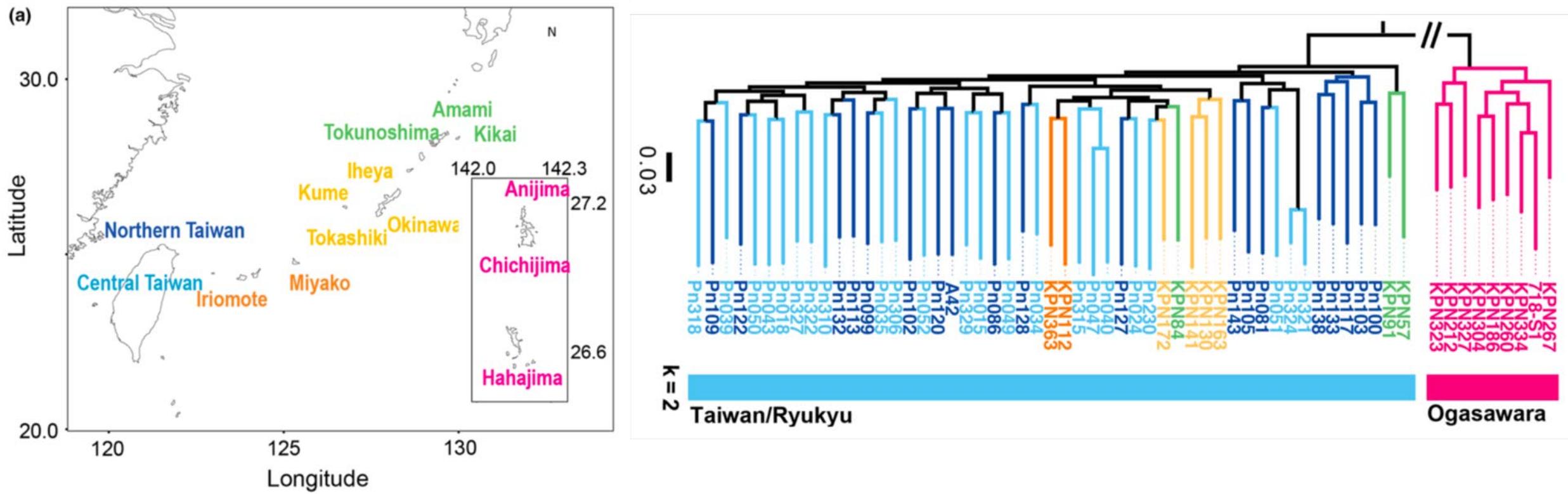
2017 – *Phellinus* genome project



Hymenochaetales



2017 – *Phellinus* genome project



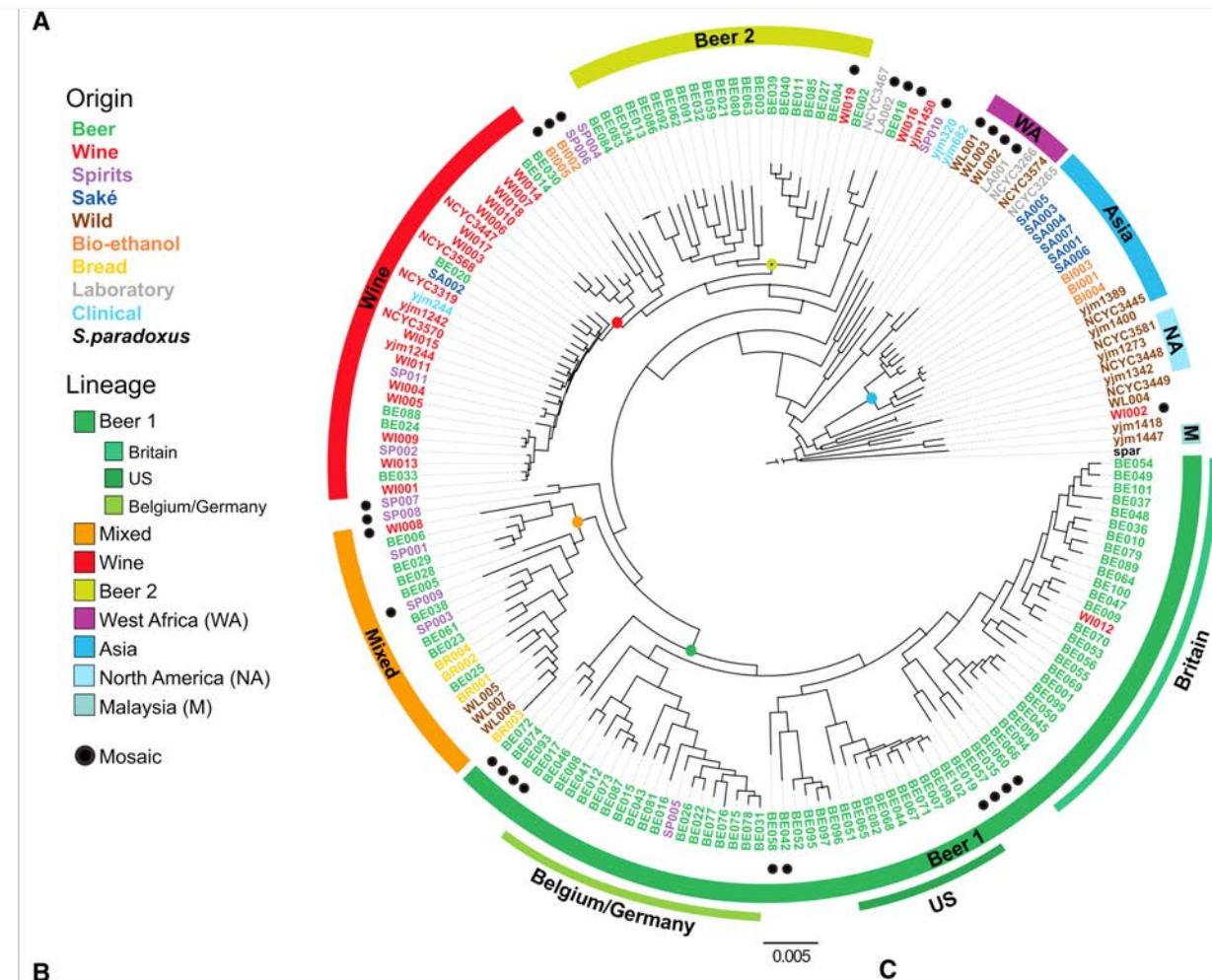
[In comparison]

Article

Cell

Domestication and Divergence of *Saccharomyces cerevisiae* Beer Yeasts

- Sequenced and phenotyped 157 *S. cerevisiae* yeasts
- Present-day industrial yeasts originate from only a few domesticated ancestors
 - Beer yeasts show strong genetic and phenotypic hallmarks of domestication
 - Domestication of industrial yeasts predates microbe discovery



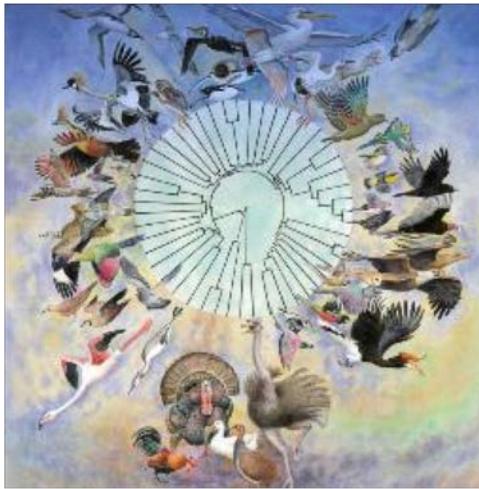
Shift in paradigm 2005-2018 (My personal take)

- A genome, a few genomes are no longer “enough”
 - ~since everybody can do it reasonably well
- Genome sequencing projects are
 - being done on a per-lab basis and no longer exclusive to sequencing centers
 - moving away from exploration to question orientated.
- Data being produced on a **much faster speed** at a **much higher throughput**, and a much **cheaper scale**
- More methods, analysis, tools, experiments...
 - Not always better

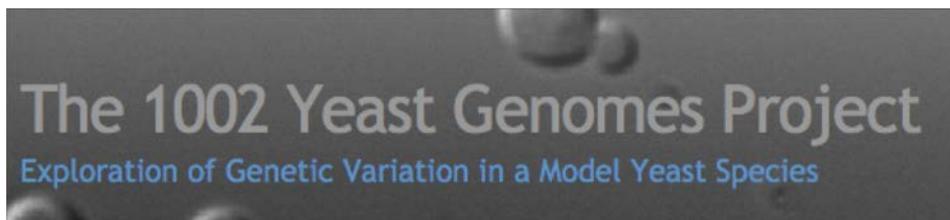
It is an exciting time to be in

Current and future

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



- Standard labs will be able to generate collections of themselves



(3 labs)

There's so much more...

- Read, read, read
- Twitter and blogs

Tweets Tweets & replies Photos & videos

You Retweeted

OfficialSMBE @OfficialSMBE · 23h
MBE latest: Large-effect beneficial synonymous mutations mediate rapid and parallel adaptation in a bacterium dlvr.it/KbShdx

6 4 ...

You Retweeted

OfficialSMBE @OfficialSMBE · 22h
GBE latest: Genome Resequencing Identifies Unique Adaptations of Tibetan Chickens to Hypoxia and High-dose... dlvr.it/KbSvMX

1 ...

You Retweeted

Justin Fay @justinfay · 19h
Check out our paper on *S. paradoxus* in Slovenian vineyards, including our first #vineyard #microbiome [journal.frontiersin.org/article/10.338...](https://journal.frontiersin.org/article/10.3389/fmicb.2018.01838)

8 9 ...

You Retweeted

Rob Waterhouse @rmwaterhouse · Feb 20
Trait databases, data quality, trees, genome structures, disease, biodiversity, @erichjarvis Ann.Rev. #birdgenomes

Erich Jarvis @erichjarvis
My perspective on questions that can be answered when all vertebrate genomes are sequenced @Genome10K @B10K_Project jarvislab.net/wp-content/upl...

1 1 ...

You Retweeted

Sujai @sujaik · Feb 20
For anyone following the ridiculousness in India, this is brilliant scroll.in/article/803856... @Sanjana2808 @karunanundy

1 ... View summary

You Retweeted

James Wasmuth @jdwaslmuth · Feb 19
Using #PacBio to gain a high-resolution phylogenetic microbial community profile bit.ly/1oR4qde

3 1 ...

Resources

Some very useful websites:

- <http://angus.readthedocs.io/en/2017/>
- <http://evomics.org/>
- <https://github.com/schatzlab/appliedgenomics2018>

Written assignment

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