Introduction

Isheng Jason Tsai

Introduction to NGS Data and Analysis
Lecture 1

Welcome!

This course is called “Introduction to Next-Generation Sequencing (NGS) Data and Analysis”

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?)

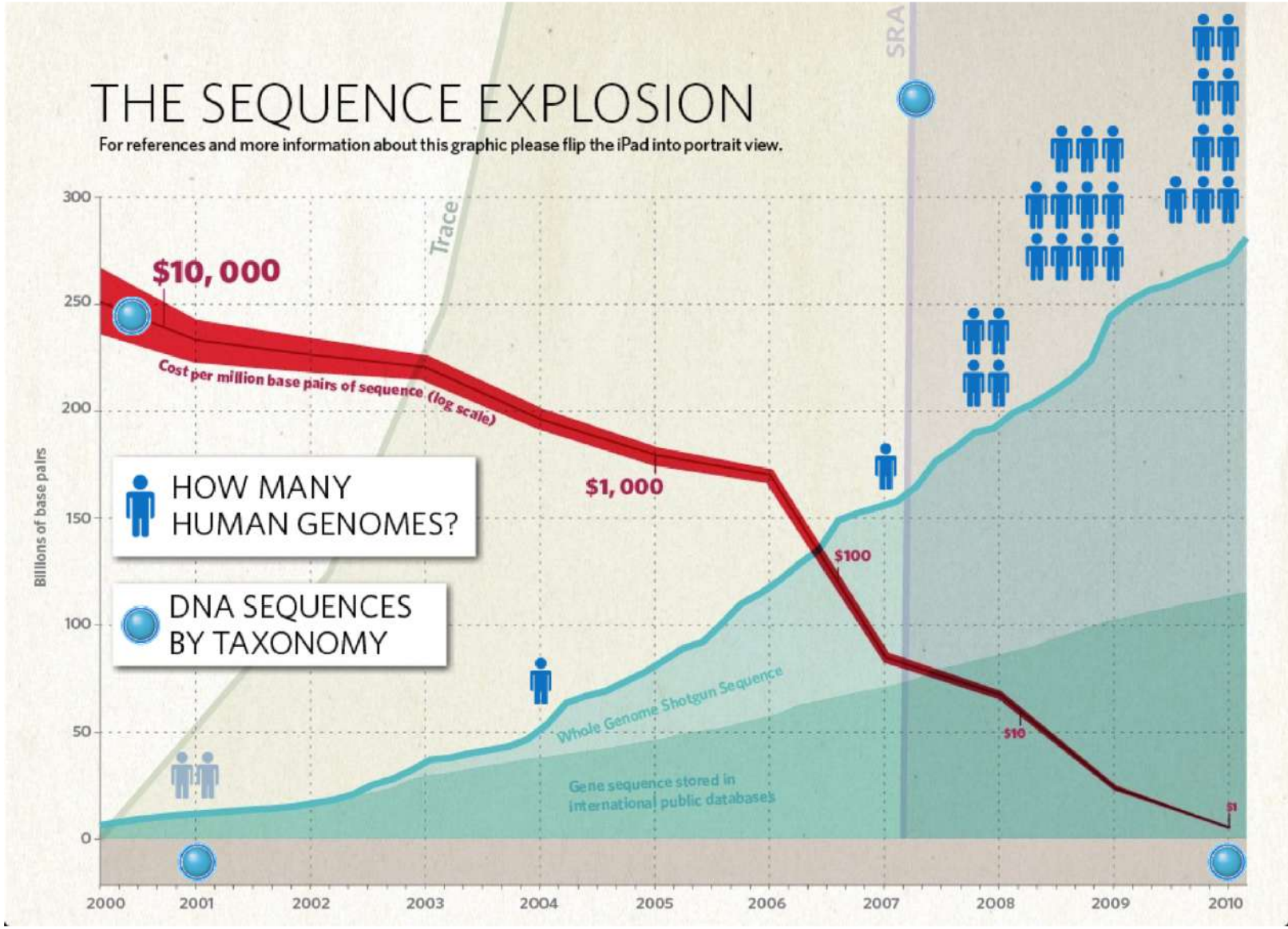
TIGP Introduction to NGS

1. Introductory lecture Download
2. Linux and R ; basic usage Download
3. Genome Assembly and case studies Download
4. Mapping and Case studies Download
5. From Alignment to phylogeny (Jiang Ming Chang) Download
6. DNA/RNA preparation and different sequencing technologies (Meiyeh Lu) Download
7. RNAseq and Genome annotation Download
8. Comparative Genomics Download
9. Population Genomics (John Wang)
10. Amplicon / Metagenomics Download
11. Practical one: Linux and R
12. Practical two: RNAseq mapping and EdgeR
13. Discussion
14. Final Report

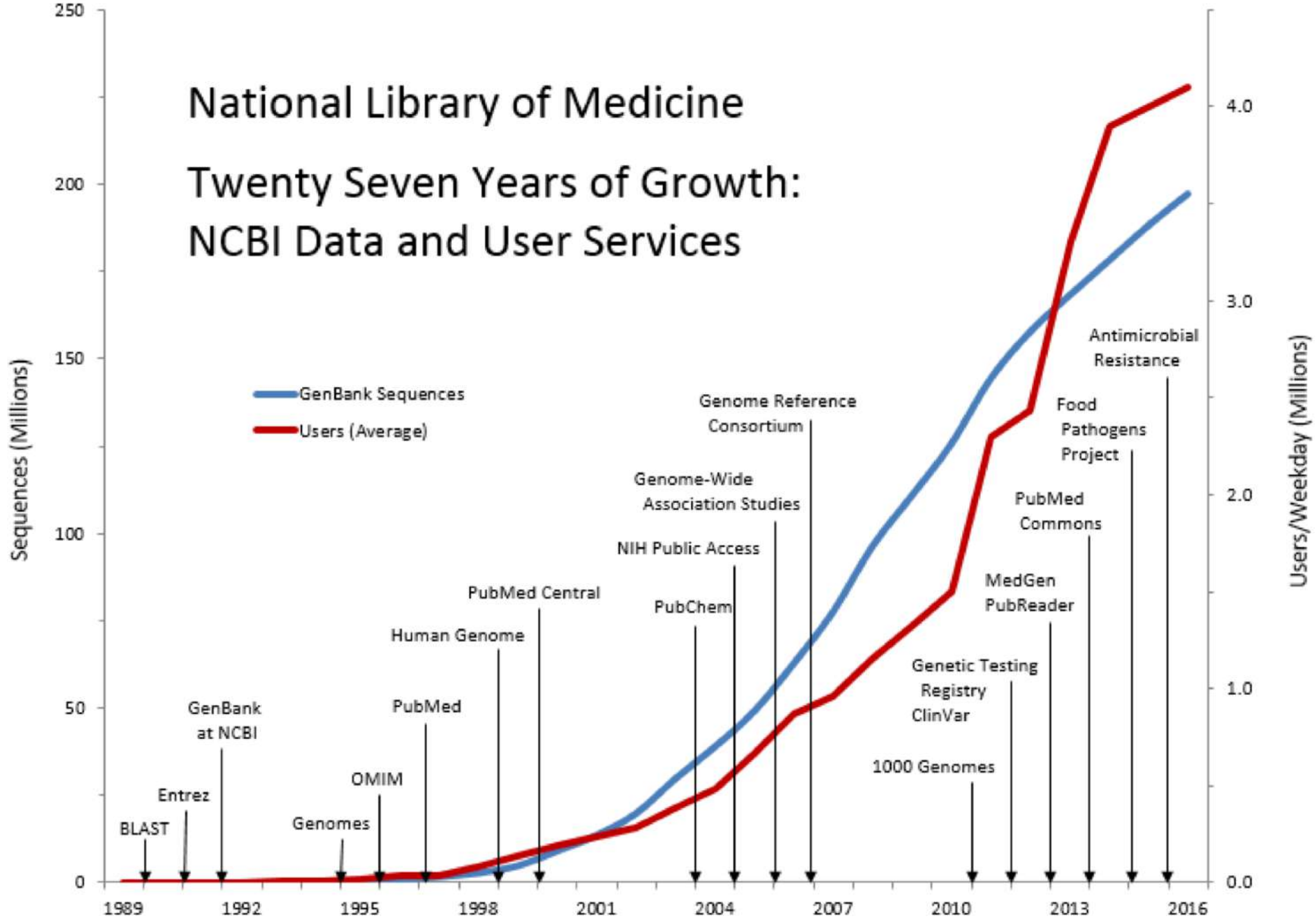
http://2015ngsintro.readthedocs.io/en/latest/lec.html

What is NGS?

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



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



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

NGS = sequencing made cheaper, faster and higher throughput

Course setup

The primary goal of this course is for students to get familiar with theories behind the sequencing field…

as well as how we do things

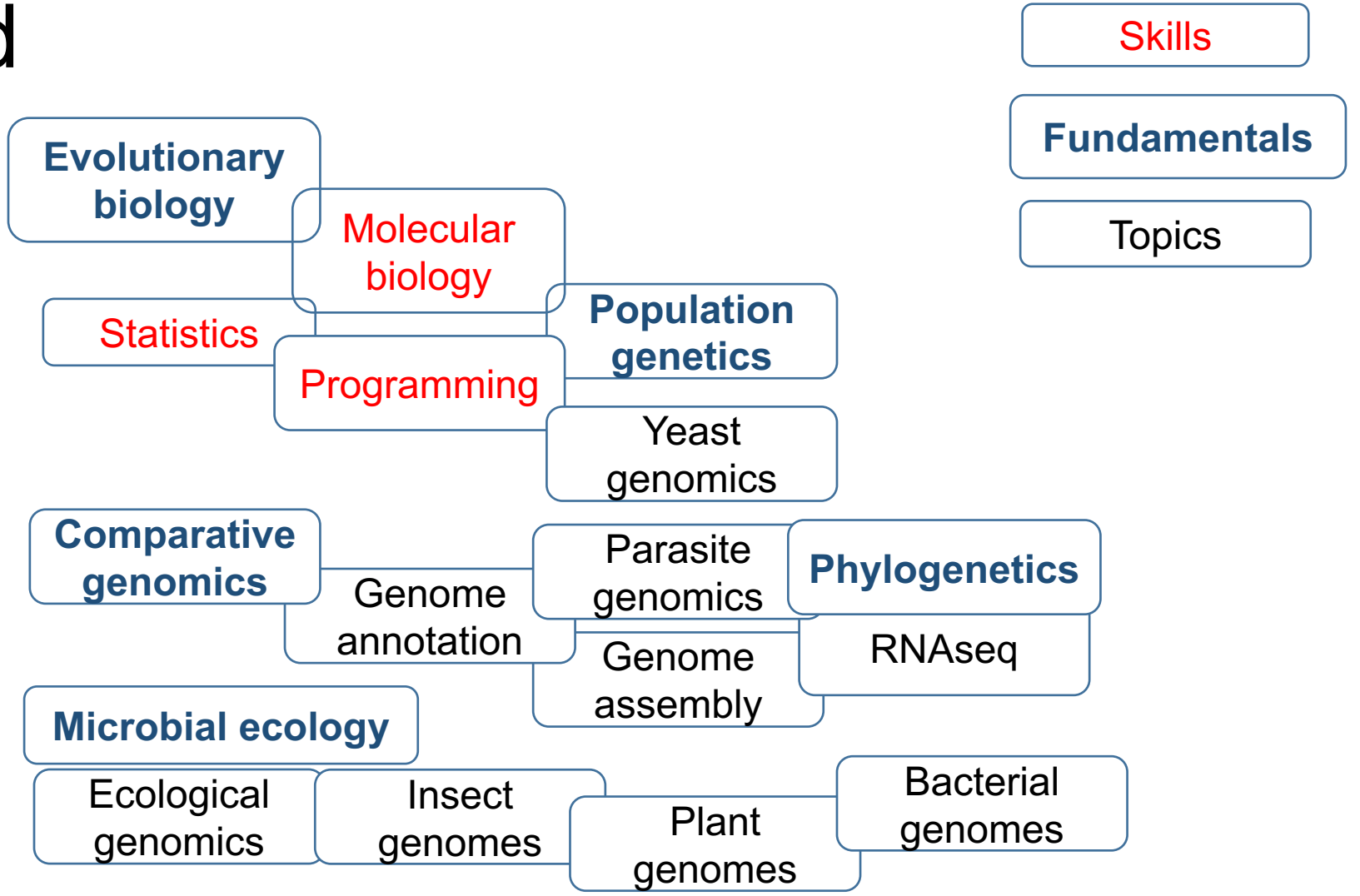
My background

Undergraduate:
Biochemistry and Genetics

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

2009-14 ; Postdoc:
Genomics & parasitology

2015 - ; Academia Sinica:
Microbial diversity & Bioinformatics



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

Goals of this course

- Expose you to all the theories behind genomics and transcriptomics
- The techniques will be applicable across all biological science disciplines
- show you how a question was formulated and how was it tackled

However, this module will **NOT**

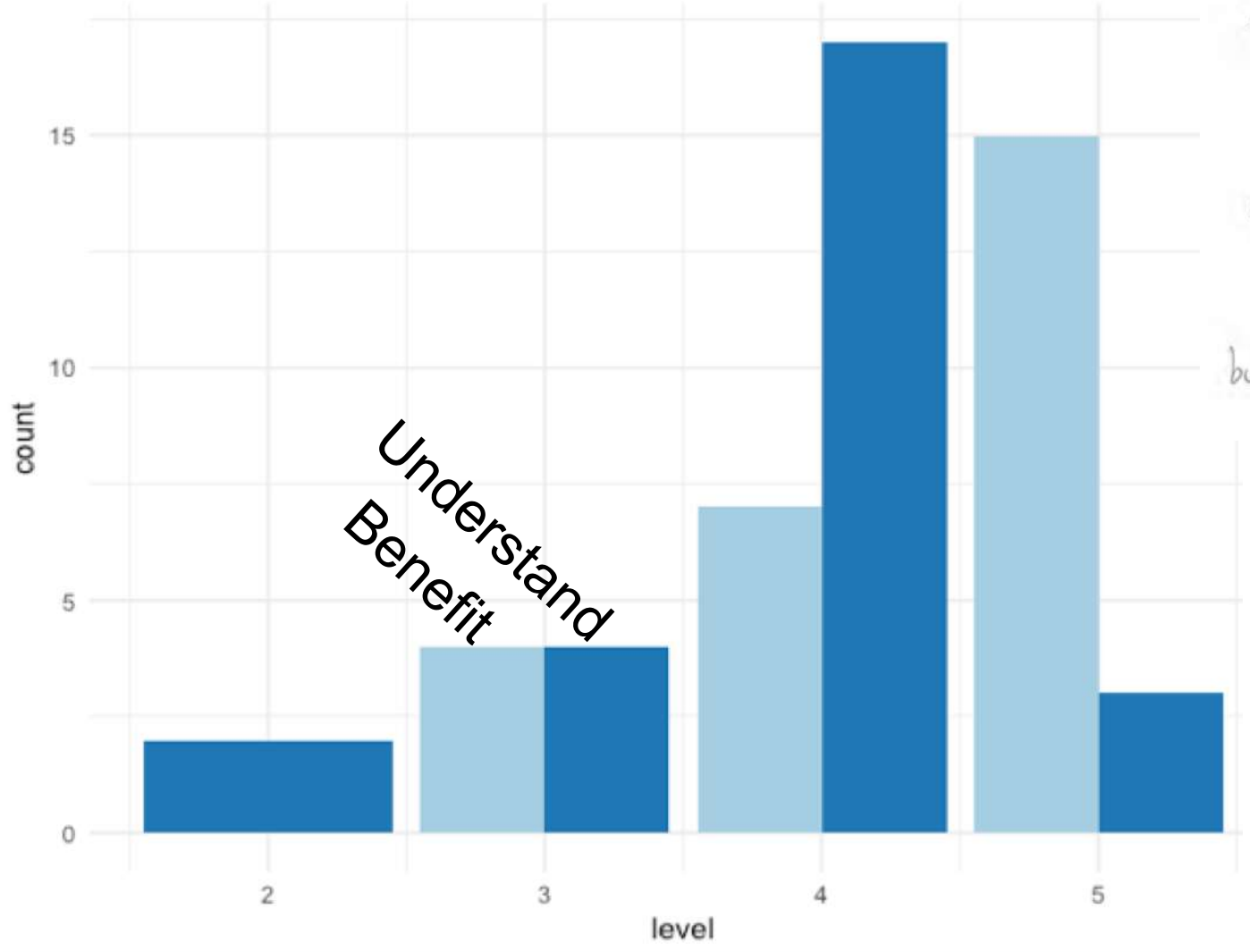
- teach you the core skills such as programming languages (although some limited exercises will be run)
- focus on human genomics but more ecological context in general

At the end of the course, students should find out how sequencing can be used in all aspects of biological sciences

I run this course every 2 years. This is the second time running

- 39 attended the first time
- 3 PIs and 6 postdocs
- 3 universities and 4 institutes

Comments last time



3. Any subject you would like to be included in this course:
more interaction or group homework.

4. How could this course be more helpful to you?
the lecture slides will are great outline of ~~paper~~ papers worth reading.

5. Constructive comments or suggestions to the lecturer/organizer:
~~maybe should have more~~
like the way prof. Tsai organized the lecture.
but if there more homework will be better.
(little, interactive one)

Grading policies

- 40% Assignments
  - 4-5 written homework to be handed in 2 weeks later (20%)
  - 1 proposal + presentation which will be marked by your own peers (20%)

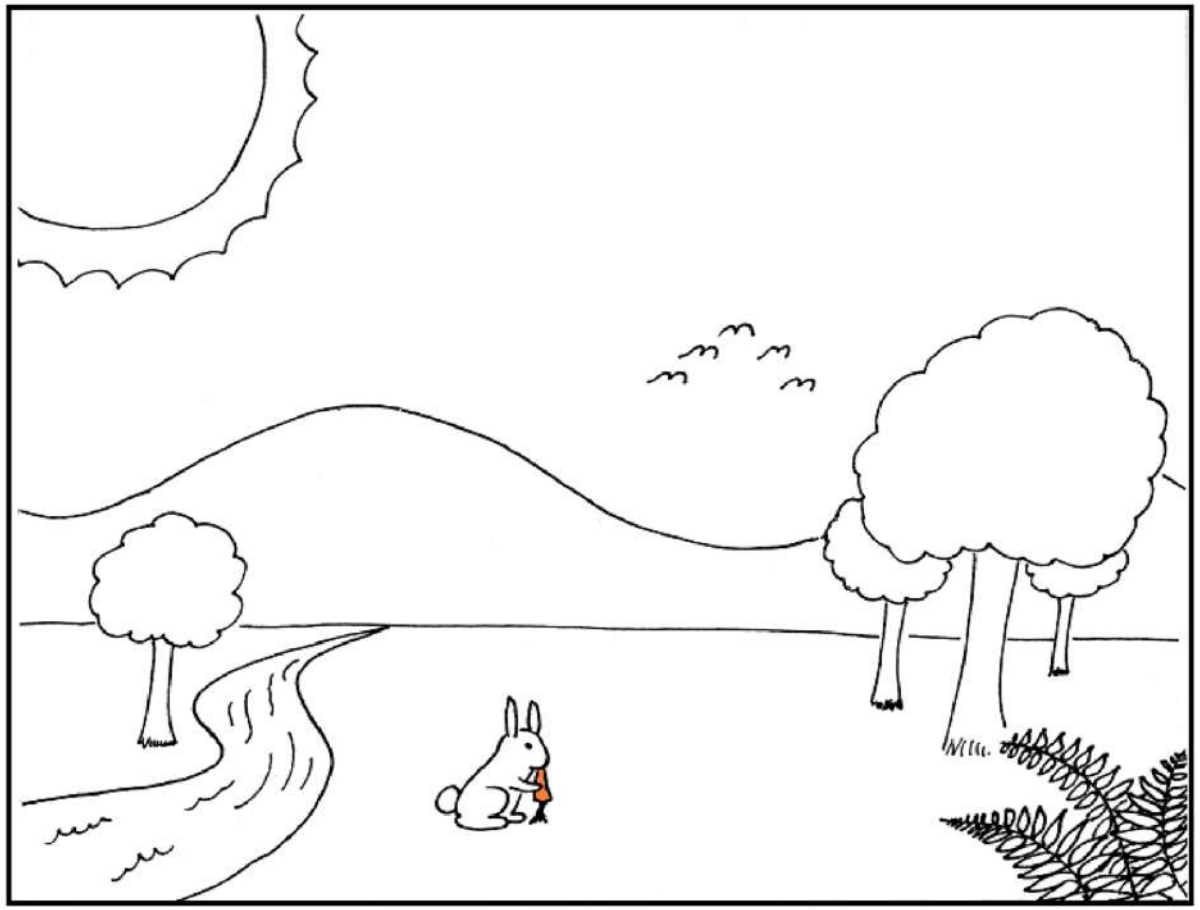
- 40% Written exam
  - Test on your understanding

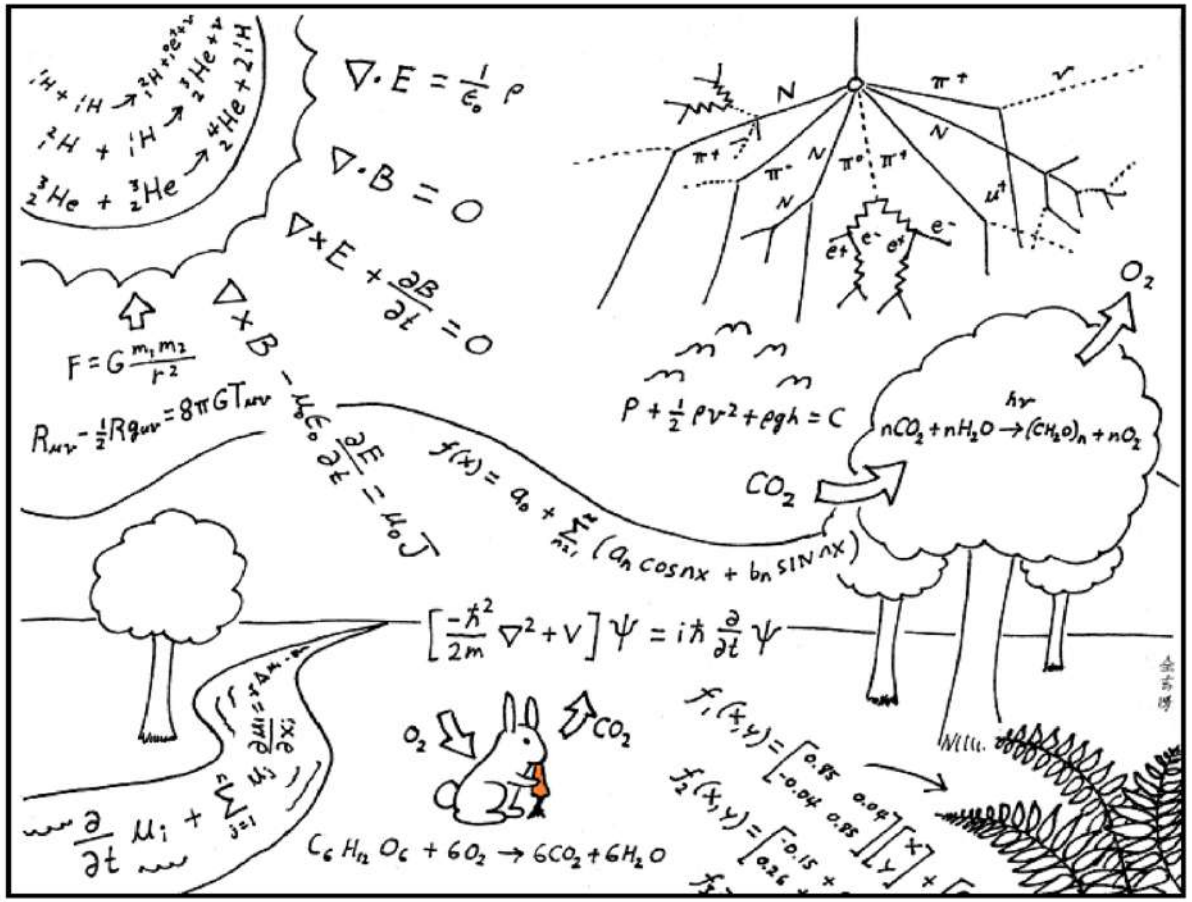
- Attendance: 20%

- Scores assigned relative to the highest points awarded

Always start with a question

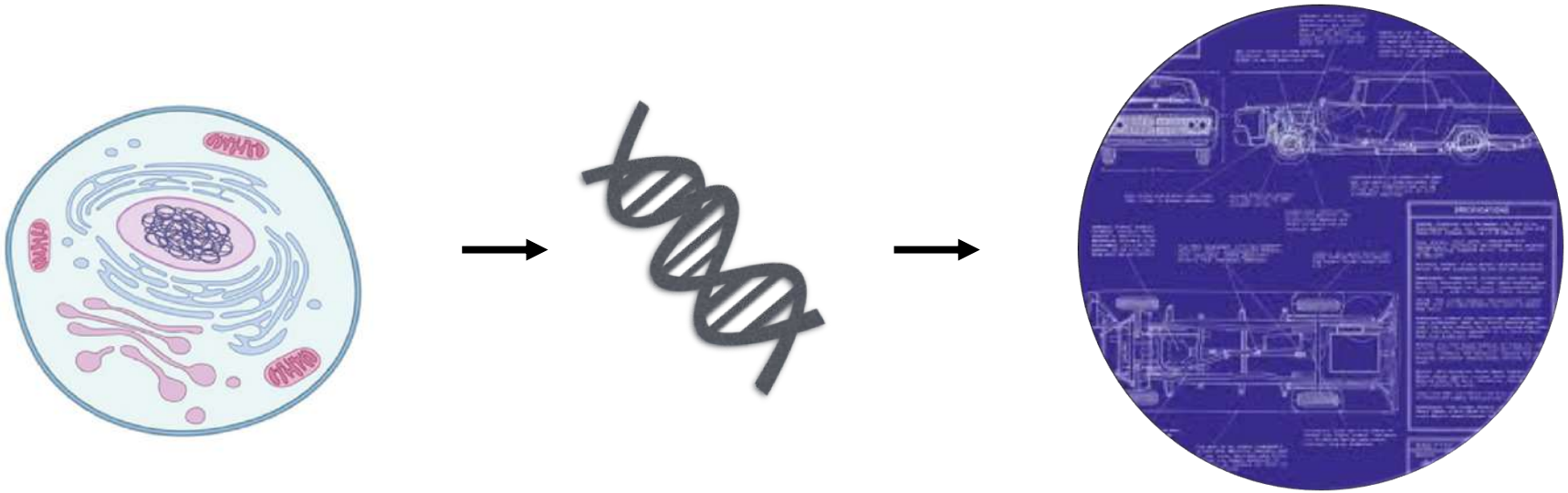
This is how scientists see the world





How? Who? Where? What?

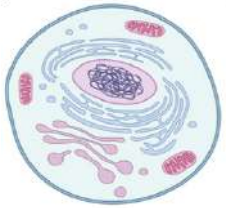
Genome



Genome = Parts list of a single genome

A genome project

Wet lab work



Extract DNA
(or RNA)

(Lecture 7)

Shear DNA
(or RNA)

Sequence

Bioinformatics

Data QC
(Lecture 4,8)

Variant
(Lecture 9-11,15)

ATCG
ATGG
ATCG

Annotation
(Lecture 8)

DNA or RNA Reads
50-500 bp

Mapping (Lecture 4)
RNAseq (Lecture 8)

Reads
50-500 bp

Assembly
(Lecture 3)

Contigs
1kb – 100 kbp

Scaffolding
(Lecture 3)

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

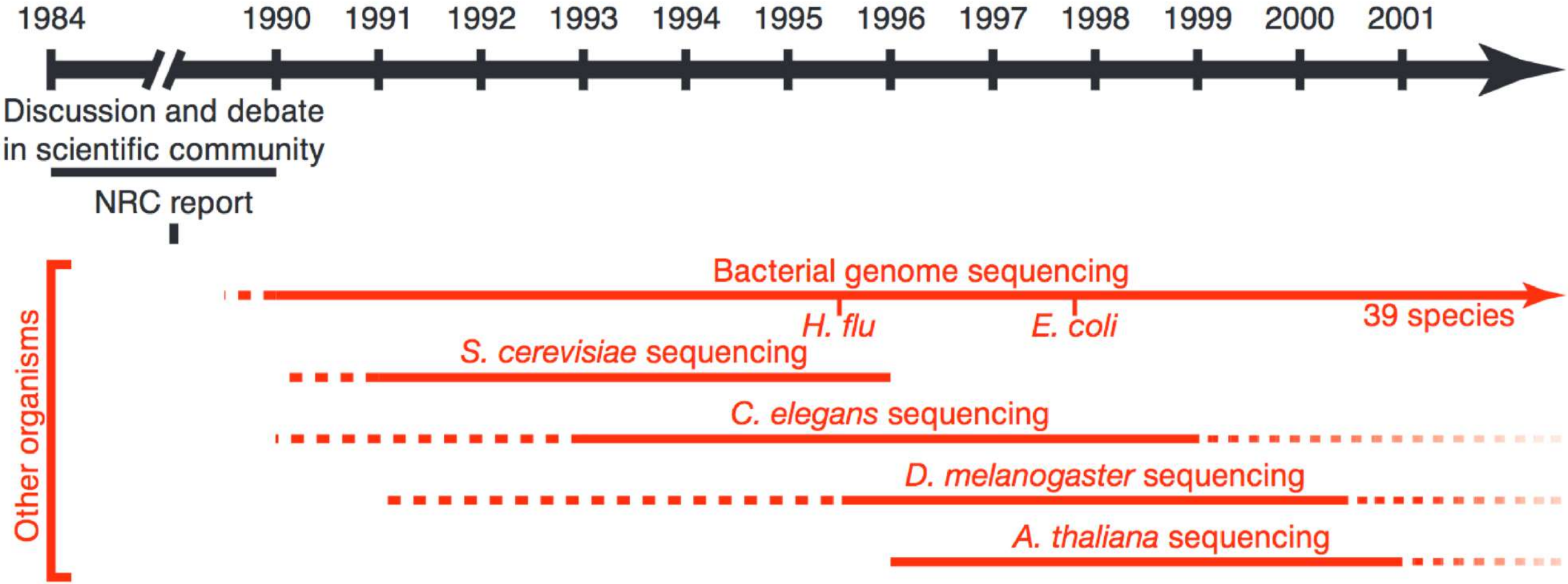
What we think we do

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

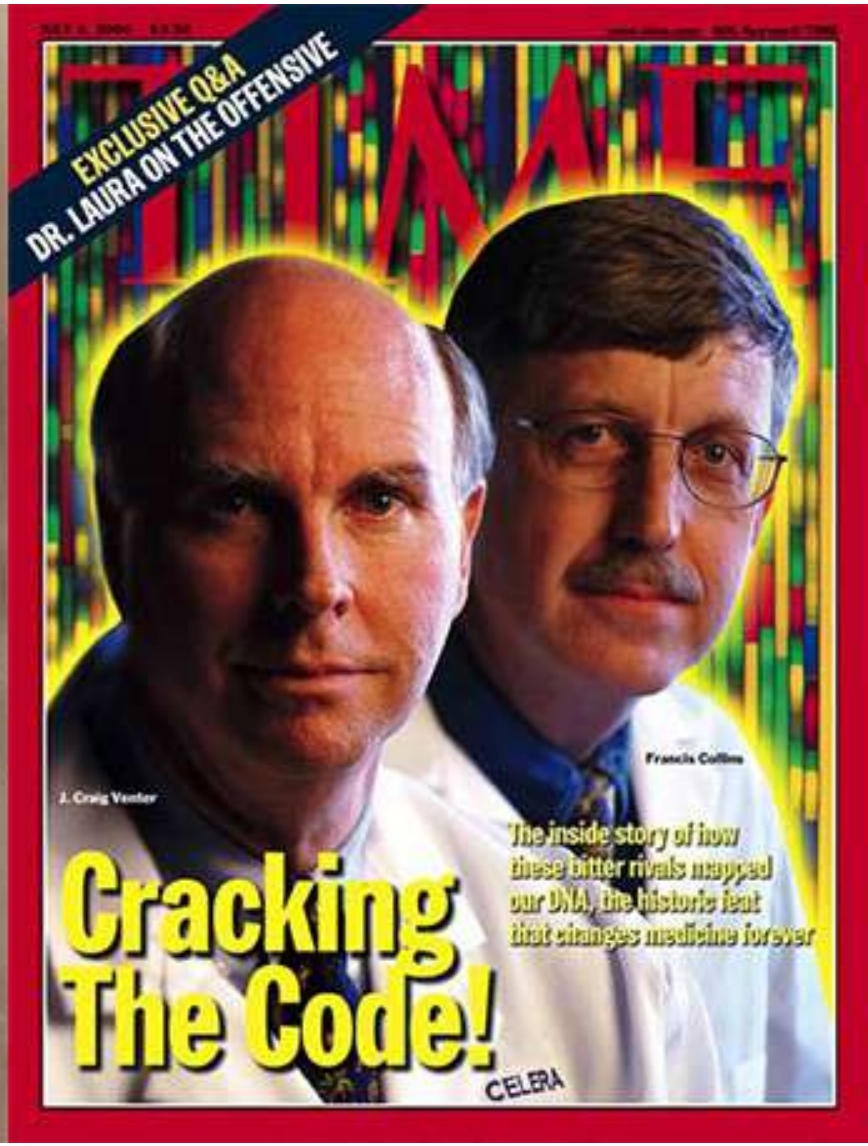
- **Help to understand biology**
- ~~Some lab just had the money~~ ; don’t do it



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


nature
the human genome


Science
THE HUMAN GENOME
AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE


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

Calculating the economic impact of the Human Genome Project

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

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

FOCUS ON GENOMES OF ICELANDERS

ARTICLES

nature genetics

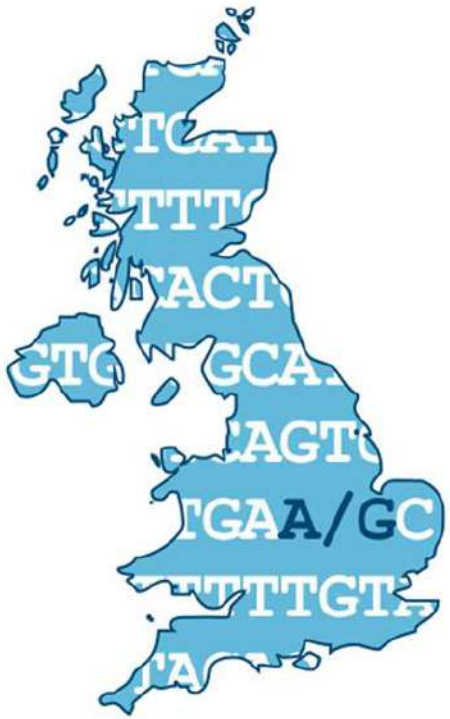
Large-scale whole-genome sequencing of the Icelandic population

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

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

The blood of a thousand Icelanders.
Photo: Chris Lund

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



UK 10K

RARE GENETIC VARIANTS IN HEALTH AND DISEASE

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

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

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

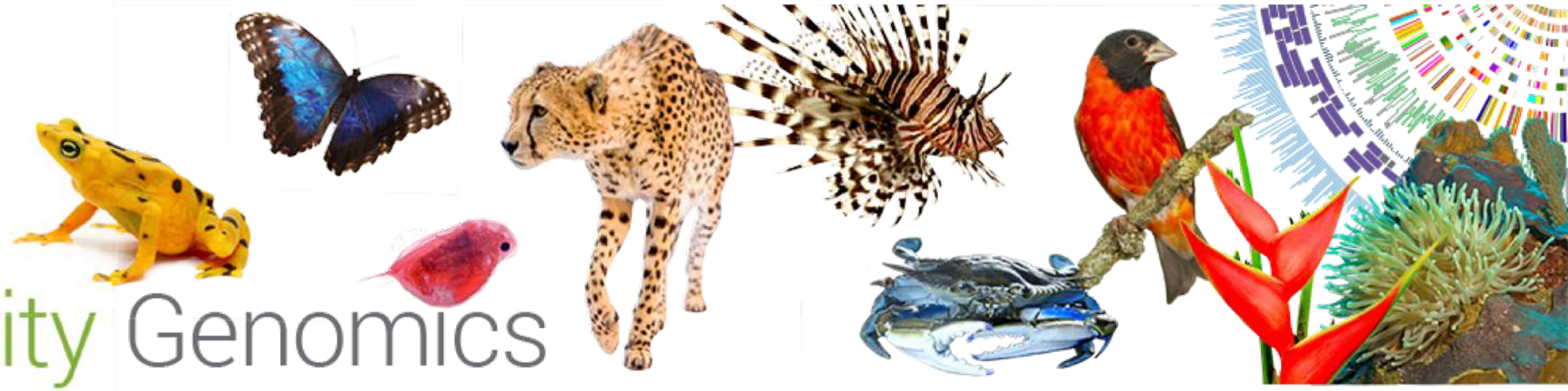
The project received a **£10.5 million** funding award from Wellcome in March 2010 and sequencing started in late 2010. For more information, please use the links on the right hand side.

https://www.uk10k.org/

Smithsonian

Institute for

Biodiversity Genomics

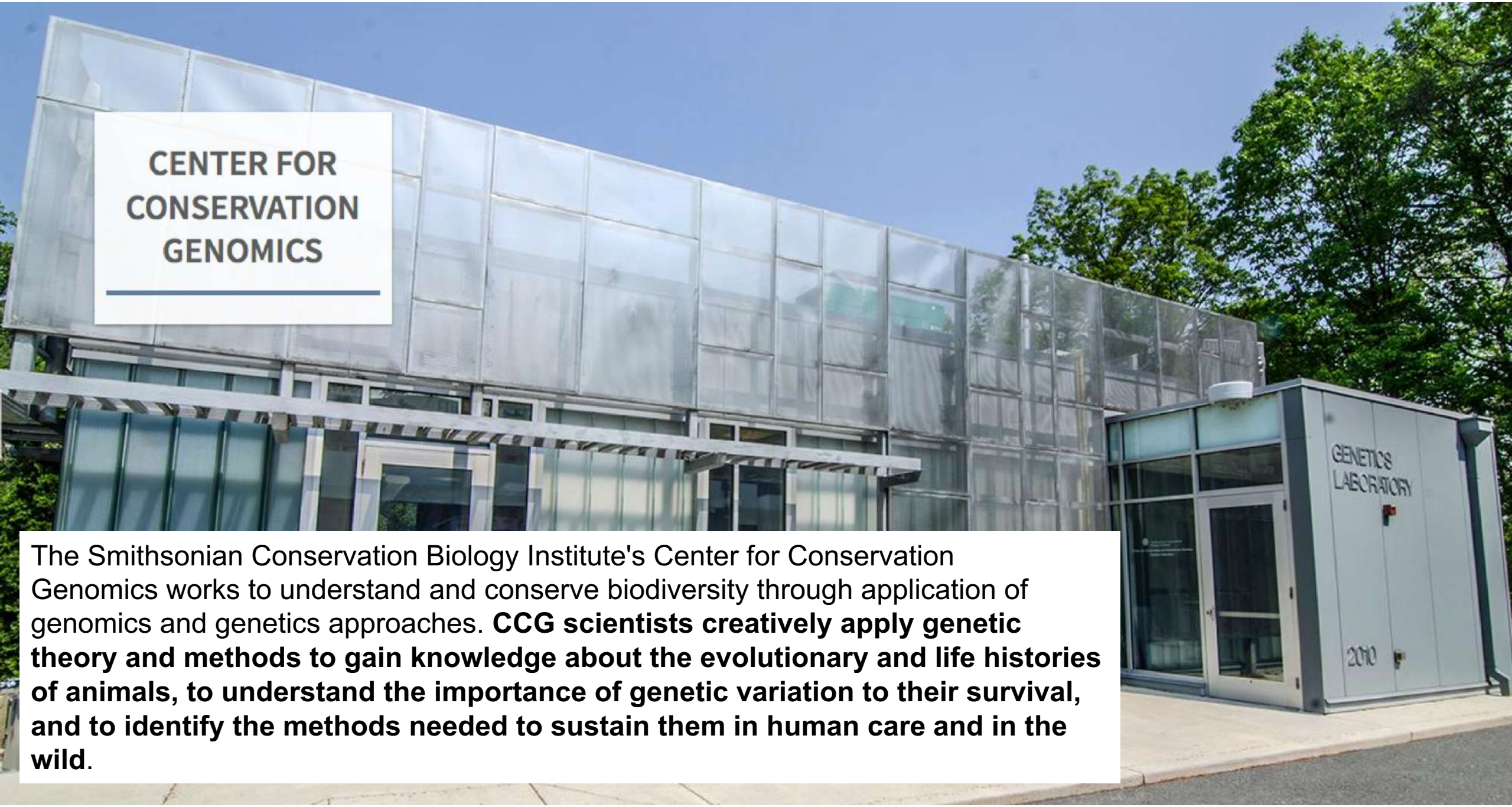


How do we sustain life on our changing planet?

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

But our biodiversity faces serious challenges.

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



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

Environmental genomics

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

EARTH BIOGENOME PROJECT

A GRAND CHALLENGE

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

A GRAND VISION

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

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.

https://thescienceweb.wordpress.com/2015/07/27/most-people-doing-genomics-not-actually-doing-genomics/

NGS dos and don't

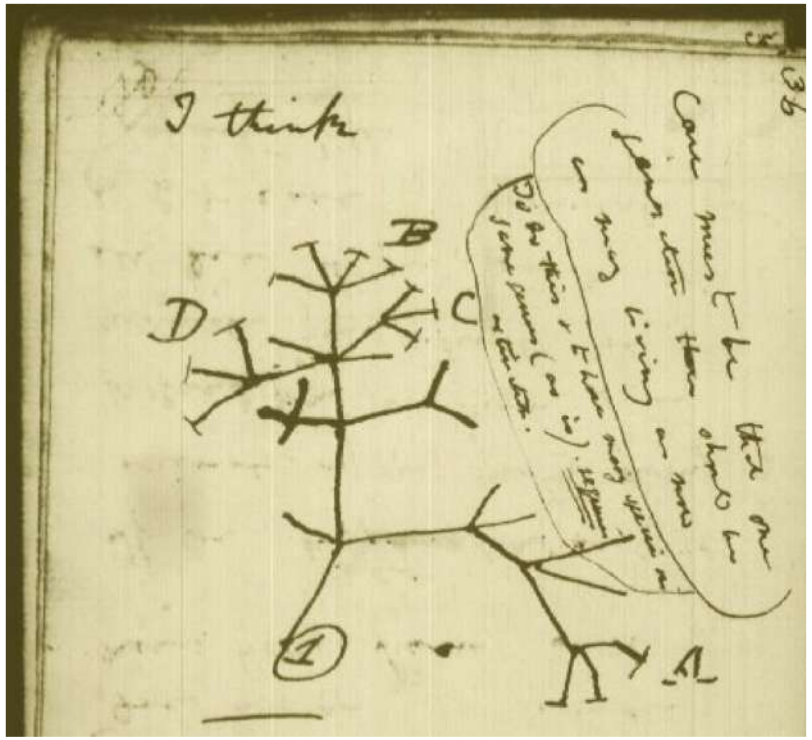
- Embrace it

- 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

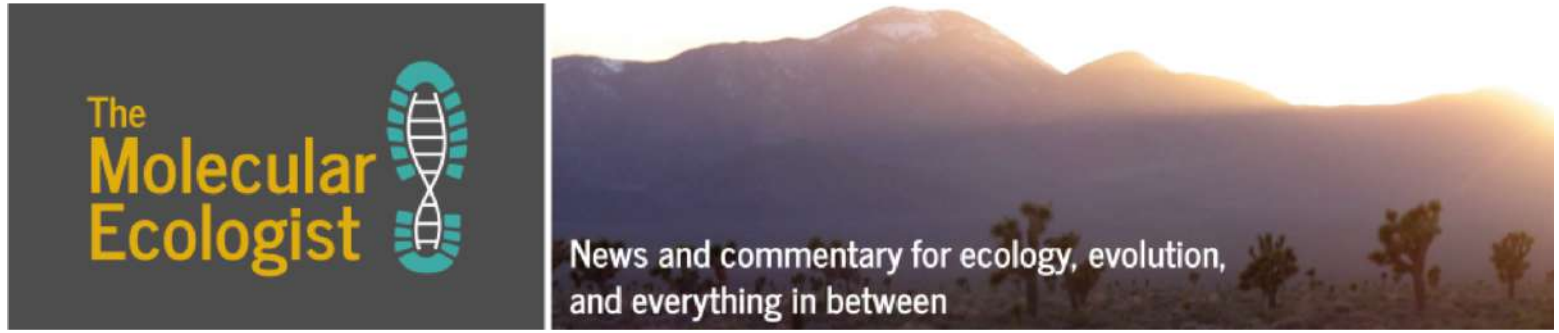
- Total number of students: **39**

- 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**

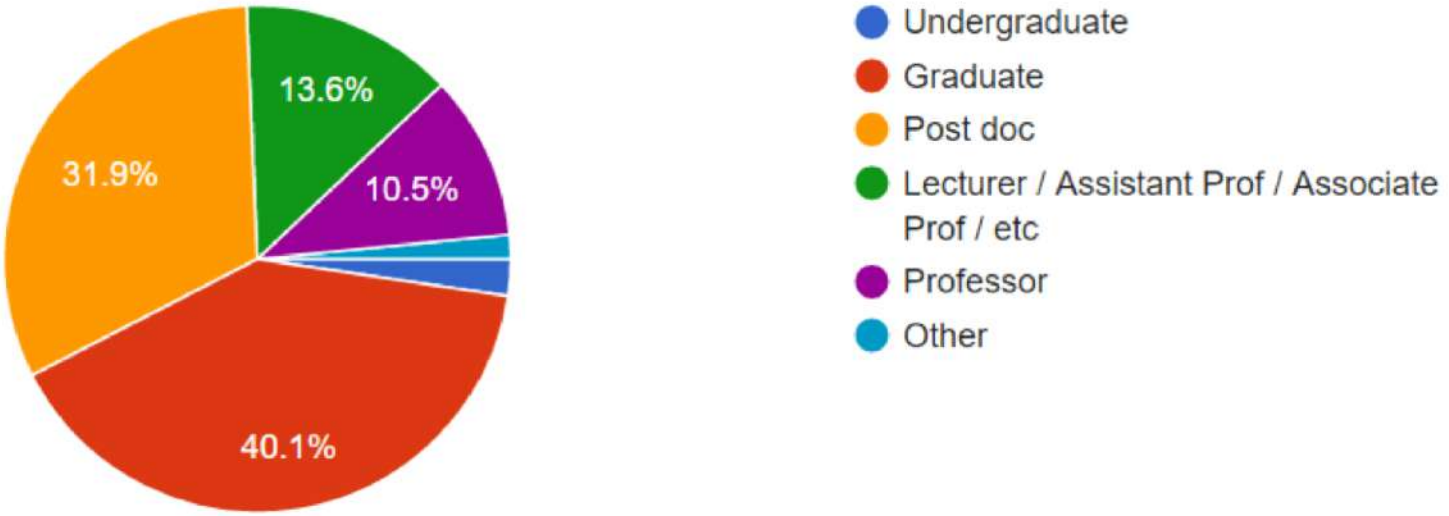
- Familiarity with Linux environment? **5**
- Programming experiences? **6**

Who's doing NGS? (Ecological context)

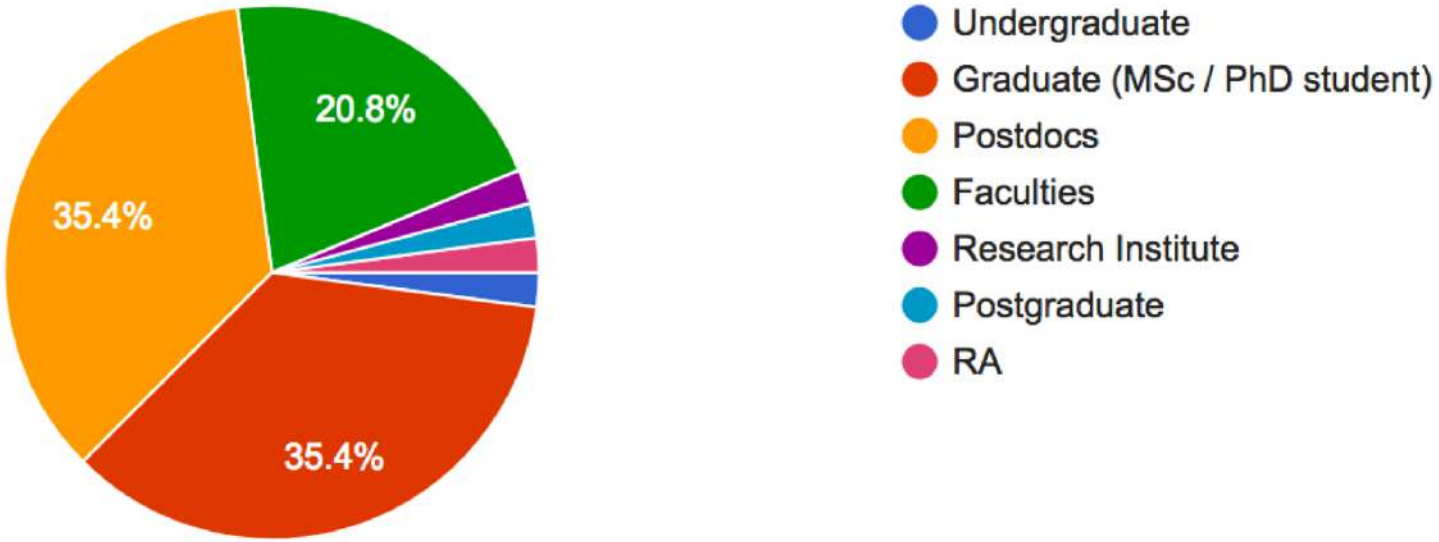


http://www.molecularecologist.com/2016/04/results-of-the-molecular-ecologists-survey-on-high-throughput-sequencing/

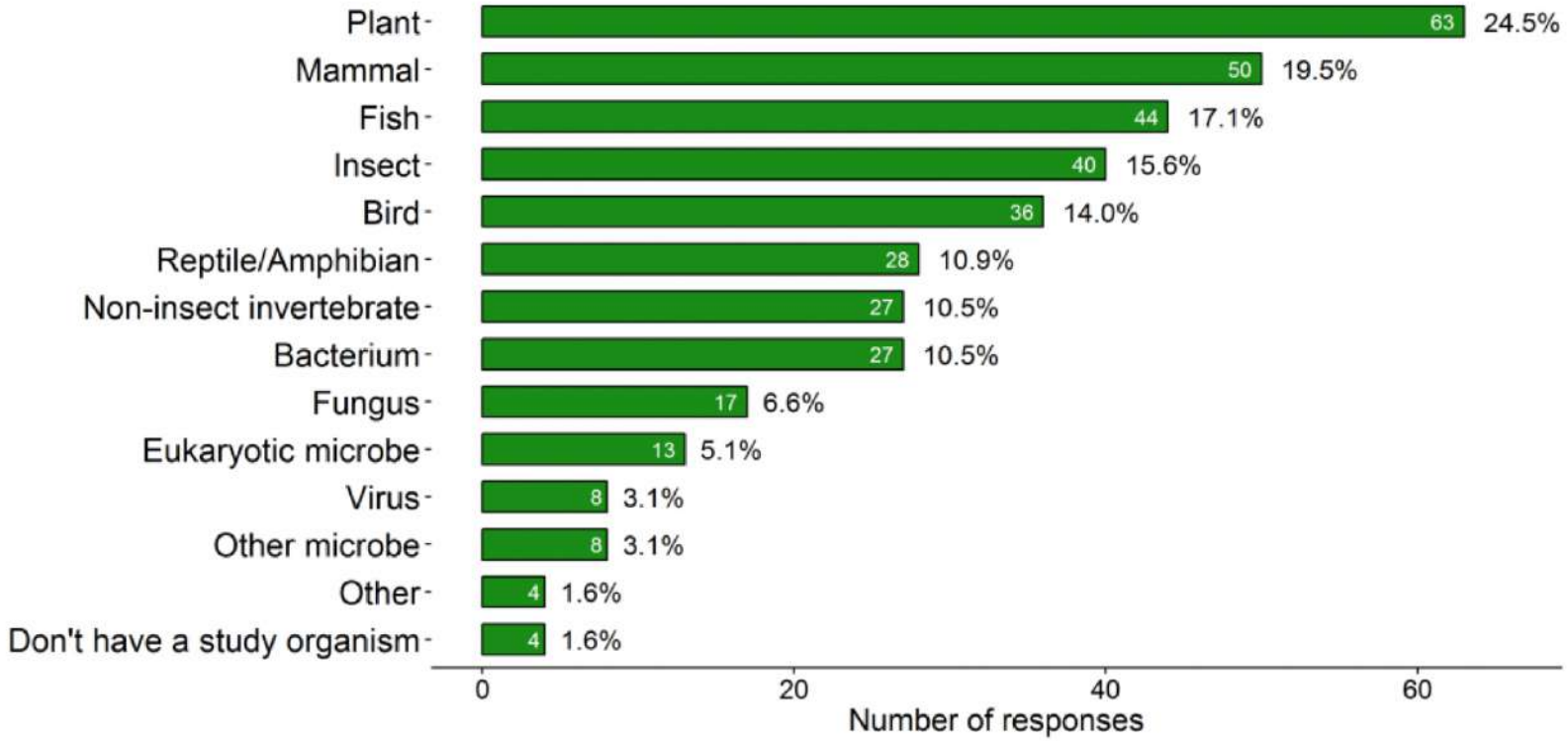
Which level are you at? (257 responses)

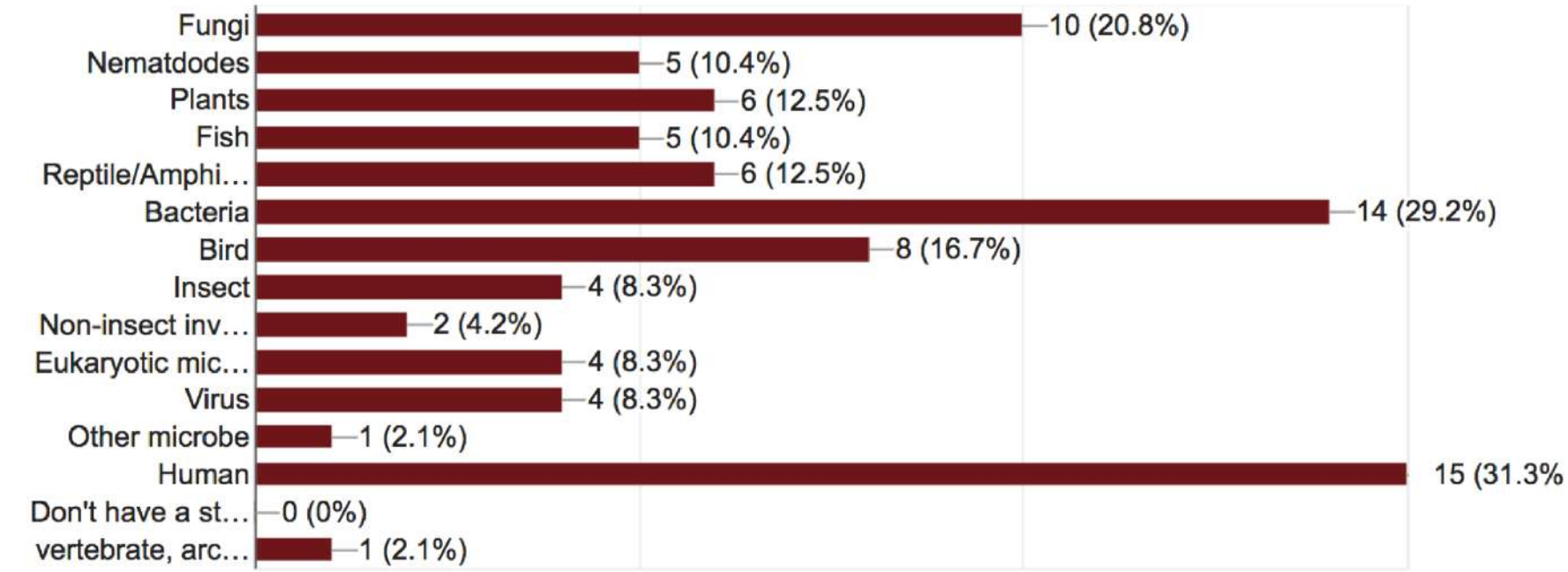


Rest of the world

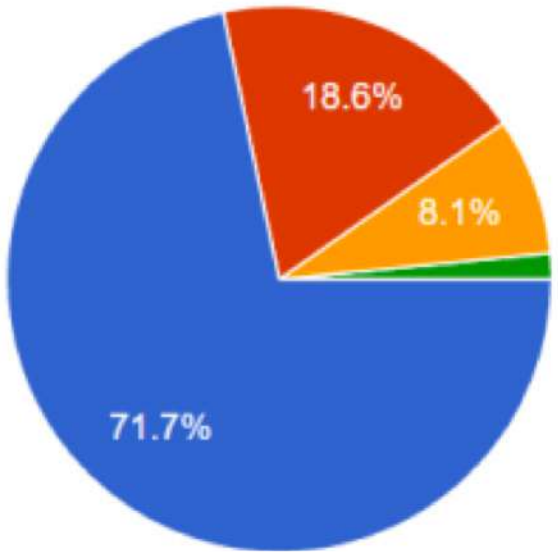


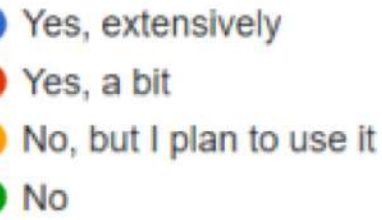
Taiwan

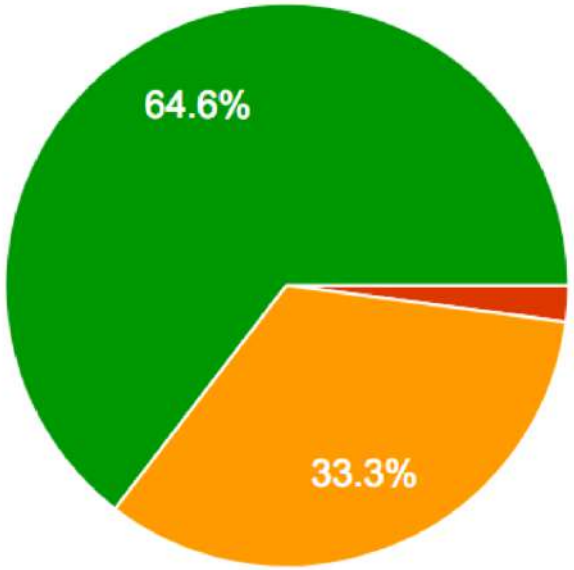

Plant
63
24.5%
Mammal
50
19.5%
Fish
44
17.1%
Insect
40
15.6%
Bird
36
14.0%
Reptile/Amphibian
28
10.9%
Non-insect invertebrate
27
10.5%
Bacterium
27
10.5%
Fungus
17
6.6%
Eukaryotic microbe
13
5.1%
Virus
8
3.1%
Other microbe
8
3.1%
Other
4
1.6%
Don't have a study organism
4
1.6%
0
20
40
60
Number of responses

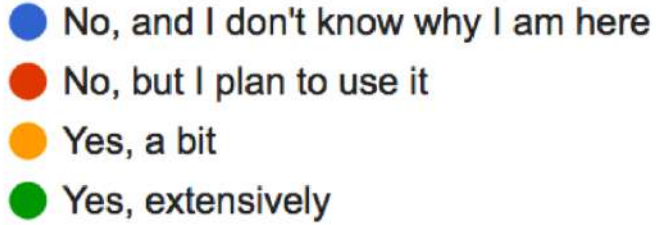

Fungi
10 (20.8%)
Nematdodes
5 (10.4%)
Plants
6 (12.5%)
Fish
5 (10.4%)
Reptile/Amphi…
6 (12.5%)
Bacteria
14 (29.2%)
Bird
8 (16.7%)
Insect
4 (8.3%)
Non-insect inv…
2 (4.2%)
Eukaryotic mic…
4 (8.3%)
Virus
4 (8.3%)
Other microbe
1 (2.1%)
Human
15 (31.3%
Don't have a st…
0 (0%)
vertebrate, arc…
1 (2.1%)

Do you use NGS techniques in your research?

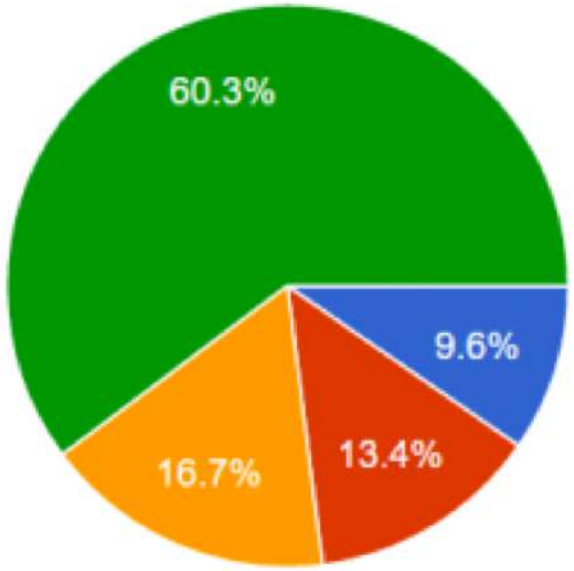


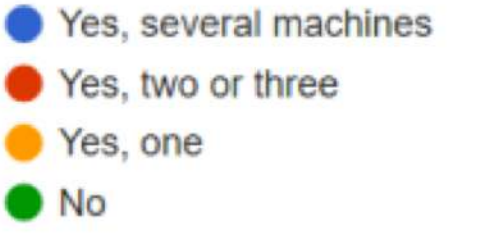


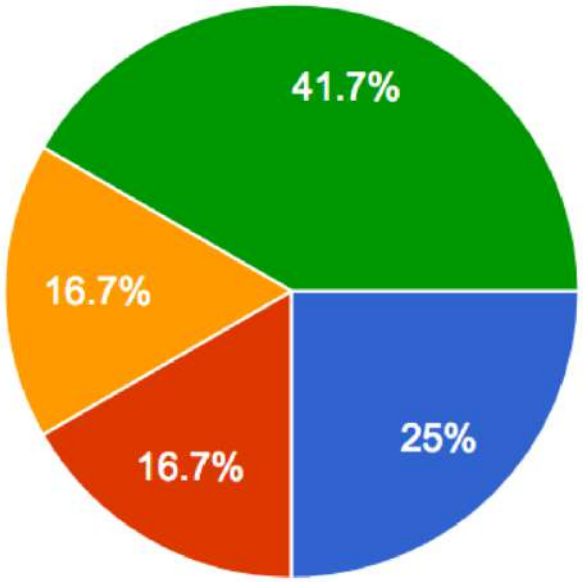


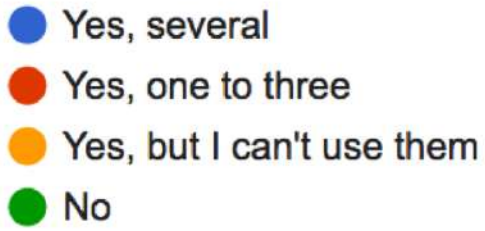


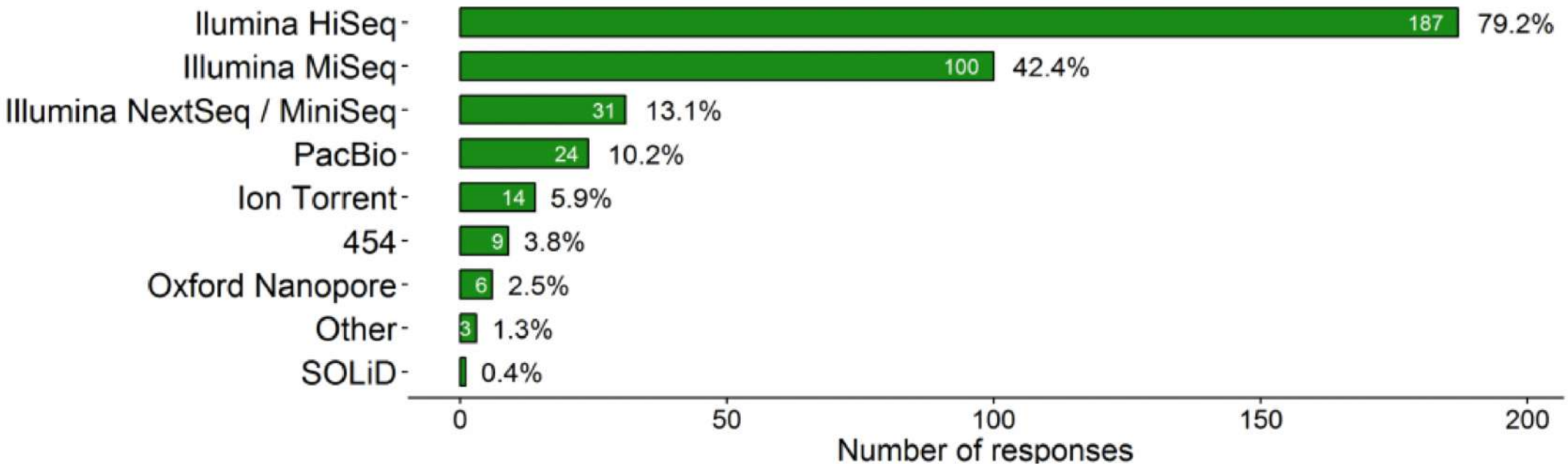
Does your lab have their own sequencing machines?

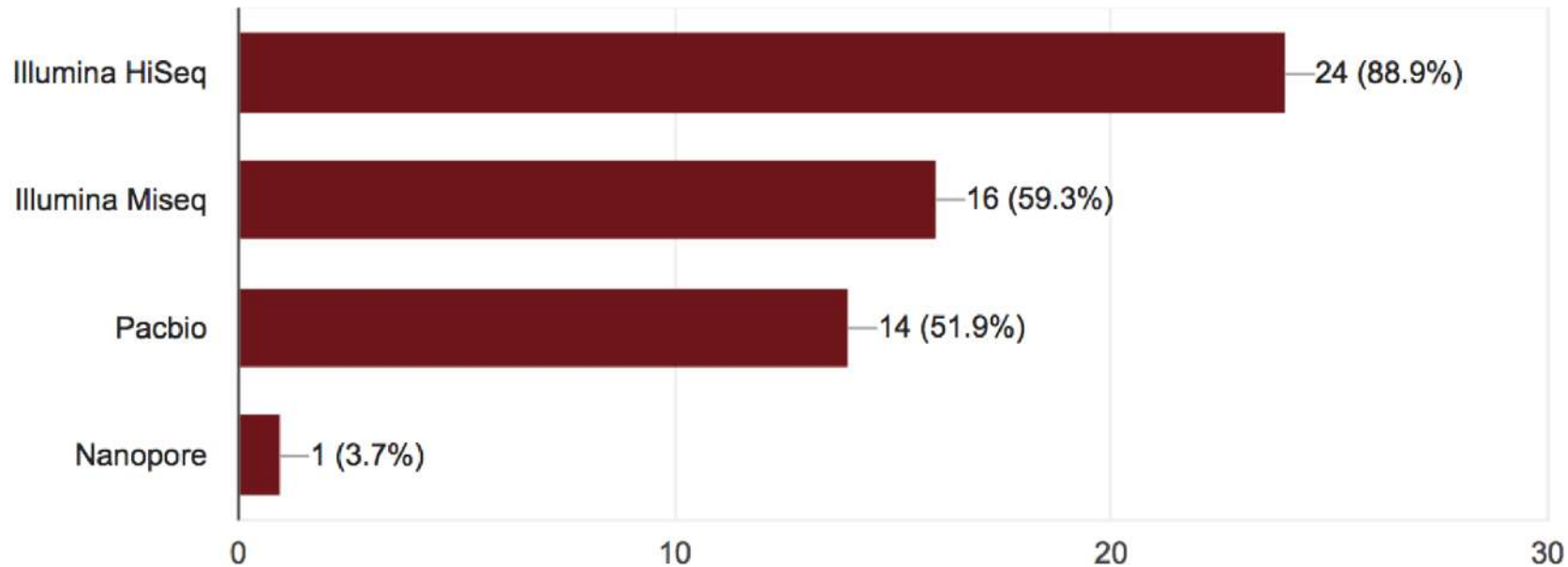


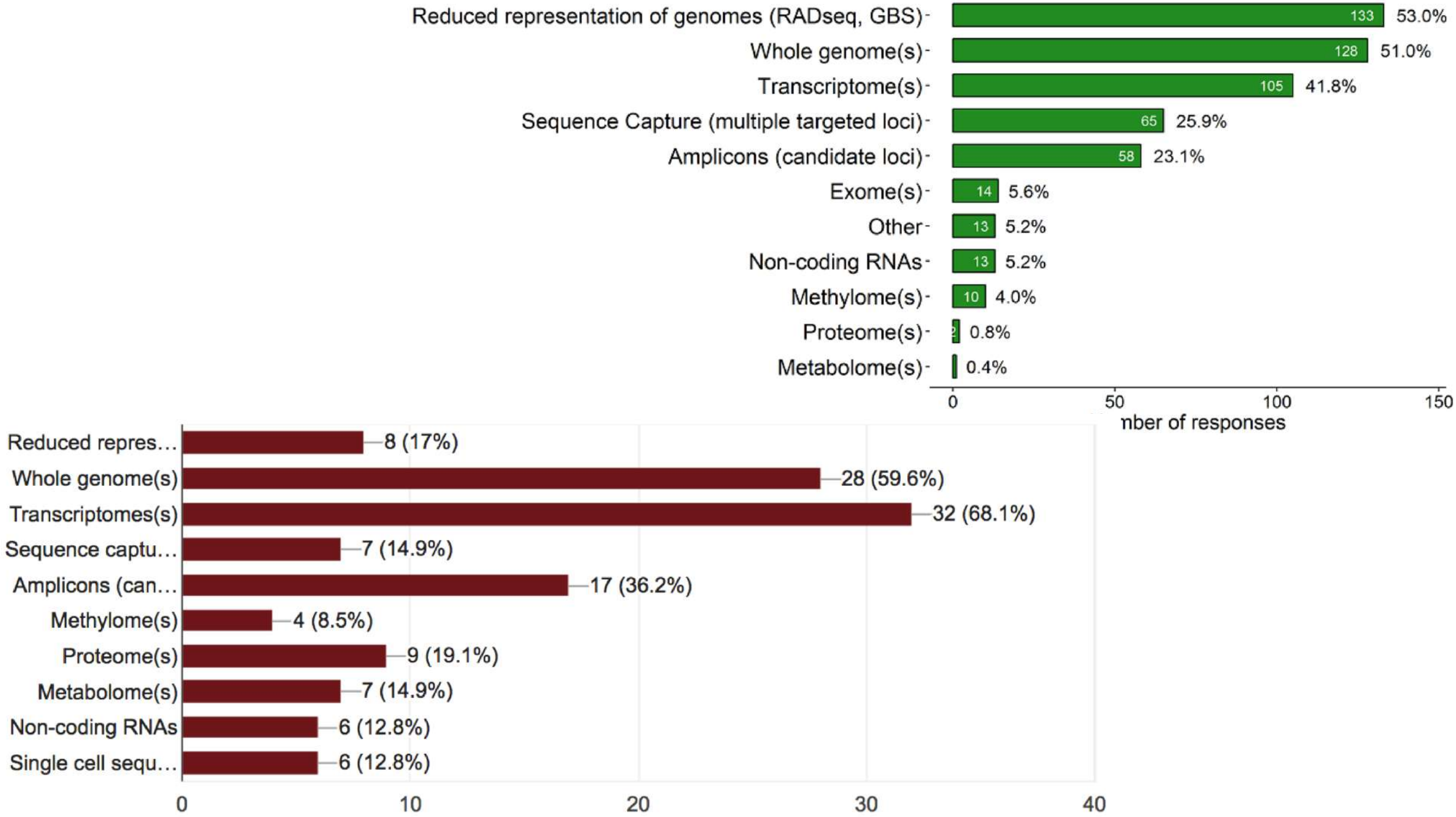




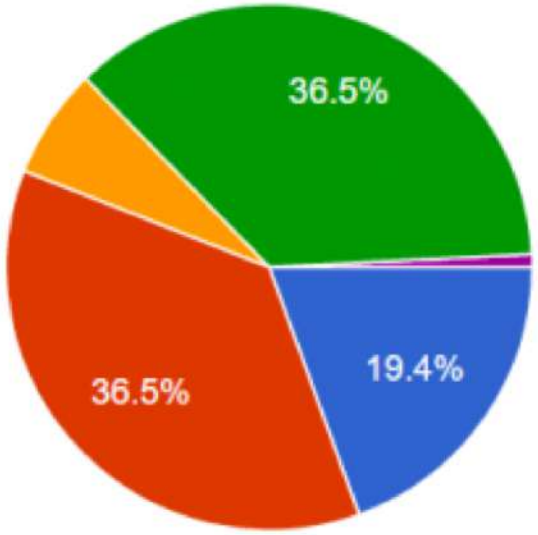



Ilumina HiSeq
Illumina MiSeq
Illumina NextSeq / MiniSeq
PacBio
Ion Torrent
454
Oxford Nanopore
Other
SOLiD
187
79.2%
100
42.4%
31
13.1%
24
10.2%
14
5.9%
9
3.8%
6
2.5%
3
1.3%
0.4%
0
50
100
150
200
Number of responses

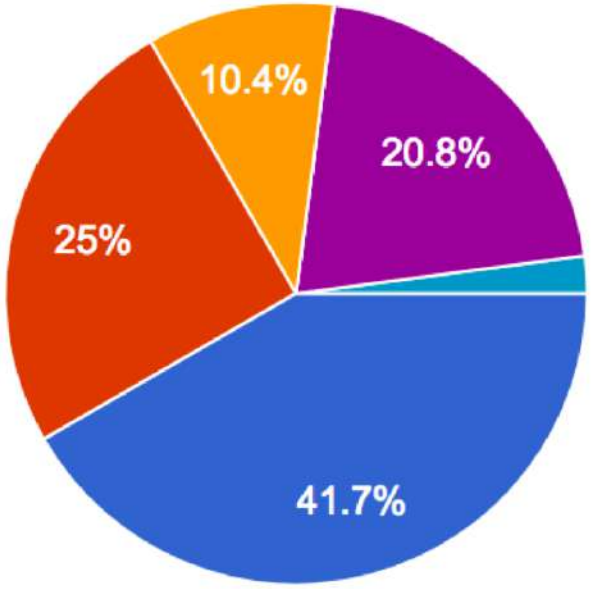

Illumina HiSeq
Illumina Miseq
Pacbio
Nanopore
24 (88.9%)
16 (59.3%)
14 (51.9%)
1 (3.7%)
0
10
20
30


Reduced representation of genomes (RADseq, GBS) 133 53.0%
Whole genome(s) 128 51.0%
Transcriptome(s) 105 41.8%
Sequence Capture (multiple targeted loci) 65 25.9%
Amplicons (candidate loci) 58 23.1%
Exome(s) 14 5.6%
Other 13 5.2%
Non-coding RNAs 13 5.2%
Methylome(s) 10 4.0%
Proteome(s) 2 0.8%
Metabolome(s) 0.4%
0
50
100
150
nber of responses
Reduced repres… 8 (17%)
Whole genome(s) 28 (59.6%)
Transcriptomes(s) 32 (68.1%)
Sequence captu… 7 (14.9%)
Amplicons (can… 17 (36.2%)
Methylome(s) 4 (8.5%)
Proteome(s) 9 (19.1%)
Metabolome(s) 7 (14.9%)
Non-coding RNAs 6 (12.8%)
Single cell sequ… 6 (12.8%)
0
10
20
30
40

Does your study organism have a reference genome?

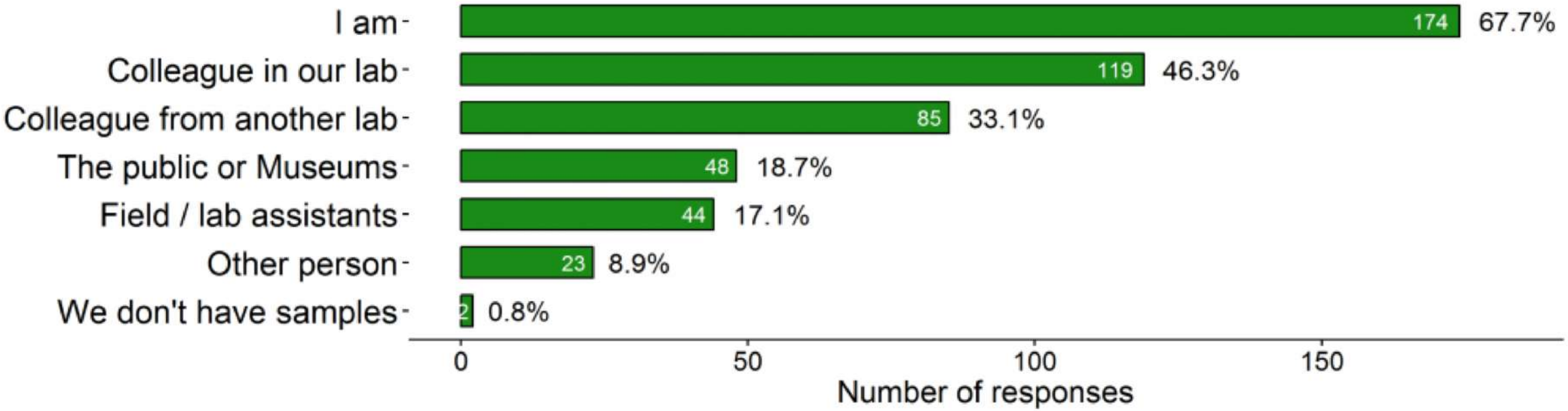


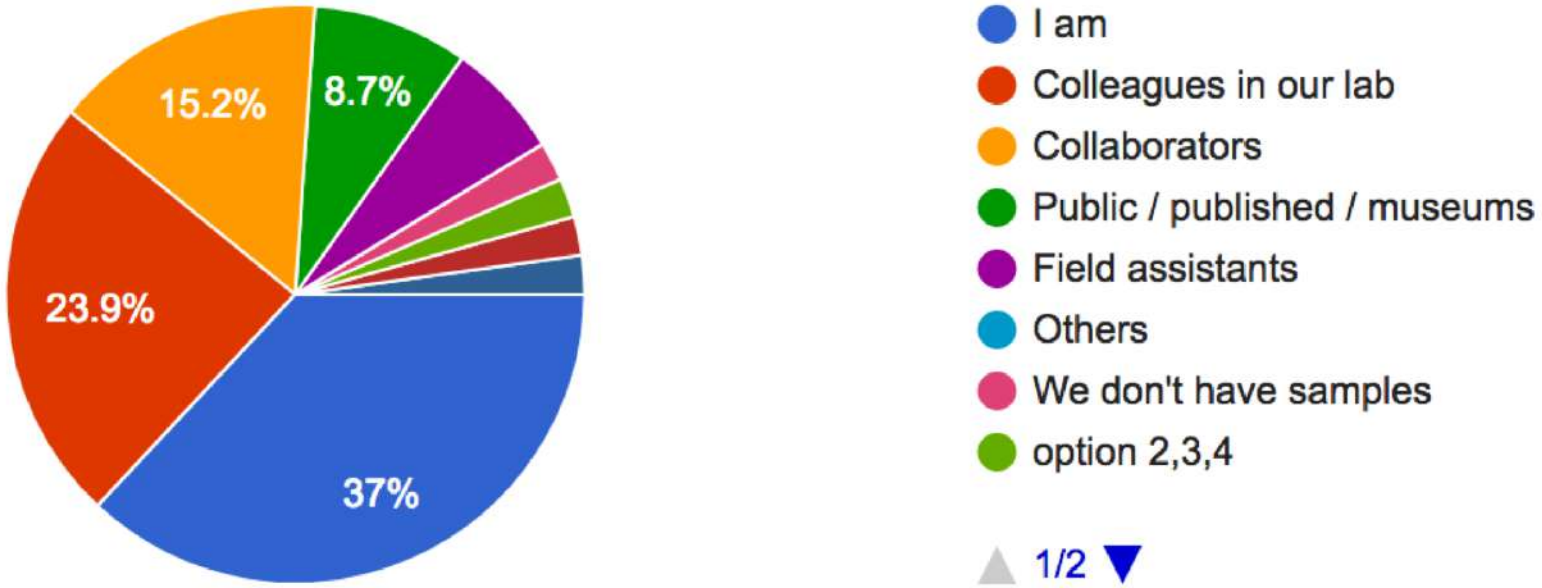
- Yes
- Partially / Some of them
- Not yet, but in the near future
- No
- Don't know



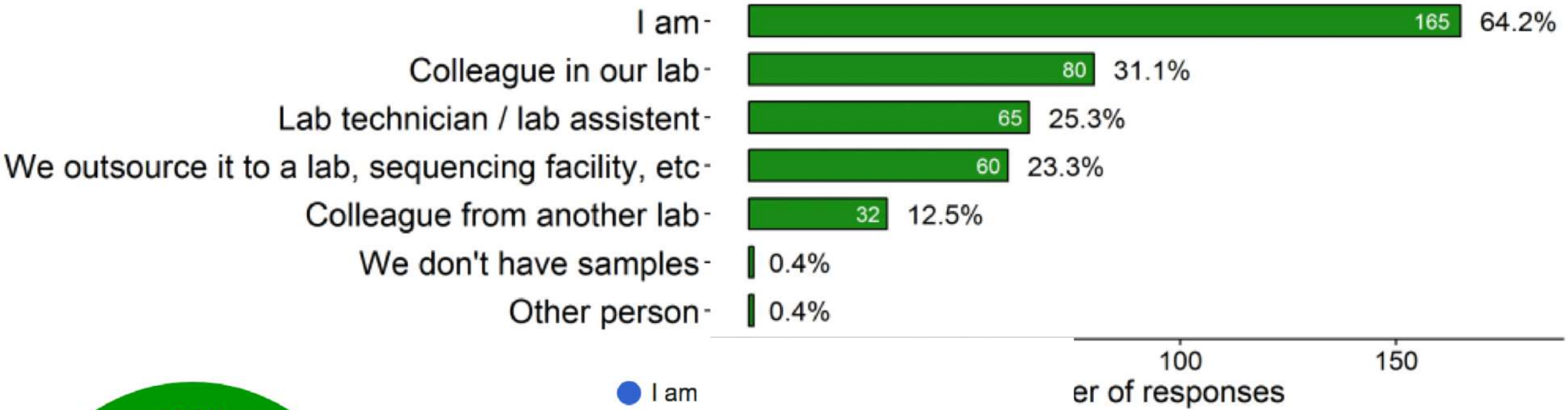
- Yes
- Partially / Some of them
- Not yet, but in the near future
- Yes, but people wouldn't release it
- No
- Don't know

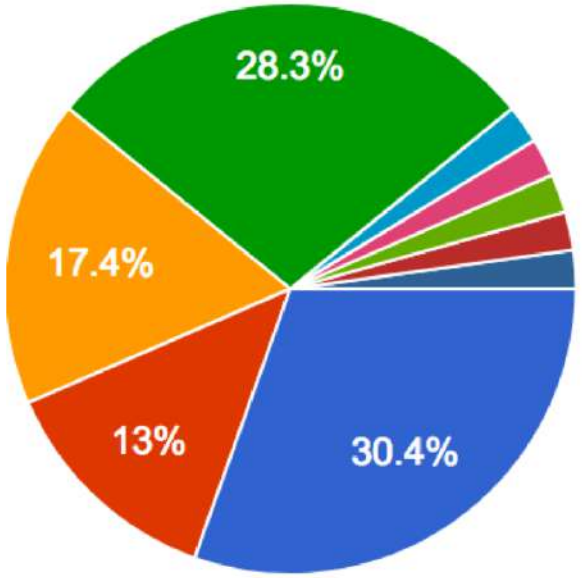
Who is collecting the samples in your main project? (field work or lab work)





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

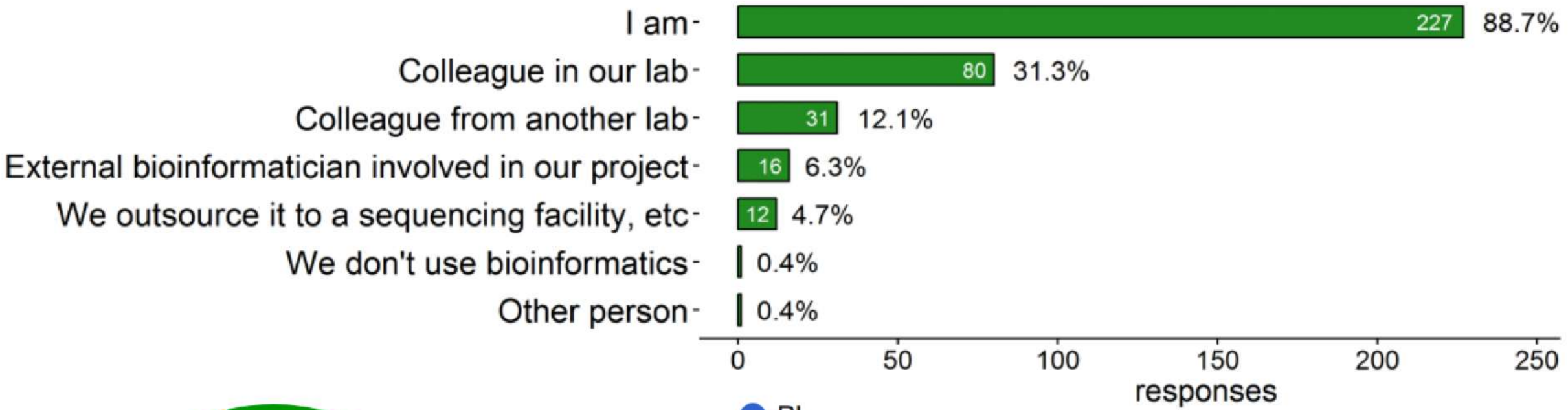


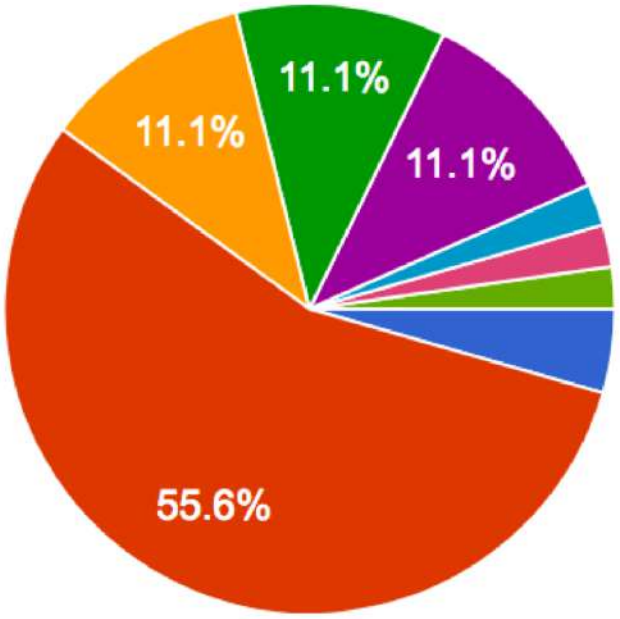


I am
Colleagues in our lab
Collaborators
Outsource to a company / core facility
Others
We don't have samples
except option #1 and may vary bas…
Lab technician

1/2

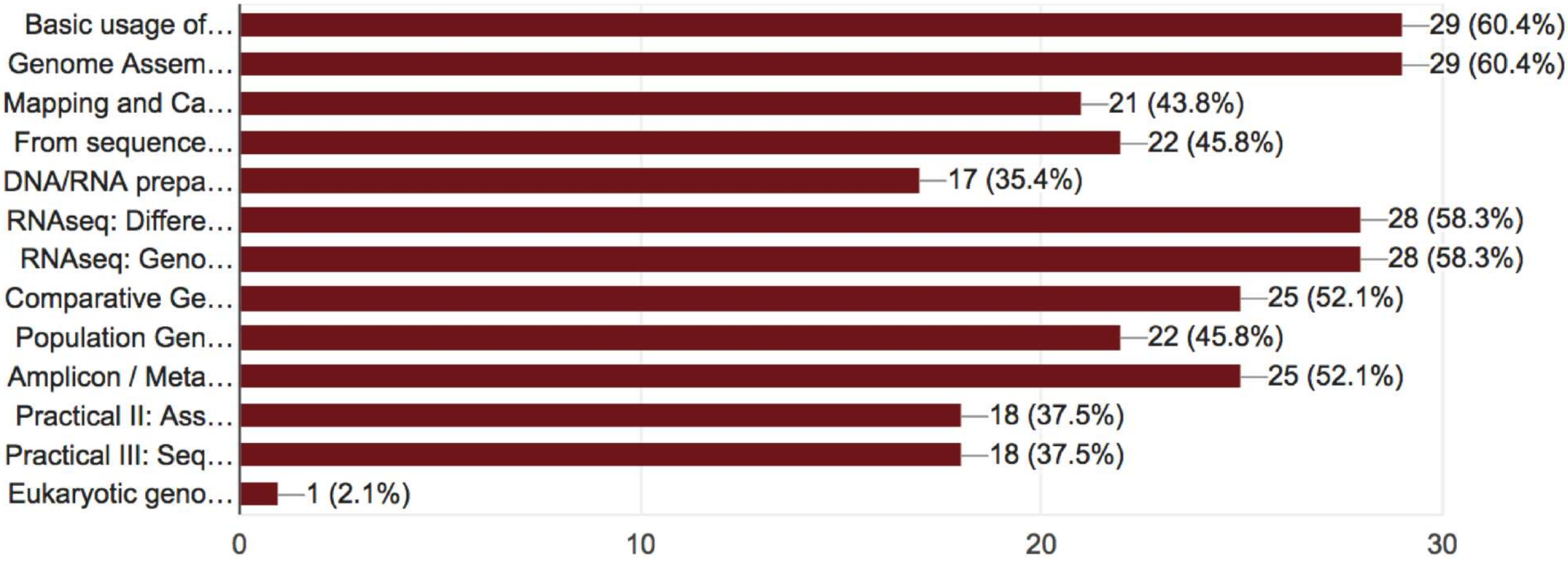
Who is doing the bioinformatic analyses in your main project?



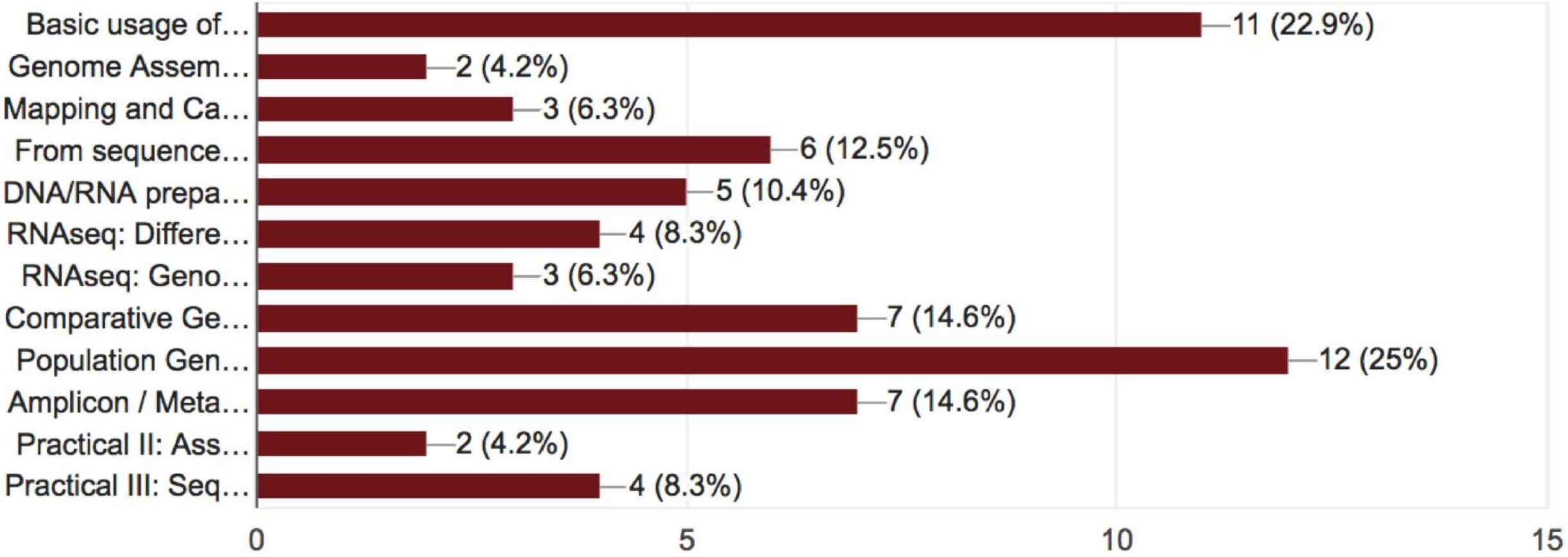


PI
I am
Colleagues in our lab
Collaborators
We outsource to a company
We don't use bioinformatics
options 2,3,4
I am and Collaborators

Most interested lectures



Most uninterested lectures



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

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

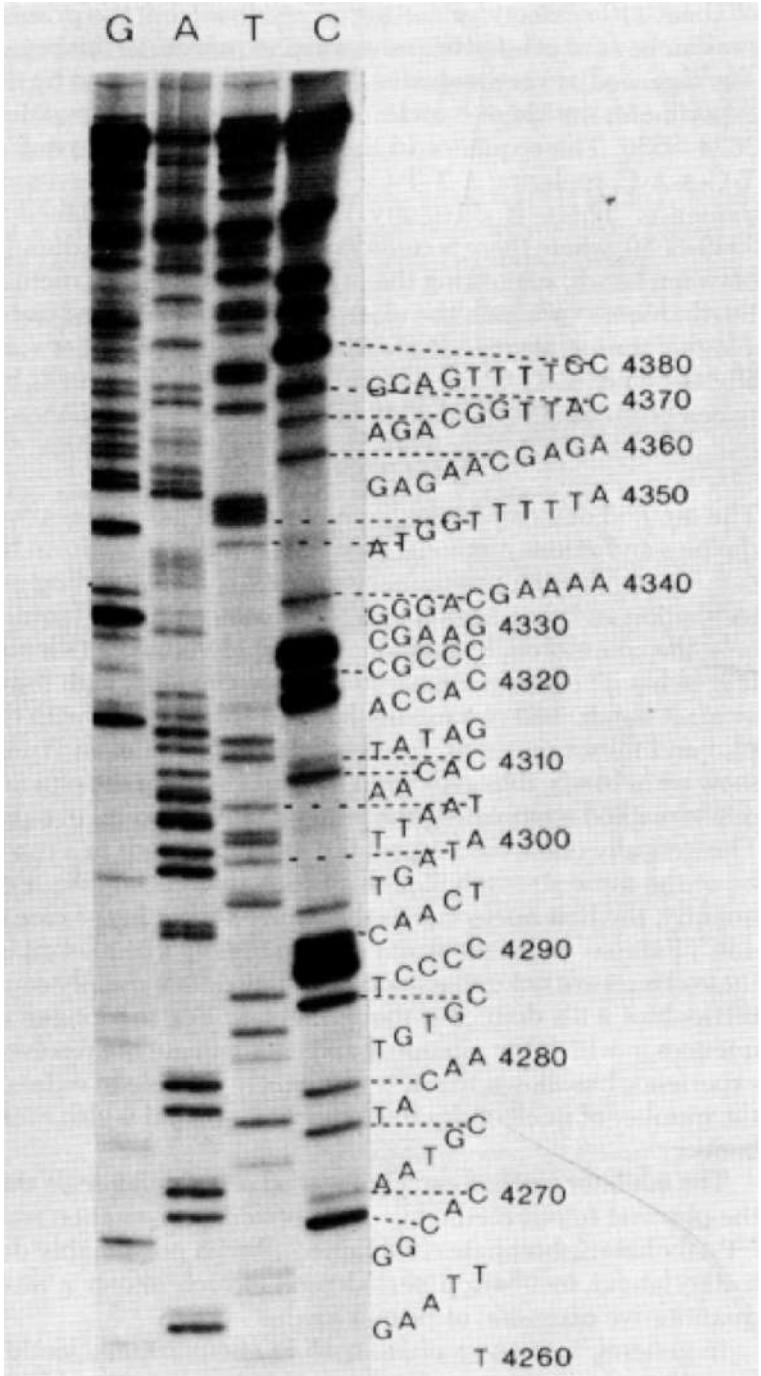
DNA sequencing with chain-terminating inhibitors

(DNA polymerase/nucleotide sequences/bacteriophage ϕX174)

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

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

Contributed by F. Sanger, October 3, 1977

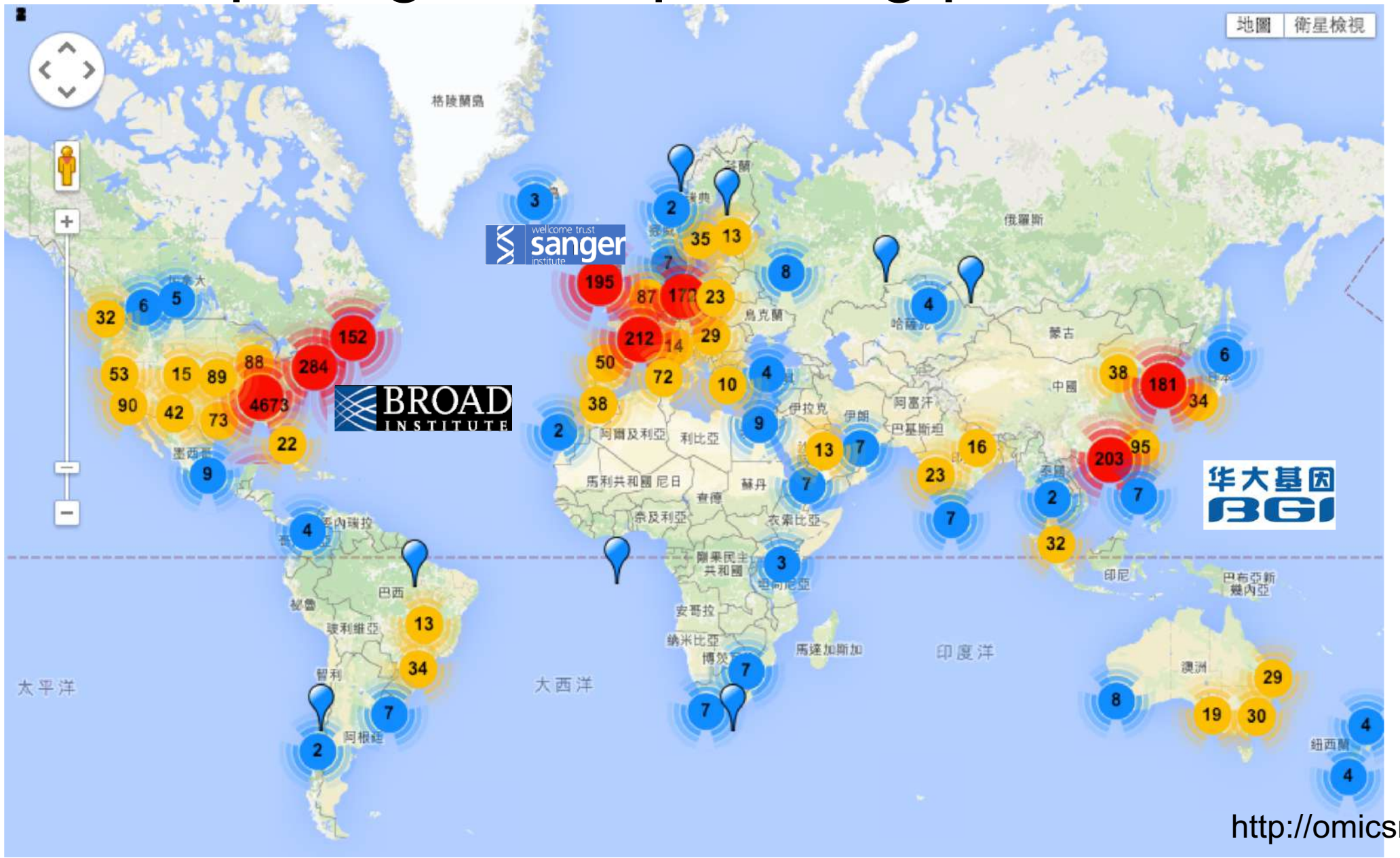


http://www.ncbi.nlm.nih.gov/pmc/articles/PMC431765/

ABI 3730xi at TIGR

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

World competing for sequencing power

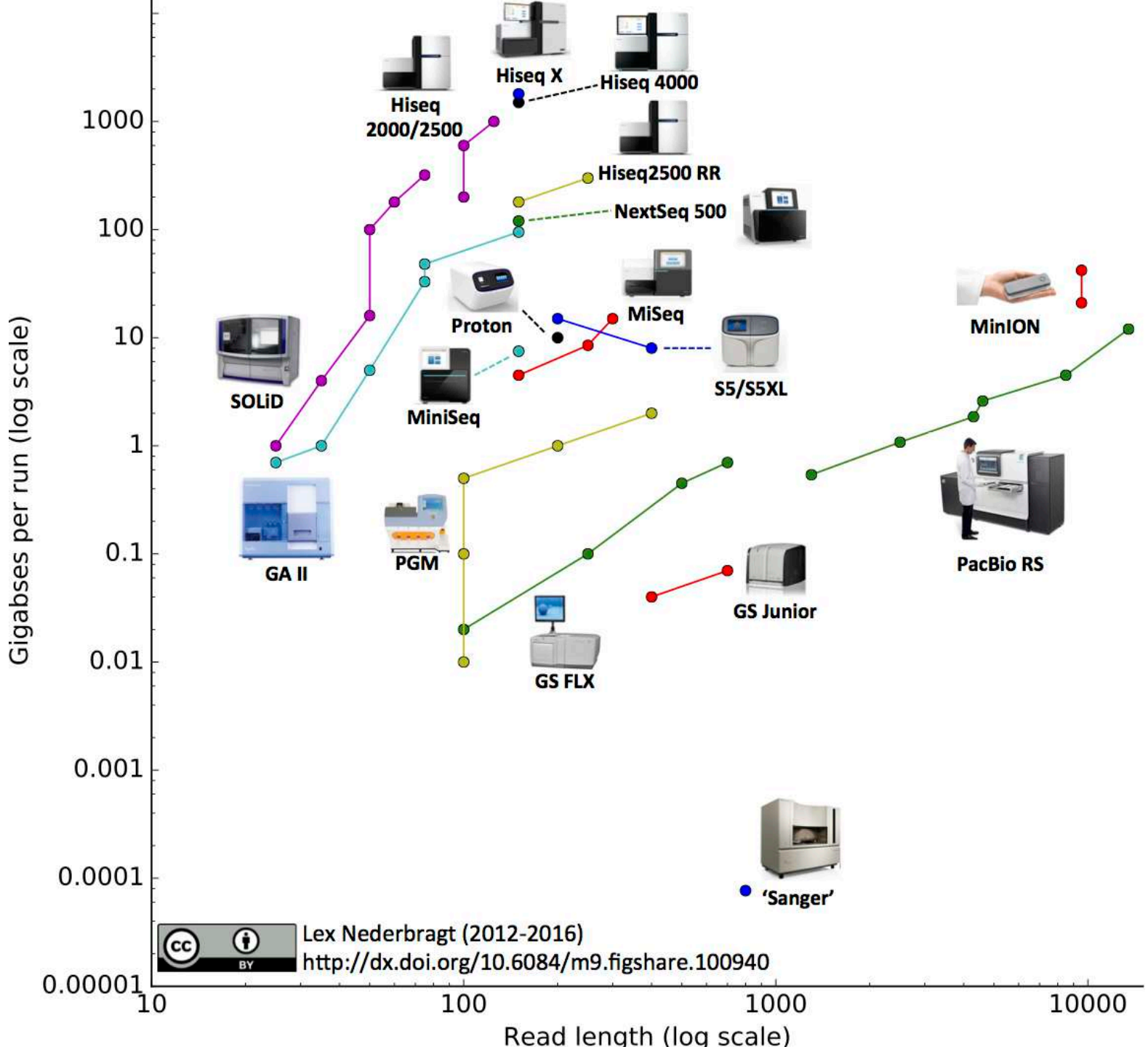


http://omicsmaps.com/

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


Hiseq X
Hiseq 4000
Hiseq 2000/2500
Hiseq2500 RR
NextSeq 500
Proton
MiSeq
MinION
SOLiD
S5/S5XL
MiniSeq
GA II
PGM
PacBio RS
GS Junior
GS FLX
'Sanger'
Gigabses per run (log scale)
Read length (log scale)
1000
100
10
1
0.1
0.01
0.001
0.0001
0.00001
10
100
1000
10000
Lex Nederbragt (2012-2016)
http://dx.doi.org/10.6084/m9.figshare.100940

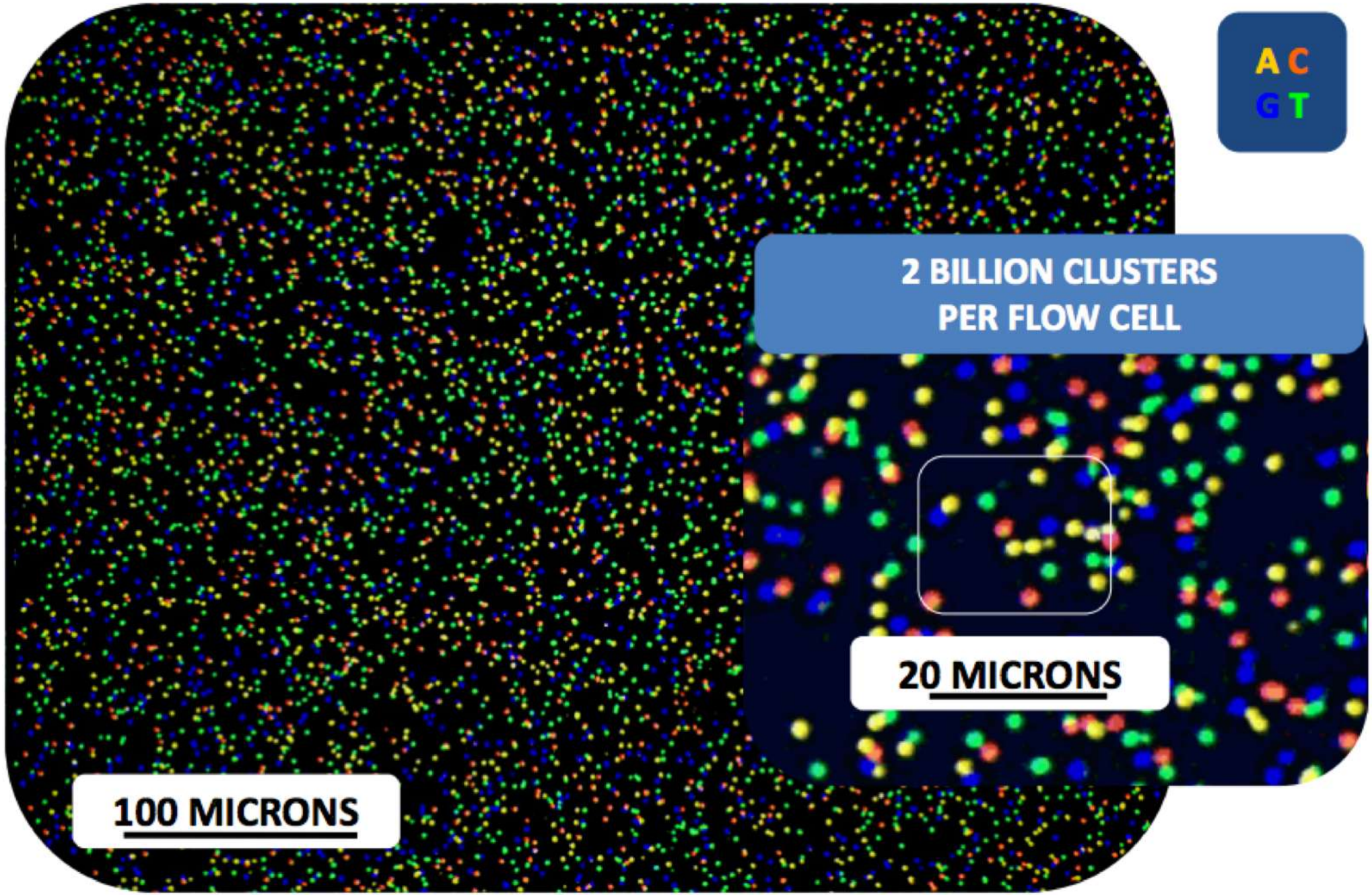
Platform	Reads x run: (M)	Read length: (paired-end*, Half of data in reads**)	Run time: (d)	Yield: (Gb)	Rate: (Gb/d)	Reagents: ($K)	per-Gb: ($)	hg-30x: ($)	Machine: ($)
iSeq 100 1fcell	4	150*	0.77	1.2	1.56	0.625	521	62500	19.9K
MiniSeq 1fcell	25	150*	1	7.5	7.5	1.75	233	28000	49.5K
MiSeq 1fcell	25	300*	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
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	3000/6000-20000	--	--	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

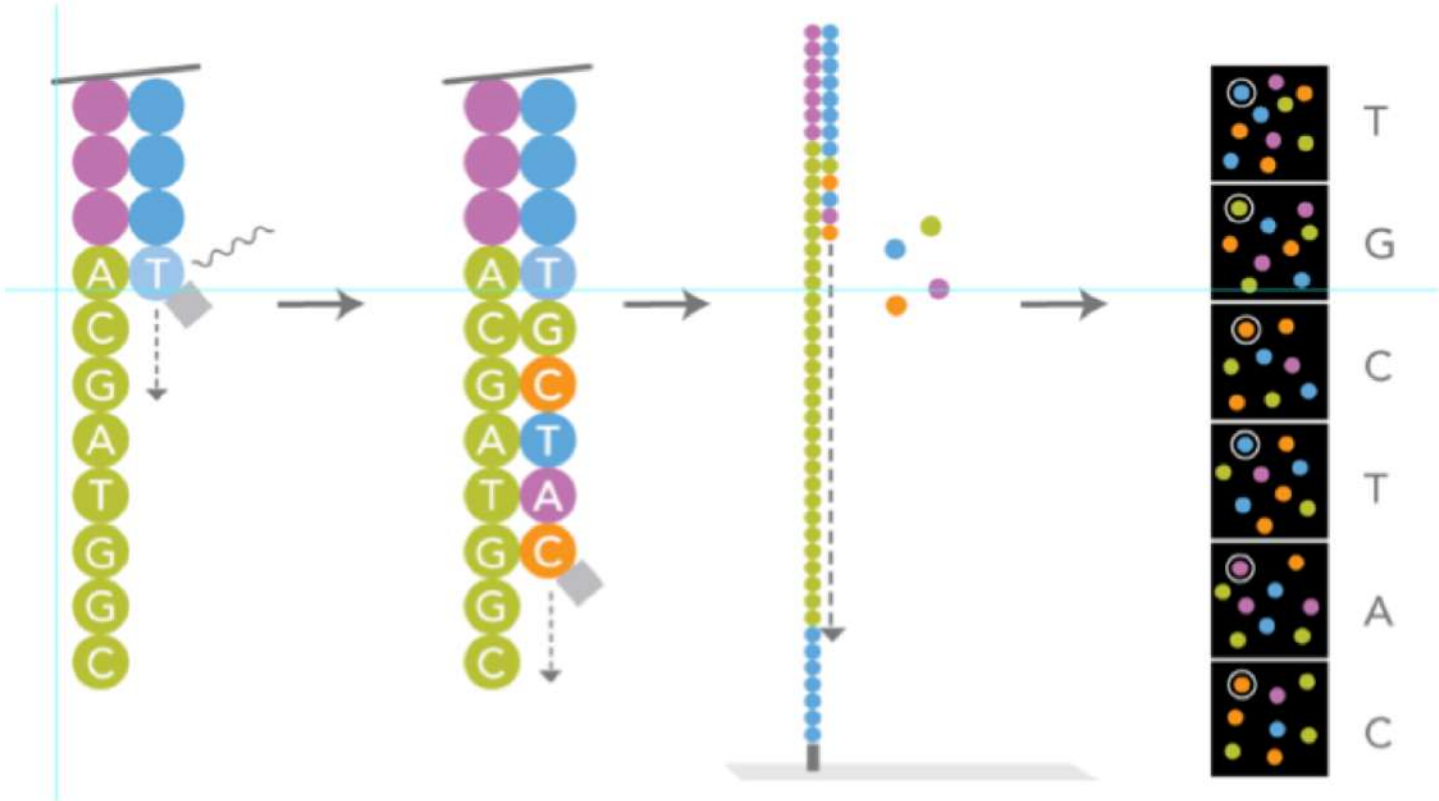
http://twitter.com/albertvilella ; Last updated: 2018.02.20

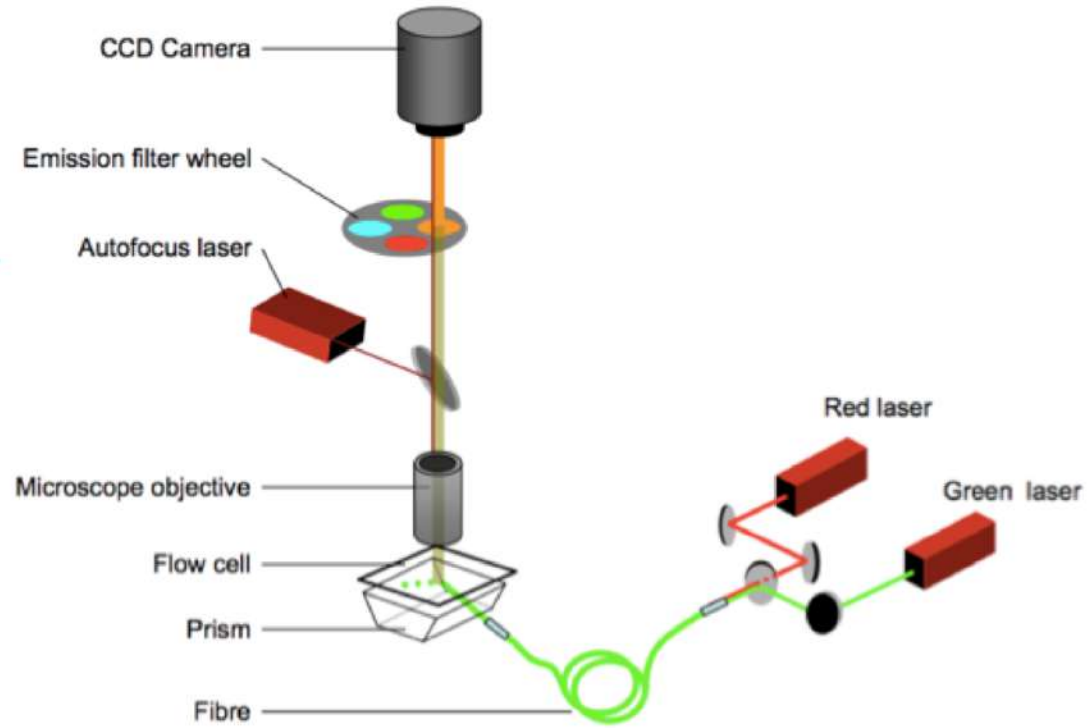
Illumina HiSeq

Sequencing by synthesis

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


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





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

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)

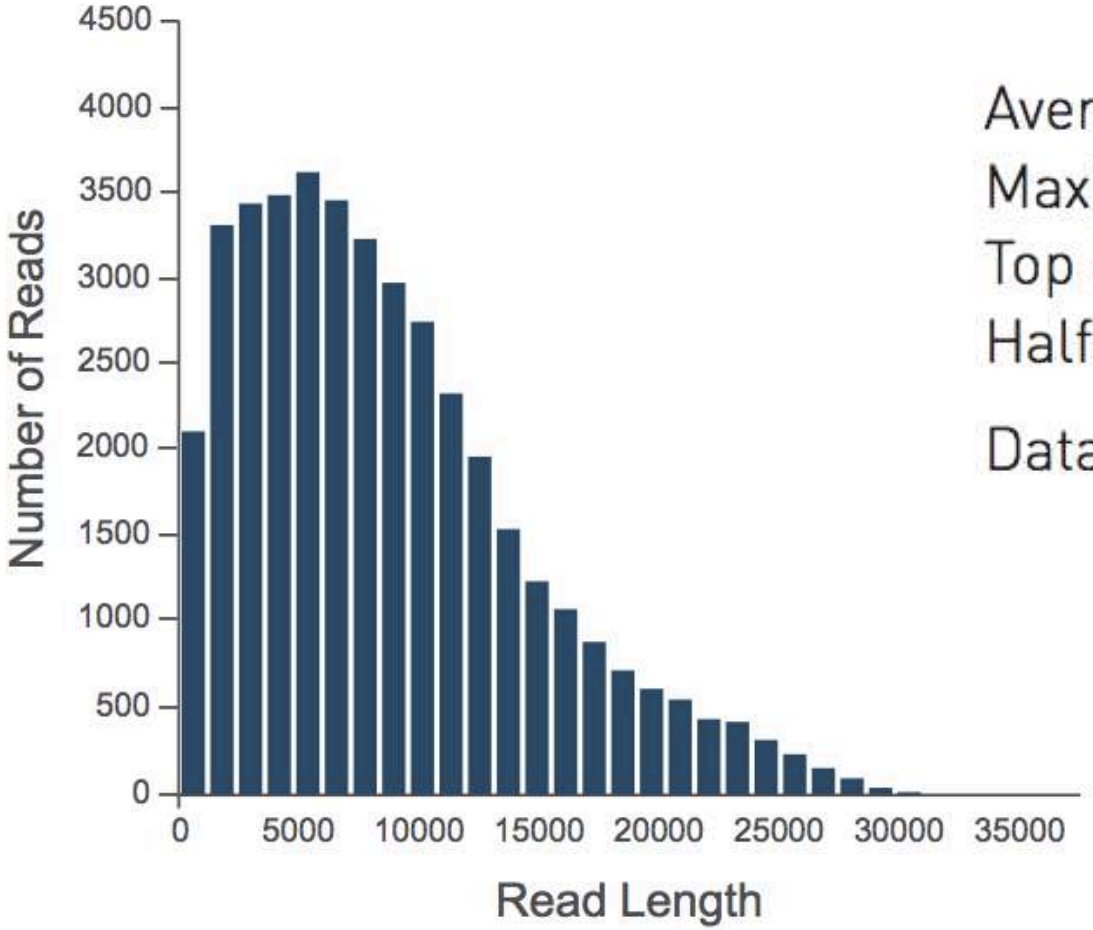
RSII

Sequel

Single molecule sequencing

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

PacBio (Pacific Biosciences)

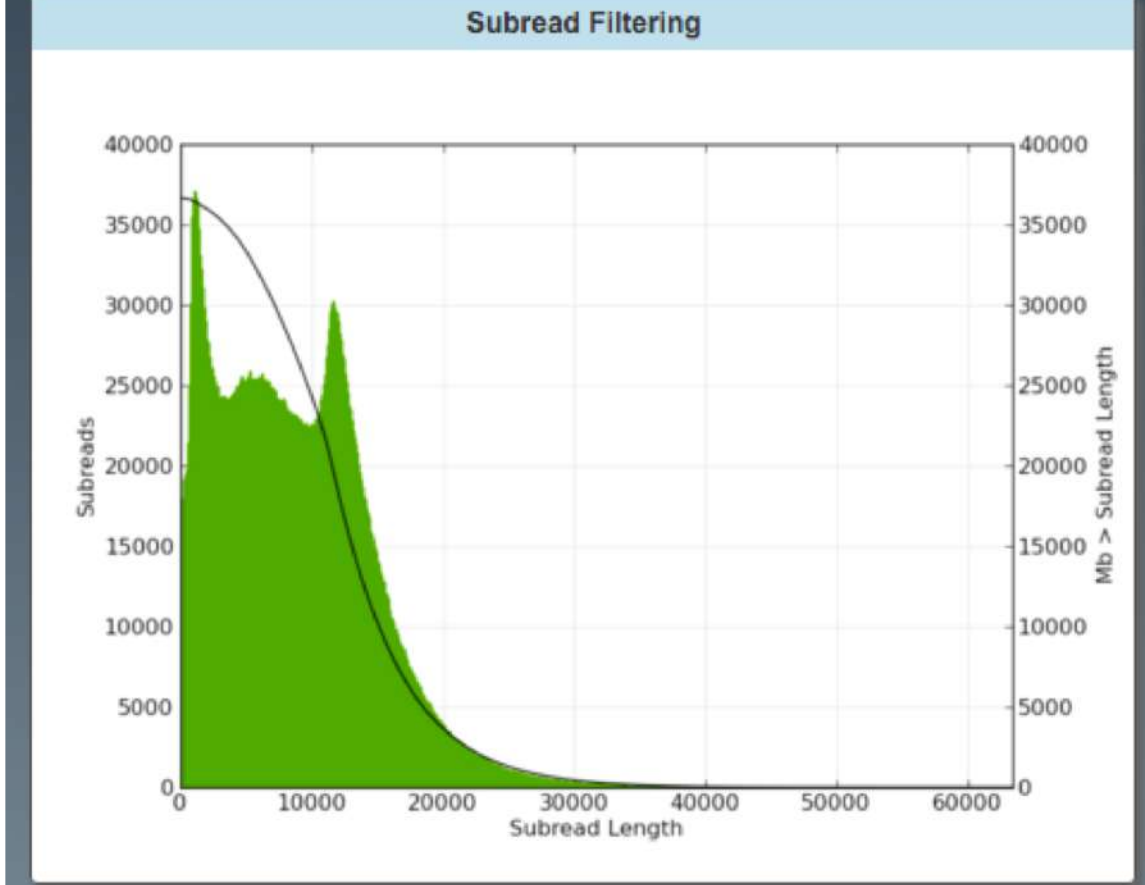


Average:	~ 8.5 kb
Maximum:	> 30 kb
Top 5% of reads:	> 18 kb
Half of data in reads:	> 10 kb
Data per SMRT® Cell:	~ 375 Mb

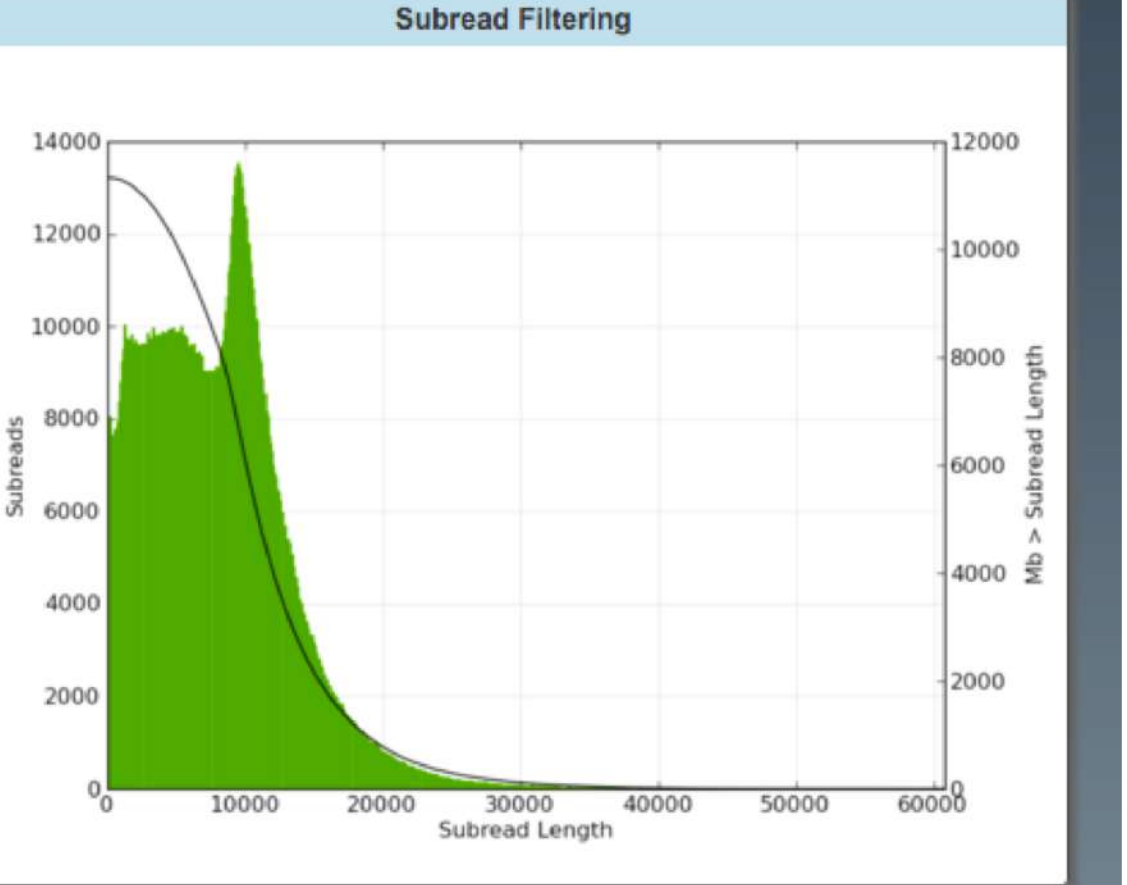
http://files.pacb.com/pdf/PacBio_RS_II_Brochure.pdf

PacBio (Pacific Biosciences)

Mean Subread length	8,722	N50	12,108
Total Number of Bases	36,850,278,578	Number of Reads	4,224,964



Mean Subread length	7,946	N50	10,452
Total Number of Bases	11,398,681,534	Number of Reads	1,434,354



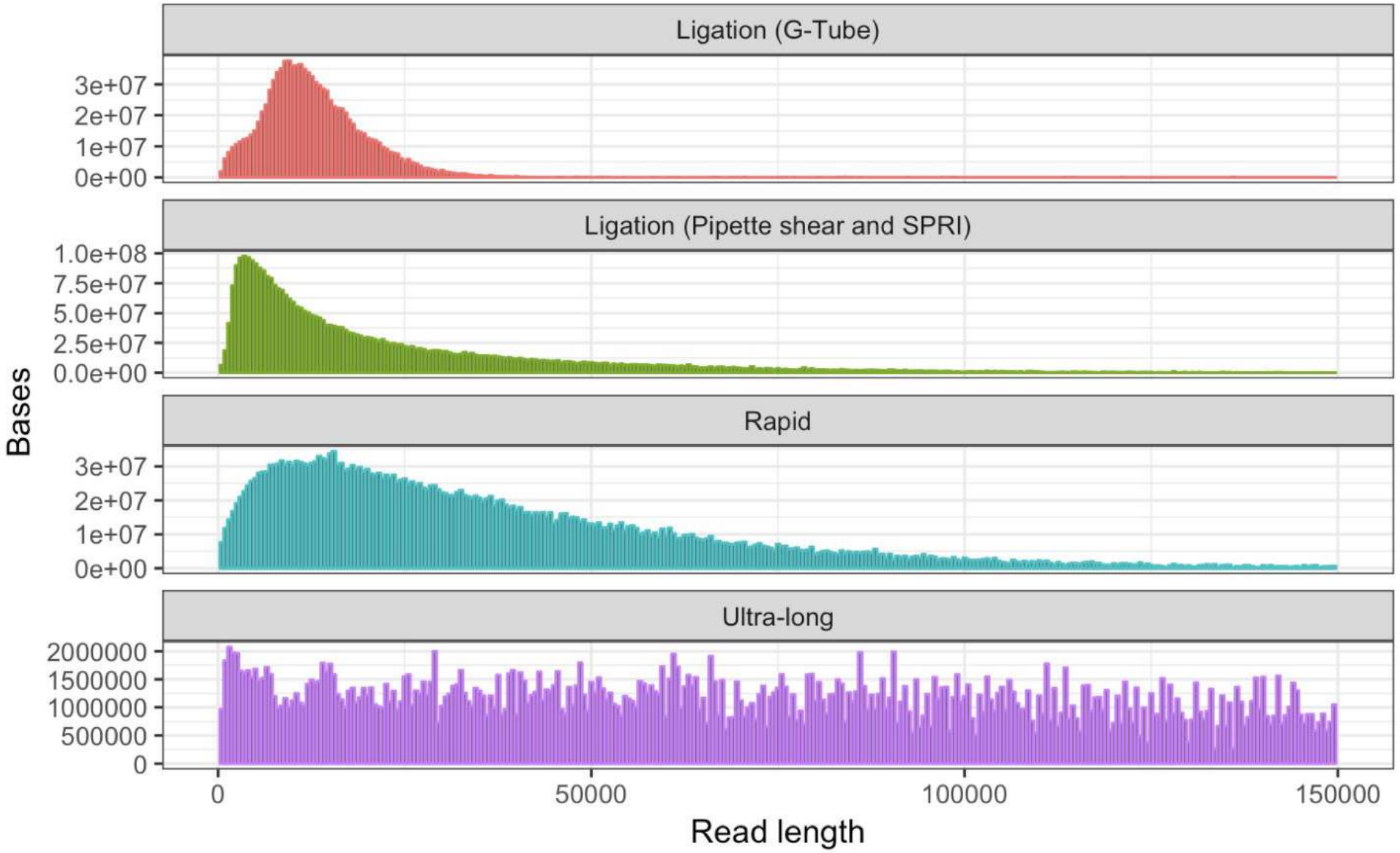
Oxford Nanopore

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

Oxford Nanopore – how it works

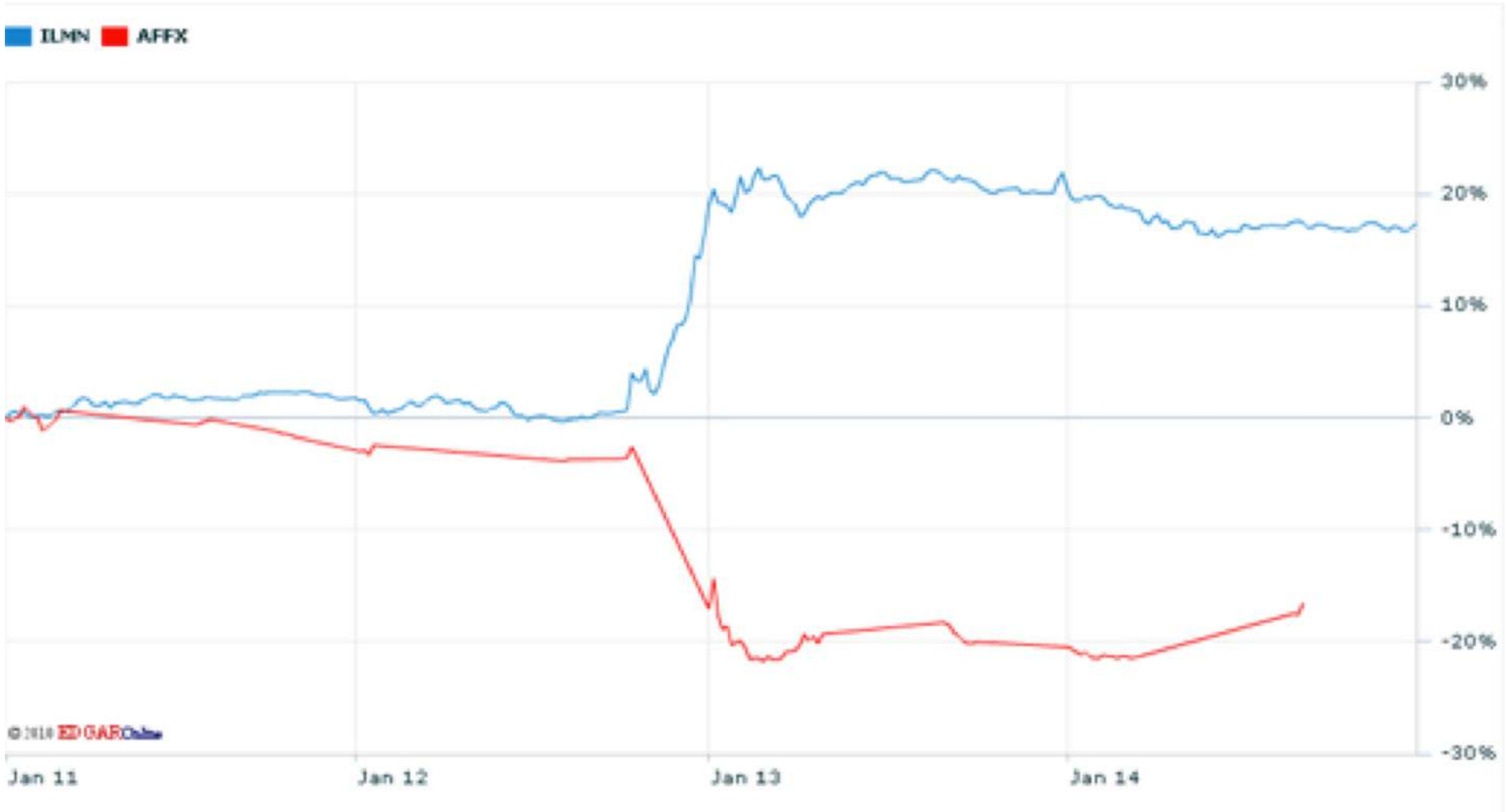
https://nanoporetech.com/how-it-works

Read length go beyond



https://www.slideshare.net/scalene/ultralong-read-methods-for-nanopore-single-moleculesequencing-88022069

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

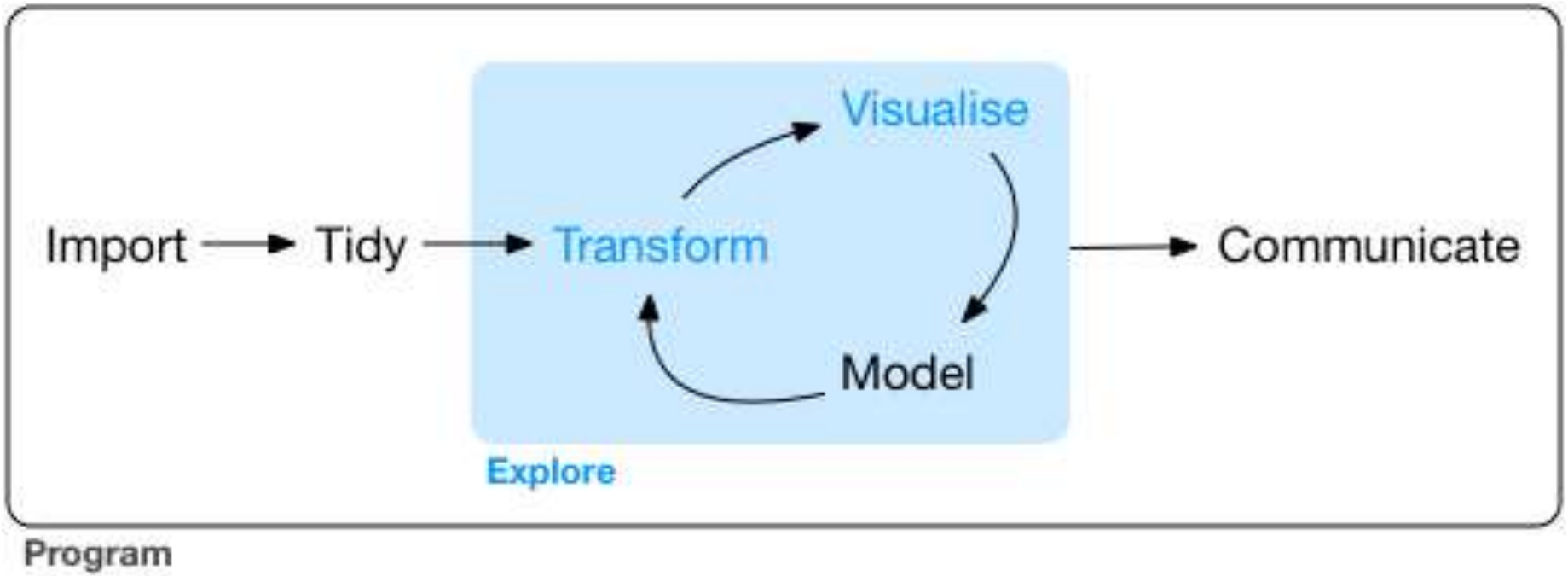
Sample Name	Sample ID	Lane ID	Yield (Mb)	# of Reads
F2-1	SG-IB01	1	11,435	75,729,838
F2-2	SG-IB02		12,014	79,561,504
F2-3	SG-IB03		11,577	76,666,714
F3-2	SG-IB05		11,119	73,638,446
F3-4	SG-IB07		10,399	68,870,380
F3-5	SG-IB08		11,671	77,292,976
F3-1	SG-IB09		12,474	82,610,516
F3-3	SG-IB10		11,916	78,915,536
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



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

More data but less people with informatics skills

- Sequencing is the result of many types of experiment

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

- **Bottleneck is bioinformatics analysis**

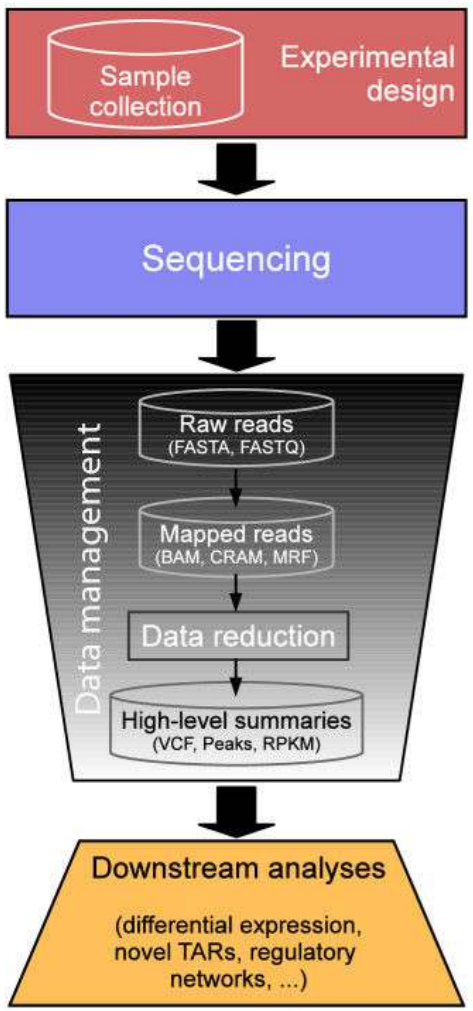
Sboner *et al. Genome Biology* 2011, **12**:125
http://genomebiology.com/2011/12/8/125

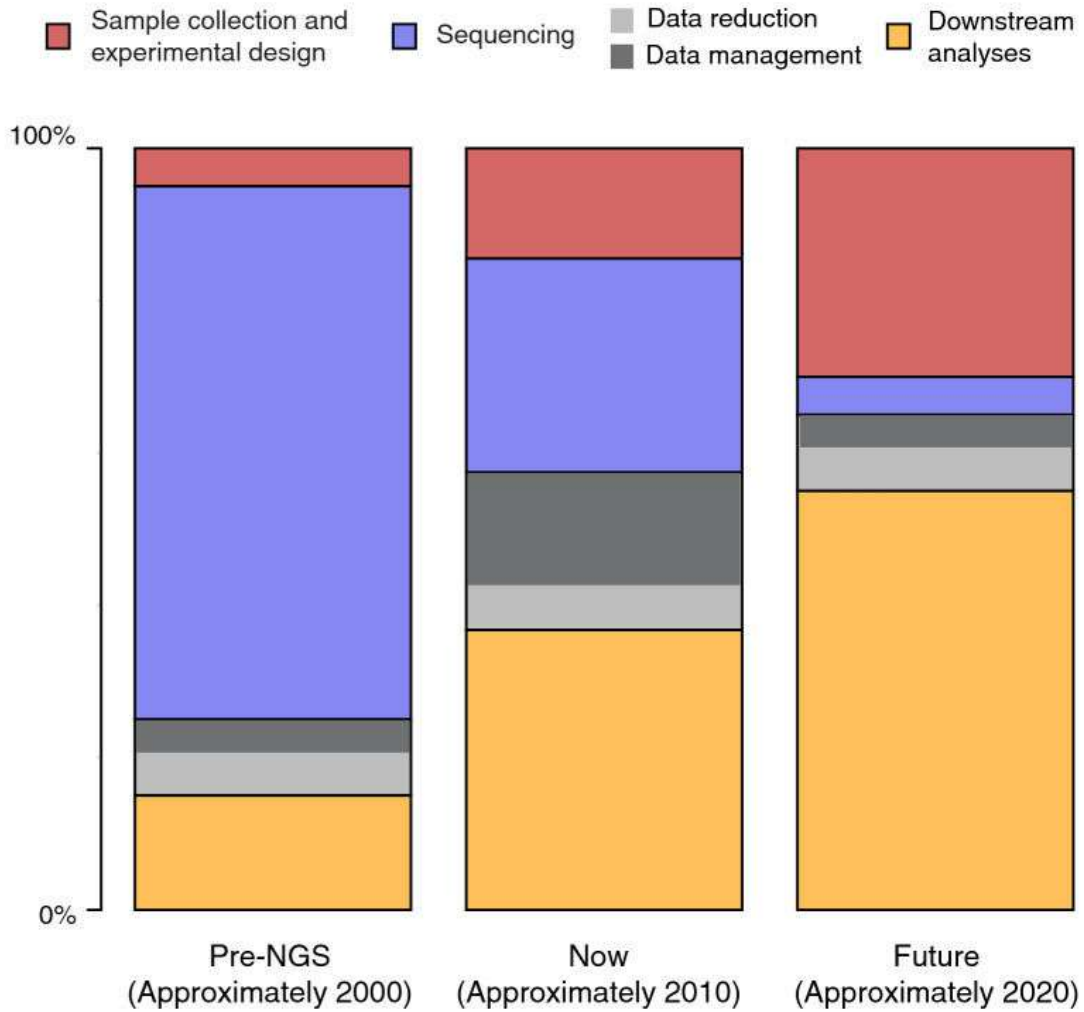
Genome Biology

OPINION

The real cost of sequencing: higher than you think!

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





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

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

Important problems:

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

No ’perfect’ pipeline – learn through experience

Always understand your data / programs

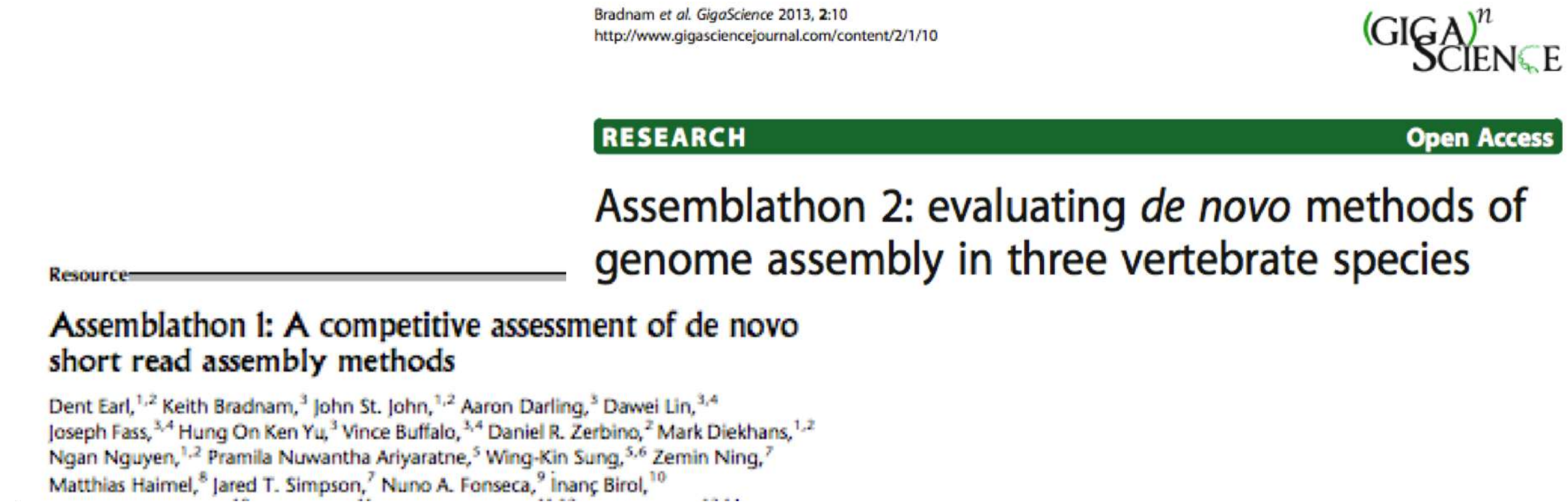
- Understand:
  - Data format
  - The nature of your data

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

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

If unsure – always check **benchmark** studies

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



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

(GIGA)^n SCIENCE

RESEARCH Open Access

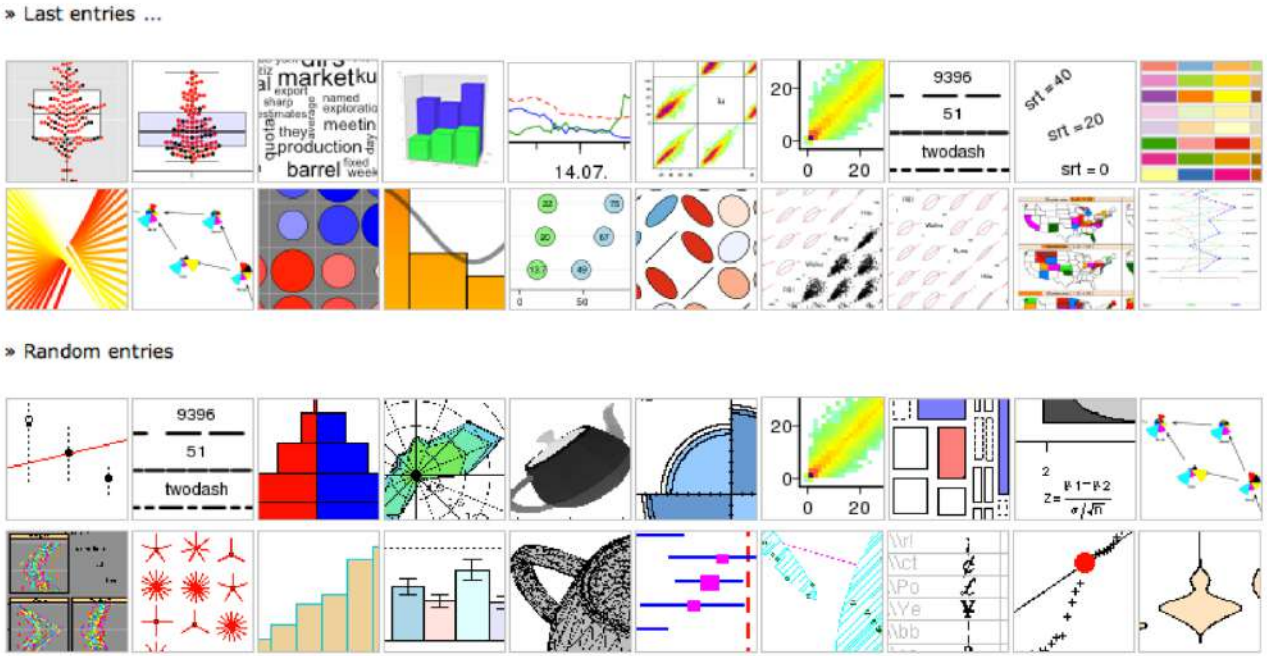
Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species

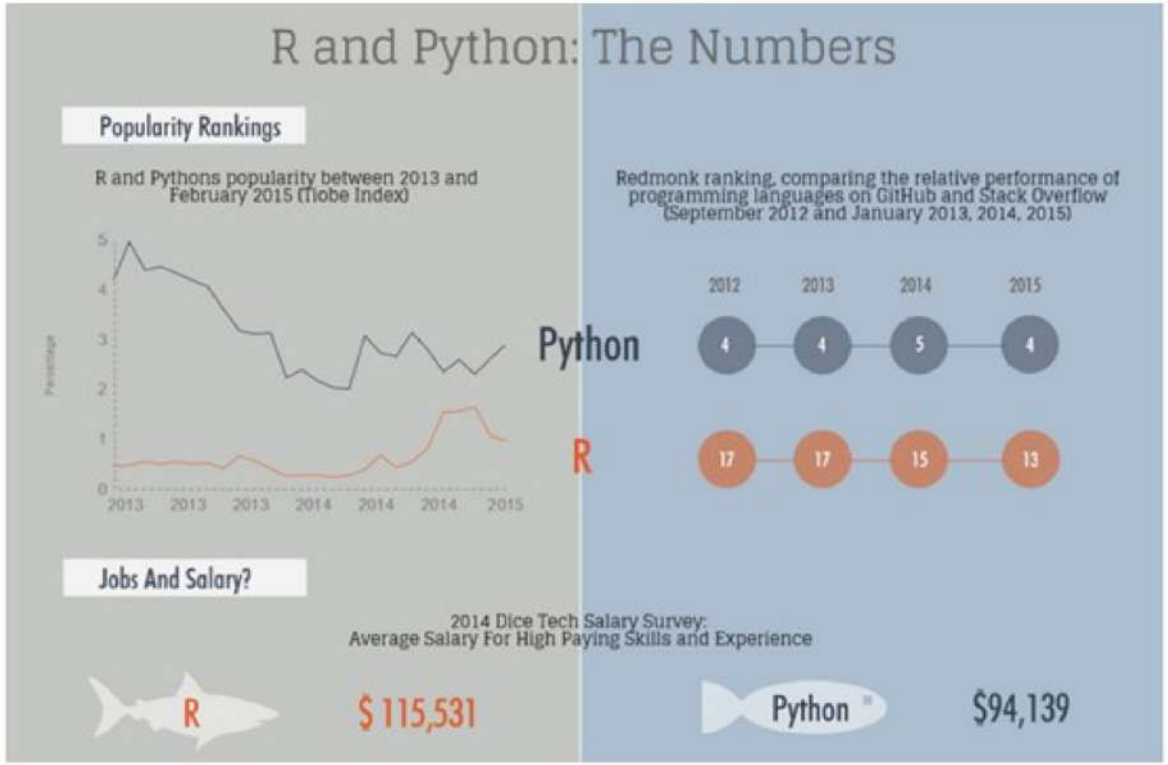
Resource

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

Dent Earl,[1,2] Keith Bradnam,[3] John St. John,[1,2] Aaron Darling,[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


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


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

FASTA format

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

Alignment format

- Some programs need slightly modified format

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

Data type keep evolving

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

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

Fastq files:

FASTQ format is a **text-based format** for storing both a biological **sequence** (usually nucleotide sequence) and its corresponding **quality scores.**

-Wikipedia

```
@SEQUENCE_ID1
ATGCGCGCGCGCGCGCGCGGGTAGCAGATGACGACACAGAGCGAGGATGCGCTGAGAGTA
GTGTGACGACGATGACGGAAAATCAGA
+
BBBBBPPPPPXXXXX^^^^^^^^^^^^^^^^^^^^^^^^__^^^^^^^^^^^^_eeeeeee
[[[[[^^^^]]]]]XXXXXPPPPPBBBBB
```

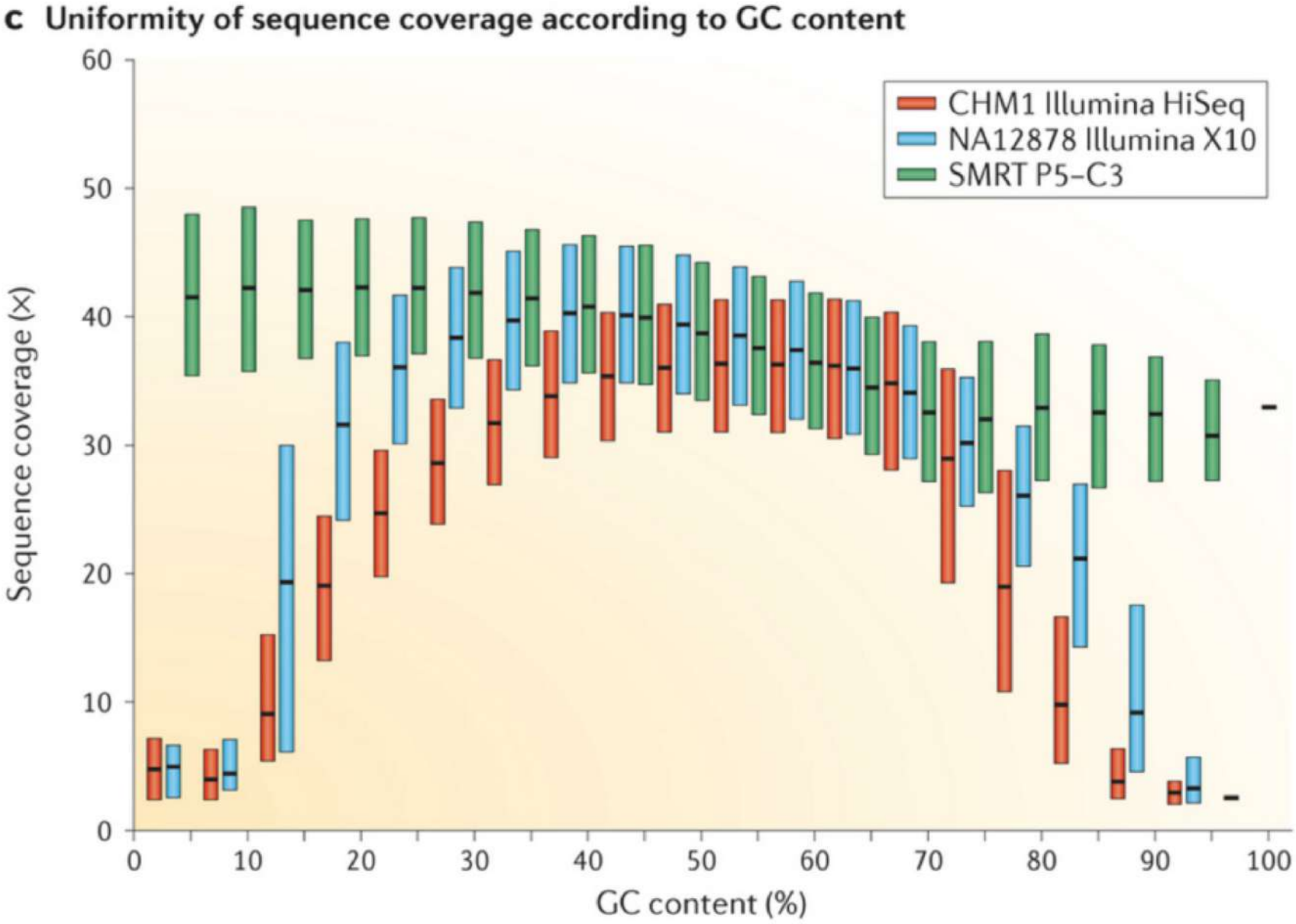
1. **Single line ID** with at symbol (“@”) in the first column.
2. There should be not space between “@” symbol and the first letter of the identifier.
3. Sequences are in multiple lines after the ID line
4. Single line with plus symbol (“+”) in the first column to represent the quality line.
5. Quality ID line can have or have not ID
6. Quality values are in multiple lines after the + line

Analysis and interpretation

Is your data good enough?

https://sequencing.qcfail.com

Sequencing Biases

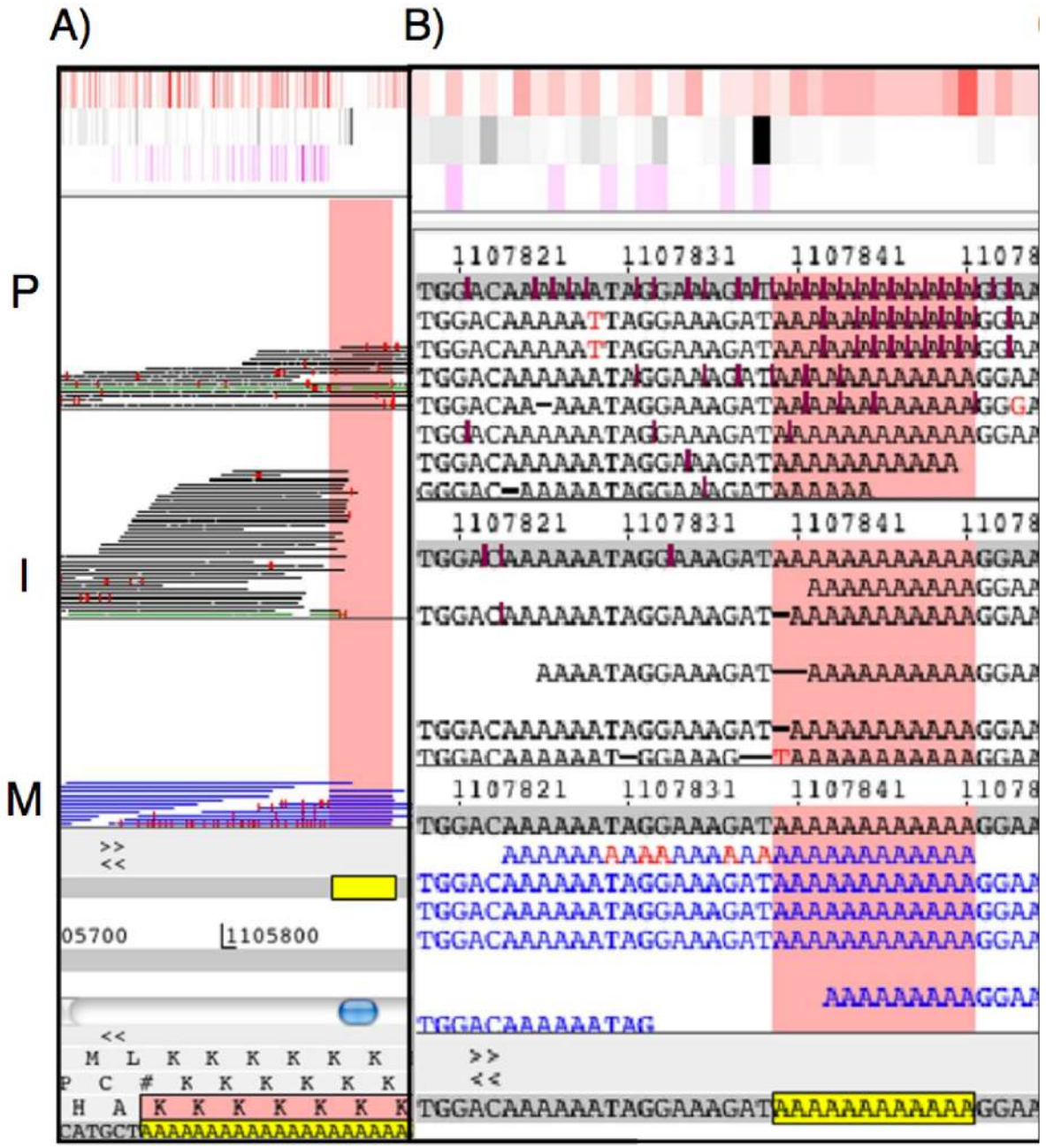


http://www.nature.com/nrg/journal/v16/n11/fig_tab/nrg3933_F2.html

Sequencing Errors

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

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



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

Three situations you are most likely to encounter

Genome reference is available (for example, humans):

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

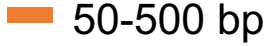
Genome reference is NOT available

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

Counting:

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

More Definition

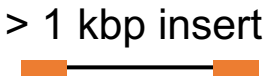


Read

A sequenced piece of DNA

Paired-end read

Sequencing both ends of a short DNA fragment



Mate-pair read

Sequencing both ends of a long DNA fragment

Insert size

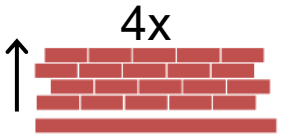
The length of the DNA fragment

Contig

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

Scaffold

Contigs separated by gaps of known length



Coverage

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

What is an alignment?

Align the following two sequences:

```
ATTGAAAGCTA
GAAATGAAAAGG
```

Scoring scheme is needed:

1 for match

-1 for mismatch

-2 for gap

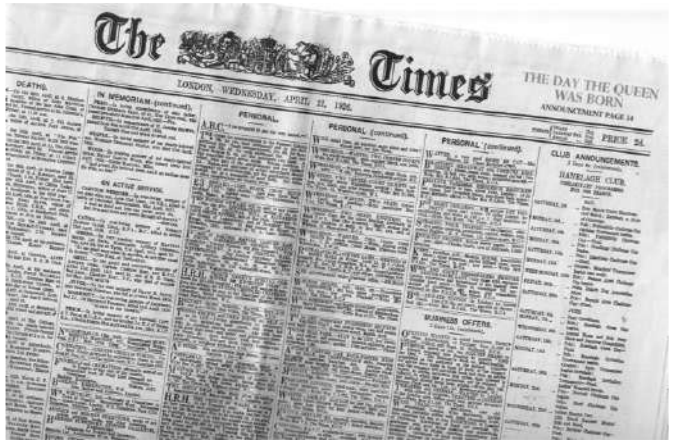
```
1:
--ATTGAAA-GCTA
  | ||||| |
GAAATGAAAAGG--

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

insertions / deletions (indels) mismatches

Which alignment is better?

Assembly (Lecture 4)

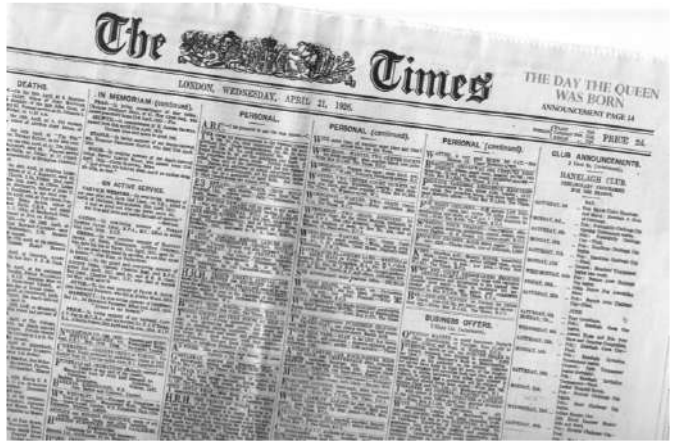


Genome
(3.000.000 letters)

Sequencing →

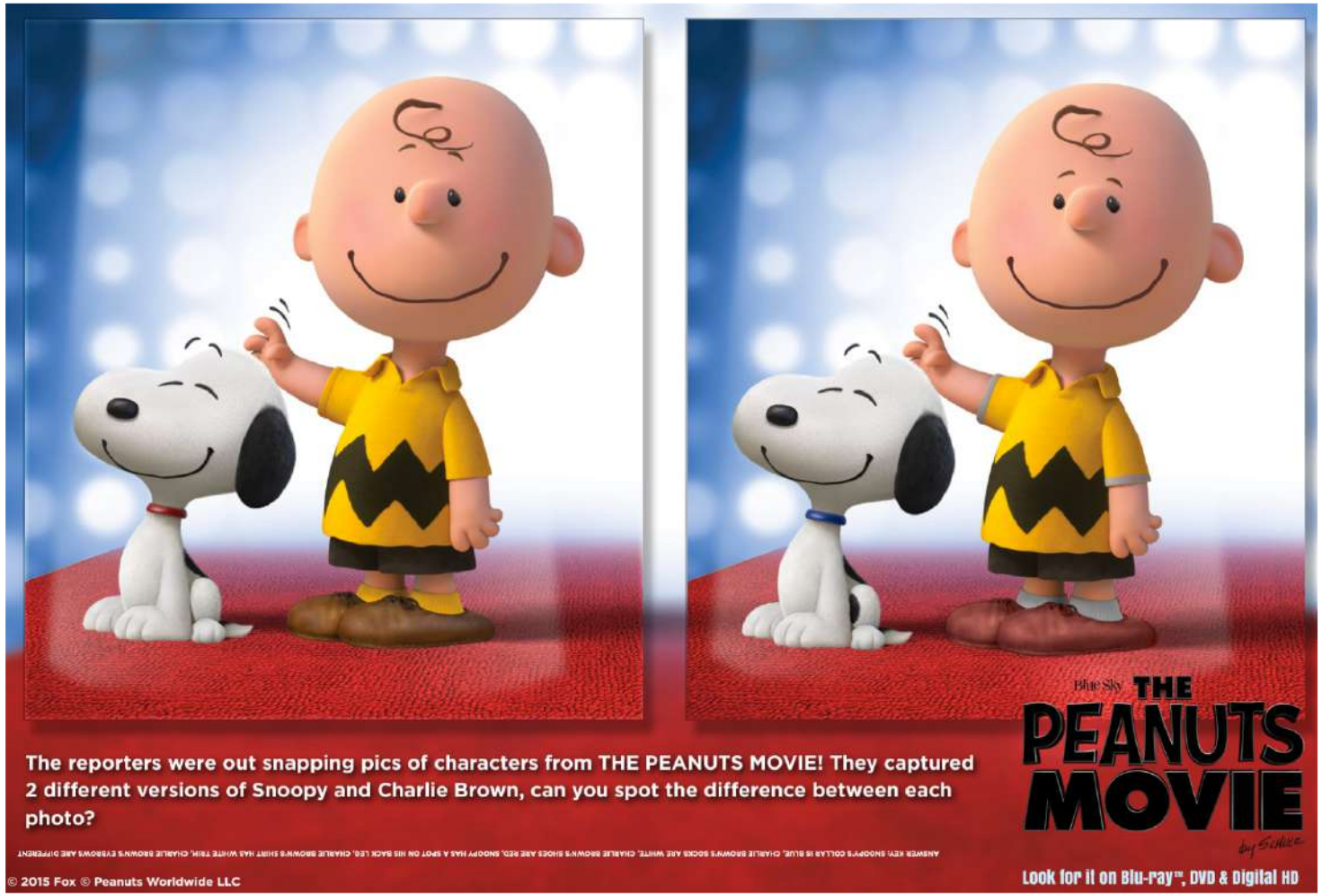
Reads
(50-500 letters each)

Assembly →

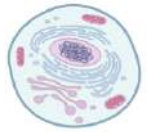


Genome
(3.000.000 letters)

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



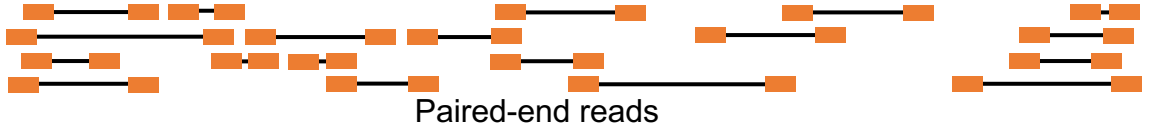
Assembly (Lecture 3)



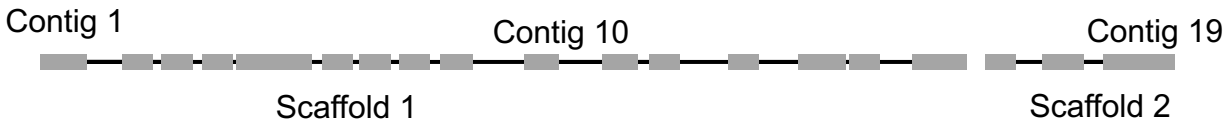
Genome

Fragment

Sequence



Assemble

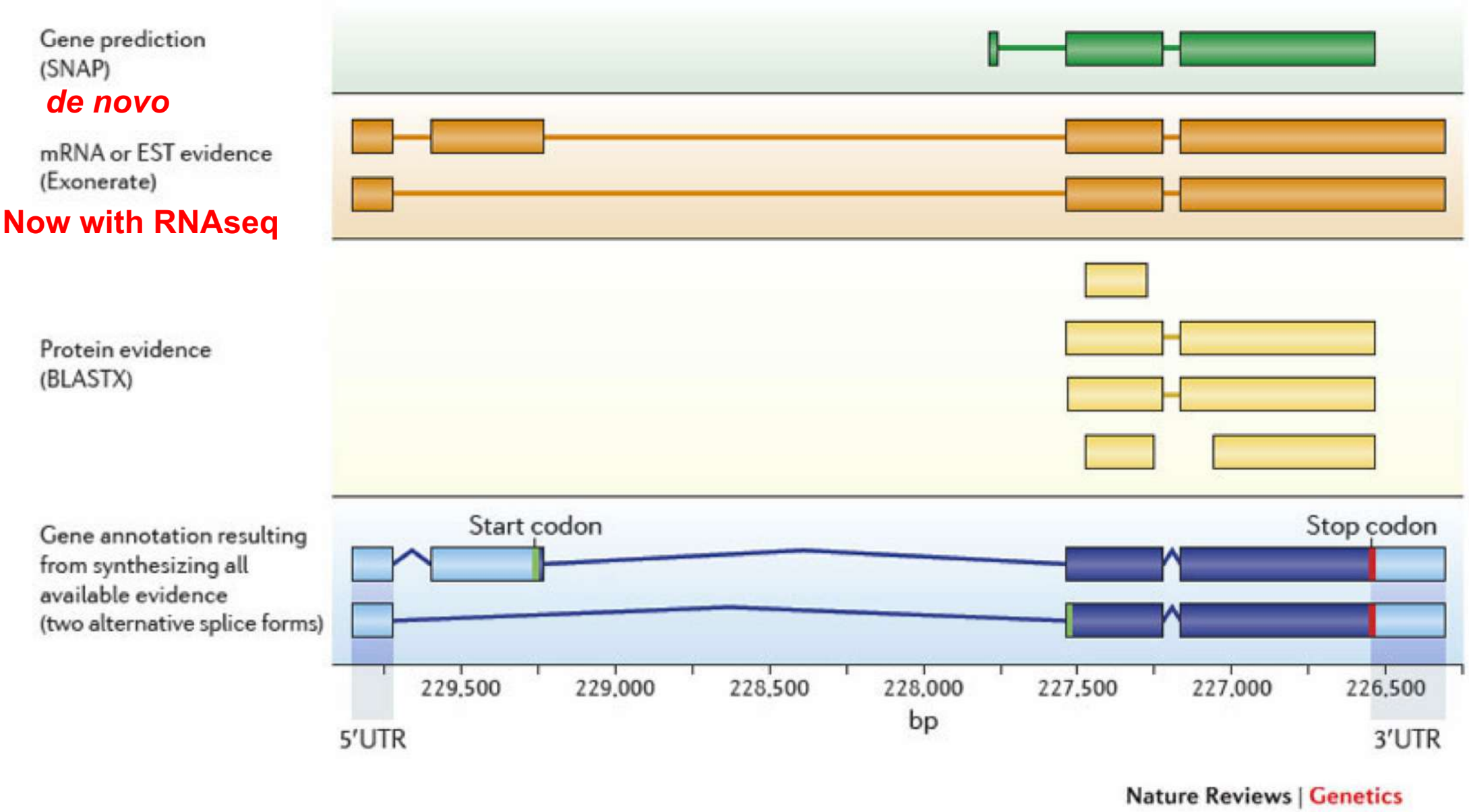


After assembly

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

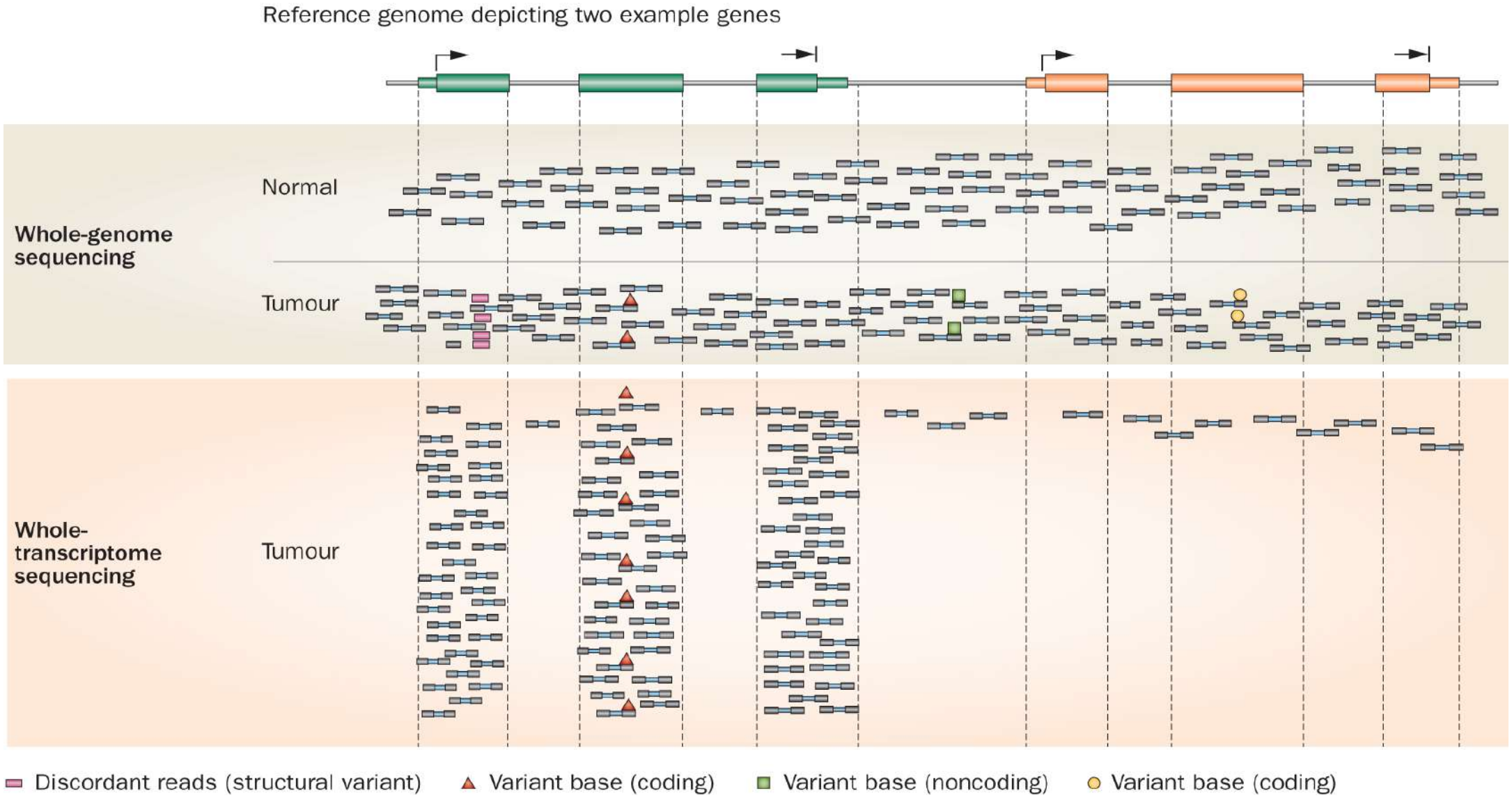
- How accurate is it?
- Have you tried different assemblers?
- Can you improve with additional data or diminishing returns?
- Is there contamination?
- How does it compare to other species?

Annotation



Yandell and Ence Nature Genetics Review (2012)

Mapping (Lecture 4)



doi:10.1038/nrgastro.2012.126

Read length matters in sequencing

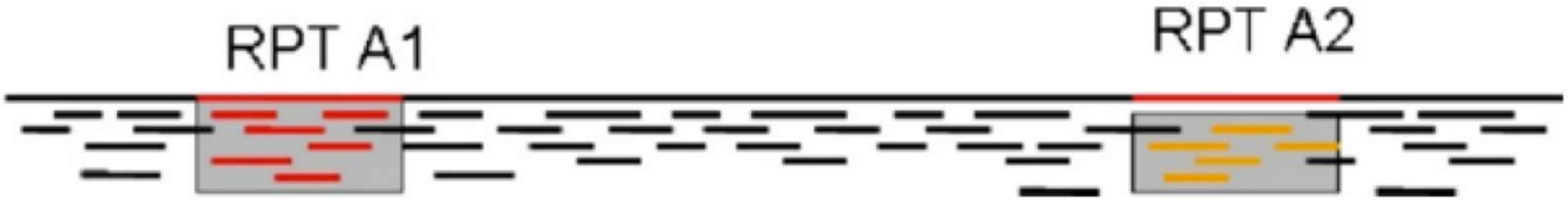


Figure 5. Two copies of a repeat along a genome. The reads colored in red and those colored in yellow appear identical to the assembly program.

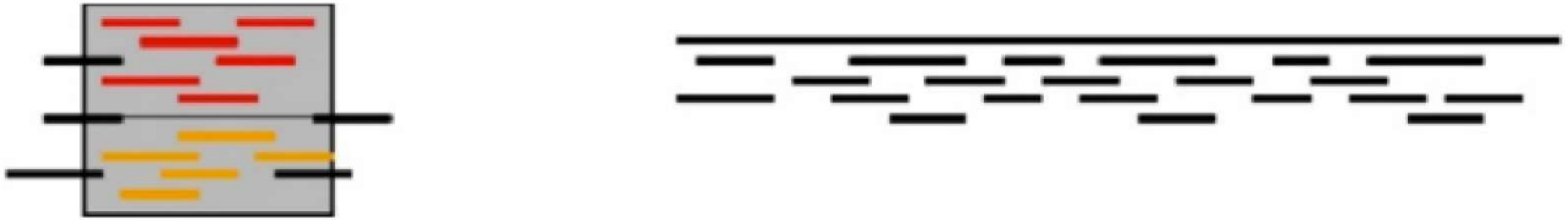
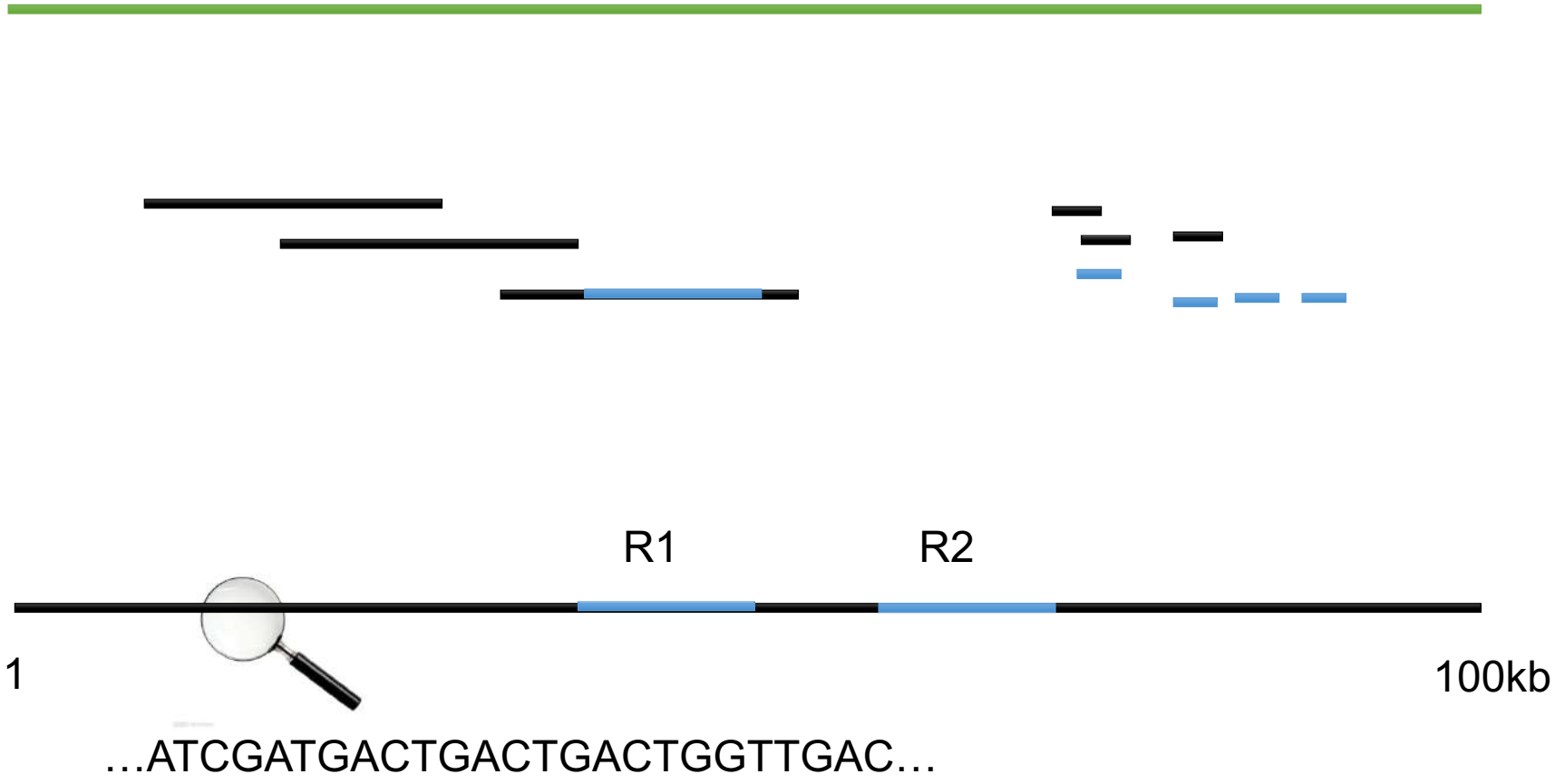


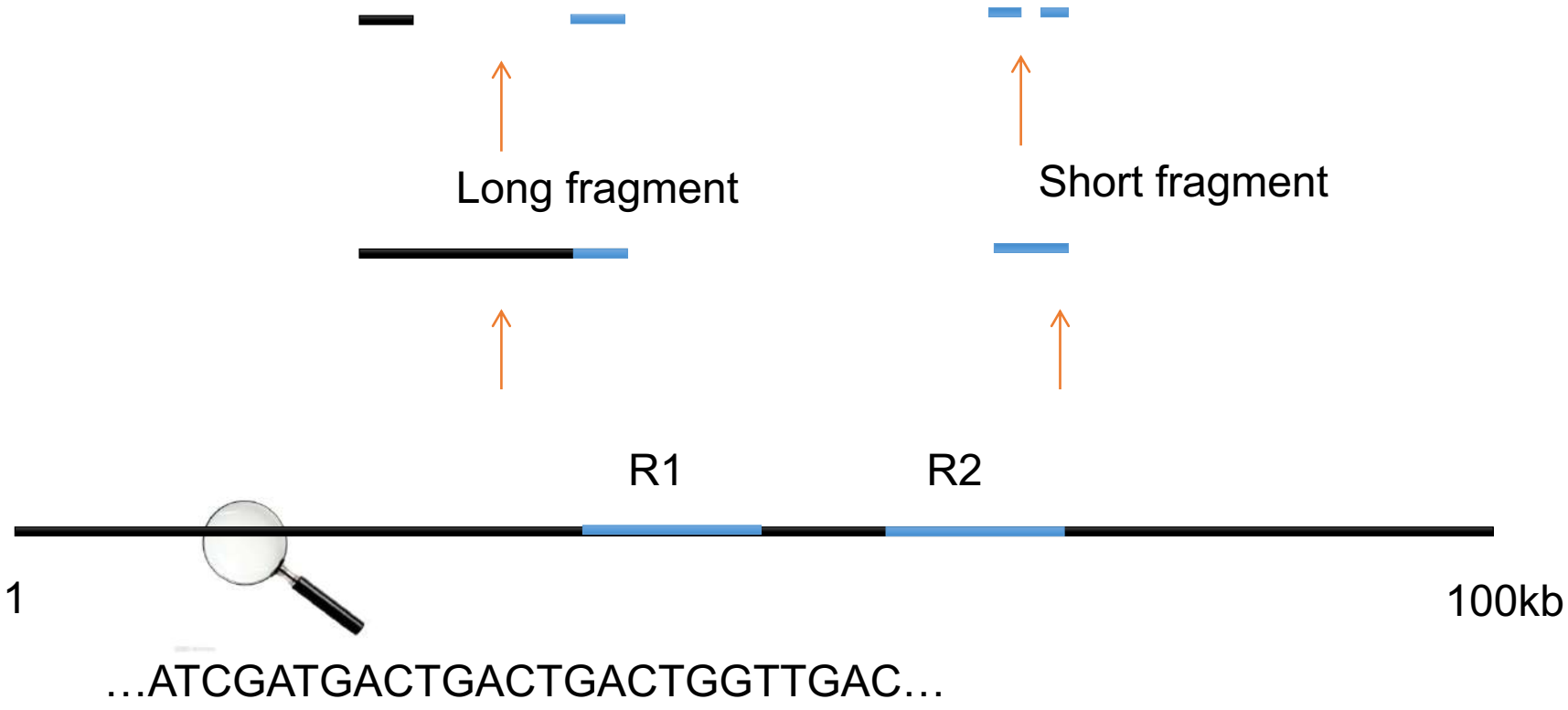
Figure 6. Genome mis-assembled due to a repeat. The assembly program incorrectly combined the reads from the two copies of the repeat leading to the creation of two separate contigs

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

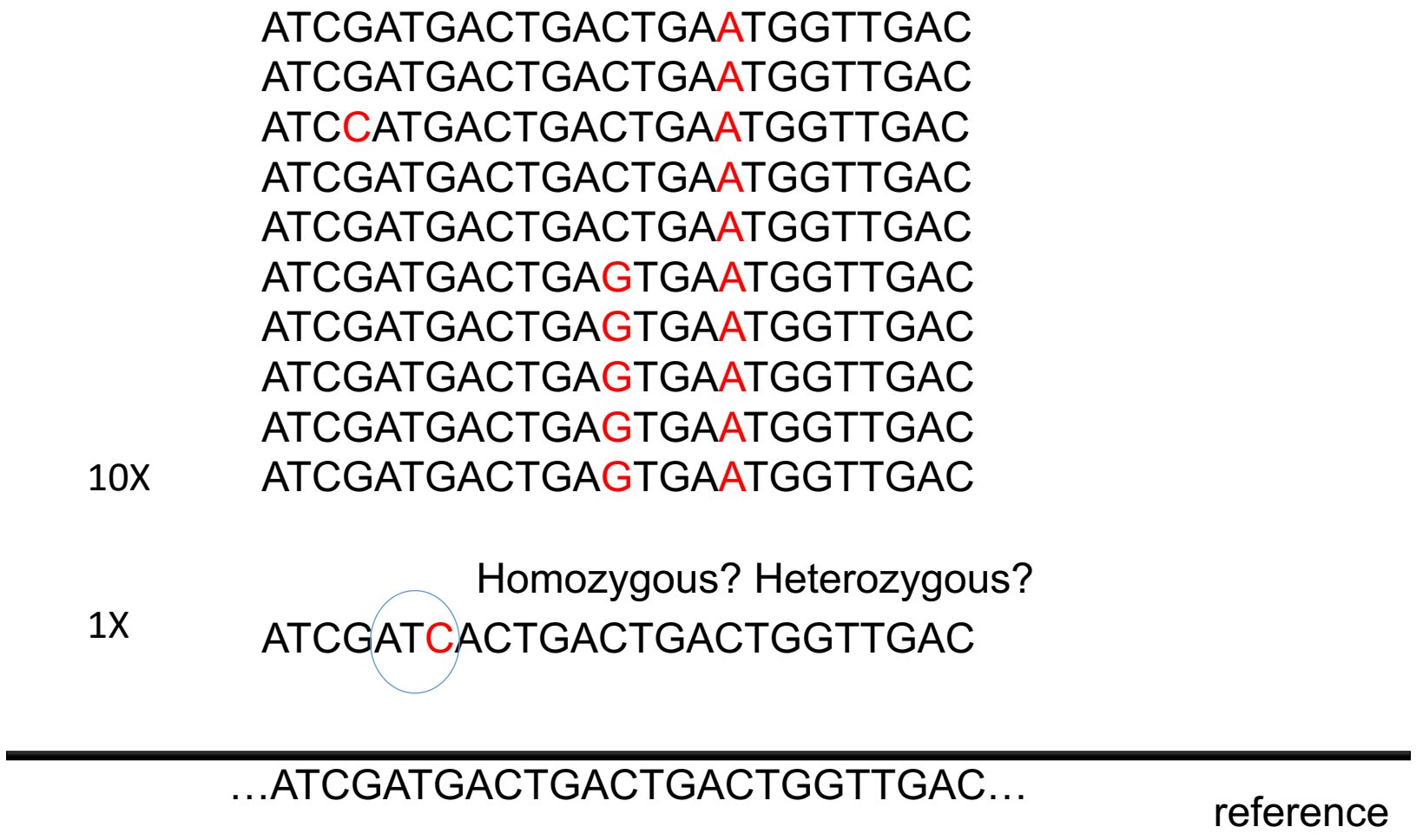
Read length matters in sequencing



Paired end and insert size matter in sequencing



Depth matters in sequencing



Case studies

Classical genetics

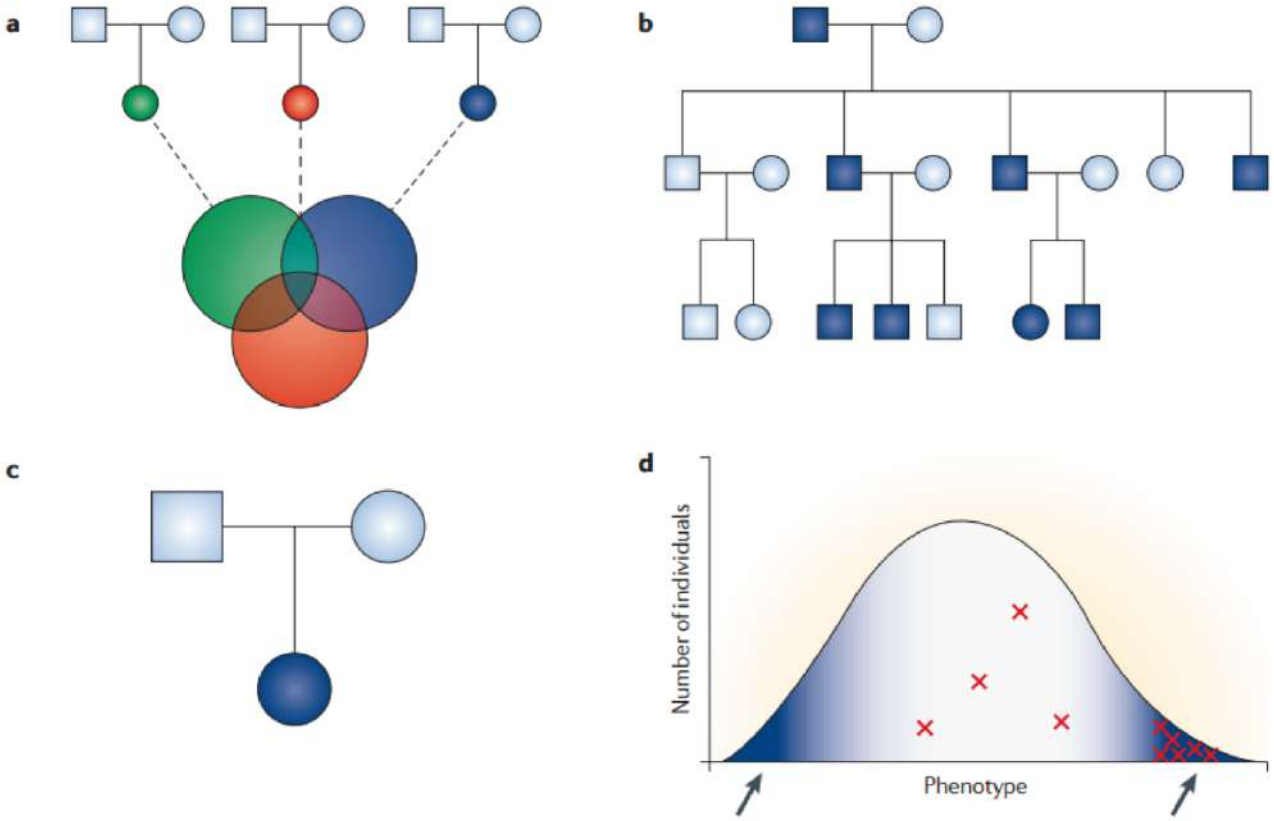
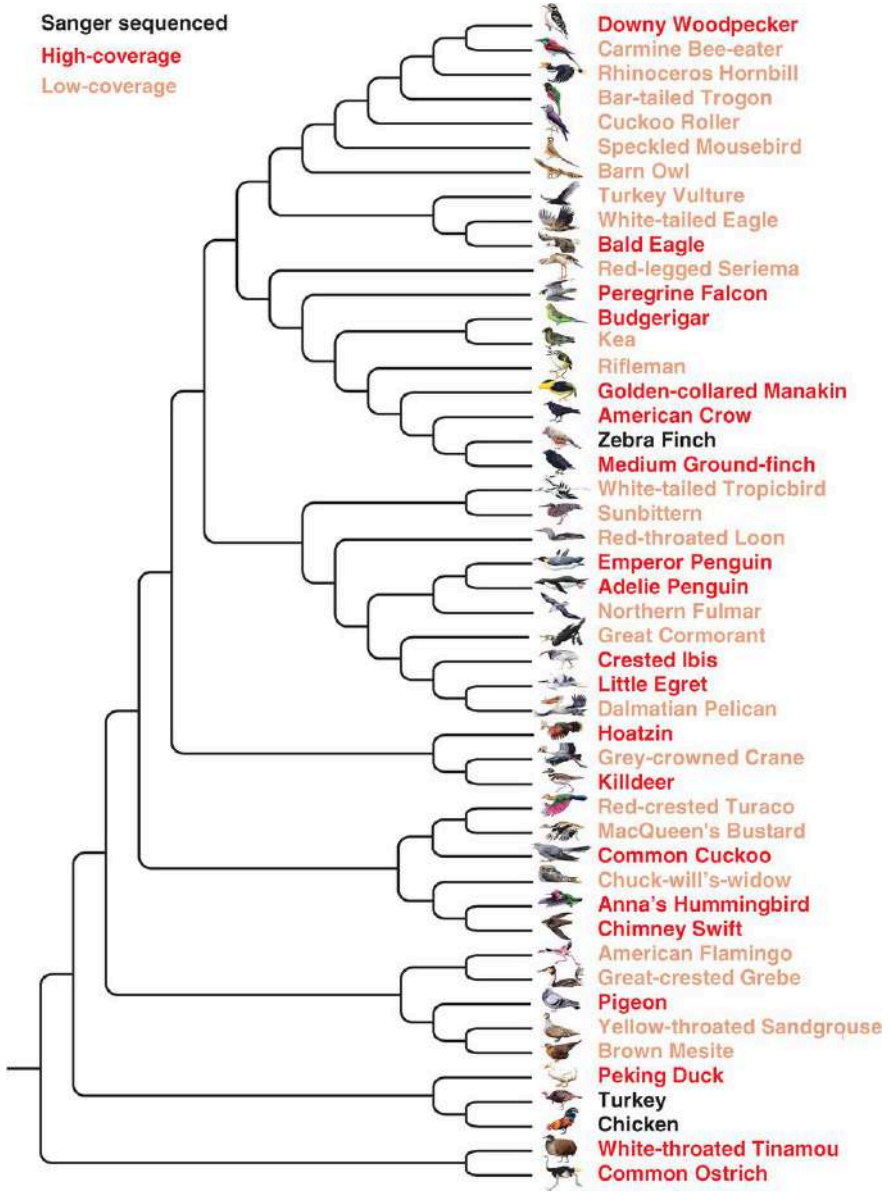


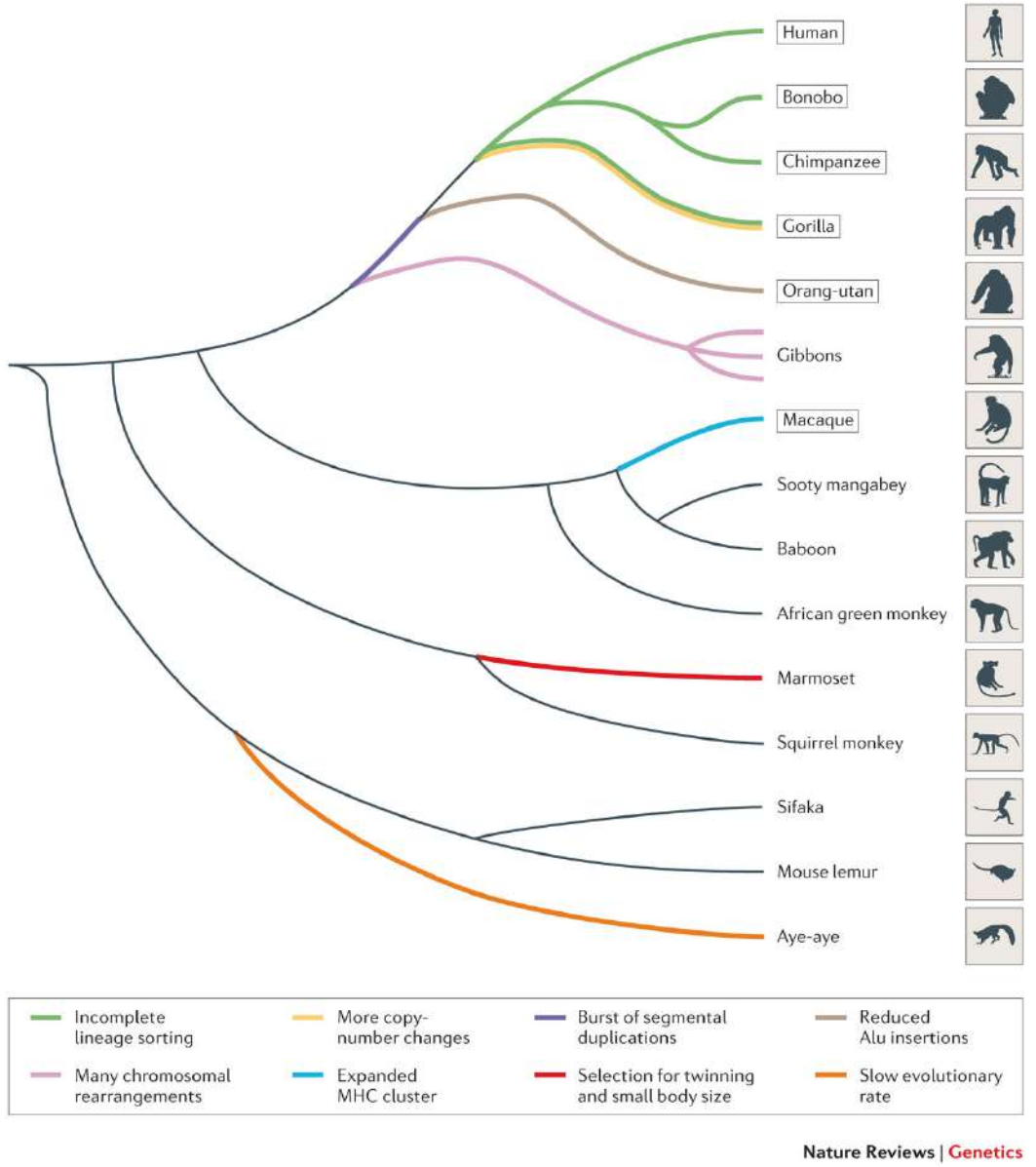
Figure 2 | **Strategies for finding disease-causing rare variants using exome sequencing.** Four main strategies are illustrated. **a** | Sequencing and filtering across multiple unrelated, affected individuals (indicated by the three coloured circles). This approach is used to identify novel variants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

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

Comparative genomics



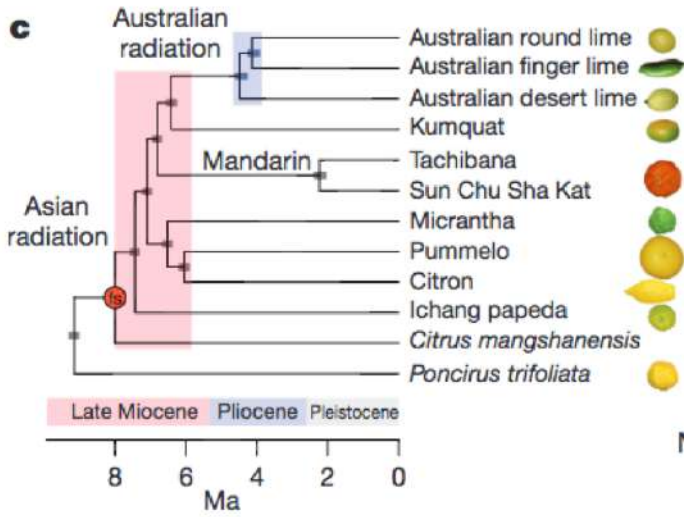
Guojie Zhang et al. Science (2014)

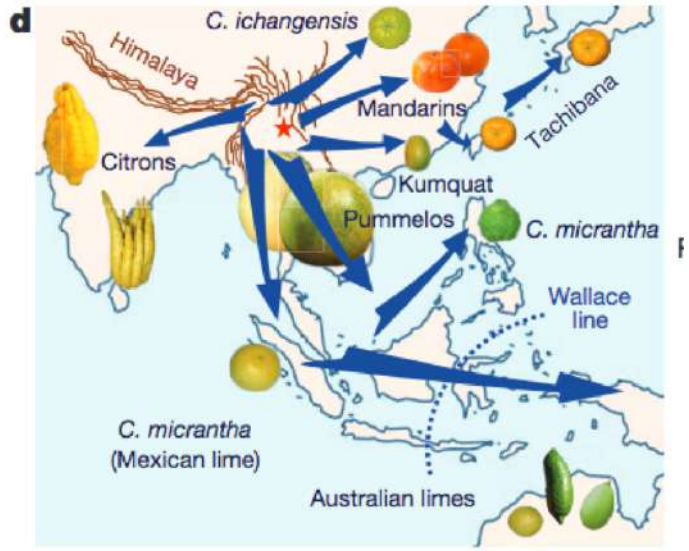


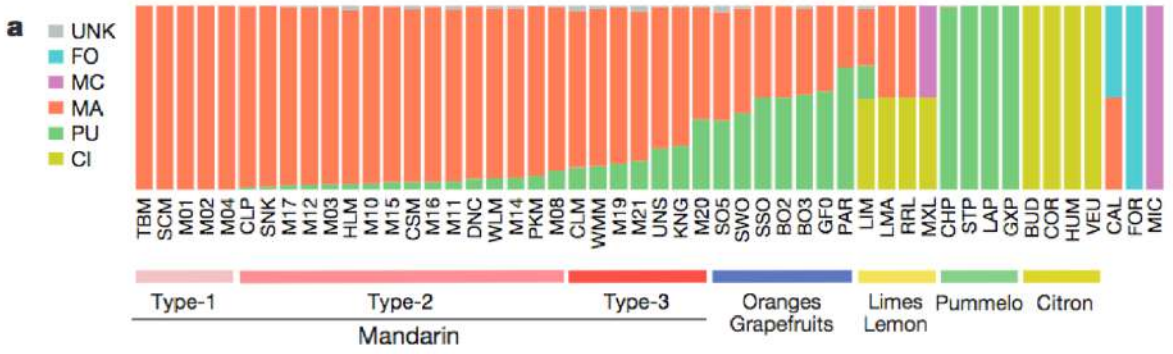
Roger & Gibbs Nature Reviews Genetics (2014)

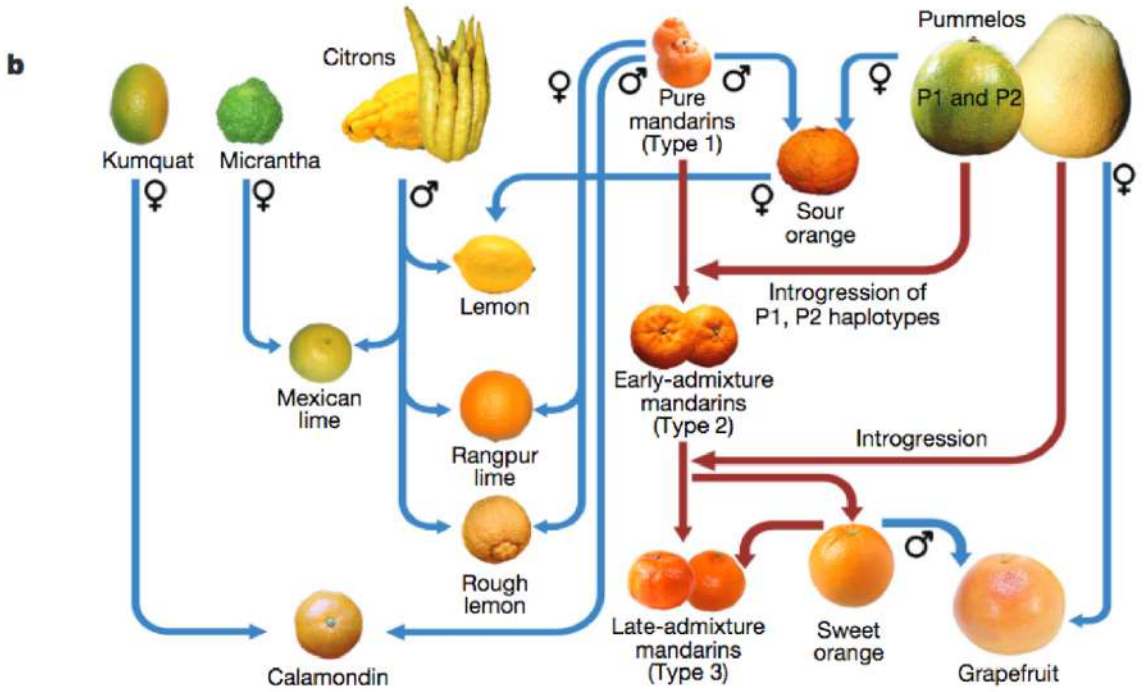
Comparative genomics

Genomics of the origin and evolution of Citrus









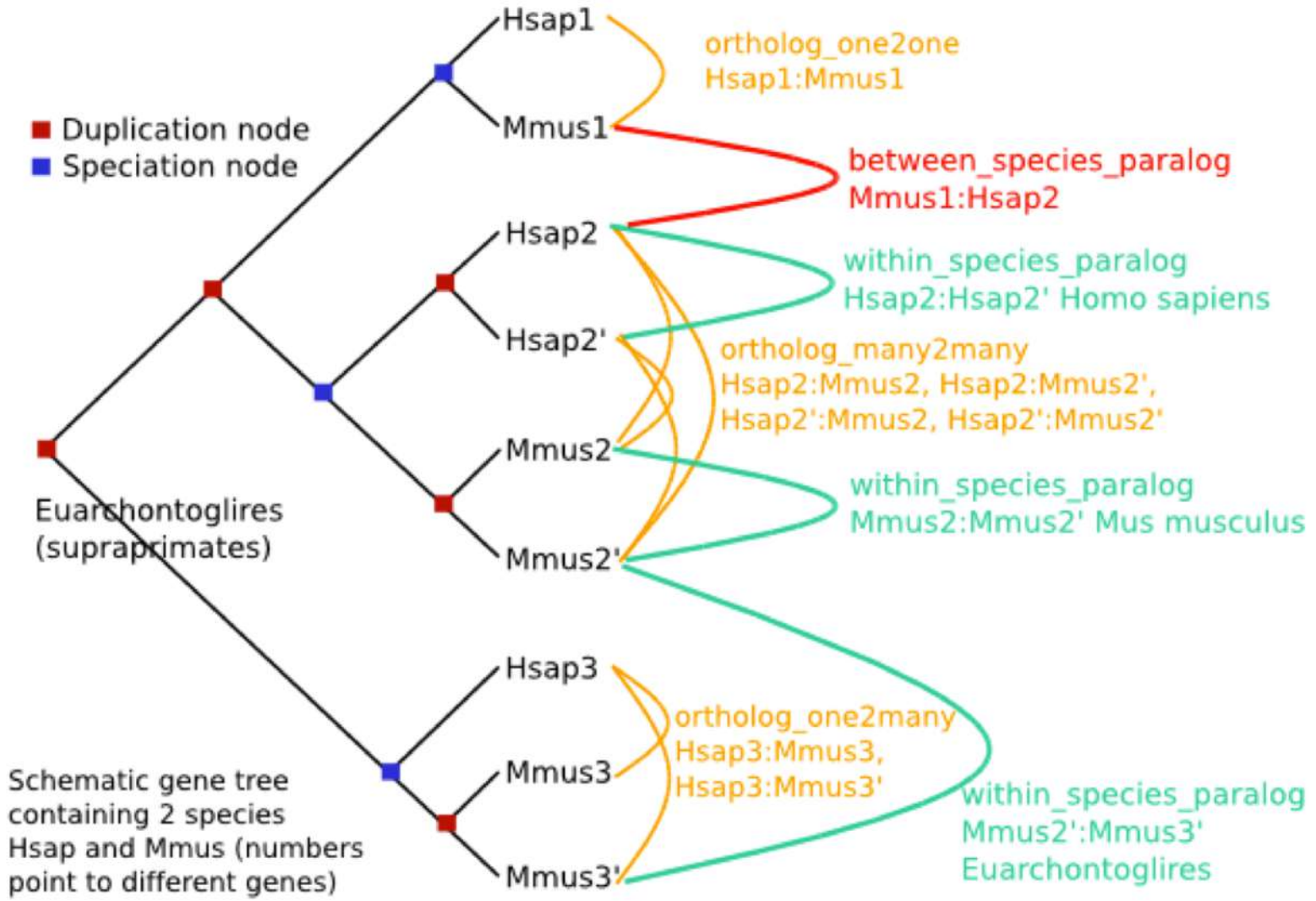
Wu *et al.*, Nature (2018)

Homologs: Orthologs and paralogs

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

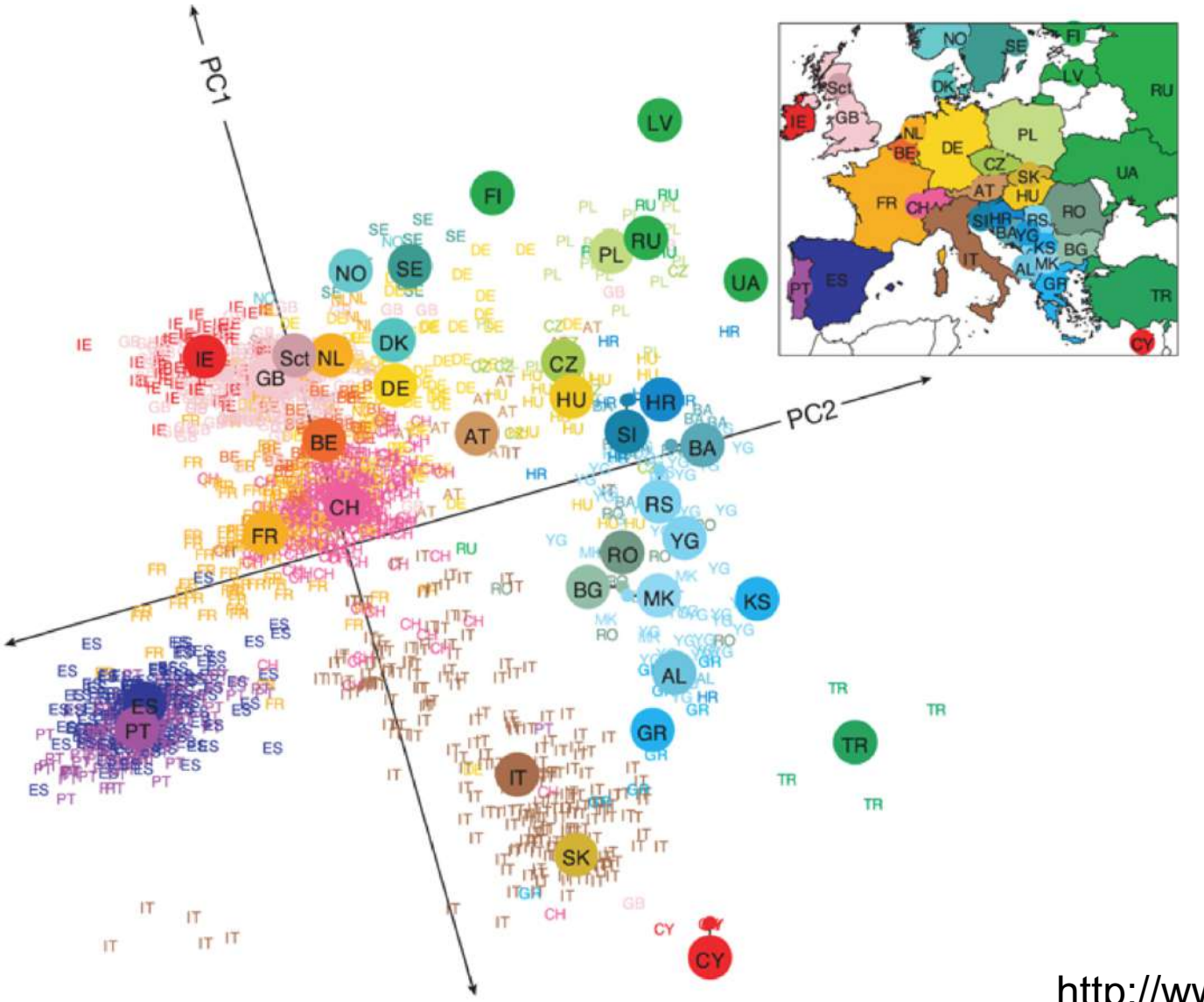
Depending on the number of genes found in each species, we differentiate among 1:1, 1:many and many:many relationships.

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

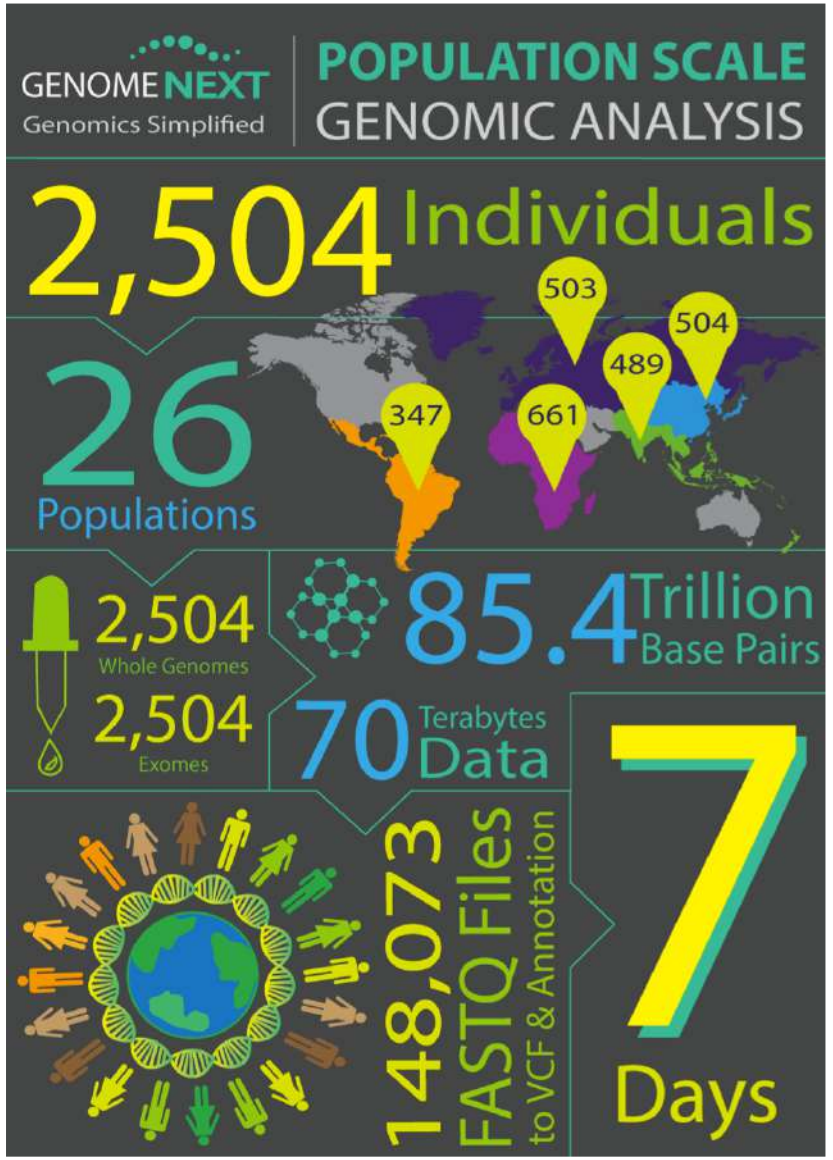


http://asia.ensembl.org/info/genome/compara/homology_method.html?redirect=no

Population genomics

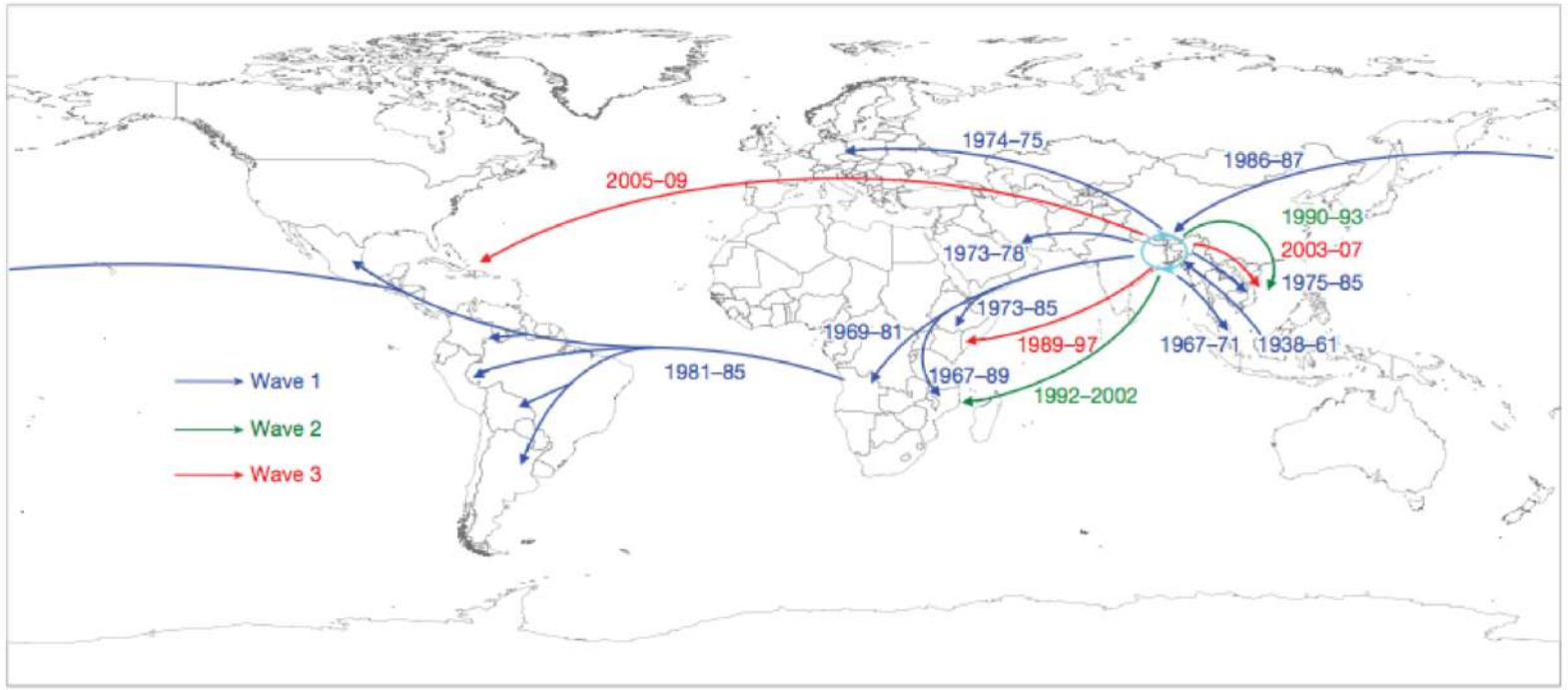


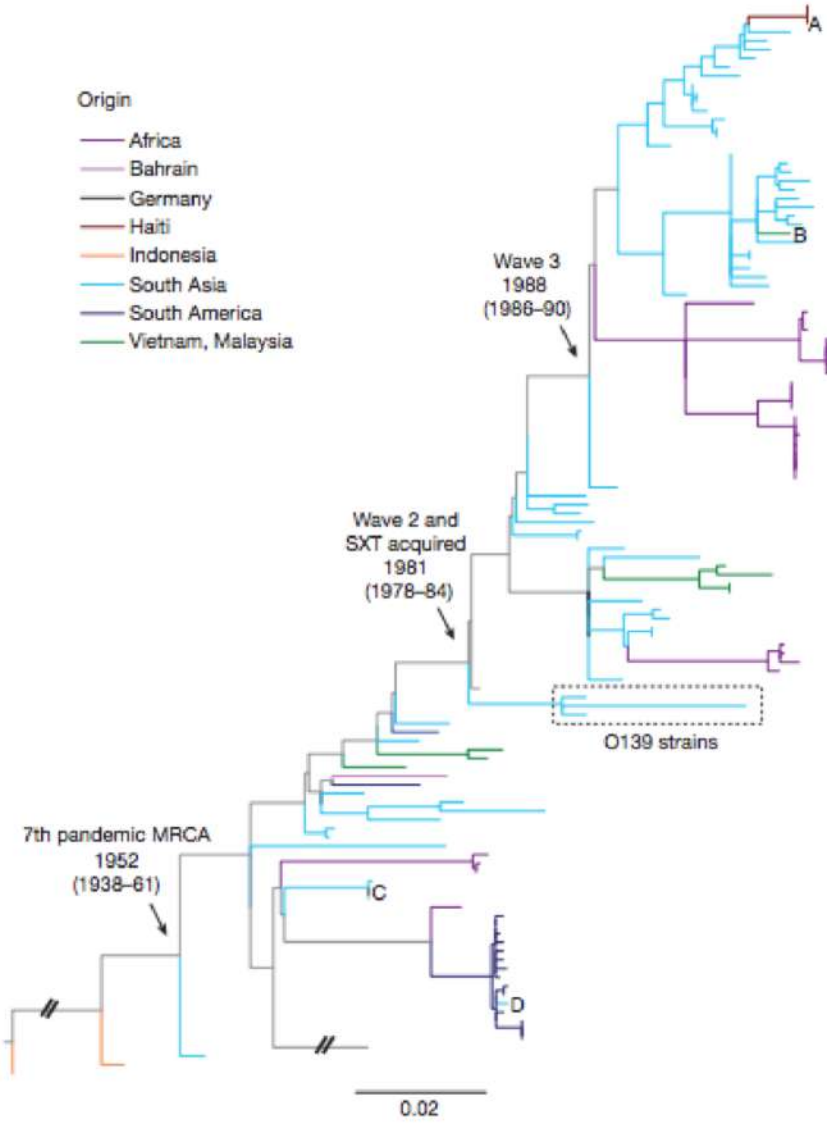
Novembre et al Nature (2008)



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

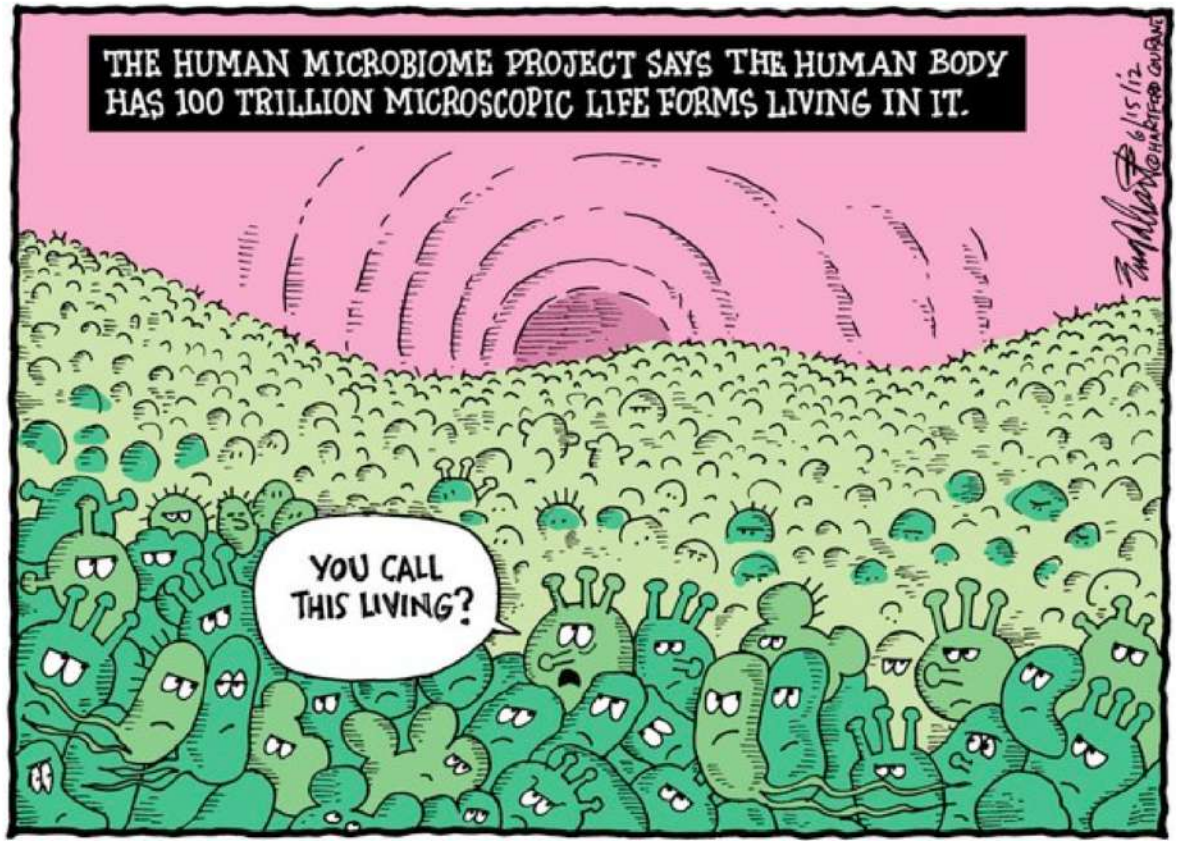
Population genomics

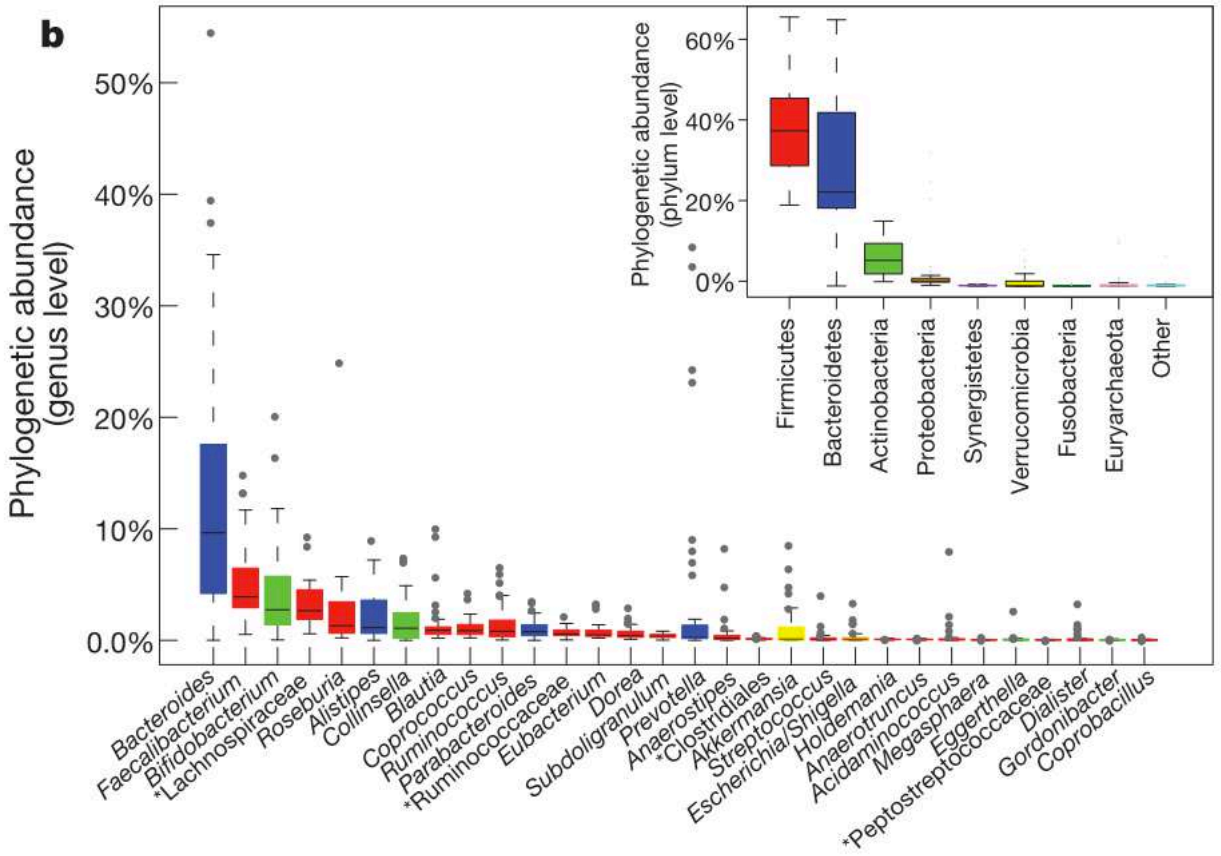




Mutreja *et al.*, Nature Genetics (2011)

Metagenomics

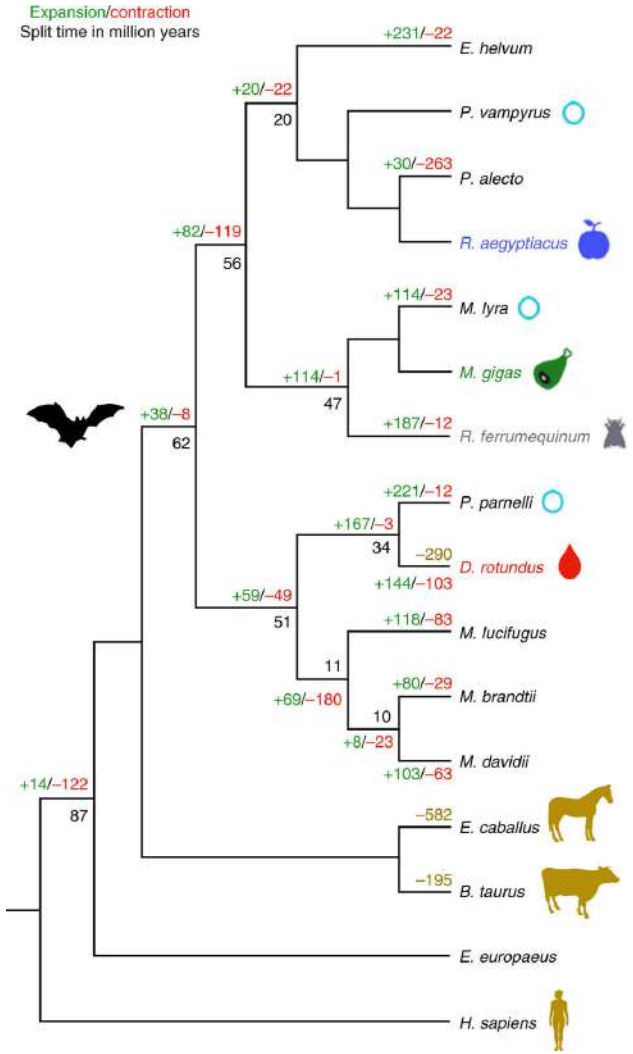


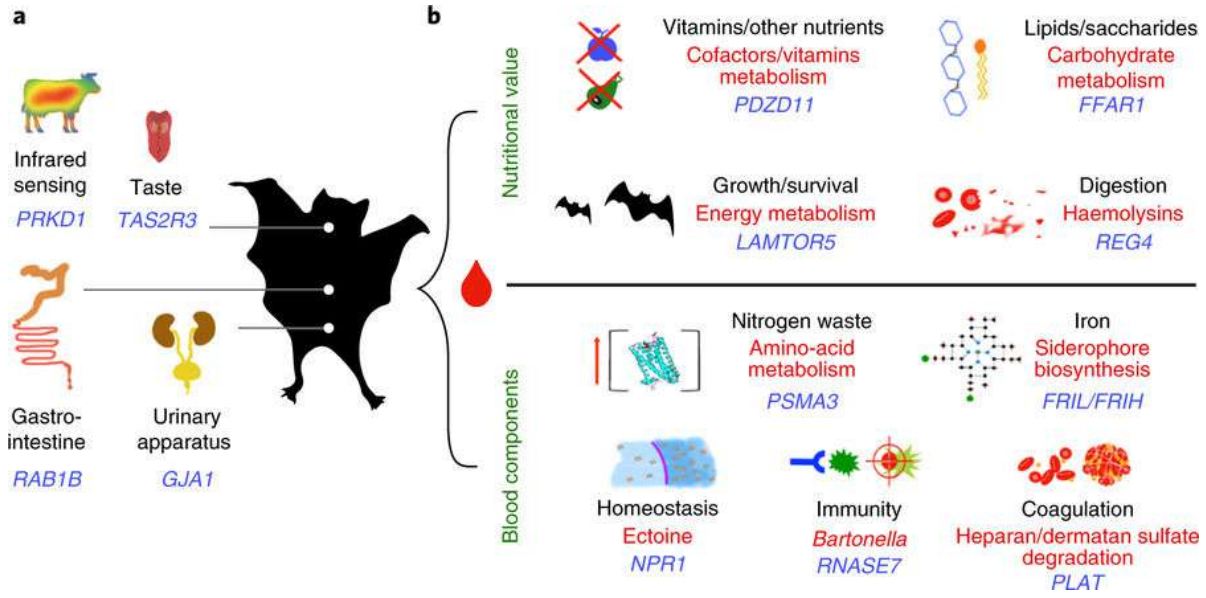


doi:10.1038/nature09944

Case study (2018)

Hologenomic adaptations underlying the evolution of sanguivory in the common vampire bat



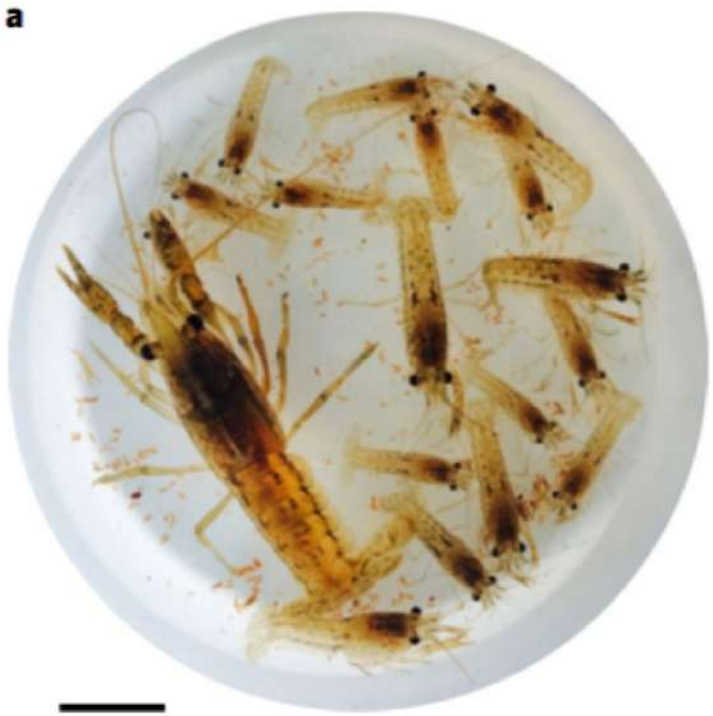


- *de novo* assembled its high-quality reference 2Gb genome (***de novo* assembly**)
- sequenced gut metagenome (**metagenomics**)
- compared them against those of insectivorous, frugivorous and carnivorous bats (**comparative genomics**)

- …we identified elements in both the host genome and microbiome that could have played relevant roles in adaptation to sanguivory.

10.1038/s41559-018-0476-8

Clonal genome evolution and rapid invasive spread of the marbled crayfish - story



1cm

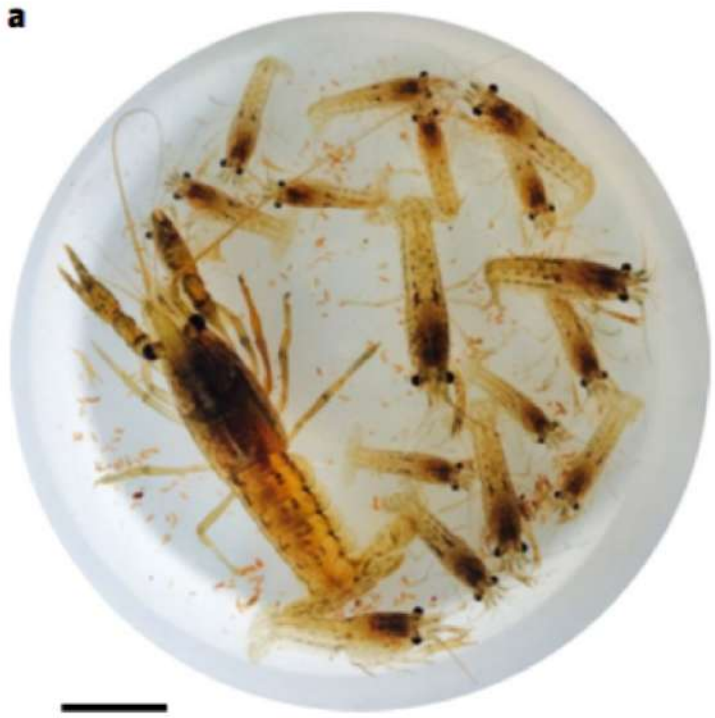
- The marbled crayfish became popular among German aquarium hobbyists in the late 1990s. The earliest report of the creature comes from a hobbyist who told Dr. Lyko he bought what were described to him as “Texas crayfish” in 1995.

- Soon the hobbyist was giving away the crayfish to his friends. And not long afterward, so-called marmorkrebs were showing up in pet stores in Germany and beyond.
- As marmorkrebs became more popular, owners grew increasingly puzzled. The crayfish seemed to be laying eggs without mating. The progeny were all female, and each one grew up ready to reproduce.

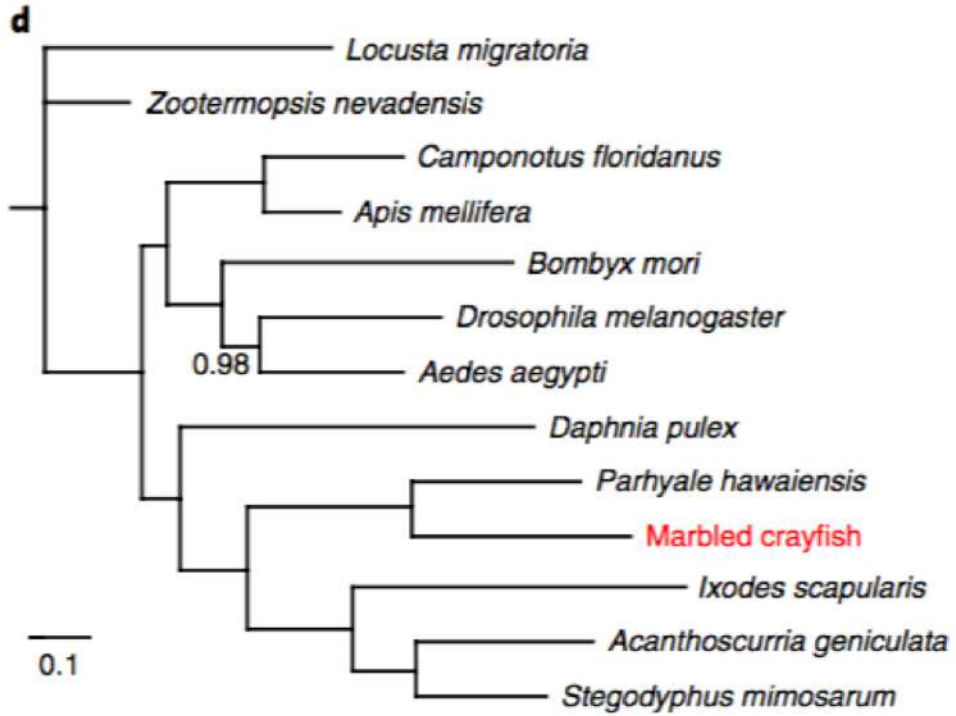
- “People would start out with a single animal, and a year later they would have a couple hundred,” said Dr. Lyko.
- Many owners apparently drove to nearby lakes and dumped their marmorkrebs. And it turned out that the marbled crayfish didn’t need to be pampered to thrive. Marmorkrebs established growing populations in the wild, sometimes walking hundreds of yards to reach new lakes and streams. Feral populations started turning up in the Czech Republic, Hungary, Croatia and Ukraine in Europe, and later in Japan and Madagascar.

10.1038/s41559-018-0476-8

Clonal genome evolution and rapid invasive spread of the marbled crayfish



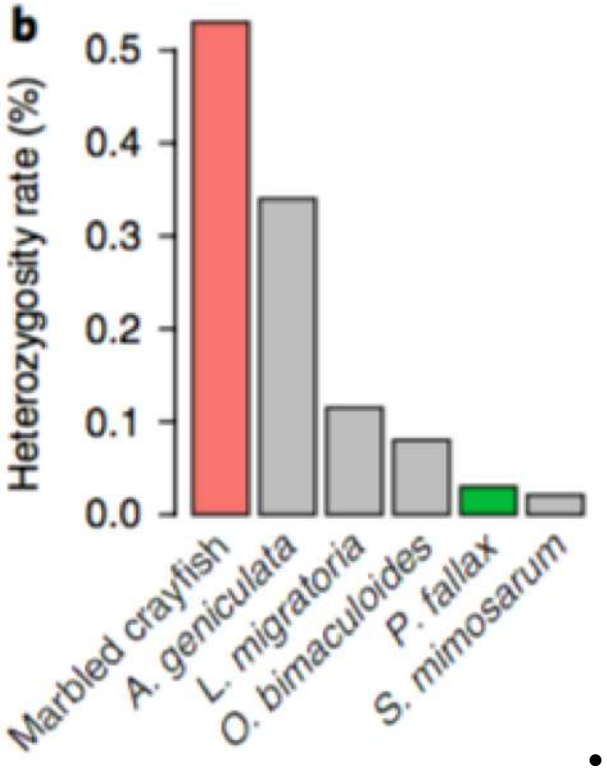
1cm

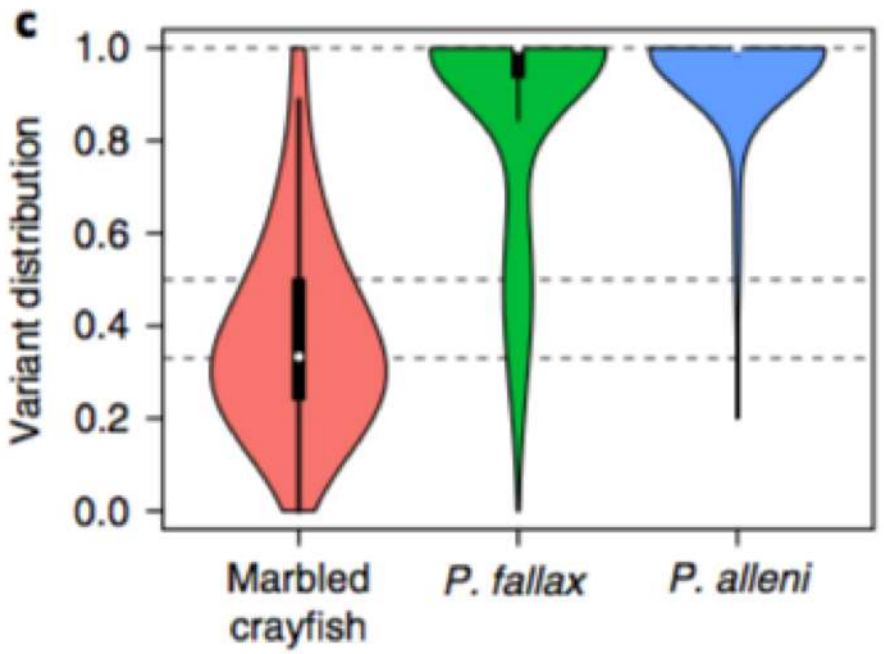


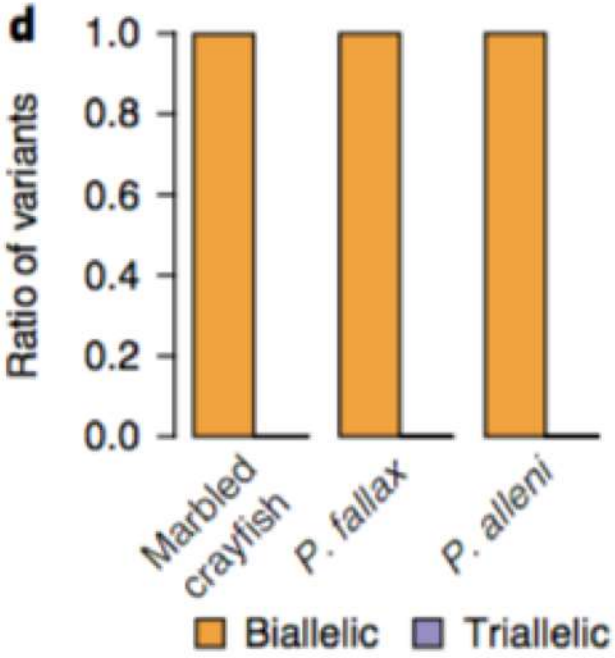
- Karyotyping shows a triploid (n=3) with **276** chromosomes
- *de novo* assembled 3.5Gb genome **(*de novo* assembly)**
- confirmed the close relationship to the genome of the slough crayfish, *Procambarus fallax ;* Male slough crayfish will readily mate with the marbled crayfish, but they never father any of the offspring.

10.1038/s41559-018-0476-8

Clonal genome evolution and rapid invasive spread of the marbled crayfish



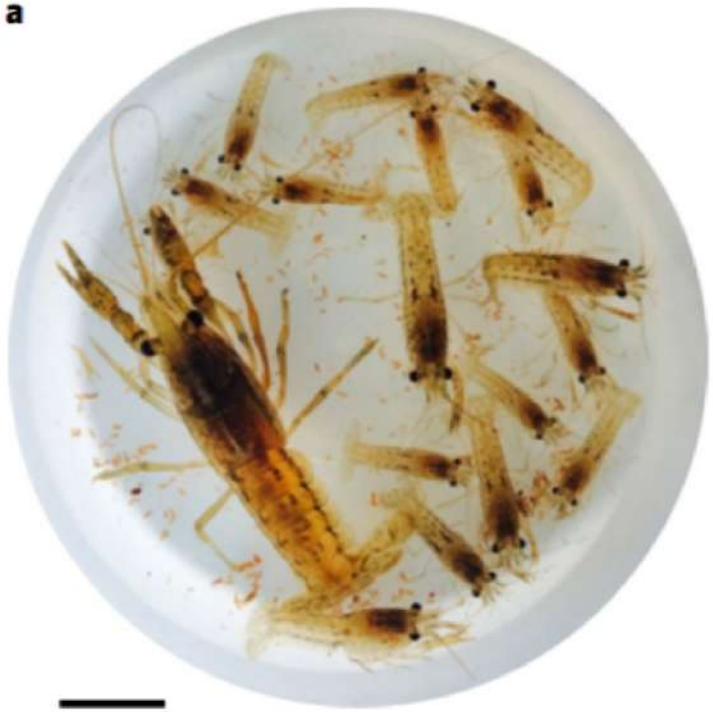




- Established a triploid AA'B genotype with a high level of heterozygosity (**population genomics**)
- The new species got its start when two slough crayfish mated. One of them had a mutation in a sex cell — whether it was an egg or sperm, the scientists can't tell. Normal sex cells contain a single copy of each chromosome. But the mutant crayfish sex cell had two.

10.1038/s41559-018-0476-8

Clonal genome evolution and rapid invasive spread of the marbled crayfish - story

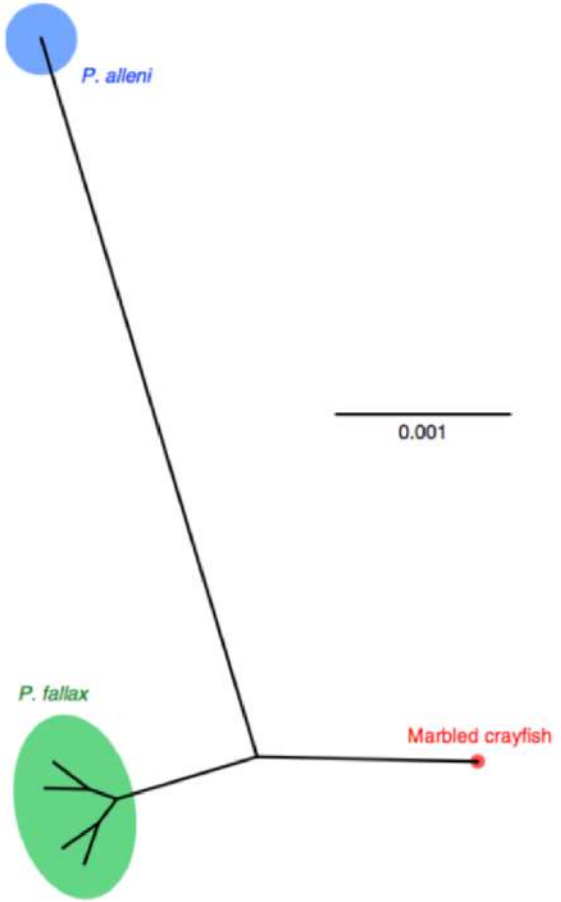


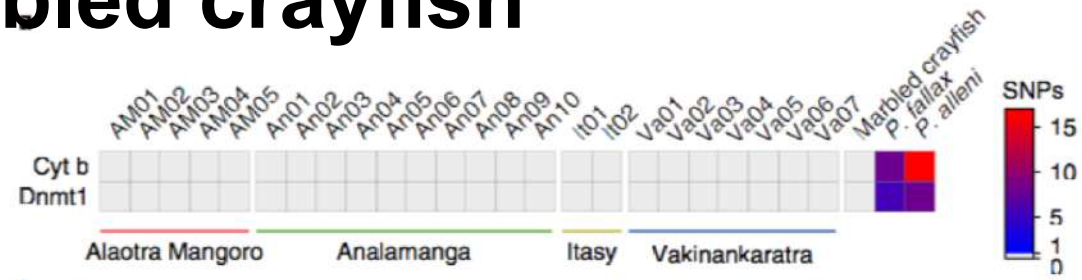
1cm

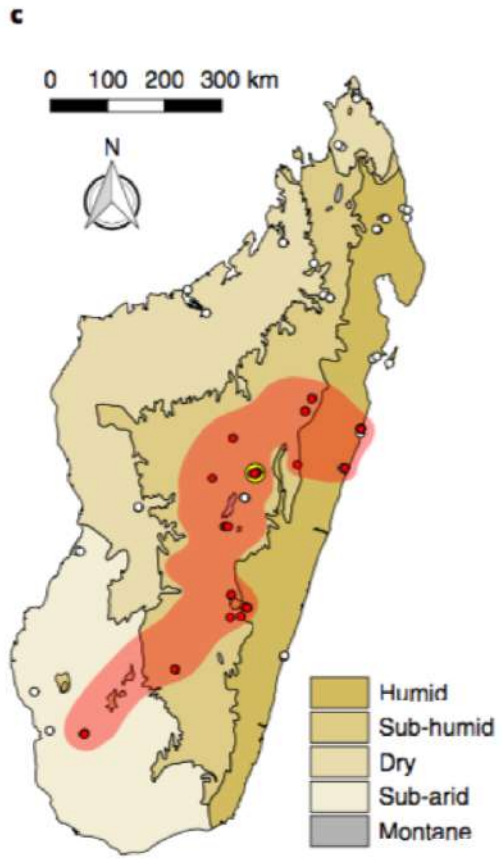
- In December 2017, Dr. Lyko and his colleagues officially declared the marbled crayfish to be a species of its own, which they named *Procambarus virginalis*. The scientists can't say for sure where the species began. There are no wild populations of marble crayfish in the United States, so **it's conceivable that the new species arose in a German aquarium.**

10.1038/s41559-018-0476-8

Clonal genome evolution and rapid invasive spread of the marbled crayfish



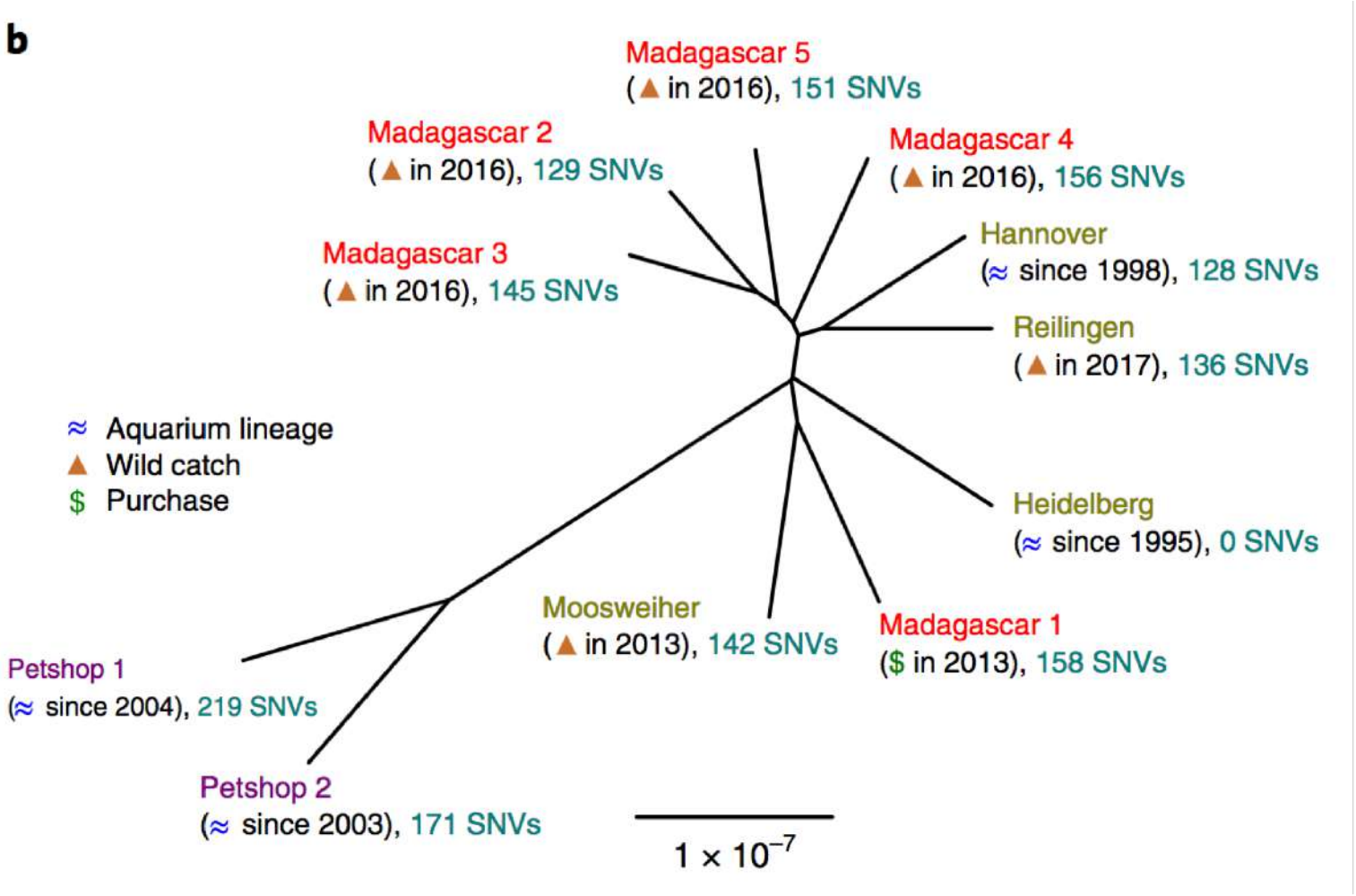




- genotyping demonstrated the rapid expansion of marbled crayfish on Madagascar and established the marbled crayfish as a potent invader of freshwater ecosystems. .**(population genomics)**
- comparative whole-genome sequencing demonstrated the clonality of the population and their genetic identity with the oldest known stock from the German aquarium trade **(population genomics)**

10.1038/s41559-018-0476-8

Clonal genome evolution and rapid invasive spread of the marbled crayfish



- comparative whole-genome sequencing demonstrated the clonality of the population and their genetic identity with the oldest known stock from the German aquarium trade **(population genomics)**

10.1038/s41559-018-0476-8

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)

PNAS

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

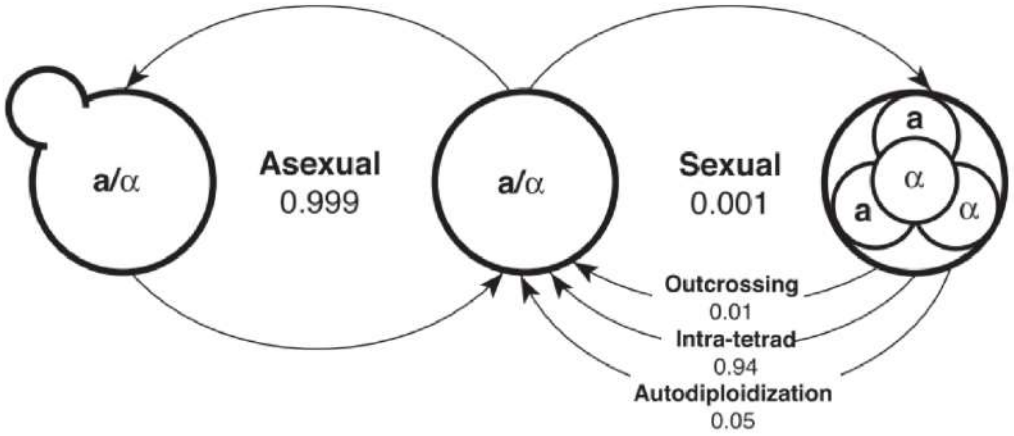
Edited by Mark Johnston, Washington University, St. Louis, MO, and accepted by the Editorial Board January 30, 2008 (received for review August 3, 2007)

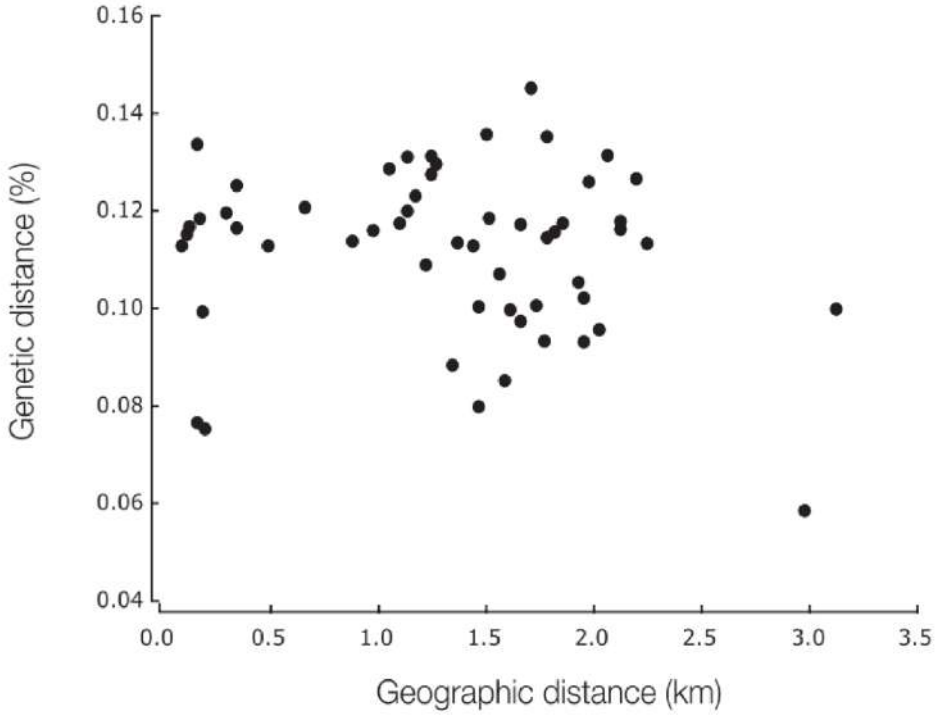
Most microbes have complex life cycles with multiple modes of reproduction that differ in their effects on DNA sequence variation. Population genomic analyses can therefore be used to estimate the

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)

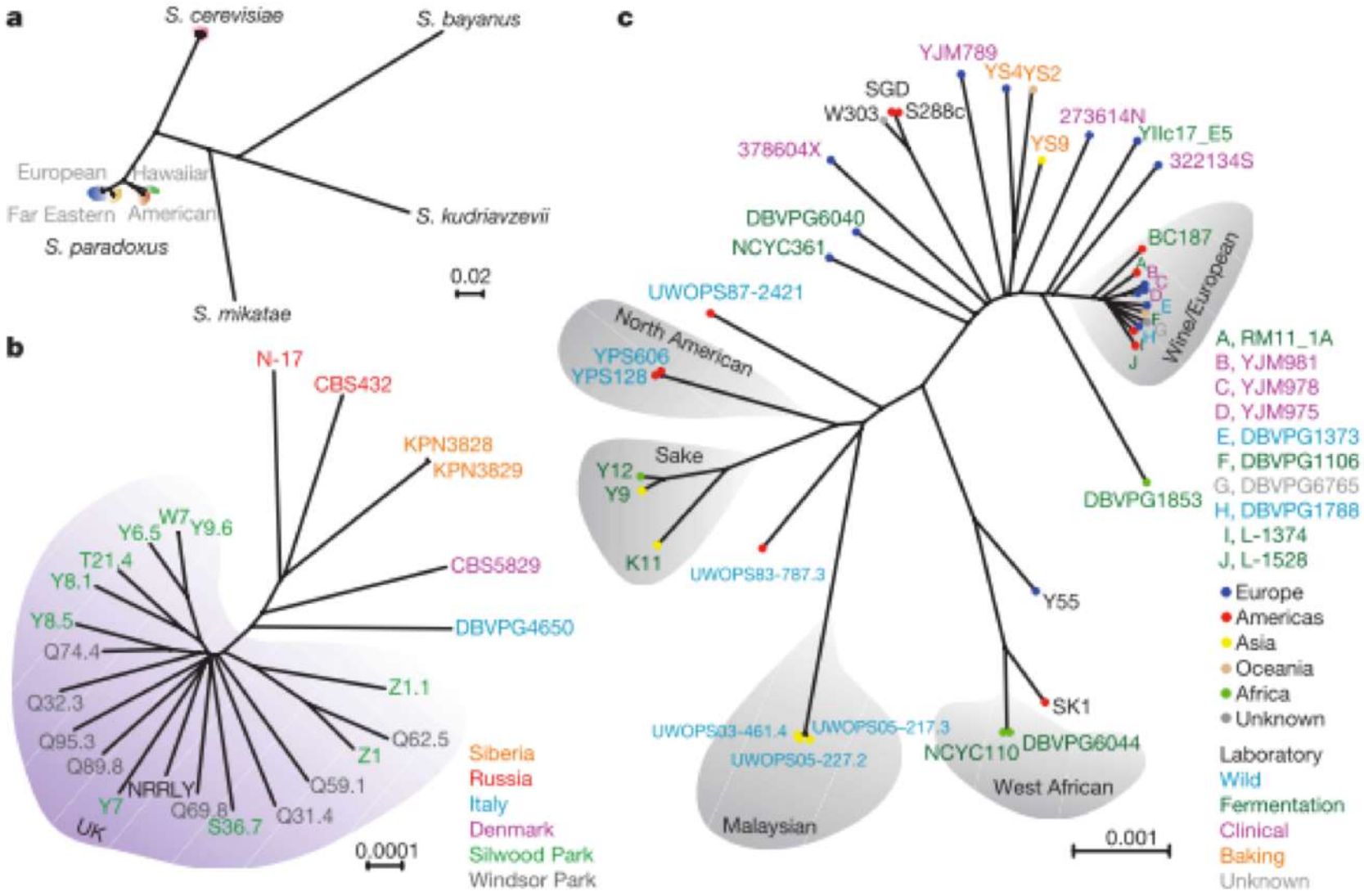
- **Now= We are collecting and sequencing hundreds of isolates in Taiwan**

Population genomics of domestic and wild yeasts

Gianni Liti[1]*, David M. Carter[2]*, Alan M. Moses[2,3], Jonas Warringer[4], Leopold Parts[2], Stephen A. James[5], Robert P. Davey[5], Ian N. Roberts[5], Austin Burt[6], Vassiliki Koufopanou[6], Isheng J. Tsai[6], Casey M. Bergman[7], Douda Bensasson[7], Michael J. T. O'Kelly[8], Alexander van Oudenaarden[8], David B. H. Barton[1], Elizabeth Bailes[1], Alex N. Nguyen Ba[3], Matthew Jones[2], Michael A. Quail[2], Ian Goodhead[2]†, Sarah Sims[2], Frances Smith[2], Anders Blomberg[4], 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

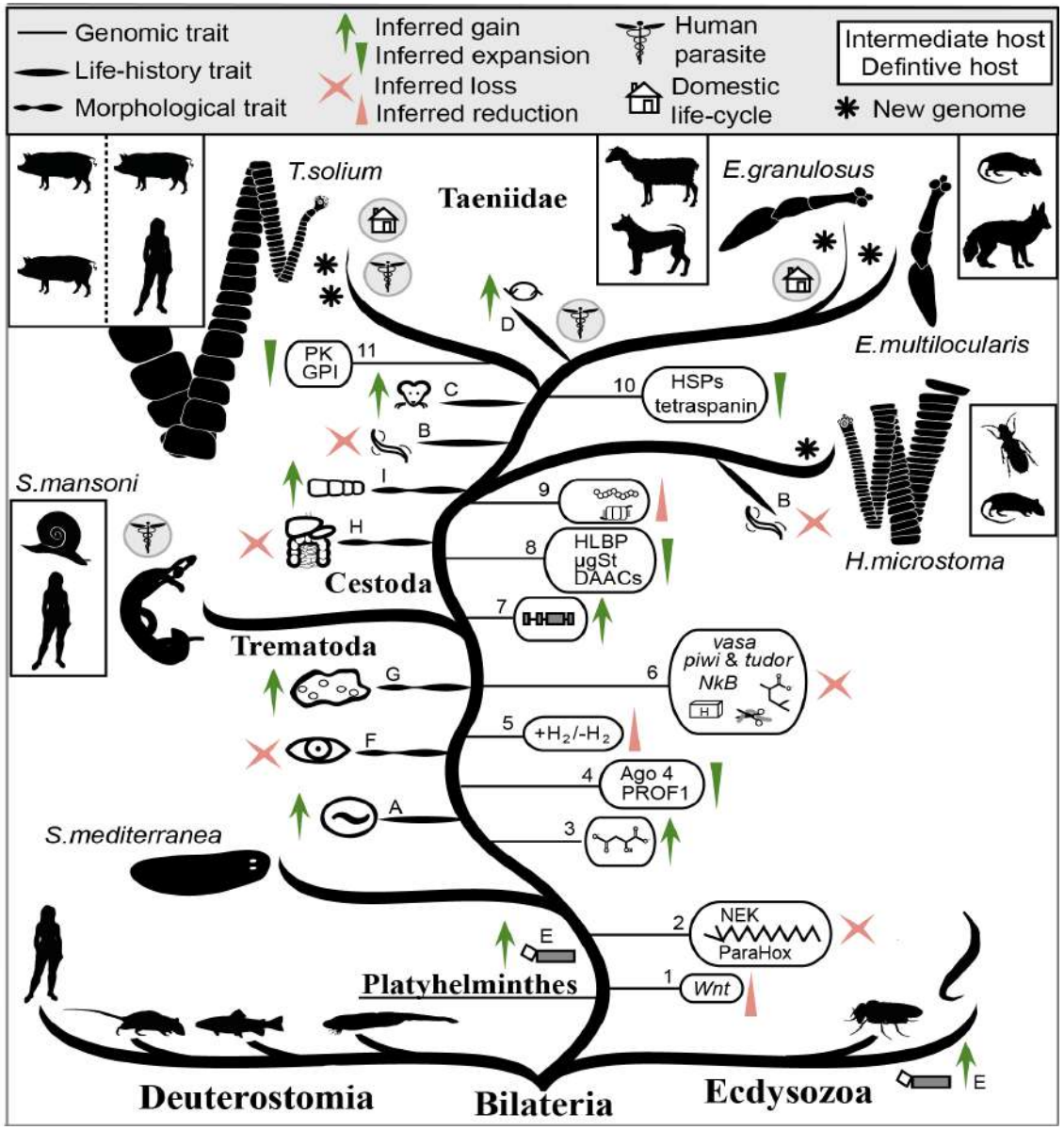
OPEN

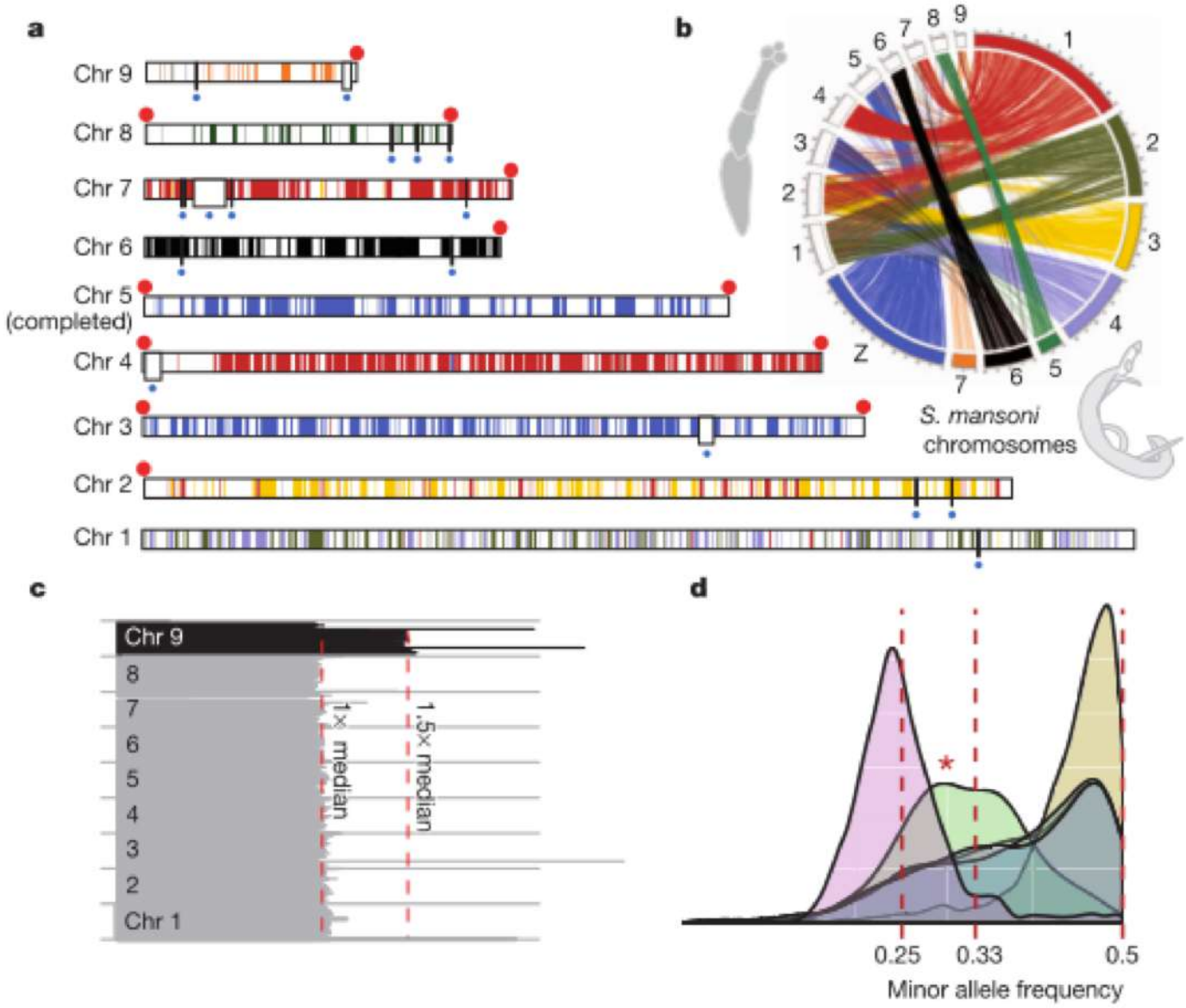
doi:10.1038/nature12031

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. Brooks[1], Alan Tracey[1], Raúl J. Bobes[4], Gladis Fragoso[4], Edda Sciutto[4], Martin Aslett[1], Helen Beasley[1], Hayley M. Bennett[1], Jianping Cai[5], Federico Camicia[6], Richard Clark[1], Marcela Cucher[6], Nishadi De Silva[1], Tim A. Day[7], Peter Deplazes[8], Karel Estrada[3], Cecilia Fernández[9], Peter W. H. Holland[10], Junling Hou[5], Songnian Hu[11], Thomas Huckvale[1], Stacy S. Hung[12], Laura Kamenetzky[6], Jacqueline A. Keane[1], Ferenc Kiss[13], Uriel Koziol[13], Olivia Lambert[1], Kan Liu[11], Xuenong Luo[5], Yingfeng Luo[11], Natalia Macchiaroli[6], Sarah Nichol[1], Jordi Paps[10], John Parkinson[12], Natasha Pouchkina-Stantcheva[14], Nick Riddiford[14,15], Mara Rosenzvit[6], Gustavo Salinas[9], James D. Wasmuth[16], Mostafa Zamanian[17], Yadong Zheng[5], The *Taenia solium* Genome Consortium†, Xuepeng Cai[5], Xavier Soberón[3,18], Peter D. Olson[14], Juan P. Laclette[4], Klaus Brehm[13] & Matthew Berriman[1]

2013 – Tapeworm genome project





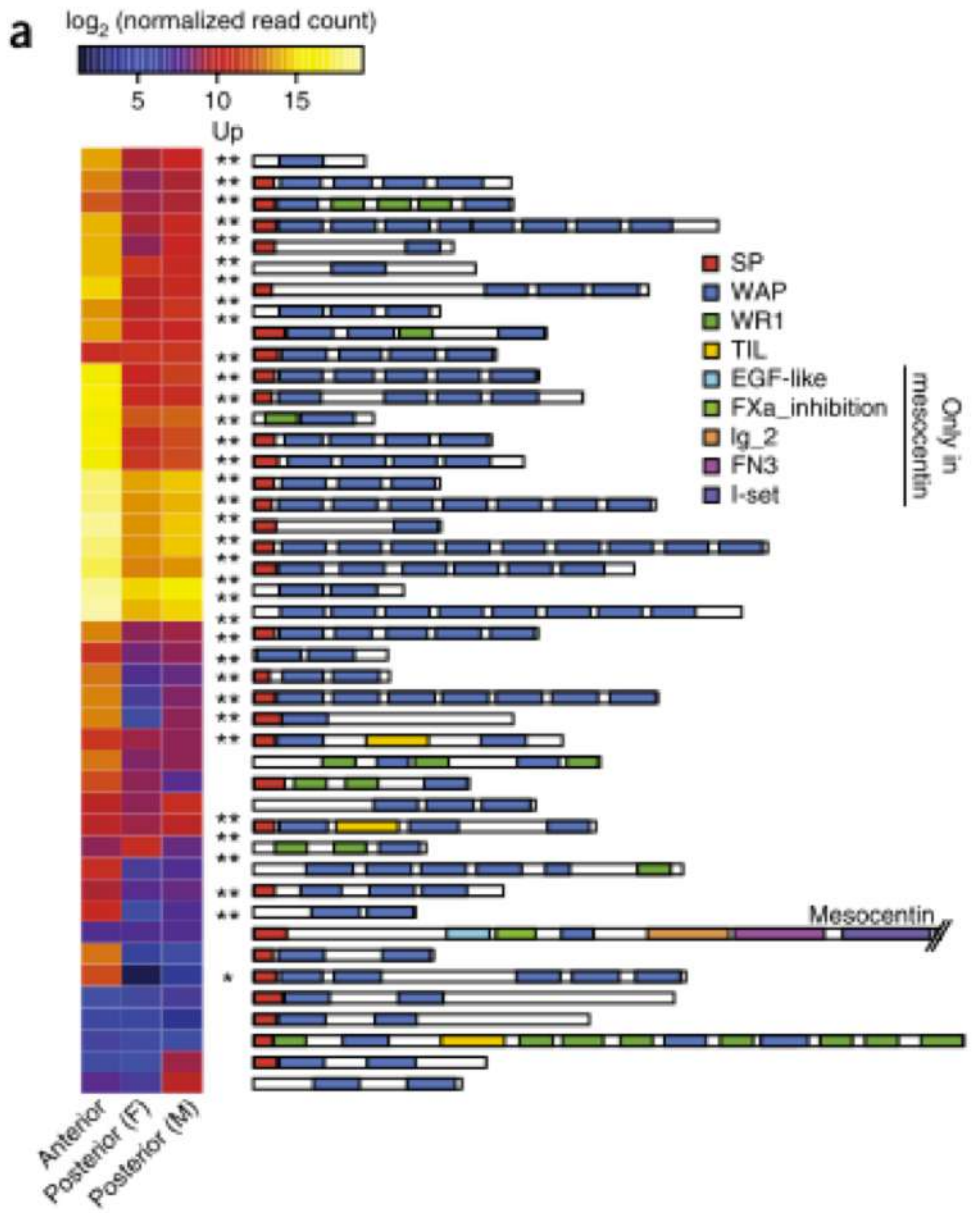
2014 – *Trichuris* genome project

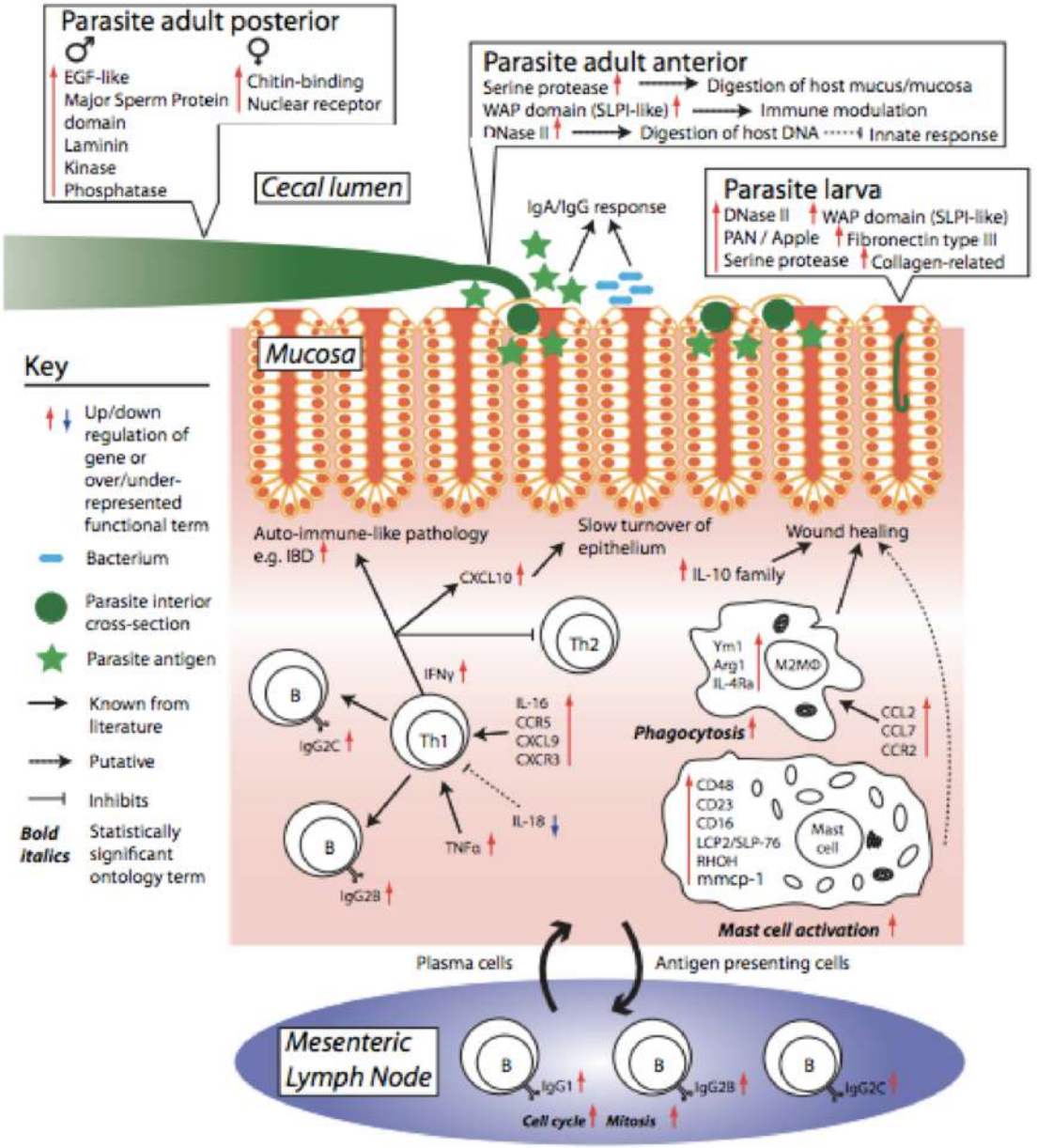
- 2 genomes probably costs less than £10,000k
- About **40 RNAseq** libraries of different life cycle stages, host infecting stages
- Paradigm shifts to RNAseq

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 Nichol[1], Alan Tracey[1], Nancy Holroyd[1], James A Cotton[1], Eleanor J Stanley[1], Magdalena Zarowiecki[1], Jimmy Z Liu[4], Thomas Huckvale[1], Philip J Cooper[5,6], Richard K Grencis[3] & Matthew Berriman[1]

2014 – *Trichuris* genome project





2014 – *Taphrina* genome project

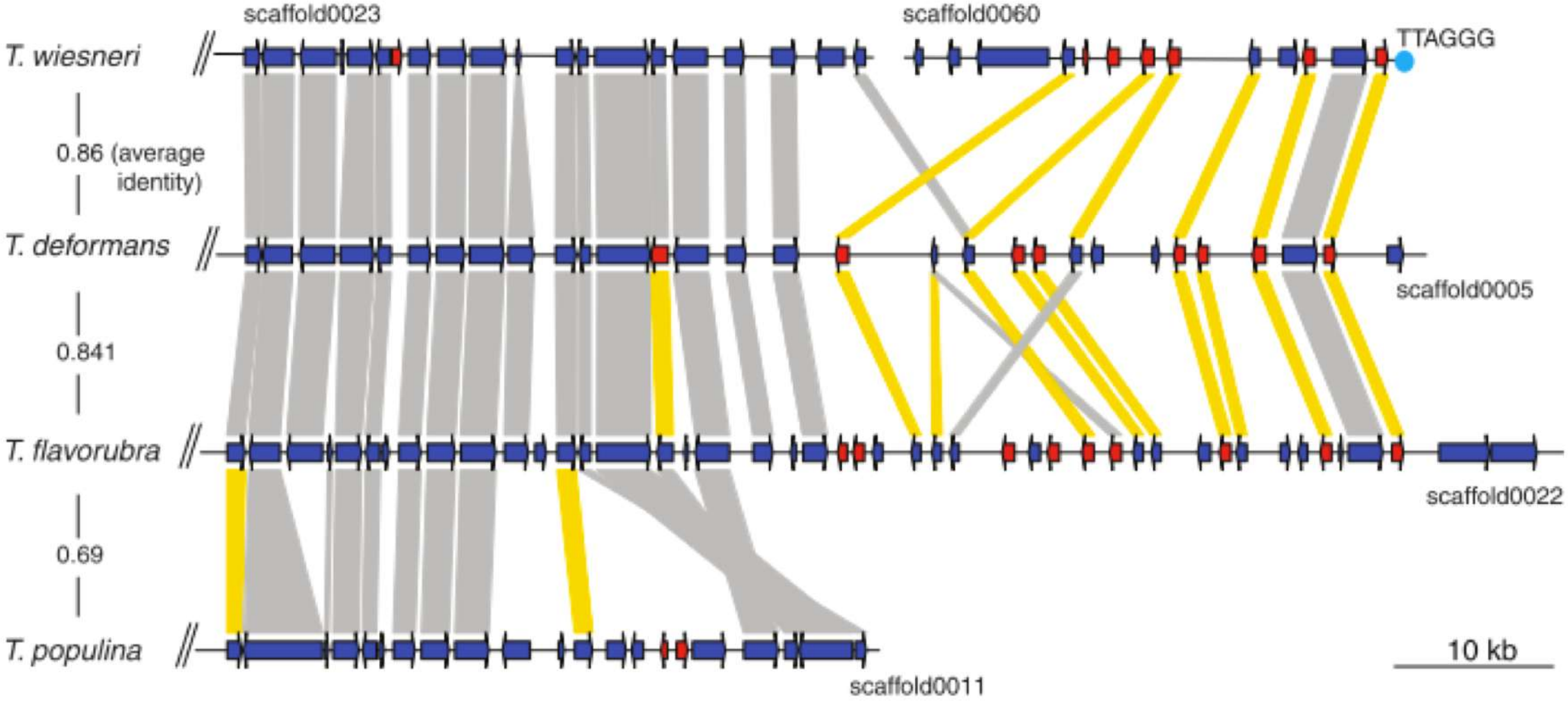
- 3 fungi genomes (~18Mb) of Illumina PE
- RNAseq for annotation purpose
- Costs probably less than 200,000 NT
- 2 months to analyse

GBE

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

Isheng J. Tsai[1,2], Eiji Tanaka[3], Hayato Masuya[4], Ryusei Tanaka[1], Yuuri Hirooka[5], Rikiya Endoh[6], Norio Sahashi[4], 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

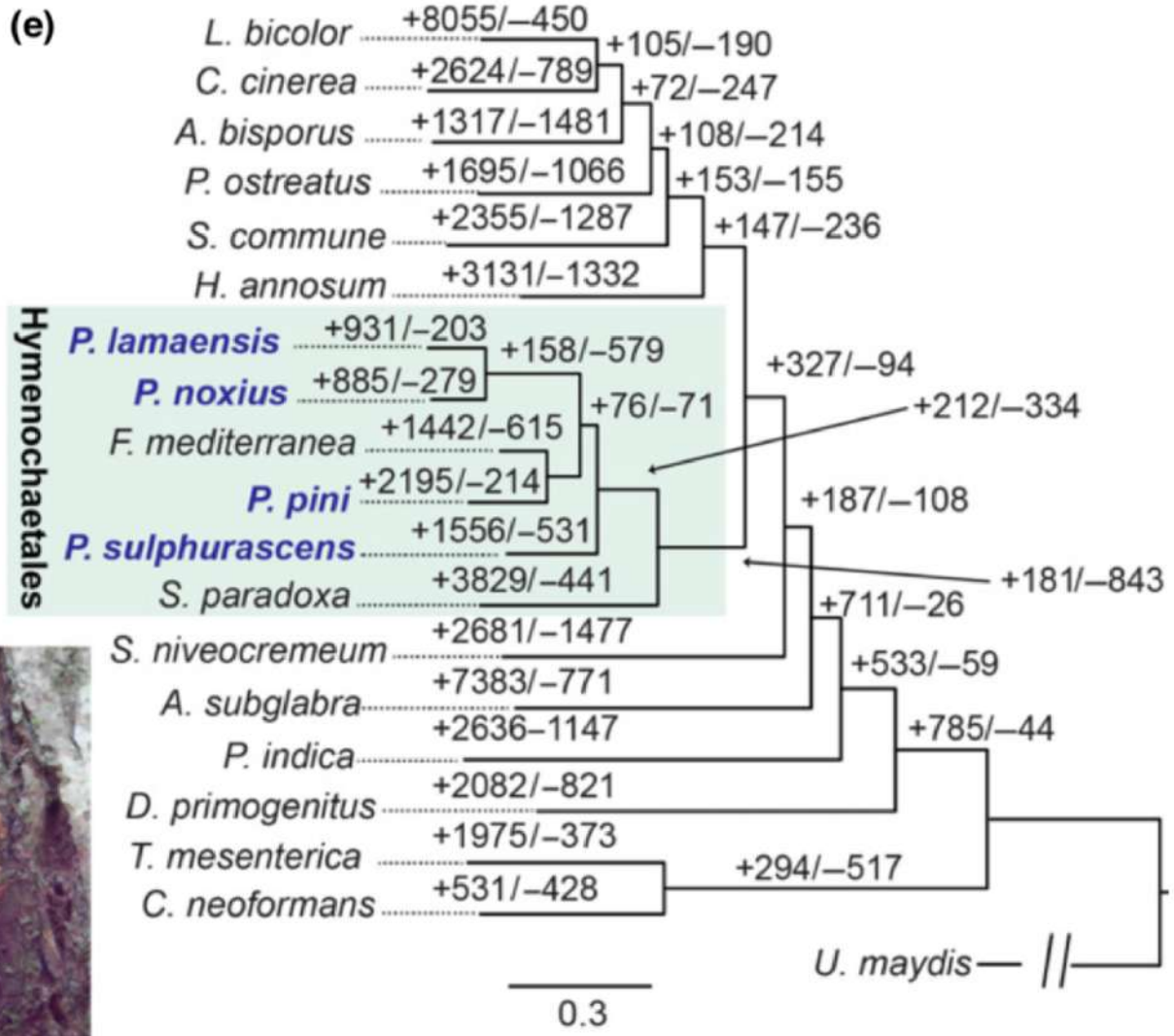
ORIGINAL ARTICLE

WILEY MOLECULAR ECOLOGY

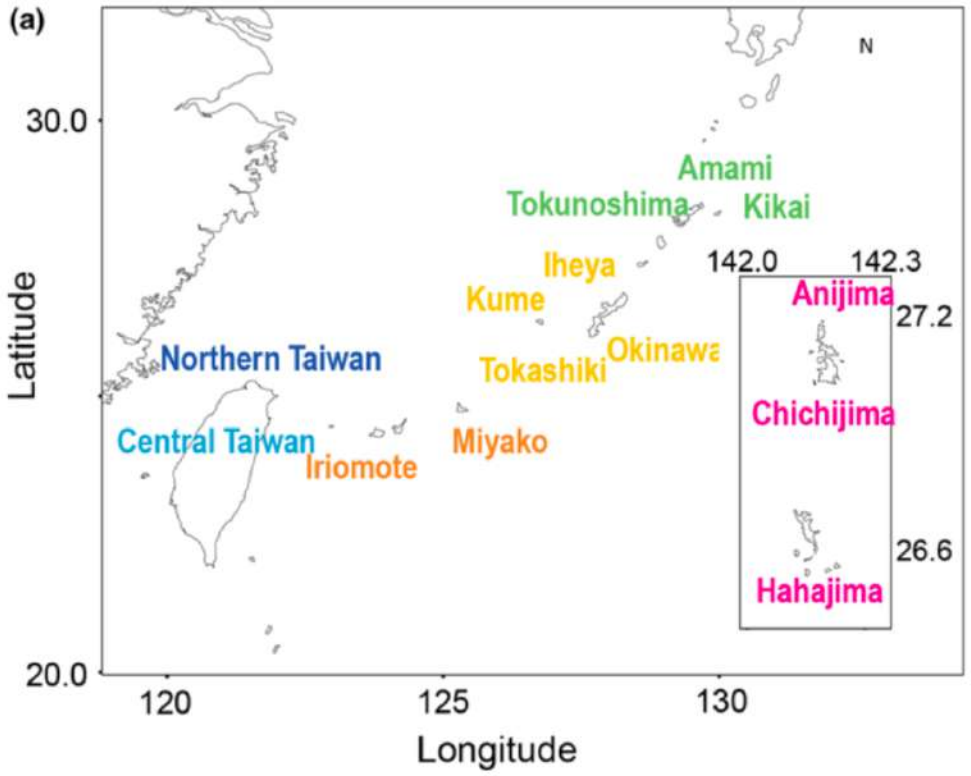
Comparative and population genomic landscape of *Phellinus noxius*: A hypervariable fungus causing root rot in trees

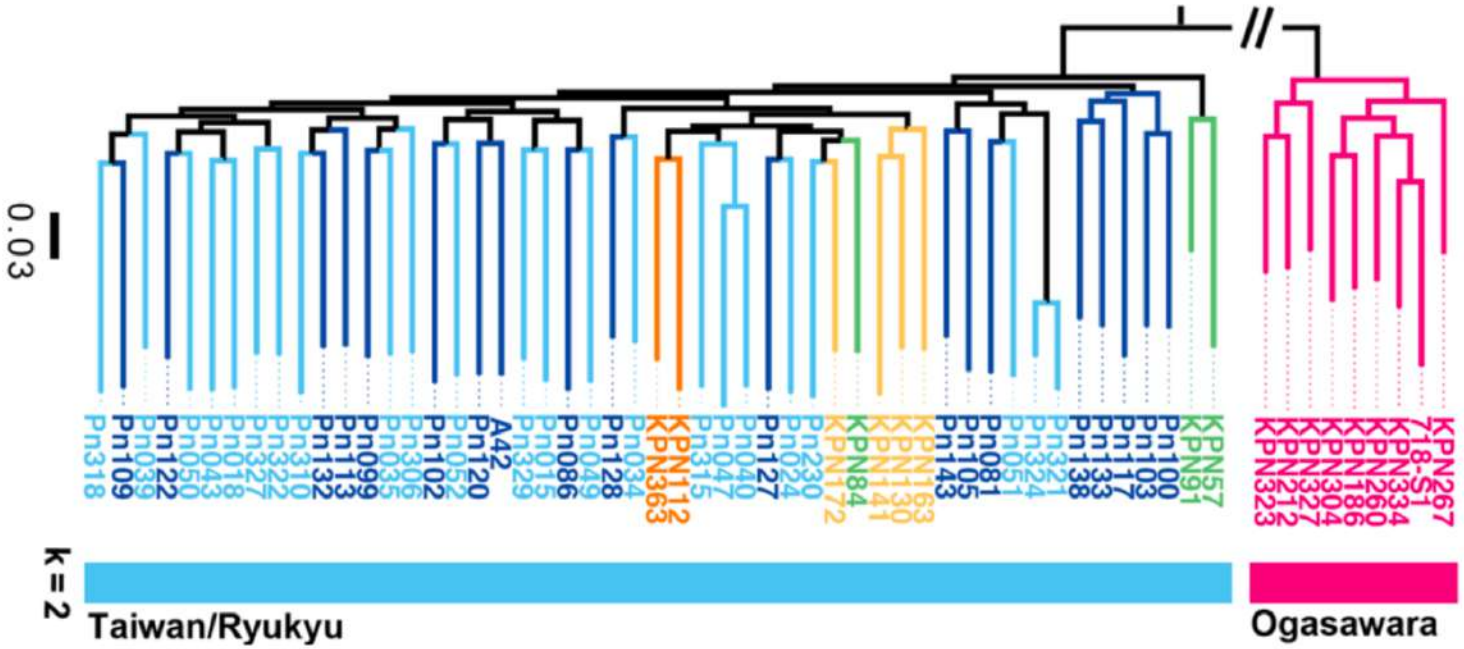
Chia-Lin Chung[1,2*] | Tracy J. Lee[3,4,5] | Mitsuteru Akiba[6] | Hsin-Han Lee[1] | Tzu-Hao Kuo[3] | Dang Liu[3,7] | Huei-Mien Ke[3] | Toshiro Yokoi[6] | Marylette B. Roa[3,8] | Mei-Yeh J. Lu[3] | Ya-Yun Chang[1] | Pao-Jen Ann[9] | Jyh-Nong Tsai[9] | Chien-Yu Chen[10] | Shean-Shong Tzean[1] | Yuko Ota[6,11] | Tsutomu Hattori[6] | Norio Sahashi[6] | Ruey-Fen Liou[1,2] | Taisei Kikuchi[12] | Isheng J. Tsai[3,4,5,7*]

2017 – *Phellinus* genome project



2017 – *Phellinus* genome project





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**

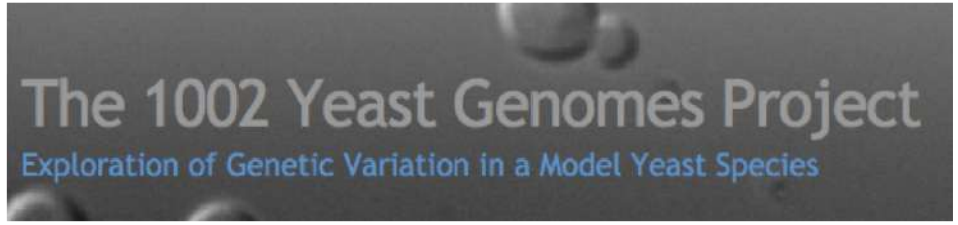
- More methods, analysis, tools, experiments…
  - Not always better

It is an exciting time to be in

Current and future

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

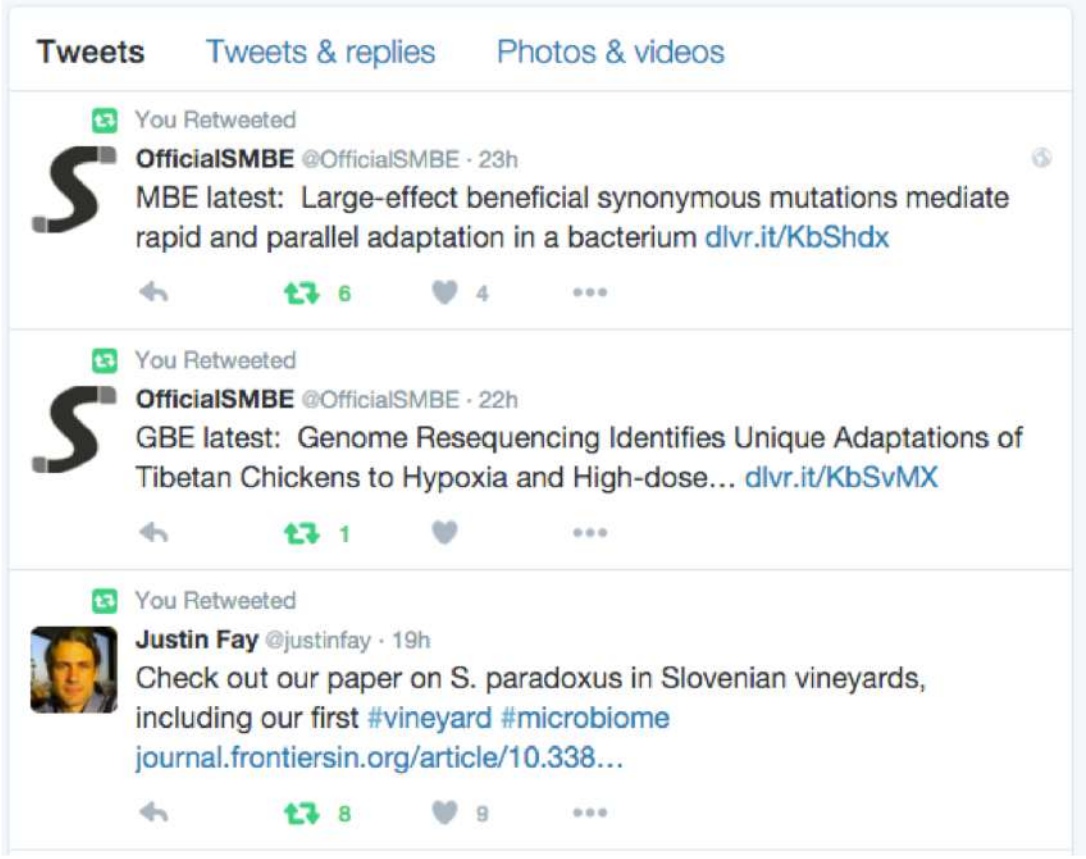
- Standard labs will be able to generate collections of themselves

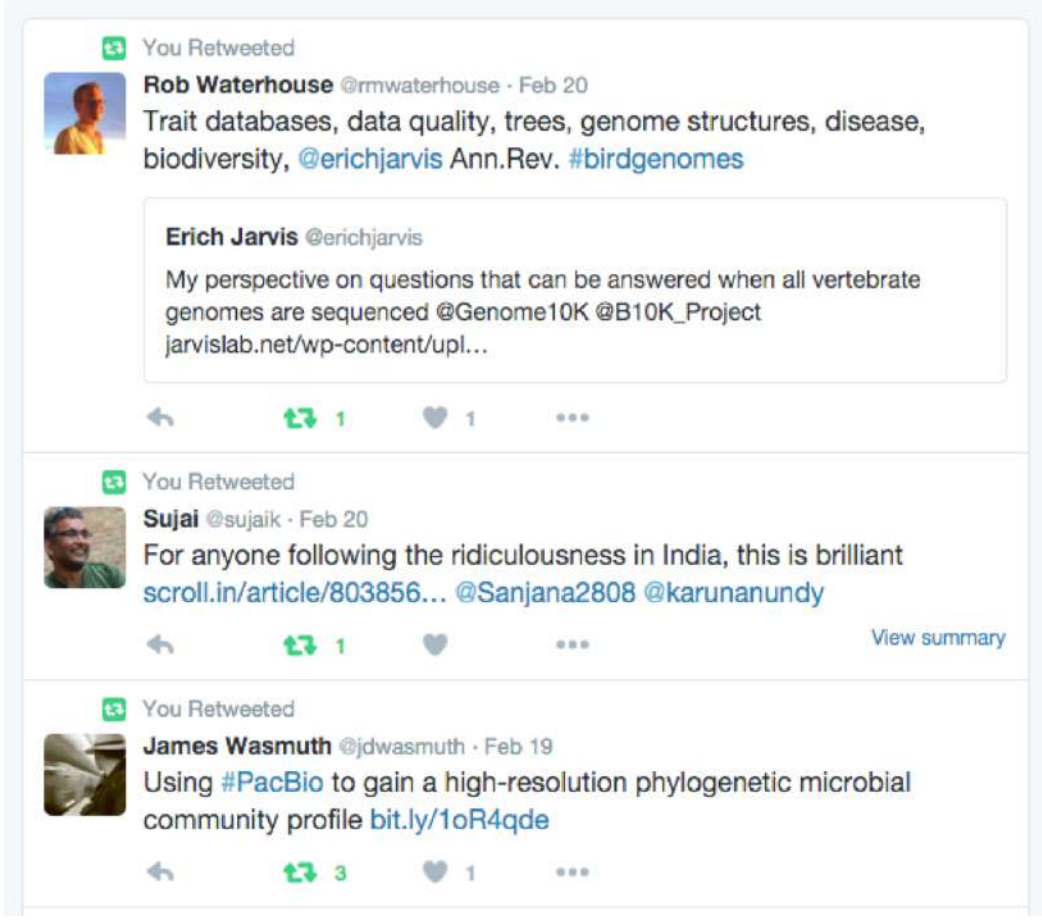


(3 labs)

There’s so much more…

- Read, read, read
- Twitter and blogs





Resources

Some very useful websites:

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

First written assignment

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

- Install R, Rstudio and
  - Give it a try
  - install package tidyverse