NGS Analytics

Isheng Jason Tsai

B2
Lecture 2

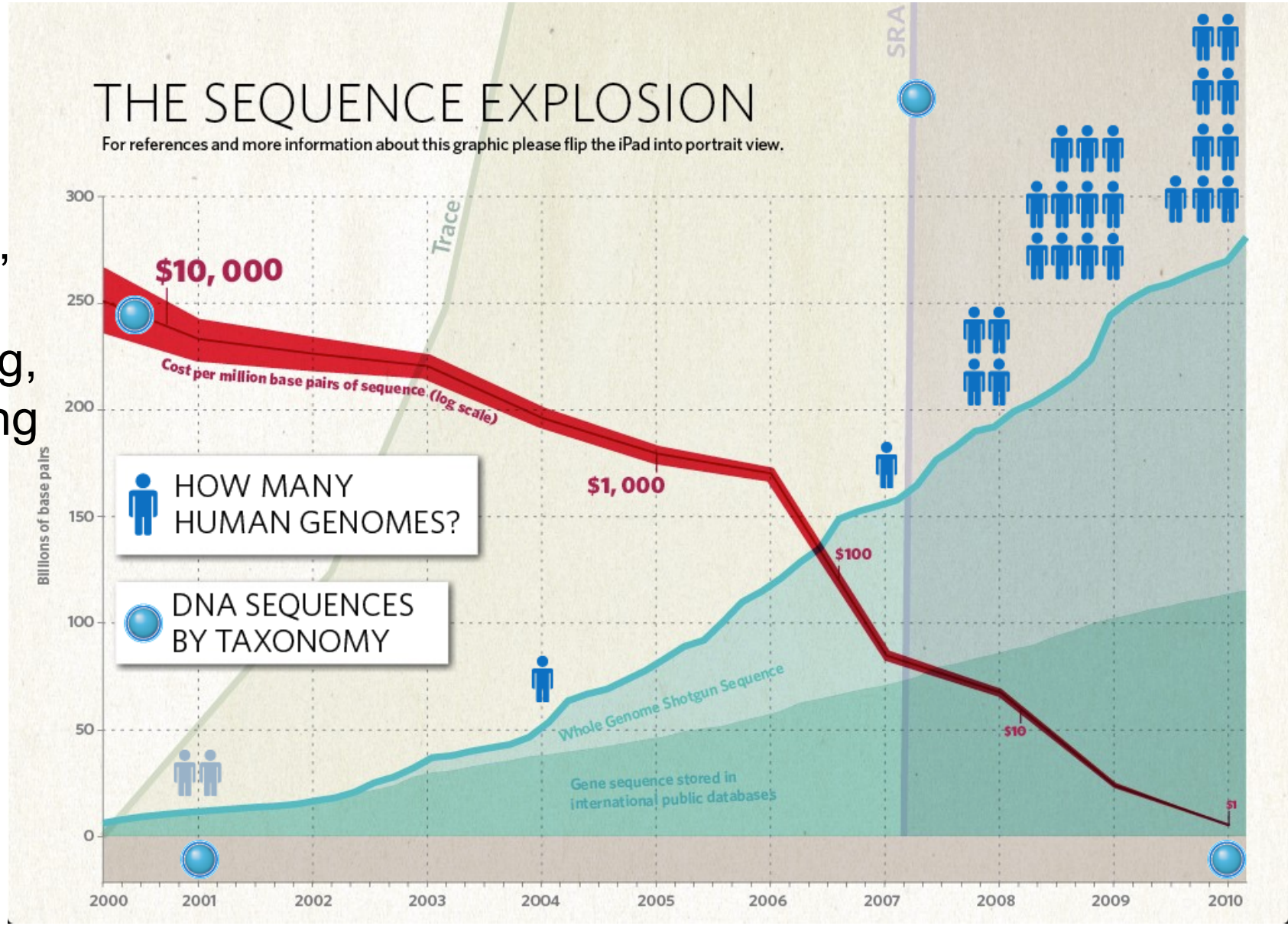
This lecture is called “Next-Generation Sequencing (NGS) Analytics”

Actually

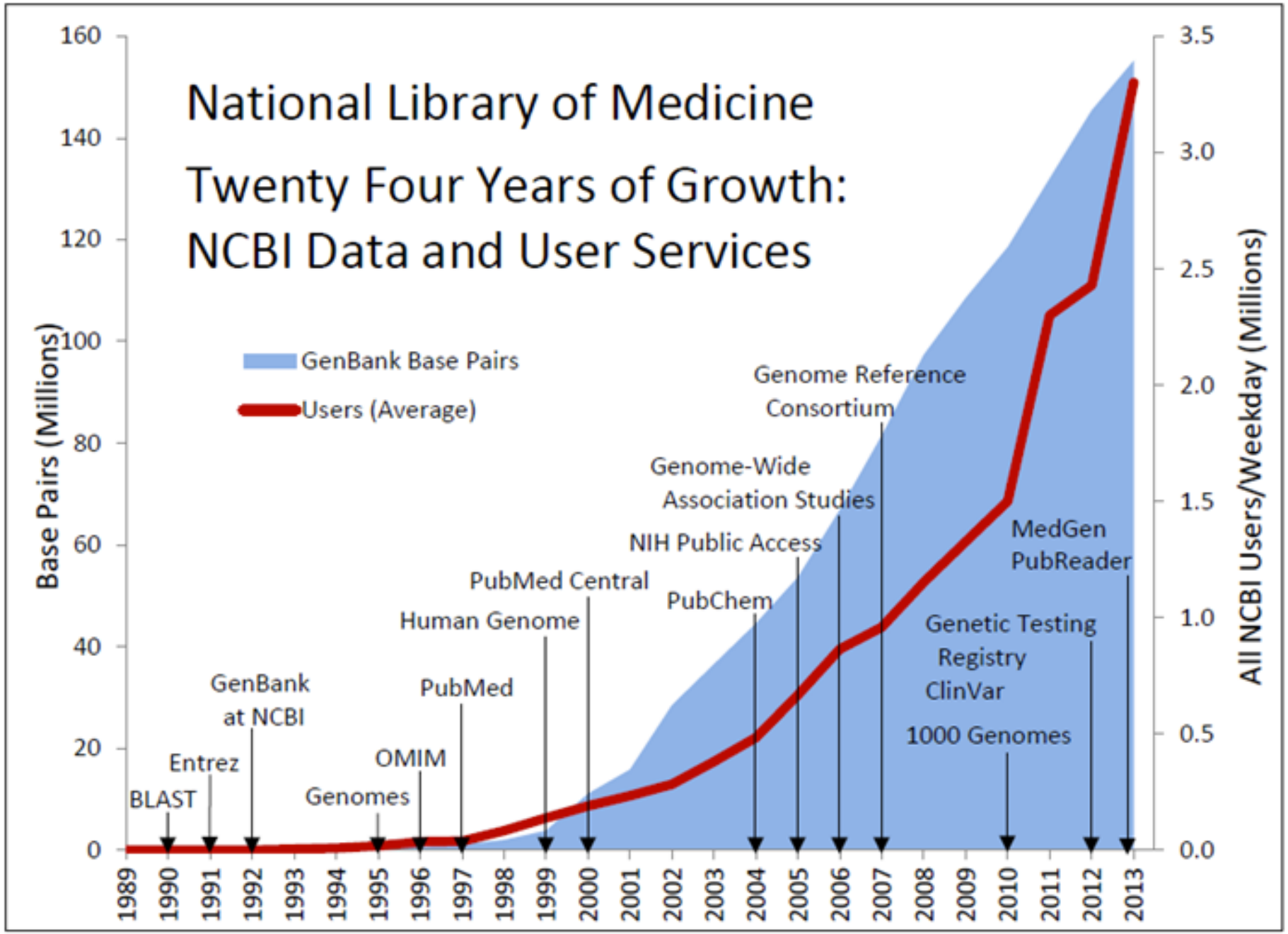
- Next Generation Sequencing is really “now” sequencing
- It won’t be so easy to tell you everything about NGS
  (it’s a bit like saying what can we do with PCR?)

What is NGS?

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



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

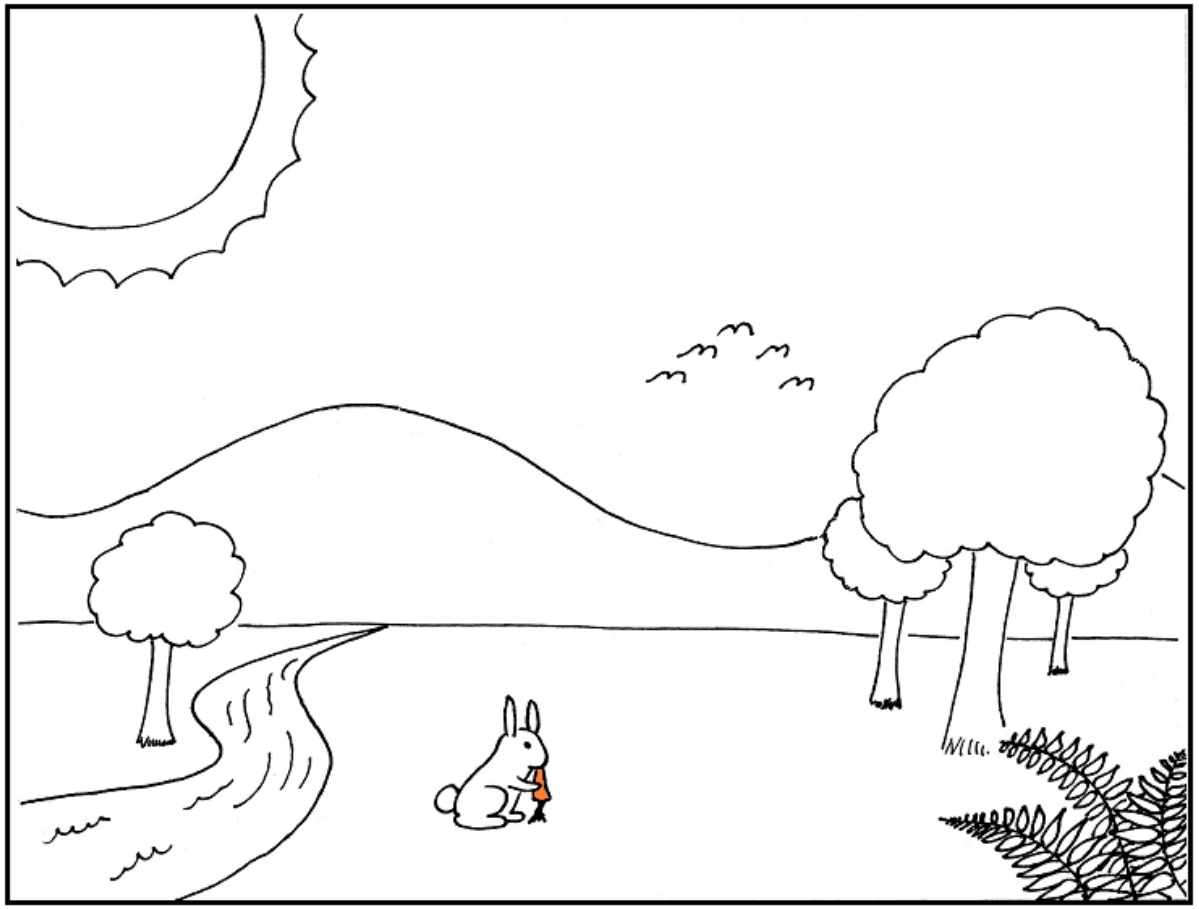


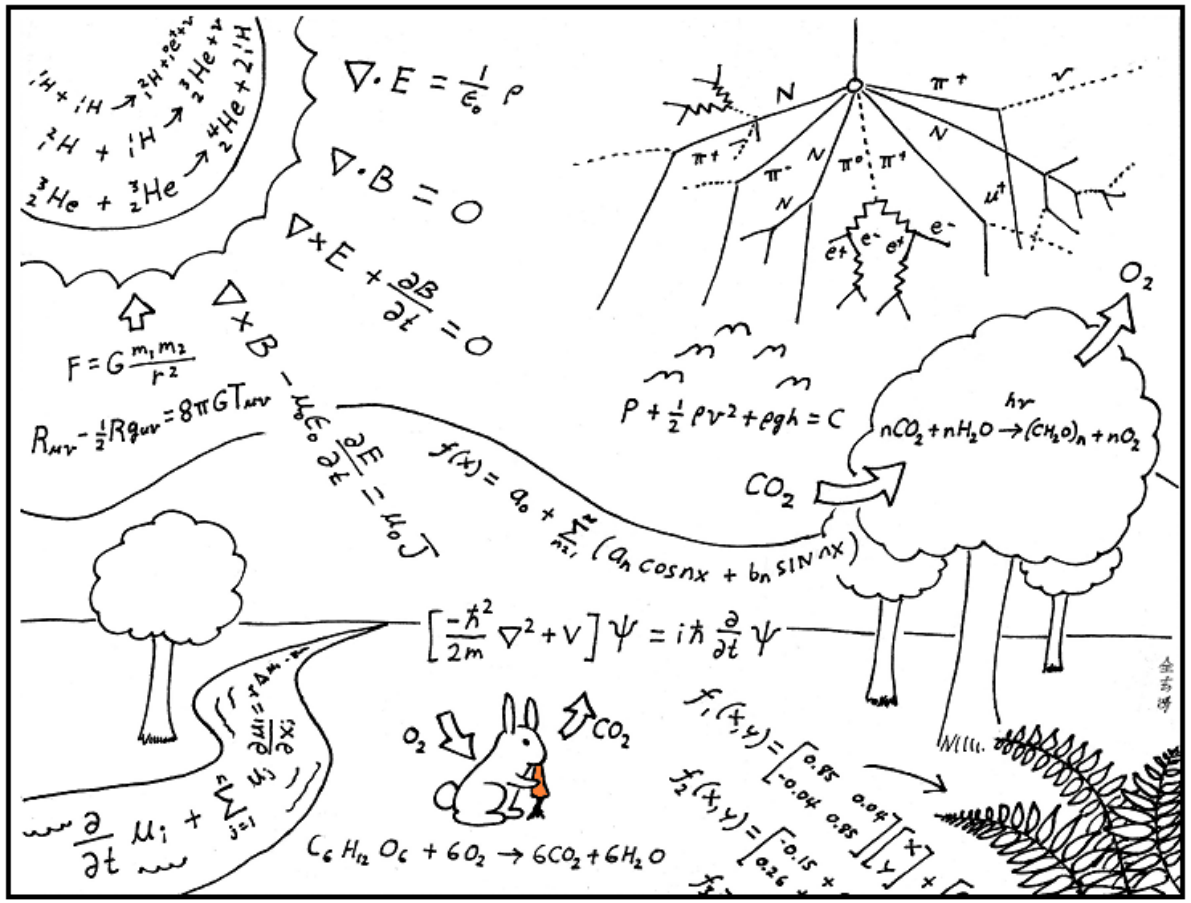
https://www.nlm.nih.gov/about/2015CJ.html

NGS = sequencing made cheaper

We start with a problem

This is how scientists see the world





How? Who? Where? What?

So why sequence?

- What's out there? Species diversity (metagenomics)
- Differences between species (comparative genomics)
- Variations between individuals (population genetics)
- Differential expression (transcriptomics)
- Annotation (transcriptomics)
- 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/

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

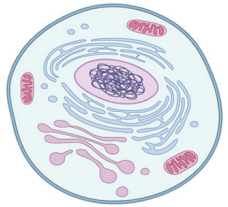
- Embrace it

- 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"

Some basics

A genome project

Wet lab work



Extract DNA
(or RNA)

Shear DNA
(or RNA)

Sequence

Bioinformatics

DNA or RNA Reads
50-500 bp

Mapping

Annotation
(Lecture 8)

Variant

ATCG
ATGG
ATCG

Data QC

Reads
50-500 bp

Assembly

Contigs
1kb – 100 kbp

Scaffolding

Scaffolds
Hopefully Mbp

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**

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

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

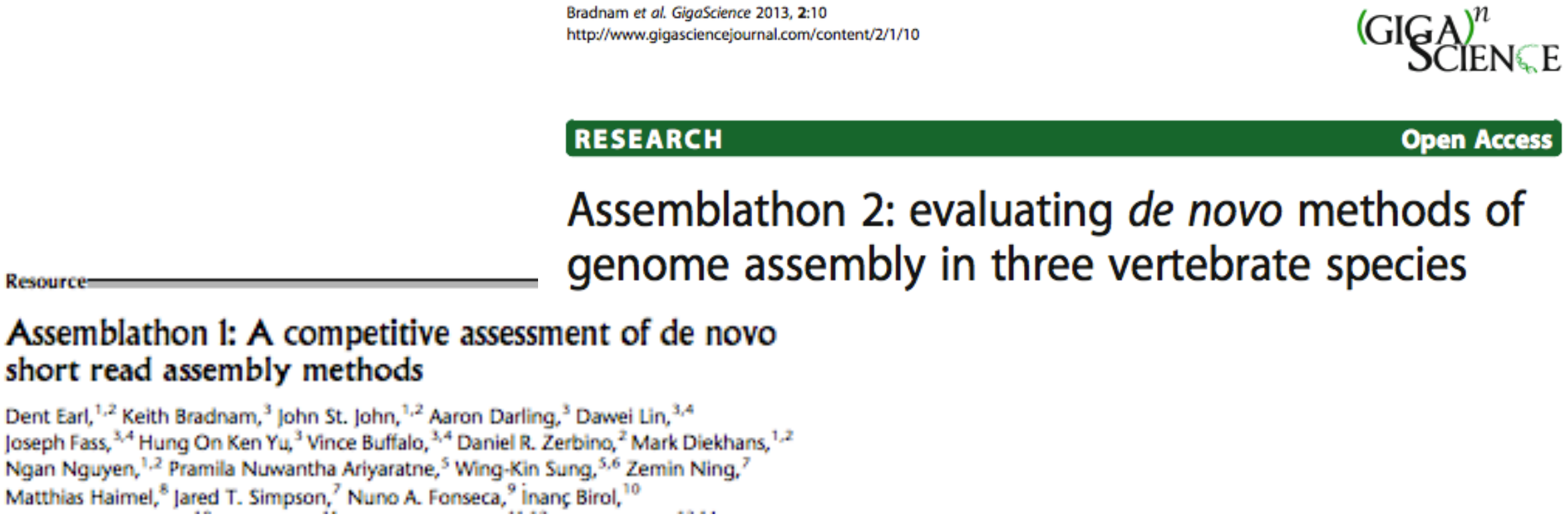
Problems:

- 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

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]

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

Run lots of programs

- Parse for information and interpret them
- Benchmark them

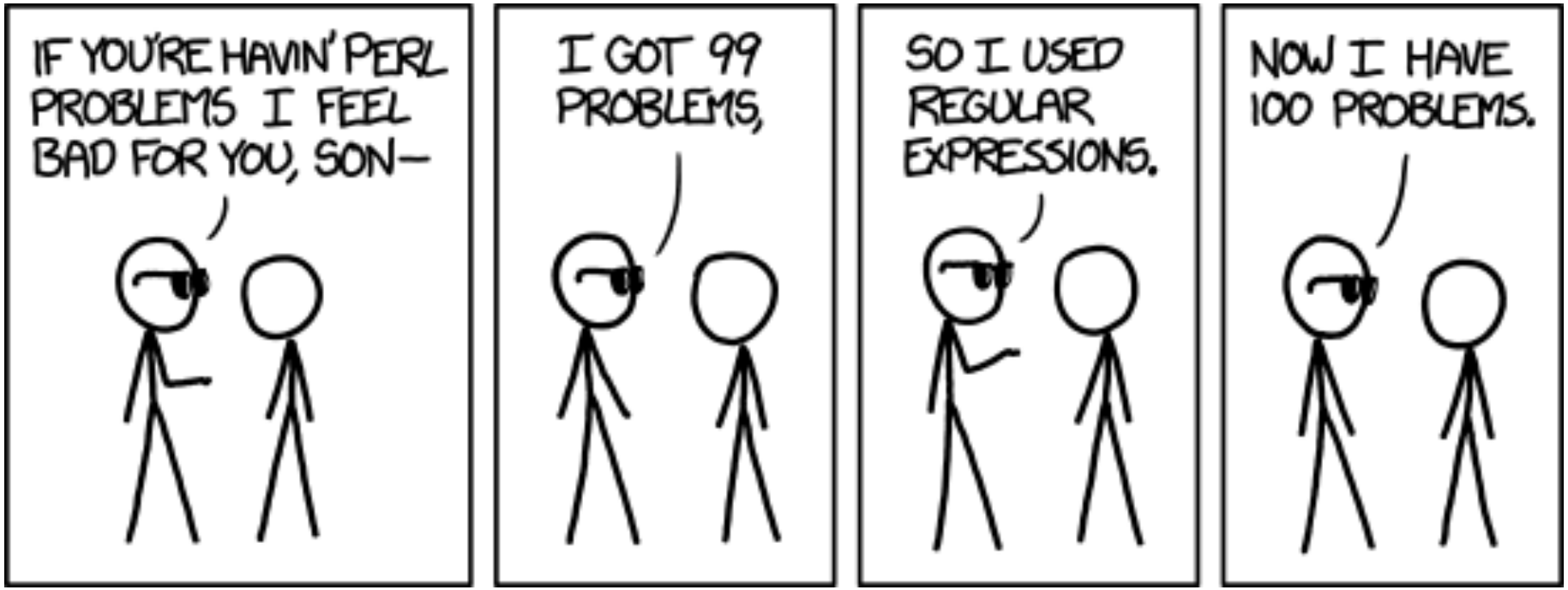
Write the programs yourself (visualisation, developing new algorithms)

Most of NGS data are **text** based

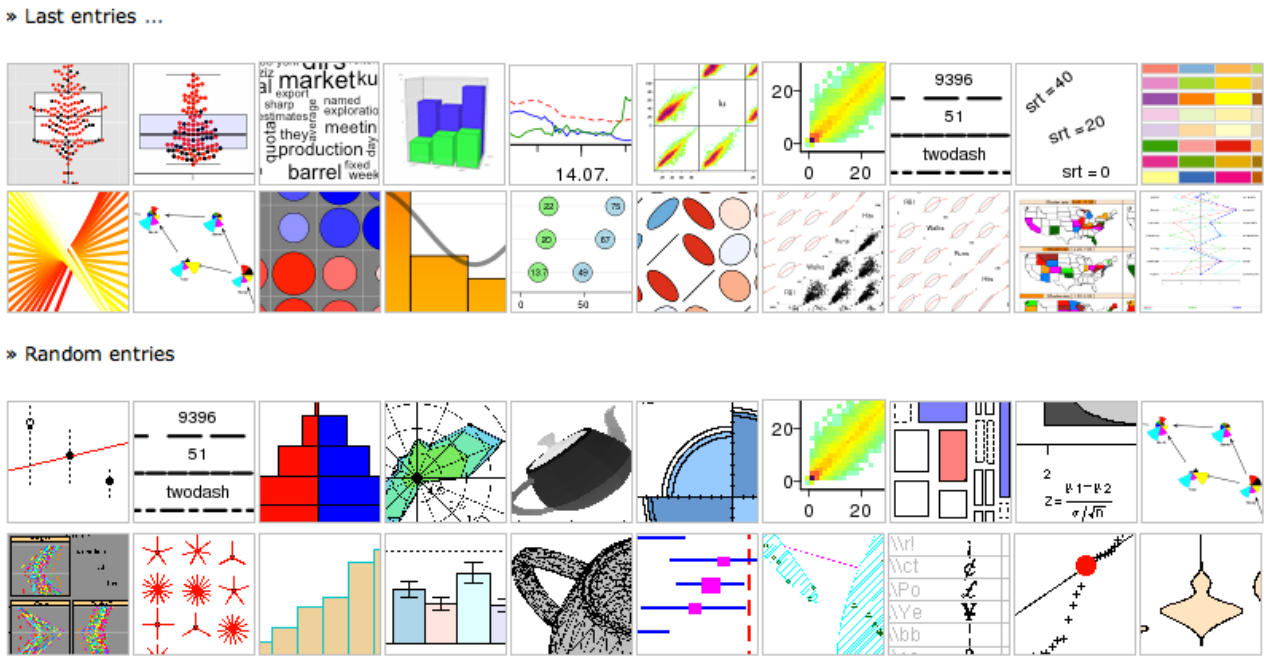
- All output = text files
- So **familiarity** to deal with large amount of data is expected
- That's why **Perl** was popular back in the early period

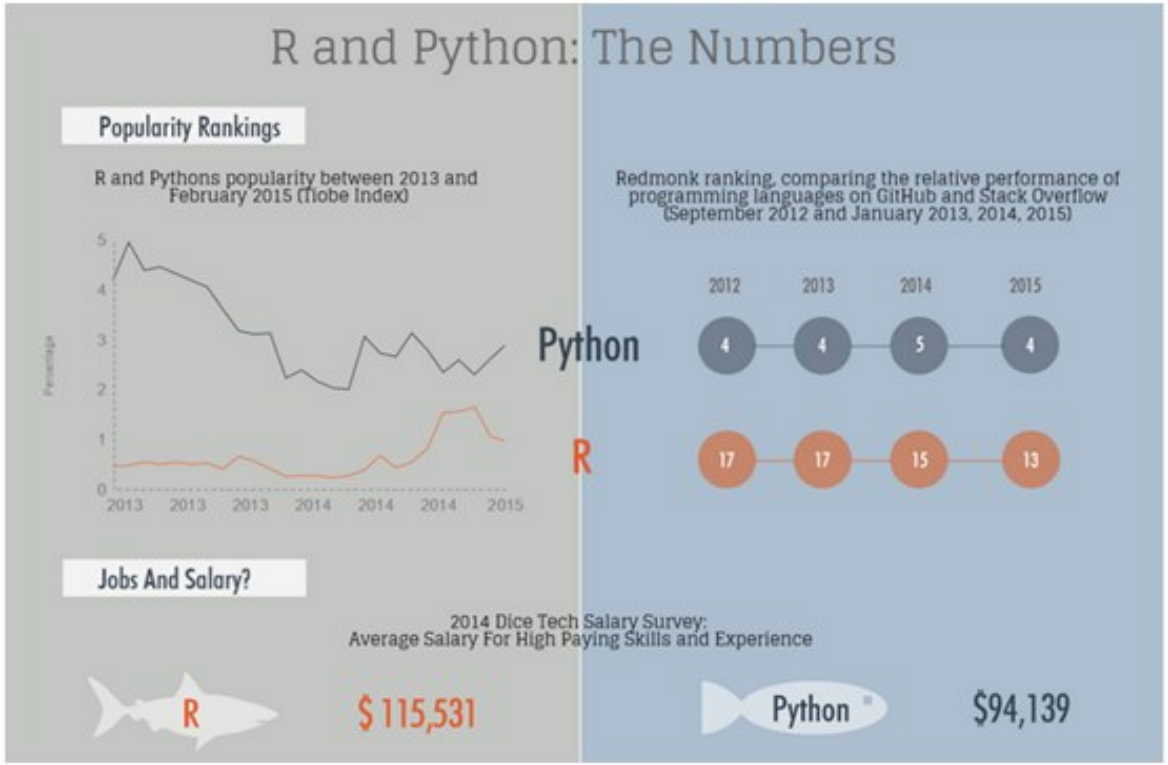
Problem with Perl

- Quick and ‘dirty’
- Data gets increasingly bigger and more complicated

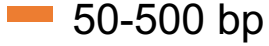


Python and R


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


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

More Definition

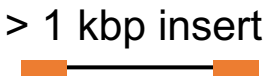


Read

A sequenced piece of DNA

Paired-end read

Sequencing both ends of a short DNA fragment



Mate-pair read

Sequencing both ends of a long DNA fragment

Insert size

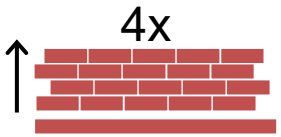
The length of the DNA fragment

Contig

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

Scaffold

Contigs separated by gaps of known length



Coverage

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

What is an alignment?

Align the following two sequences:

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

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

Scoring scheme is needed:

1 for match

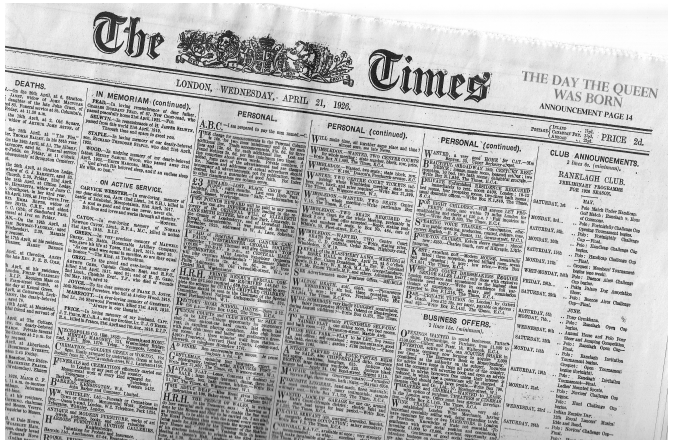
-1 for mismatch

-2 for gap

insertions / deletions (indels) mismatches

Which alignment is better?

Assembly

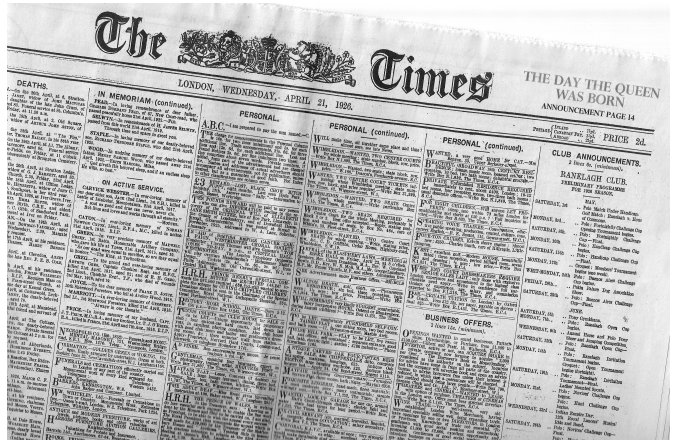


Genome
(3.000.000 letters)

Sequencing →

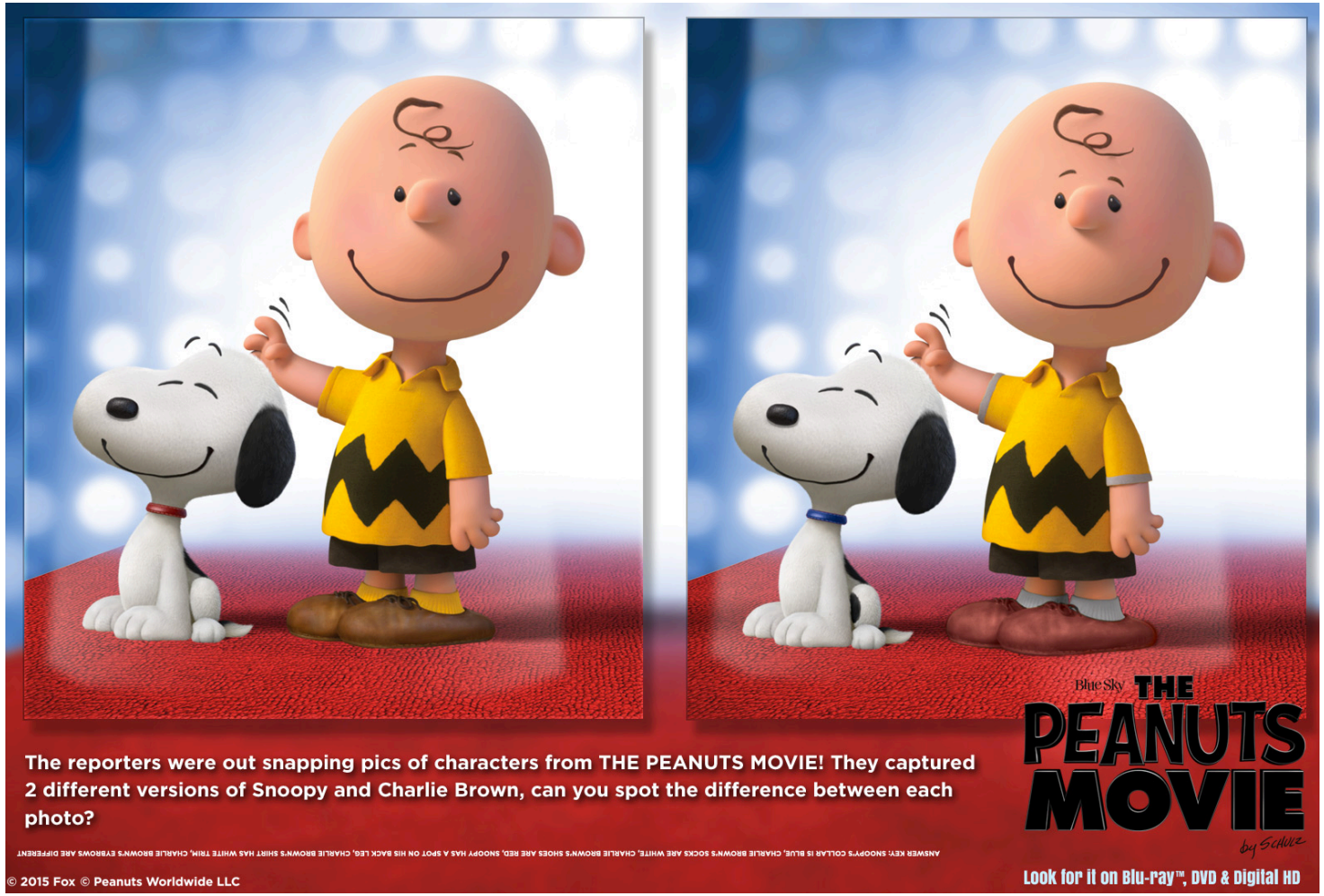
Reads
(50-500 letters each)

Assembly →



Genome
(3.000.000 letters)

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

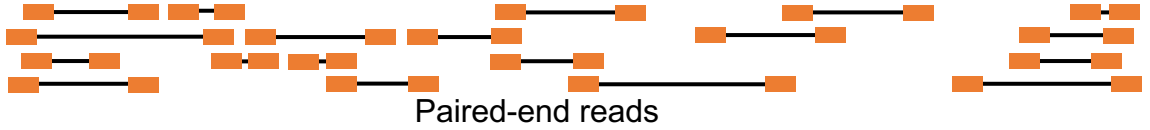


Assembly

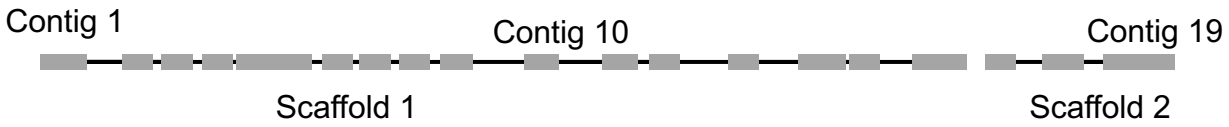
Genome

Fragment

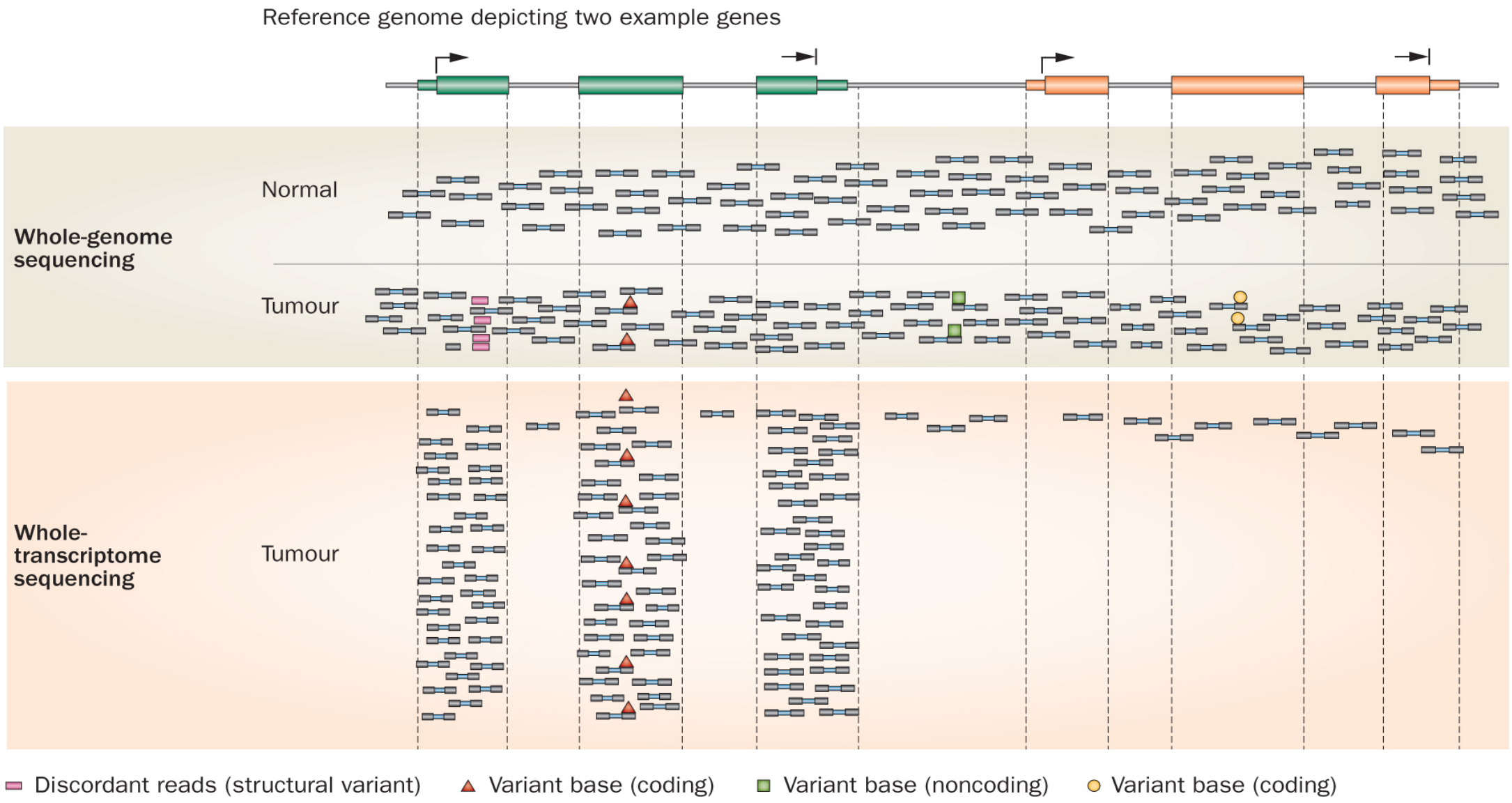
Sequence



Assemble



Mapping



doi:10.1038/nrgastro.2012.126

Read length matters in sequencing

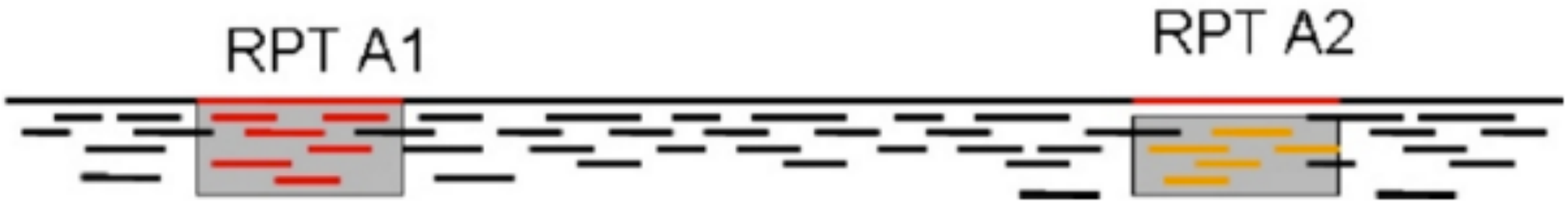


Figure 5. Two copies of a repeat along a genome. The reads colored in red and those colored in yellow appear identical to the assembly program.

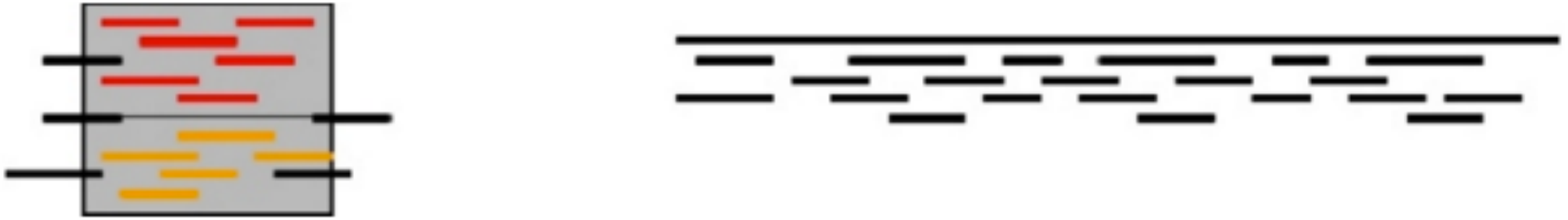
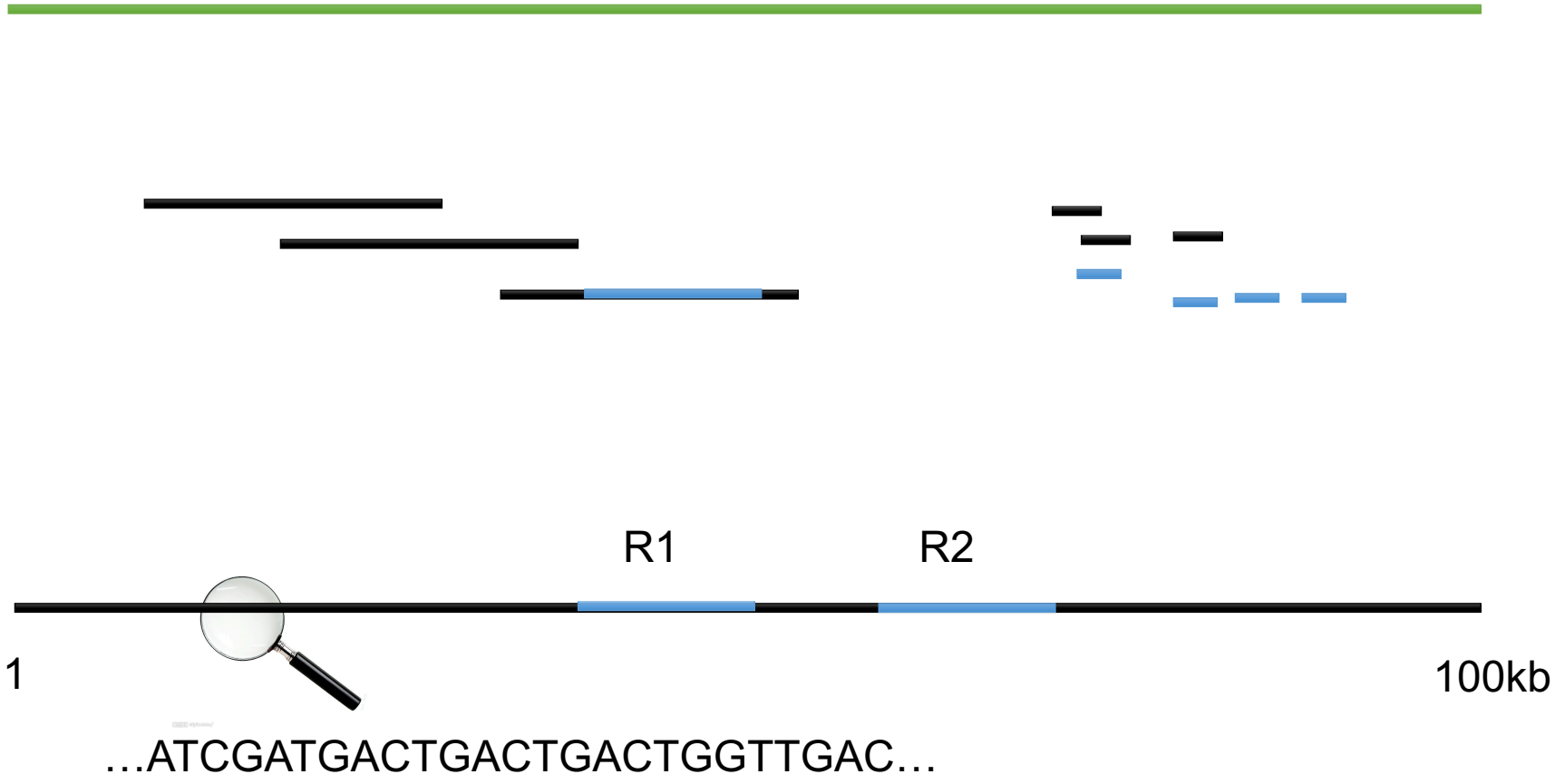


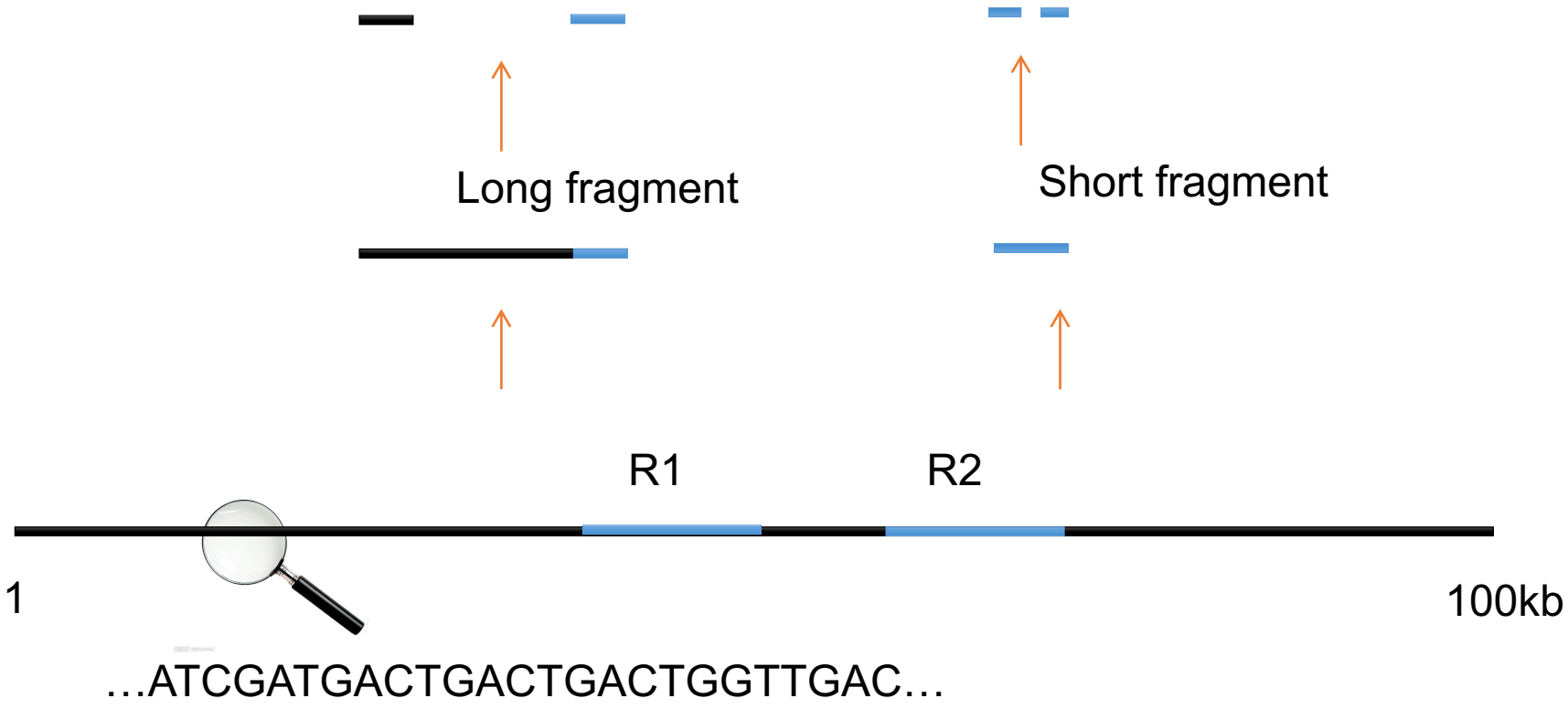
Figure 6. Genome mis-assembled due to a repeat. The assembly program incorrectly combined the reads from the two copies of the repeat leading to the creation of two separate contigs

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

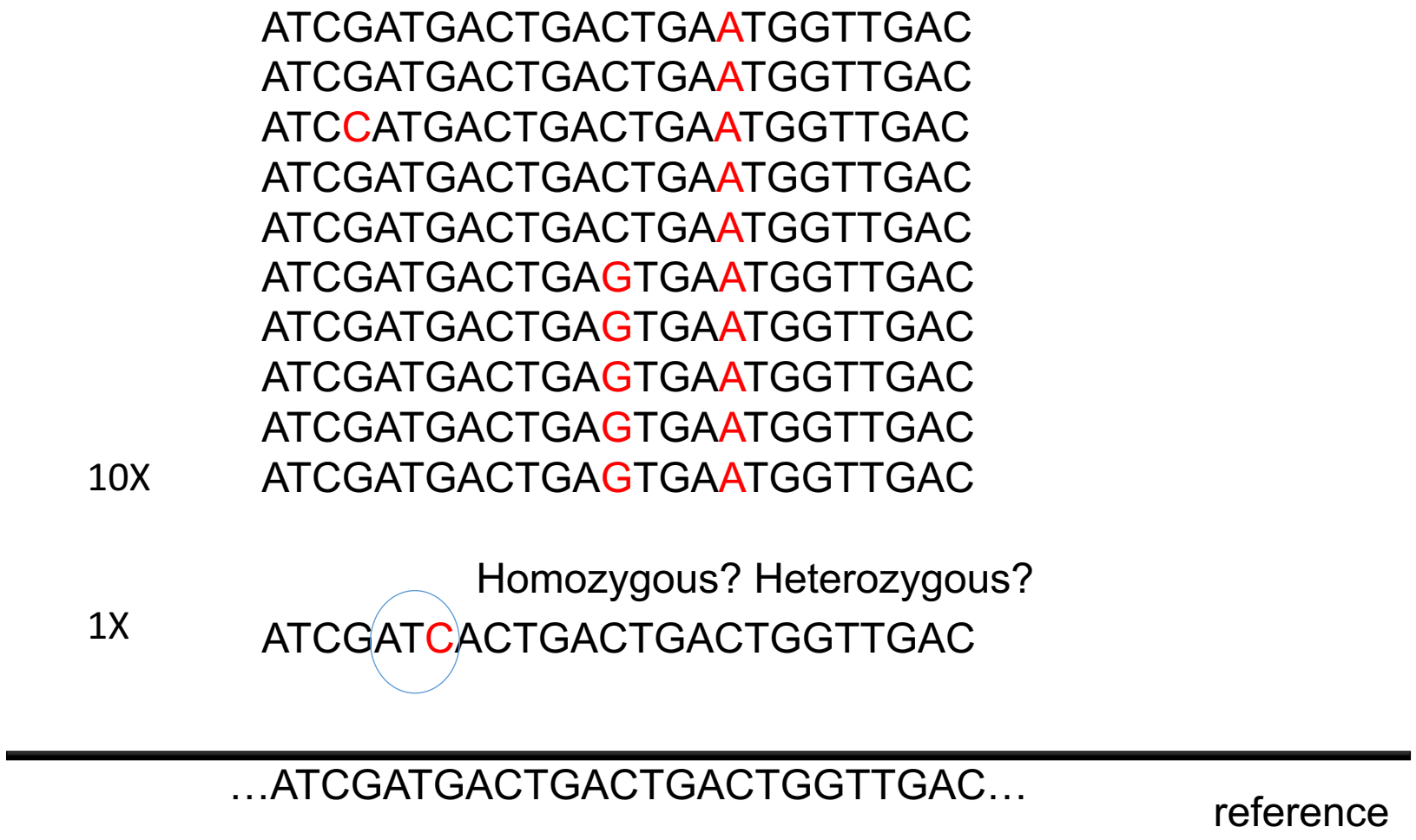
Read length matters in sequencing

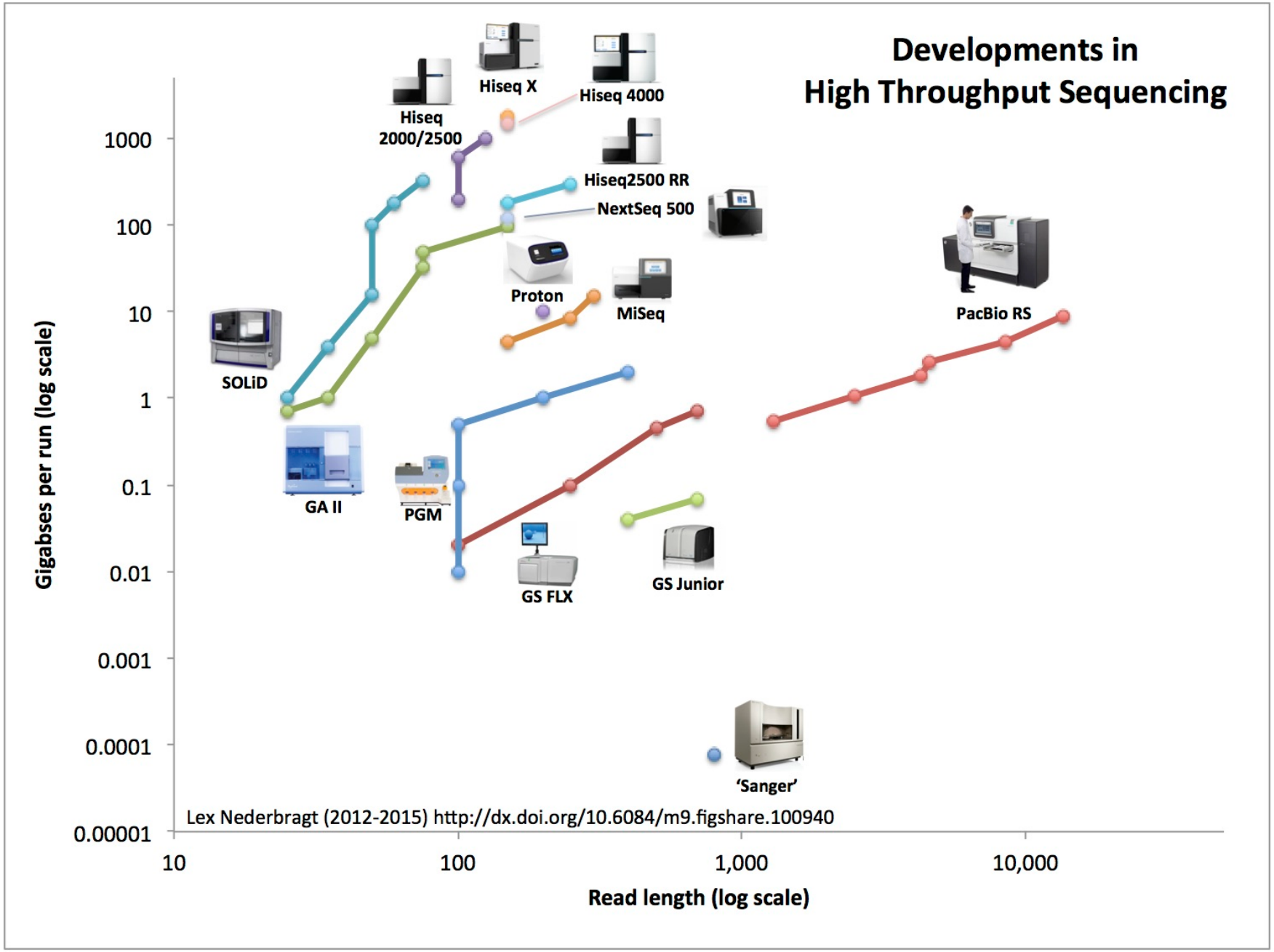


Paired end and insert size matter in sequencing



Depth matters in sequencing




Developments in
High Throughput Sequencing
Hiseq X
Hiseq 4000
Hiseq
2000/2500
Hiseq2500 RR
NextSeq 500
Proton
MiSeq
PacBio RS
SOLiD
GA II
PGM
GS FLX
GS Junior
'Sanger'
Gigabses per run (log scale)
1000
100
10
1
0.1
0.01
0.001
0.0001
0.00001
10
100
1,000
10,000
Read length (log scale)
Lex Nederbragt (2012-2015) http://dx.doi.org/10.6084/m9.figshare.100940

Different sequencing platforms

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

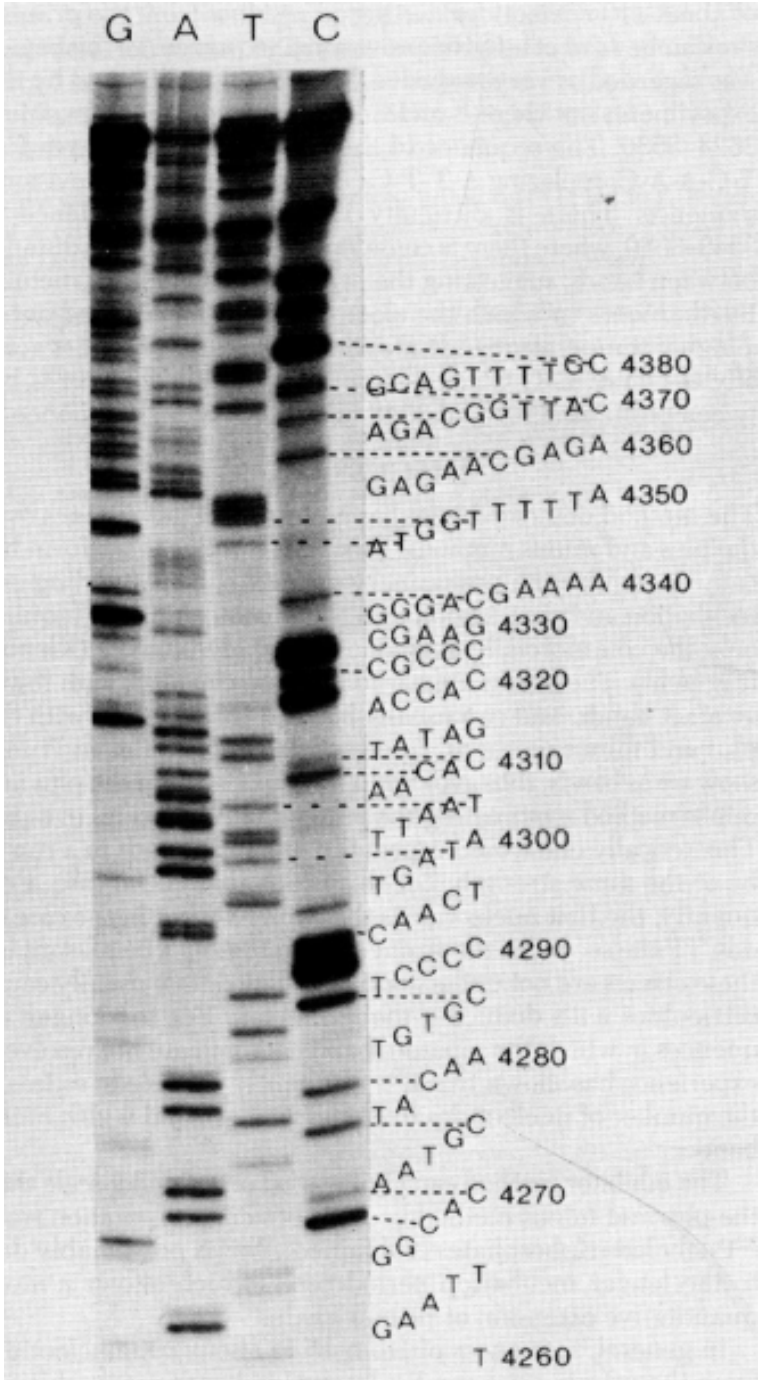
DNA sequencing with chain-terminating inhibitors

(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

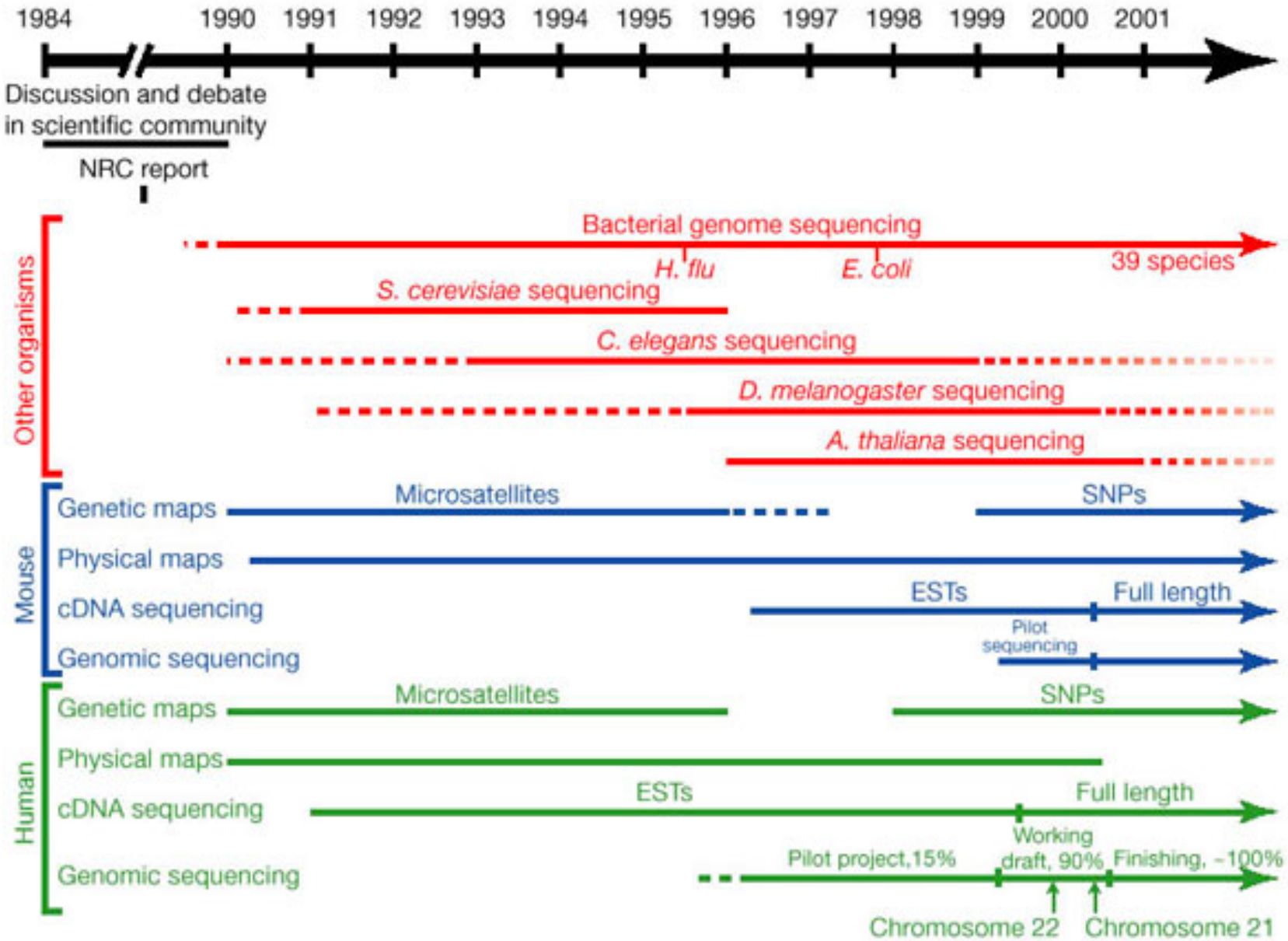
Contributed by F. Sanger, October 3, 1977



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

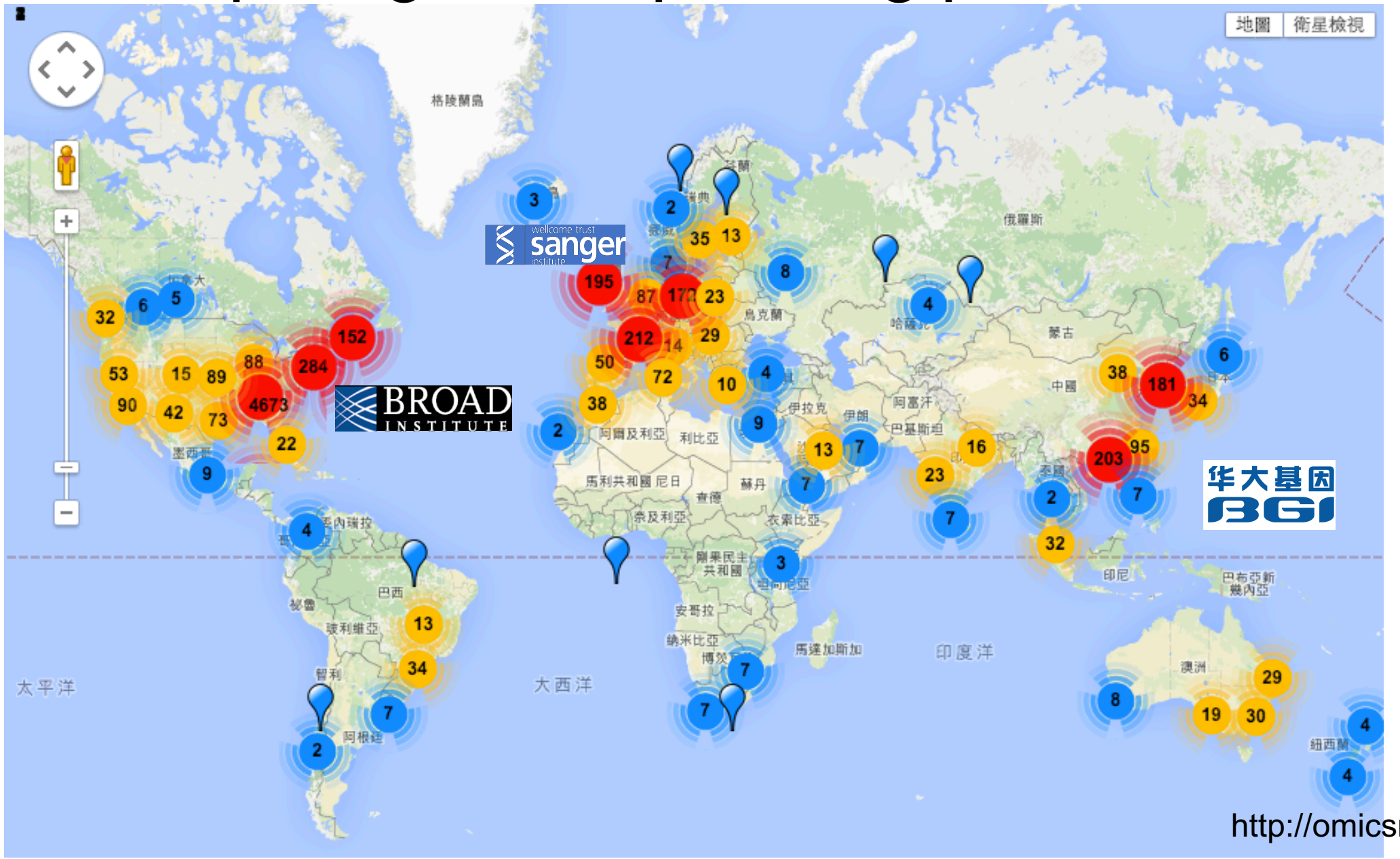
ABI 3730xi at TIGR

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



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

World competing for sequencing power

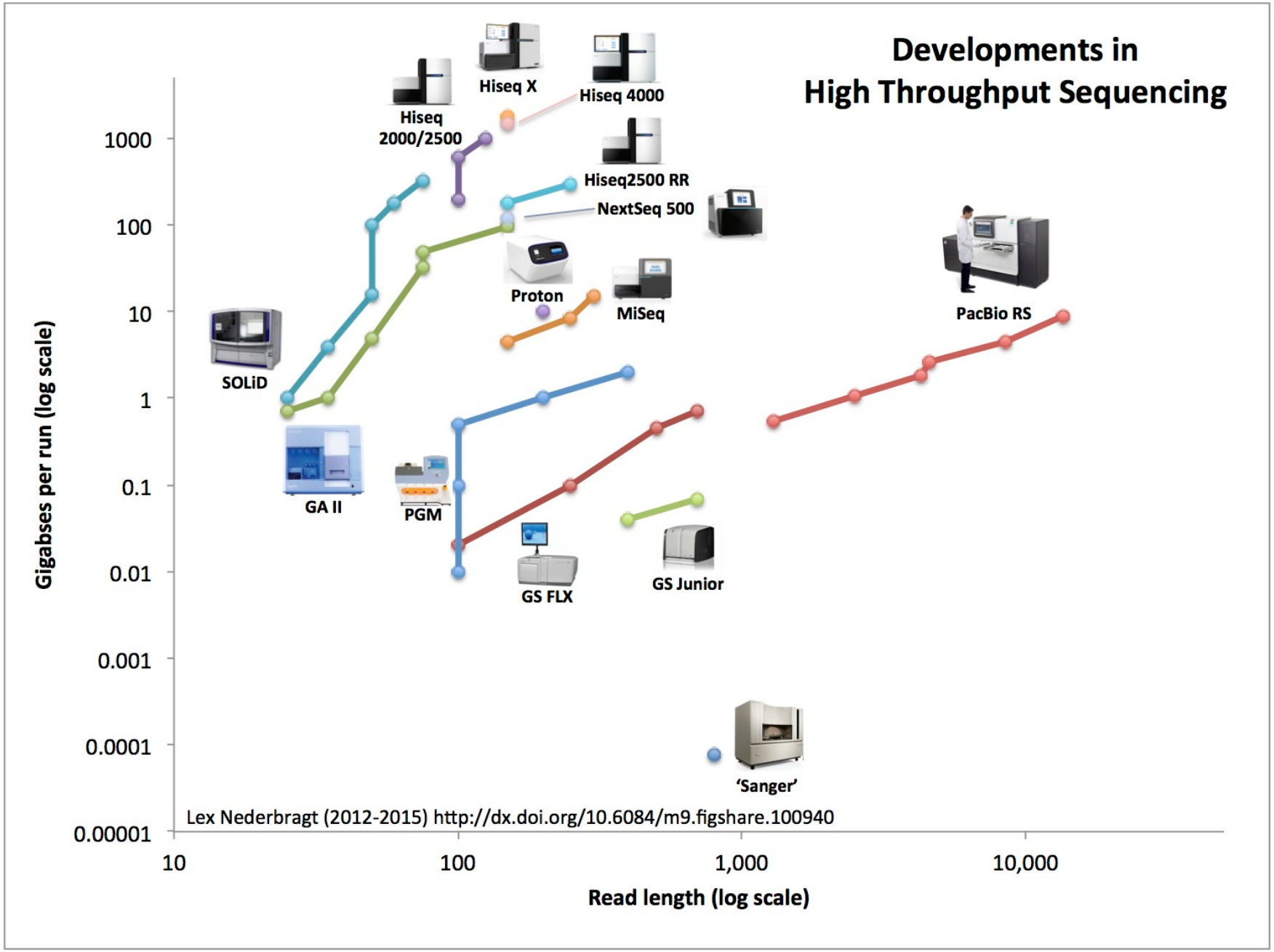


http://omicsmaps.com/

Sequencing Platforms

- Short reads
  1. ~~Genome Analyzer IIx (GAIIx)~~ – Illumina
  2. HiSeq2000, HiSeq2500, MiSeq – Illumina

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


Developments in
High Throughput Sequencing
Hiseq X
Hiseq 4000
Hiseq
2000/2500
Hiseq2500 RR
NextSeq 500
Proton
MiSeq
PacBio RS
SOLiD
GA II
PGM
GS FLX
GS Junior
'Sanger'
Gigabses per run (log scale)
1000
100
10
1
0.1
0.01
0.001
0.0001
0.00001
10
100
1,000
10,000
Read length (log scale)
Lex Nederbragt (2012-2015) http://dx.doi.org/10.6084/m9.figshare.100940

Platform	GS jnr	FLX plus	MiSeq	Next Seq 500	HiSeq 2500 RR	Hiseq 2500 V3	HiSeq 2500 V4	HiSeq 4000	HiSeq X	SOLiD 4	5000 XL	318 HiQ 520	Ion 530	Ion Proton P1	PGM HiQ 540	RS P6-C4	Sequel	Mini ION	Prome thION	QiaGen Gene Readr	BGI SEQ 500	#
Reads: (M)	0.1	1.25	25	400	600	3000	4000	5000	6000	1400	--	5	15-20	165	60-80	5.5	38.5	0.05	--	400	--	--
Read length: (paired-end*)	400	700	300*	150*	100*	100*	125*	150*	150*	50	75	200 400	200 400	200	220	15K	12K	10K	10K	--	--	--
Run time: (d)	0.4	0.9	2	1.2	1.125	11	6	3.5	3	12	7	0.37	--	--	--	4.3	4.3	2	--	--	1	--
Yield: (Gb)	0.035	0.7	15	120	120	600	1000	1500	1800	100	180	1.2-2	6-8	10	10-15	12	84	0.5	600	80	200	--
Rate: (Gb/d)	0.2	0.75	7.5	100	106.6	55	166	400	600	8.3	30	--	--	--	--	2.8	19.5	0.25	--	--	--	--
Reagents: ($K)	1.1	6.2	1	4	6.145	23.47	29.9	29.9	12.75	9	10.5	0.6	--	1	1.2	2.4	11.2	1	--	0.5	--	--
per-Gb: ($)	31K	8K	93	33.3	51.2	39.1	29.9	20	7	90	58.33	--	--	100	--	200	80	2000	20	--	--	--
Machine: ($)	110K	500K	99K	250K	740K	690K	690K	900K	1M	500K	595K	50K	65K	243K	242K	695K	350K	1000	30K	--	--	--

#Page maintained by http://www.vilellagenomics.com #Editable version: tinyurl.com/ngsspecsshared

#curl "https://docs.google.com/spreadsheets/d/1GMMfhyLK0-q8Xklo3YxlWaZA5vVMuhU1kg41g4xLkXc/export?gid=4&format=csv" | grep -v '^#' | column -t -s\, | less -S

Albert Vilella @AlbertVilella · 18h

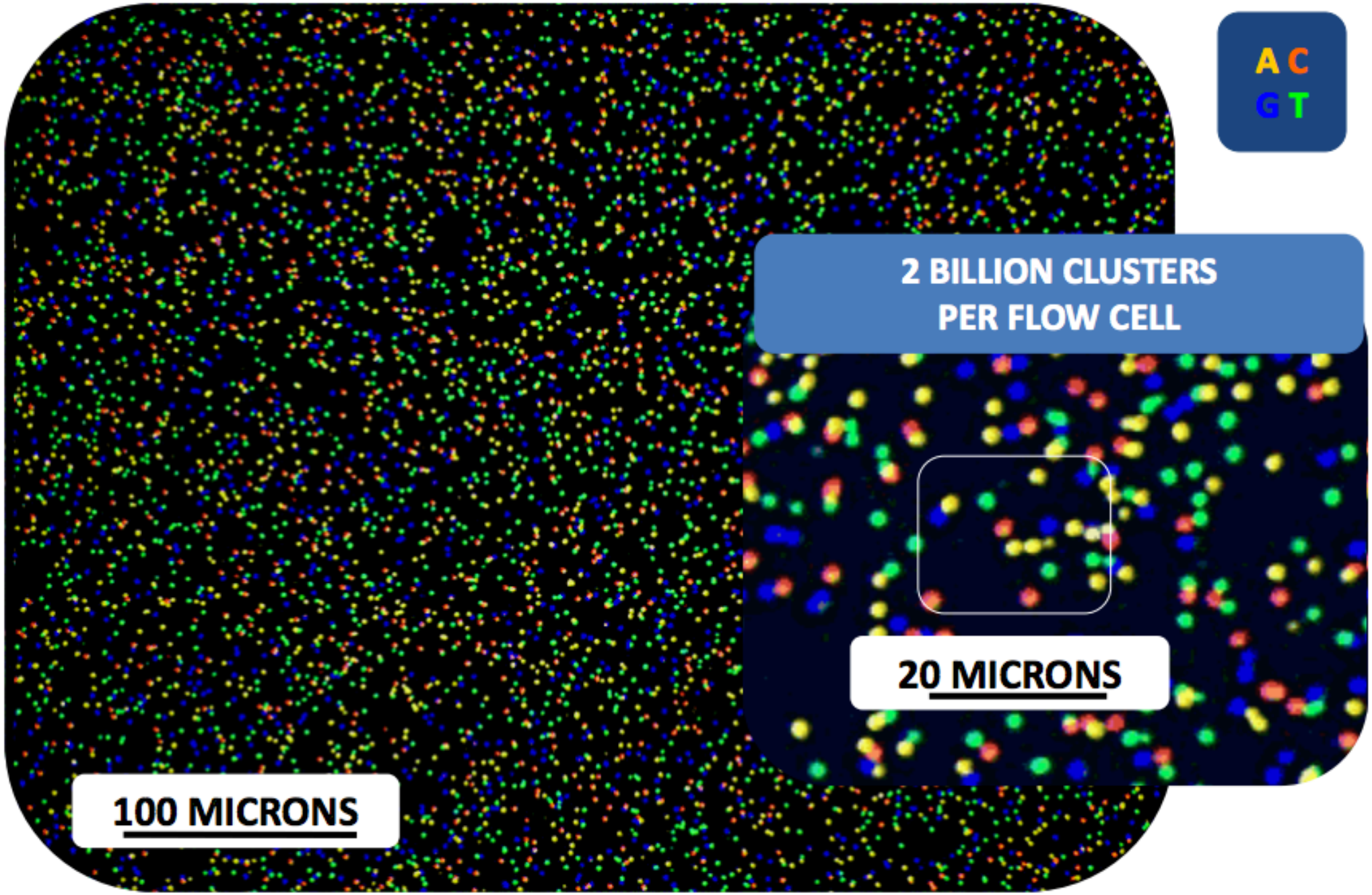
Updated NGS specs @BGI_Genomics @PacBio @illumina @thermofisher @nanopore @QIAGENscience tinyurl.com/ngsspecs

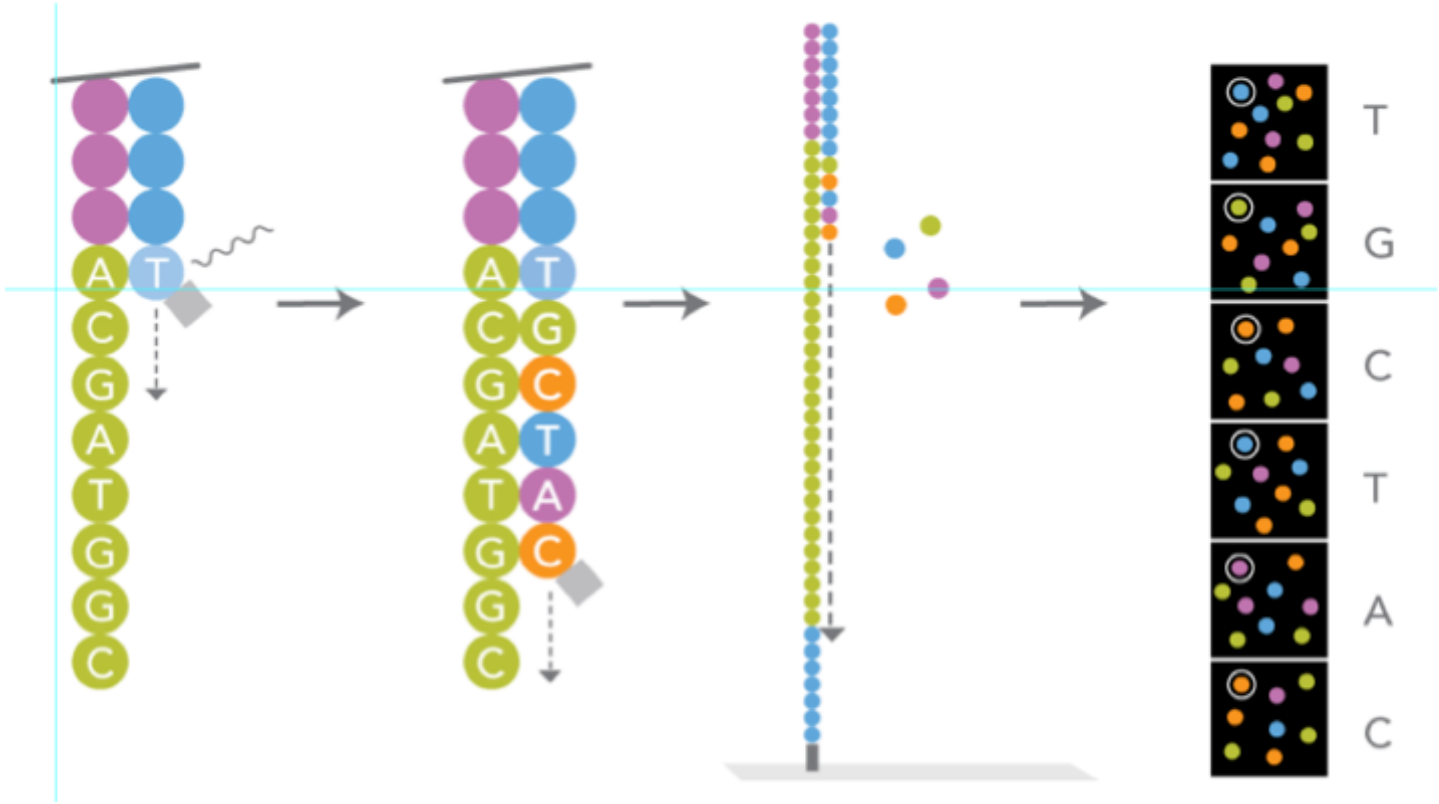
1 5

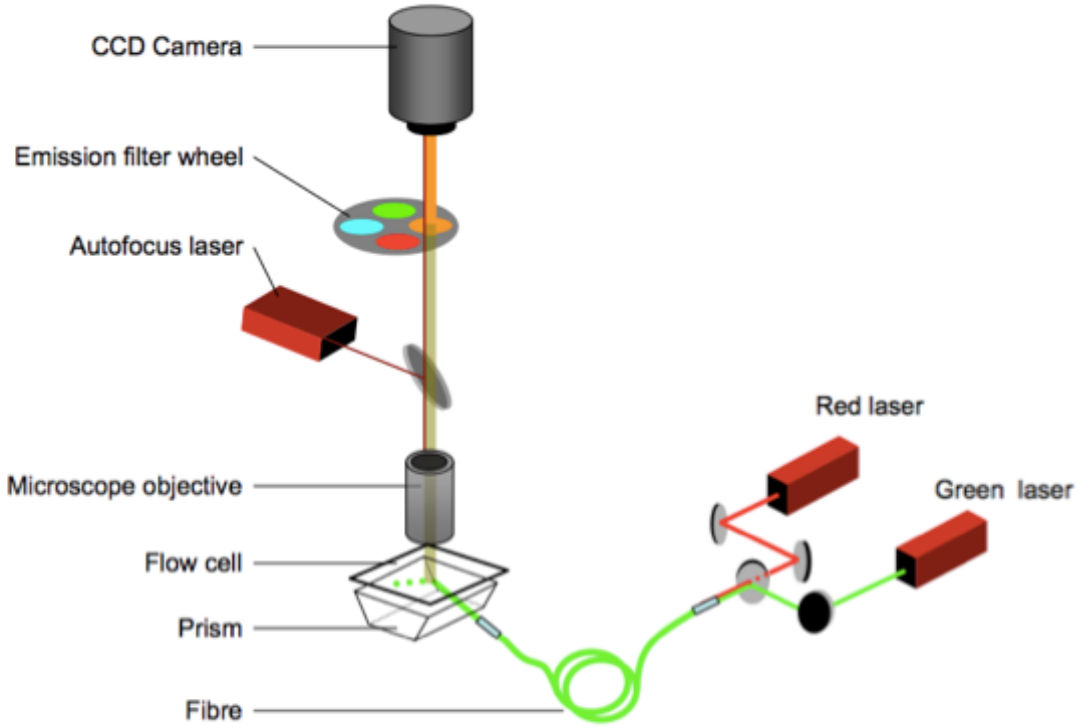
Illumina HiSeq

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

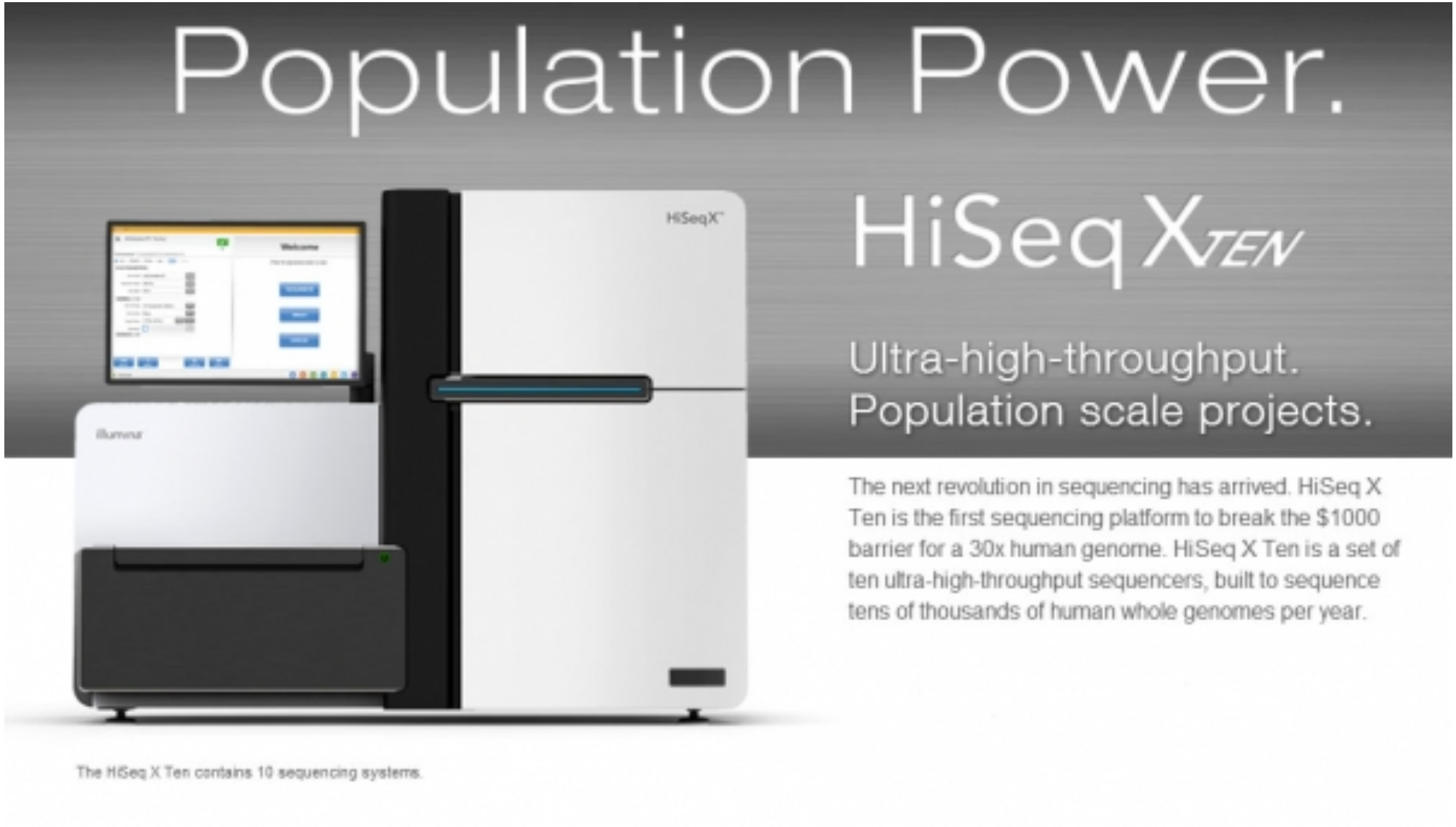
HiSeq 2000

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

HiSeq X Ten!



http://www.illumina.com/systems/hiseq-x-sequencing-system.ilmn

Illumina platform comparison

	Run type	Max read length (per end)	Max read length (combined	run time (hr)	runs per year	costs per M reads	costs per GB
MiSeq	Nano	300	300	21	417		
	Micro	300	300	22	398		
	V3 150 cycle	300	300	20	438	£ 22.32	£ 74.40
	V3 600 cycle	600	600	55	159	£ 39.00	£ 65.00
NextSeq	500 Mid-output	300	300	26	337	£ 8.00	£ 26.68
	500 High-output	300	300	29	302	£ 6.67	£ 22.22
HiSeq	2000 High-output	100	200	264	33	£ 6.12	£ 30.60
	2500 Rapid	250	500	60	146	£ 7.05	£ 14.09
	2500 High-output	125	250	144	61	£ 5.05	£ 20.21
	4000 High-output	150	300	120	73	£ 4.85	£ 16.17
	HiSeq X	150	300	72	122	£ 2.00	£ 6.67
Proton	PI	200	200	16	548	£ -	£ -
Proton	PII	100	200	16	548	£ -	£ -

Third generation sequencing

PacBio (Pacific Biosciences)

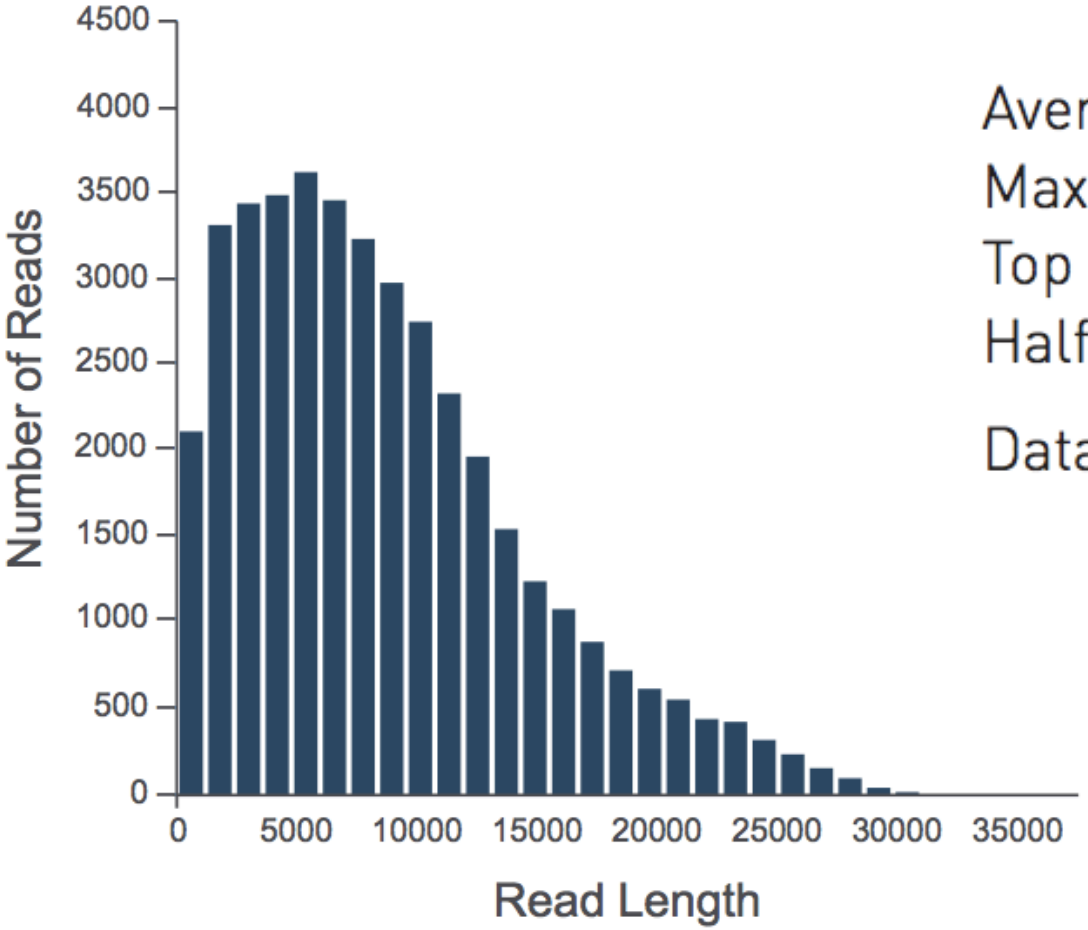
RSII

Sequel

Single molecule sequencing

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

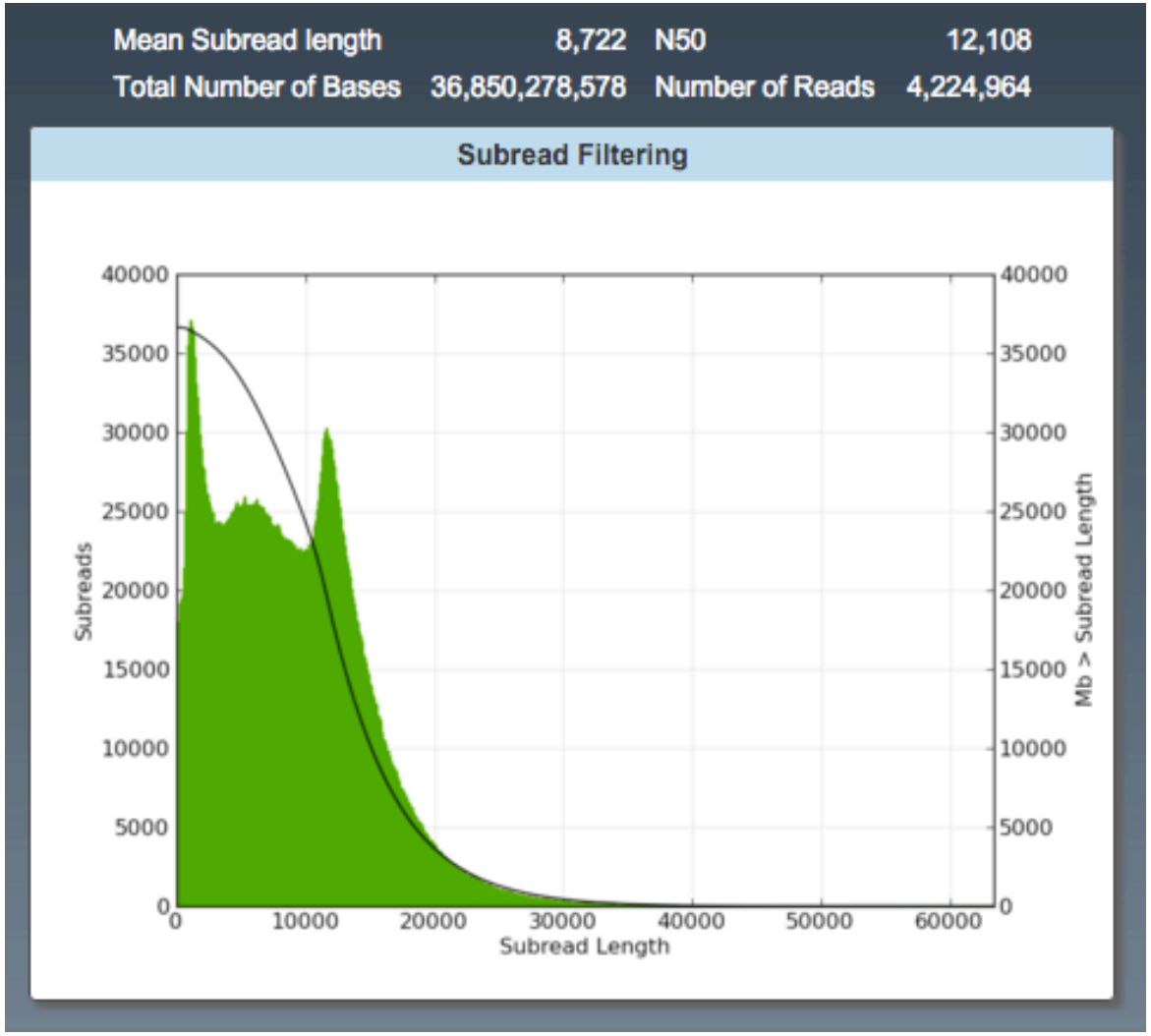
PacBio (Pacific Biosciences)

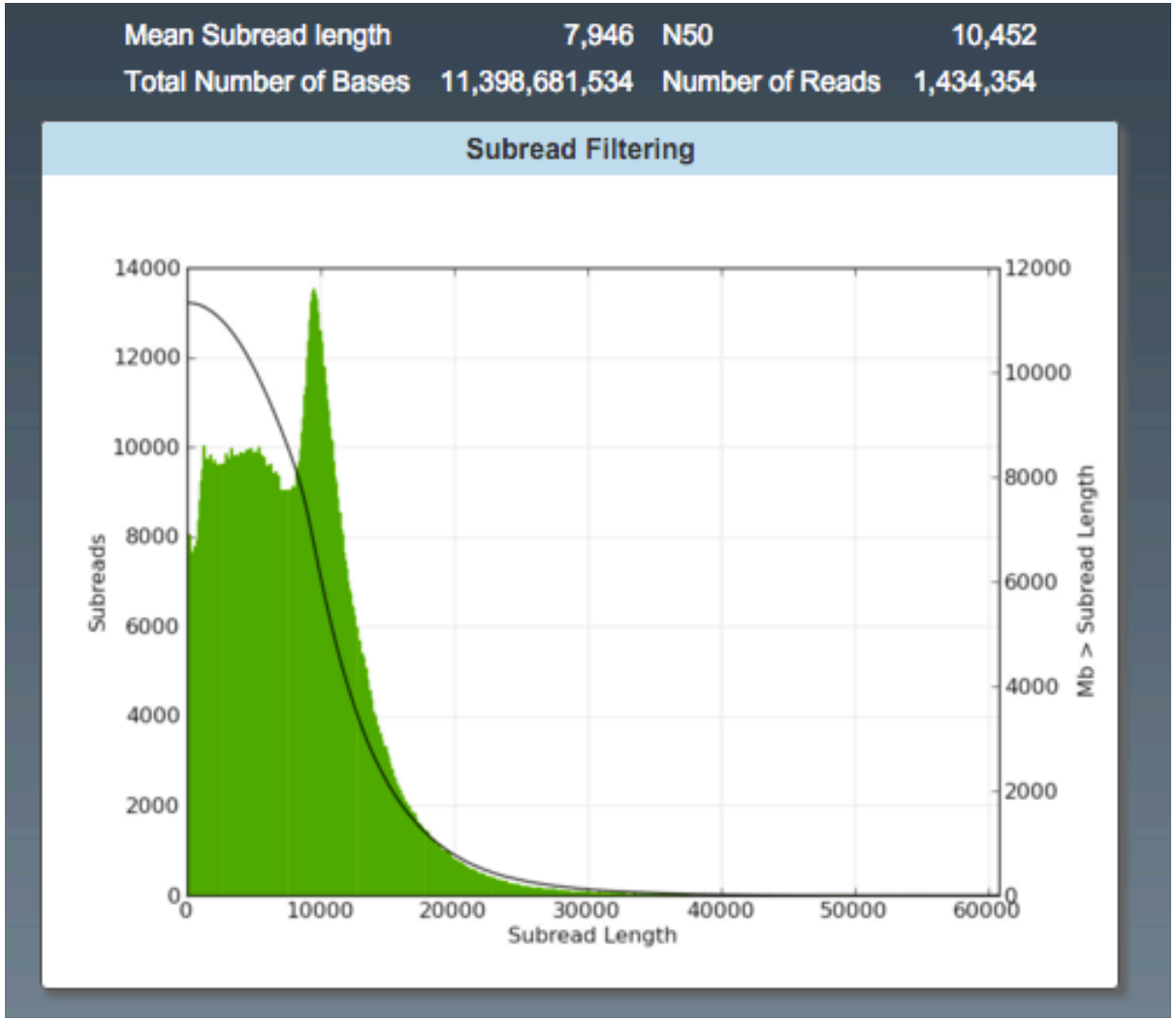


Average:	~ 8.5 kb
Maximum:	> 30 kb
Top 5% of reads:	> 18 kb
Half of data in reads:	> 10 kb
Data per SMRT® Cell:	~ 375 Mb

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

PacBio (Pacific Biosciences)



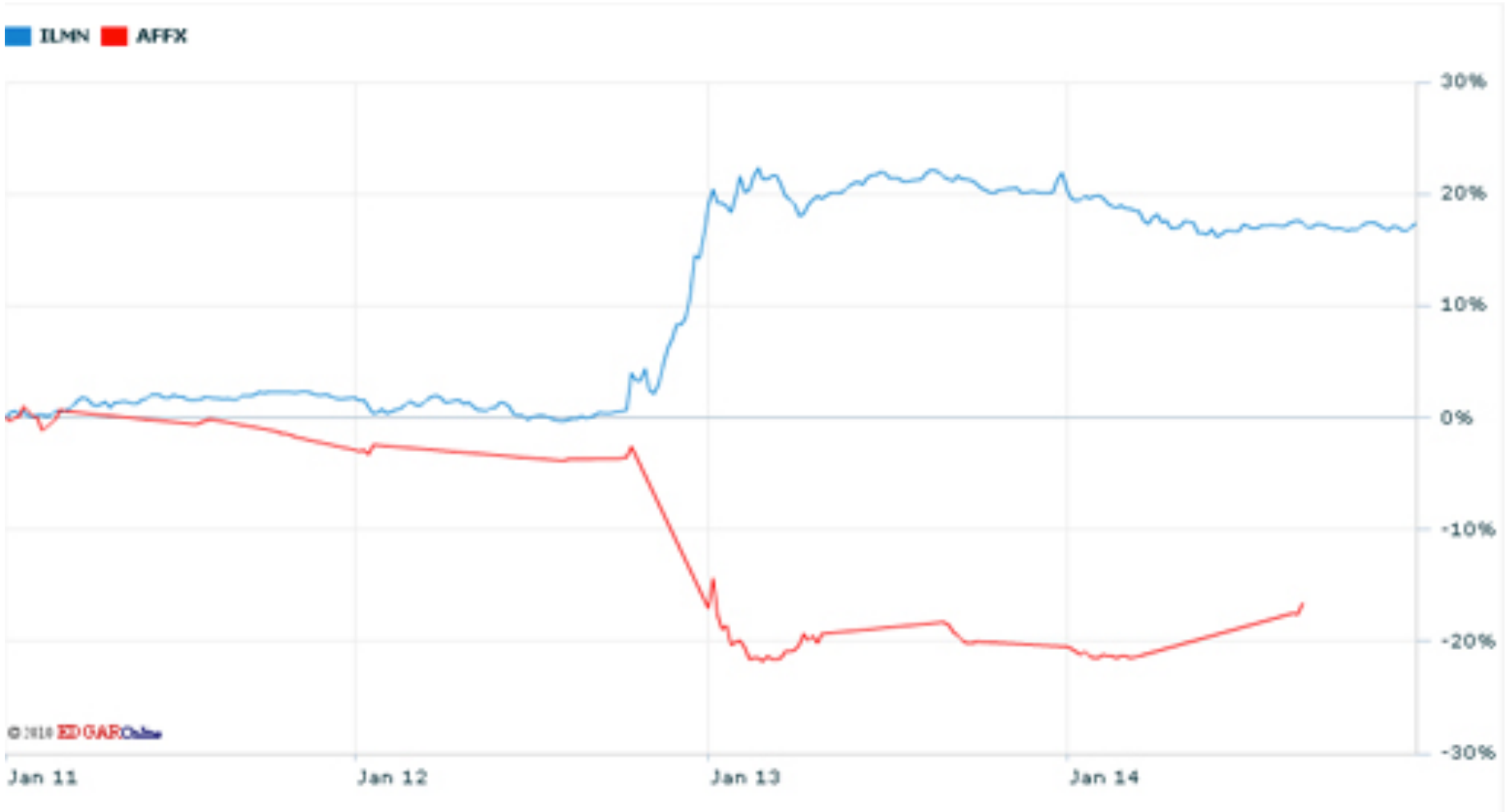


Oxford Nanopore

- 150Mb per run
- Tested 48kb read length
-$900 per instrument
-500 pores per device

- XXXMb per run
- Tested 48kb read length
-$XXX per instrument
-2000 pores per device, soon 8000 pores
-Cost per human genome $1500.

Come and go of technologies



Break here

NGS Data types

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?

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 assmeblies 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

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.**

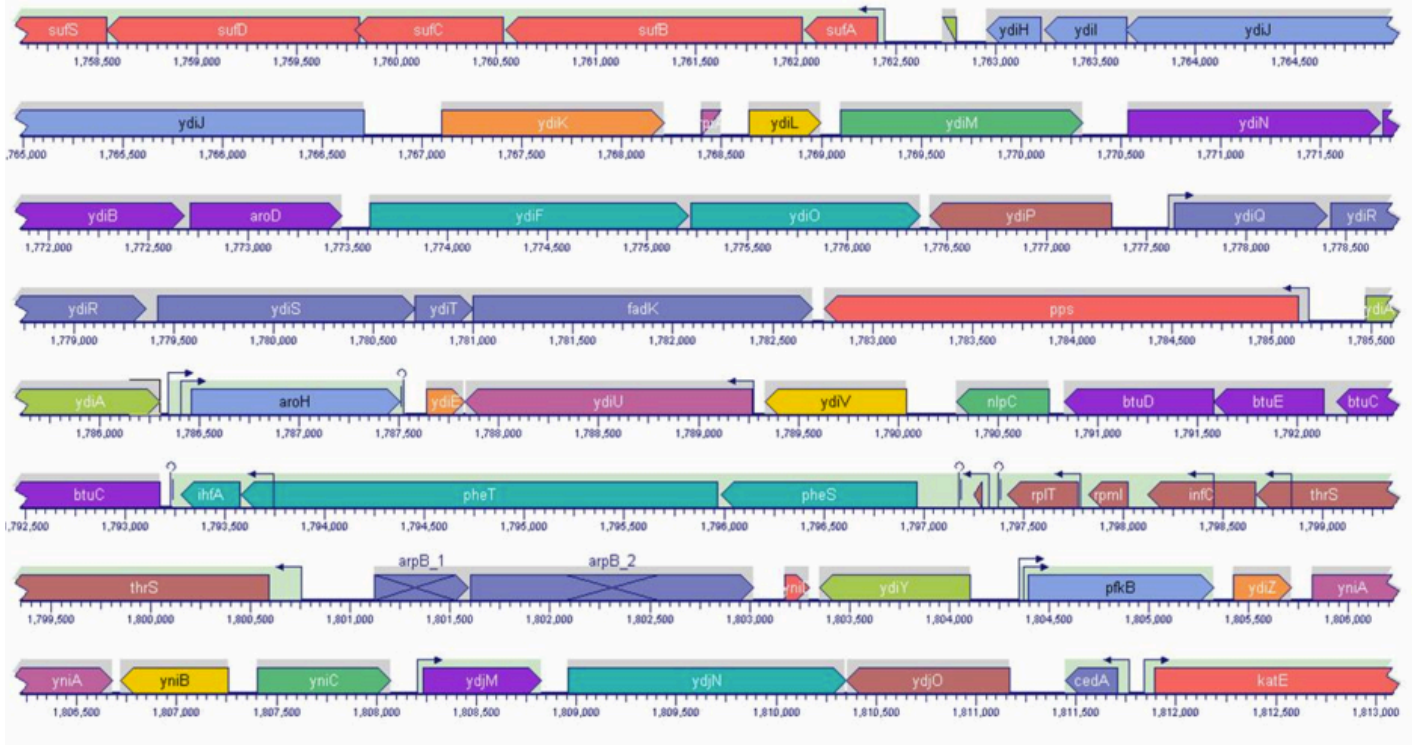
-Wikipedia

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

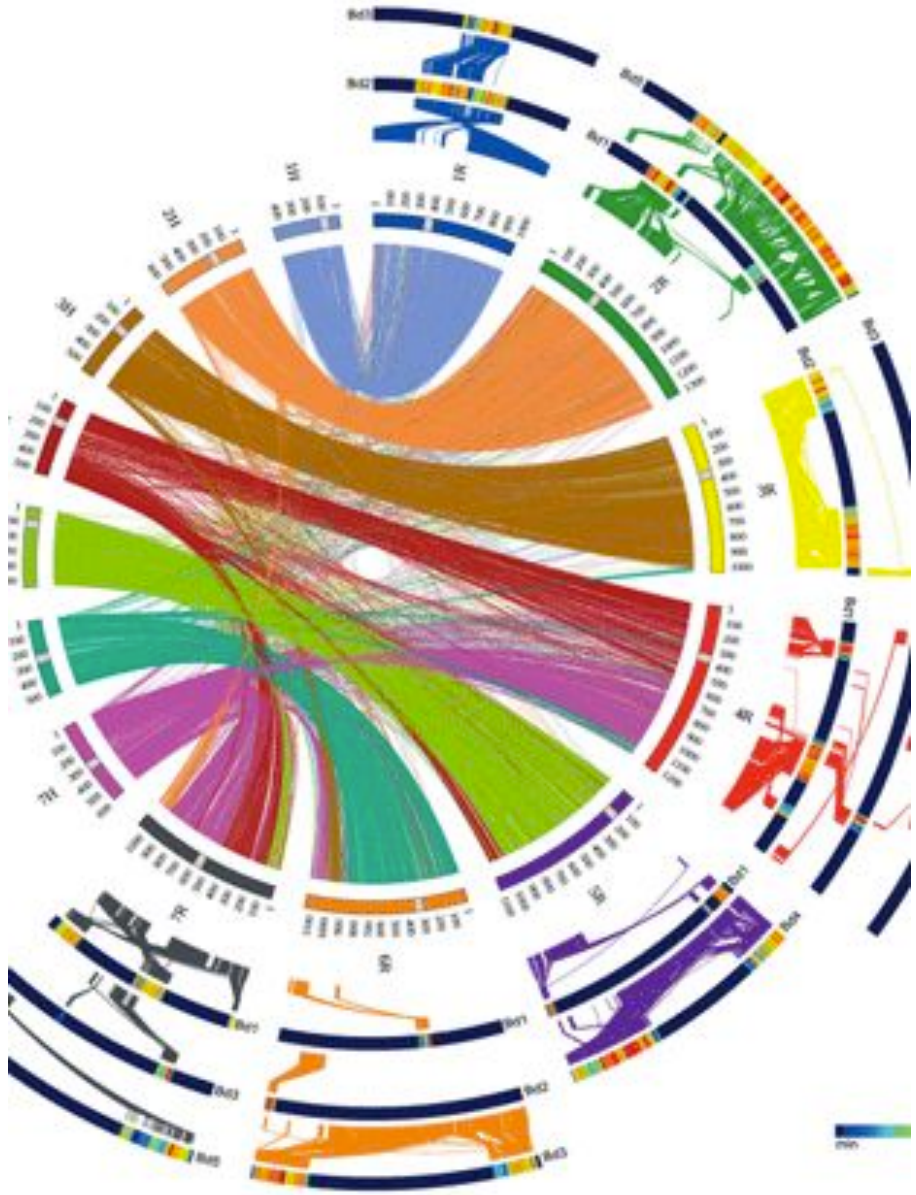
1. **Single line ID** with at symbol (“@”) in the first column.
2. There should be not space between “@” symbol and the first letter of the identifier.
3. Sequences are in multiple lines after the ID line
4. Single line with plus symbol (“+”) in the first column to represent the quality line.
5. Quality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

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

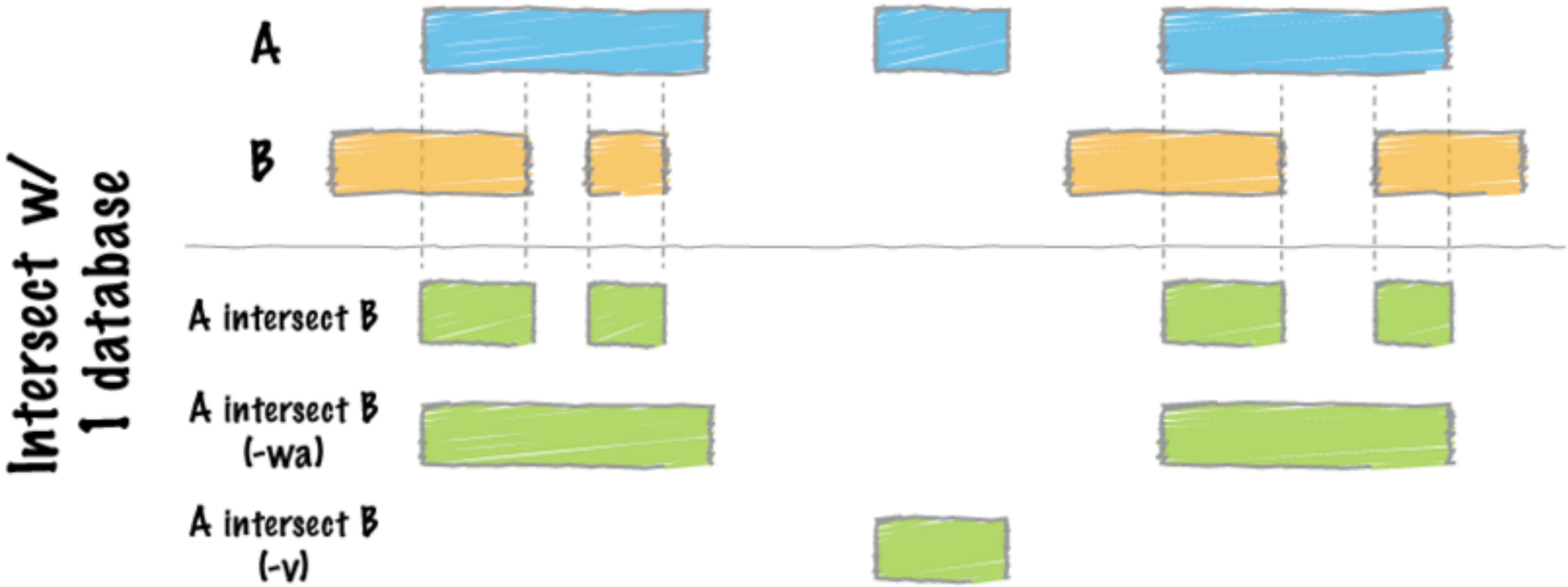
```
chr7    127471196  127472363
chr7    127472363  127473530
chr7    127473530  127474697
chr7    127474697  127475864
chr7    127475864  127477031
chr7    127477031  127478198
chr7    127478198  127479365
chr7    127479365  127480532
chr7    127480532  127481699
```

```
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
```

http://genome.ucsc.edu/FAQ/FAQformat#format1

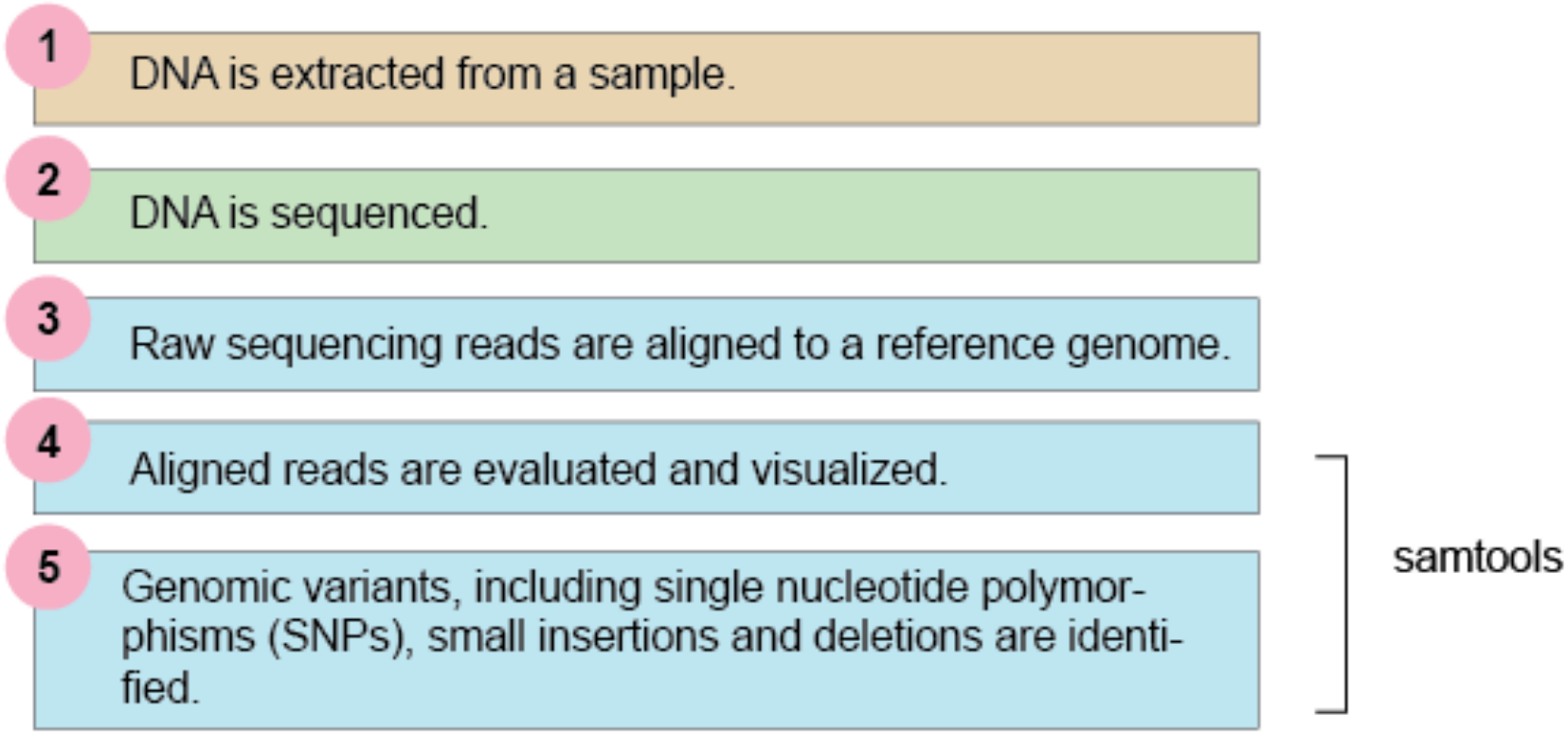
http://gmod.org/wiki/GFF2

Bedtools – extremely useful



http://bedtools.readthedocs.org/en/latest/

SAM format



http://biobits.org/samtools_primer.html

SAM format

- Everything in one line

HISEQ:134:C6H9FANXX:8:1110:10236:94013 **99** chr1 11844 0 150M = 12057 363
GGTATCATTCACCATTTTTCTTTTCGTTAACTTGCCGTCAGCCTTTTCTTTGACCTCTTCTTTCTGTTC
ATGTGTATTTGCTGTCTCTTAGCCCAGACTTCCCGTATCCTTTCCACCGGGCCTTTGAGAGGTCACA
GGGTCTTGATGCTG
>A=>AFDEEGEGGEFFFFFFFFFFGCDFBEGFFHFGCDGEHGGFFFFGFFEEDGGFGFFGFFFDFEDF
GCFHCHDBEFFHFEGCFEFED@CEEEBEADCBBCB>?,?AA@@?@?>@?;?@??===?=?;<=;?@GEH
GFDGFFHAC=?=@ MC:Z:150M
BD:Z:NNOOPSQQNOPMNOOMGGGNMMGGNONNLMNONOPOMNPOPQOONHGONNHOPOOOO
NNONNHONPMNOPPPMNMONNHOPPPMQNONNM

https://broadinstitute.github.io/picard/explain-flags.html

http://www.htslib.org/doc/samtools.html

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 POSistion
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

https://broadinstitute.github.io/picard/explain-flags.html

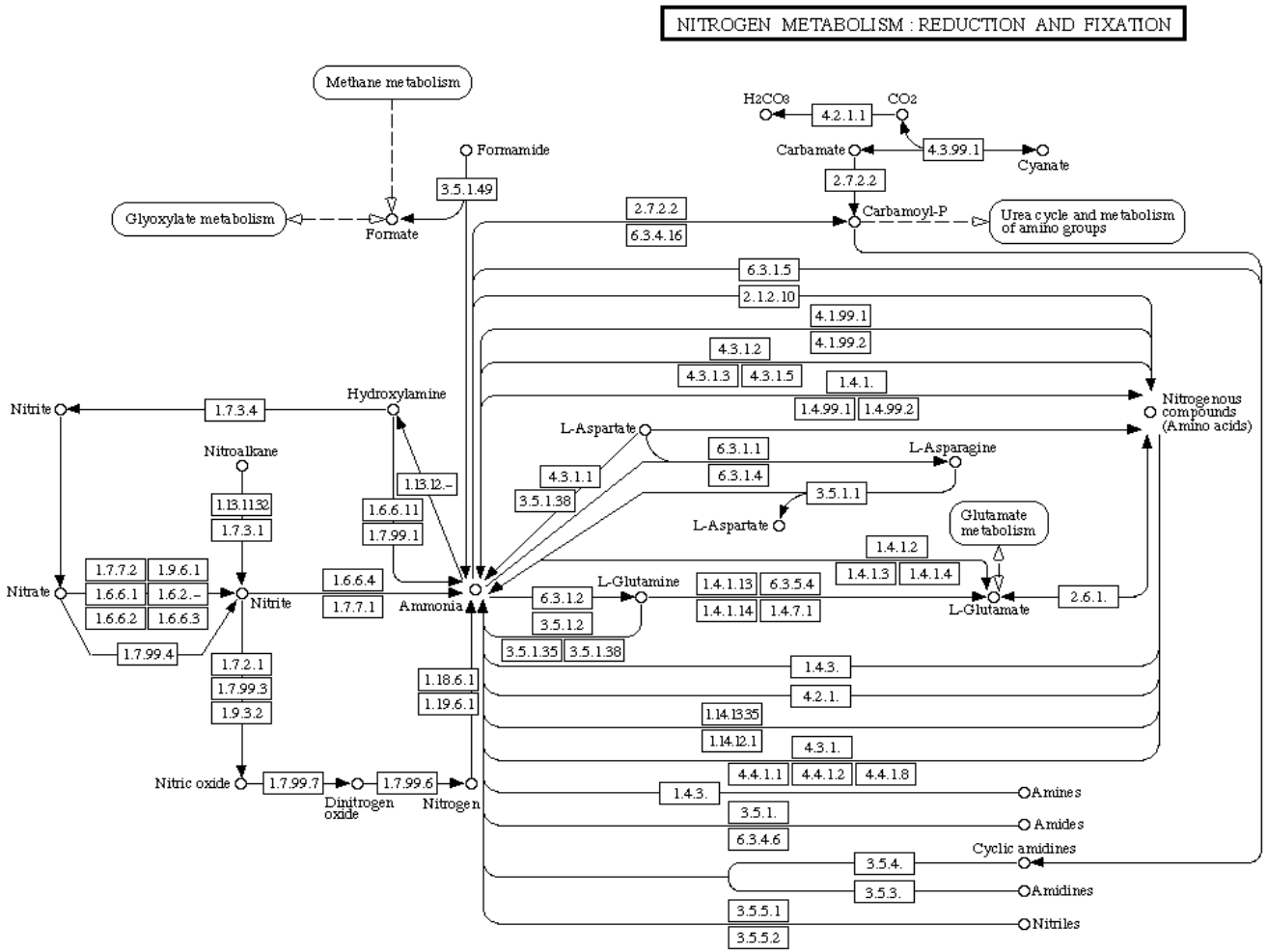
http://www.htslib.org/doc/samtools.html

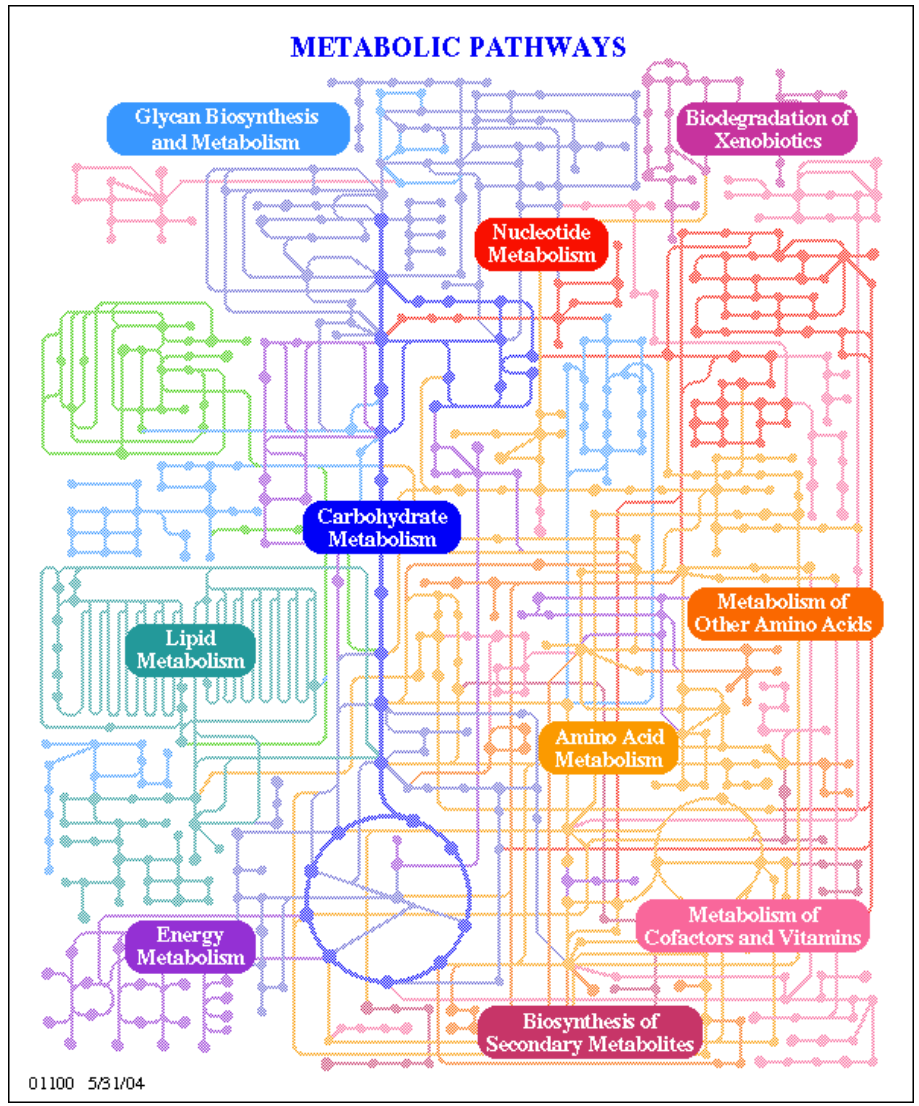
Public databases

- 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?

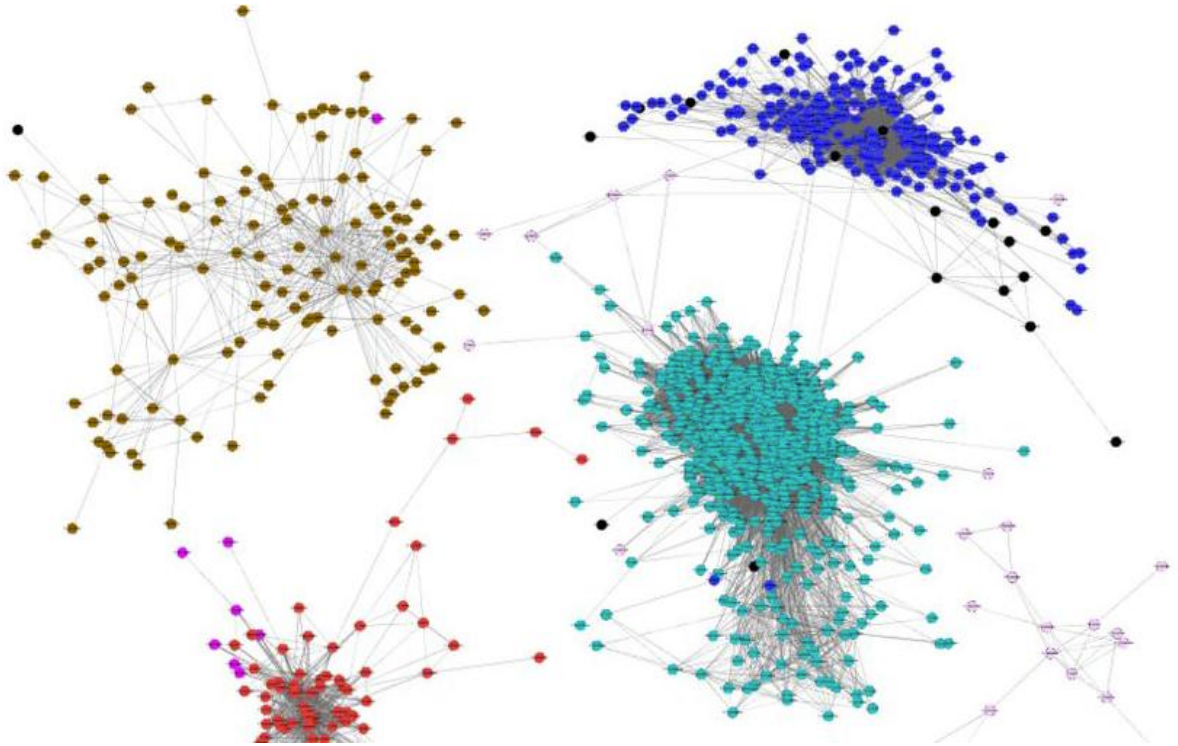
- Remember to deposit your own data to contribute

KEGG: Kyoto Encylopedia of Genes and Genomes





Importance of networks in biology

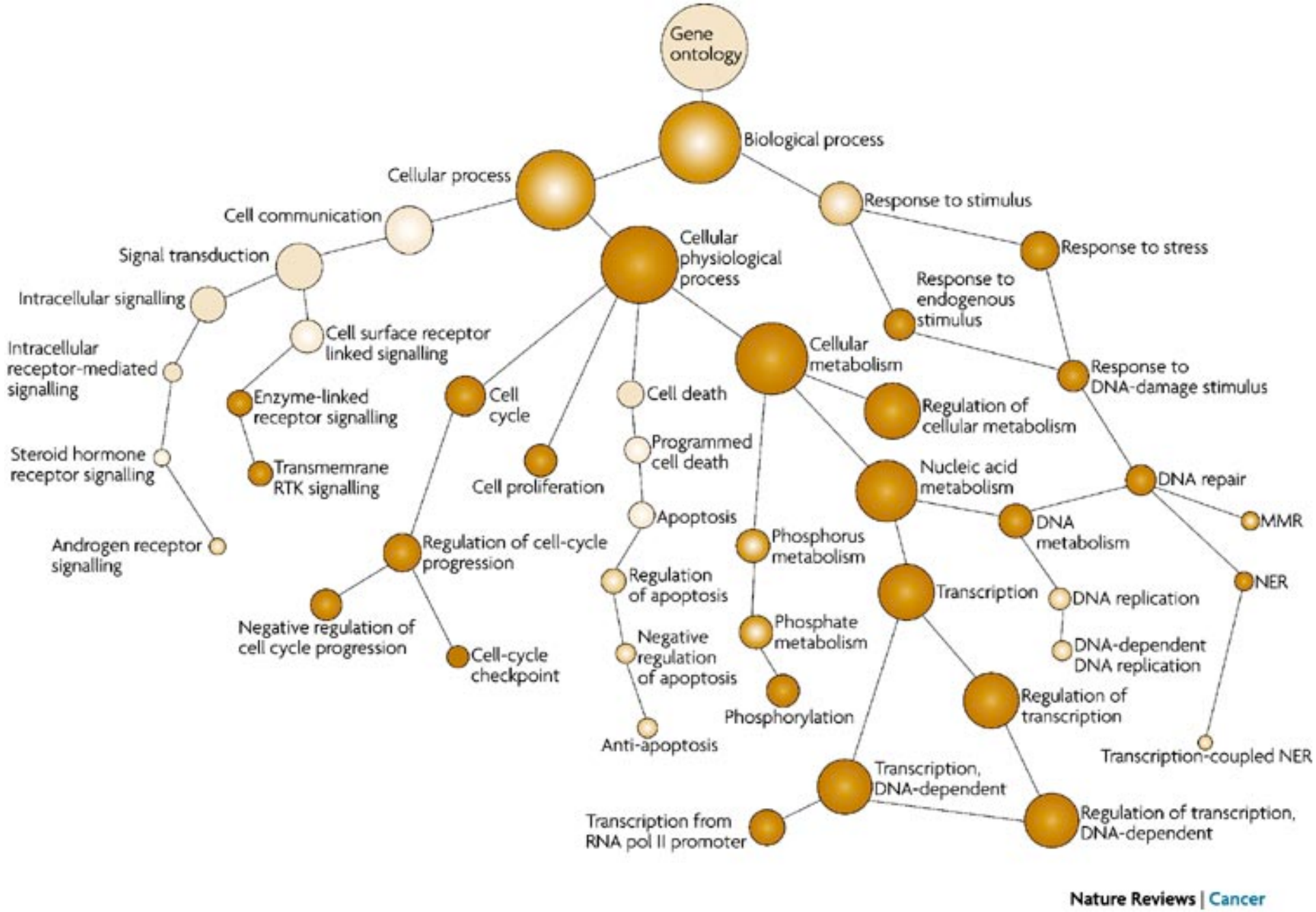


Gene interaction networks

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 biologial goals - e.g cell division)
  - Cellular component (location)

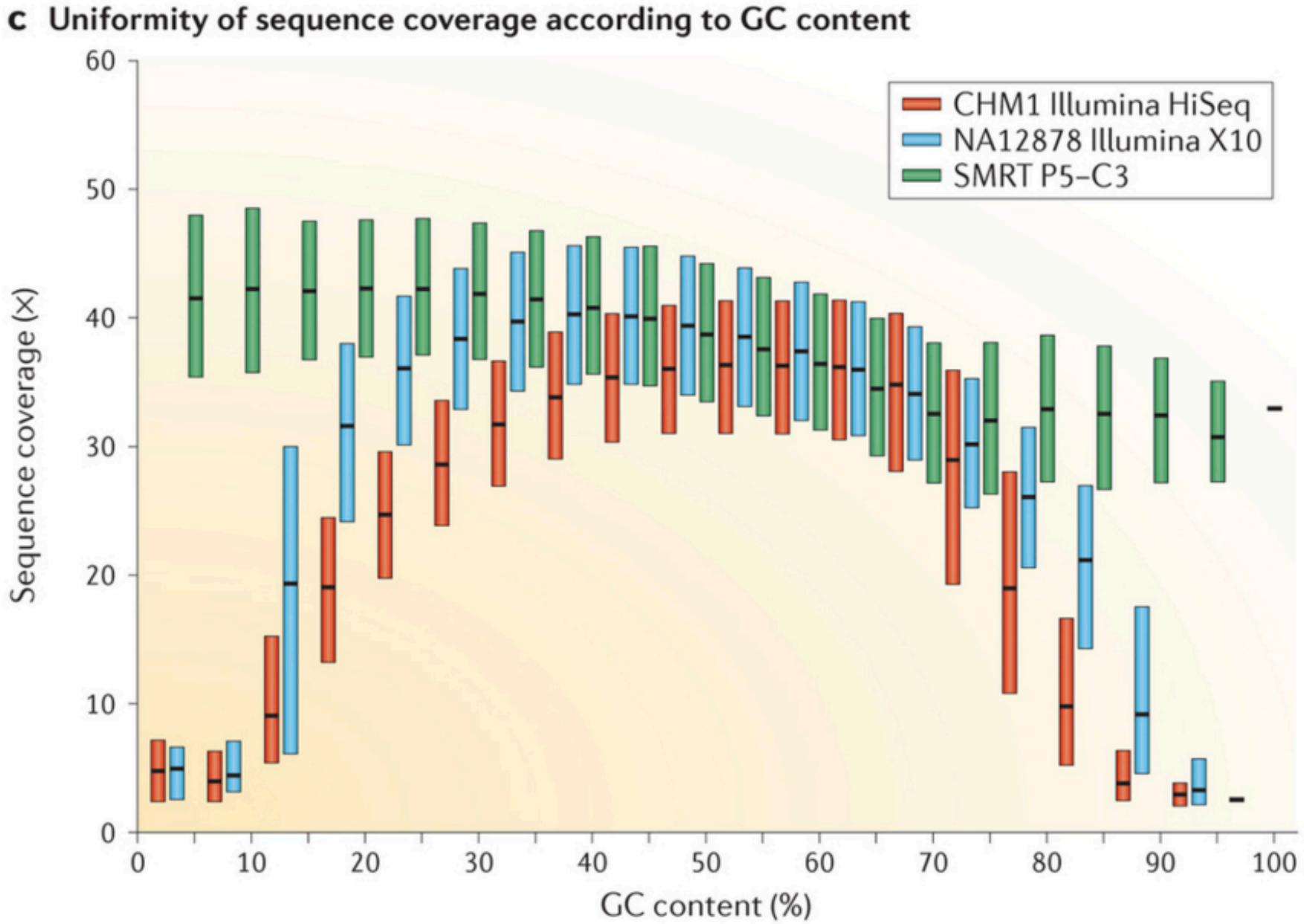


Nature Reviews | Cancer

http://www.nature.com/nrc/journal/v7/n1/fig_tab/nrc2036_F1.html

Analysis and interpretation

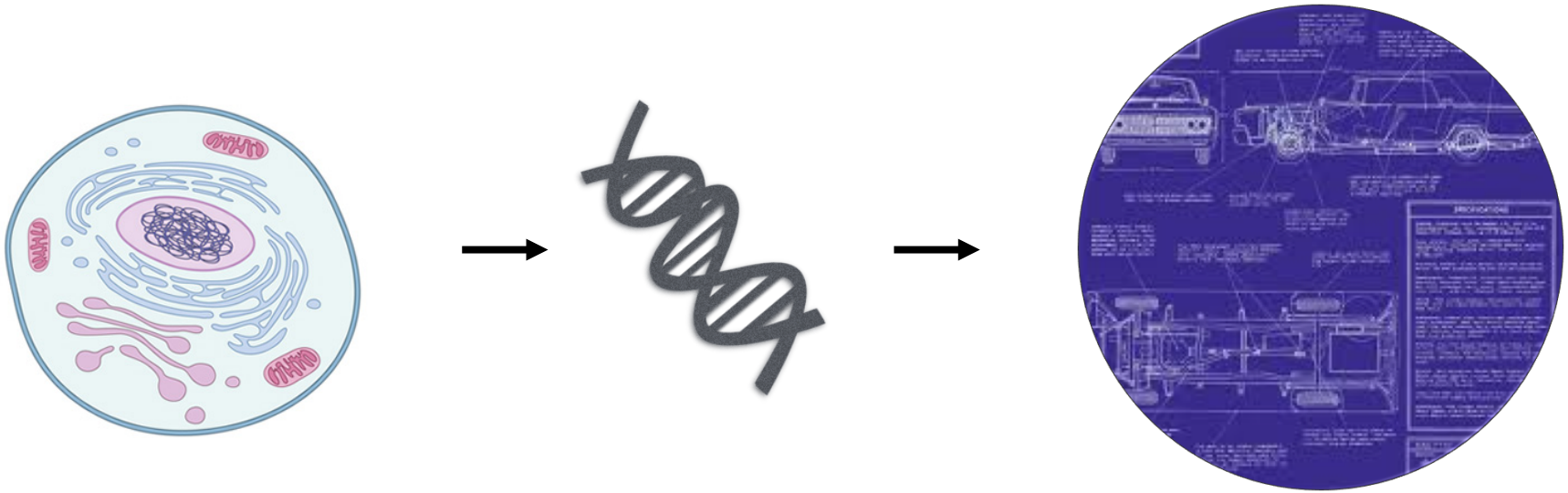
Biases



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

Genomics

Genome



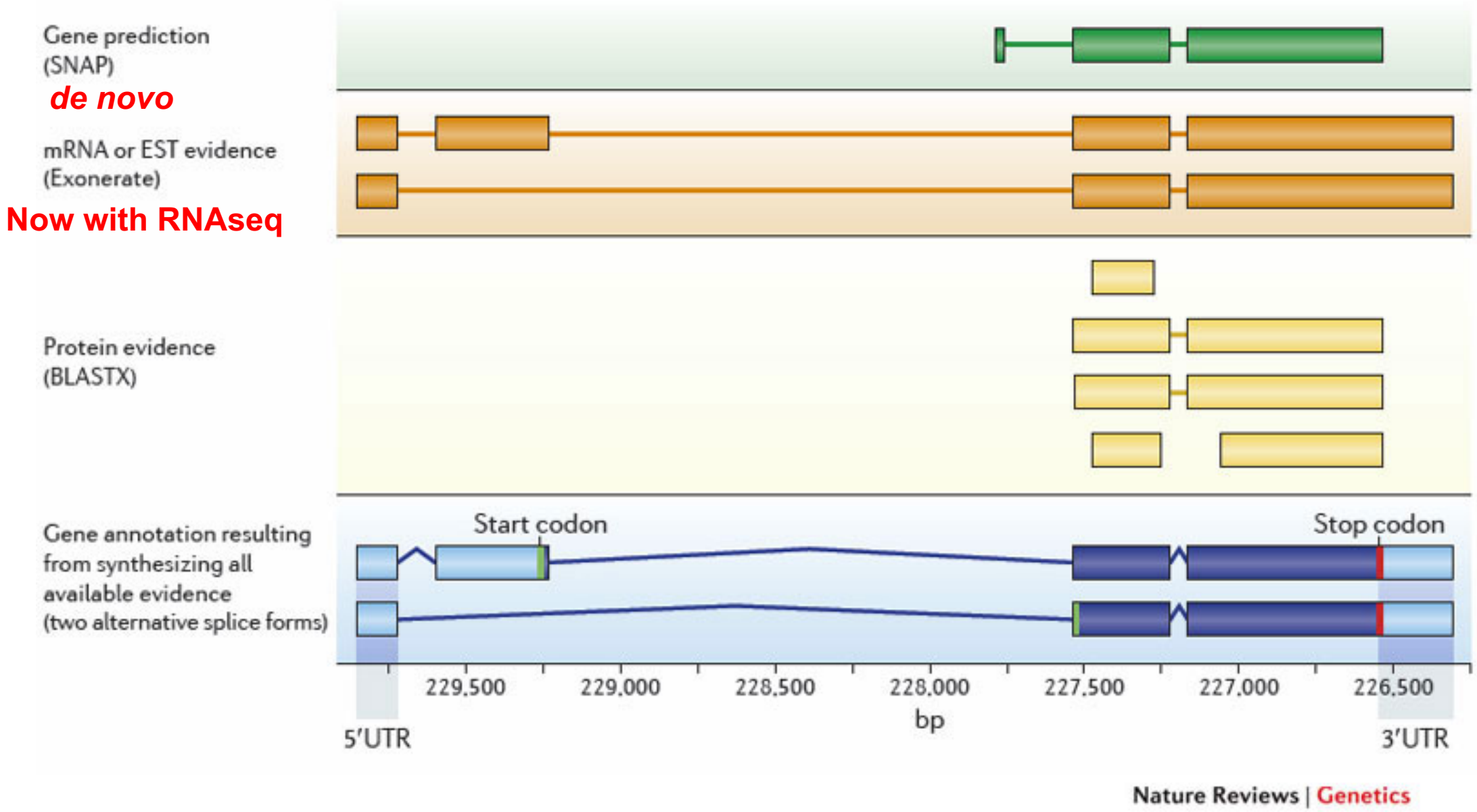
Genome = Parts list of a single genome

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)

Classical genetics

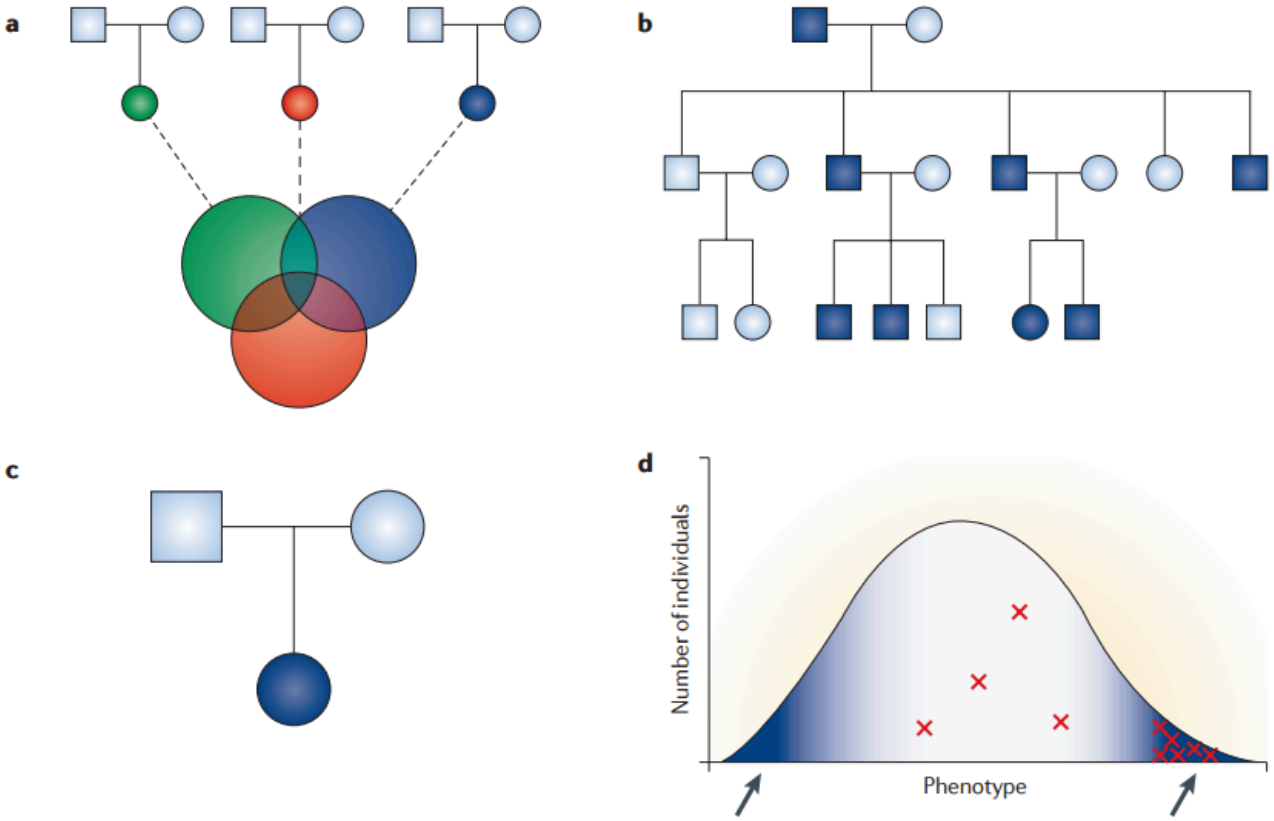
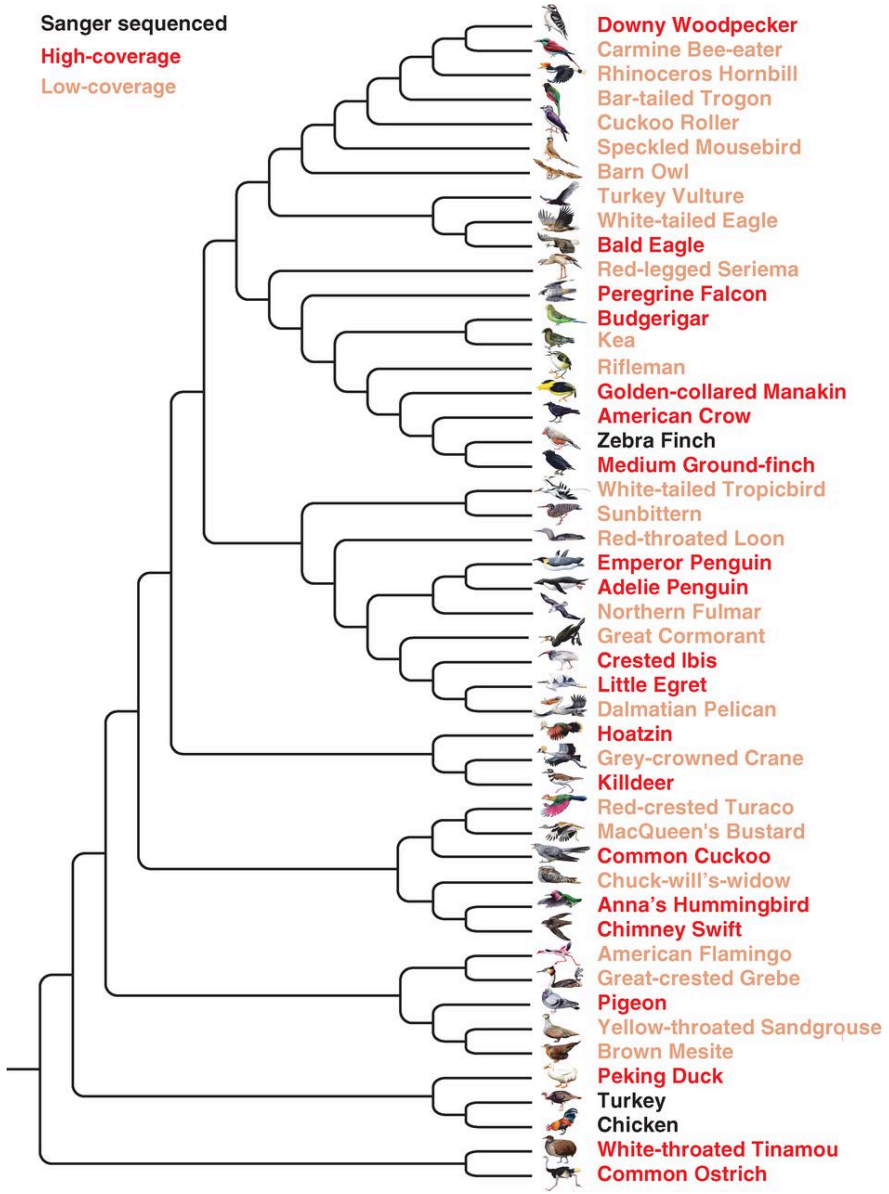


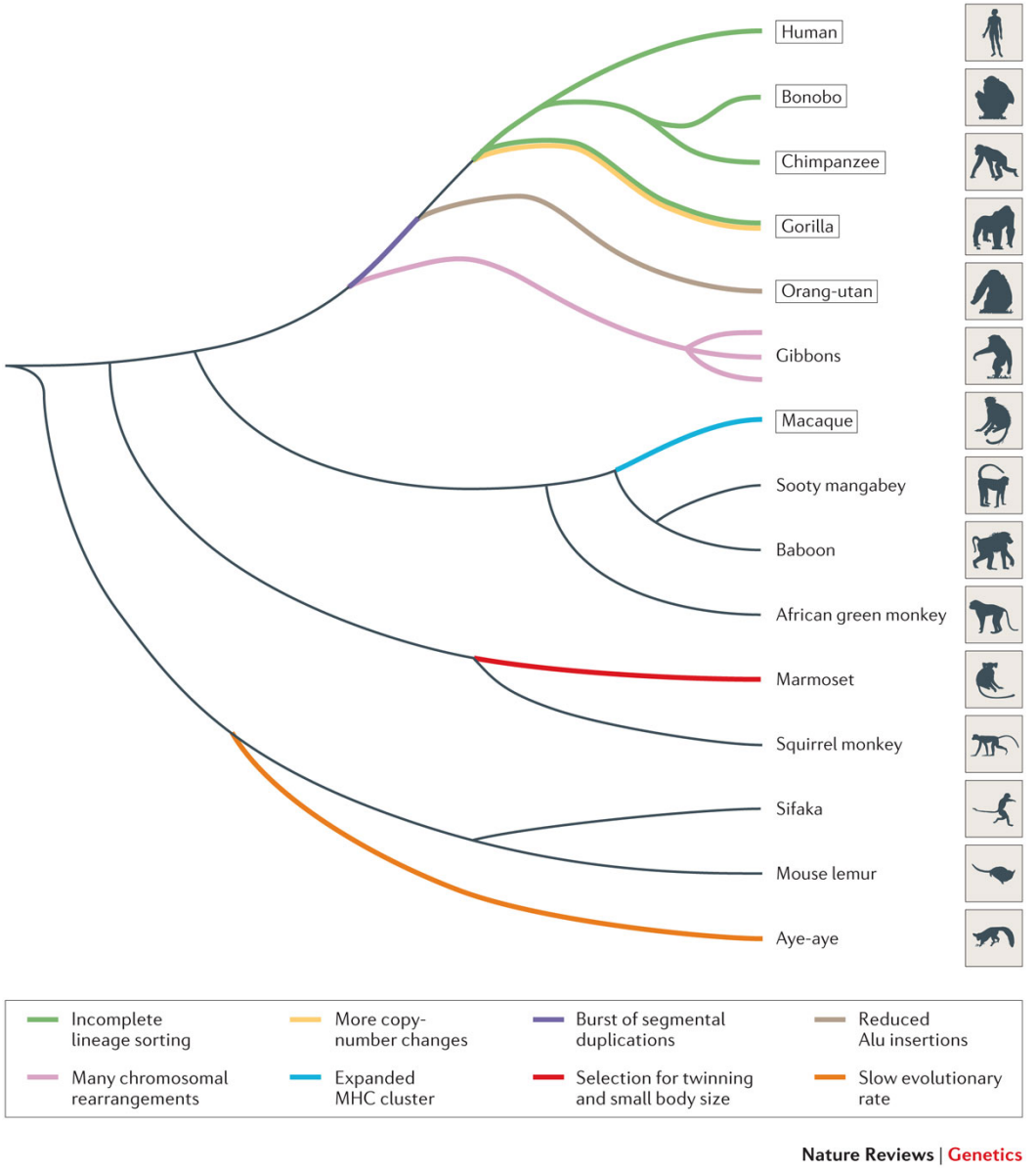
Figure 2 | **Strategies for finding disease-causing rare variants using exome sequencing.** Four main strategies are illustrated. **a** | Sequencing and filtering across multiple unrelated, affected individuals (indicated by the three coloured circles). This approach is used to identify novel variants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



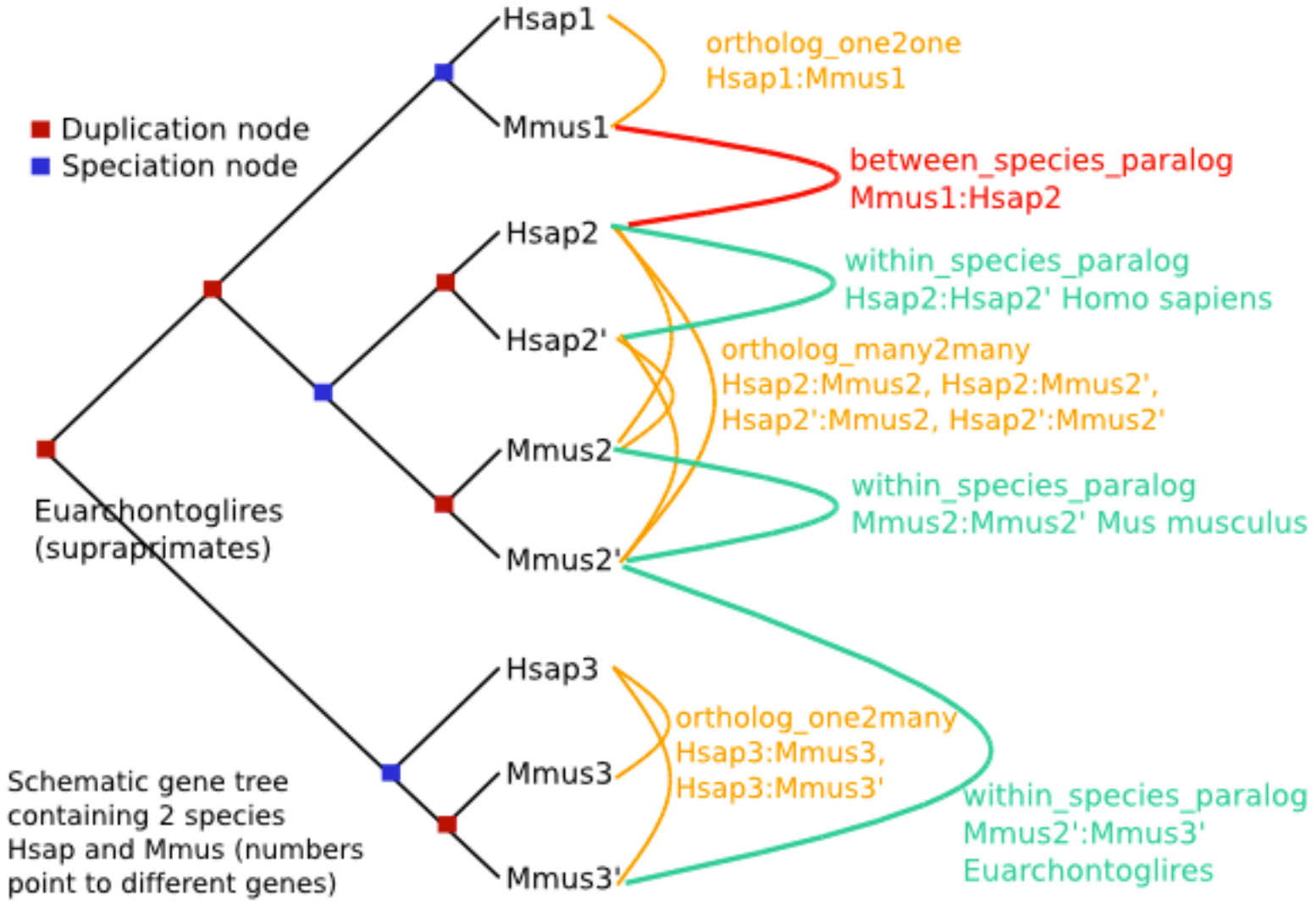
Roger & Gibbs Nature Reviews Genetics (2014)

Homologs: Orthologs and paralogs

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

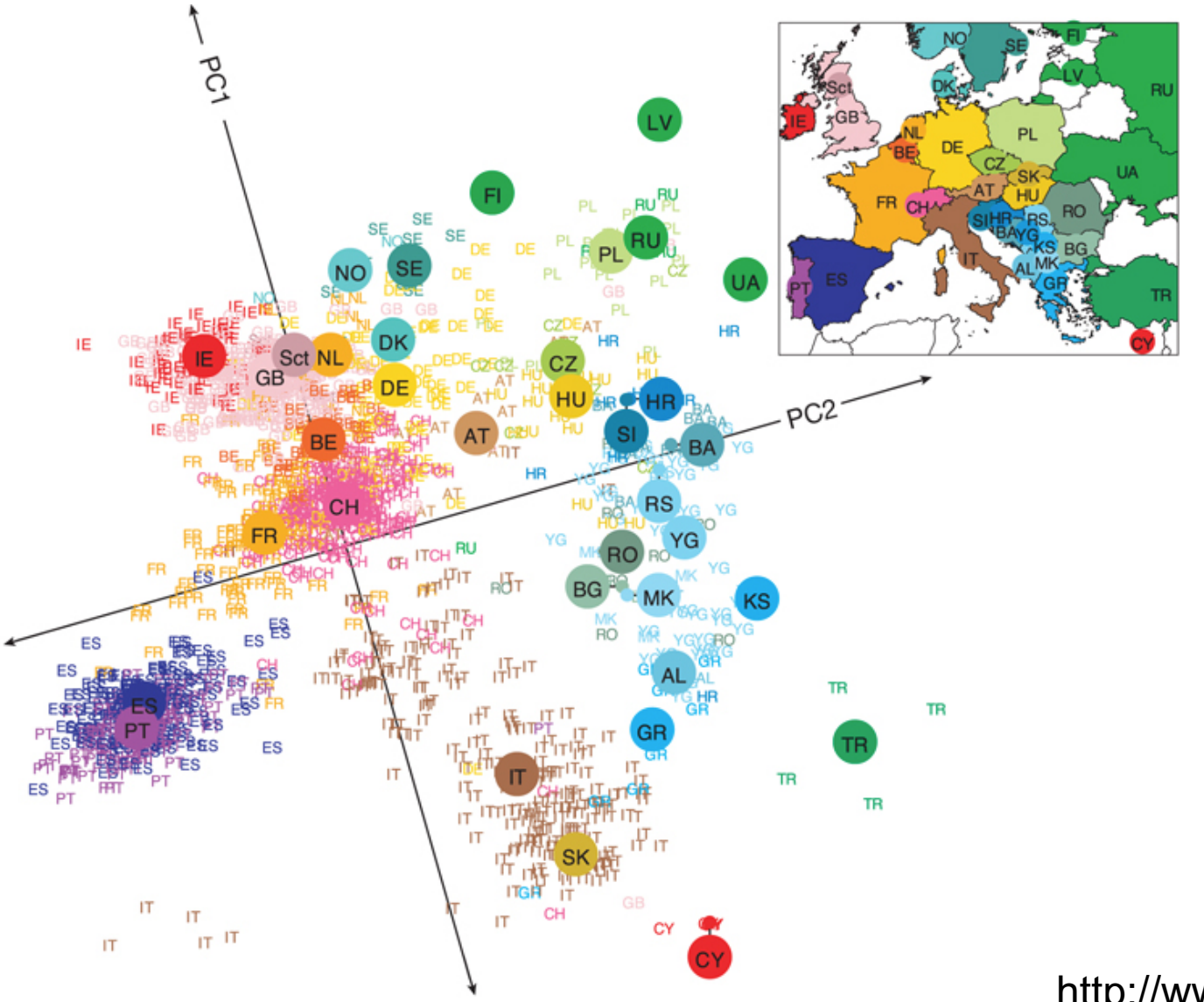
Depending on the number of genes found in each species, we differentiate among 1:1, 1:many and many:many relationships.

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

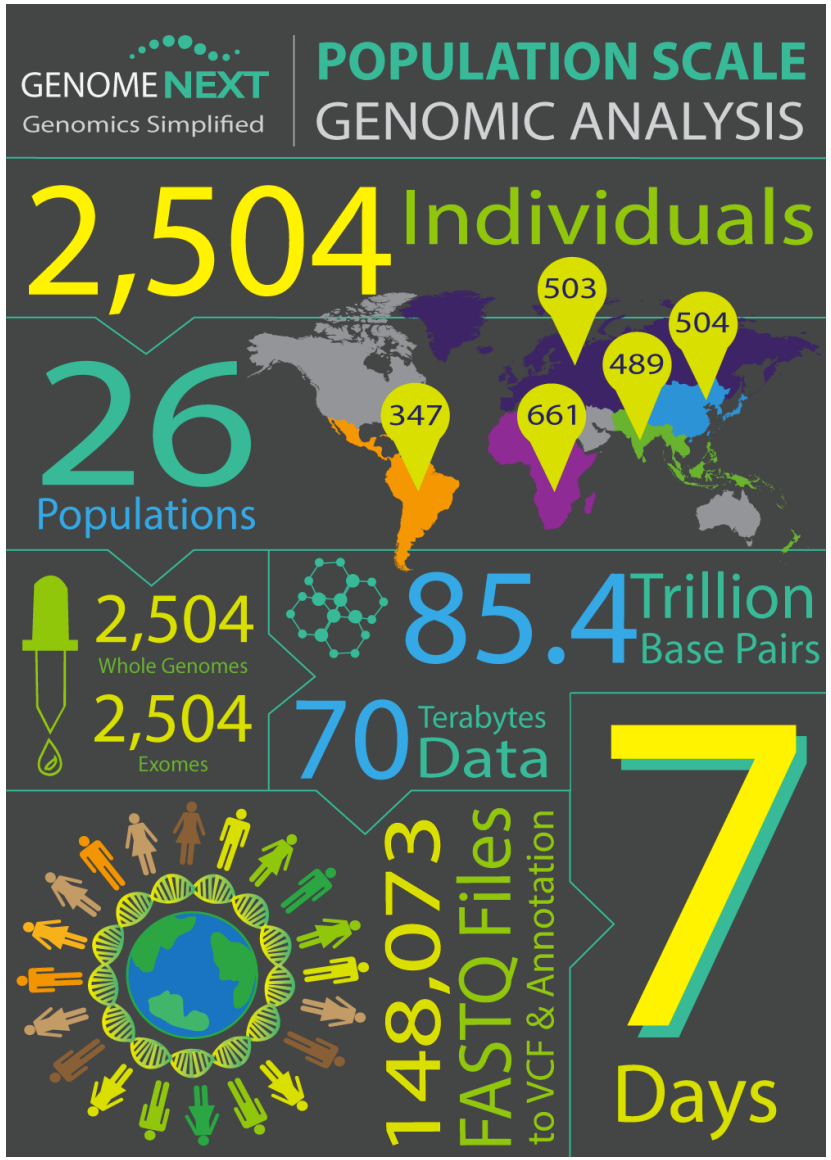


http://asia.ensembl.org/info/genome/compara/homology_method.html?redirect=no

Population genomics

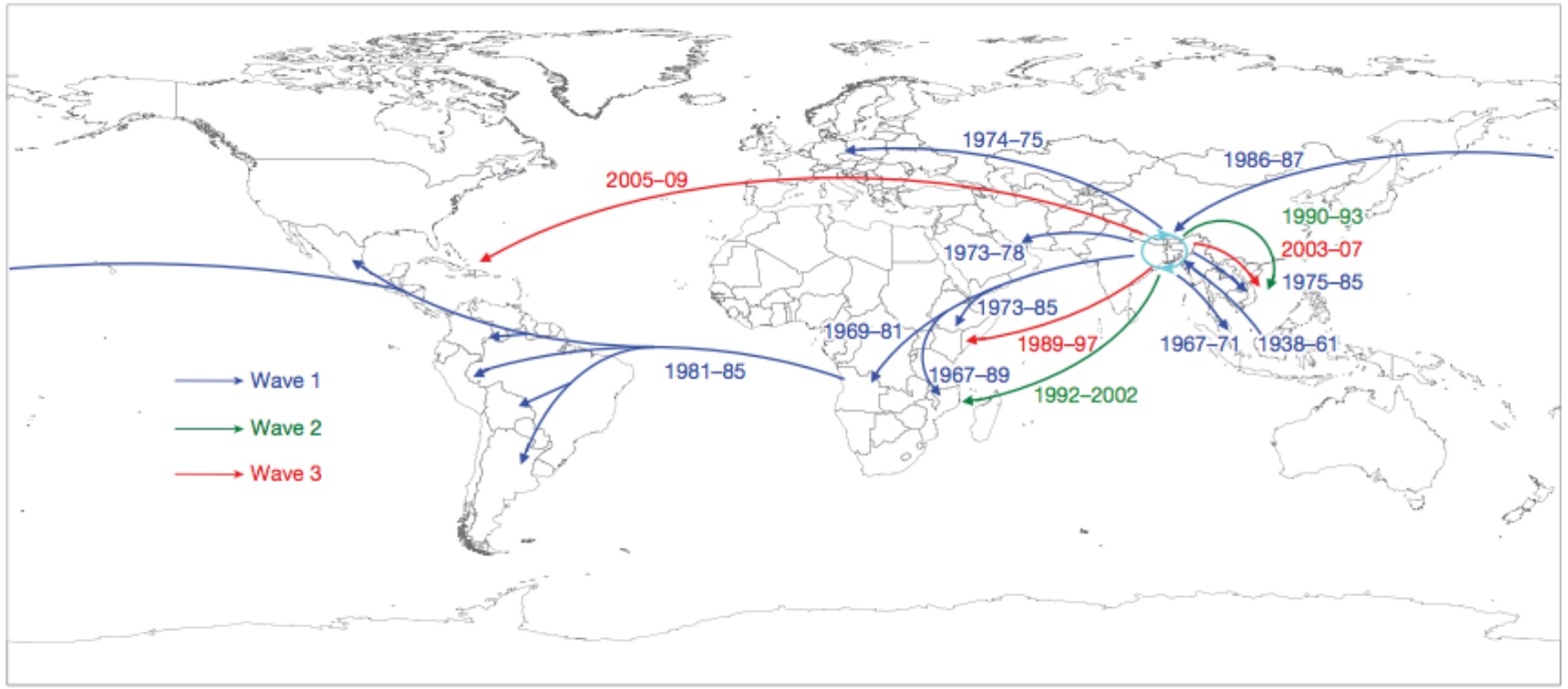


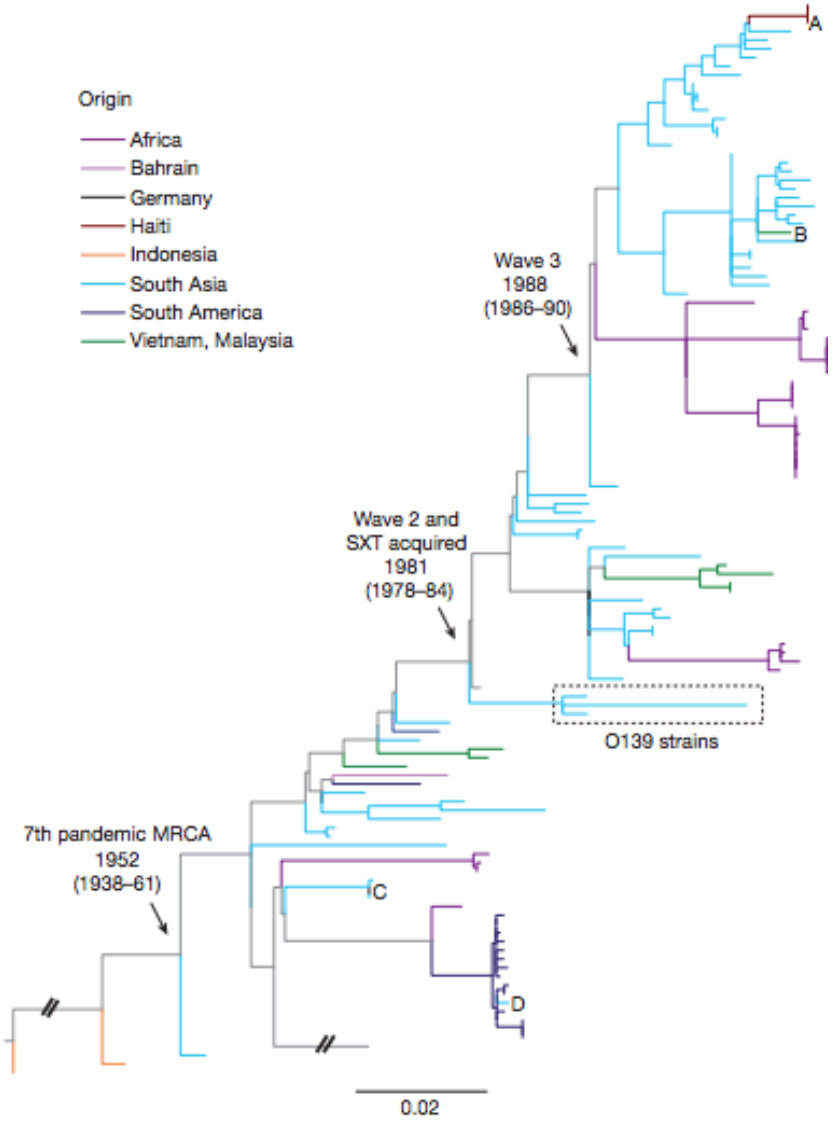
Novembre et al Nature (2008)



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

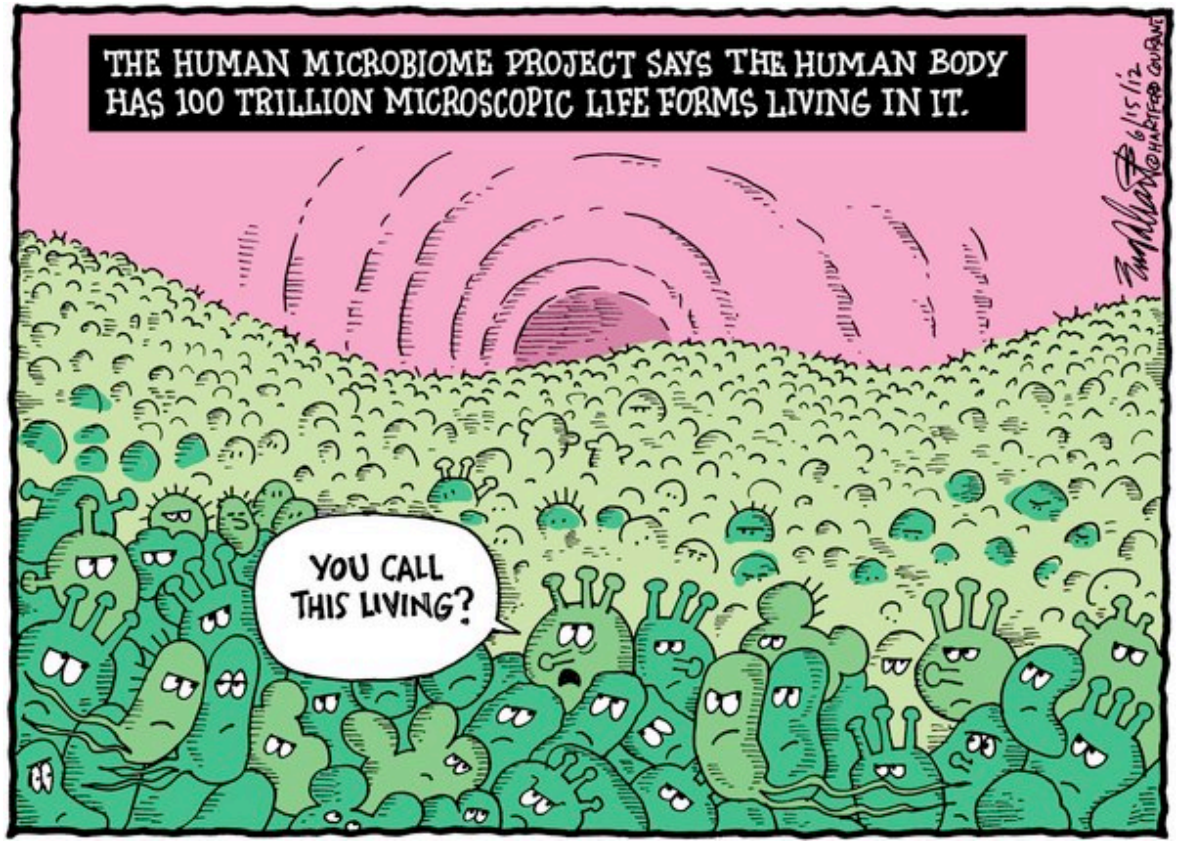
Population genomics

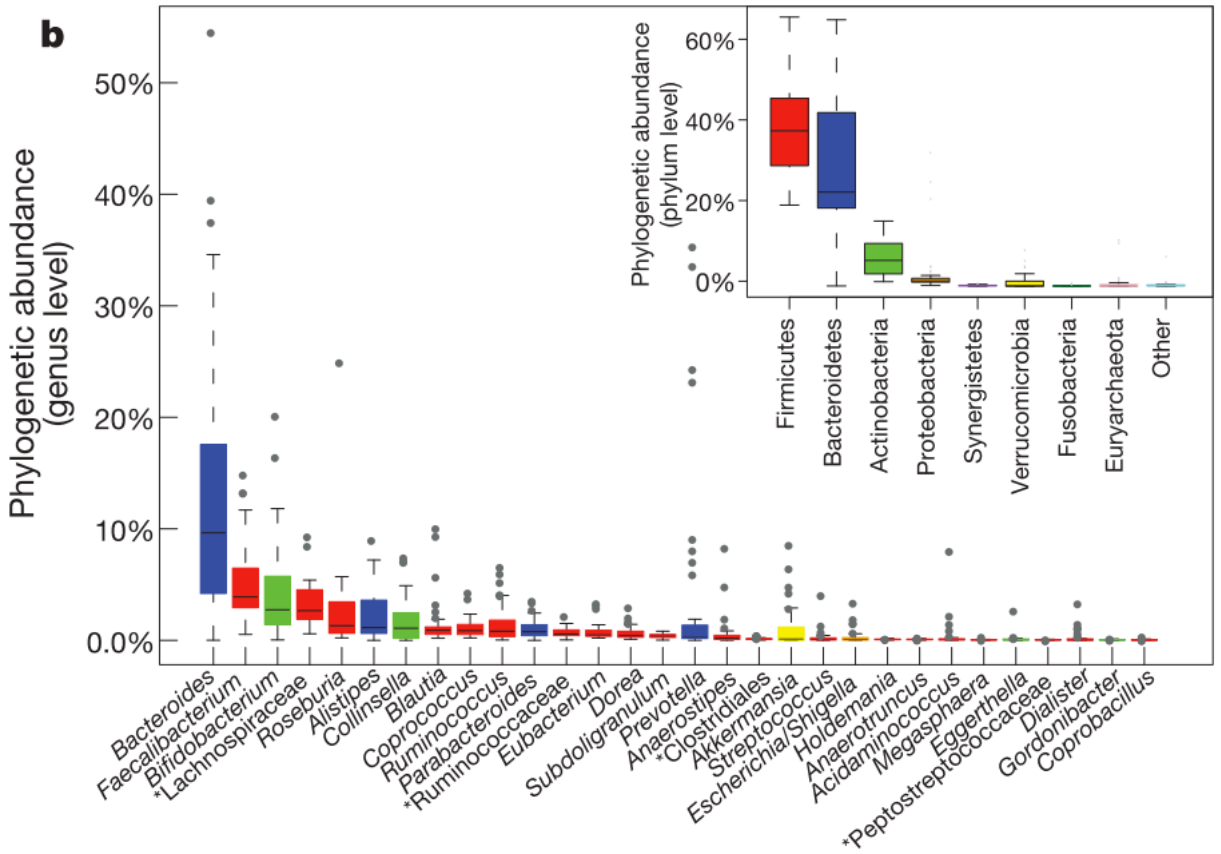




Mutreja *et al.*, Nature Genetics (2011)

Metagenomics





doi:10.1038/nature09944

Break

Transcriptomics / RNAseq

Applications of RNAseq

Discovery / Annotation

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

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

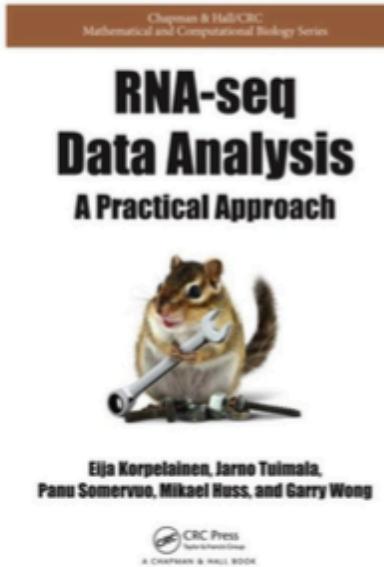
- expression
- Isoform abundance, splice patterns, transcript boundaries

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

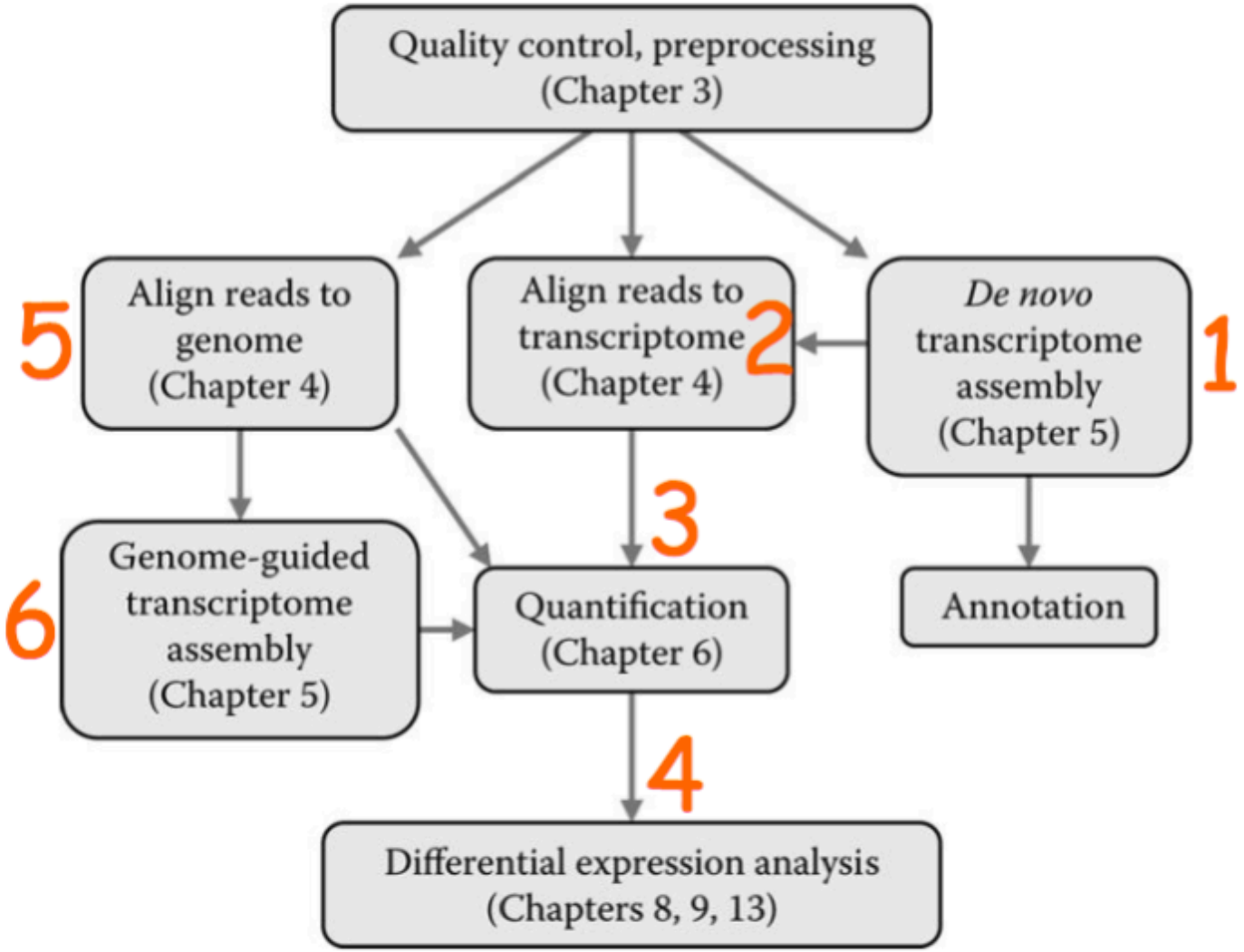


FIGURE 2.1 Possible paths in RNA-seq data analysis.

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

Next Generation Sequencing technologies

	Strenghs	Weaknesses
454 Pyrosequencing (GS FLX Titanium XL+)	– Long reads (450/700 bp). – Long insert for mate pair libraries (20Kb). – Low observed raw error rate (0.1%) – Low percentage of PCR duplications for mate pair libraries	– Homopolymer error. – Low sequence yield per run (0.7 Gb). – Preferred assembler (gsAssembler) uses overlapping methodology.
Illumina (HiSeq 2500)	– High sequence yield per run (600 Gb) – Low observed raw error rate (0.26%)	– High percentage of PCR duplications for mate pair libraries. – Long run time (11 days) – High instrument cost (~ $650K)
Illumina (MiSeq)	– Medium read size (250 bp) – Faster run than Illumina HiSeq	– Medium sequence yield per run (8.5 Gb)
SOLID (5500xl system)	– 2-base encoding reduce the observed raw error rate (0.06%)	– 2-base color coding makes difficult the sequence manipulation and assembly. – Short reads (75 bp)
Ion Torrent (Ion Proton I)	– Fast run (2 hours) – Low instrument cost ($80K). – Medium read size (200 bp)	– Medium sequence yield per run (10 Gb) – Medium observed raw error rate (1.7%)
PacBio (PacBioRS)	– Long reads (3000 bp) – Fast run (2 hours)	– Really high observed raw error rate (12.7%) – High instrument cost (~ $700K) – No pair end/mate pair reads



Next Generation Sequencing technologies

	Inputs	Outputs
454 Pyrosequencing (GS FLX Titanium XL+)	– Single Reads Library. – Pair End Library (3 to 20 Kb insert size). – Multiplexed sample.	– sff files – (fasta and fastq files)
Illumina (HiSeq 2500)	– Single Reads Library. – Pair End Library (170-800 bp insert size). – Mate Pair Library (2 to 10 Kb insert Size) – Multiplexed sample.	– fastq files (Phred+64) – fastq files (Phred+33, Illumina 1.8+)
Illumina (MiSeq)		
SOLID (5500xl system)	– Single Reads Library. – Mate Pairs Library (0.6 to 6 Kb insert size). – Multiplexed sample.	– fastq files (Phred+33)
Ion Torrent (Ion Proton I)	– Single Reads Library. – Pair End Library (0.6 to 6 Kb insert size). – Multiplexed sample.	
PacBio (PacBioRS)	– Single Reads Library.	

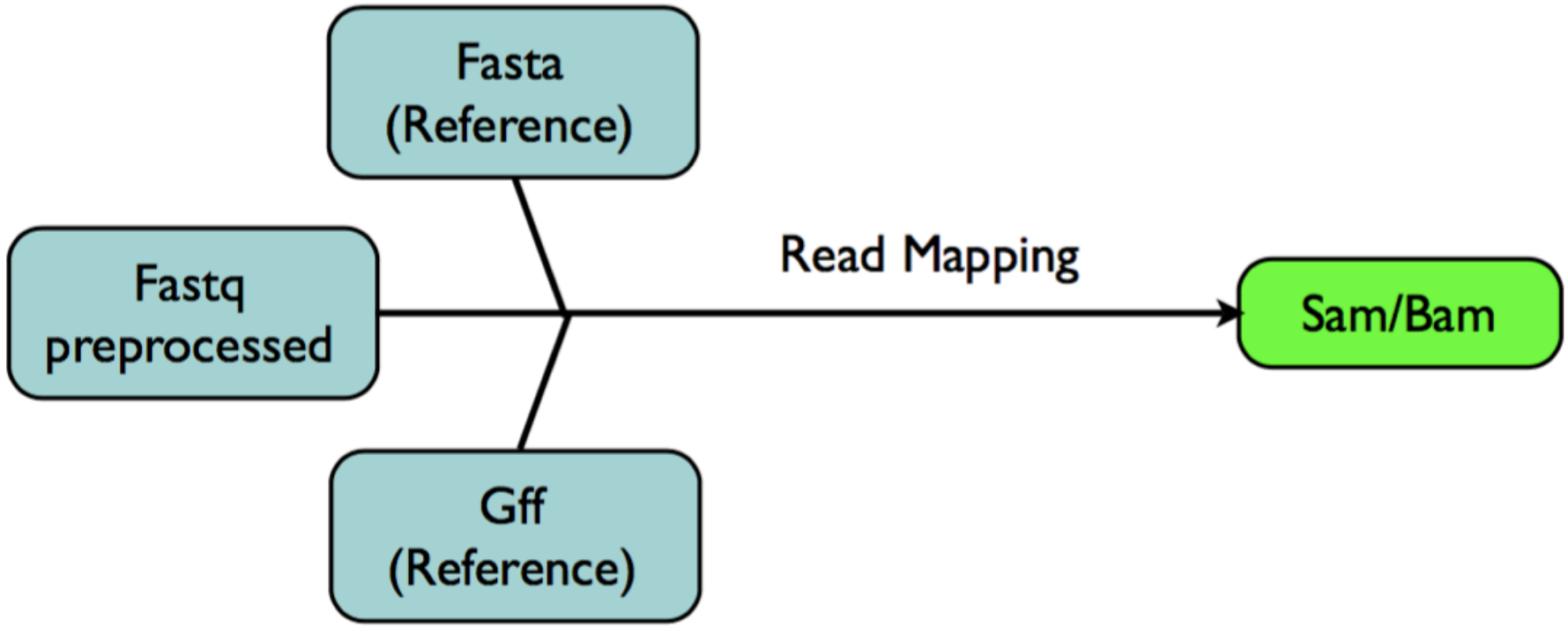
Once you have the raw data: **map** the RNAseq reads

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

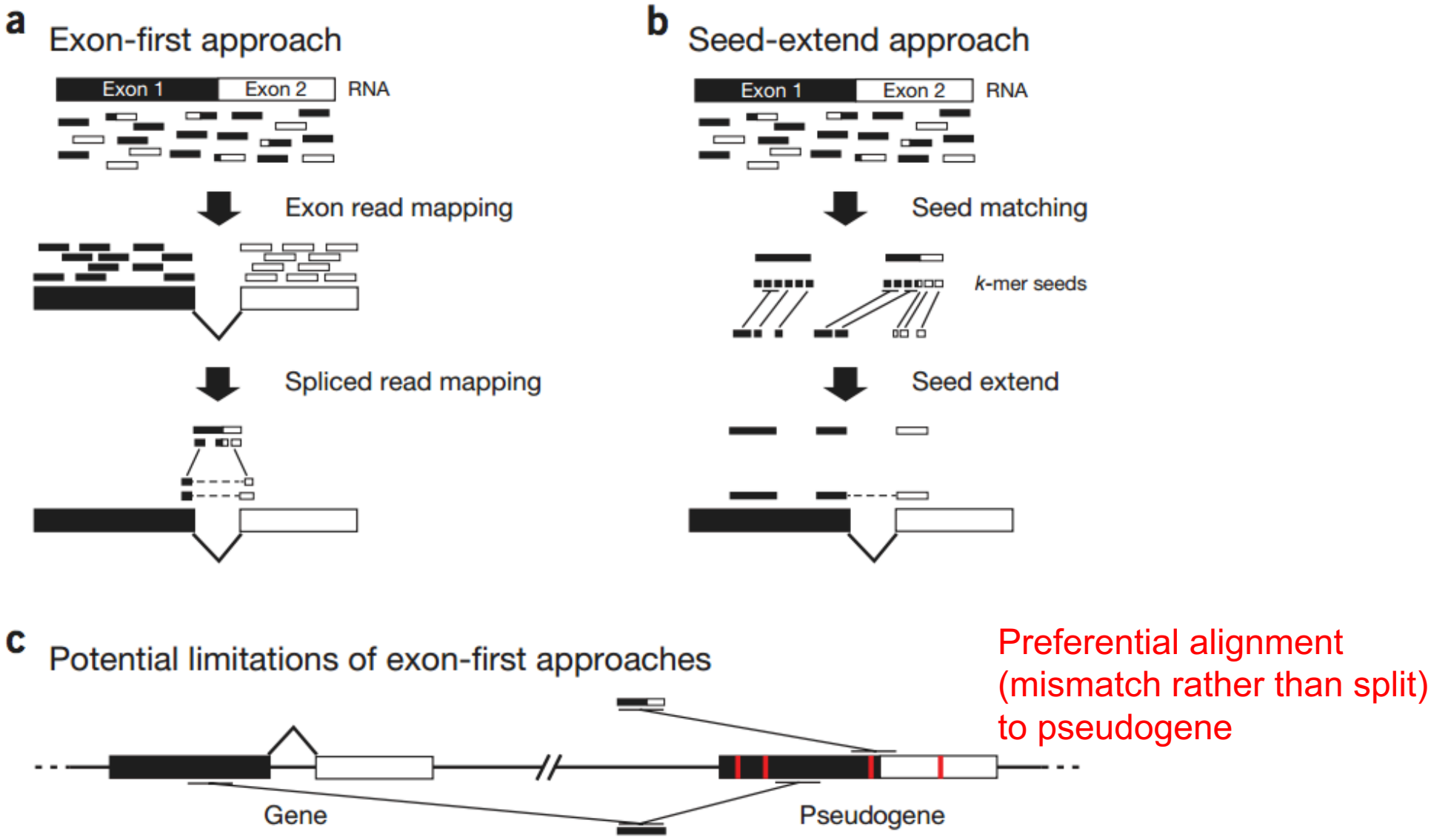


Note about the hardware and mapping software:

+ Bigger is the reference, more memory the programs needs
  (example: Bowtie2 ~2.1 Gb for human genome with 3 Gb)
+ Longer are the reads, more time the program needs for the mapping.

RNAseq aligners

Strategies for gapped alignments of RNAseq reads



Preferential alignment (mismatch rather than split) to pseudogene

Garber et al (2010)

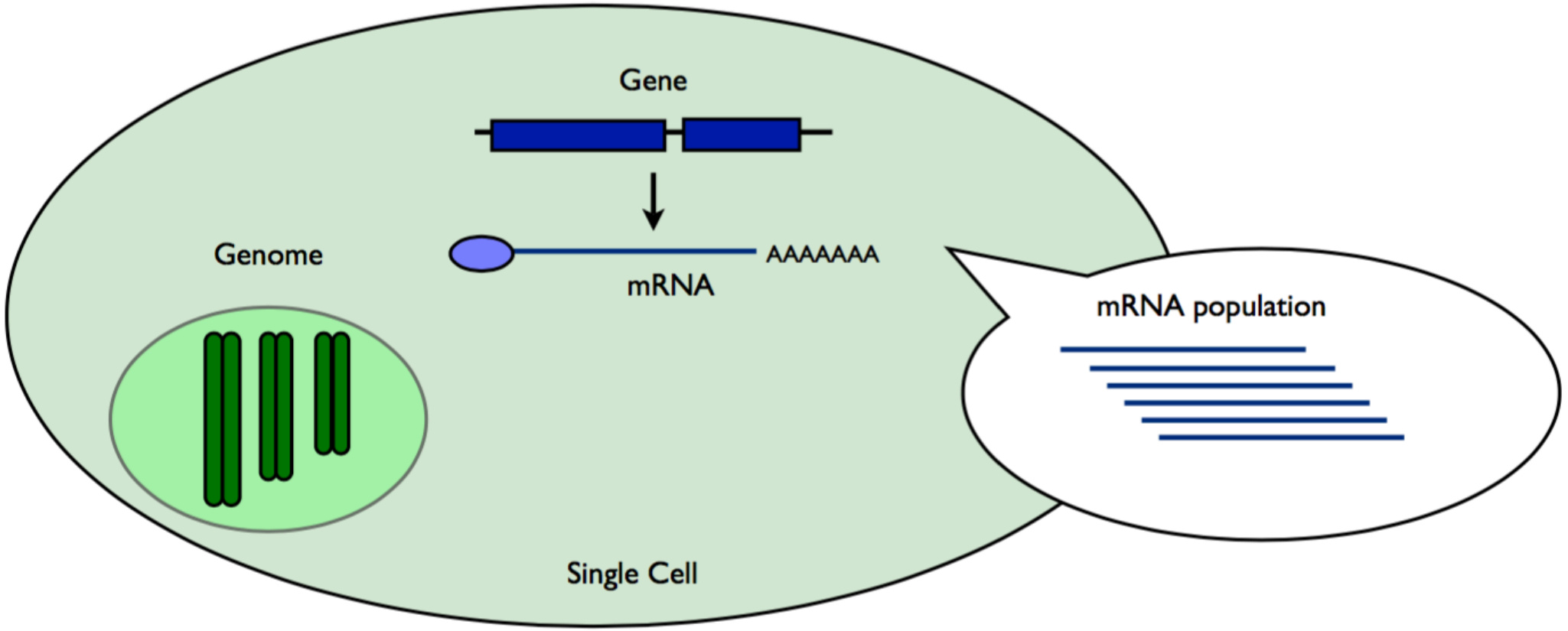
Differential expression

Types of experiments

Transcriptome Complexity:

Simple System:

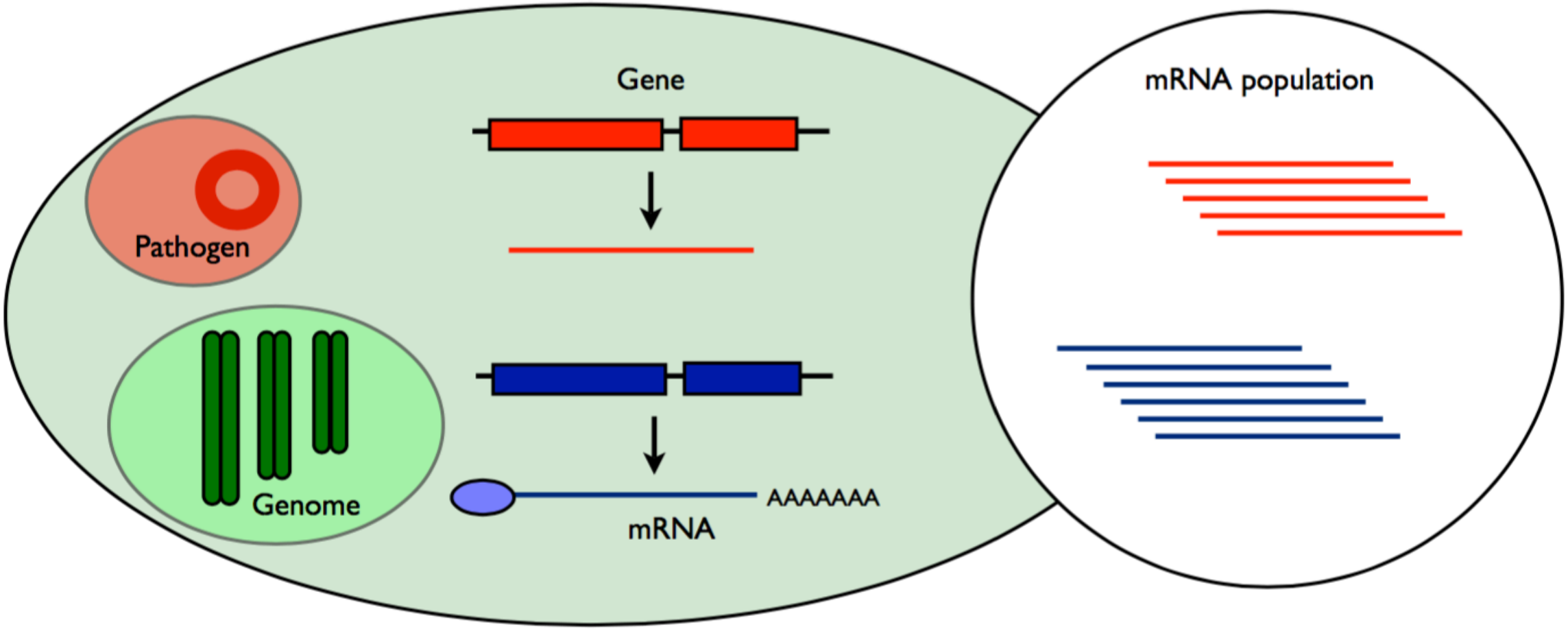
One Genome => Gene 1 copy => Single mRNA



How many species we are analyzing ?

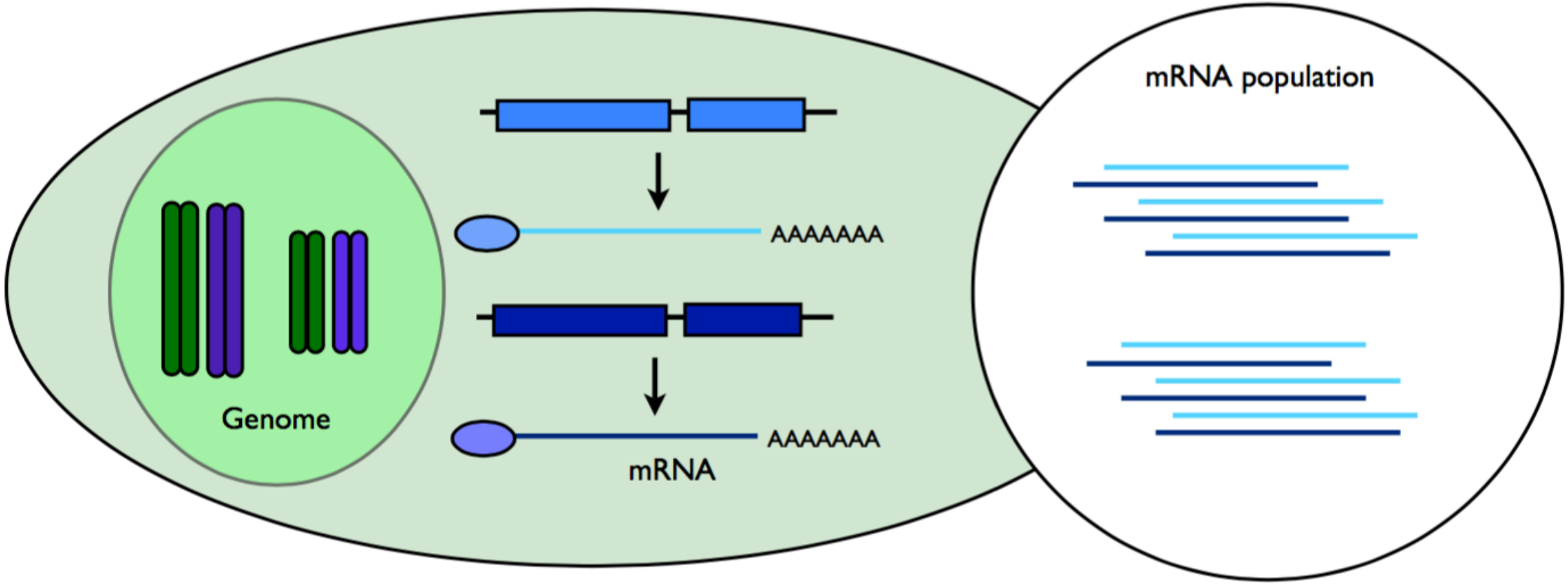
1) Problems to isolate a single species (rhizosphere)

2) Species interaction study (plant-pathogen)



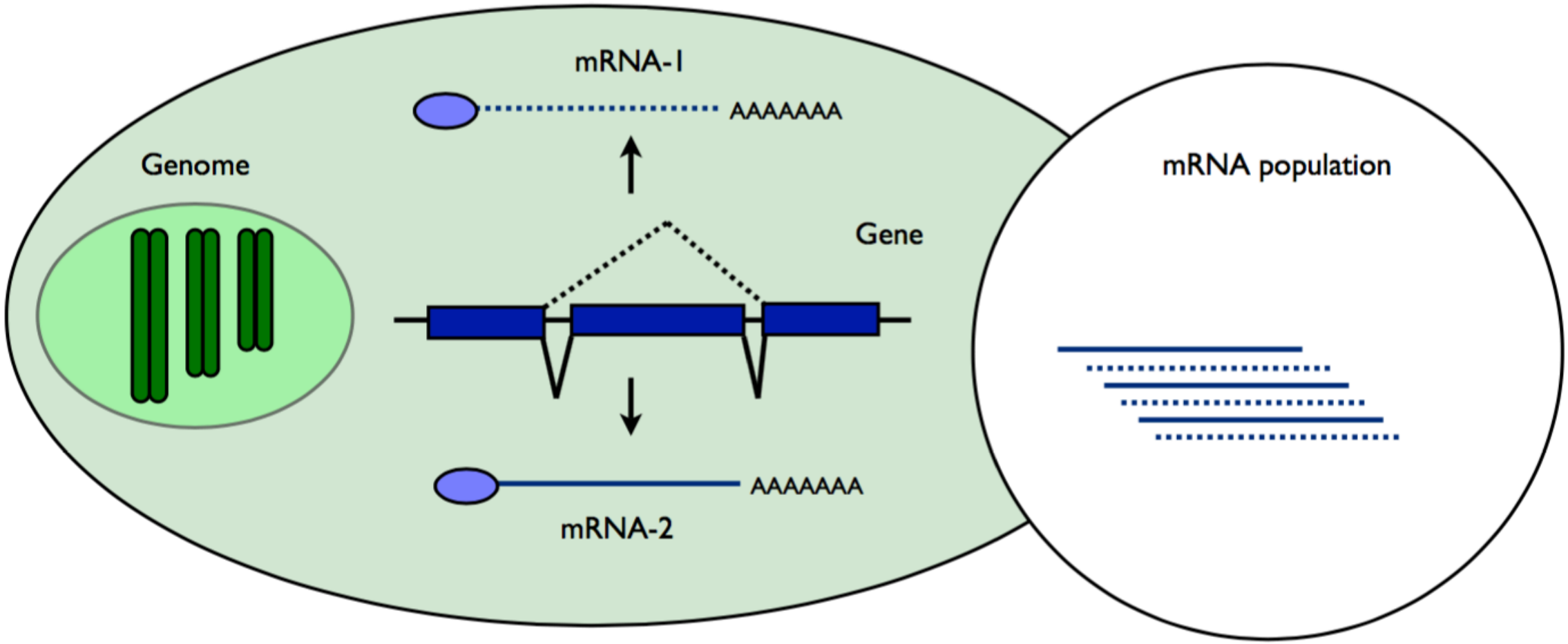
How many possible alleles we expect per gene ?

1) **Polyploids (autopolyploids, allopolyploids).**
2) **Heterozygosity**
3) **Complex Gene Families (tandem duplications)**



How many isoforms we expect for each allele ?

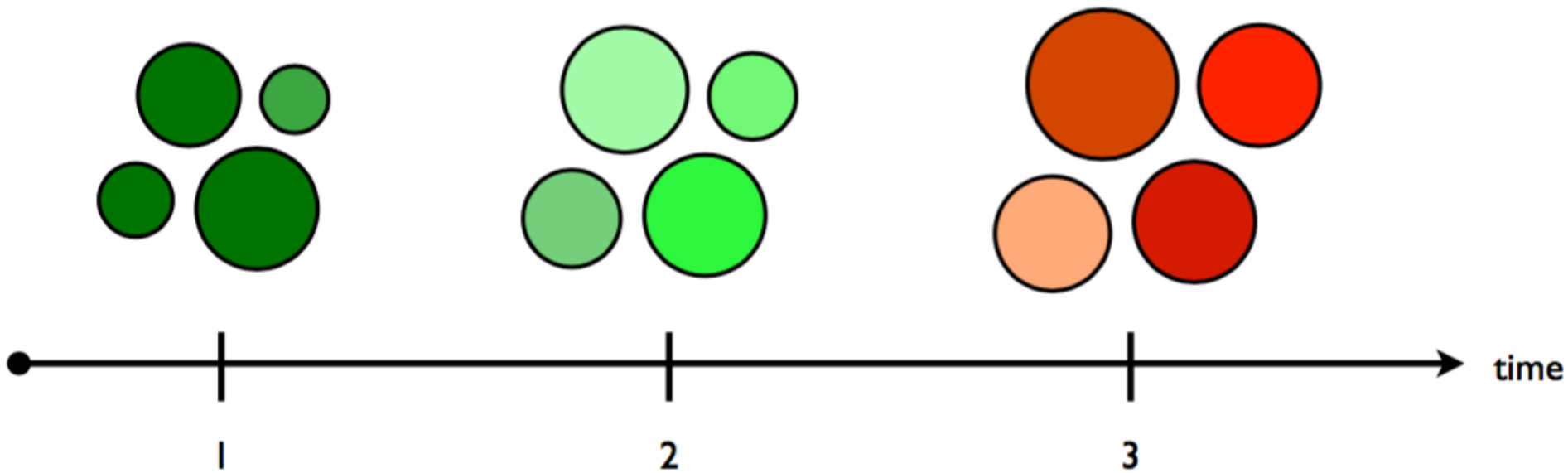
1) Alternative splicings



Is the study performed at different time points?

1) Developmental stages (difficult to select the same)

2) Response to a treatment

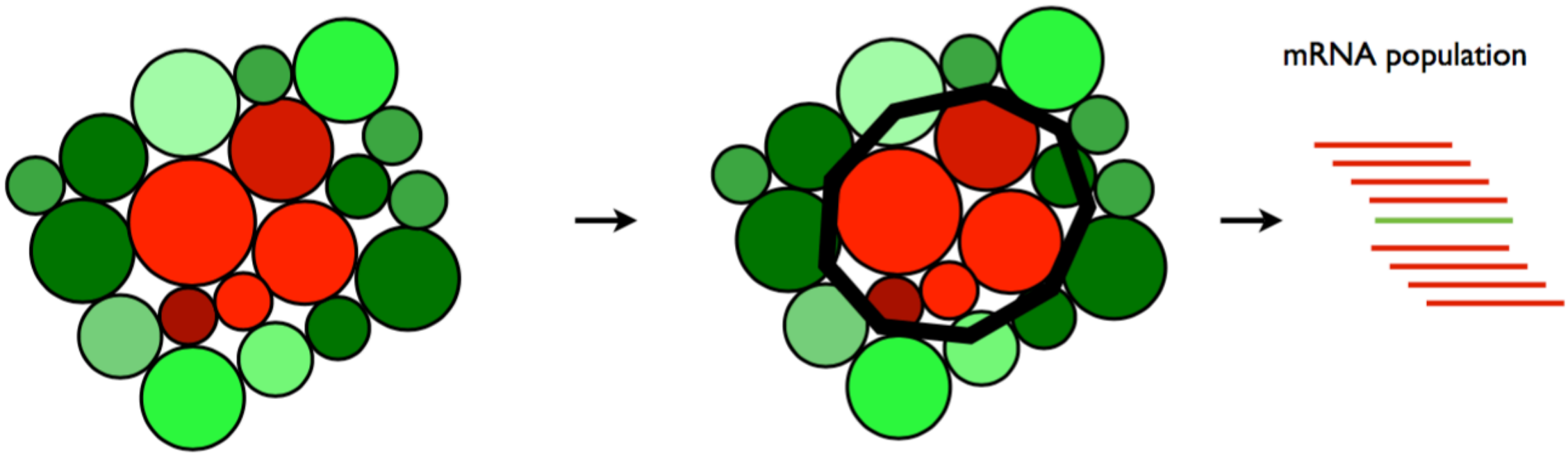


Is the study performed with different parts?

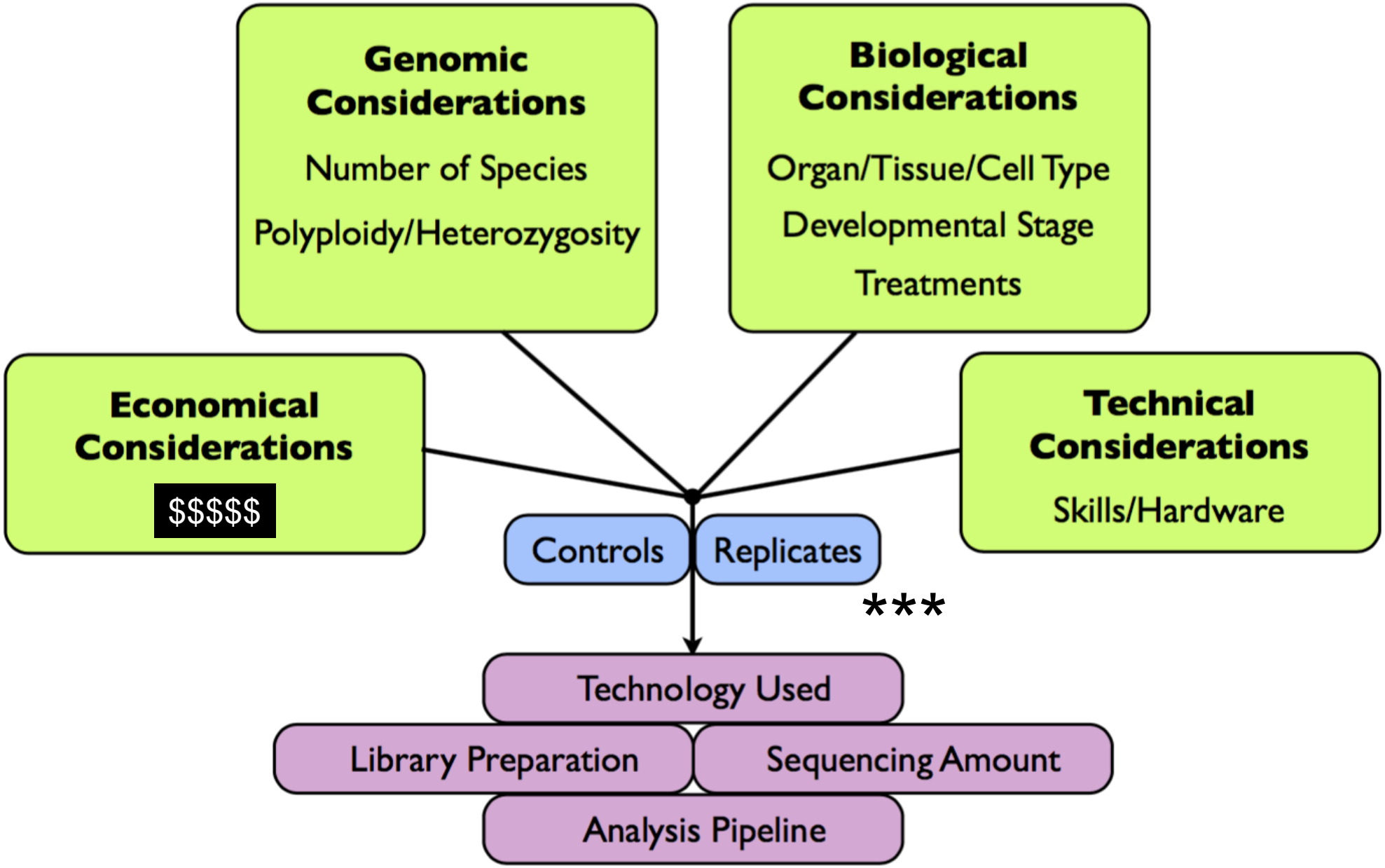
1) Organ specific

2) Tissue/Cell type specific

(Laser Capture Microdissection, LCM)



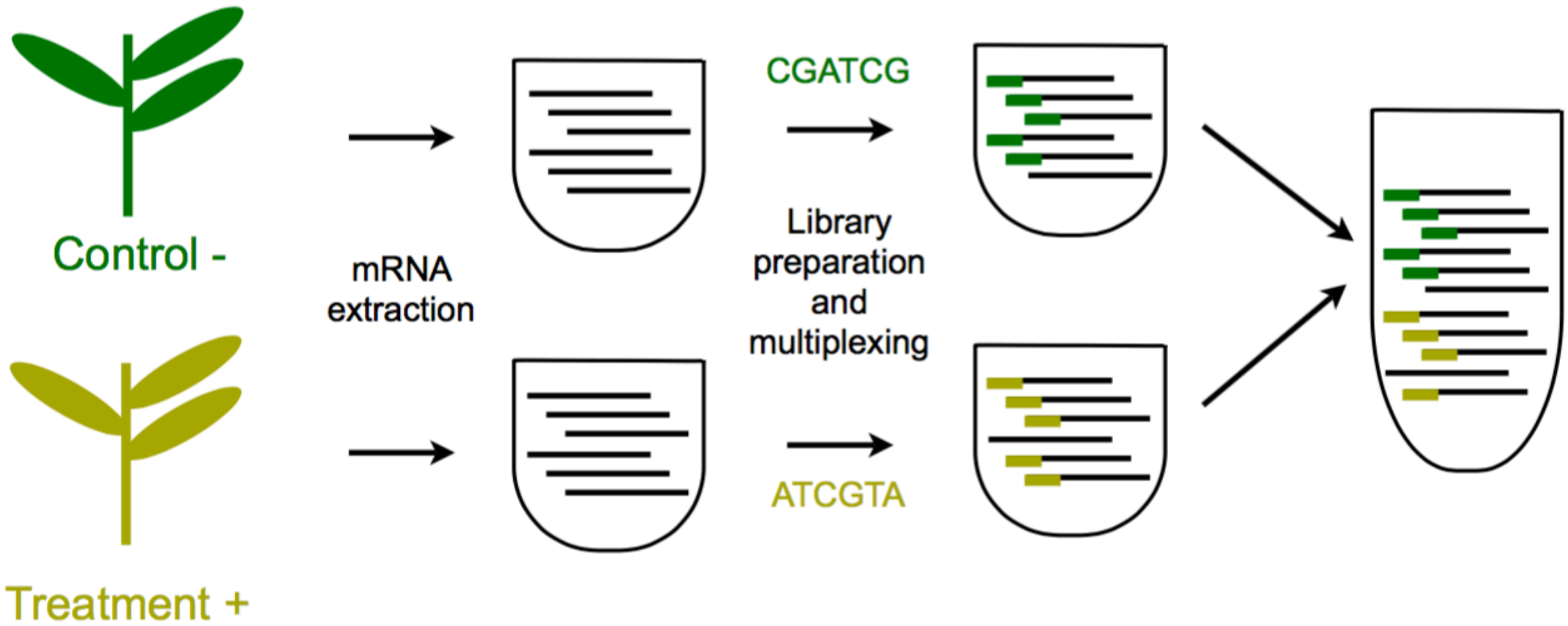
Experimental design



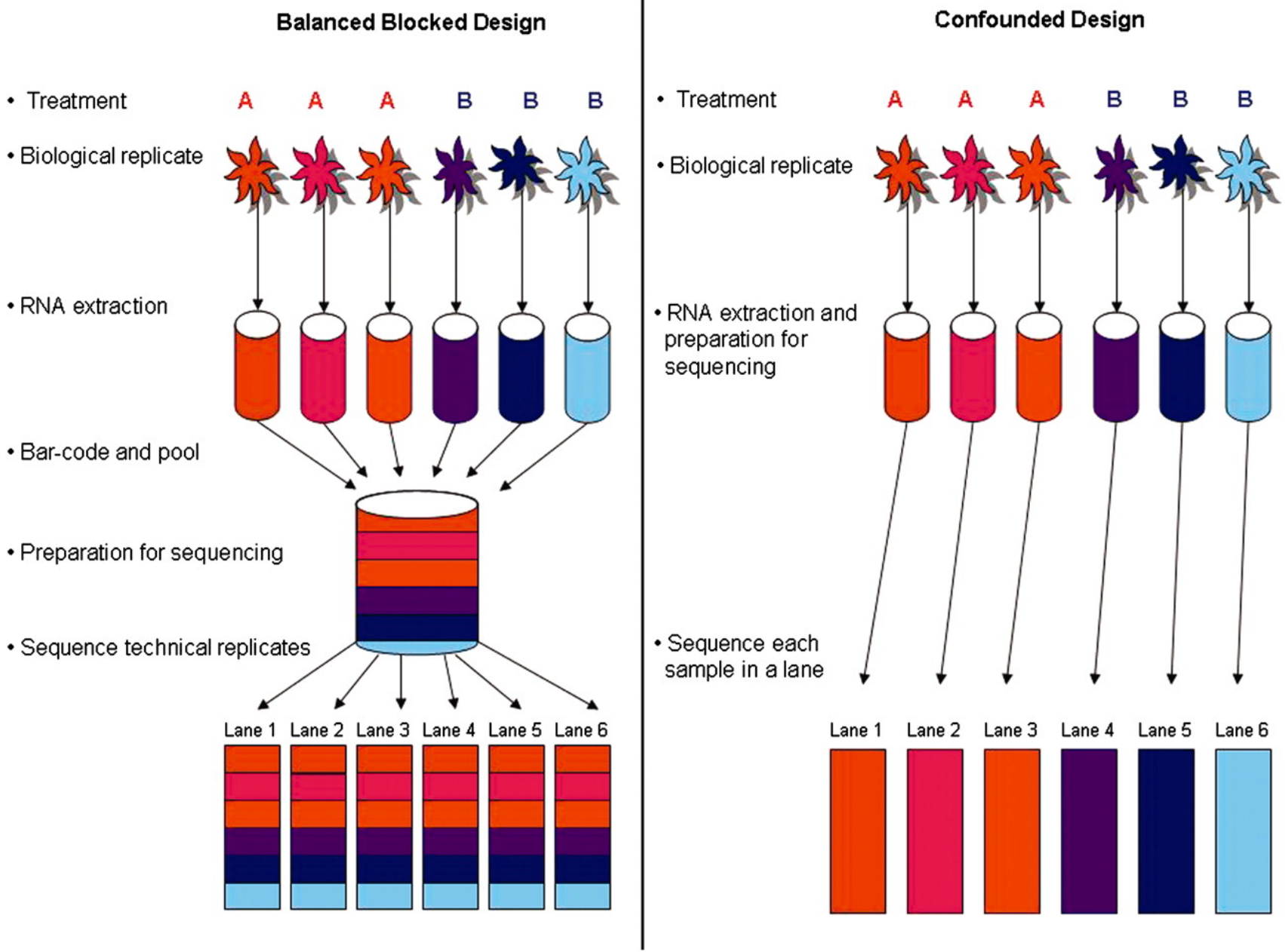
Prep and treatment

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

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

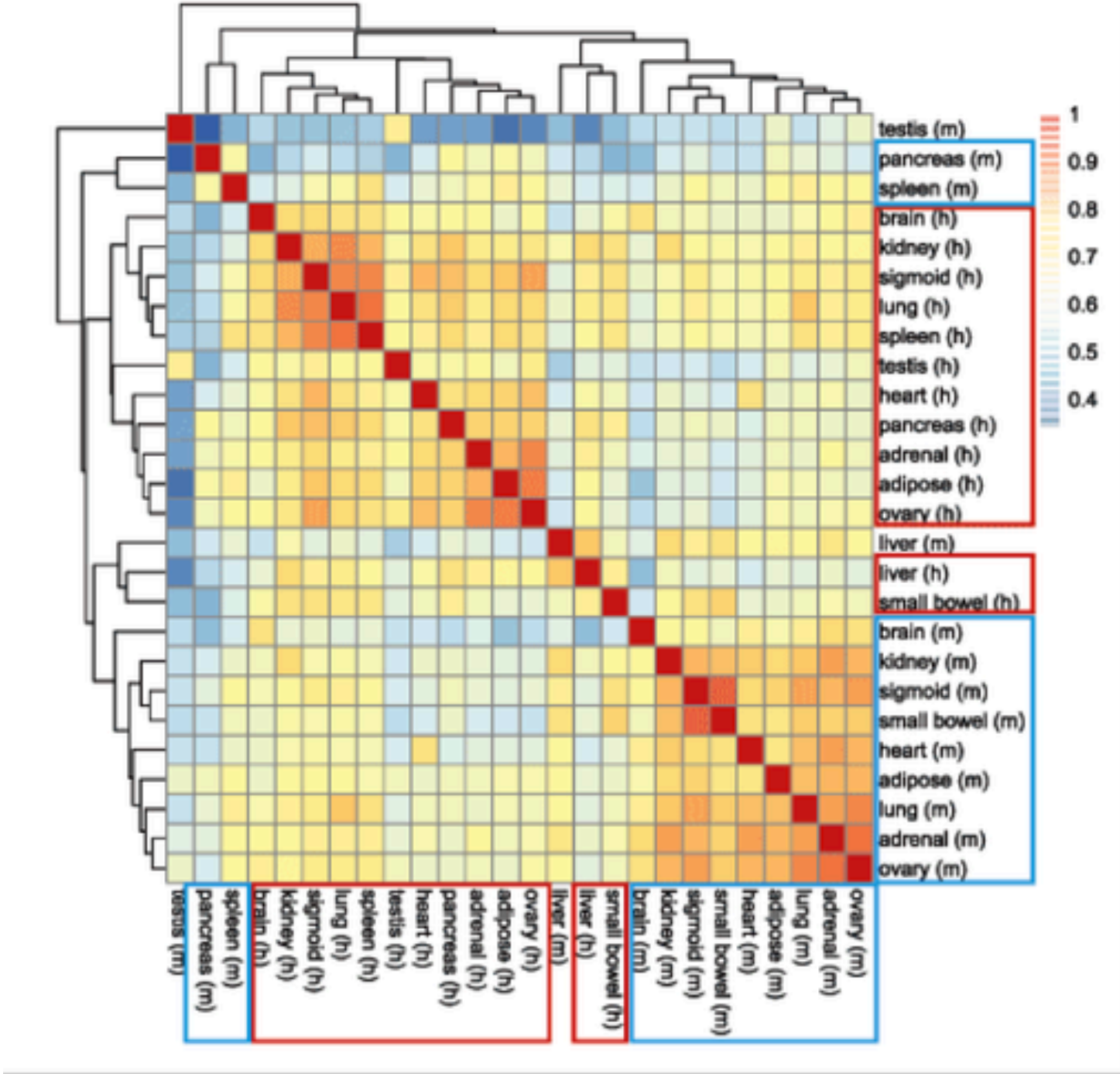


You need to design experiment carefully



Auer *et al.,* Genetics (2010)

Example of batch effect:



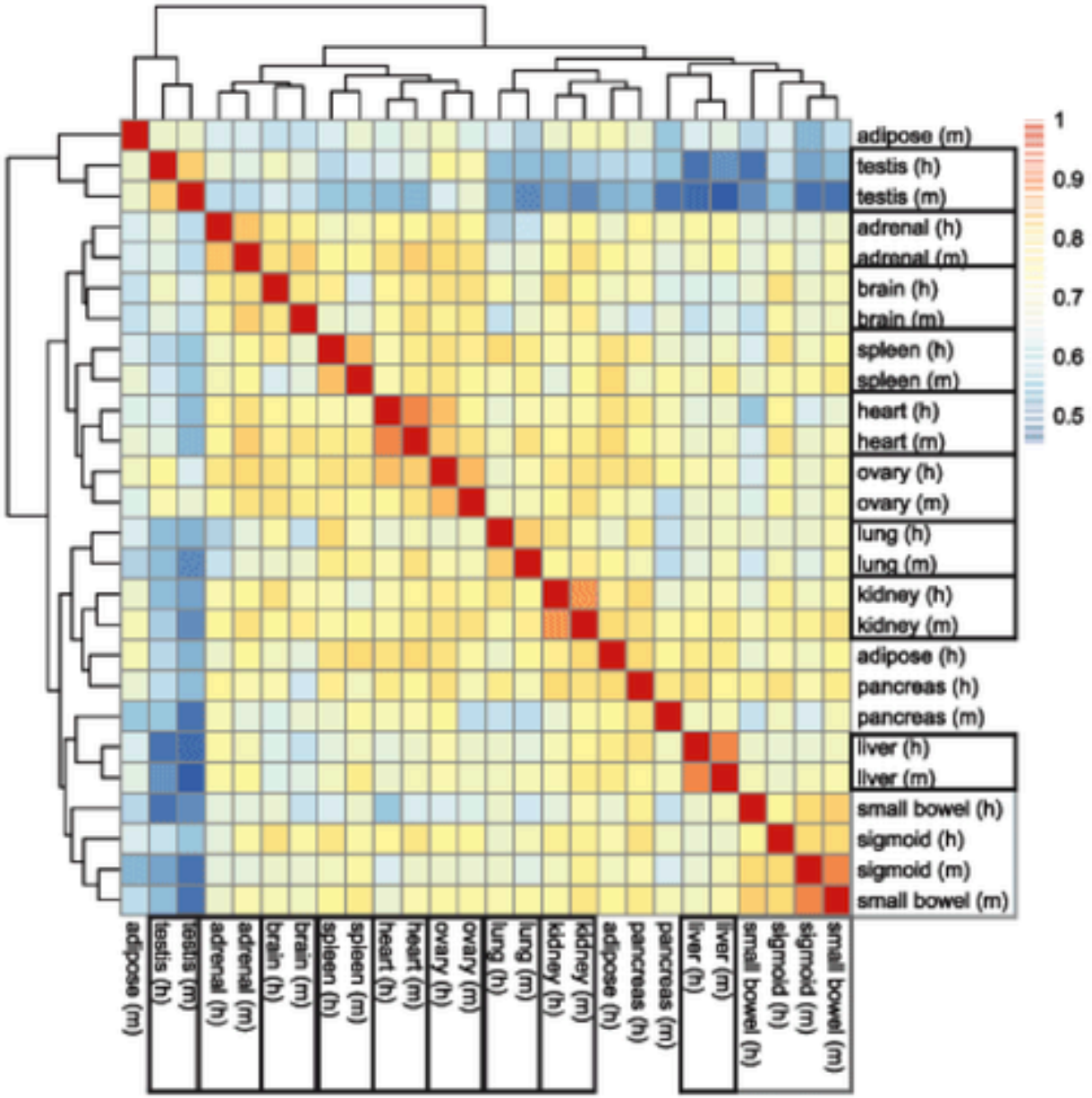
Lin *et al.,* (PNAS) 2014

Example of batch effect:

Yoav Gilad
@Y_Gilad

Following

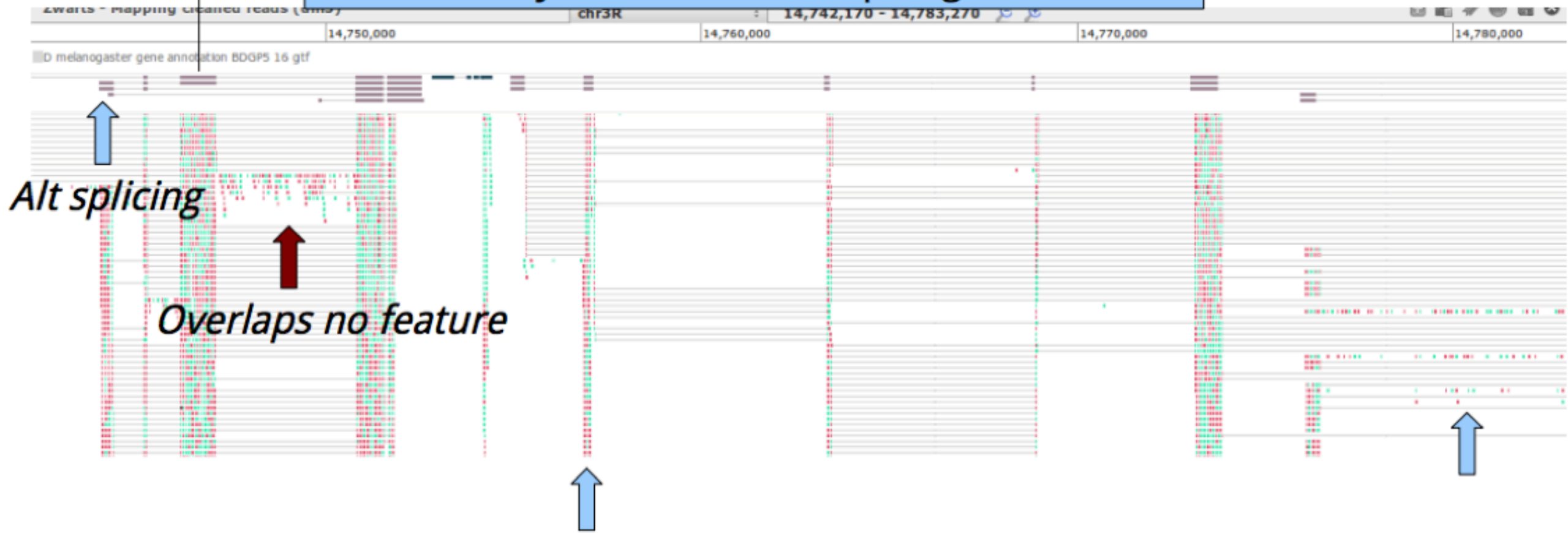
We reanalyzed the data from pnas.org/content/111/48… and found the following:



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

Once you have mappings, you can start counting

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



Concept:

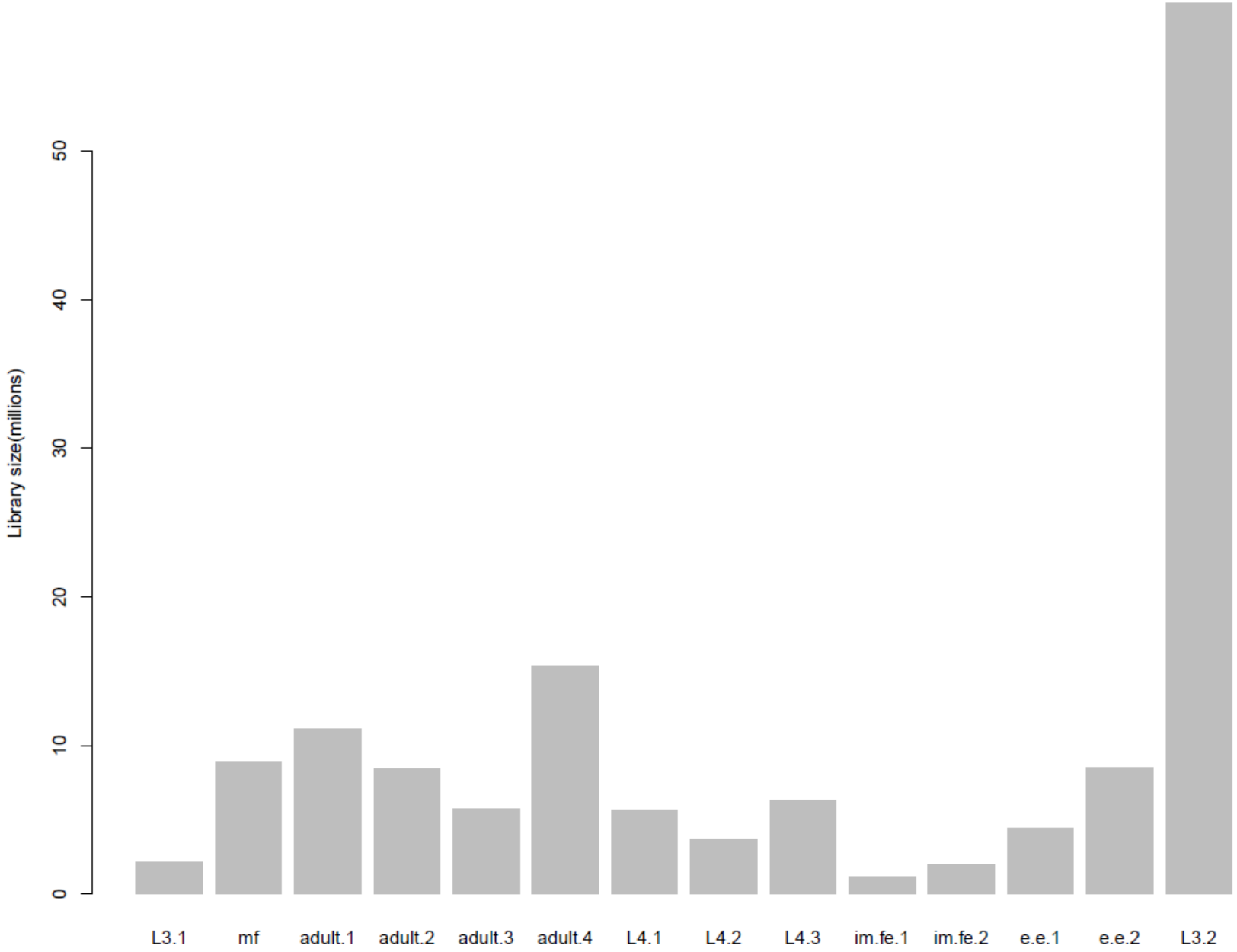
GeneA = exon 1 + exon 2 + exon 3 + exon 4 = 215 reads

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

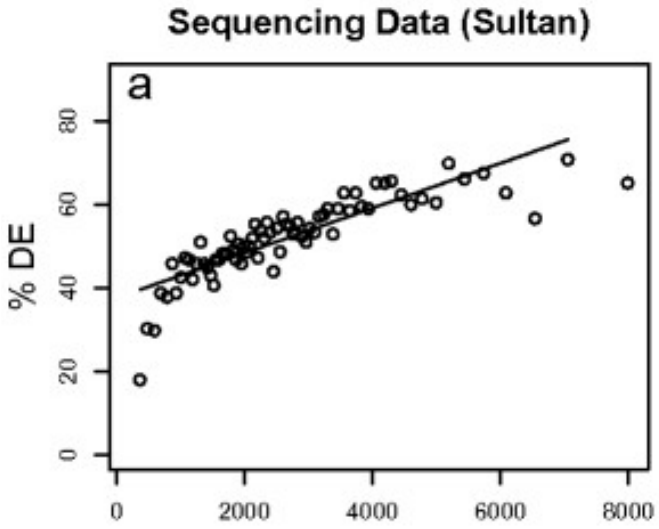
This is the bit we care about!

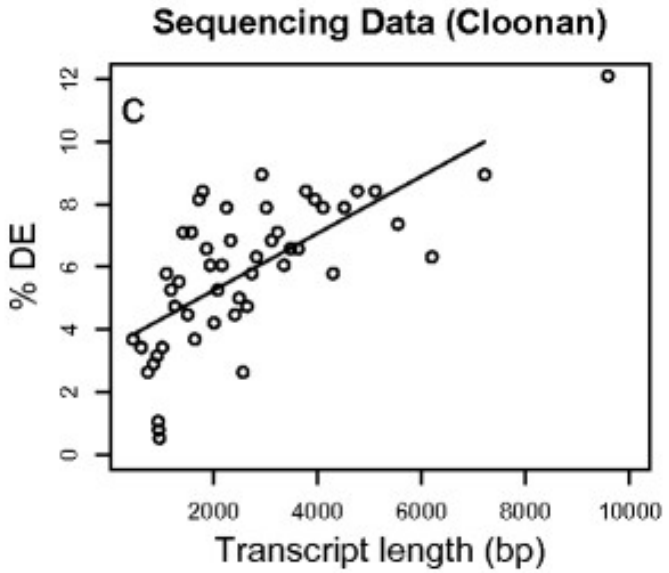
Counts of the gene depends on **expression** ,transcript length ,sequencing depth and simply chance

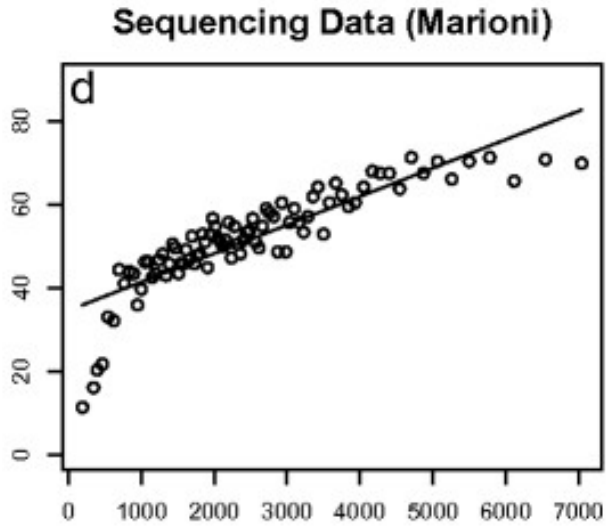
Higher sequencing depth equals more counts

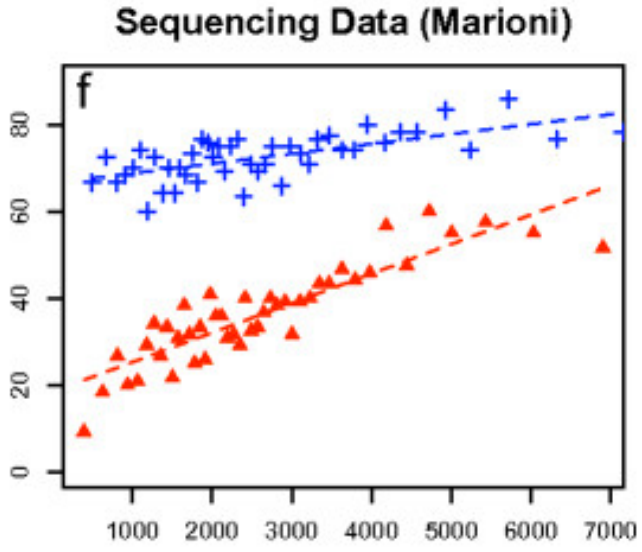


Counts are proportional to the transcript length x mRNA expression level









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

Oshlack and Wakefield (2009) Biology Direct.

Normalization: different goals

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

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

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

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

Optimal Scaling of Digital Transcriptomes

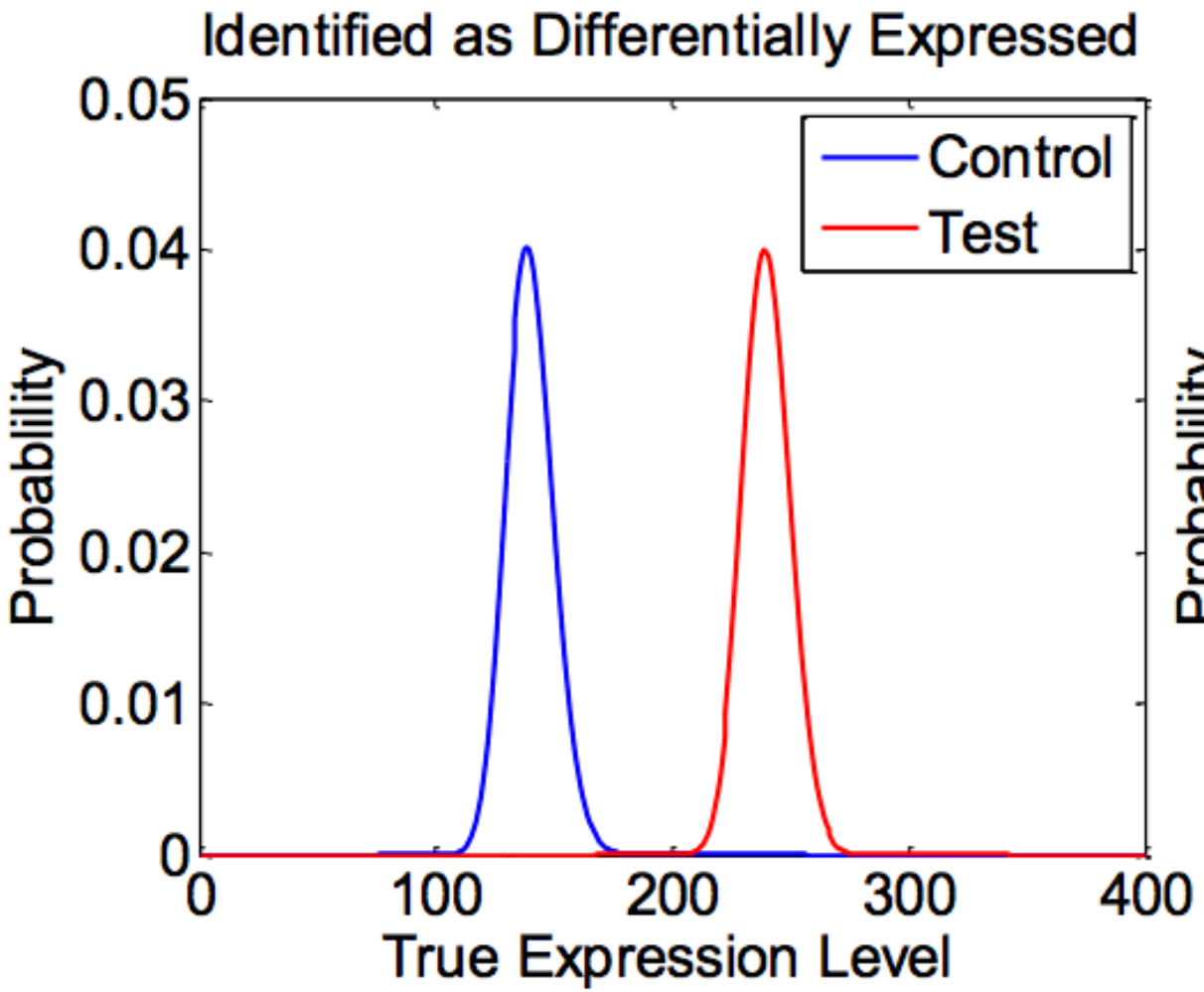
Gustavo Glusman, Juan Caballero, Max Robinson, Burak Kutlu, Leroy Hood

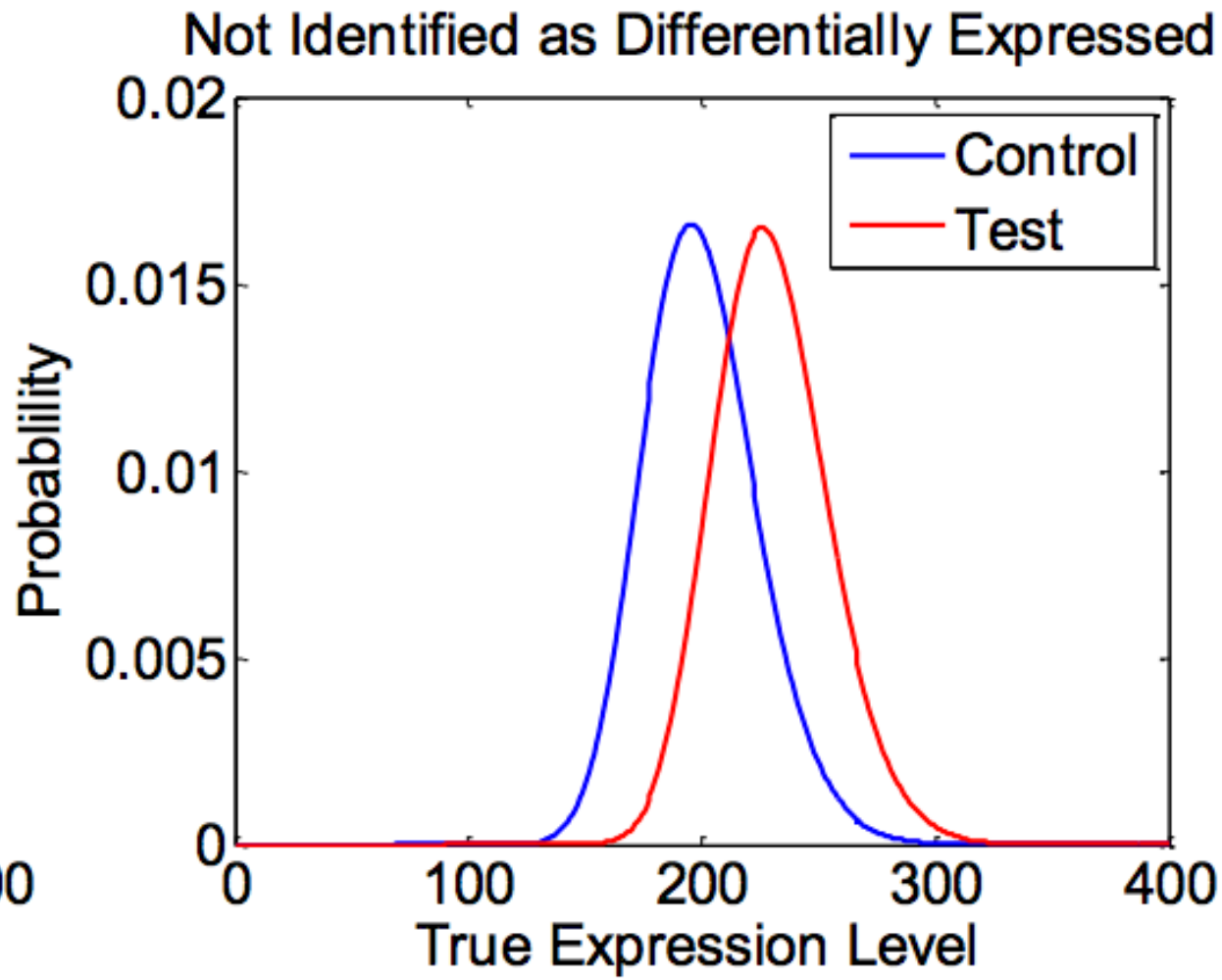
Published: Nov 06, 2013 • DOI: 10.1371/journal.pone.0077885

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

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

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

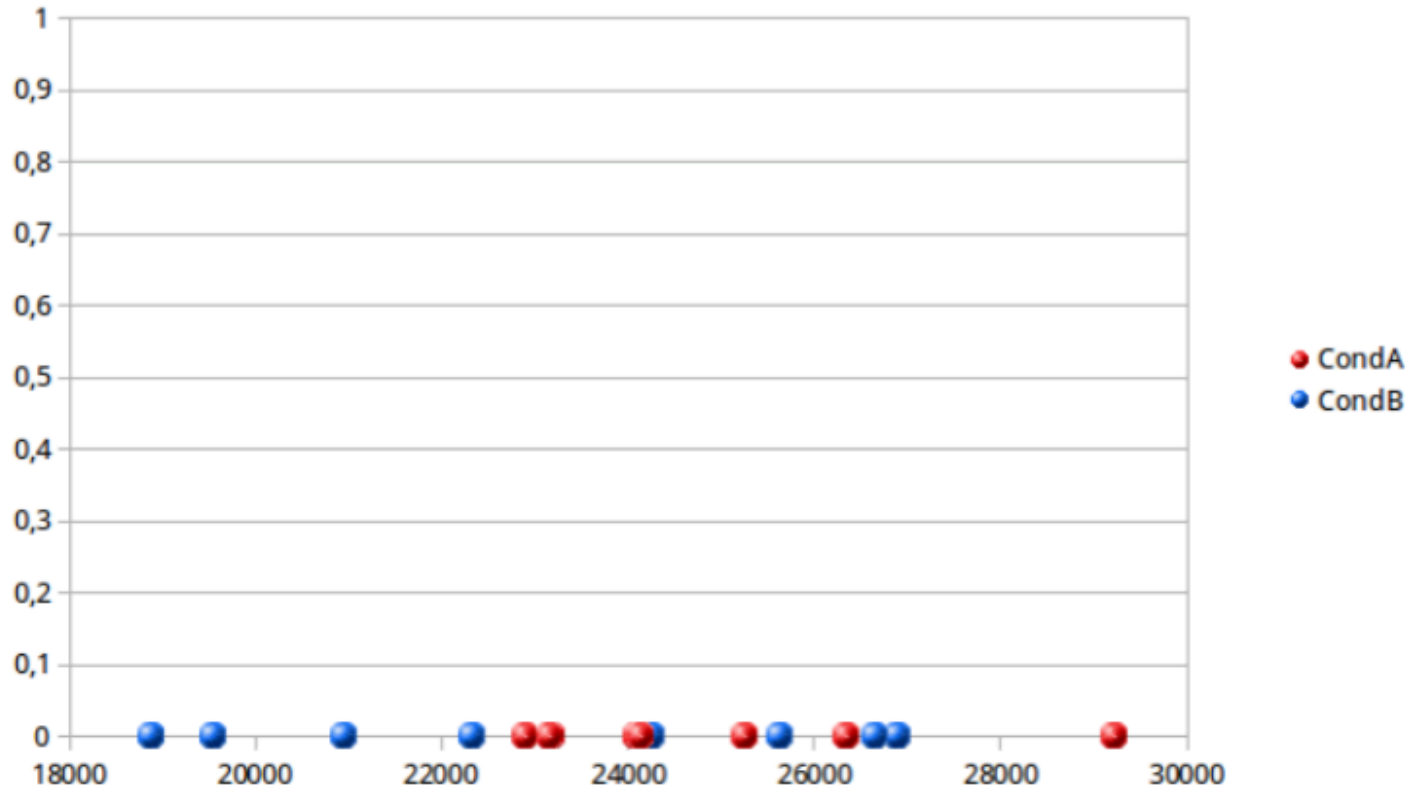




. Busby *et al.,* Bioinformatics (2013)

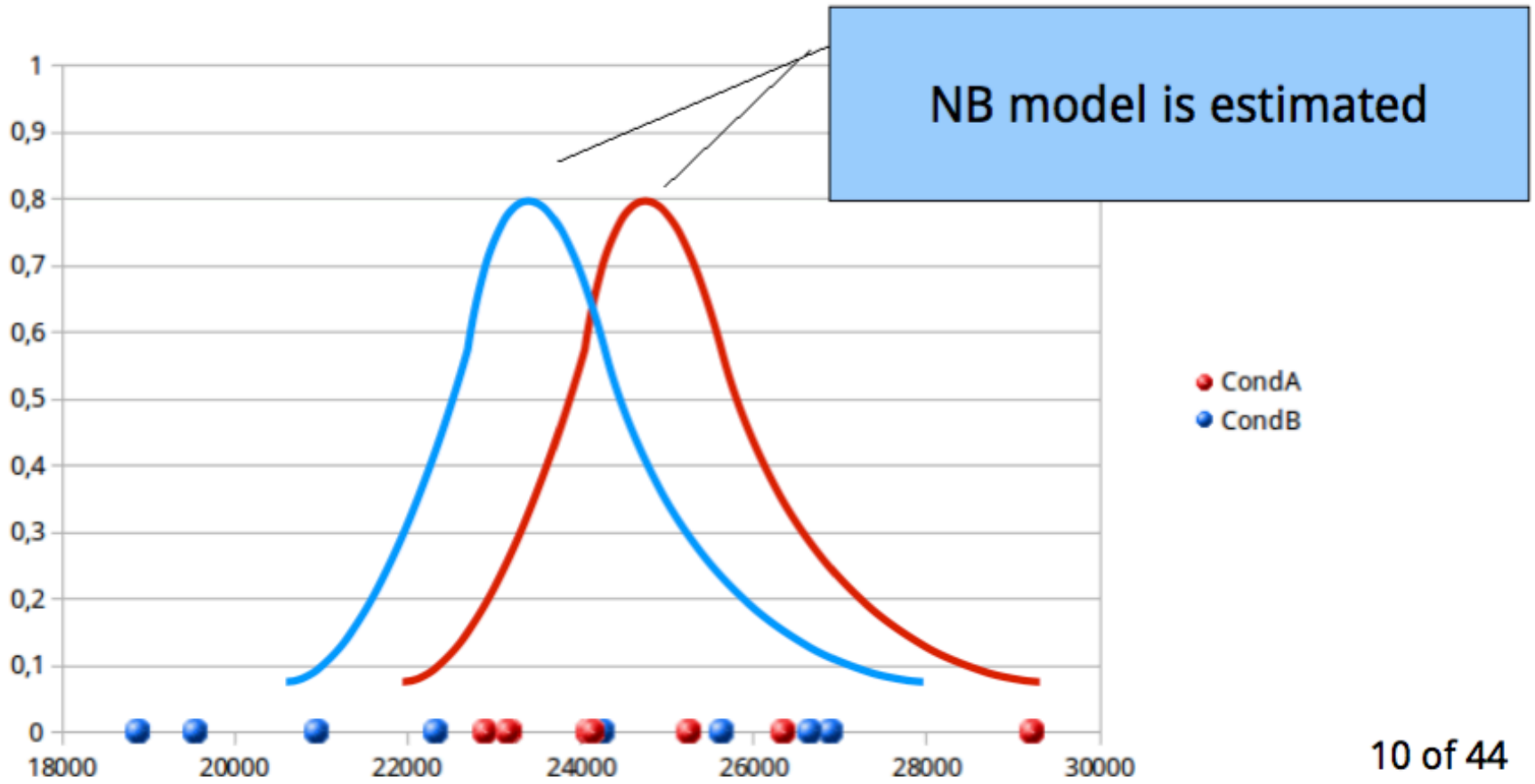
Scenario

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



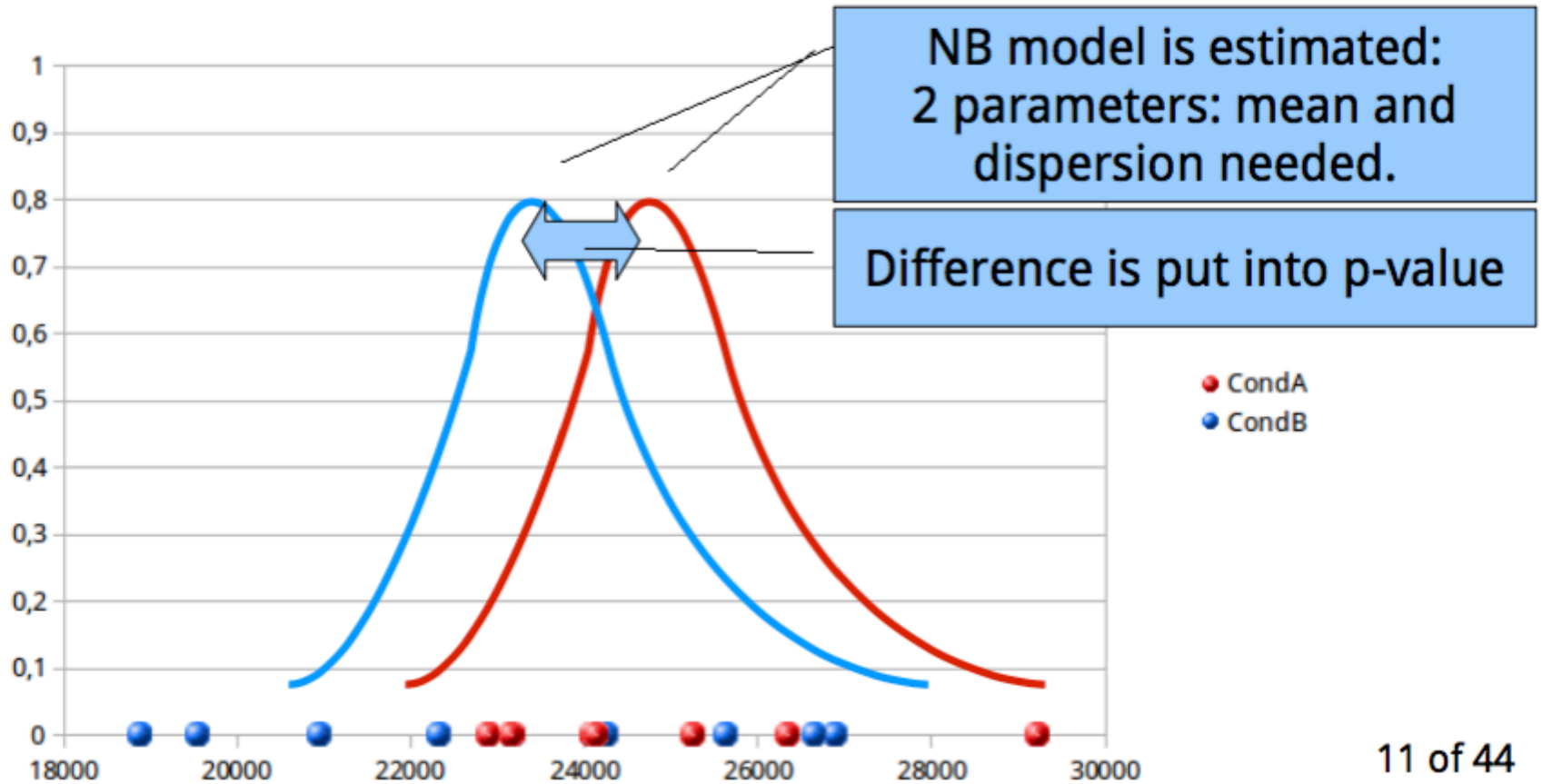
Scenario

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

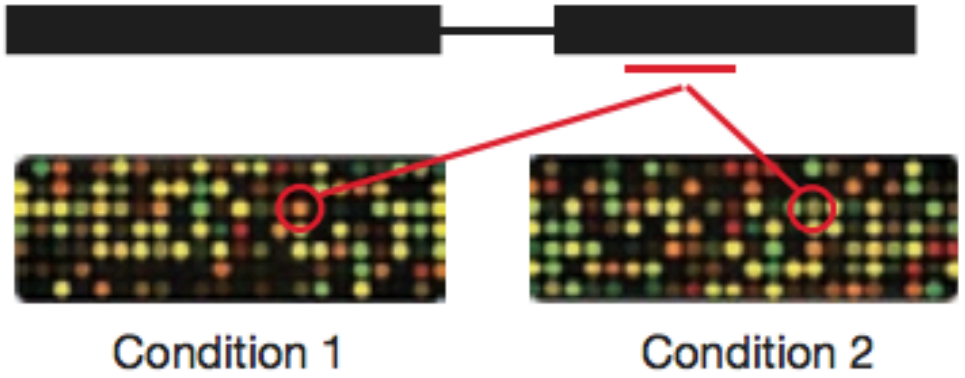


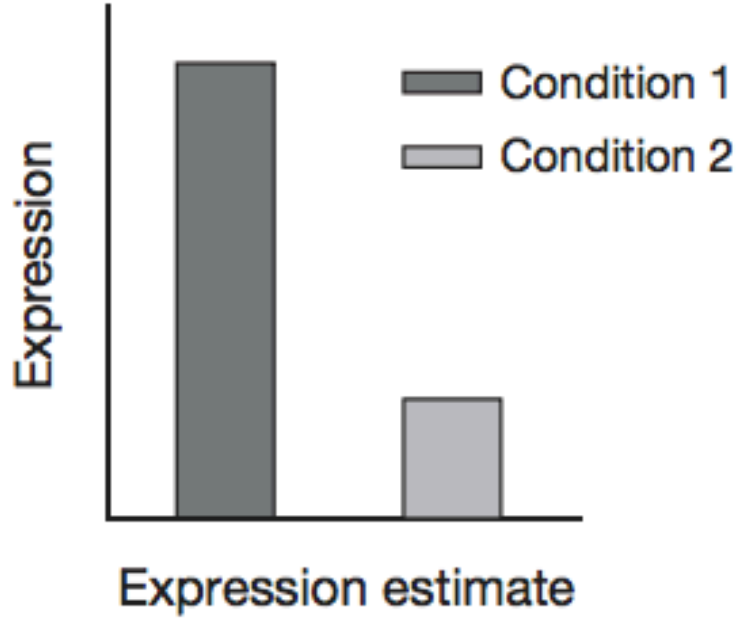
Scenario

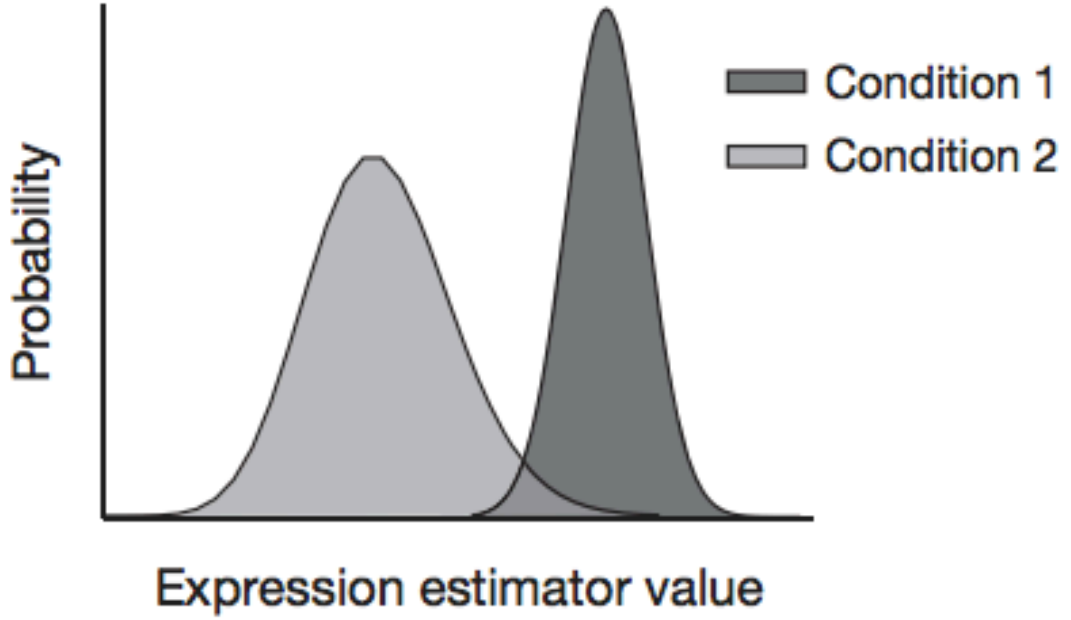
gene_id	CAF0006876							
Condition A	sample1 23171	sample2 22903	sample3 29227	sample4 24072	sample5 23151	sample6 26336	sample7 25252	sample8 24122
Condition B	Sample9 19527	sample10 26898	sample11 18880	sample12 24237	sample13 26640	sample14 22315	sample15 20952	sample16 25629



RNAseq vs Microarray

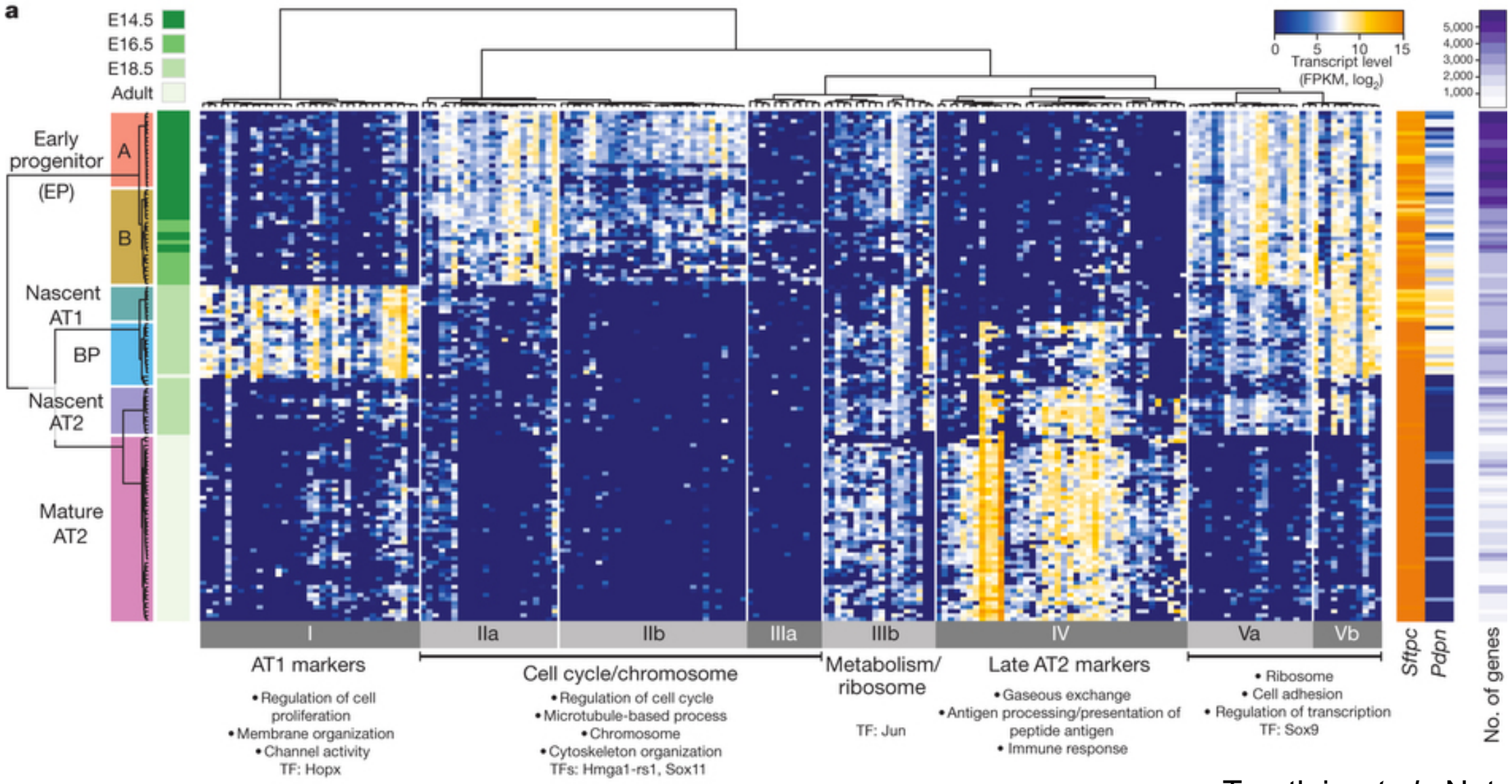






Once you have set of differentially expressed genes

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



Treutlein *et al.,* Nature (2014)

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

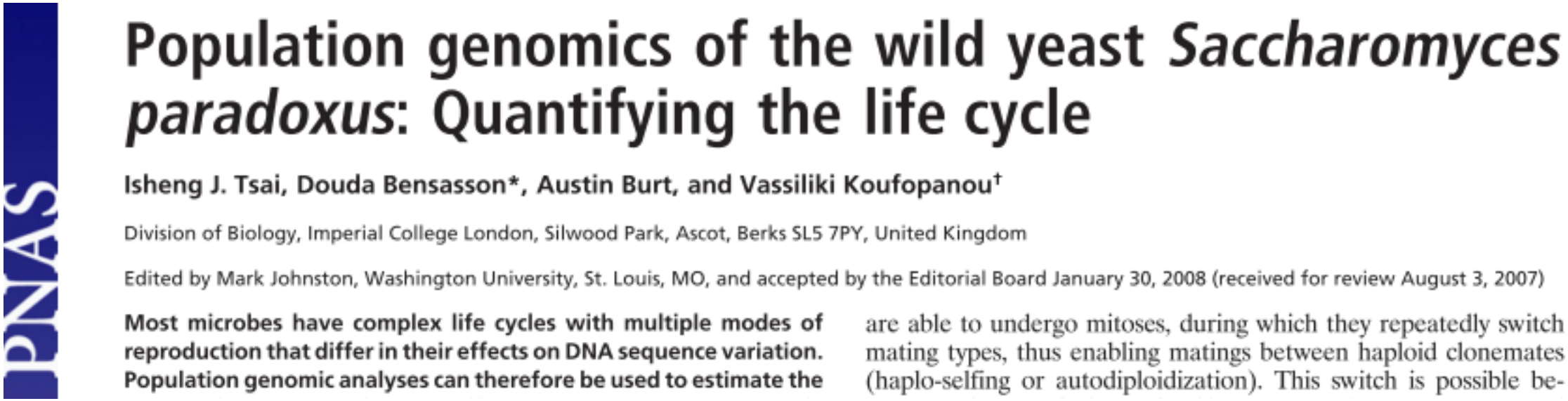
- A lof of omics not covered..
  - Single cell sequencing
  - New technologies
  - And much more..

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)



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

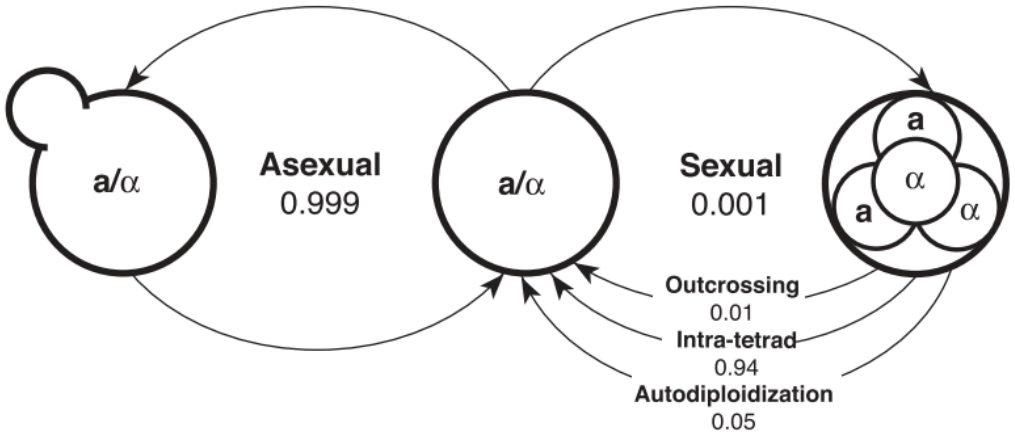
Edited by Mark Johnston, Washington University, St. Louis, MO, and accepted by the Editorial Board January 30, 2008 (received for review August 3, 2007)

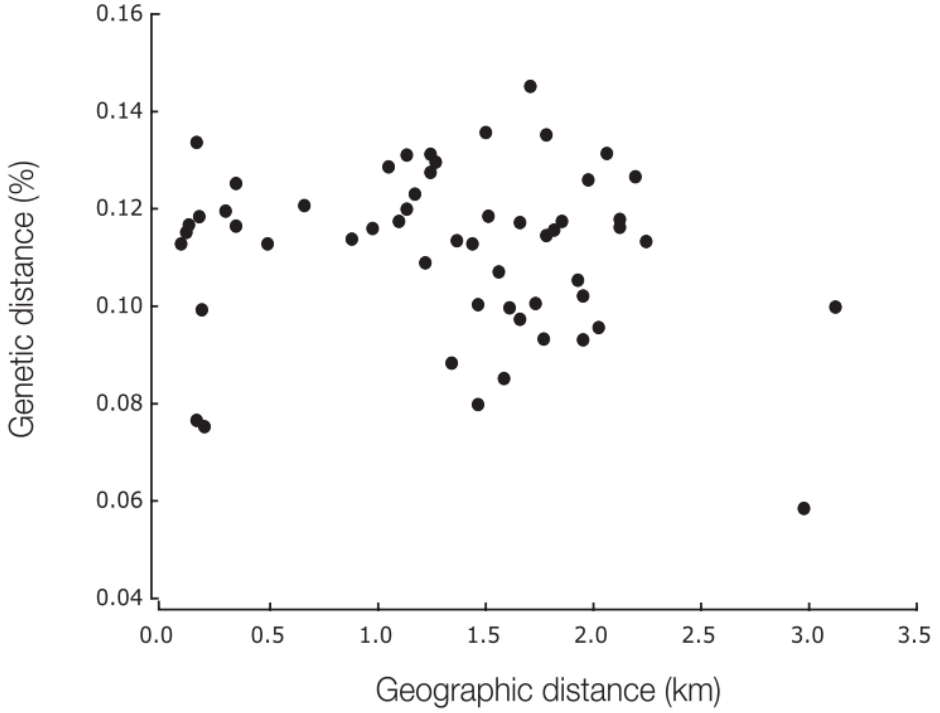
Most microbes have complex life cycles with multiple modes of reproduction that differ in their effects on DNA sequence variation. Population genomic analyses can therefore be used to estimate the

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)

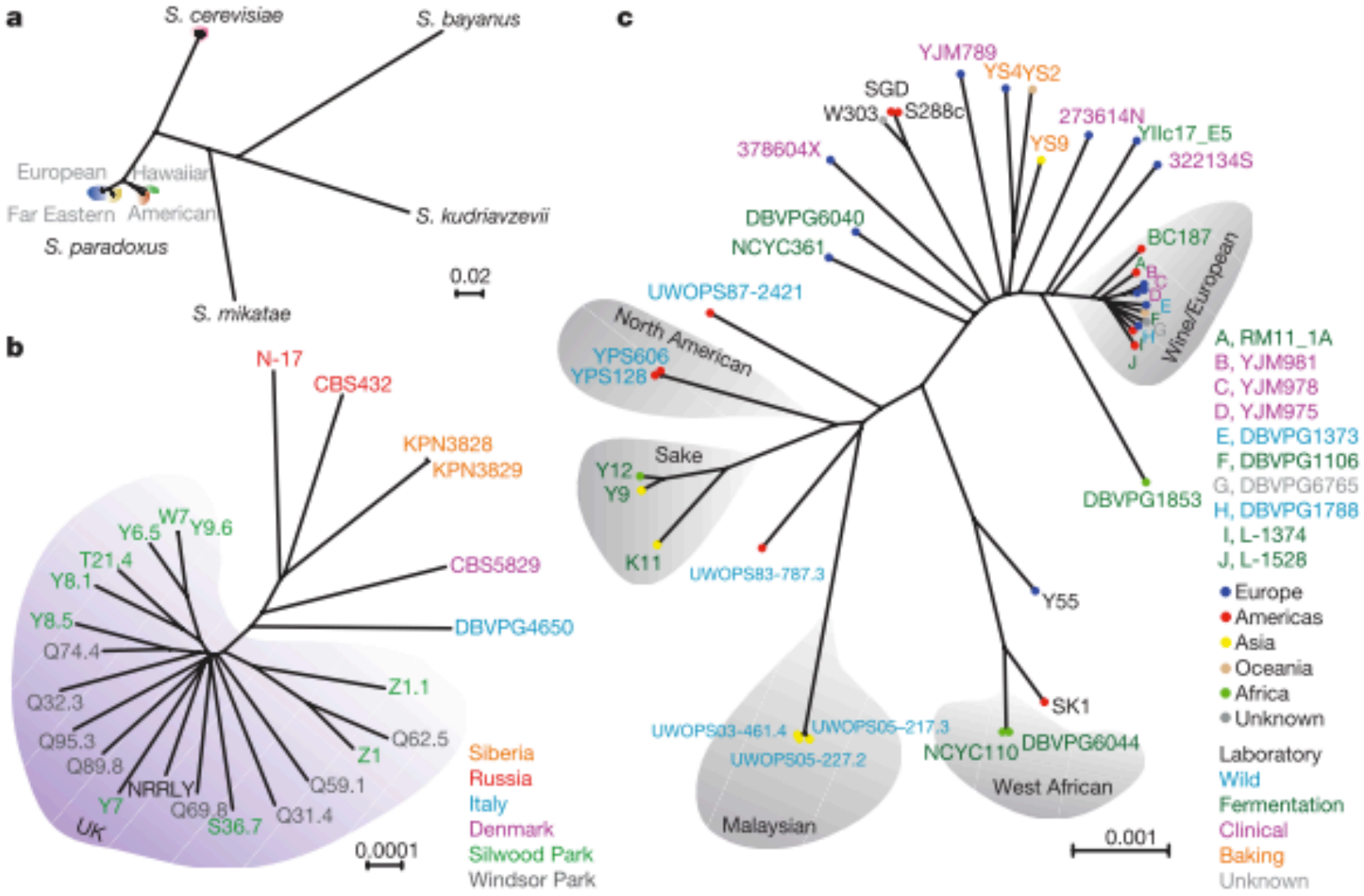
- **Now = this can probably be done in 3 months by my RA**

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

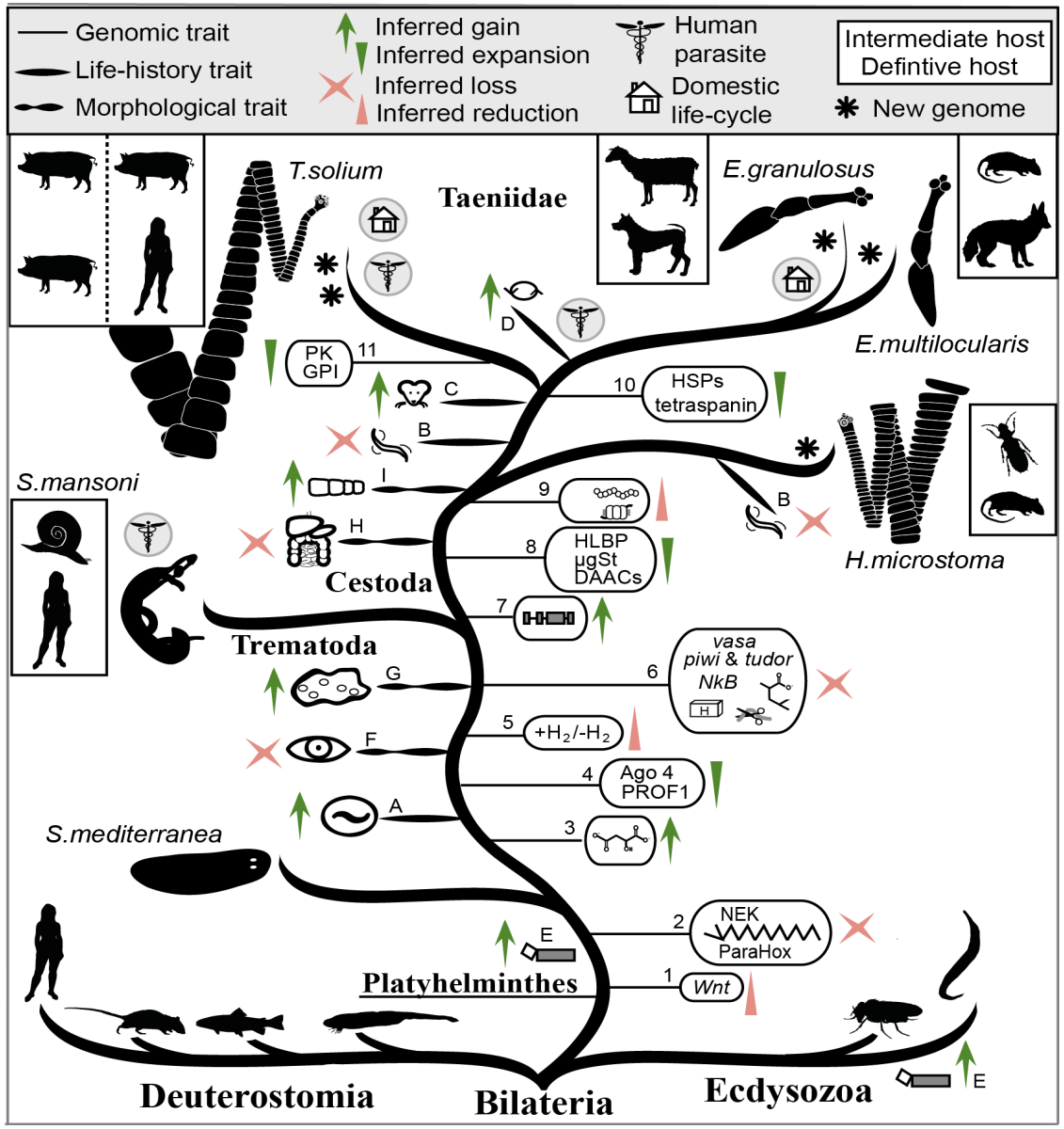
OPEN

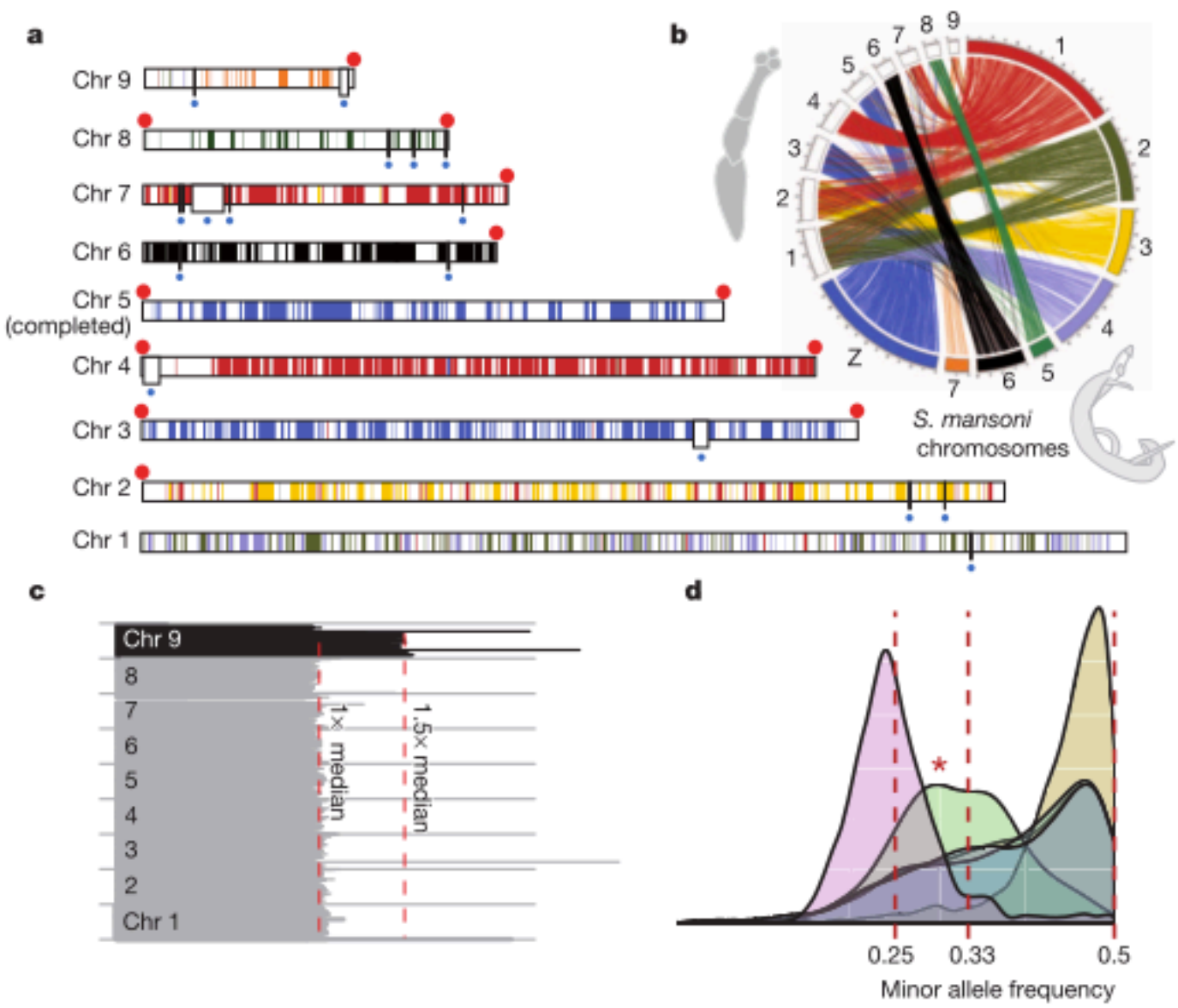
doi:10.1038/nature12031

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





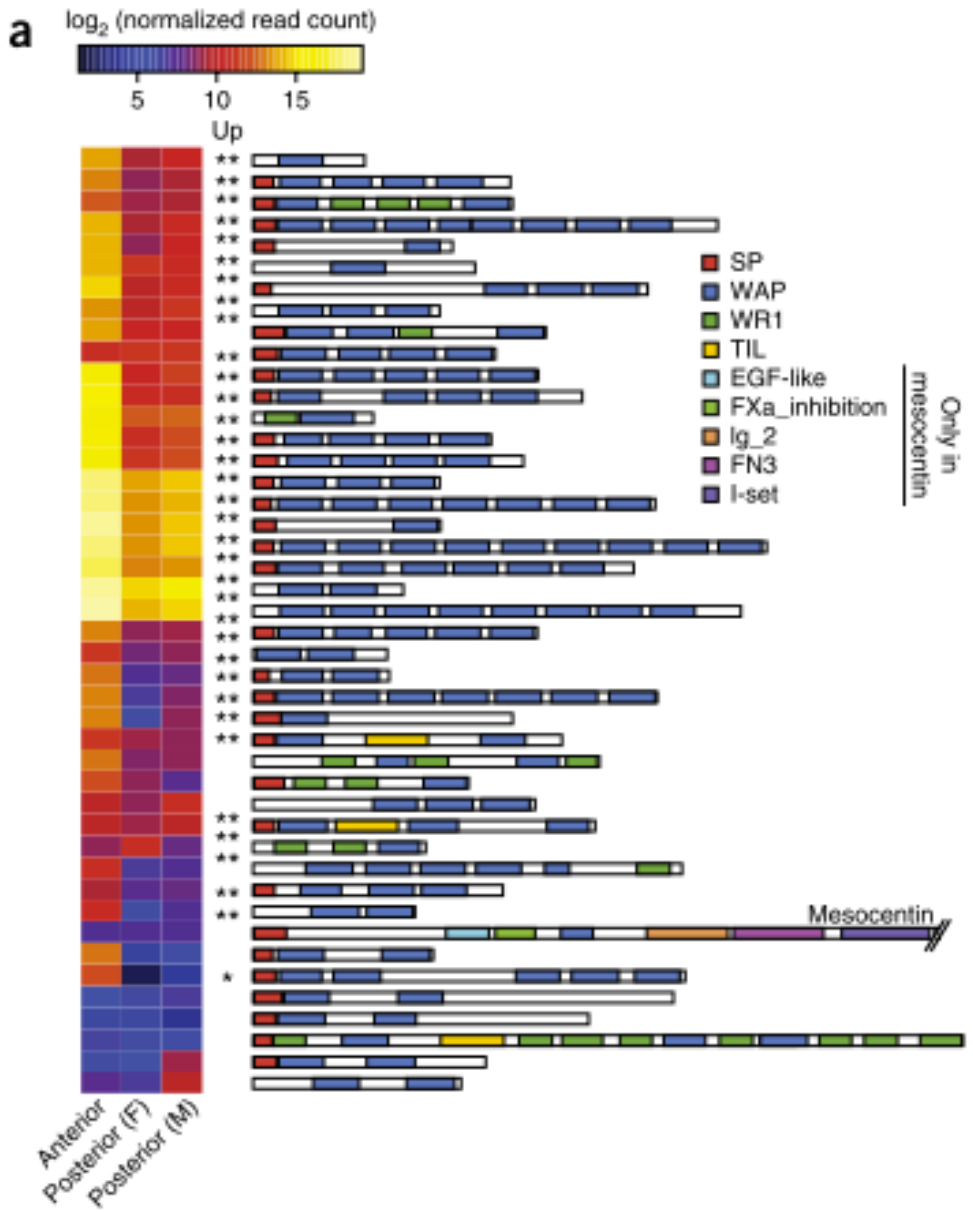
2014 – *Trichuris* genome project

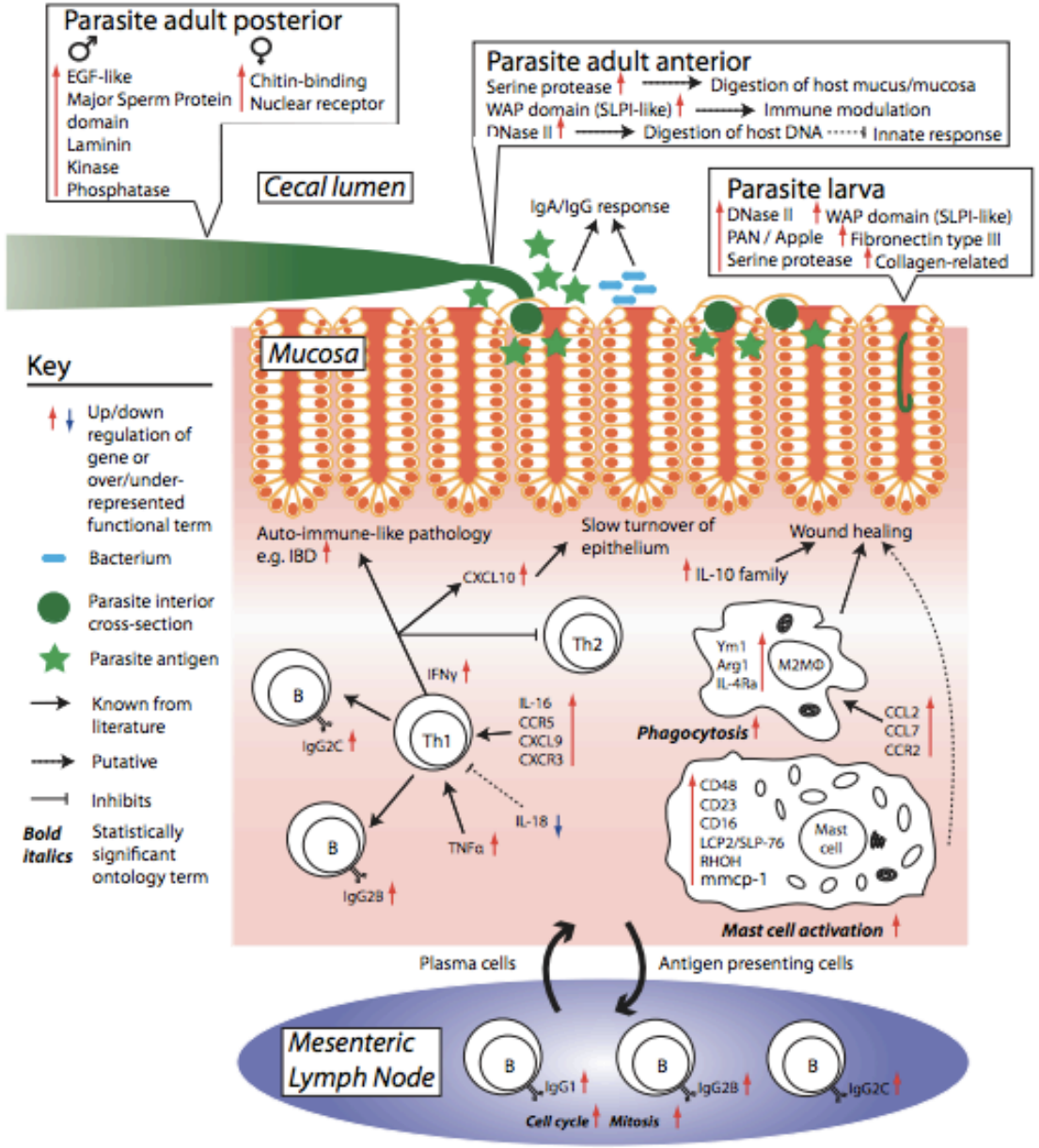
- 2 genomes probably costs less than £10,000k
- About **40 RNAseq** libraries of different life cycle stages, host infecting stages
- Paradigm shifts to RNAseq

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





[In comparison]

Cell

Article

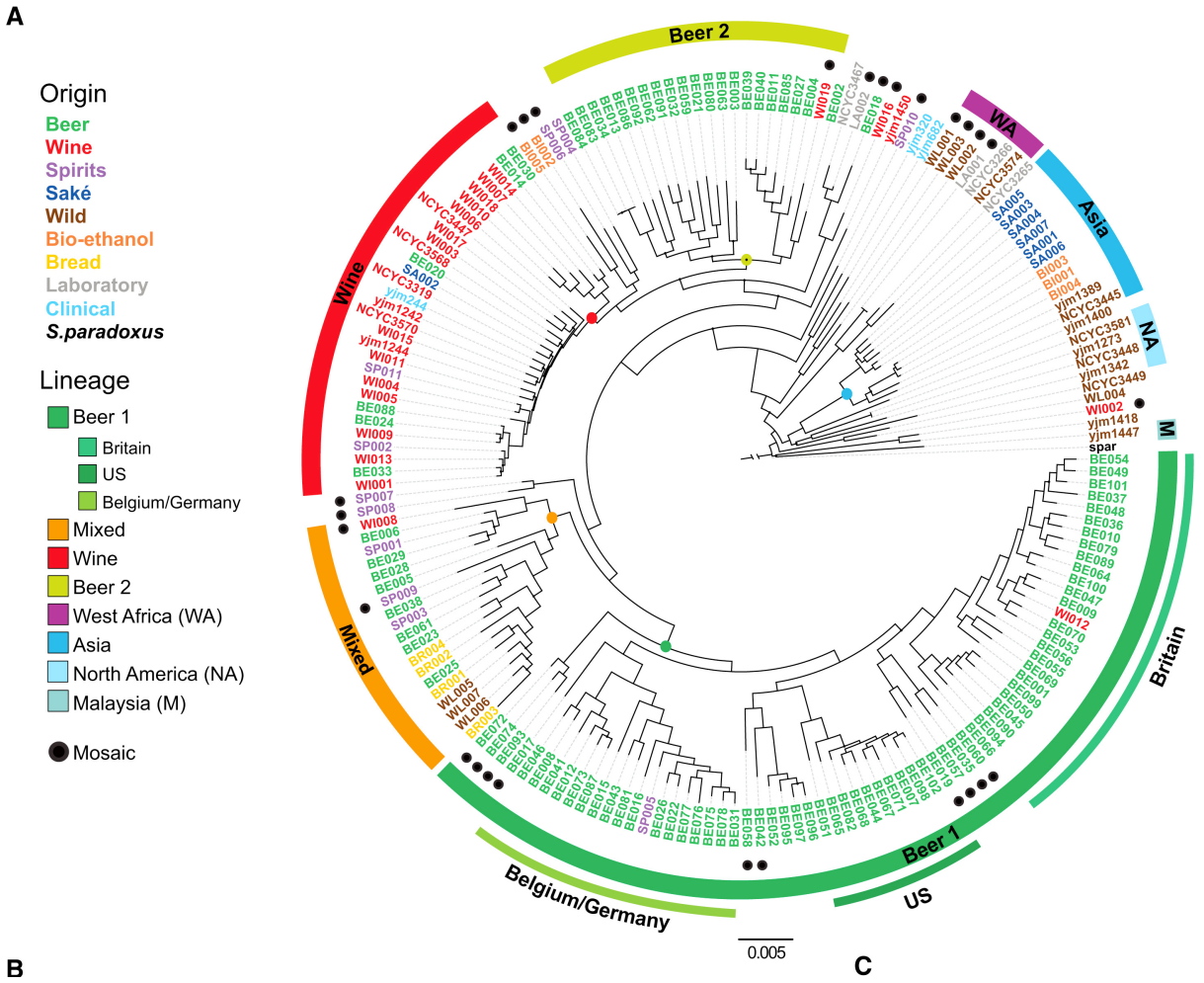
Domestication and Divergence of *Saccharomyces cerevisiae* Beer Yeasts

••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 2005-2016 (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
  - No longer exclusive to sequencing centres
  - But it also means some rubbish is being produced..

- 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

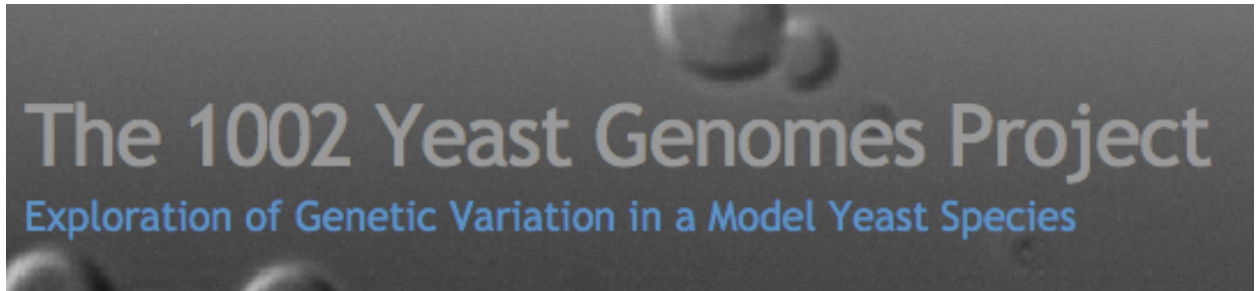
Current and future

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

GENOME 10K©

Genomics england

- Standard labs will be able to generate collections of themselves



(3 labs)

Resources

Some very useful websites:

- http://angus.readthedocs.org/en/2015/#
- http://www.pasteur.fr/~tekaia/BCGA2014/BCGA2014_Prog.html
- http://evomics.org/
- http://molb7621.github.io/workshop/
- https://sequencing.qcfail.com
- http://schatzlab.cshl.edu/teaching/

It’s a big world out there

- Read, read, read
- Setup twitter and follow what others are doing