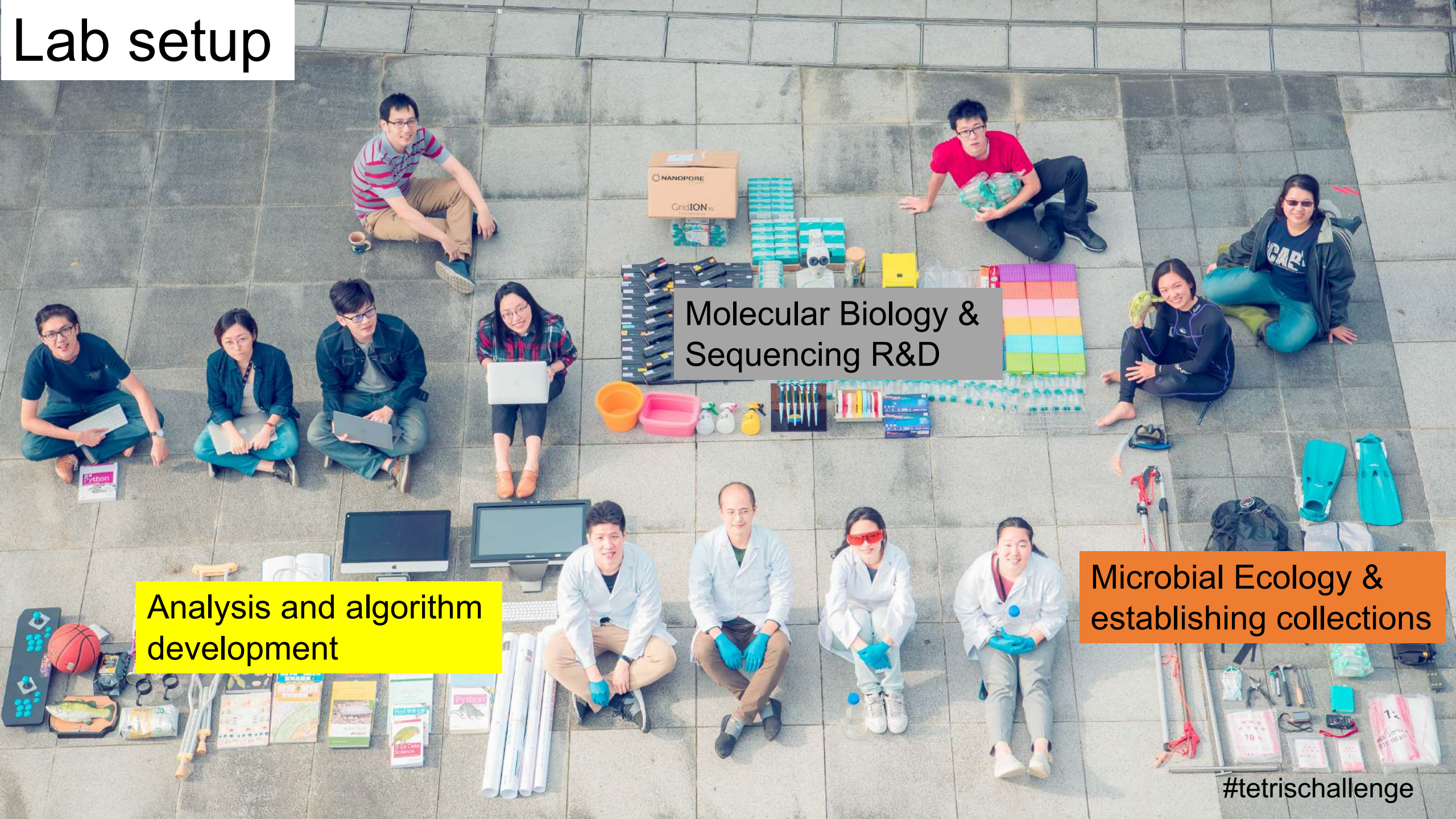
Recent updates on genomics and transcriptomics.

Isheng Jason Tsai

生圖教育訓練課程
2021.01.11


NANOPORE
GridION
BOOM CHICKA POP
nature
Python
#tetrischallenge

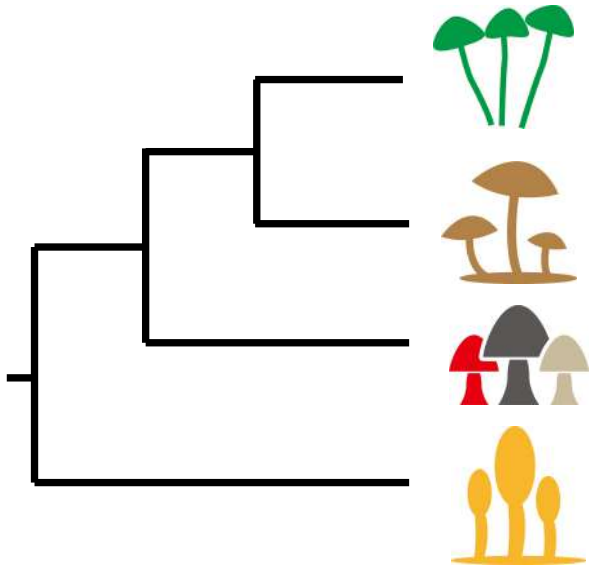
Lab setup


NANOPORE
GridION
Molecular Biology & Sequencing R&D
Analysis and algorithm development
Microbial Ecology & establishing collections
#tetrischallenge

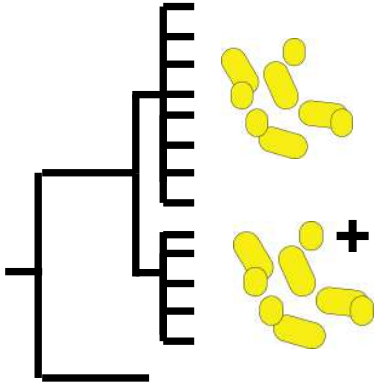
Evolutionary genomics

million years ago

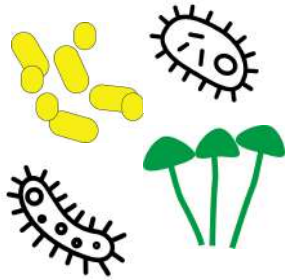
Comparative genomics/transcriptomics



Population genomics



Metagenomics/ metatranscriptomics



Tsai *et al* (2008, 2010) PNAS
Liti *et al.,* Nature (2009)
Tsai *et al.*, Nature (2013)
Valentim *et al.,* Science (2013)
Foth and Tsai *et al.*, Nature Genetics (2014)
Hunt and Tsai *et al.,* Nature Genetics (2016)

Natsumi and Tsai *et al.,* Nature Communications (2018)
Coghlan *et al.,* Nature Genetics (2018)
Chaw *et al.,* Nature Plants (2019)
Sung *et al.,* CMGH (2019)
Ke *et al.,* PNAS (2020)
Lin *et al* Gut Microbes (2021)

Recent updates in…

Genomics

"is an interdisciplinary field of biology focusing on the structure, function, evolution, mapping, and editing of genomes." (wiki)

Transcriptomics

"the study of the complete set of RNA transcripts that are produced by the genome, under specific circumstances or in a specific cell—using high-throughput methods, such as microarray analysis." (nature portfolio)

Systems Biology

"…is an approach in biomedical research to understanding the larger picture—be it at the level of the organism, tissue, or cell—by putting its pieces together."[1] (opposite to reductionist view of taking things apart)

Bioinformatics

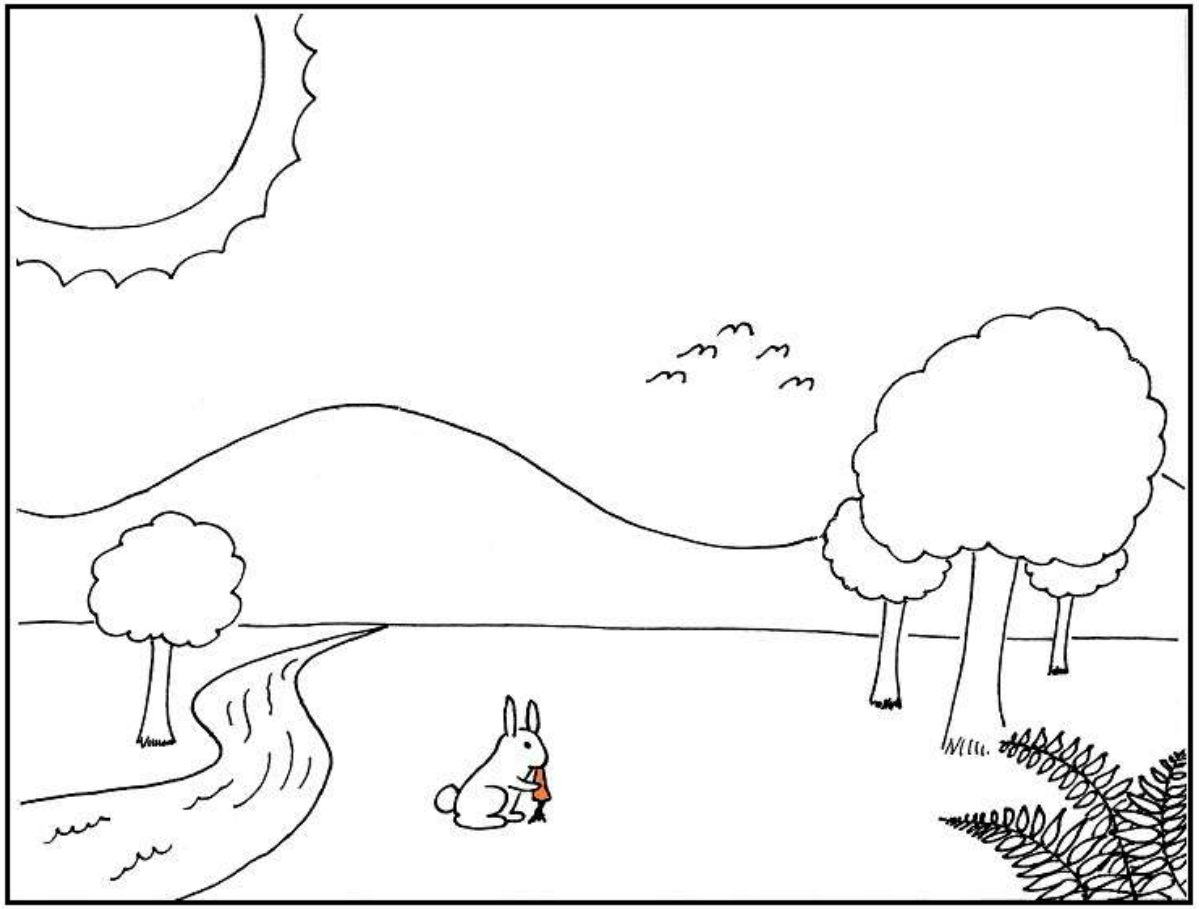
"is a subdiscipline of biology and computer science concerned with the acquisition, storage, analysis, and dissemination of biological data, most often DNA and amino acid sequences."

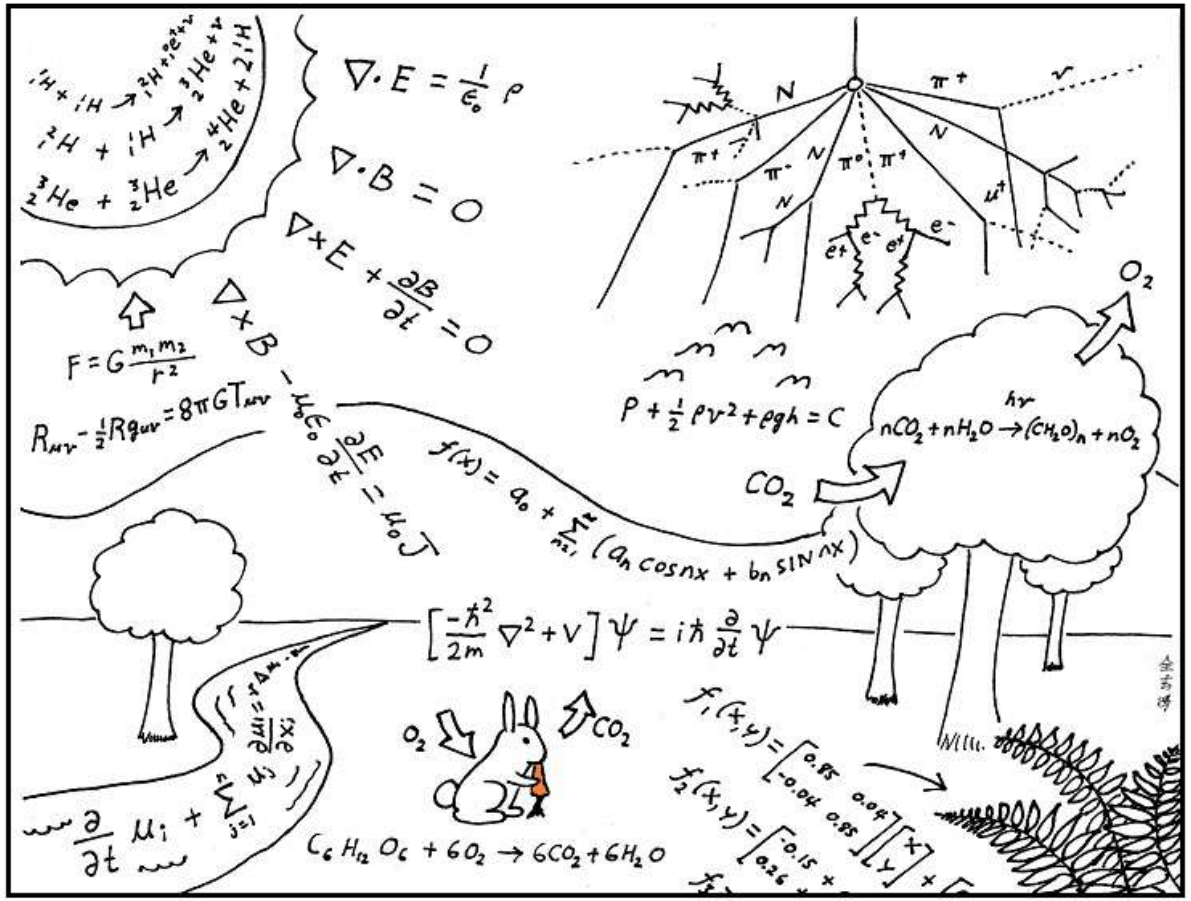
https://www.genome.gov/genetics-glossary/Bioinformatics

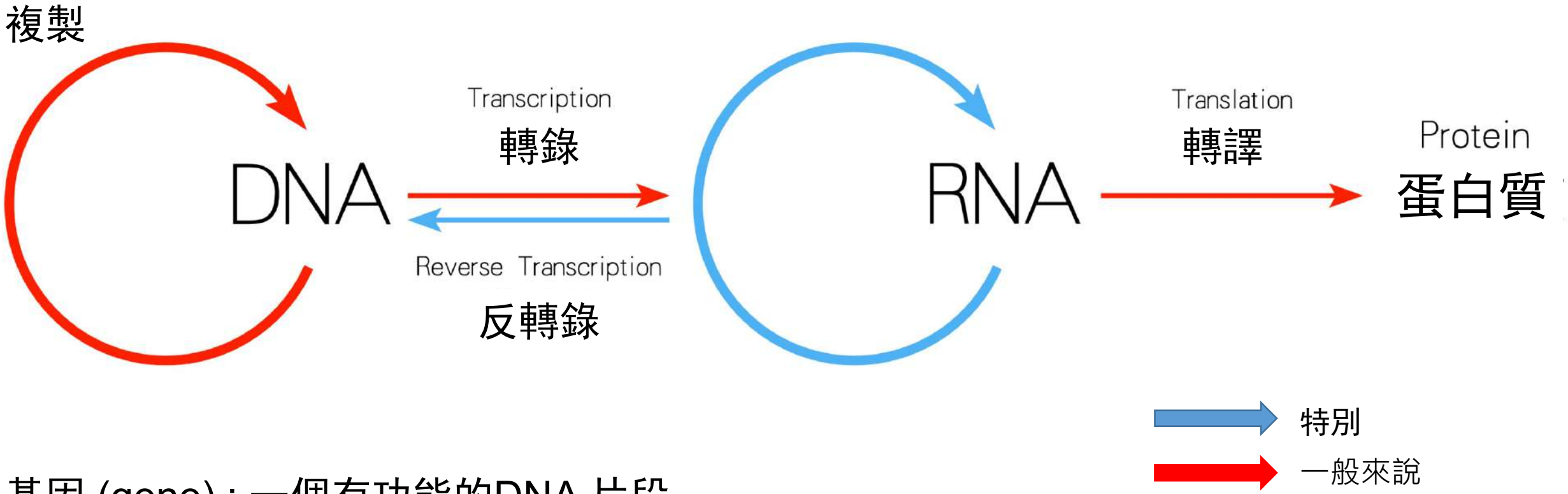
Lecture outline

1. Introduction
2. Brief history of sequencing
3. dawn of third gen sequencing
4. Case studies - genomics
5. Case studies - transcriptomics

This is how scientists see the world

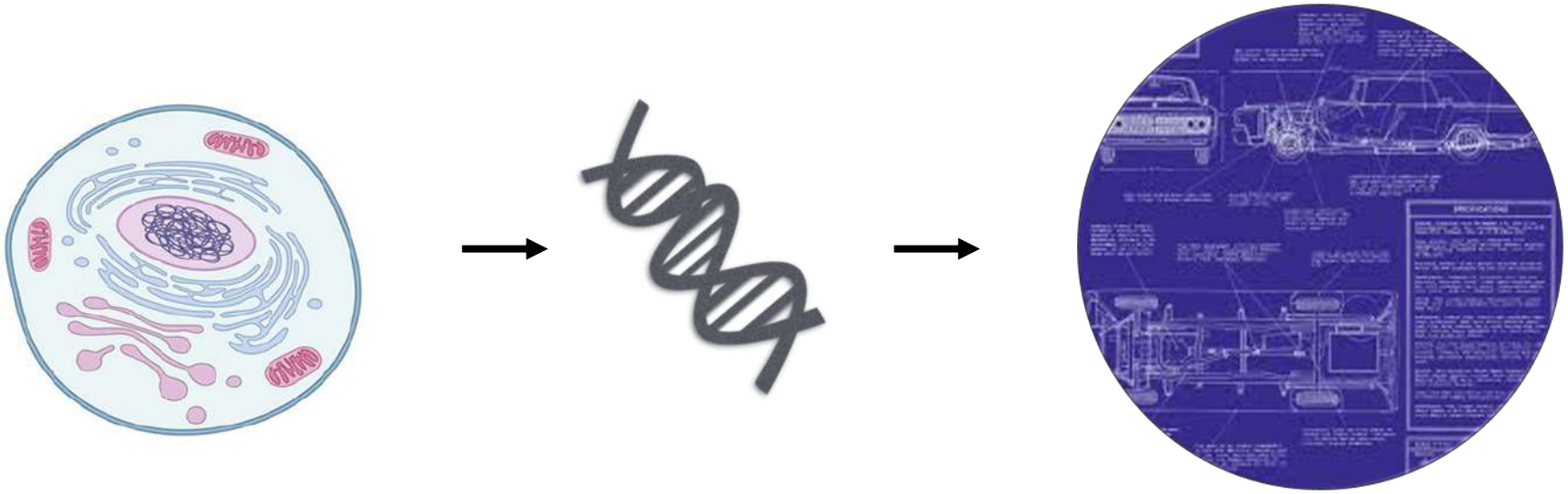






基因 (gene) : 一個有功能的DNA 片段
coding DNA: 可以轉譯成蛋白質的 DNA #noncoding
基因體 (genome): 物種一個細胞核內所有的DNA
定序 (sequencing) : 解析出DNA 序列 [ATCGTGACGTGACGTAC…]

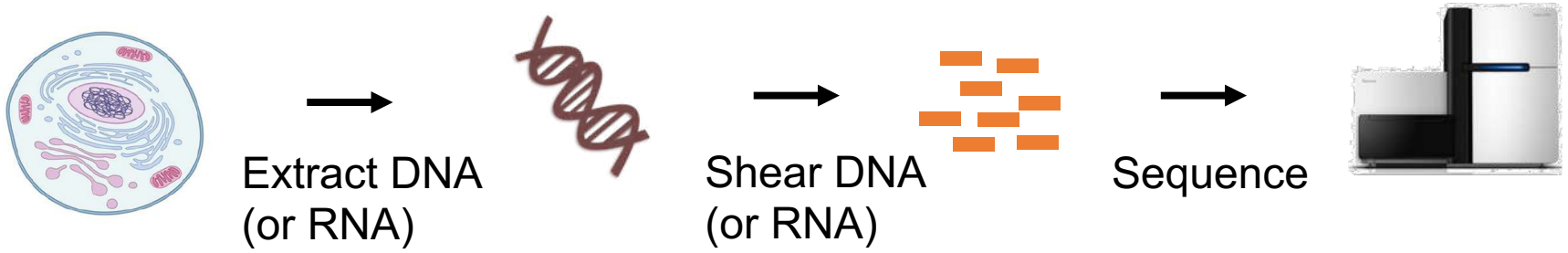
Genome



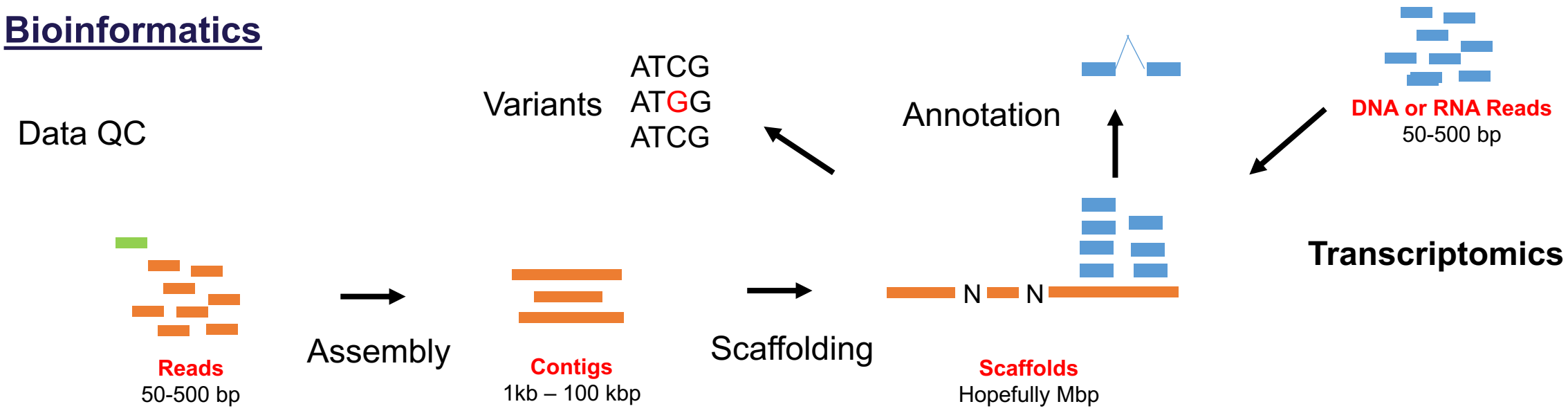
Genome = Parts list of a single genome

A typical genome/transcriptomic project

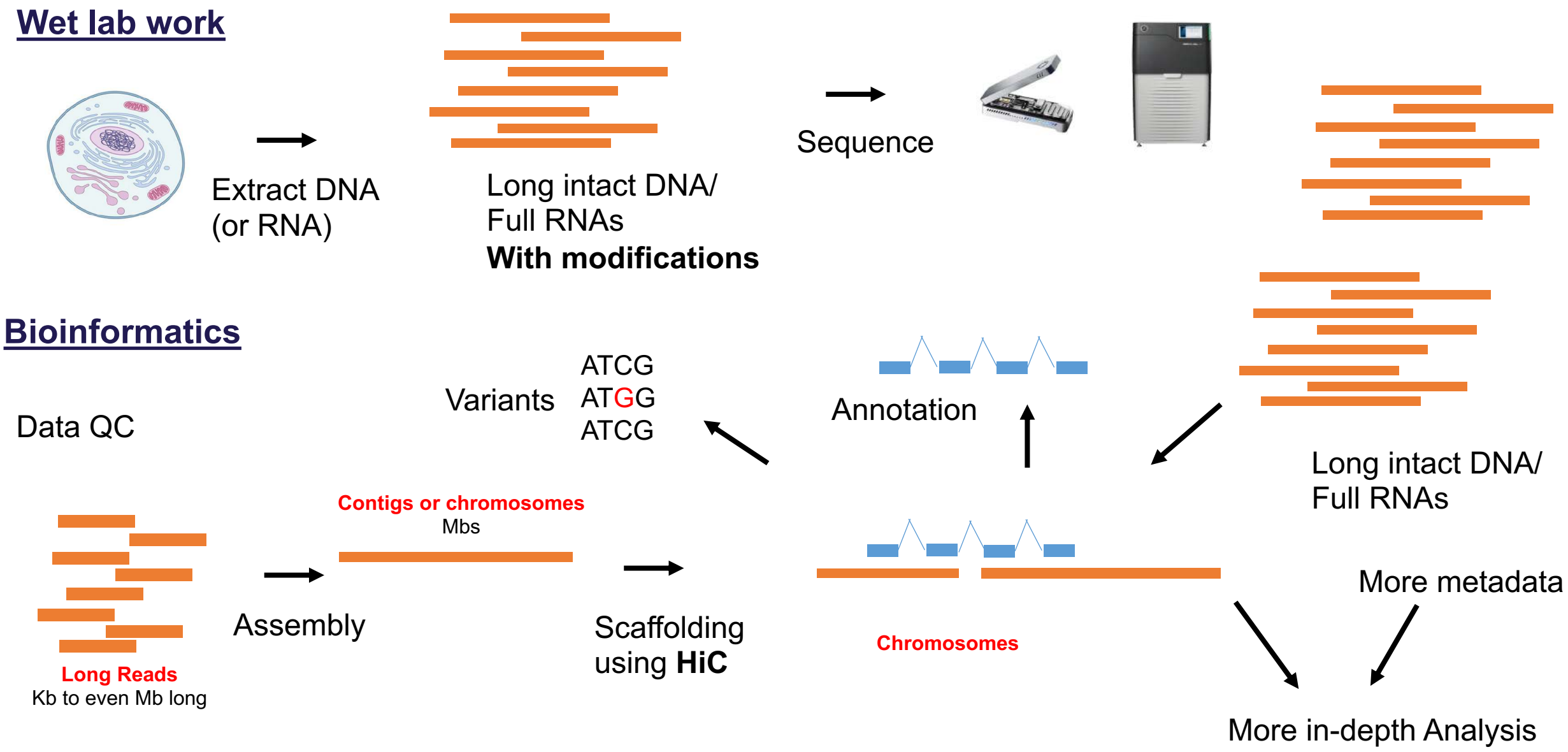
Wet lab work



Bioinformatics



A 2021 genome/transcriptomic project (take home message)



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**

Things to consider in sequencing

1. Length
2. Depth
3. Biases
4. Errors

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.

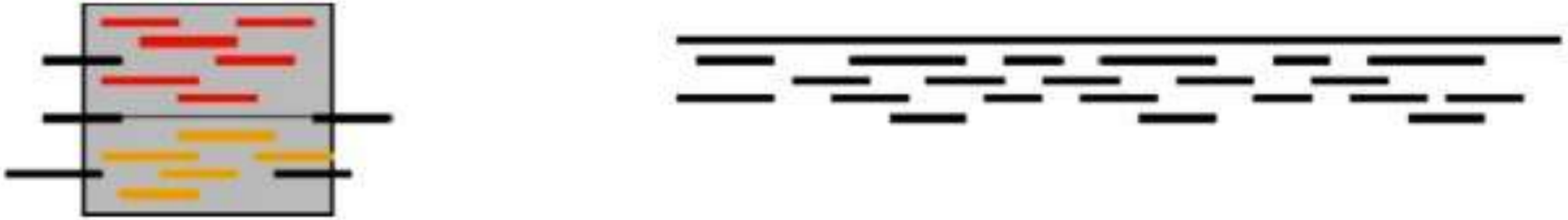
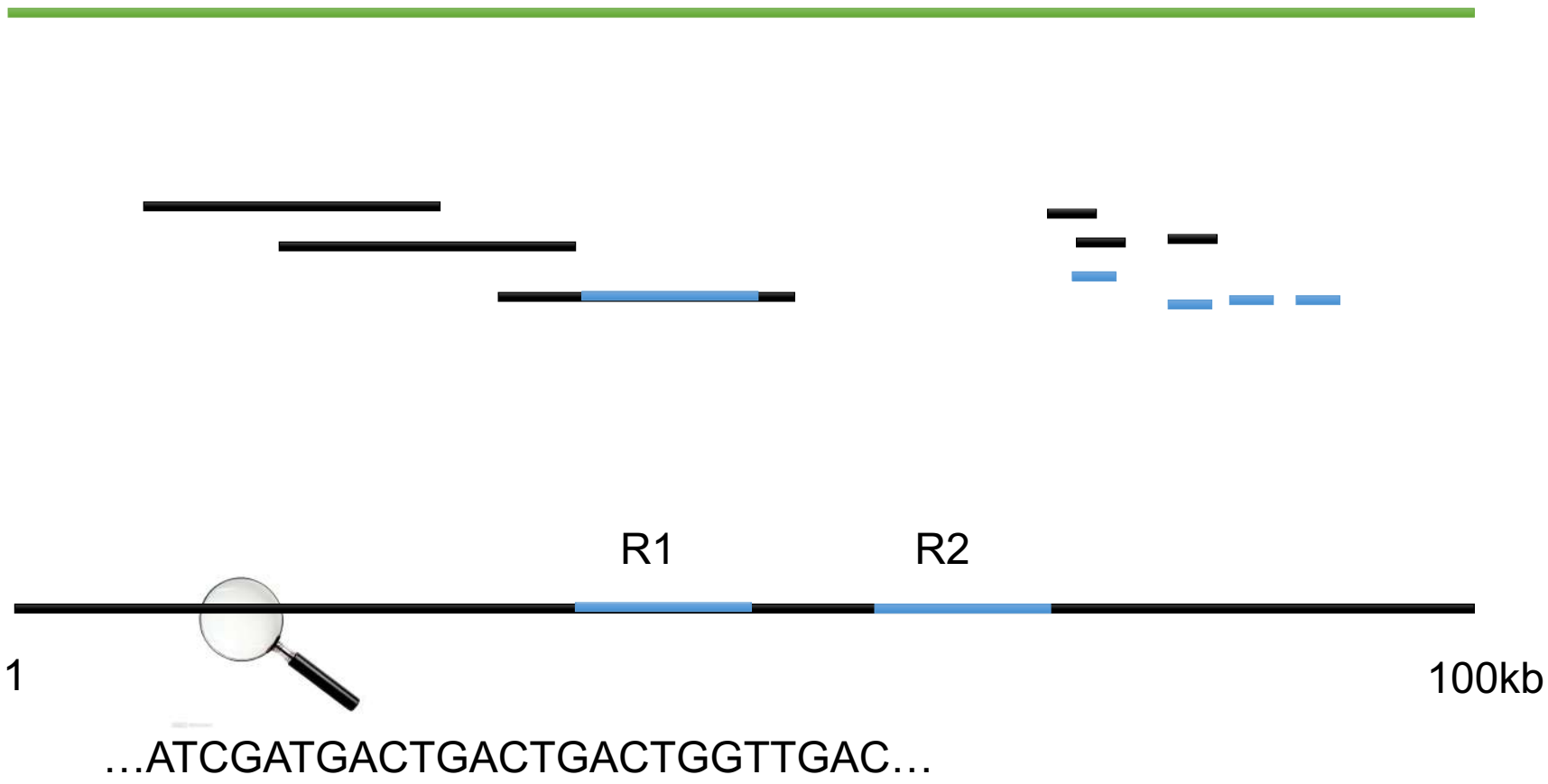


Figure 6. Genome mis-assembled due to a repeat. The assembly program incorrectly combined the reads from the two copies of the repeat leading to the creation of two separate contigs

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

Read length matters in sequencing



Depth matters in sequencing

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

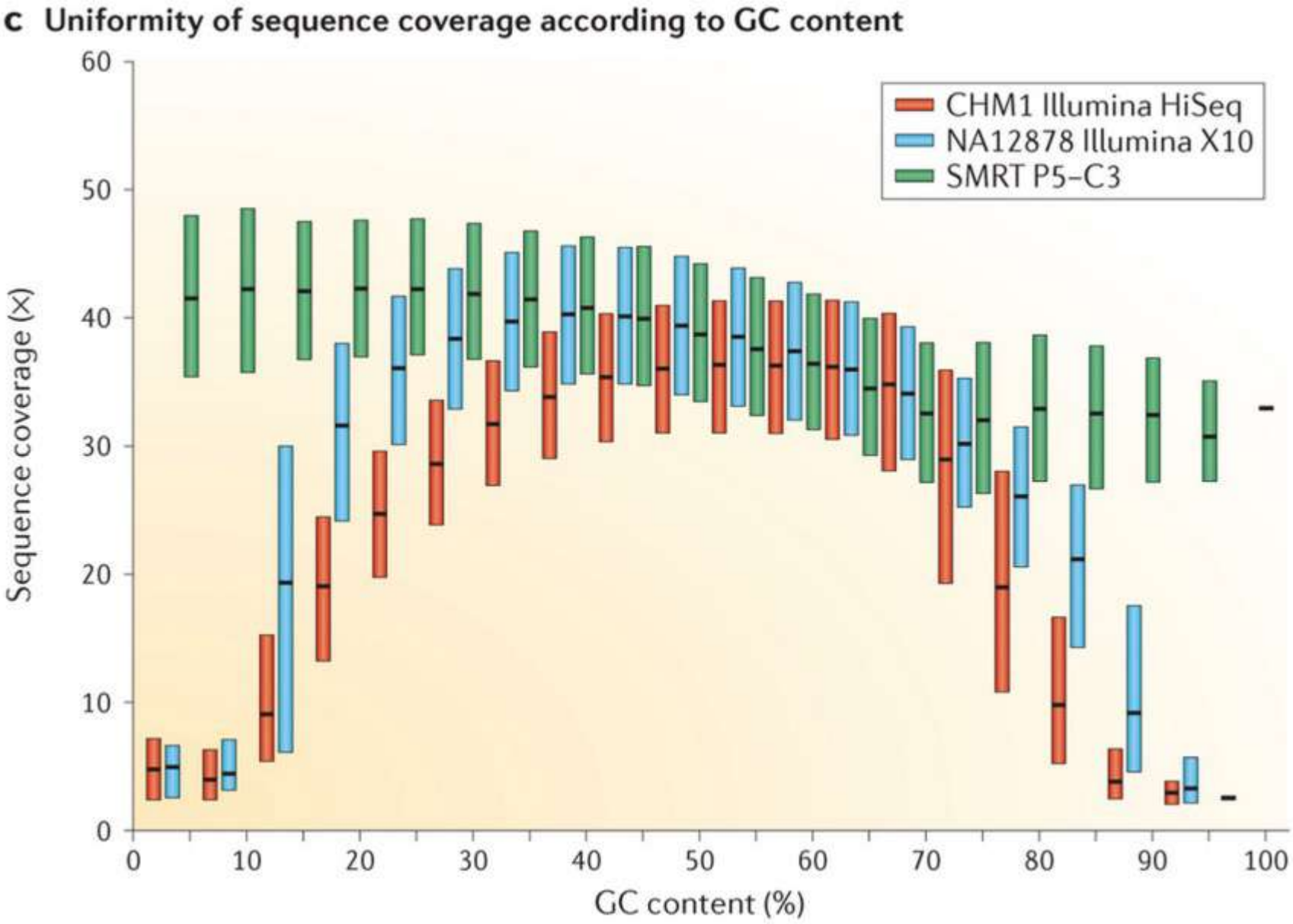
Homozygous? Heterozygous?

1X ATCGATCACTGACTGACTGGTTGAC

…ATCGATGACTGACTGACTGGTTGAC…

reference

Sequencing Biases

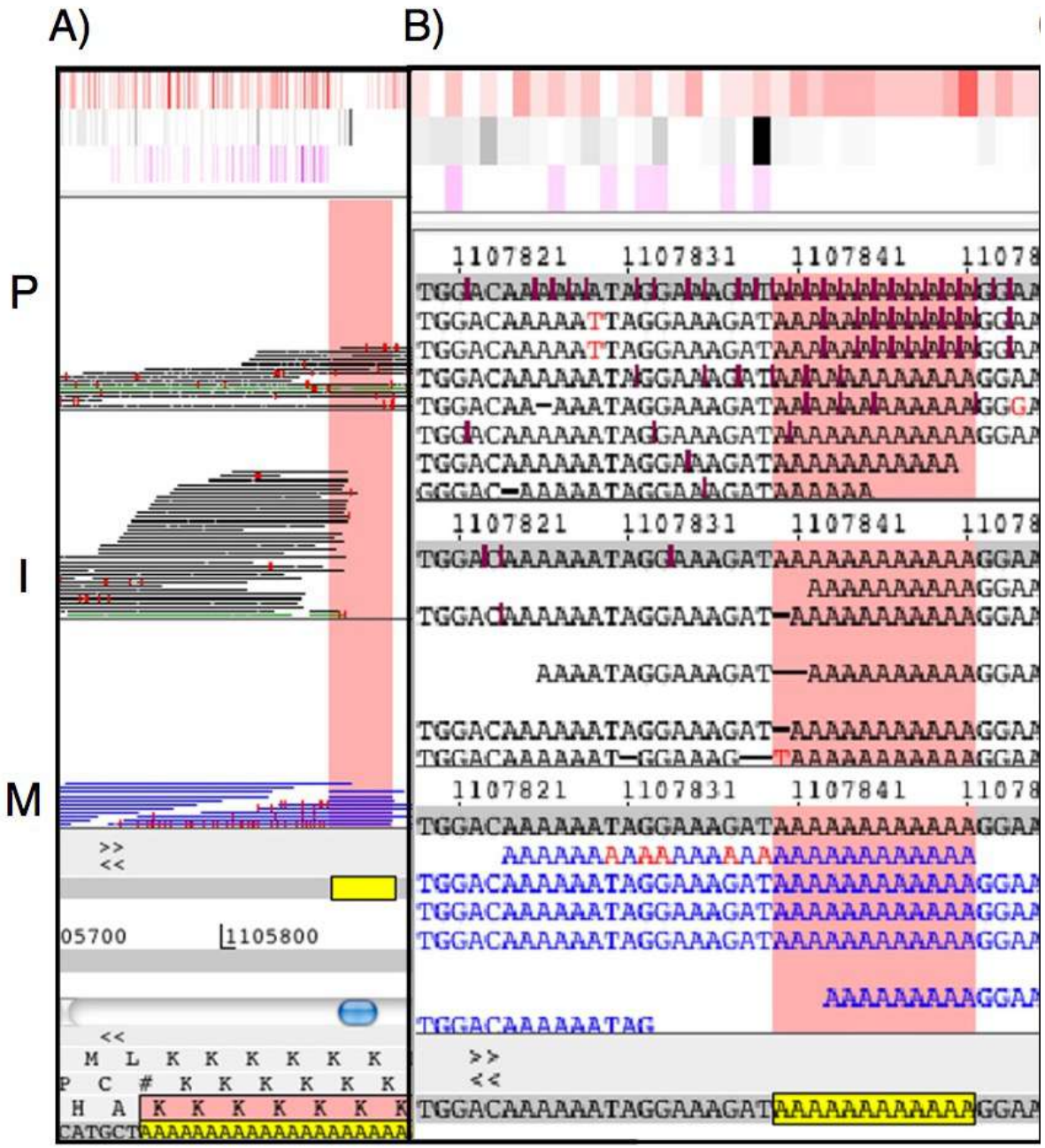


http://www.nature.com/nrg/journal/v16/n11/fig_tab/nrg3933_F2.html

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

Sequencing – a brief history

nature milestones

Genomic sequencing

1. **The Human Genome Project**
2. Sequencing the unculturable majority
3. **Sequencing — the next generation**
4. ChIP–seq captures the chromatin landscape
5. The dawn of personal genomes
6. A sequencing revolution in cancer
7. Transcriptomes — a new layer of complexity
8. **Long reads become a reality**
9. Exploring whole exomes
10. **Probing nuclear architecture with Hi-C**
11. **Sequencing one cell at a time**
12. Waking the dead: sequencing archaic hominin genomes
13. Cataloguing a public genome
14. Our most elemental encyclopaedia
15. Pan-genomes: moving beyond the reference
16. Genomes go platinum
17. **Filling in the gaps telomere to telomere**

10 February 2021

https://www.nature.com/immersive/d42859-020-00099-0/index.html

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

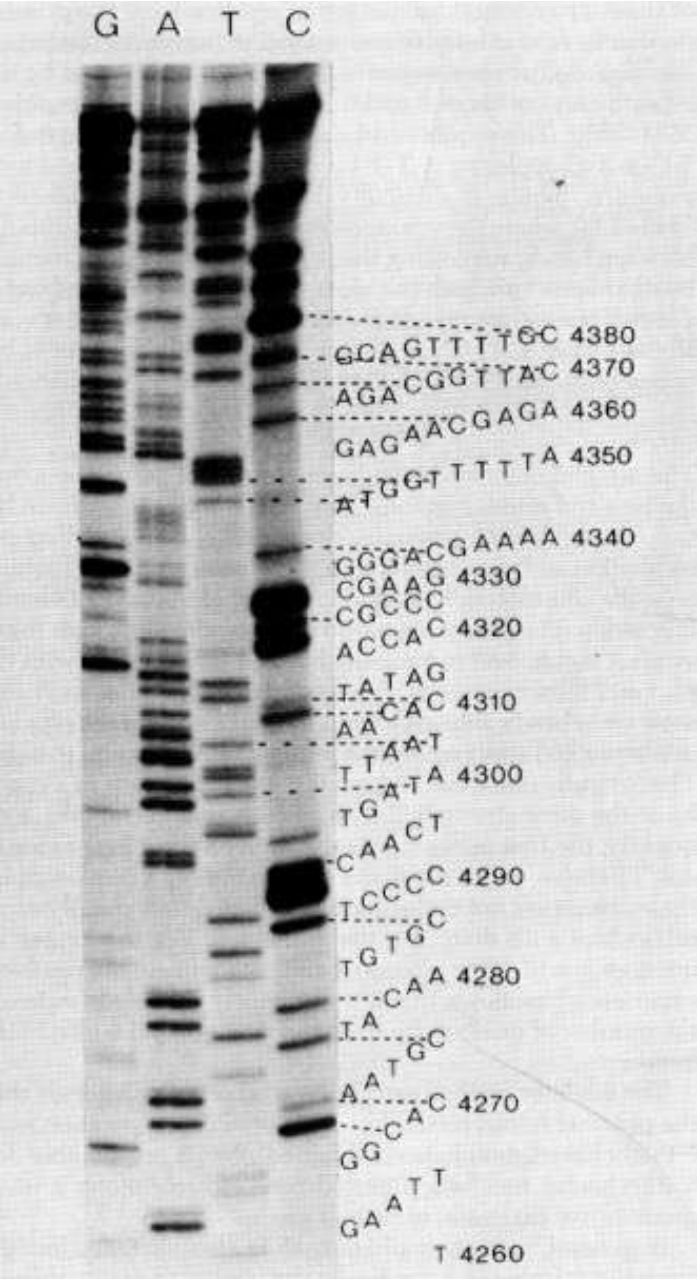
DNA sequencing with chain-terminating inhibitors

(DNA polymerase/nucleotide sequences/bacteriophage ϕX174)

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

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

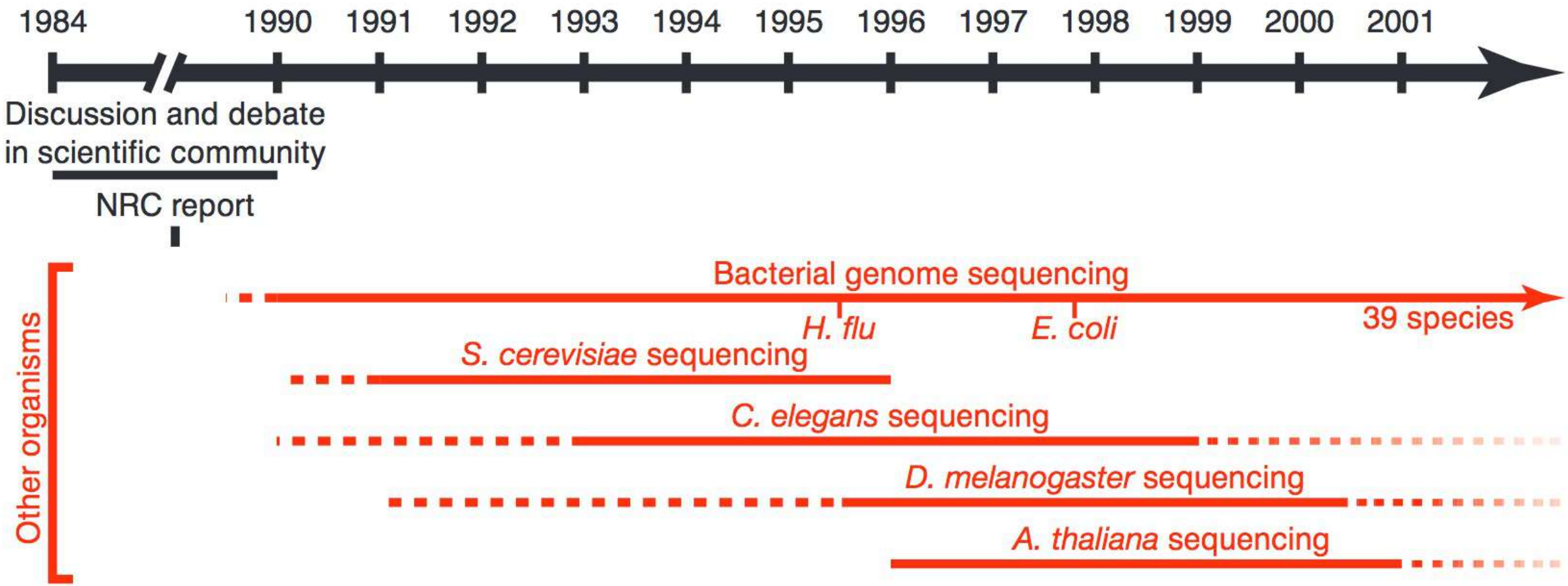
Contributed by F. Sanger, October 3, 1977



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

ABI 3730xi at TIGR (1.6Mb per day)

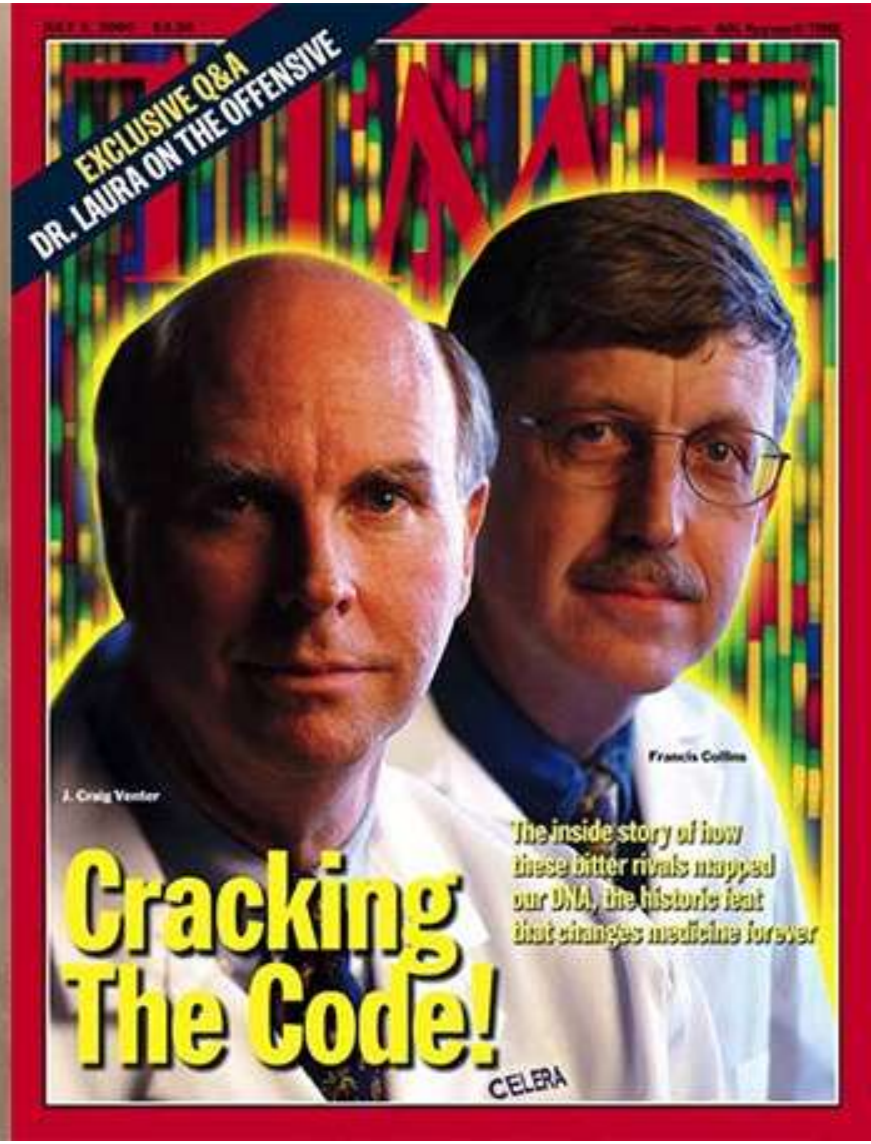
https://www.flickr.com/photos/jurvetson/57080968



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

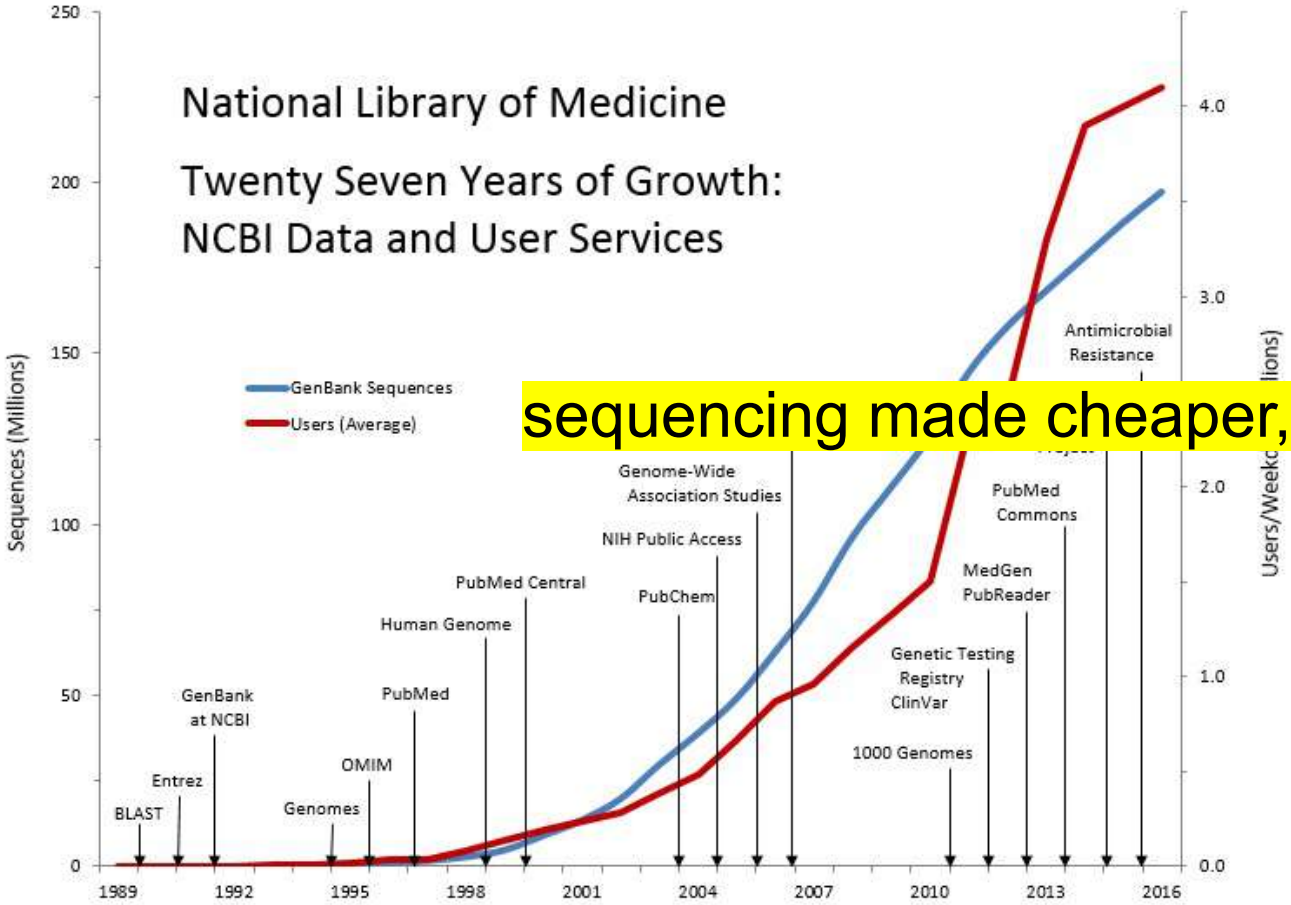

EXCLUSIVE Q&A
DR. LAURA ON THE OFFENSIVE
J. Craig Venter
Francis Collins
The inside story of how these bitter rivals mapped our DNA, the historic feat that changes medicine forever
Cracking The Code!
CELERA

Calculating the economic impact of the Human Genome Project

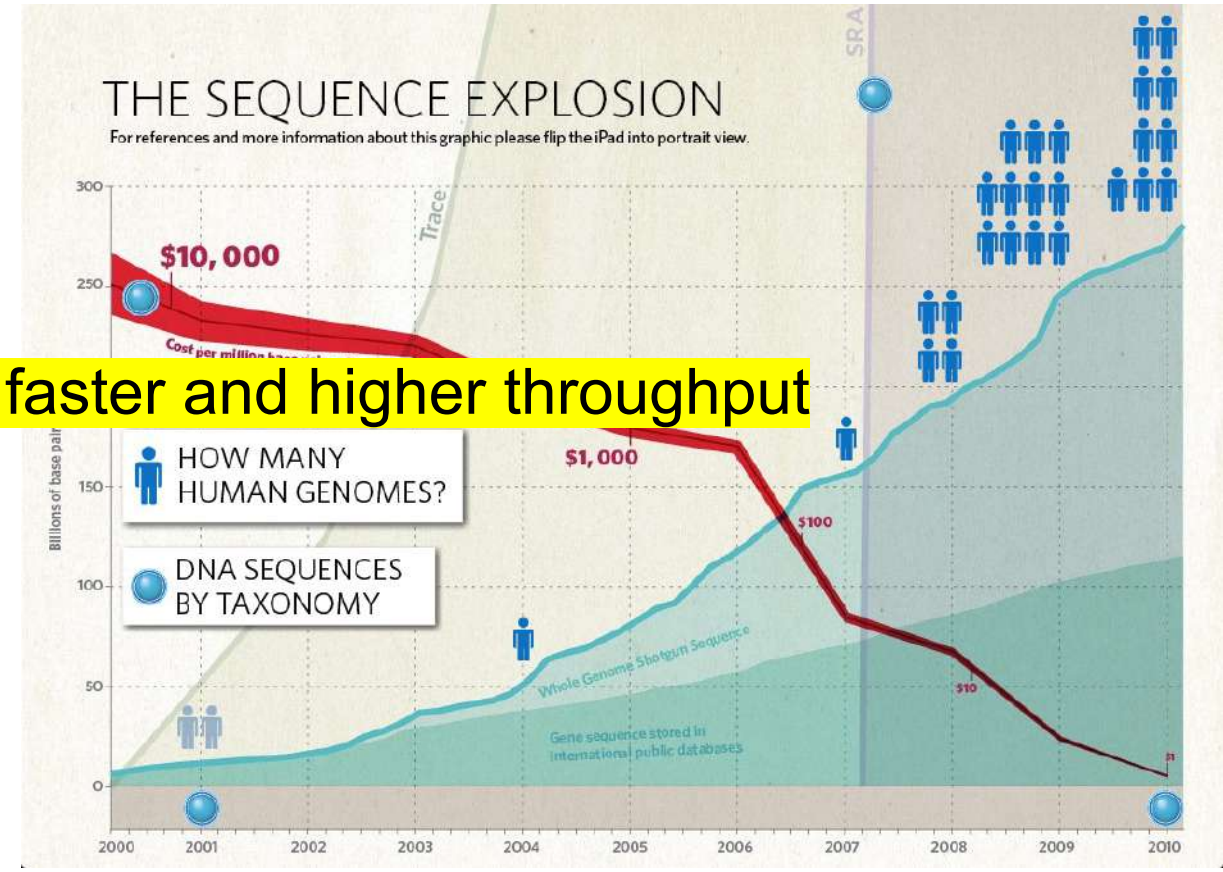
Public funding of scientific R&D has a significant positive impact on the wider economy, but quantifying the exact impact of research can be difficult to assess. A new report by research firm Battelle Technology Partnership Practice estimates that **between 1988 and 2010, federal investment in genomic 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/

2000-2010s – Second generation sequencing and associated challenges



sequencing made cheaper, faster and higher throughput



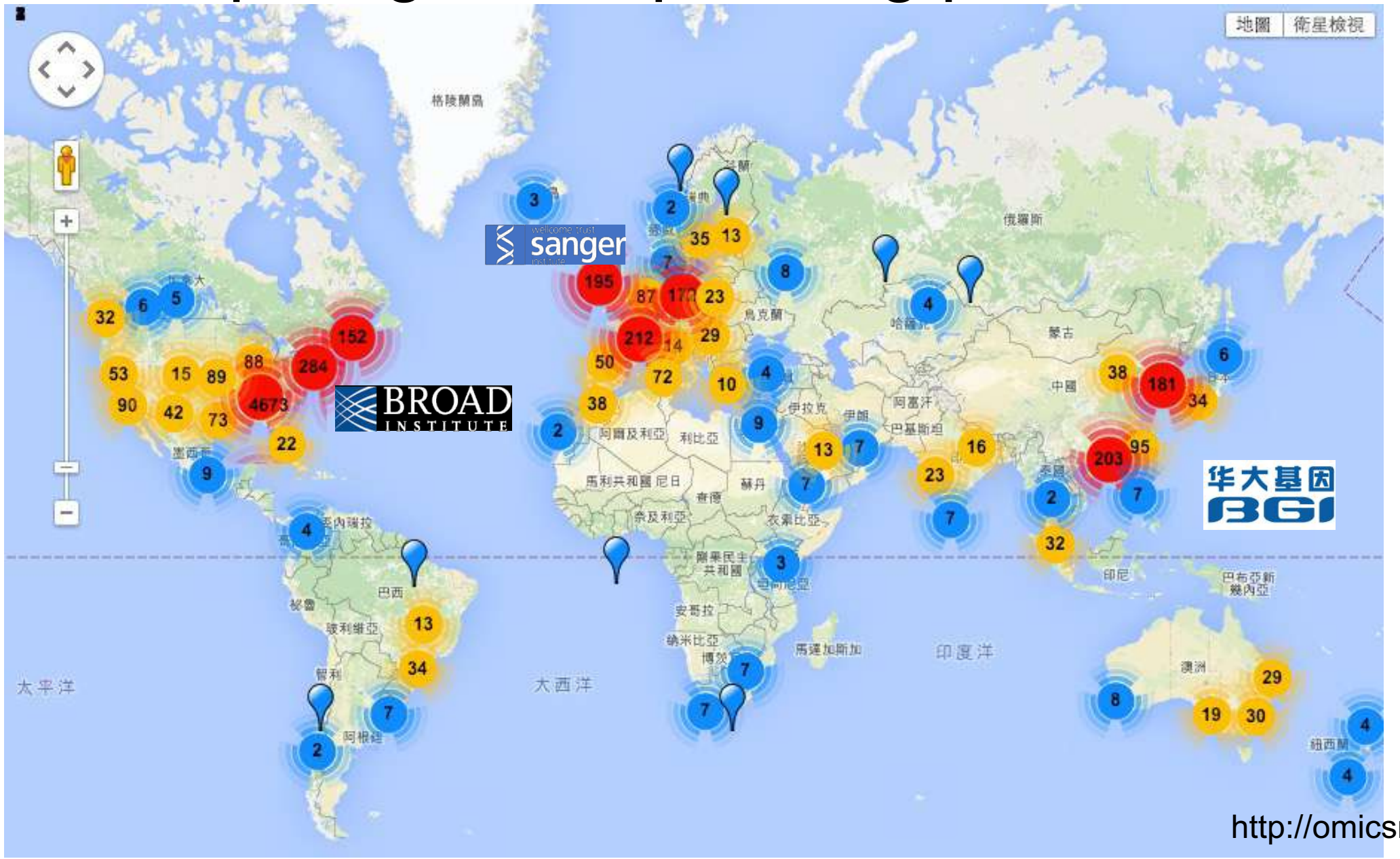
https://www.nlm.nih.gov/about/https://www.nlm.nih.gov/about/2018CJ.html

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

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

NGS = sequencing made cheaper, faster and higher throughput

World competing for sequencing power

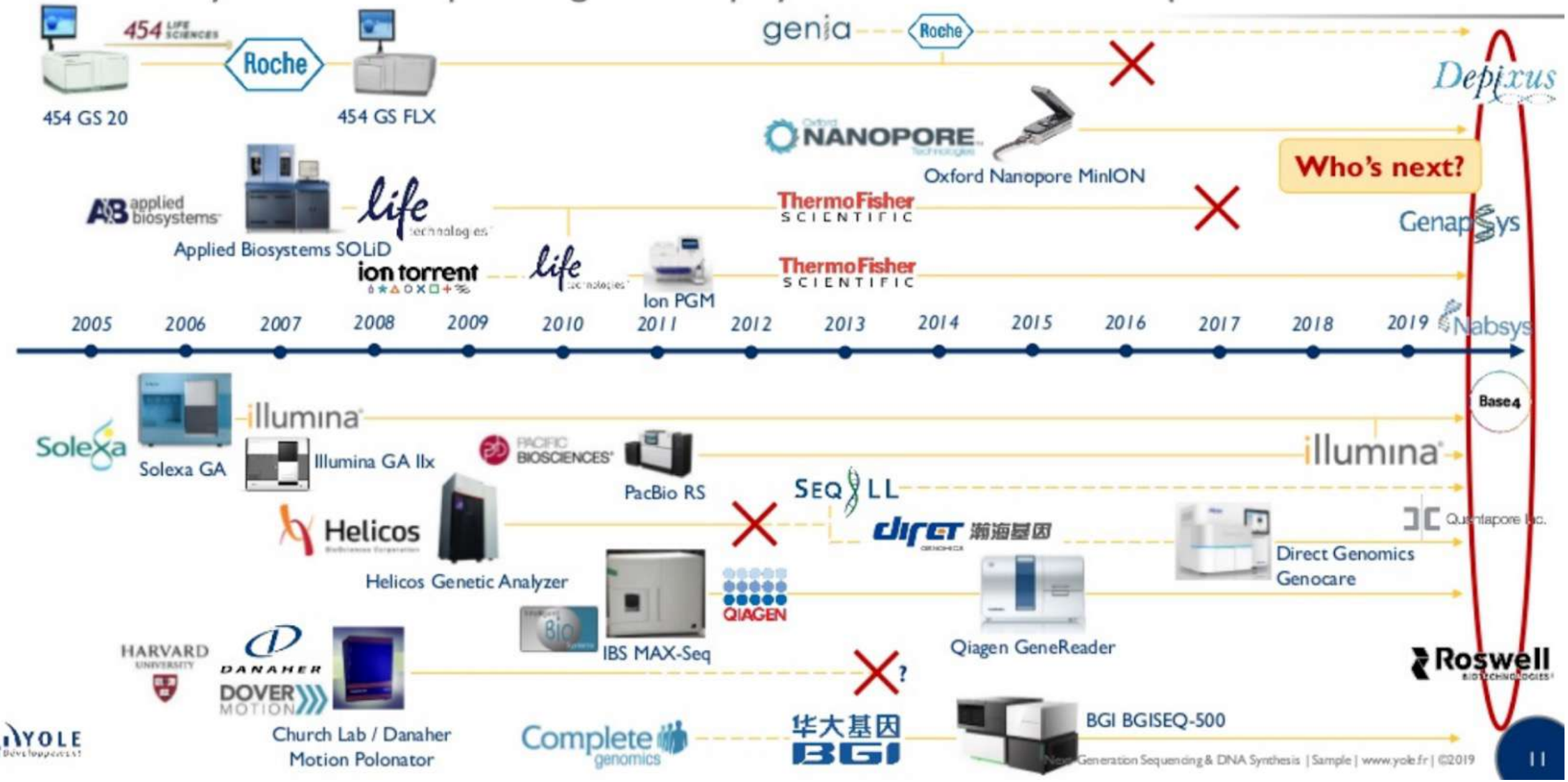


http://omicsmaps.com/

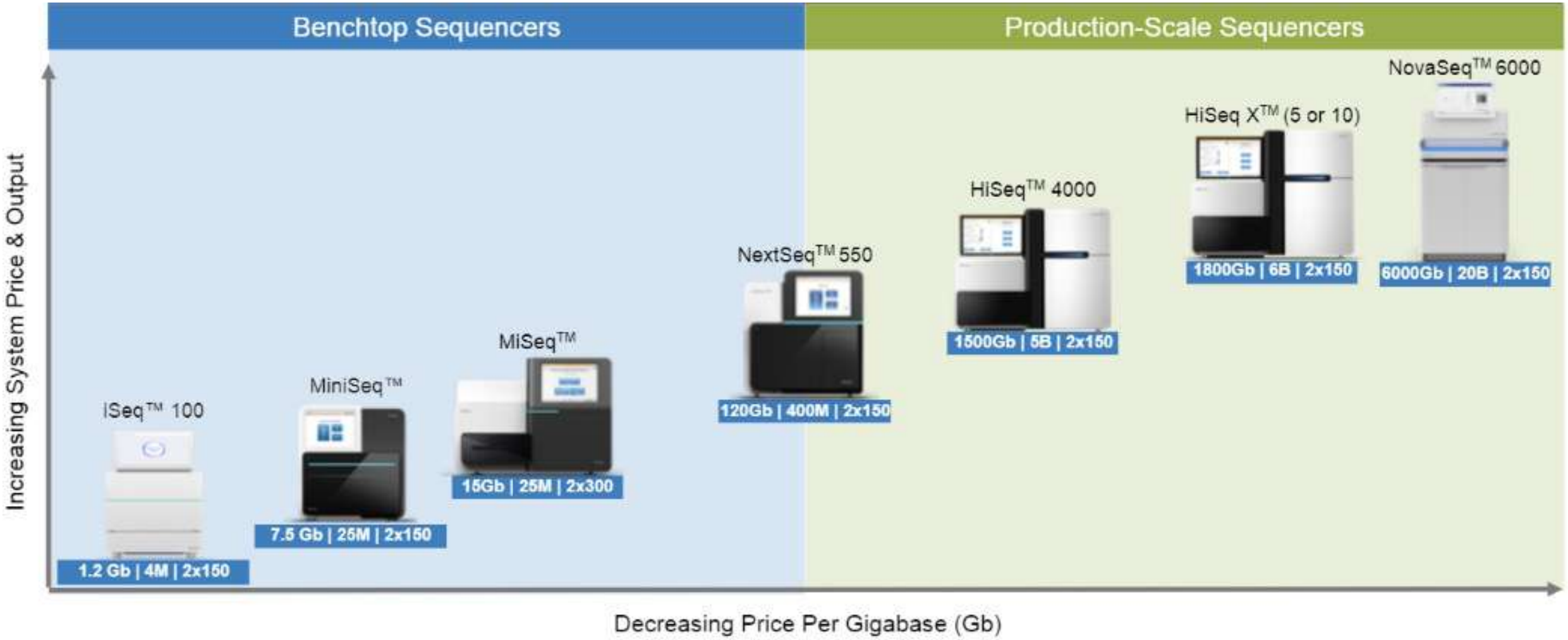
INTRODUCTION

Clip slide

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

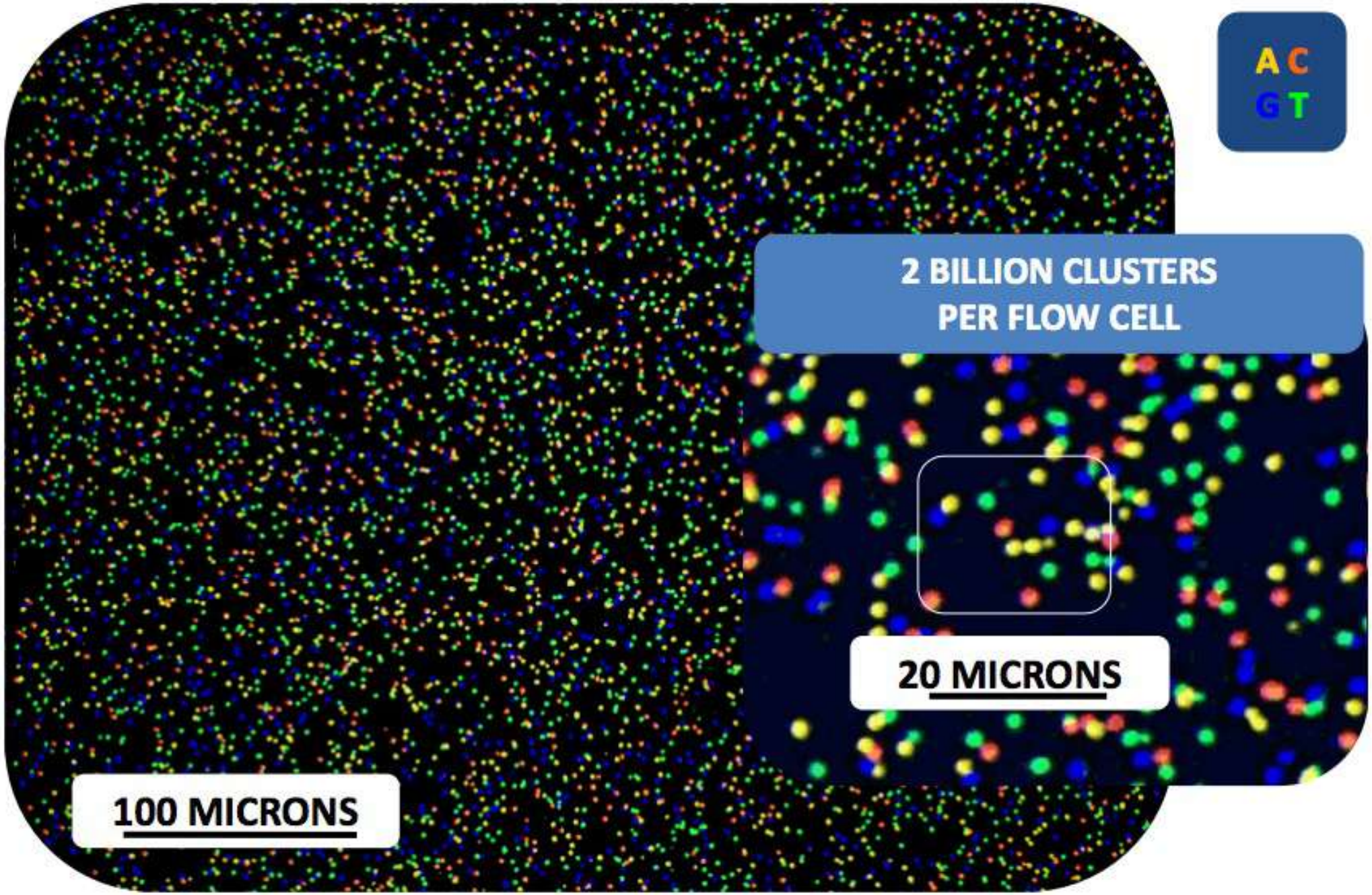


Illumina machines



Illumina: sequencing by synthesis

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


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

Illumina platform comparison

And the arrival of 3rd generation sequencing…

(much longer read lengths)

PacBio (Pacific Biosciences)

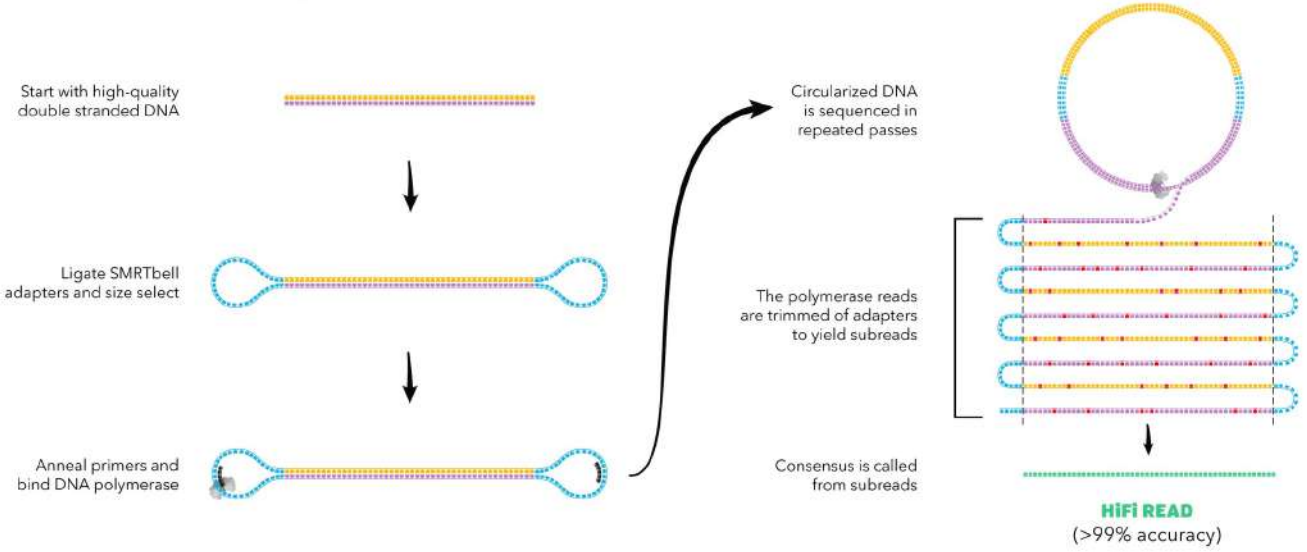
RSII

Sequel II

PacBio (Pacific Biosciences)

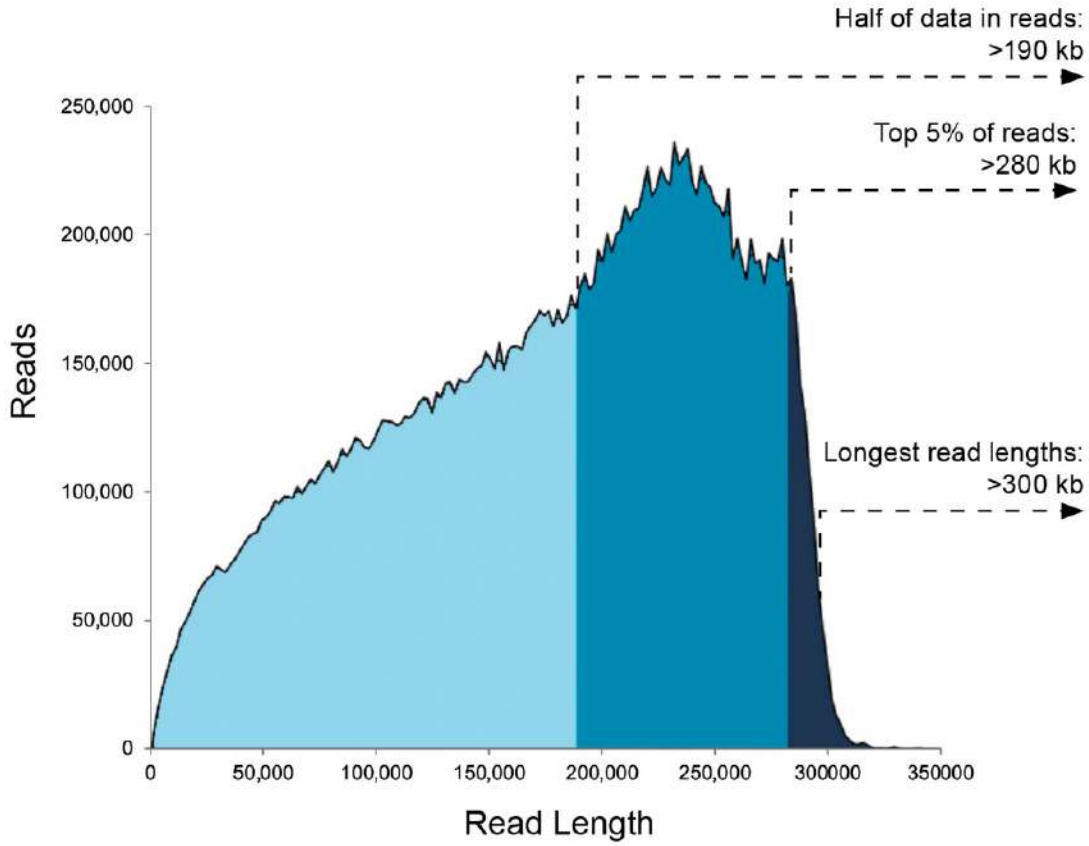
HiFi READS

Produce HiFi reads using the circular consensus sequencing (CCS) mode to provide base-level resolution with >99% single-molecule read accuracy for the detection of all variant types from single nucleotide to structural variants. Learn more about the advantages of long reads with high accuracy.



Half of data in reads: >190 kb

Data per SMRT Cell: Up to 50 Gb



https://www.pacb.com/smrt-science/smrt-sequencing/smrt-sequencing-modes/

Oxford Nanopore





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

Oxford Nanopore – how it works

Introduction to nanopore
https://vimeo.com/297106166

Voltrax
https://vimeo.com/297106291

Sequencing for farmers
https://vimeo.com/294216876

@ Oceans
https://vimeo.com/294744892

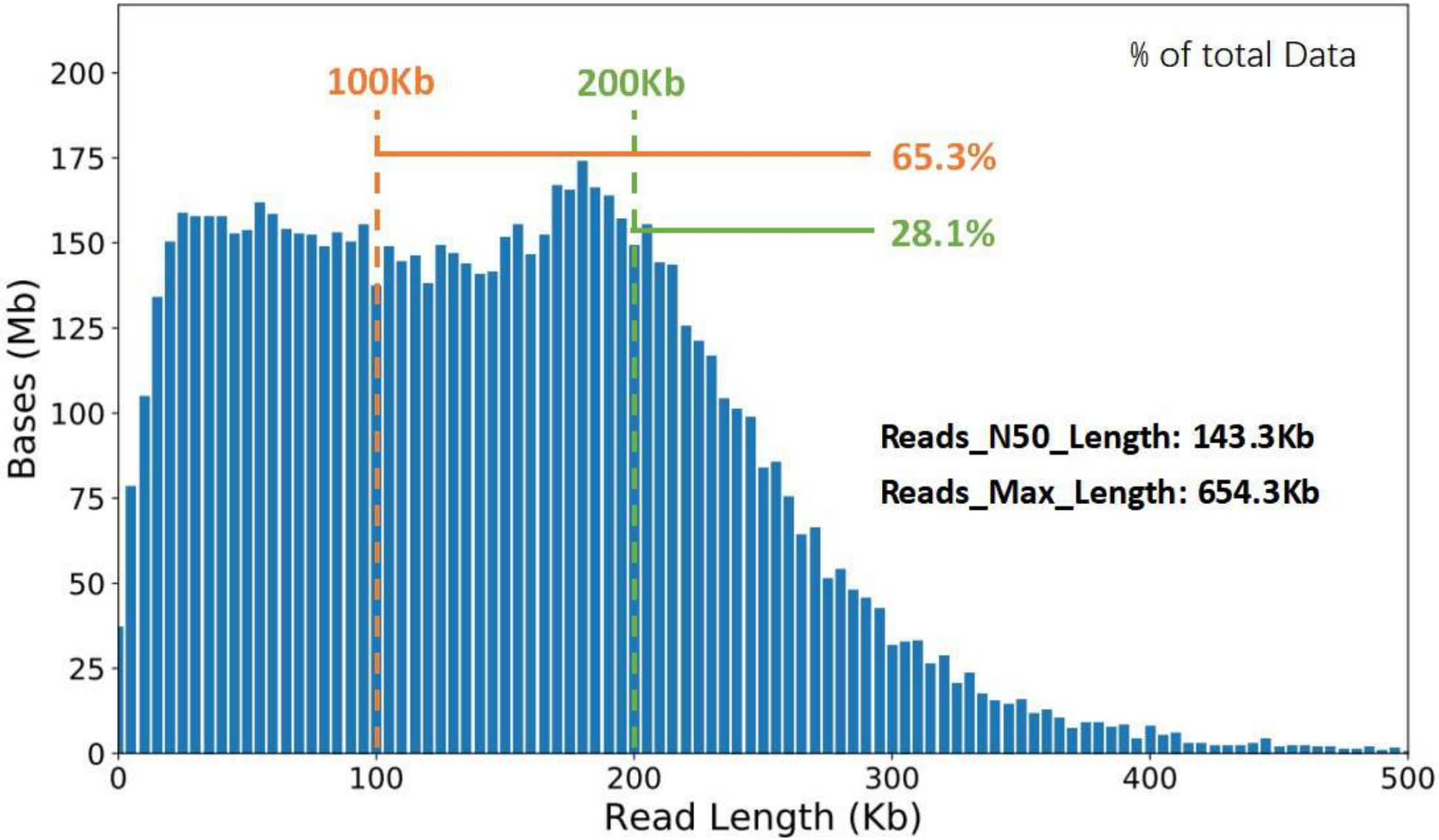
Rainforest
https://www.youtube.com/watch?v=6RRSxWtJPUw

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

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

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

Read length and capacity go beyond

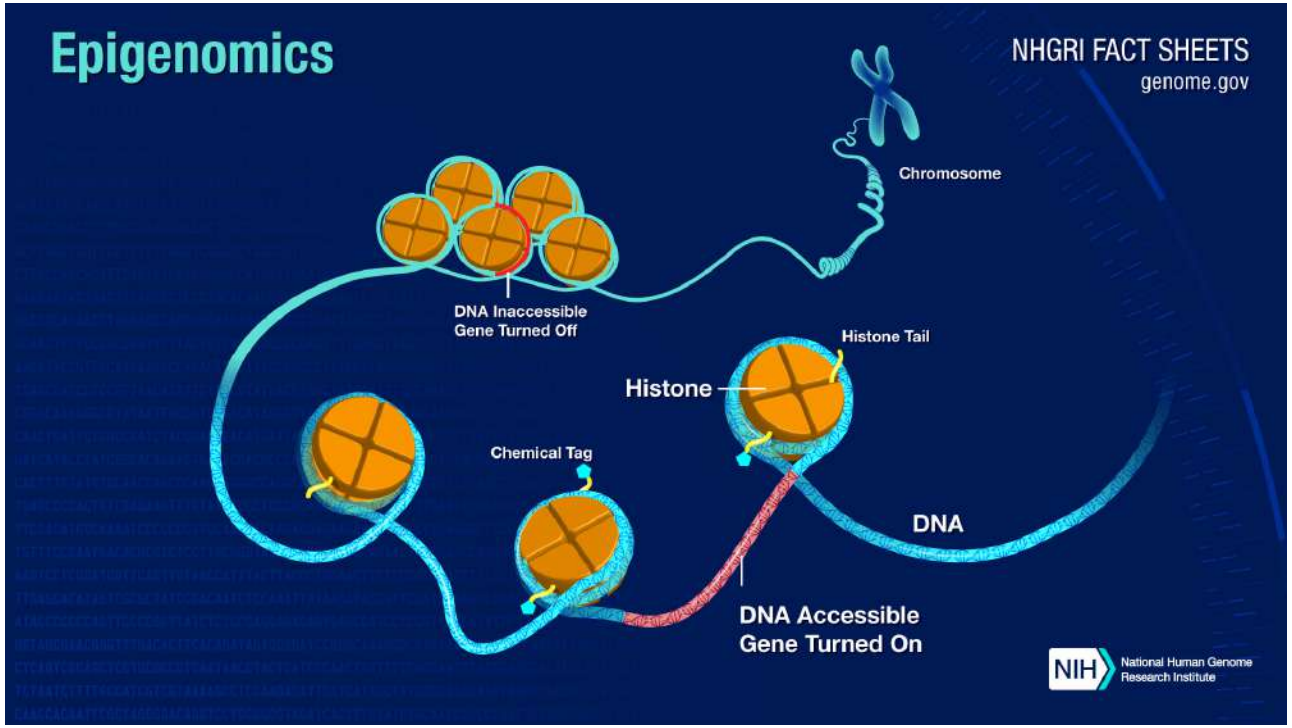


https://twitter.com/GrandOmics_Intl/status/1188724248480608257/photo/1

Epigenomics

The *epigenome* is a multitude of chemical compounds that can tell the *genome* what to do.

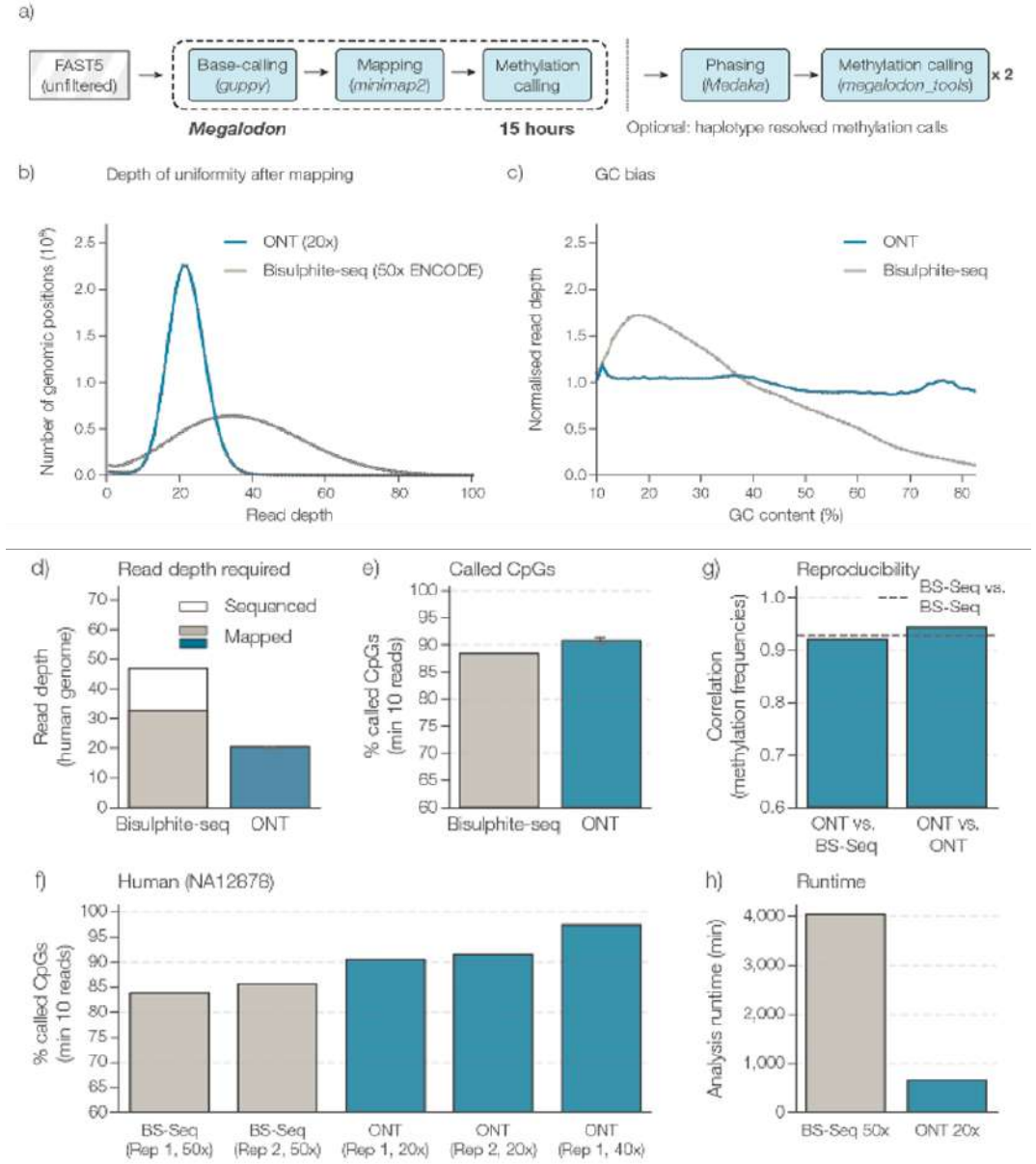
1. Histone modifications
2. DNA Methylation



Directly detect DNA and RNA methylation with high reproducibility and low bias

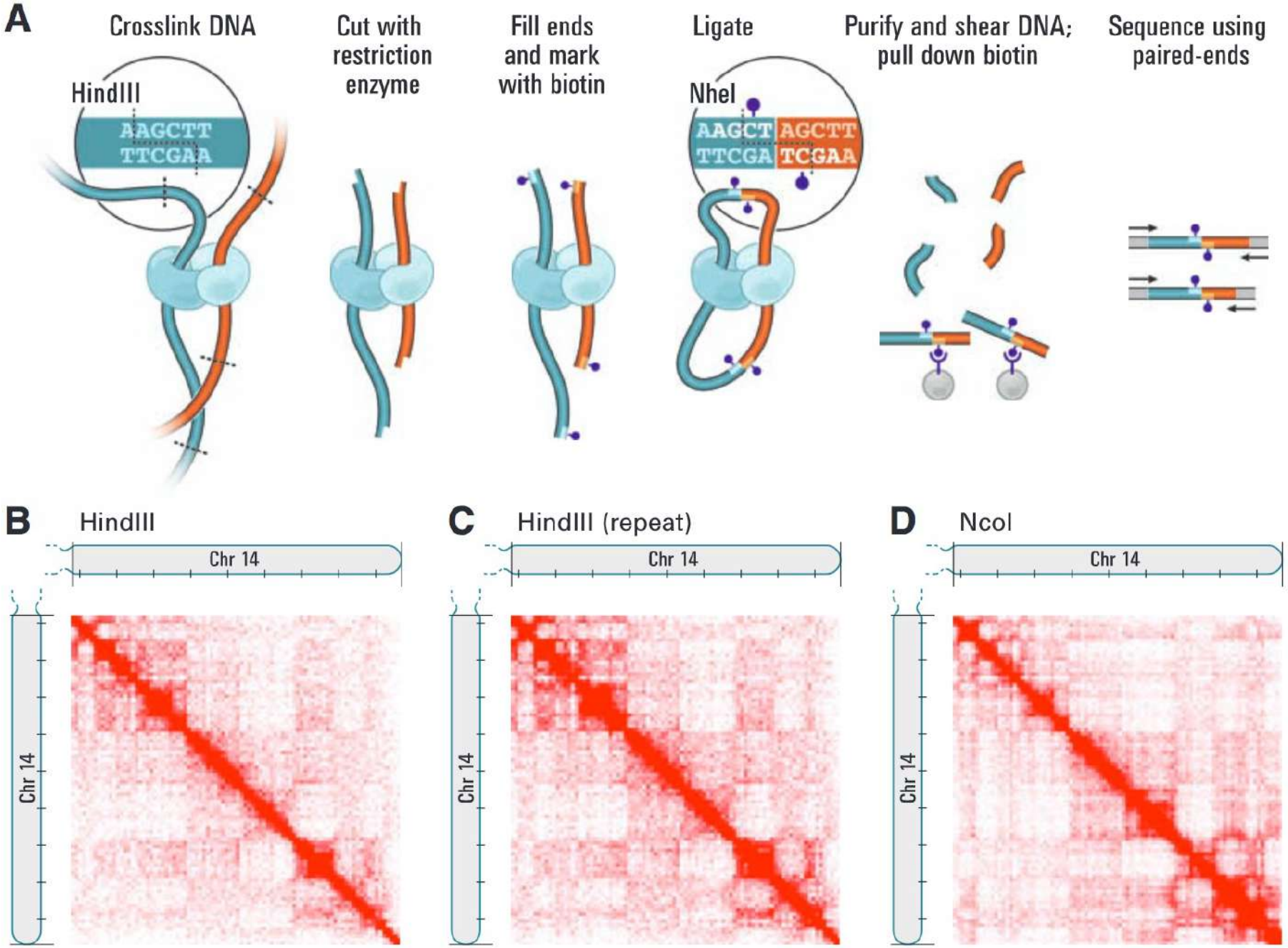
Using nanopore sequencing, researchers have directly identified DNA and RNA base modifications at nucleotide resolution, including 5mC, 5hmC, 6mA, and BrdU in DNA, and m6A in RNA, with detection of other natural or synthetic epigenetic modifications possible through training basecalling algorithms.

One of the most widespread genomic modifications is 5-methylcytosine (5mC), which most frequently occurs at CpG dinucleotides. Compared to whole-genome bisulfite sequencing, the traditional method of 5mC detection, nanopore technology calls a higher number of CpG positions in the genome, requires less sequencing data, and shows more even genomic coverage with considerably lower GC bias; analysis runtime is also significantly shorter (Figure 1).

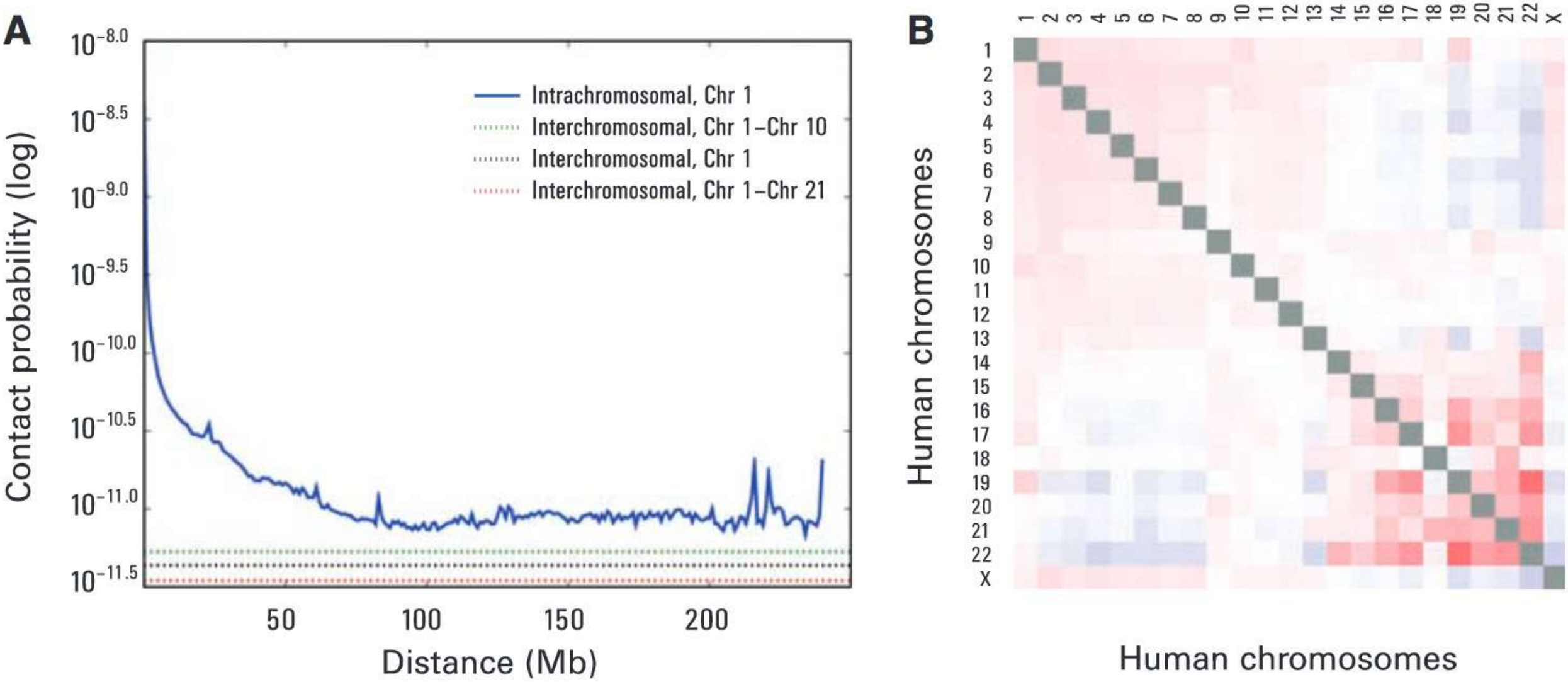


https://nanoporetech.com/applications/investigation/epigenetics-and-methylation-analysis

~~New~~ techniques that transformed genomics in addition to long reads: Chromosome conformation capture (previously used epigenetic dimension of chromosomes)



Lieberman-Aiden (2009) Science 10.1126/science.1181369



Lieberman-Aiden (2009) Science 10.1126/science.1181369

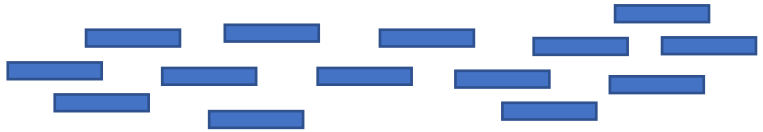
Summary: evolution of sequence reads

DNA fragment (300-600bp)

Paired end reads ; ~150bp sequenced on both ends of a DNA fragment. We know these read pairs belong to the same fragment.

Illumina ; high accuracy ; “short” reads
Still needed with its **high depth and accuracy**

DNA fragment (1-30+++ kb)



All these reads belong to the same fragment

10X technology sequenced in Illumina ; high accuracy ; ‘linked’ reads

Whole read (very long) sequenced!

Oxford Nanopore or Pacbio ; still errorneous; **contain modification information**

Analysis approaches

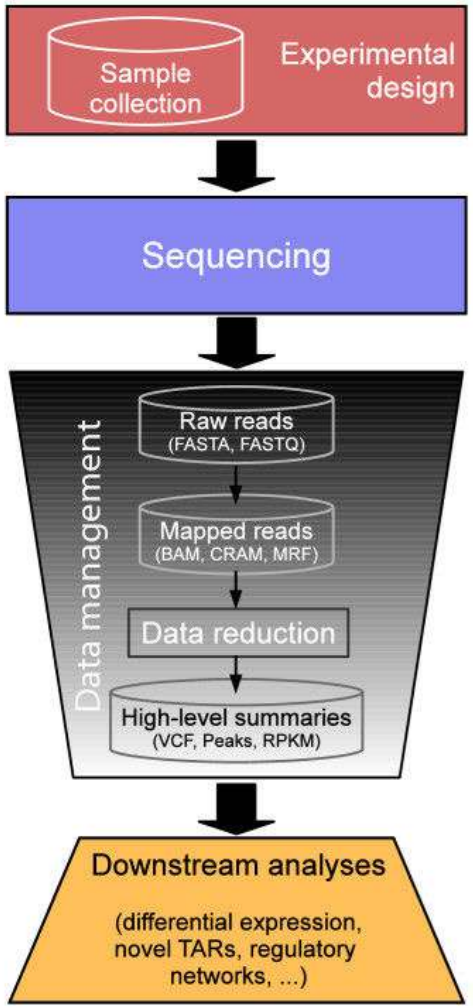
Sboner *et al. Genome Biology* 2011, **12**:125
http://genomebiology.com/2011/12/8/125

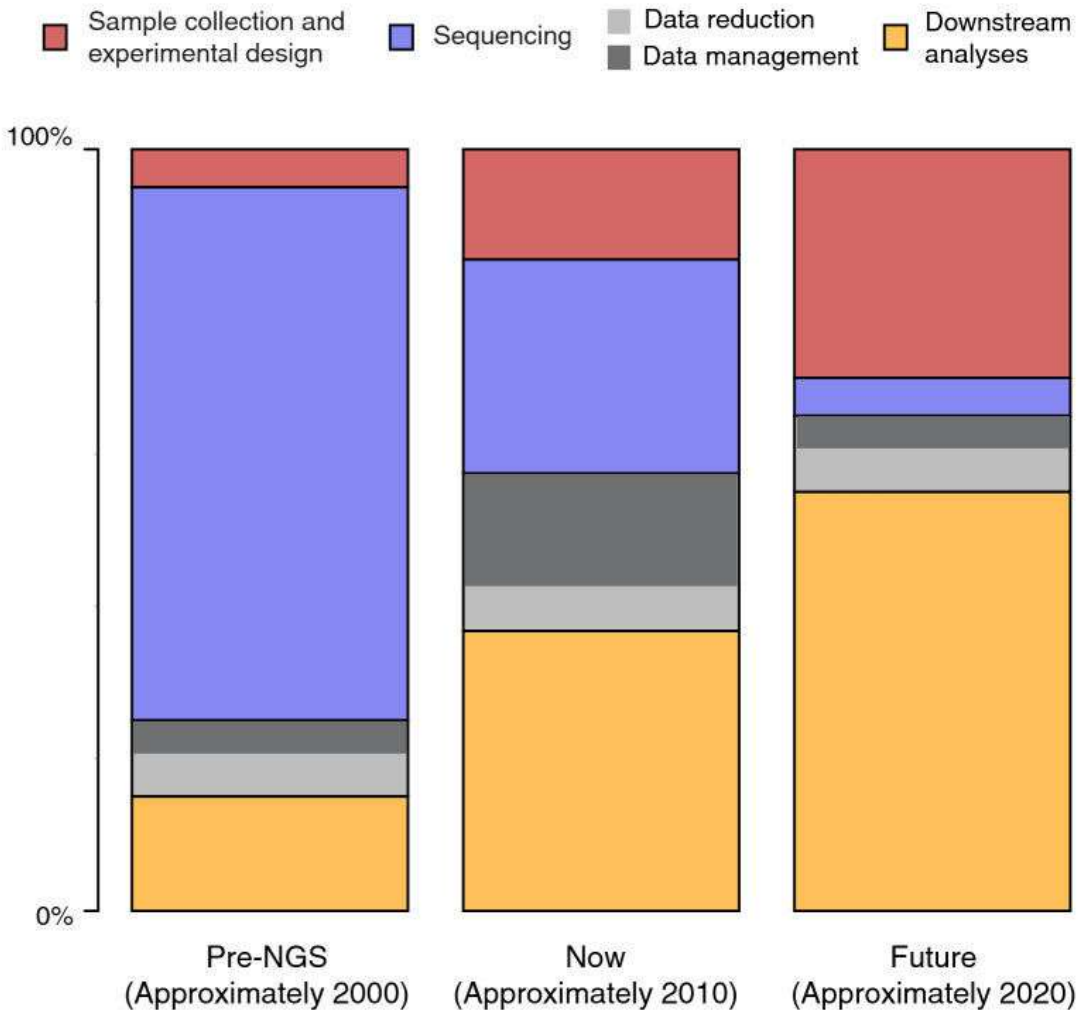
Genome Biology

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]





Four 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

- **Statistics**

What is an alignment? (mapping)

Align the following two sequences:

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

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

Scoring scheme is needed:

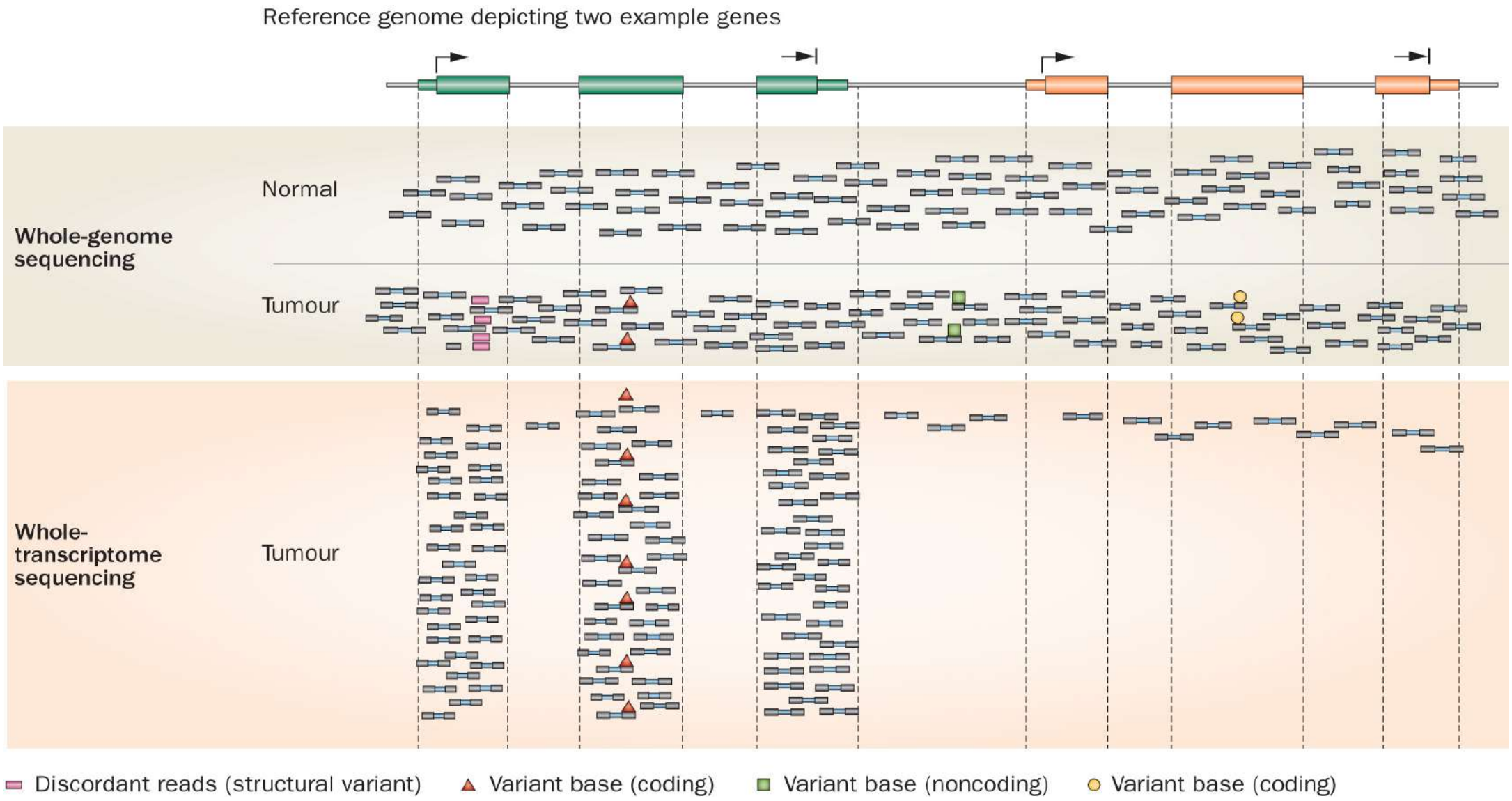
1 for match

-1 for mismatch

-2 for gap

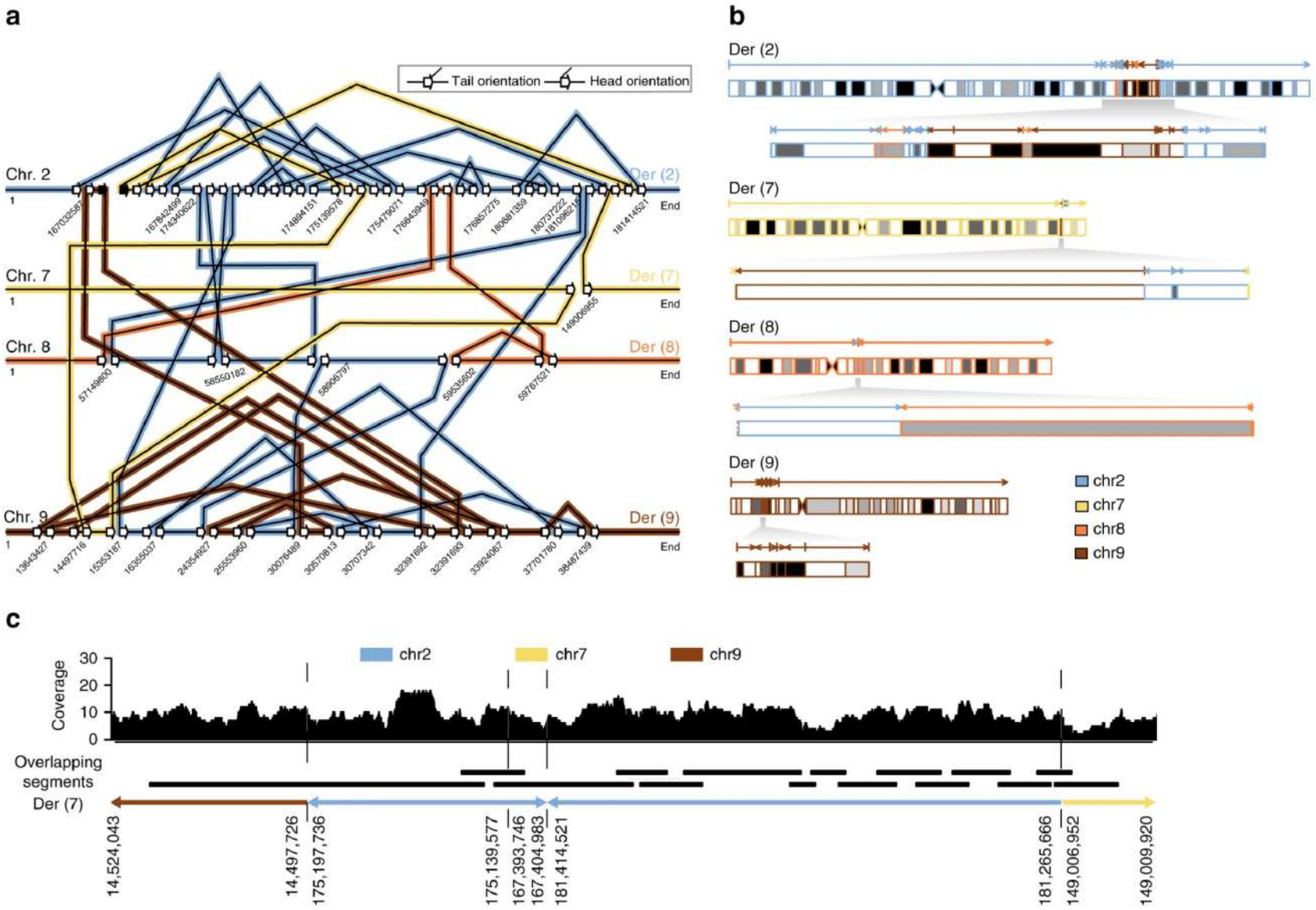
```
insertions / deletions (indels) mismatches
Which alignment is better?
```

Mapping



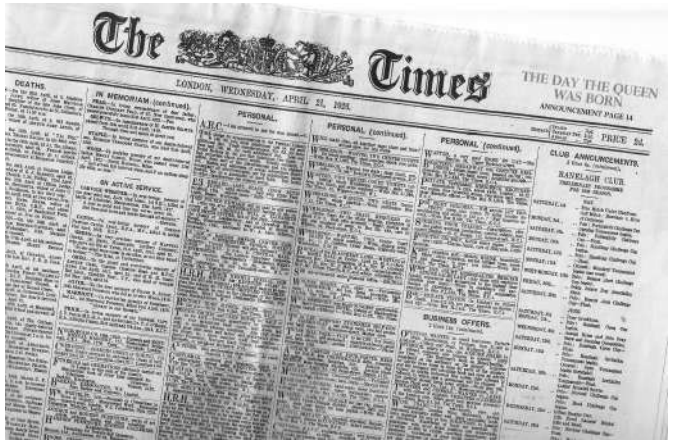
doi:10.1038/nrgastro.2012.126

Long read able to uncover long SV



https://www.nature.com/articles/s41467-017-01343-4

Assembly

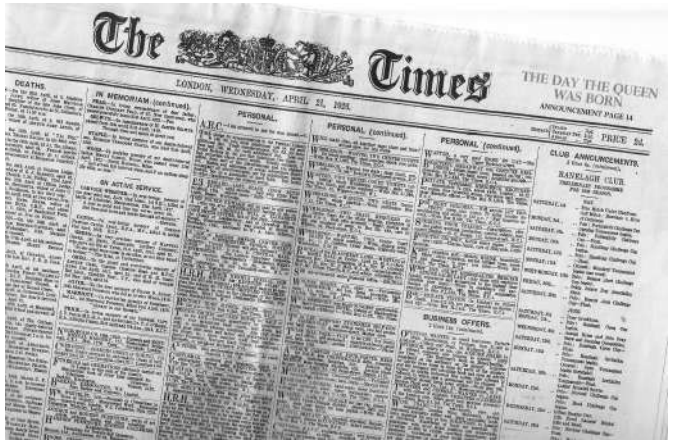


Genome
(3.000.000 letters)

Sequencing →

Reads
(50-500 letters each)

Assembly →



Genome
(3.000.000 letters)

Long read + HiC to produce a reference assembly (since 2017)

a

Most genomes in this N50 range

Contig N50

Number of vertebrate genomes		Less than 10 kb	10 kb to 100 kb	100 kb to 1 Mb	1 Mb to 10 Mb	Greater than 10 Mb
Diploid human	21	–	8	6	3	4
Non-human mammal	196	34	34	14	1	3
Non-mammal	193	43	133	14	3	0
Total	410	77	285	34	7	7

Genomes with highest N50

b

	Goat CHIR_1.0	Goat ARS1	Human GRCh38
Total sequence length	2.6 Gb	2.9 Gb	3.2 Gb
Total assembly gap length	140 Mb	38 Mb	160 Mb
Gaps between scaffolds	411	0	349
Number of scaffolds	77,431	29,907	735
Scaffold N50	14 Mb	87 Mb	67 Mb
Number of contigs	337,494	30,399	1,385
Contig N50	18.9 kb	26.2 Mb	56.4 Mb
Number of chromosomes and plasmids	30	31	25

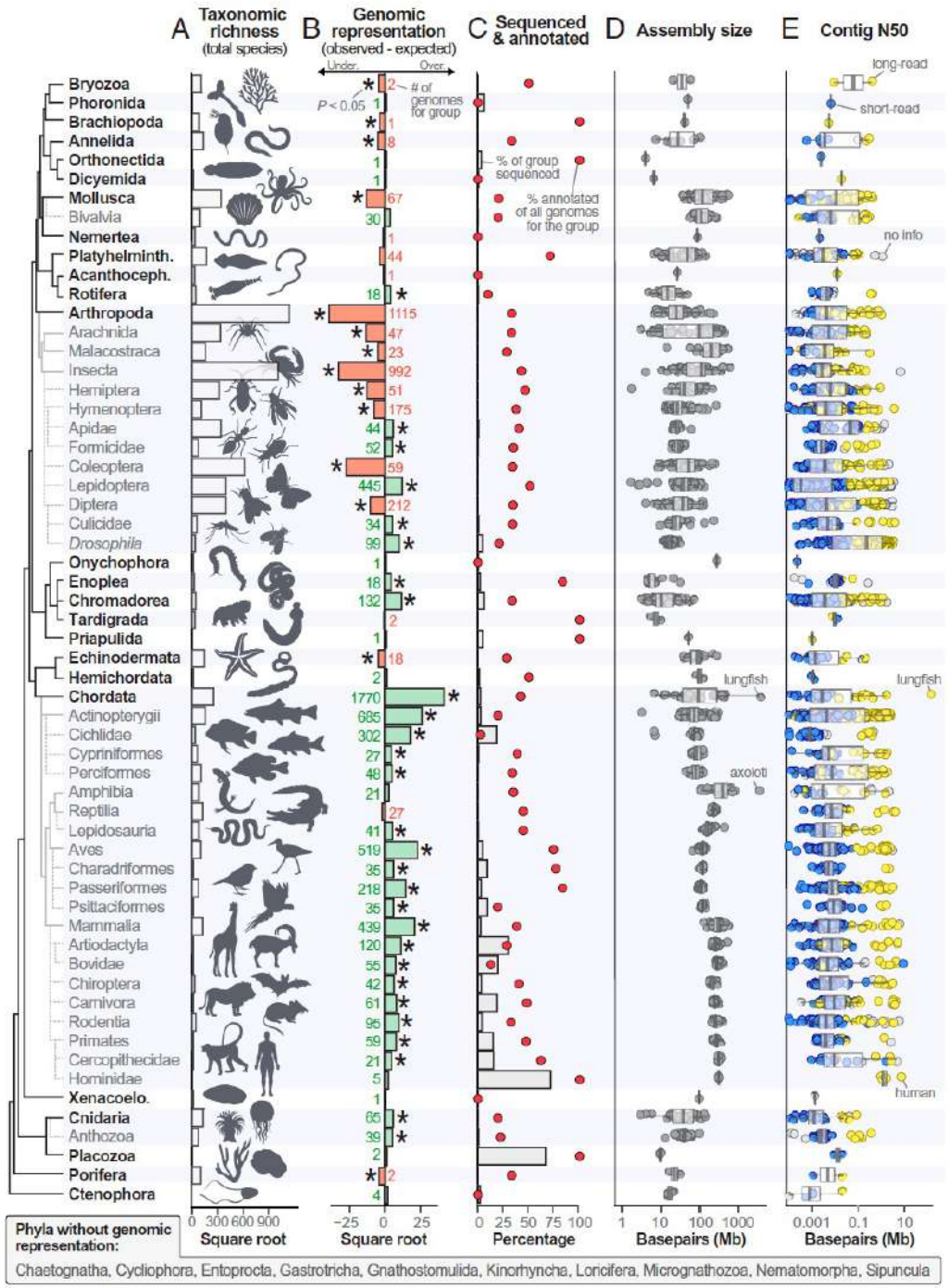
c

https://www.nature.com/articles/ng.3824

Toward a genome sequence for every animal: Where are we now?

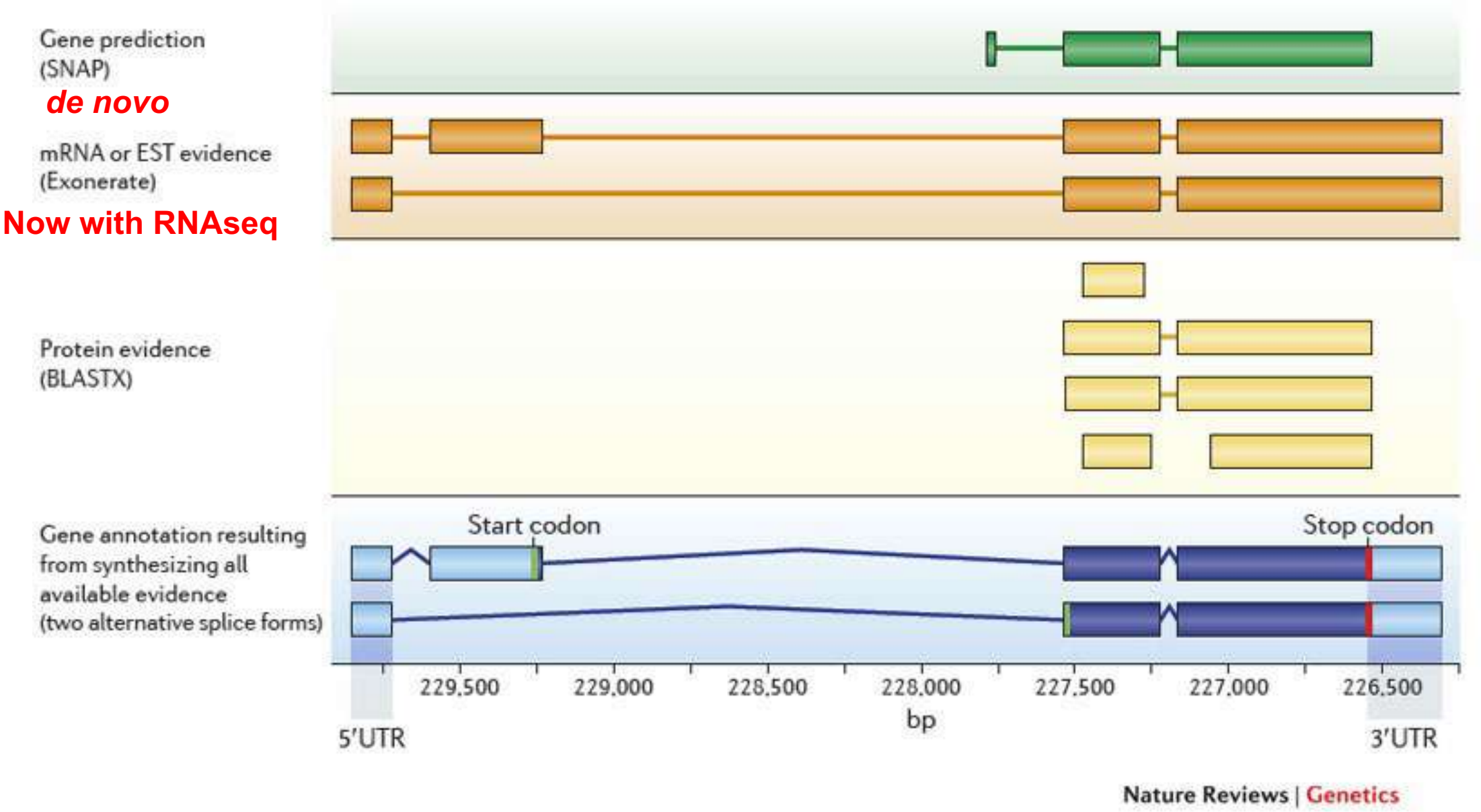
Scott Hotaling[a,1], Joanna L. Kelley[a], and Paul B. Frandsen[b,c,d,1]

Edited by Gene E. Robinson, University of Illinois at Urbana–Champaign, Urbana, IL, and approved October 28, 2021 (received for review August 4, 2021)

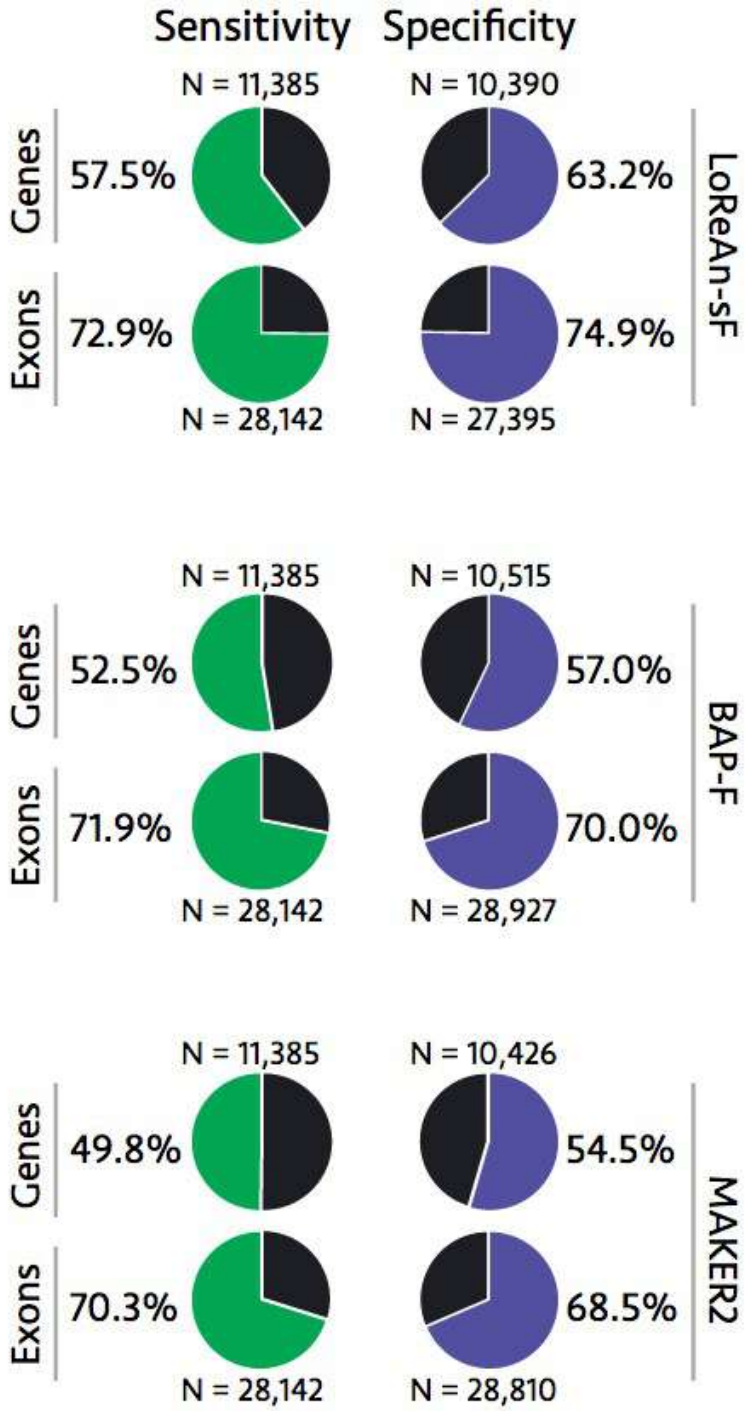


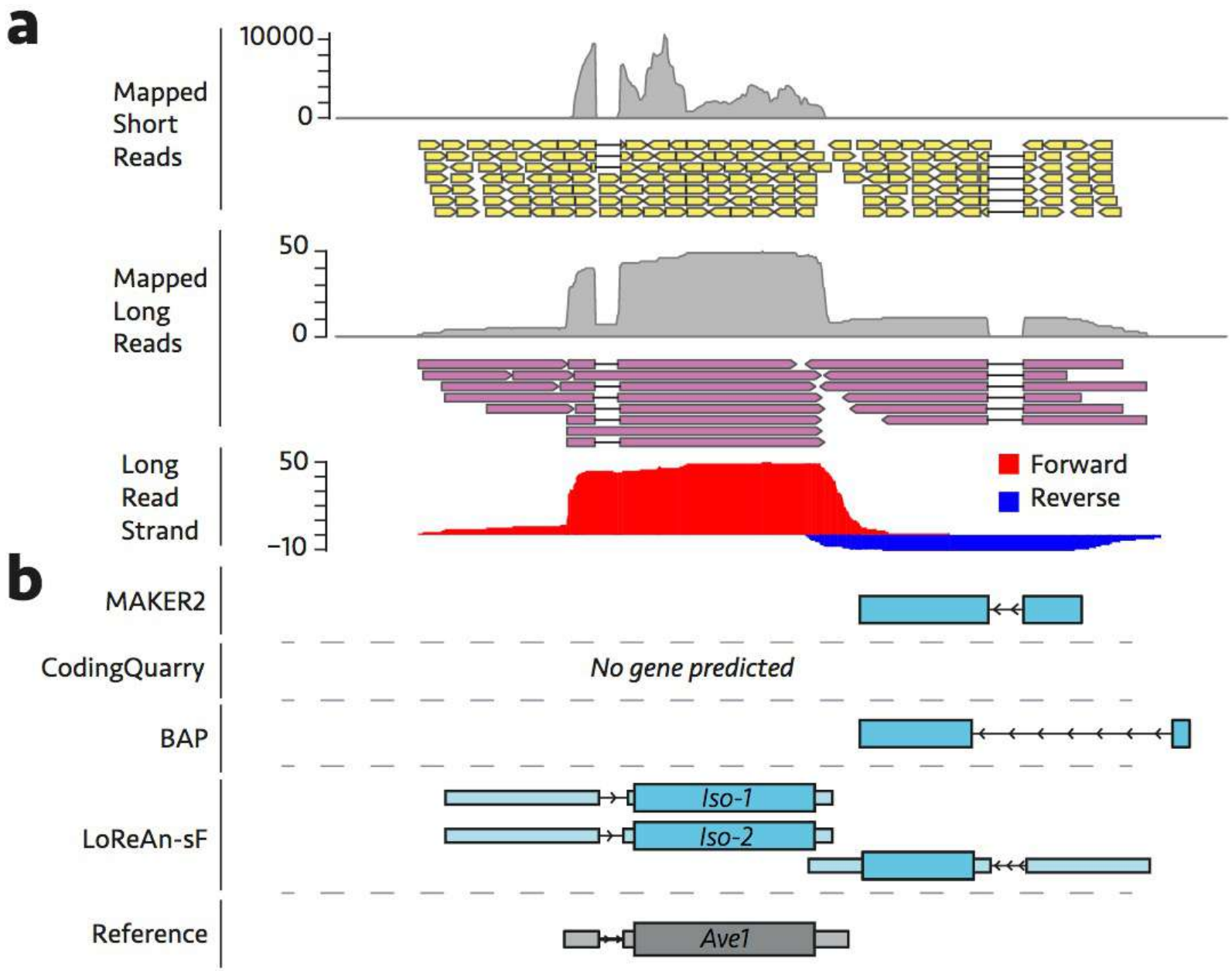
https://www.pnas.org/content/118/52/e2109019118

Annotation



Yandell and Ence Nature Genetics Review (2012)





Cook *et al* (2017) Plant Physiology

Case studies - genomics

Scenarios now and then

1. [lab/hospital/mountain/sea] Collect samples (1.1, 1.2, 1.3…)
2. [lab/hospital] Extract DNA (2.1, 2.2, 2.3…)
3. [lab/hospital/company] Sequencing (3.1, 3.2, 3.3…)
4. [lab/company] Analysis
5. [lab/hospital] Report

Weeks

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

Minutes

(diagnostic, cheaper, larger-scale..)

Current and future

Darwin TREE of LIFE

EARTH BIOGENOME PROJECT

- 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)

Classical genetics

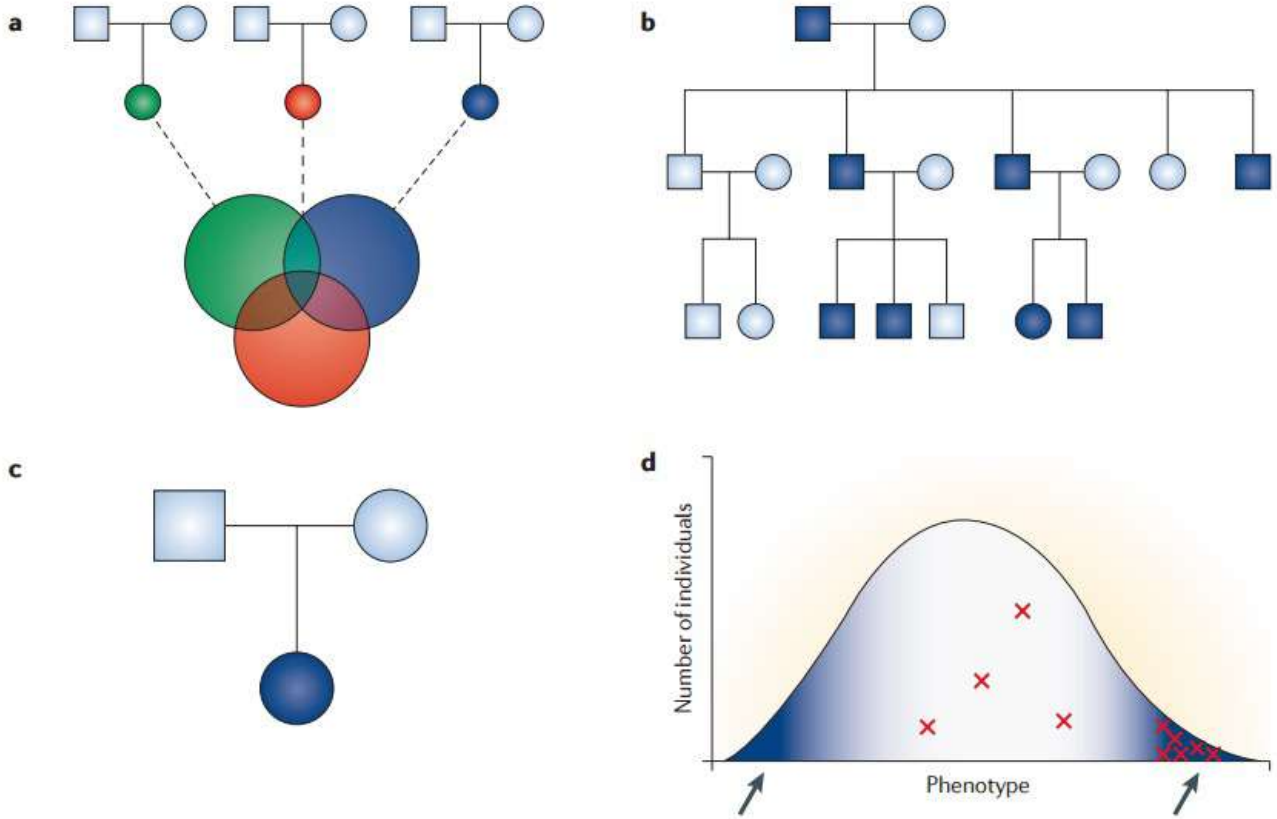
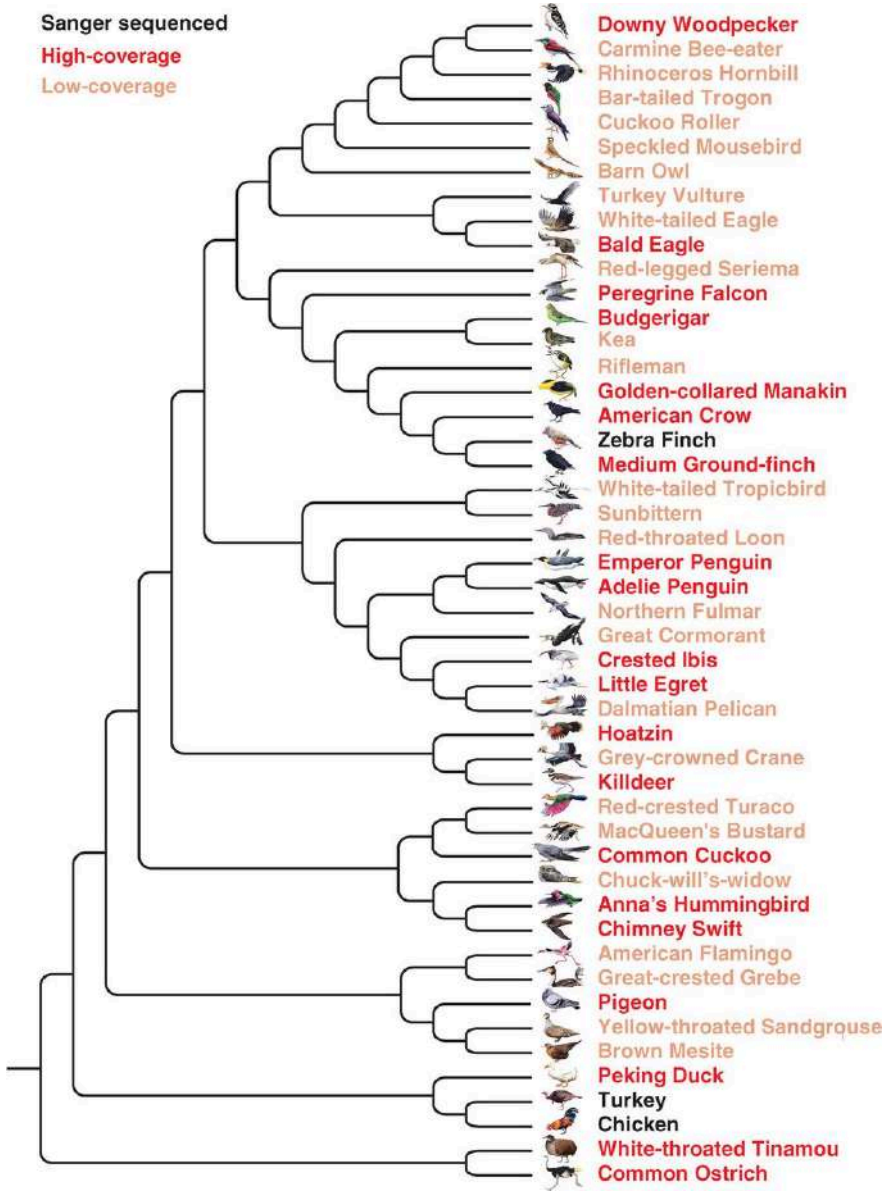


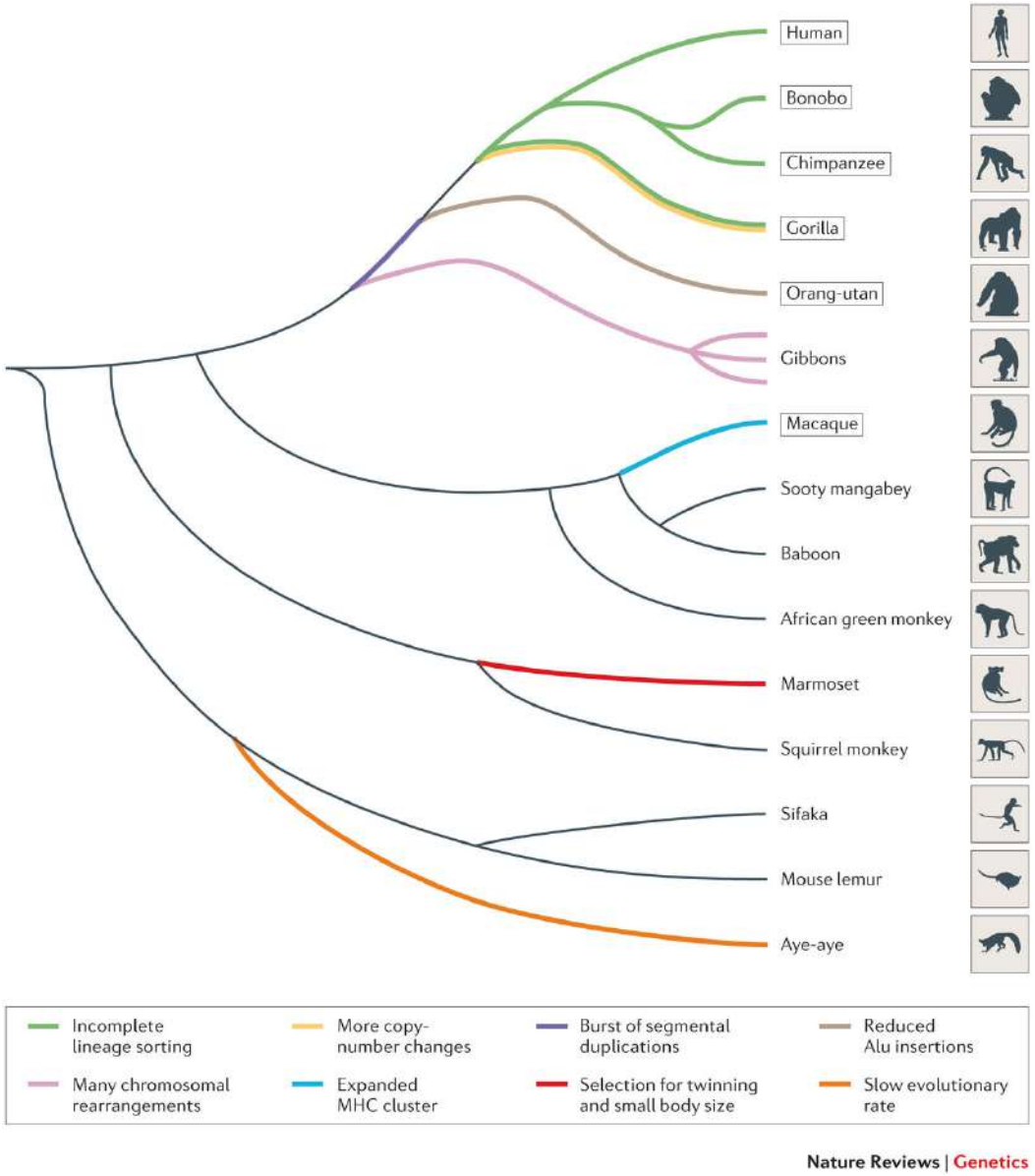
Figure 2 | **Strategies for finding disease-causing rare variants using exome sequencing.** Four main strategies are illustrated. **a** | Sequencing and filtering across multiple unrelated, affected individuals (indicated by the three coloured circles). This approach is used to identify novel variants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



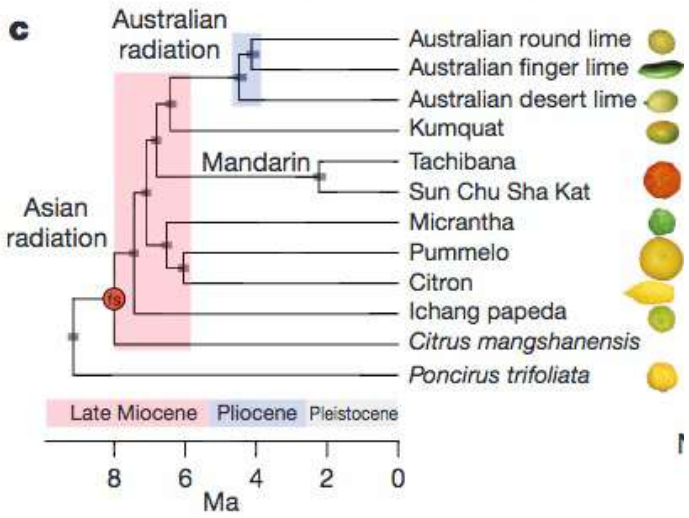
Guojie Zhang et al. Science (2014)

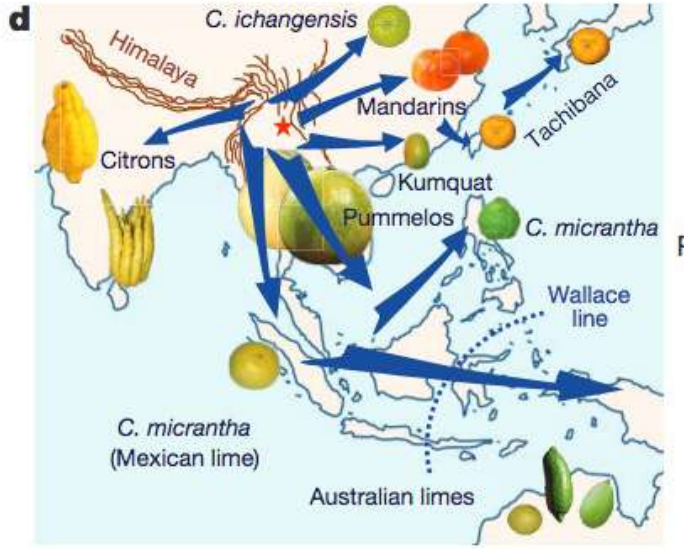


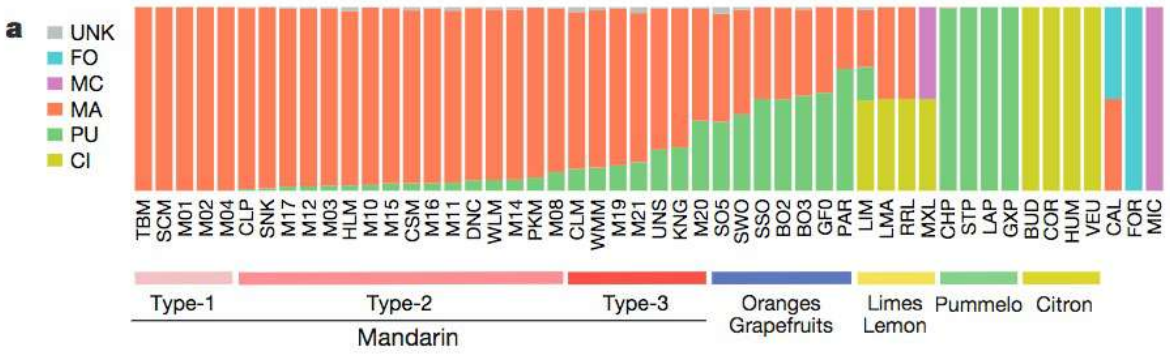
Roger & Gibbs Nature Reviews Genetics (2014)

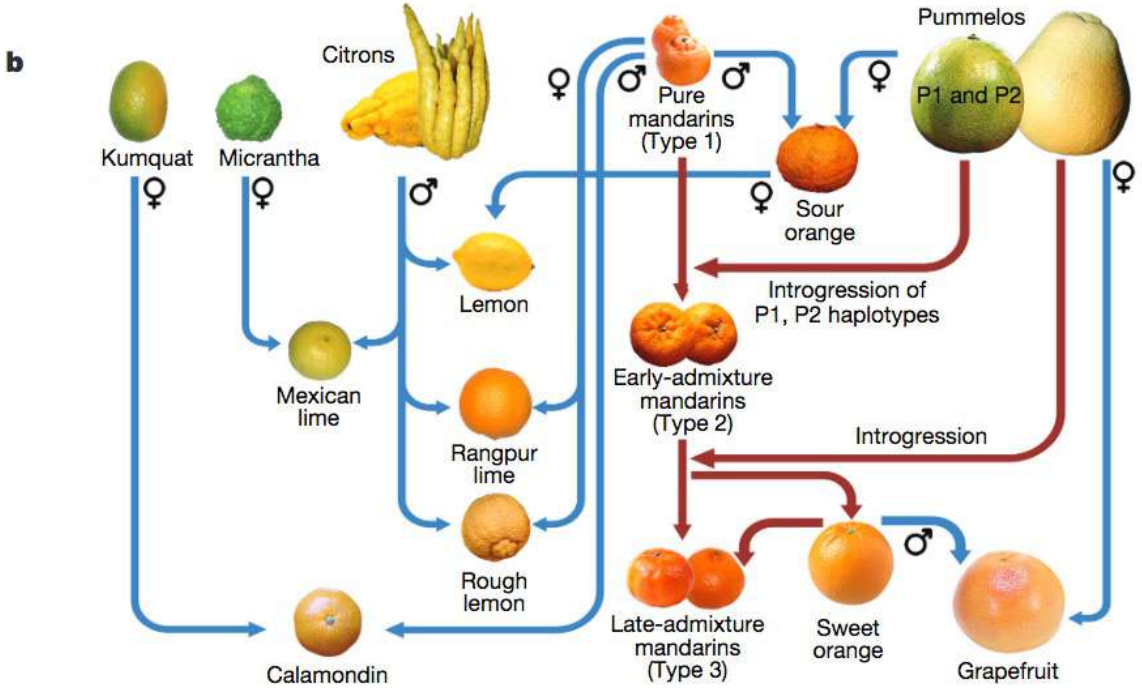
Comparative genomics

Genomics of the origin and evolution of Citrus



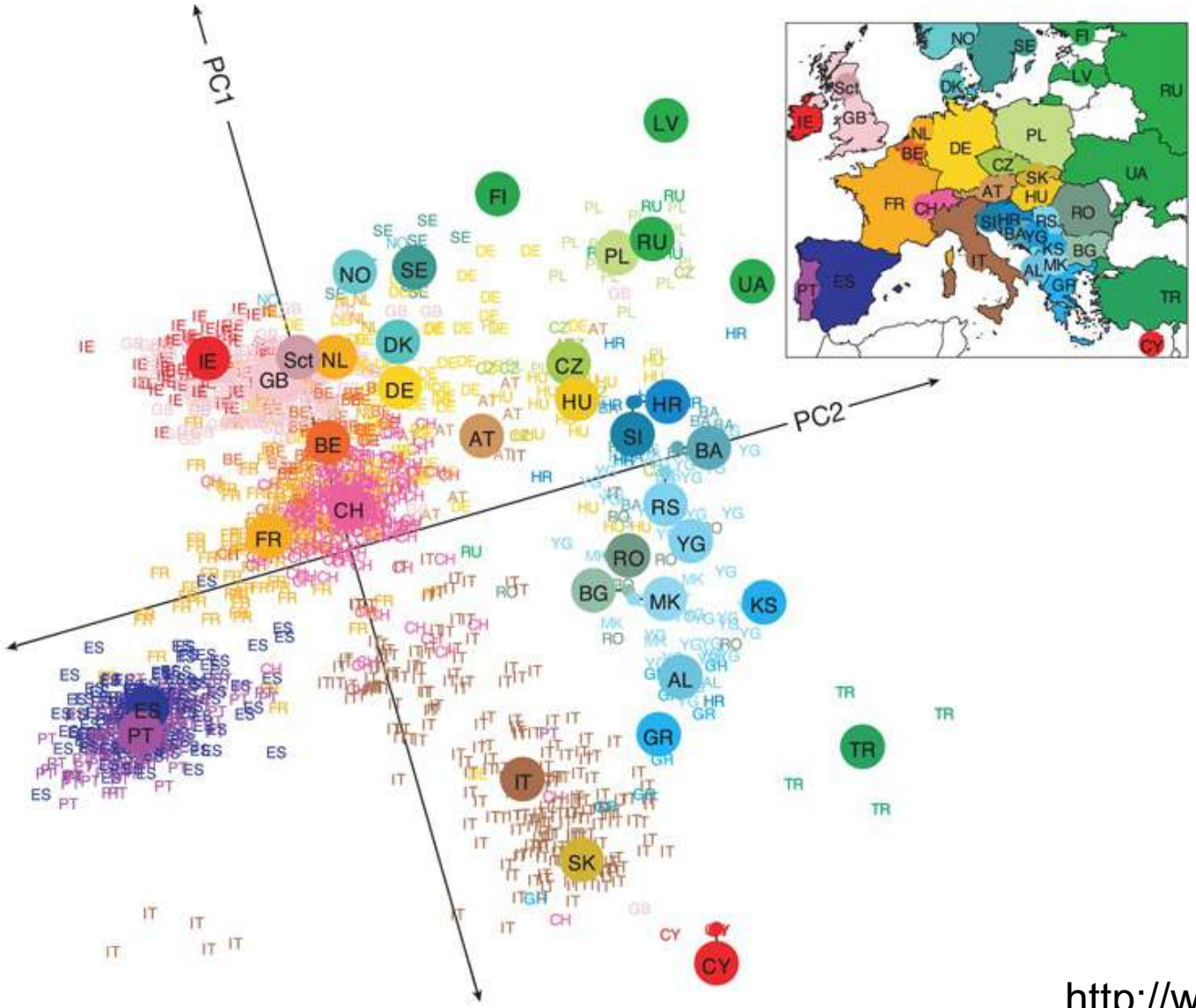




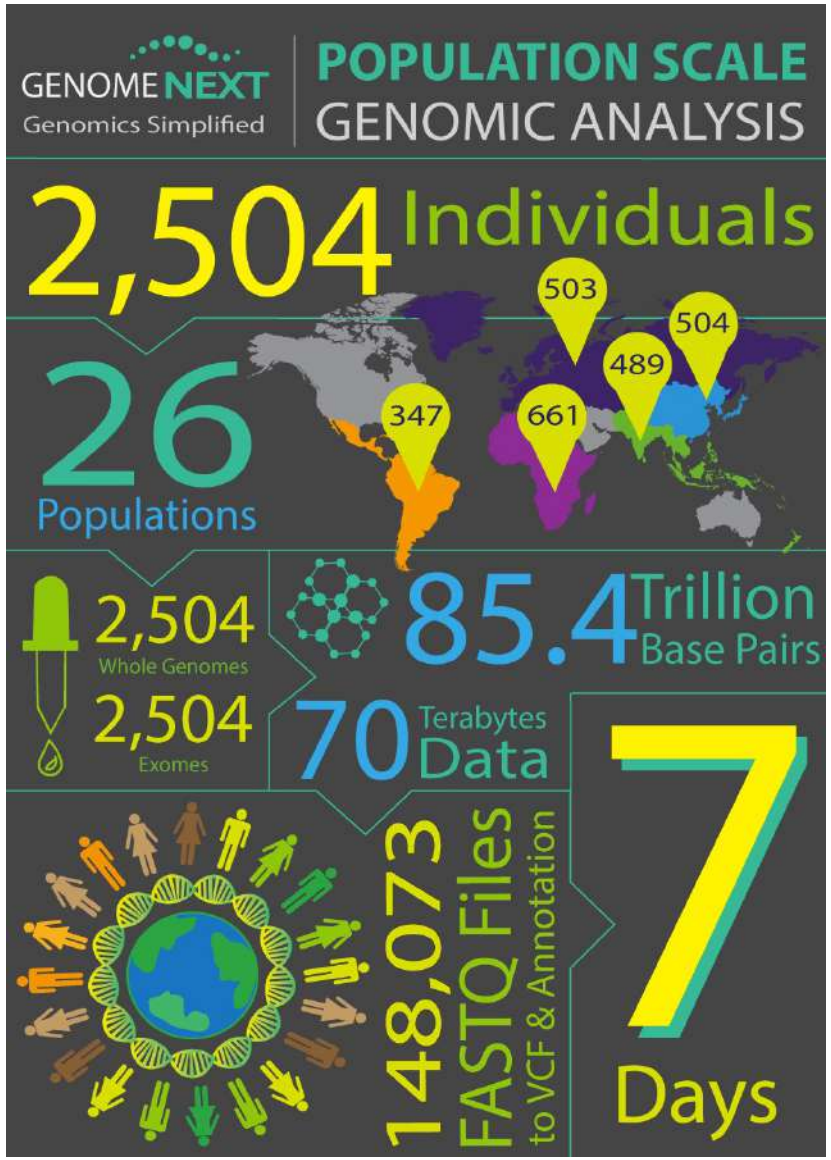


Wu *et al.*, Nature (2018)

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/

FOCUS ON GENOMES OF ICELANDERS

ARTICLES

nature genetics

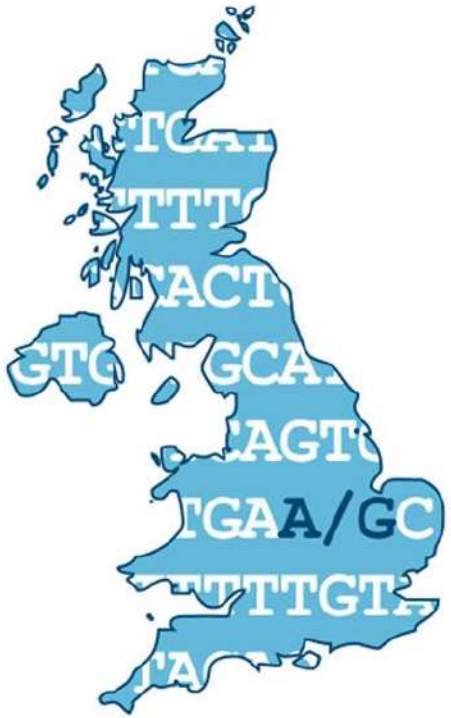
Large-scale whole-genome sequencing of the Icelandic population

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

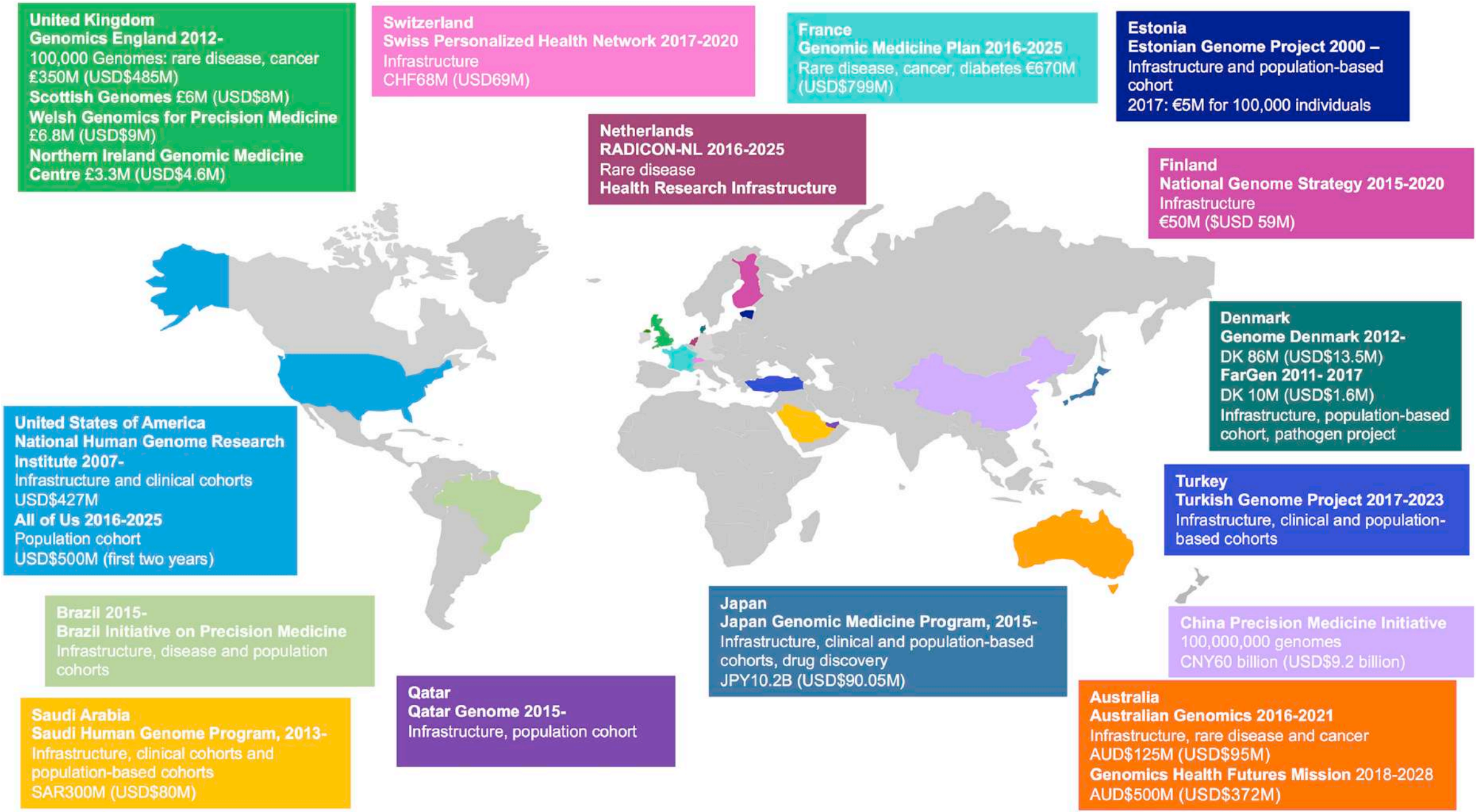
https://www.uk10k.org/

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.



Stark *et al* (2019) AJHG

臺灣
Taiwan Biobank

The Cumulative 累計收案數

統計至2019年01月31日止(請按此)

社區民眾收案數

109,059
參與個案總數

22,502
完成第一輪追蹤個案總數

醫學中心患者收案數

1,862
參與個案總數

320
完成第一輪追蹤個案總數

8
完成第二輪追蹤個案總數

The Cumulative 累計收案數

統計至2019年07月31日止(請按此)

社區民眾收案數

118,548
參與個案總數

24,936
完成第一輪追蹤個案總數

醫學中心患者收案數

3,145
參與個案總數

659
完成第一輪追蹤個案總數

104
完成第二輪追蹤個案總數

累計收案數

統計至2021年08月31日止(請按此)

社區民眾收案數

151,406
參與個案總數

37,508
完成第一輪追蹤個案總數

醫學中心患者收案數

7,387
參與個案總數

1,418
完成第一輪追蹤個案總數

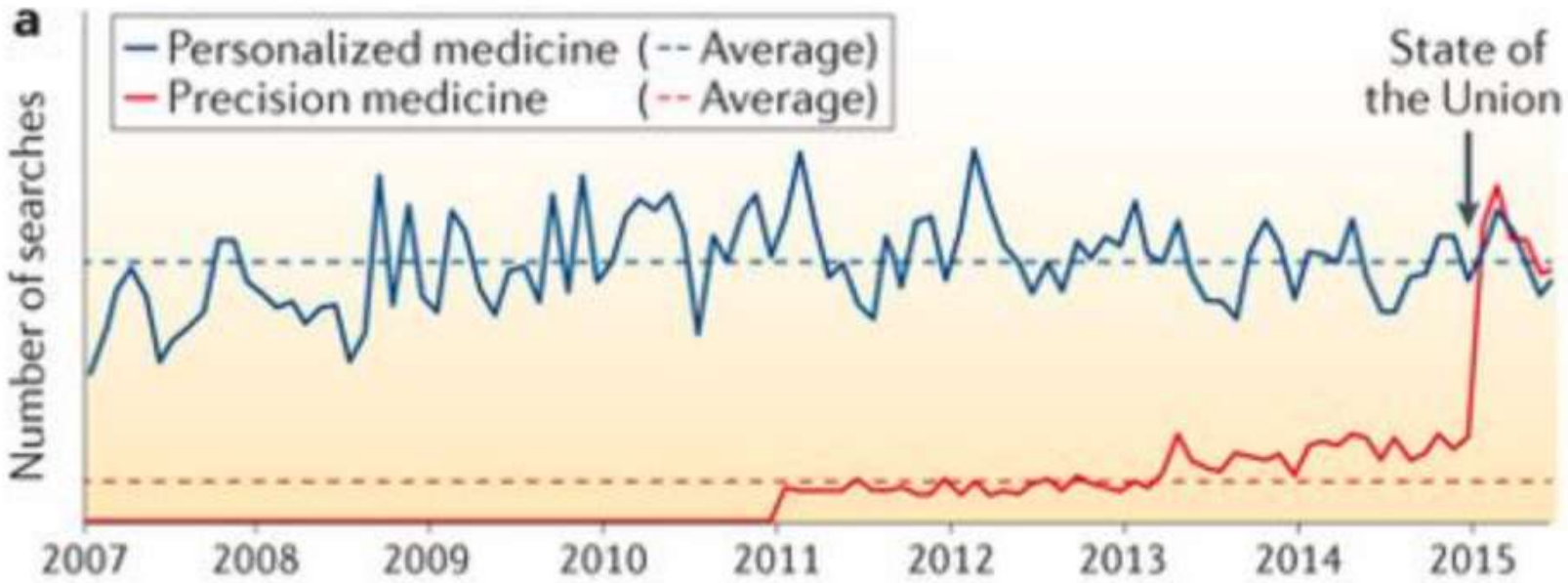
422
完成第二輪追蹤個案總數

98
完成第三輪追蹤個案總數

8
完成第四輪追蹤個案總數

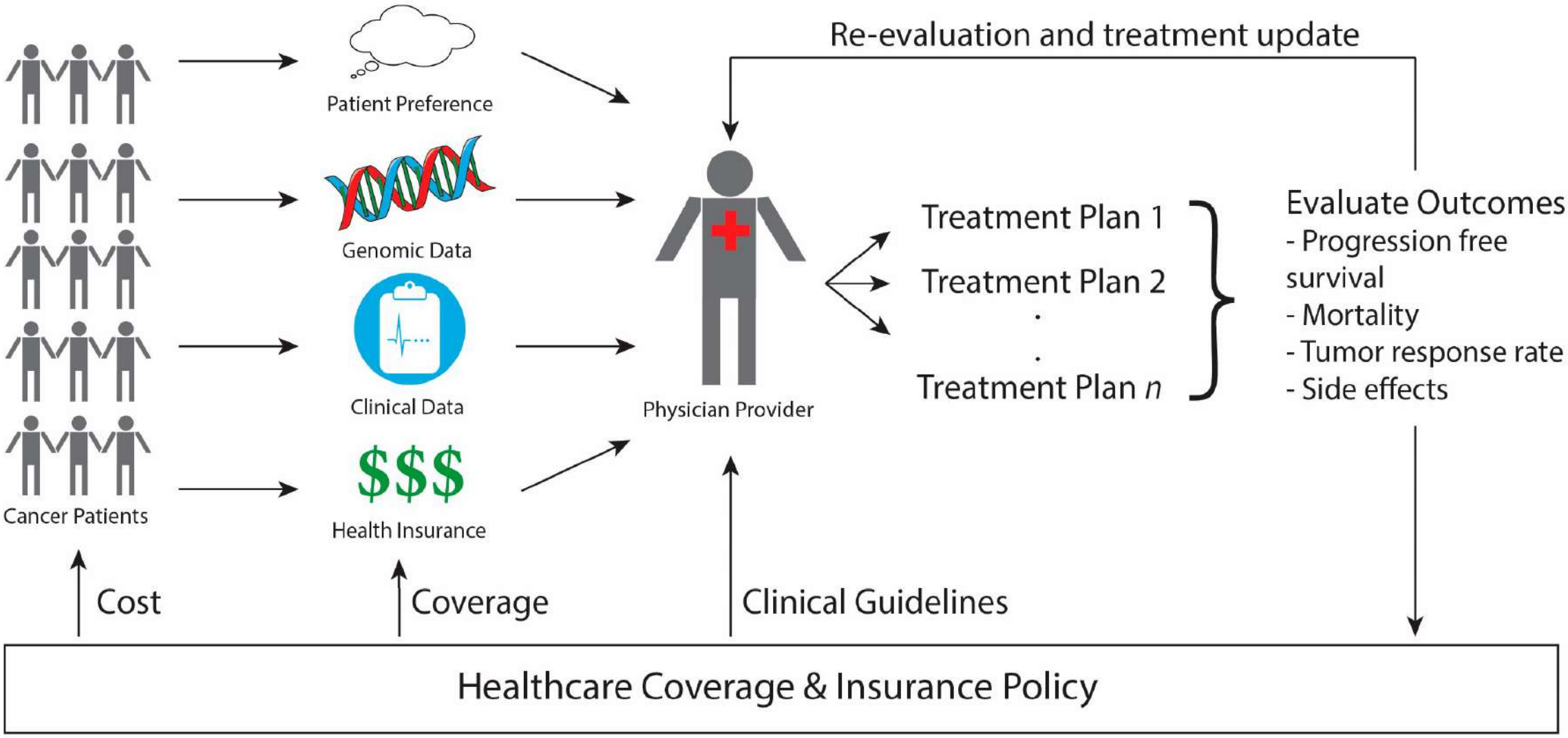
https://www.twbiobank.org.tw/new_web/index.php

Precision medicine
精準醫學



Ashley (2016) Nature Review Genetics

Outline of precision medicine



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

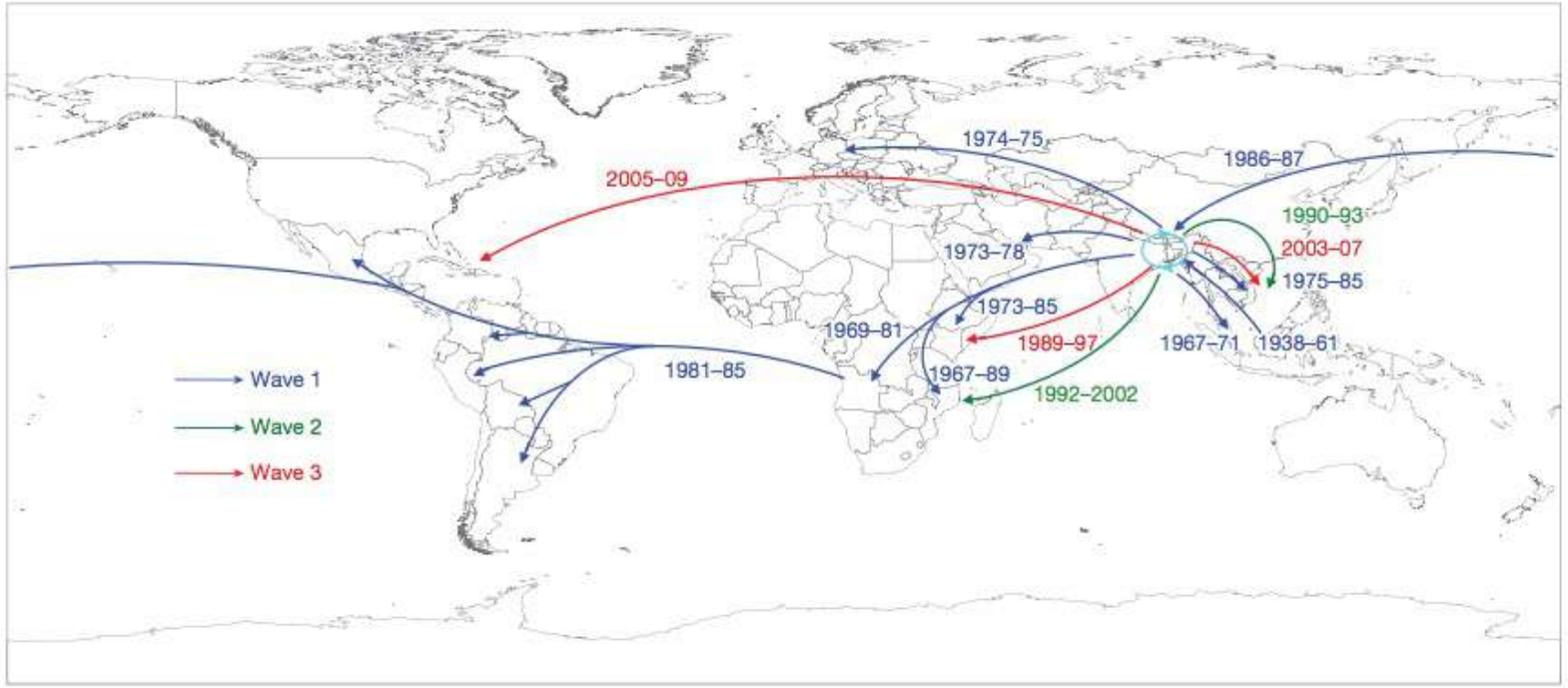
Summary of outcomes in Oncology PM Studies

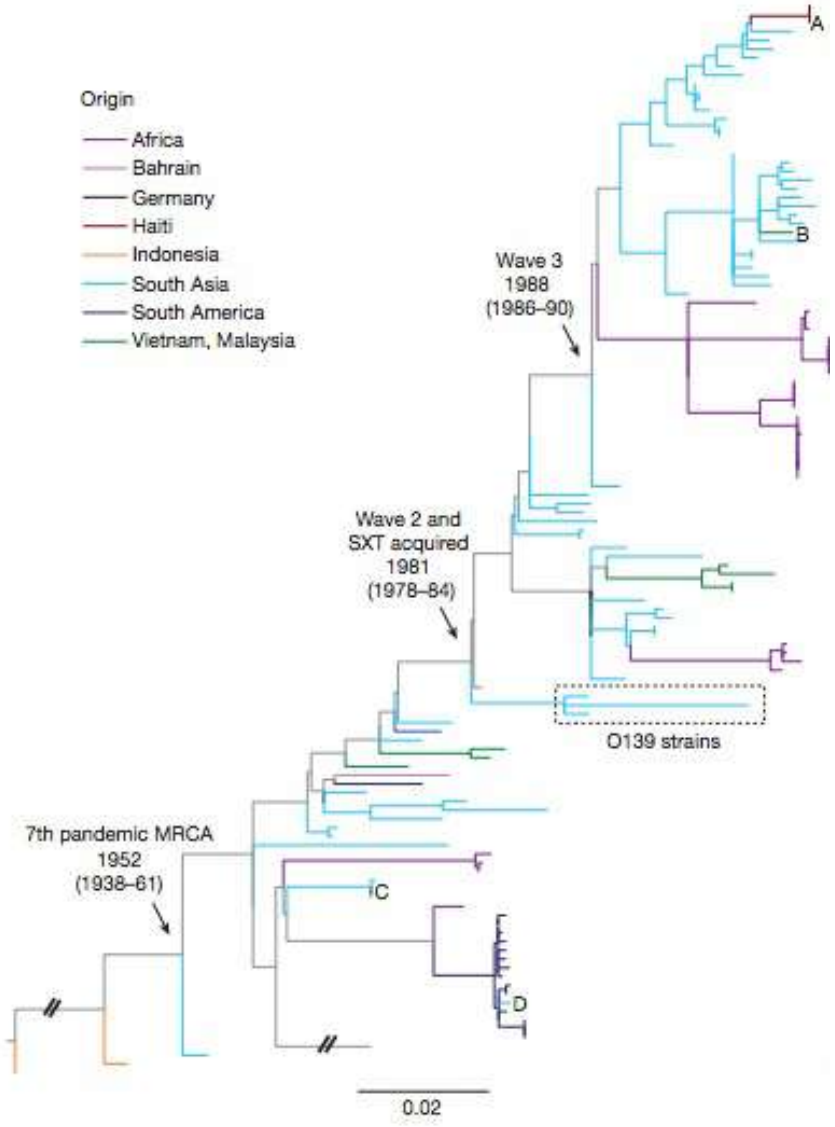
Study	Sample Size	Most Prevalent Tumor Types	Outcomes Reported
Tsimberidou et al. *Clin. Cancer Res.* **2012** [5]	291 patients with one molecular aberration (175 treated with matched therapy, 116 control)	Colorectal, melanoma, lung, ovarian	Matched group had improved ORR (27% vs. 5%), TTF (median 5.2 vs. 2.2 month), OS (median 13.4 vs. 9.0 month)
Radovich et al. *Oncotarget* **2016** [6]	101 patients with sequencing and follow up (44 treated with matched therapy, 57 control)	Soft tissue sarcoma, breast, colorectal	Matched group had improved PFS (86 vs. 49 days)
Schwaederle et al. *Mol. Cancer Ther.* **2016** [7]	180 patients with sequencing and follow up (87 treated with matched therapy, 93 control)	Gastrointestinal, breast, brain	Matched group had improved PFS (4.0 vs. 3.0 month), TRR (34.5% vs. 16.1% achieving SD/PR/CR)
Kris et al. *JAMA* **2014** [8]	578 patients with oncogenic driver and followup (260 with matched therapy, 318 control)	Lung only	Matched group had improved survival (median 3.5 vs. 2.4 years)
Aisner et al. *J. Clin. Oncol.* **2016** [9]	187 patients with targetable alteration and follow up (112 with matched therapy, 74 control)	Lung only	Matched group had improved survival (median 2.8 vs. 1.5 years)
Stockley et al. *Genome Med.* **2016** [10]	245 patients with sequencing matched to clinical trials (84 on matched trial, 161 control)	Gynecological, lung, breast	Matched group had improved ORR (19% vs. 9%)
LeTourneau et al. *Lancet Oncol.* **2015** [11]	RCT with 195 patients with molecular aberration (99 treated with matched therapy, 96 control)	Gastrointestinal, breast, brain	No difference in PFS between groups

ORR = overall response rate, TTF = time to treatment failure, OS = overall survival, PFS = progression free survival, TRR = tumor response rate, SD = stable disease, PR = partial response, CR = complete response, RCT = randomized controlled trial. Matched group indicates patients matched to a therapy based on sequencing results.

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

Population genomics of pathogens





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

2013

8000家庭破碎 聯合國遭控傳染霍亂

2013-10-11 by：阿嘞 5725

一直以來，聯合國給世人的印象多是促進世界永續發展的正面印象，但對海地居民來說，聯合國卻成了當地人最恐懼的劊子手，最新報導就指出，因聯合國駐軍而散佈的霍亂已經造成8,000人死亡。

BBC綜合報導，聯合過派駐的維和部隊(UN peacekeepers)意外將細菌帶到海地境內，在當地造成霍亂大流行，自2010年爆發至今，霍亂已經在海地造成8,000人病死，這也讓海地成為目前全世界霍亂疫病最嚴重的地區。

聯合國是兇手

儘管許多調查指出聯合國就是霍亂源頭，但海地數度請願要求補償未果，現在海地的代表律師團就上訴紐約法院，控告聯合國是造成海地霍亂疫情的元凶。

2016

聯合國坦承：我們將霍亂帶進了海地

2016-08-19 by：泥仔 15040

將近六年的時間，聯合國終於承認海地的霍亂疫情與他們有關。到目前為止，已經有數十萬名居民感染上霍亂、一萬名海地人因霍亂而去世。

維和部隊惹的禍？

由於海地過去都沒有類似霍亂症狀的疾病，部分專家也發現海地的霍亂細菌種類與尼泊爾的種類是一樣的，因此懷疑是聯合國在尼泊爾的維和部隊將霍亂弧菌帶進海地。但將近六年來，聯合國一直都否認這樣的指控。

聯合國坦承與疫情爆發有關

在本周三(17)，聯合國副發言人哈奇(Farhan Haq)聲明：「過去幾年來，聯合國有鑑於海地初期的瘟疫爆發與我們有些關係，聯合國決定要多做些什麼。」他也強調聯合國會在接下來兩個月內有所行動。

ARTIC NETWORK

About

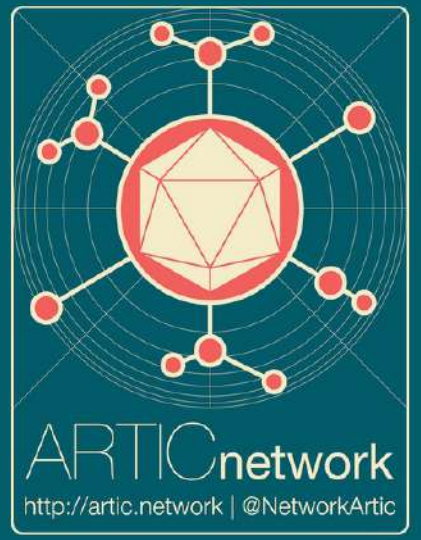
The Project

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

https://artic.network/ncov-2019

ARTIC NETWORK

nCoV-2019



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

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

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

https://artic.network/ncov-2019

Nextstrain

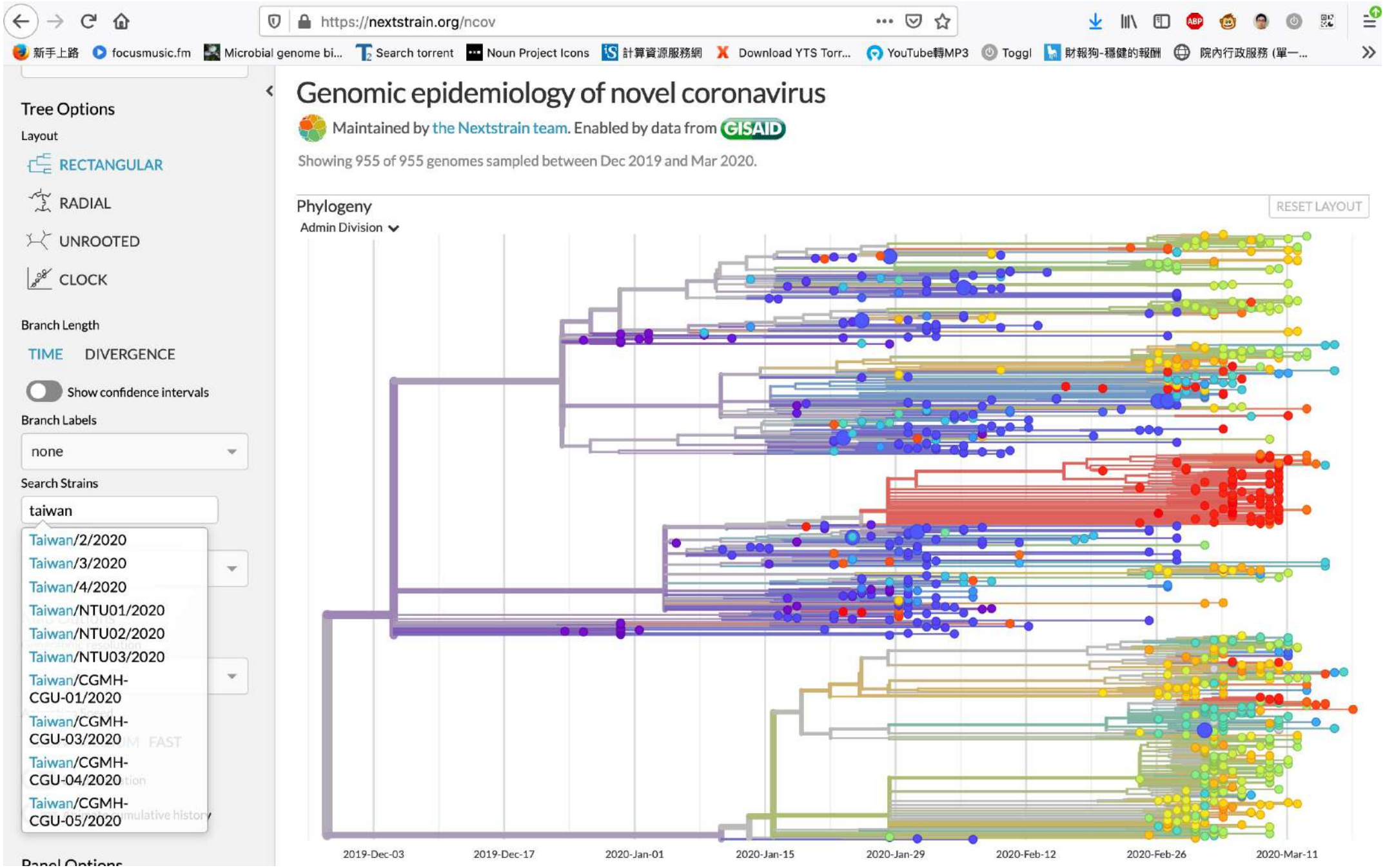
HELP DOCS BLOG LOGIN

Nextstrain

Real-time tracking of pathogen evolution

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

https://nextstrain.org/



https://nextstrain.org/

GISAID Initiative Retweeted

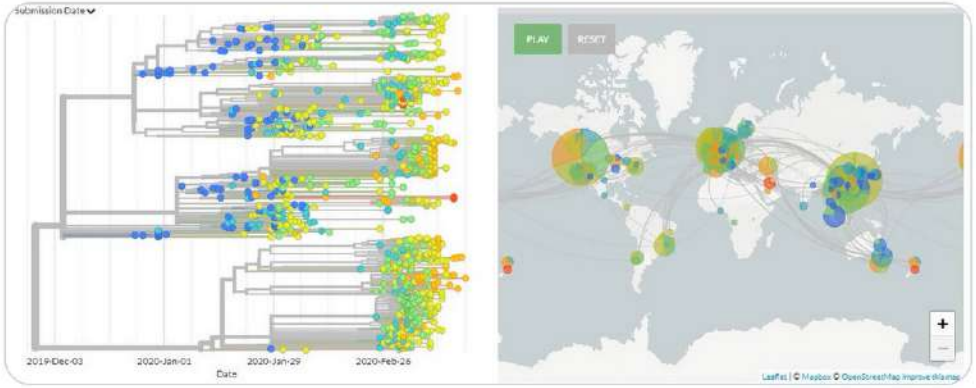
Nextstrain @nextstrain · Mar 6

nextstrain.org/ncov now updated with the first sequence from New Zealand 🇳🇿. The NZ seq groups with other sequences with travel history to Iran (circled), as expected. Sequenced by @ESRNewZealand @MathStorey @Joepdl @sciolato using @NetworkArtic & #RAMPART. Data via @GISAID

Nextstrain @nextstrain · 2h

Thanks to #opendata sharing by @dasmaninstitute @KUWAIT_MOH @FahdAlMulla @KATarinambraun @GageKMoreno @tcflab @dho_lab @ESRNewZealand @MathStorey @Joepdl @sciolato & @GISAID nextstrain.org/ncov is updated with 7 new sequences from Kuwait, Wisconsin, & New Zealand!

#COVID19



Catherine Moore @SmallRedOne · Mar 7

Replying to @SmallRedOne

Yesterday, the Public Health Wales Specialist virology Centre passed the first two positive samples in Wales to the team in the Pathogen Genomics Unit to perform whole genome sequencing

1 2 13

Catherine Moore @SmallRedOne · Mar 7

Today...

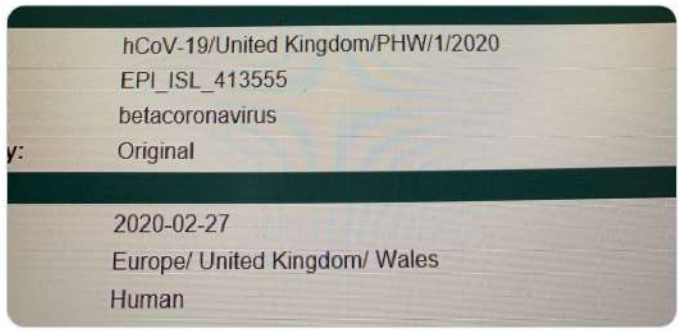
3 1 20

Catherine Moore @SmallRedOne · Mar 7

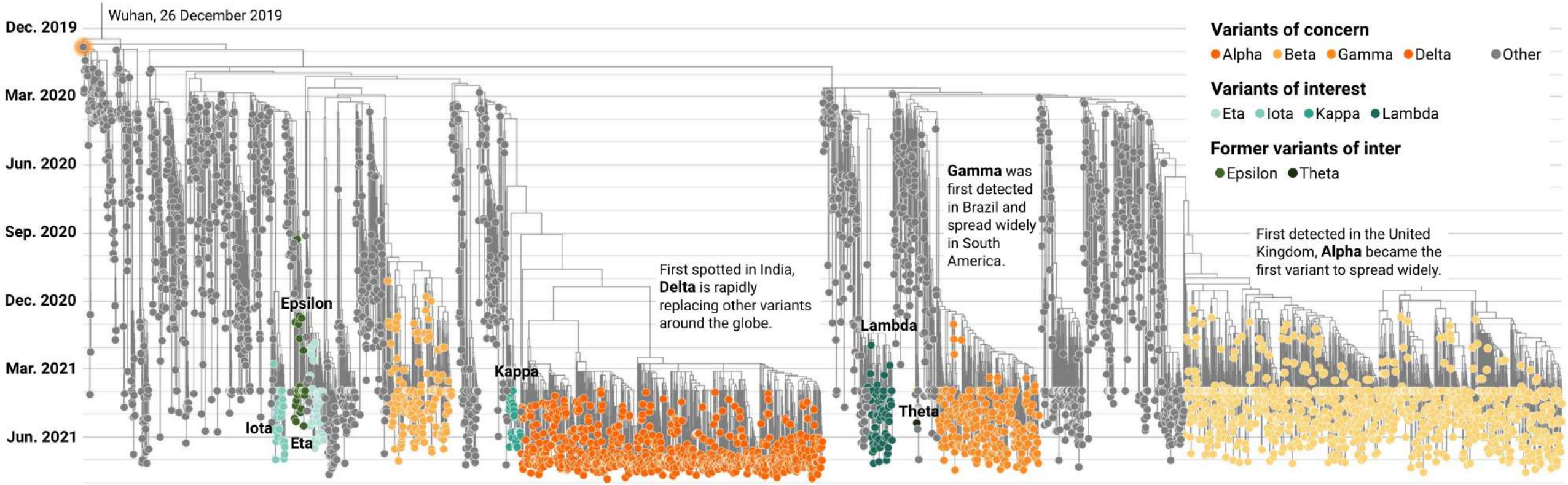
@tomrconnor tells me that our sequences are almost ready for public release through #GISAID for addition to the global dataset. This is incredible.

Catherine Moore @SmallRedOne · Mar 7

And we're live



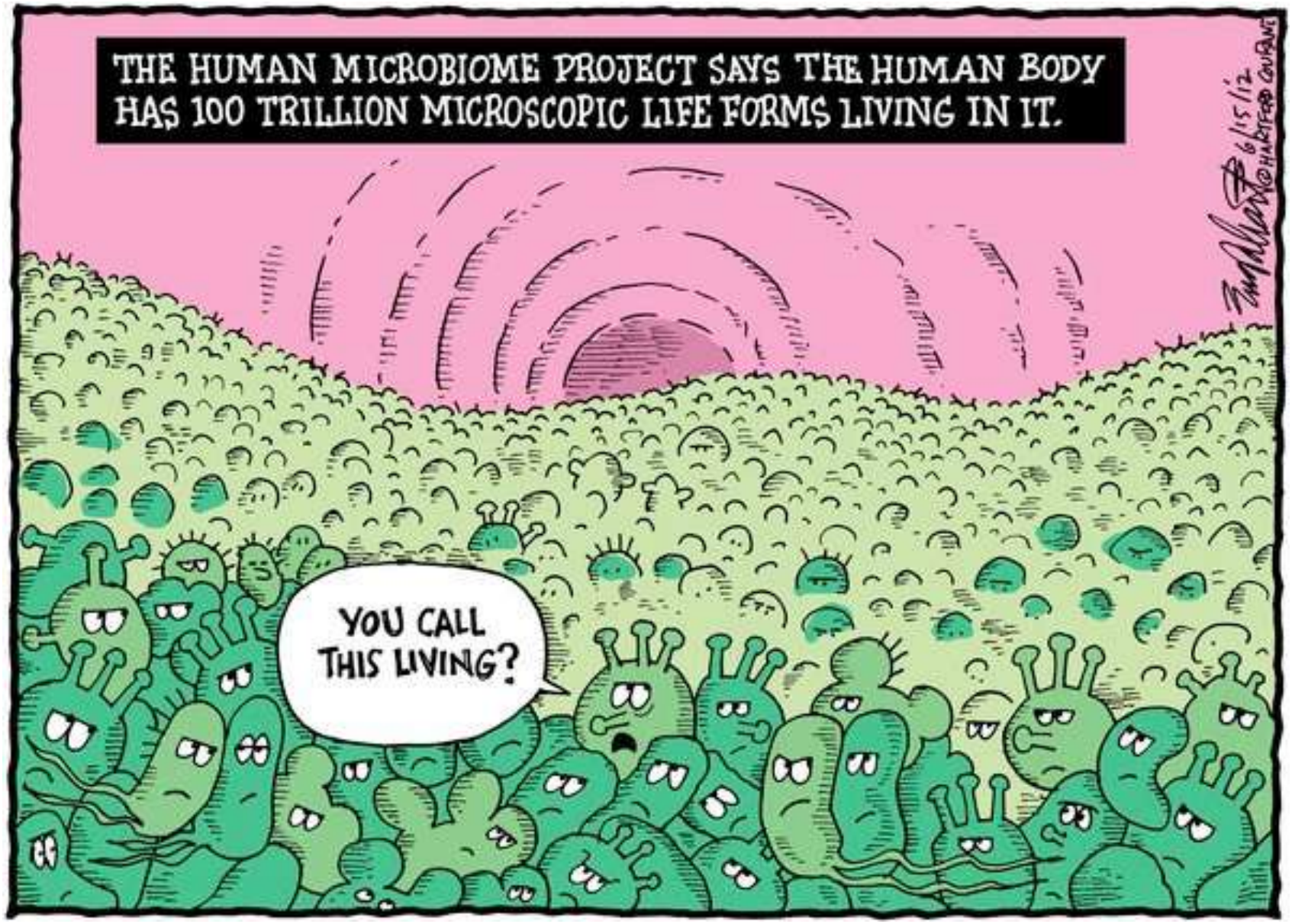
https://nextstrain.org/narratives/ncov/sit-rep/en/2020-03-20



https://www.science.org/news/2021/08/new-sars-cov-2-variants-have-changed-pandemic-what-will-virus-do-next
https://www.science.org/doi/epdf/10.1126/science.acx8919

20 August 2021

Human gut microbiome



Human gut microbiome

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

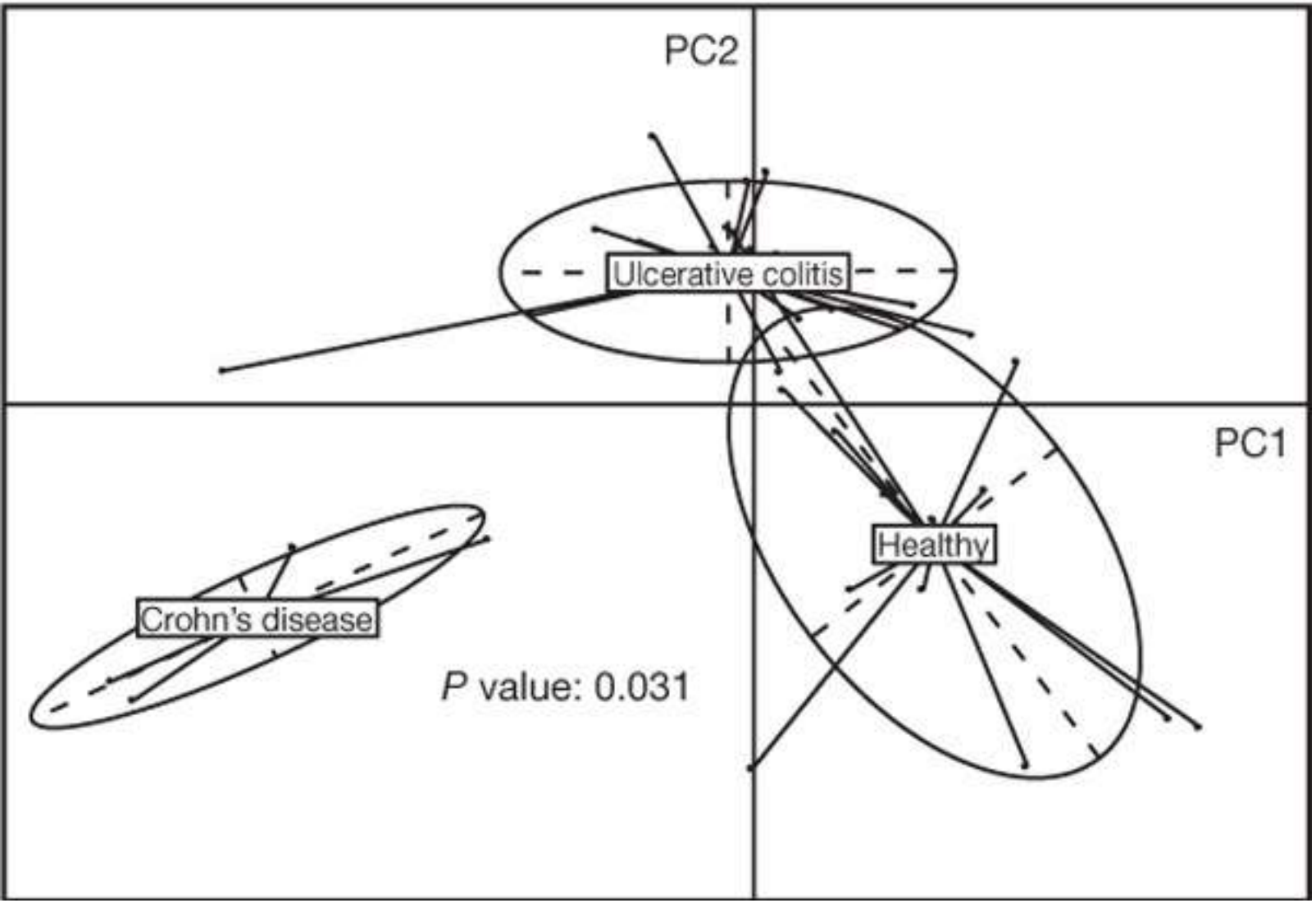
ARTICLES

A human gut microbial gene catalogue established by metagenomic sequencing

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

doi:10.1038/nature08821

Human gut microbiome

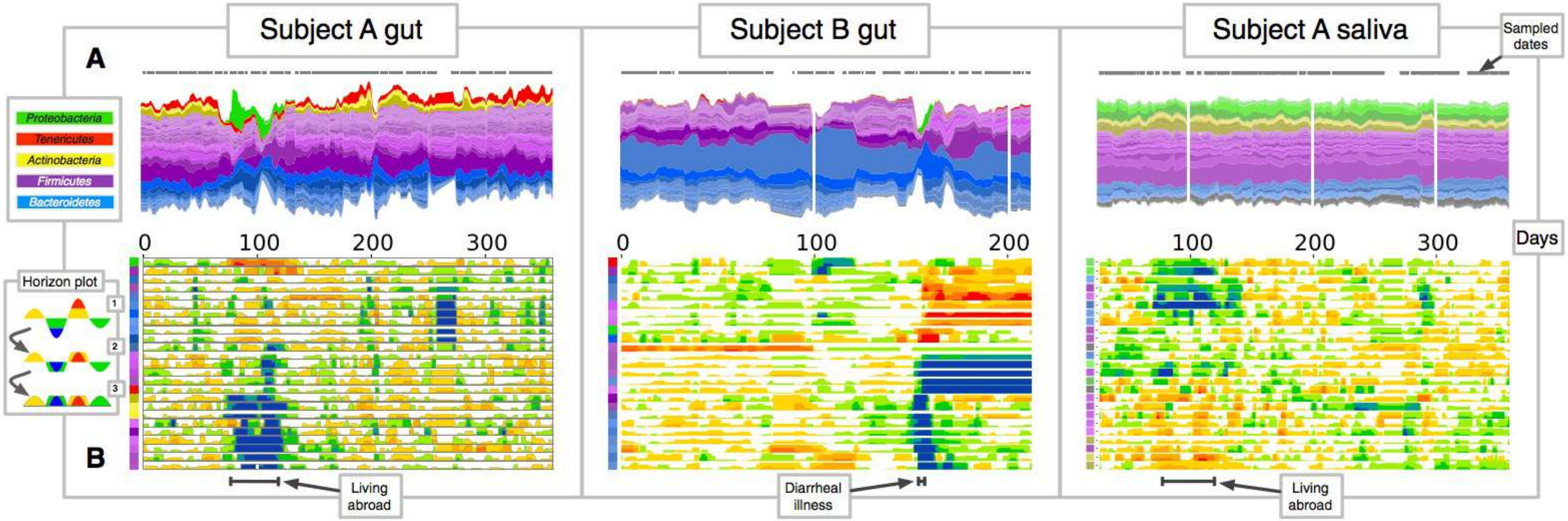


We can check which OTUs constitute the clustering (and separation) patterns

-> Biology
-> Biomarkers

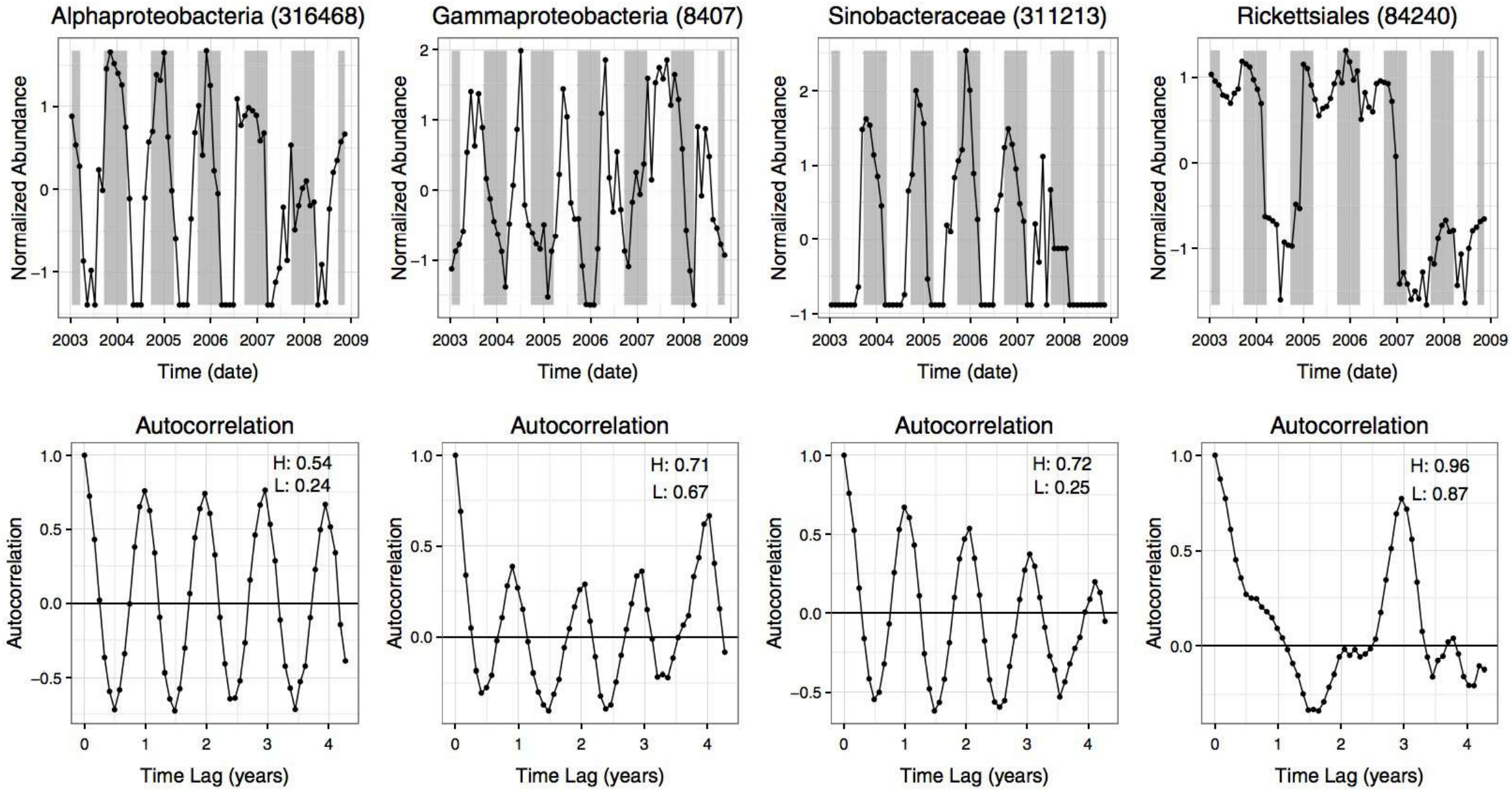
doi:10.1038/nature08821

Tracking microbiome on a daily scale



David et al. Genome Biology 2014, 15:R89

Tracking microbiome spanning 6 years

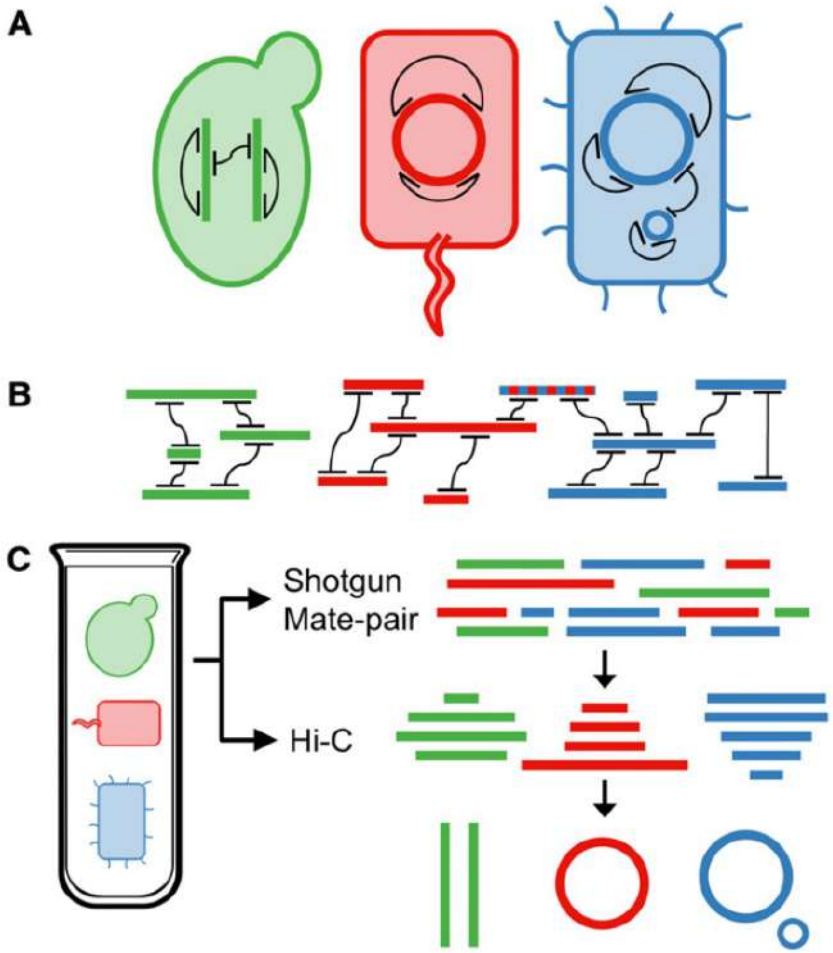


Faust et al 2015

Current Opinion in Microbiology

Species-Level Deconvolution of Metagenome Assemblies with Hi-C–Based Contact Probability Maps

Joshua N. Burton,[1] Ivan Liachko,[1] Maitreya J. Dunham,[2] and Jay Shendure[2]
Department of Genome Sciences, University of Washington, Seattle, Washington 98195-5065



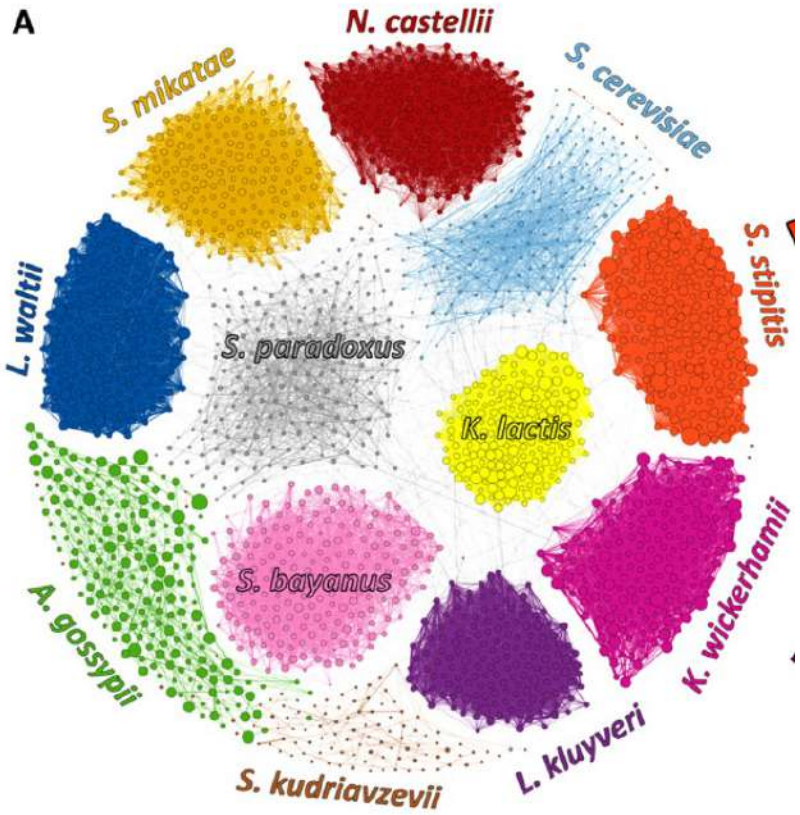
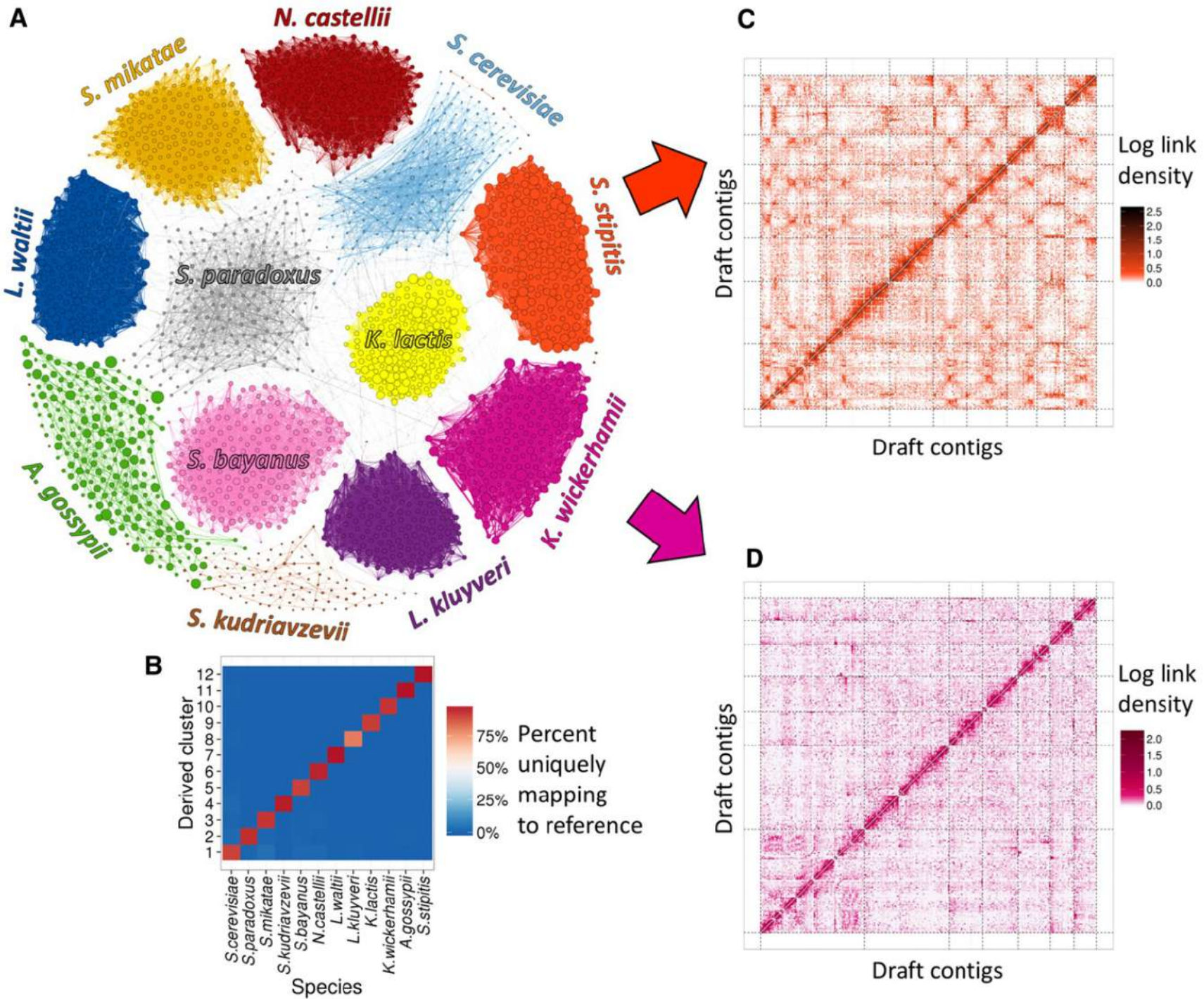


Table 2 Sequencing libraries used in MetaPhase analyses

Sample	Library Type	Read Length, bp	Read Pairs, millions
M-Y	Shotgun	101	85.7
	Mate-pair	100	9.2
	Hi-C	100	81.0


A
N. castellii
S. mikatae
S. cerevisiae
S. stipitis
L. waltii
S. paradoxus
K. lactis
K. wickerhamii
S. bayanus
A. gossypii
L. kluyveri
S. kudriavzevii
B
Derived cluster
12
11
10
9
8
7
6
5
4
3
2
1
S.cerevisiae
S.paradoxus
S.mikatae
S.kudriavzevii
S.bayanus
N.castellii
L.waltii
L.kluyveri
K.lactis
K.wickerhamii
A.gossypii
S.stipitis
Species
75%
50%
25%
0%
Percent uniquely mapping to reference
C
Draft contigs
Draft contigs
Log link density
2.5
2.0
1.5
1.0
0.5
0.0
D
Draft contigs
Draft contigs
Log link density
2.0
1.5
1.0
0.5
0.0

Other fields transformed by genomics

ancient DNA

ORIGINAL ARTICLES Green, R. E. et al. A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010) | Meyer, M. et al. A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226 (2012) | Rasmussen, M. et al. Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463**, 757–762 (2010)

FURTHER READING Slon, V. et al. The genome of the offspring of a Neanderthal mother and a Denisovan father. *Nature* **561**, 113–116 (2018) | Haak, W. et al. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207–211 (2015) | Sankararaman, S. et al. The date of interbreeding between Neandertals and modern humans. *PLoS Genet.* **8**, e1002947 (2012)

“Ancient DNA research has been limited only by the technology, and never by a lack of interesting questions to be asked”

nature milestones

Genomic sequencing

1. **The Human Genome Project**
2. Sequencing the unculturable majority
3. **Sequencing — the next generation**
4. ChIP–seq captures the chromatin landscape
5. The dawn of personal genomes
6. A sequencing revolution in cancer
7. Transcriptomes — a new layer of complexity
8. **Long reads become a reality**
9. Exploring whole exomes
10. **Probing nuclear architecture with Hi-C**
11. **Sequencing one cell at a time**
12. Waking the dead: sequencing archaic hominin genomes
13. Cataloguing a public genome
14. Our most elemental encyclopaedia
15. Pan-genomes: moving beyond the reference
16. Genomes go platinum
17. **Filling in the gaps telomere to telomere**

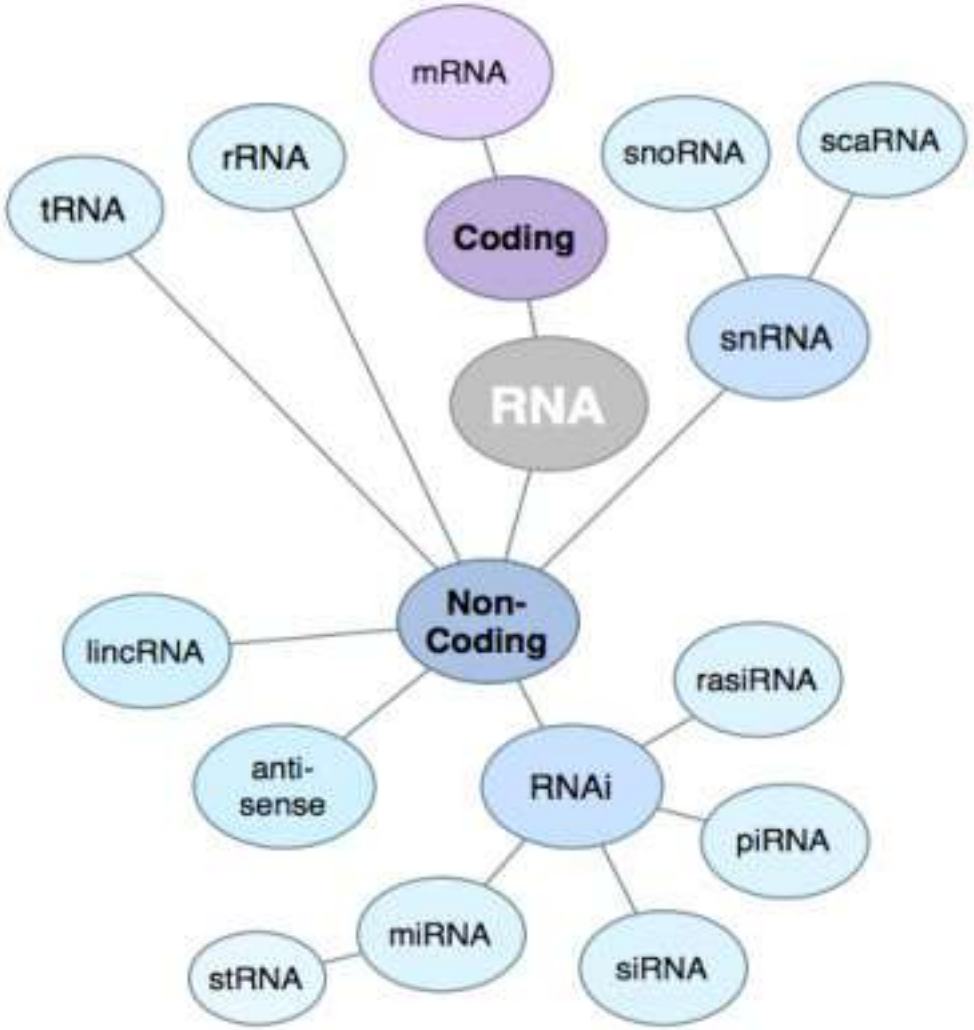
10 February 2021

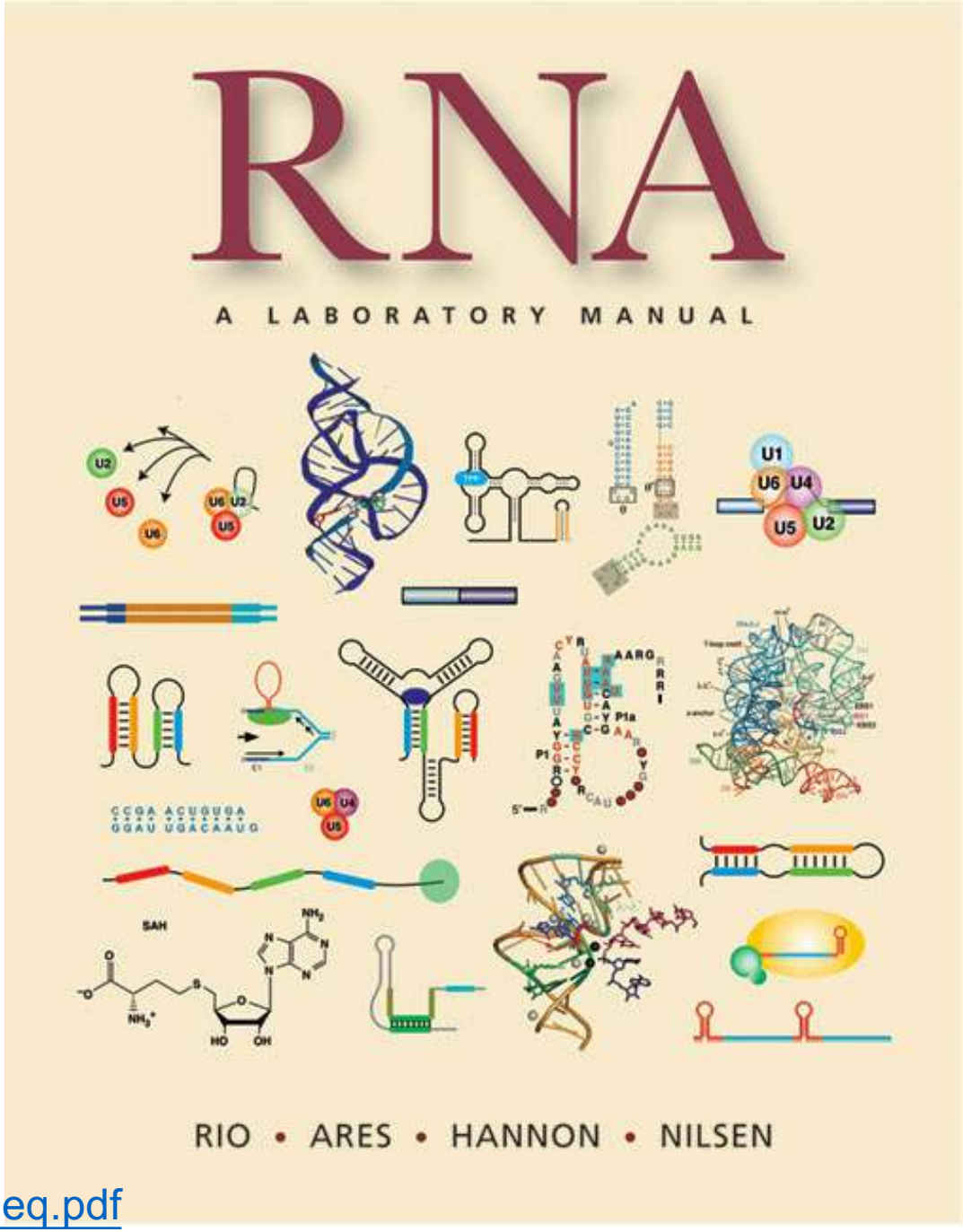
https://www.nature.com/immersive/d42859-020-00099-0/index.html

Break here

Transcriptomics / RNAseq

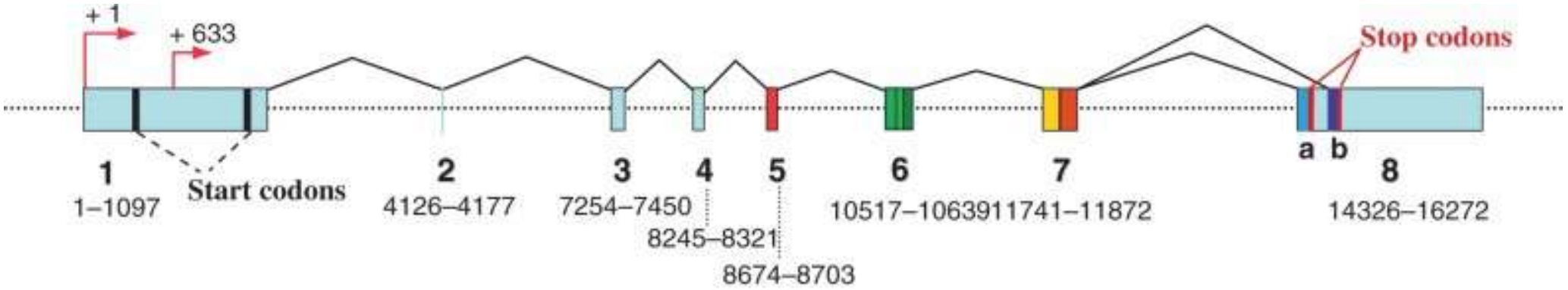
Types of RNA

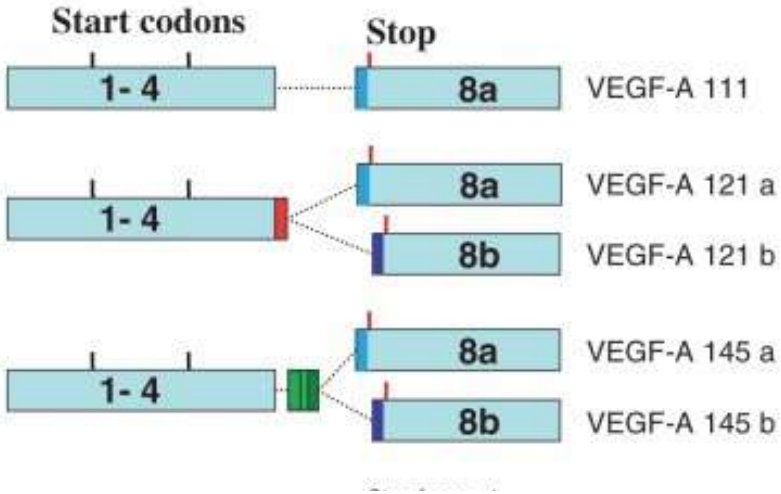


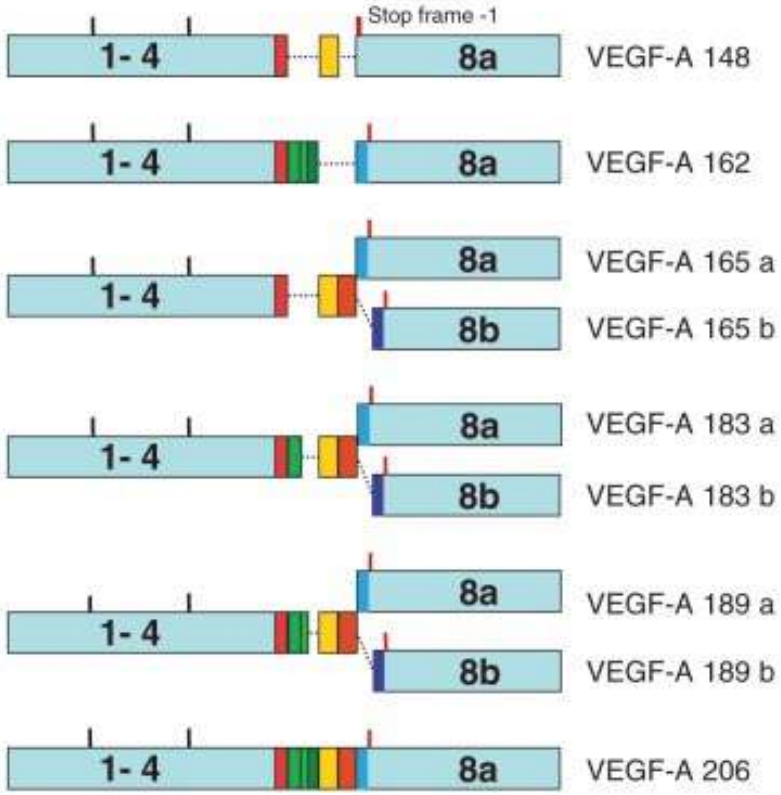


http://jura.wi.mit.edu/bio/education/hot_topics/RNAseq/RNA_Seq.pdf

Gene and isoforms







Arcondeguy et al (2015)

Conesa *et al. Genome Biology* (2016) 17:13
DOI 10.1186/s13059-016-0881-8

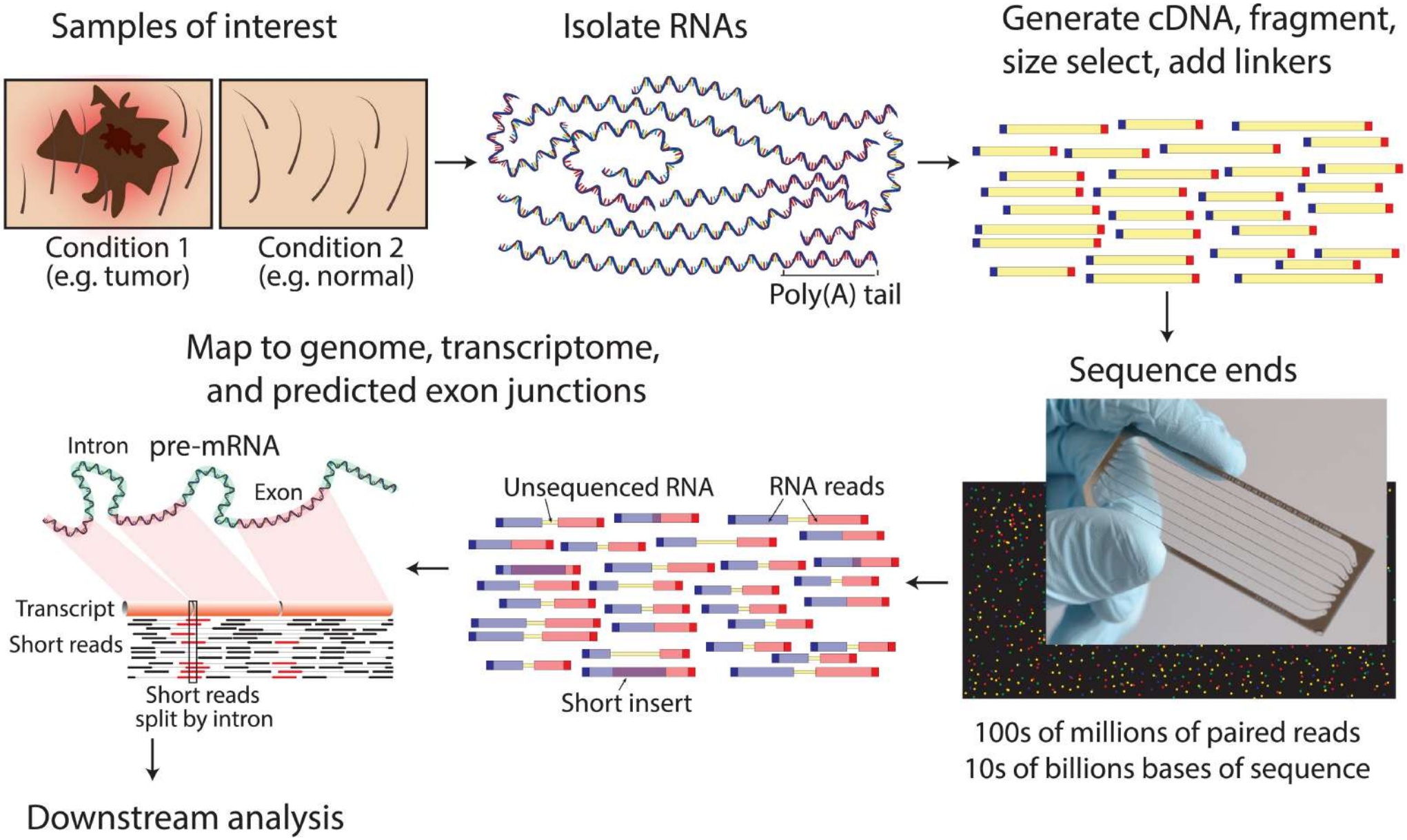
Genome Biology

REVIEW Open Access

A survey of best practices for RNA-seq data analysis

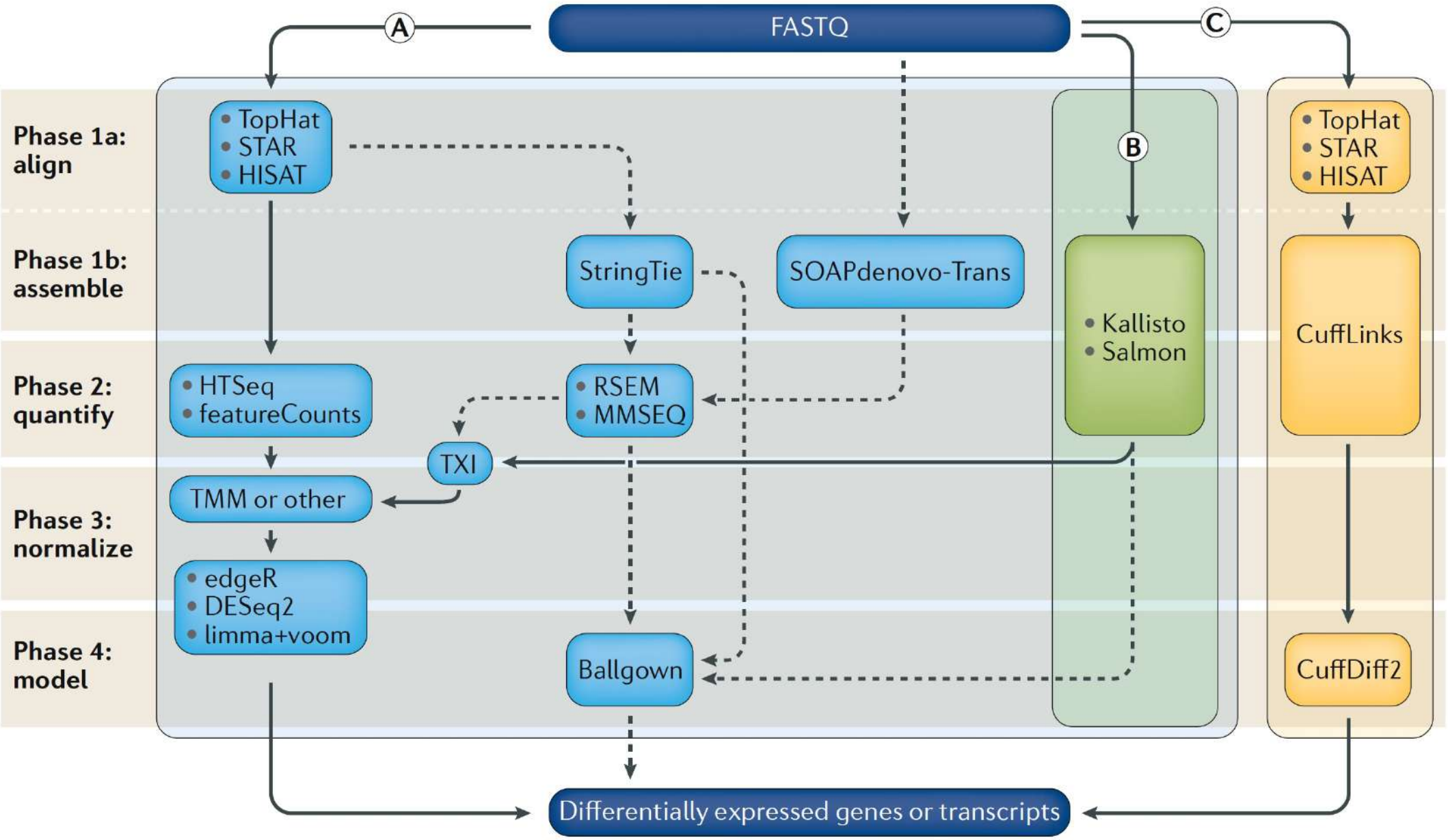
Ana Conesa[1,2*], Pedro Madrigal[3,4*], Sonia Tarazona[2,5], David Gomez-Cabrero[6,7,8,9], Alejandra Cervera[10], Andrew McPherson[11], Michał Wojciech Szcześniak[12], Daniel J. Gaffney[3], Laura L. Elo[13], Xuegong Zhang[14,15] and Ali Mortazavi[16,17*]

RNA-seq data generation



Griffith et al (2015) Plos Computational Biology

RNAseq analysis workflow for differential expression (generalized)



Stark, Grzelak and Hadfield (2019) Nature Reviews Genetics

REVIEWS

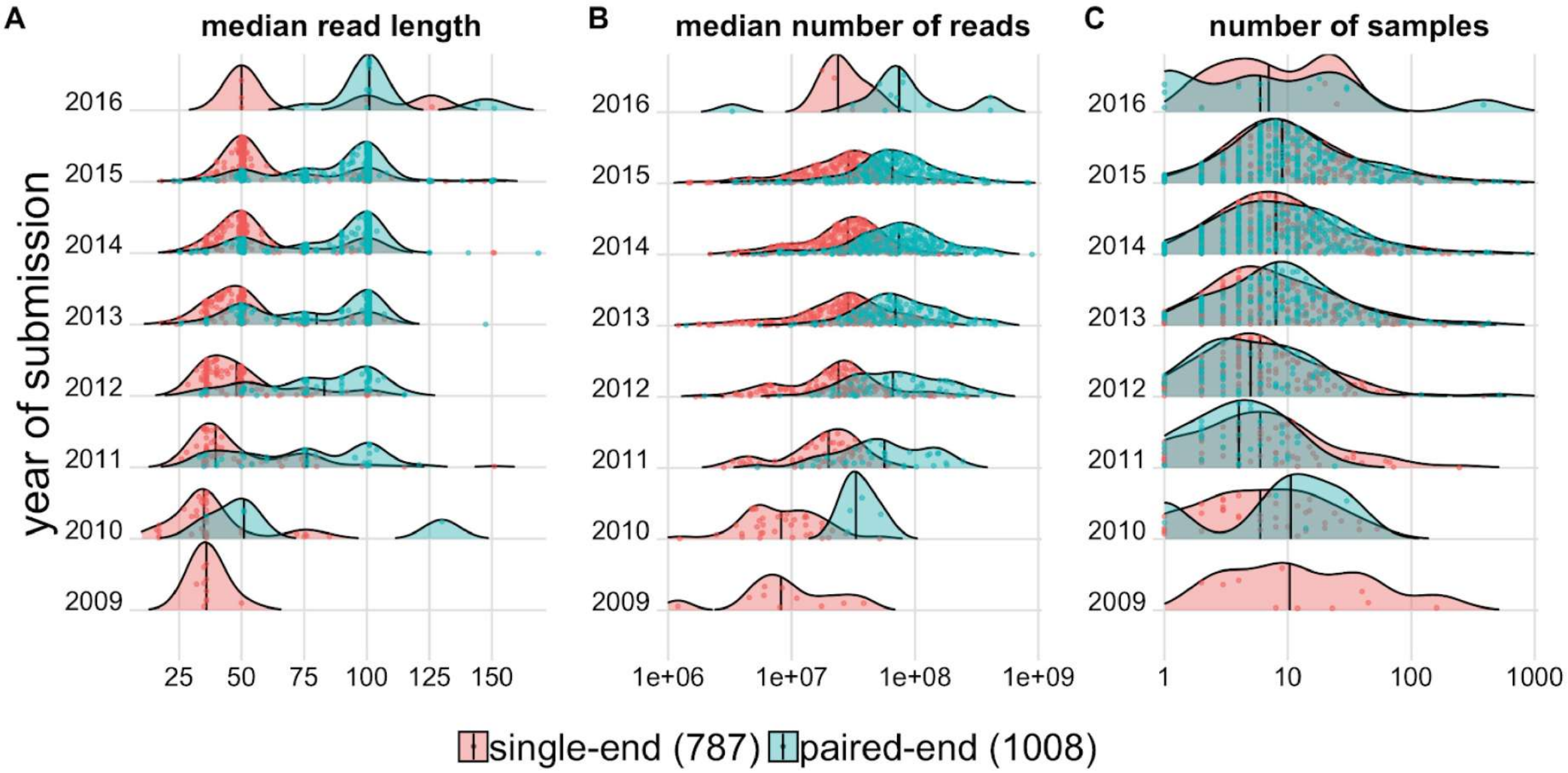
RNA sequencing: the teenage years

Rory Stark[1], *Marta Grzelak*[1] *and James Hadfield*[2]*

Abstract | Over the past decade, RNA sequencing (RNA-seq) has become an indispensable tool for transcriptome-wide analysis of differential gene expression and differential splicing of mRNAs. However, as next-generation sequencing technologies have developed, so too has RNA-seq. Now, RNA-seq methods are available for studying many different aspects of RNA biology, including single-cell gene expression, translation (the translatome) and RNA structure (the structurome). Exciting new applications are being explored, such as spatial transcriptomics (spatialomics). Together with new long-read and direct RNA-seq technologies and better computational tools for data analysis, innovations in RNA-seq are contributing to a fuller understanding of RNA biology, from questions such as when and where transcription occurs to the folding and intermolecular interactions that govern RNA function.

Stark, Grzelak and Hadfield (2019) Nature Reviews Genetics

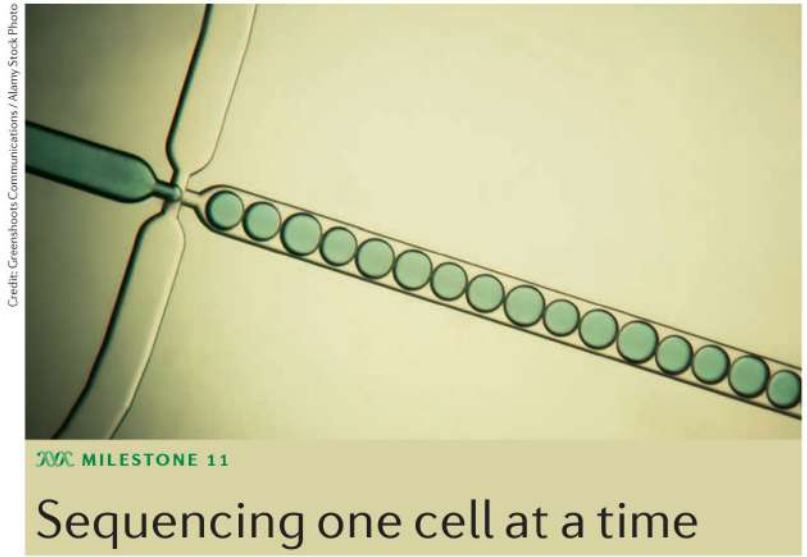
Evolution of RNAseq over time (from SRA)



Berge *et al.,* Annual Review of Biomedical Data Science (2019)

Further advances
Single cell RNAseq (ScRNAseq)

single cell sequencing

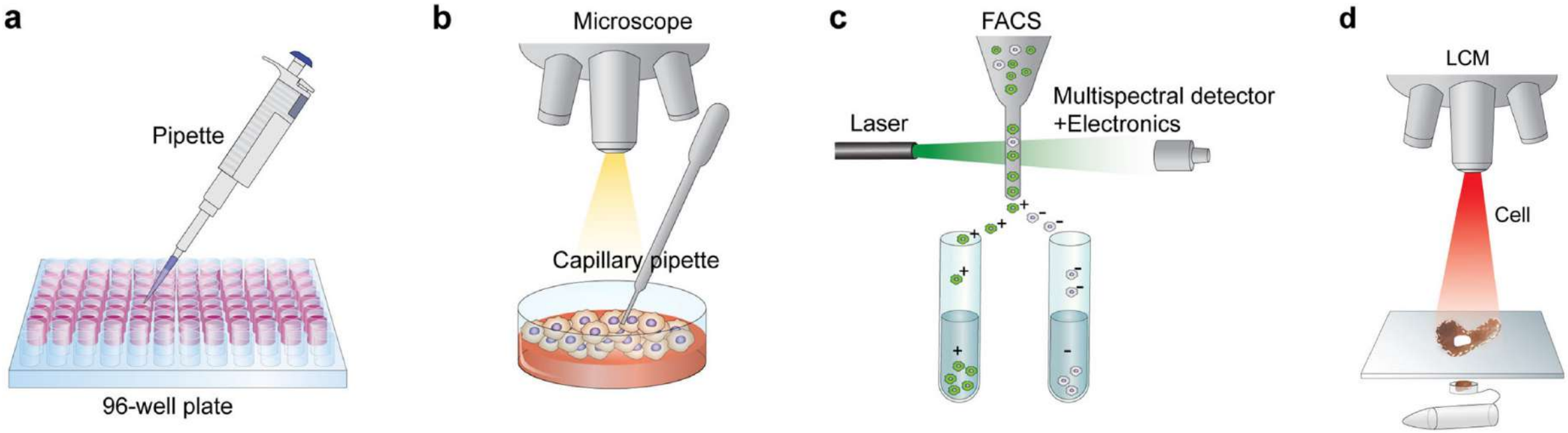


ORIGINAL ARTICLE Tang, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009)

FURTHER READING Islam, S. et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* **21**, 1160–1167 (2011) | Ramsköld, D. et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012) | Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015) | Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015) | Haber, A. L. et al. A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333–339 (2017) | Vento-Tormo, R. et al. Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature* **563**, 347–353 (2018)

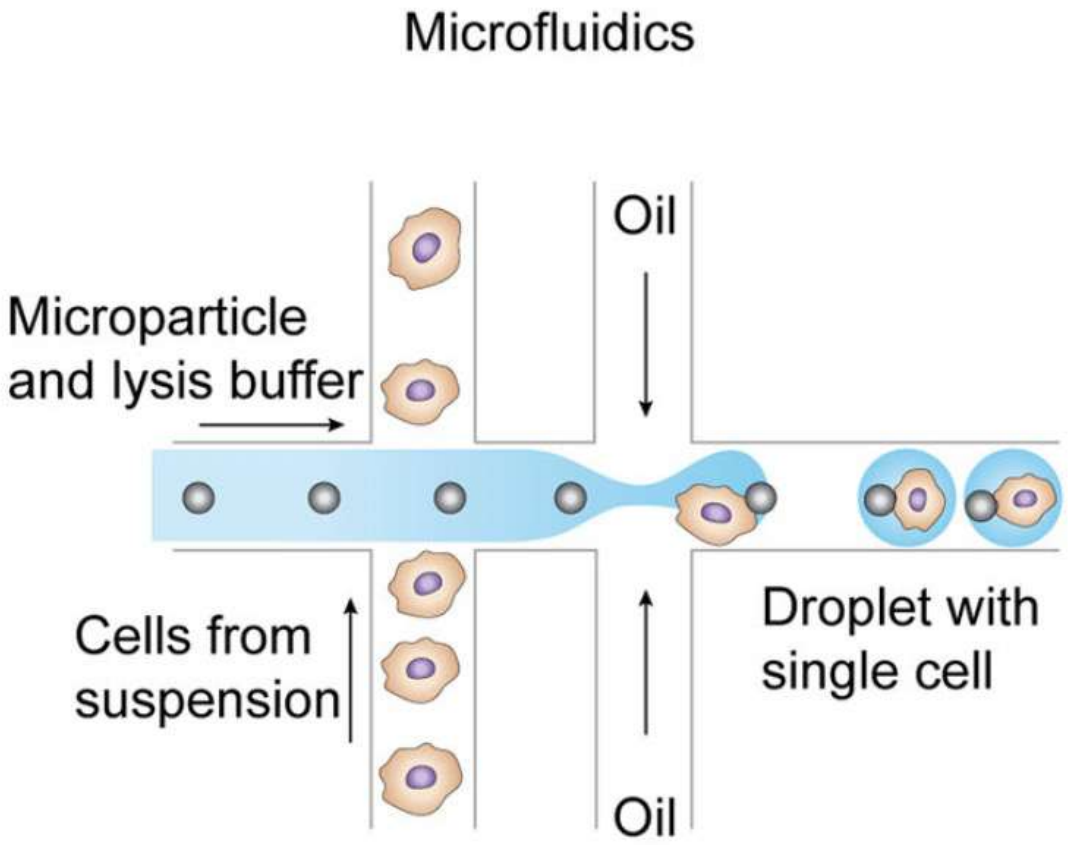
> “single-cell techniques [have] enabled analysis of cell states, … diseased states and tumour heterogeneity”

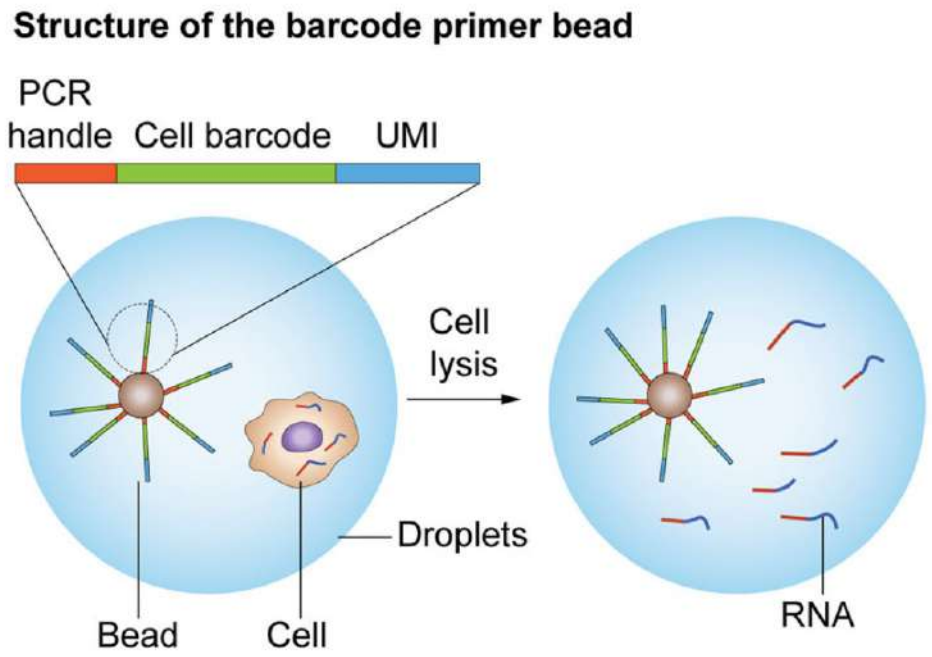
Evolution of single-cell isolation

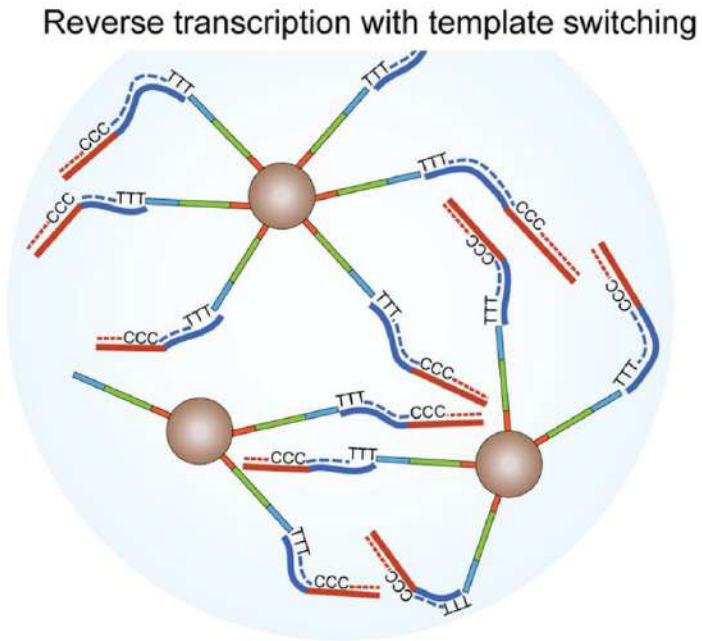


Huang *et al.,* Experimental & Molecular Medicine (2018) 50:96

Microfluidic isolation in reagent- filled droplets



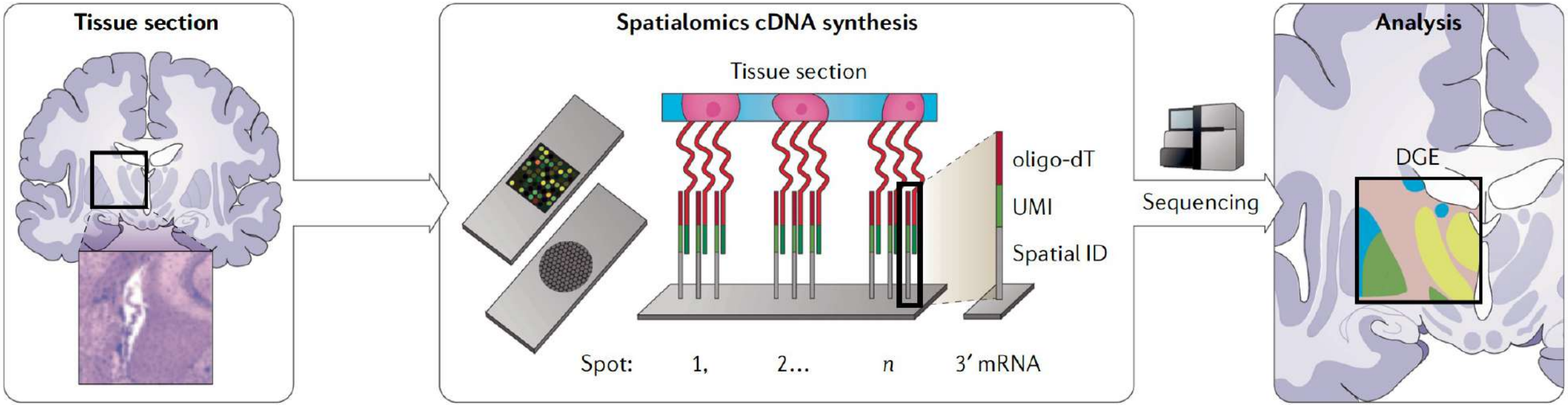




Huang *et al.,* Experimental & Molecular Medicine (2018) 50:96

Spatialomics

1. Spatial encoding requires a frozen tissue section to be applied to oligo- arrayed microarray slides or to 'pucks' of densely packed oligo- coated beads.

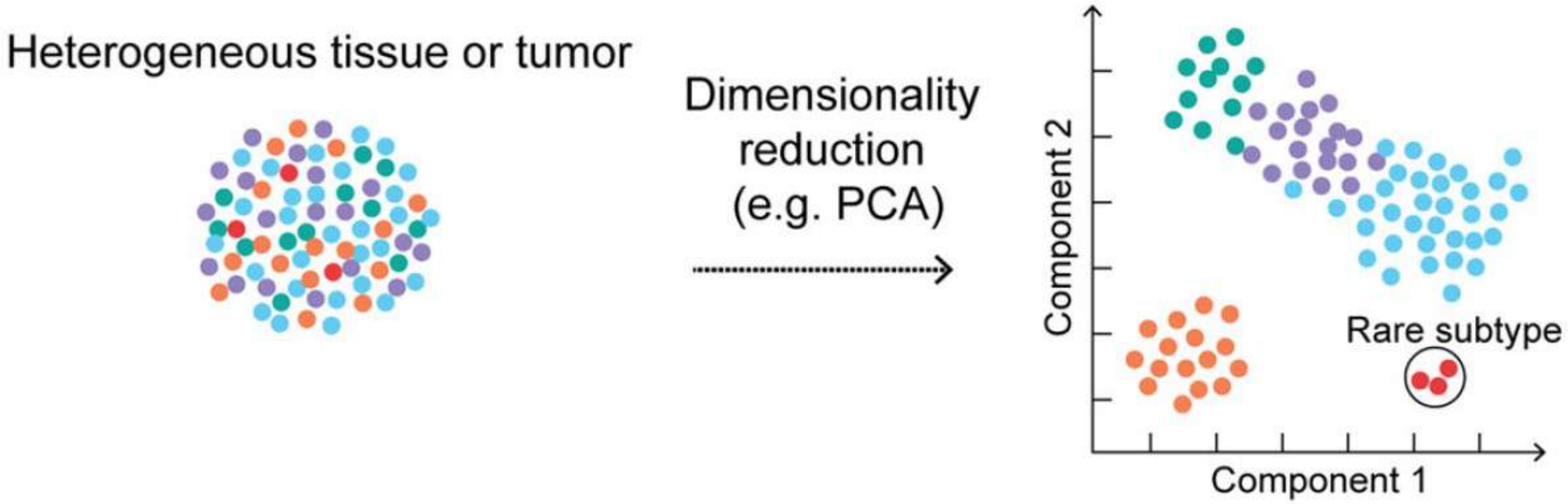


2. The mRNA diffuses to the slide surface and hybridizes to oligo- dT cDNA synthesis primers that encode UMIs and spatial barcodes. It is then reverse transcribed to produce cDNA, which is pooled for library preparation and sequencing.

3. Computational analysis of the spatialomics data maps sequence reads back to their spatial coordinates after DGE analysis and allows differential spatial expression to be visualized.

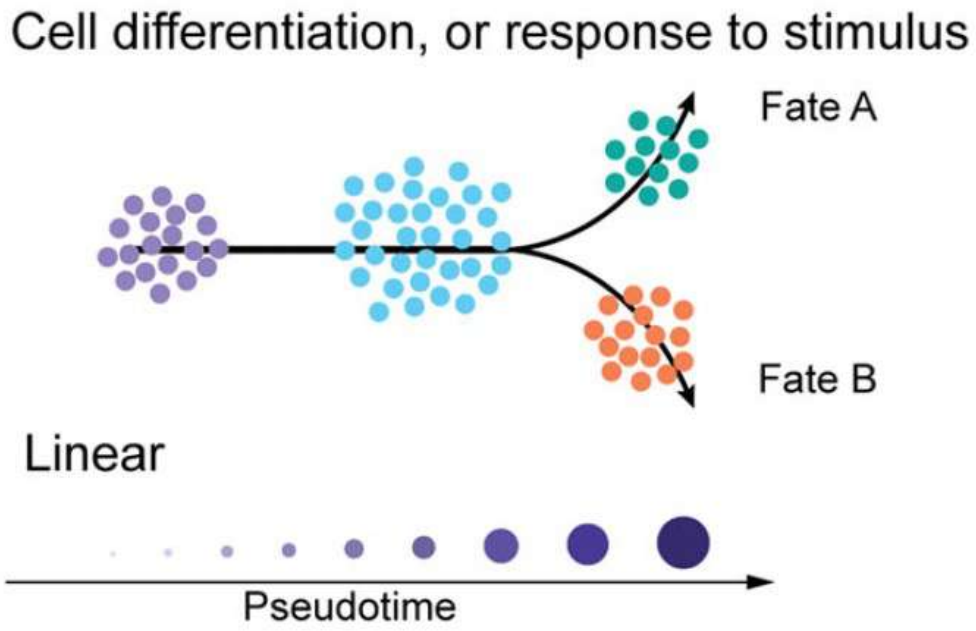
Stark *et al.,* Nature Review Genetics (2019)

Applications of scRNAseq computational approaches
1. Cell type identification



Huang *et al.,* Experimental & Molecular Medicine (2018) 50:96

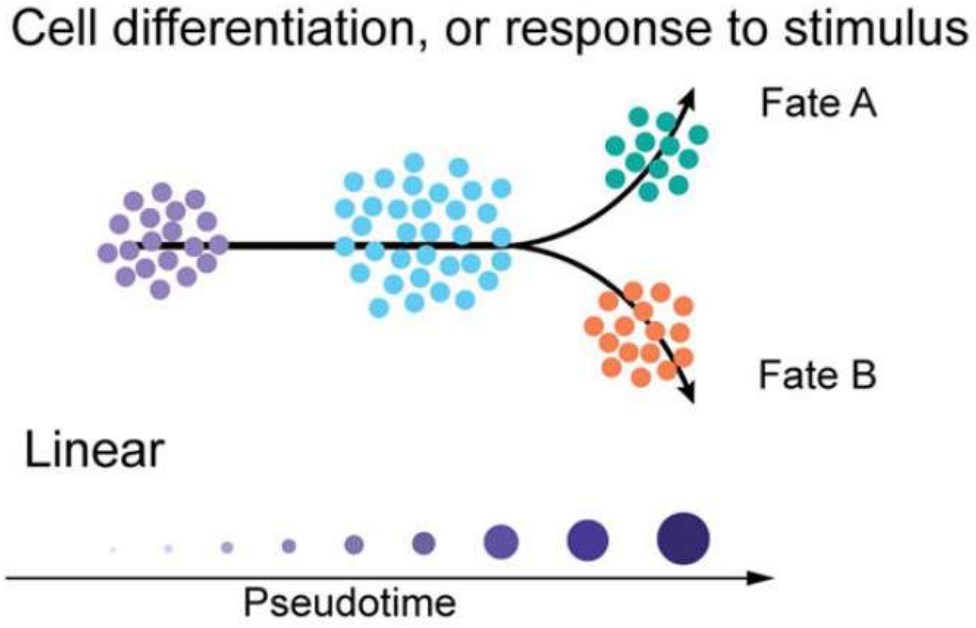
Applications of scRNAseq computational approaches
2. Cell hierarchy reconstruction

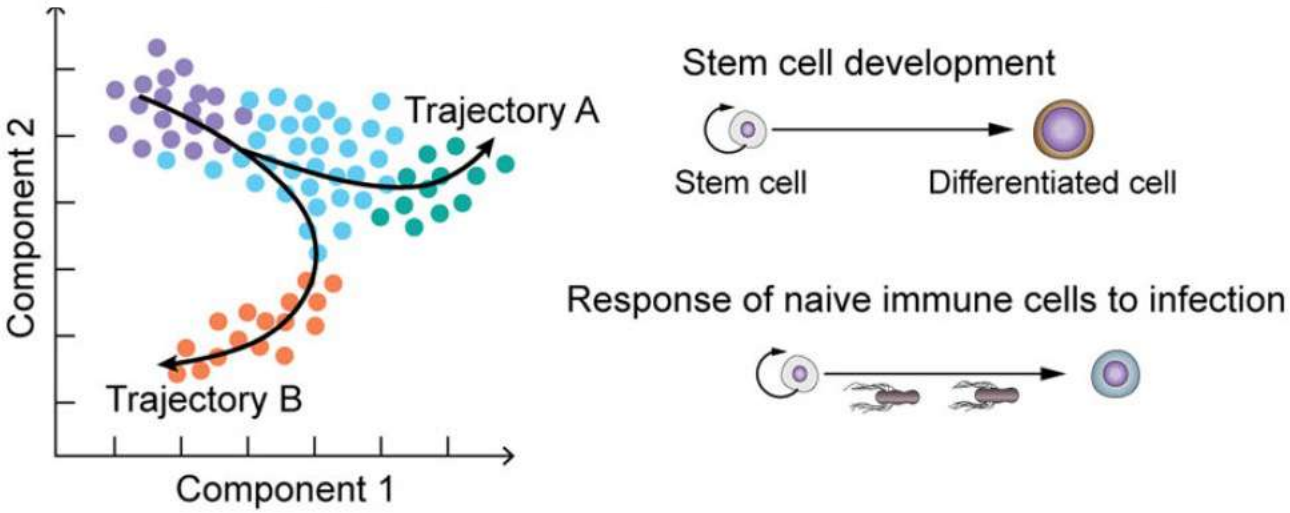


Trajectory analysis pipeline (Monocle)

Huang *et al.,* Experimental & Molecular Medicine (2018) 50:96

Applications of scRNAseq computational approaches
2. Cell hierarchy reconstruction

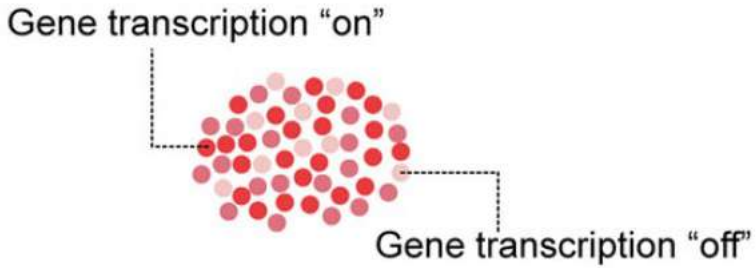




Trajectory analysis pipeline (Monocle)

Huang *et al.,* Experimental & Molecular Medicine (2018) 50:96

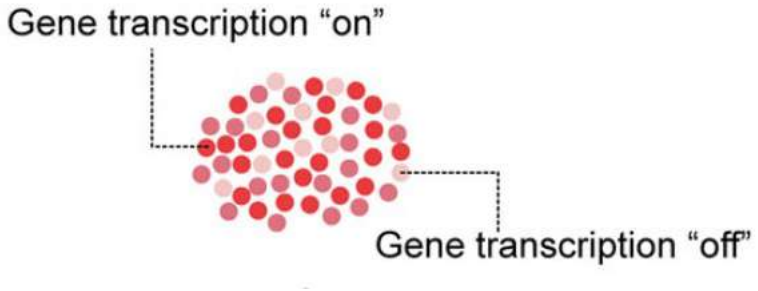
Applications of scRNAseq computational approaches
3. Inferring regulatory networks



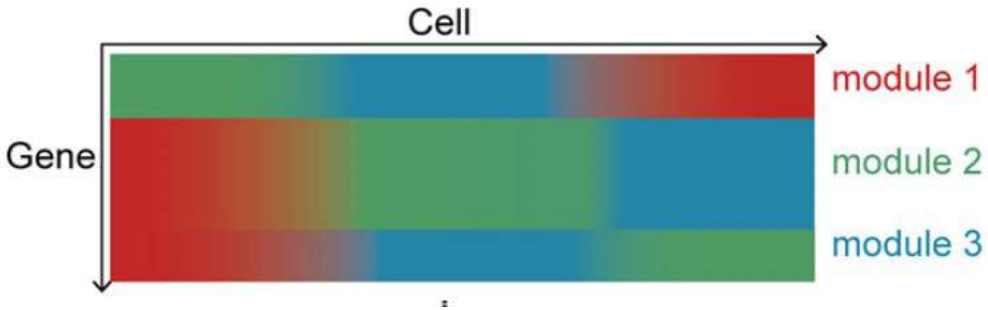
Identifying modules of co-regulated genes

Huang *et al.,* Experimental & Molecular Medicine (2018) 50:96

Applications of scRNAseq computational approaches
3. Inferring regulatory networks

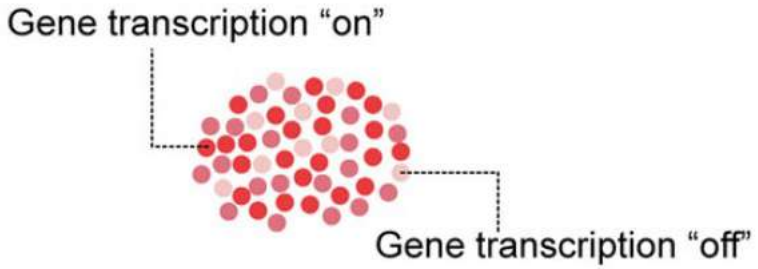


Identifying modules of co-regulated genes

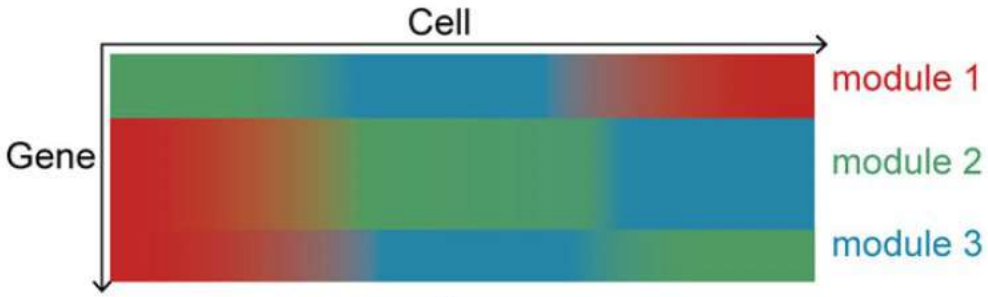


Huang *et al.,* Experimental & Molecular Medicine (2018) 50:96

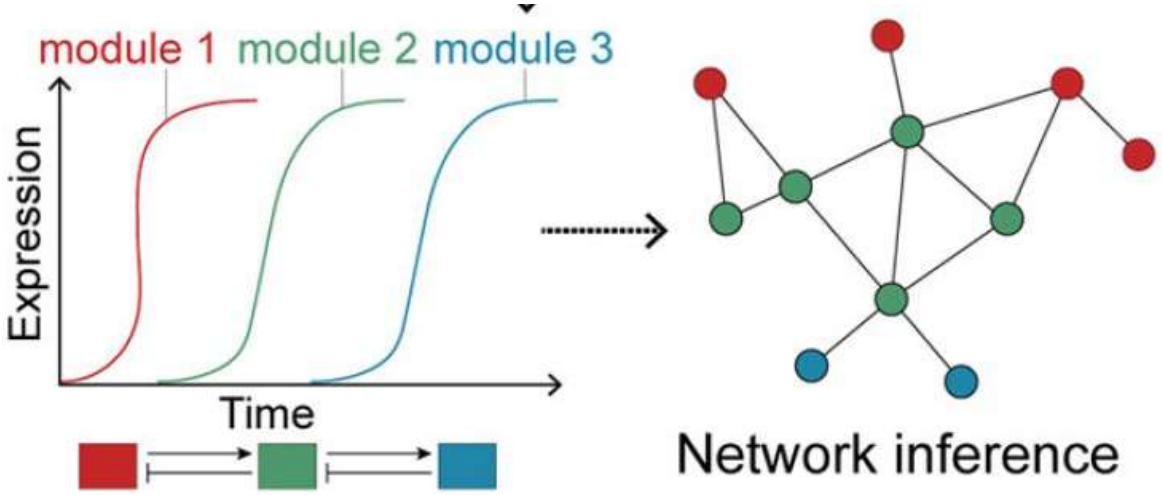
Applications of scRNAseq computational approaches
3. Inferring regulatory networks



Identifying modules of co-regulated genes



Pseudotime + Clustering

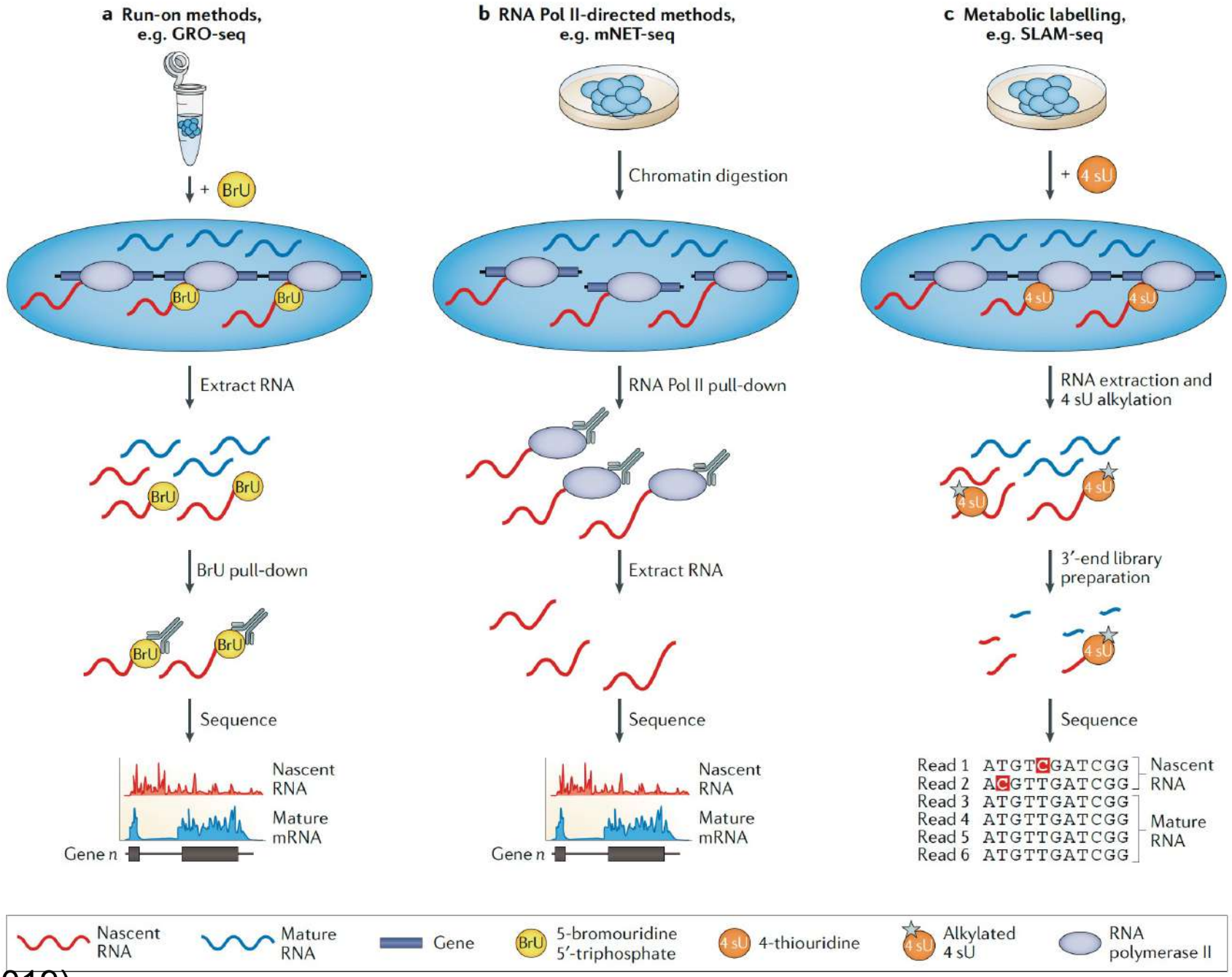


Huang *et al.,* Experimental & Molecular Medicine (2018) 50:96

Other approaches

nascent RNA

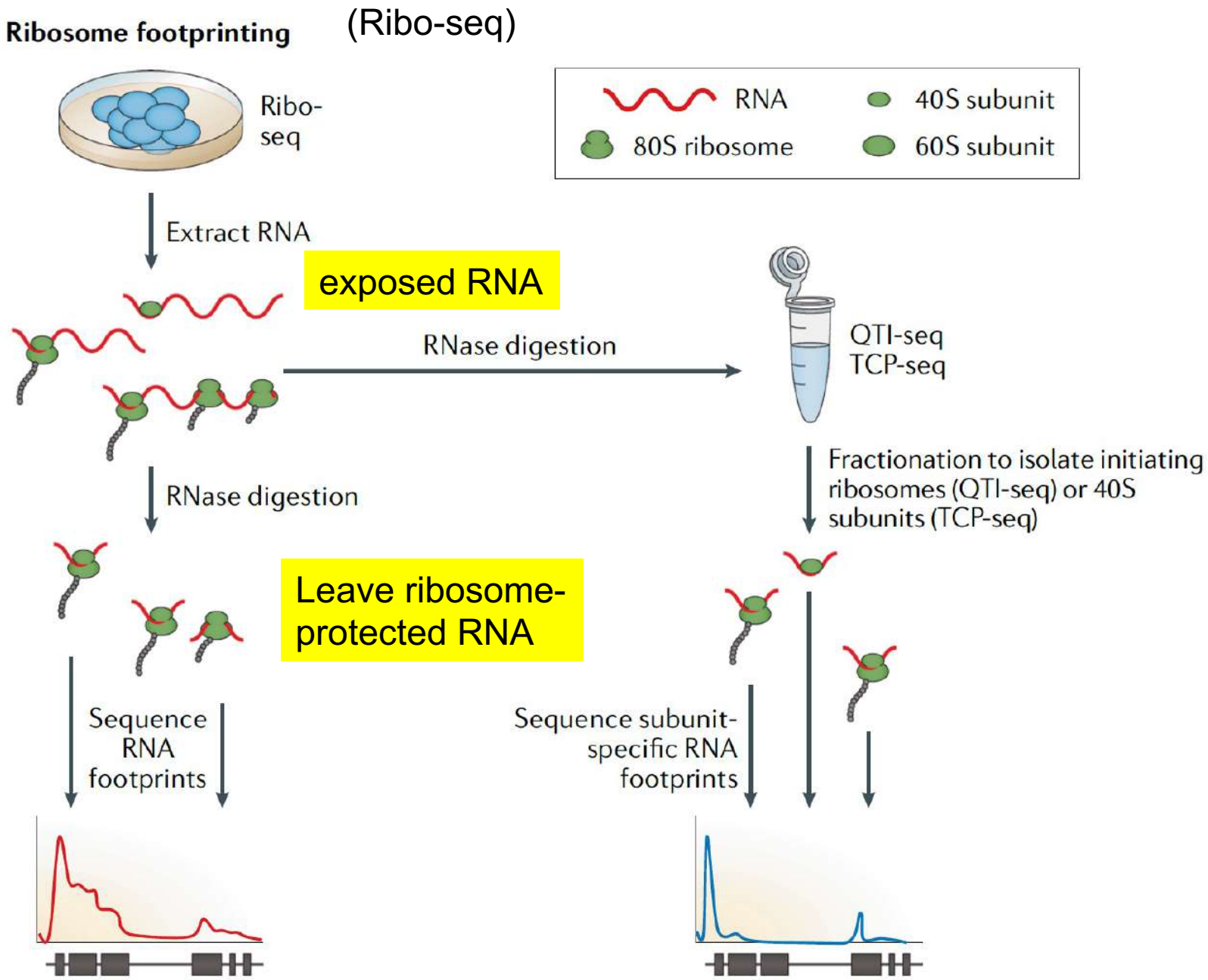
Essentially enrich newly
transcribed RNAs in a cell
and compare to control (mature RNA)



Stark *et al.,* Nature Review Genetics (2019)

translatome

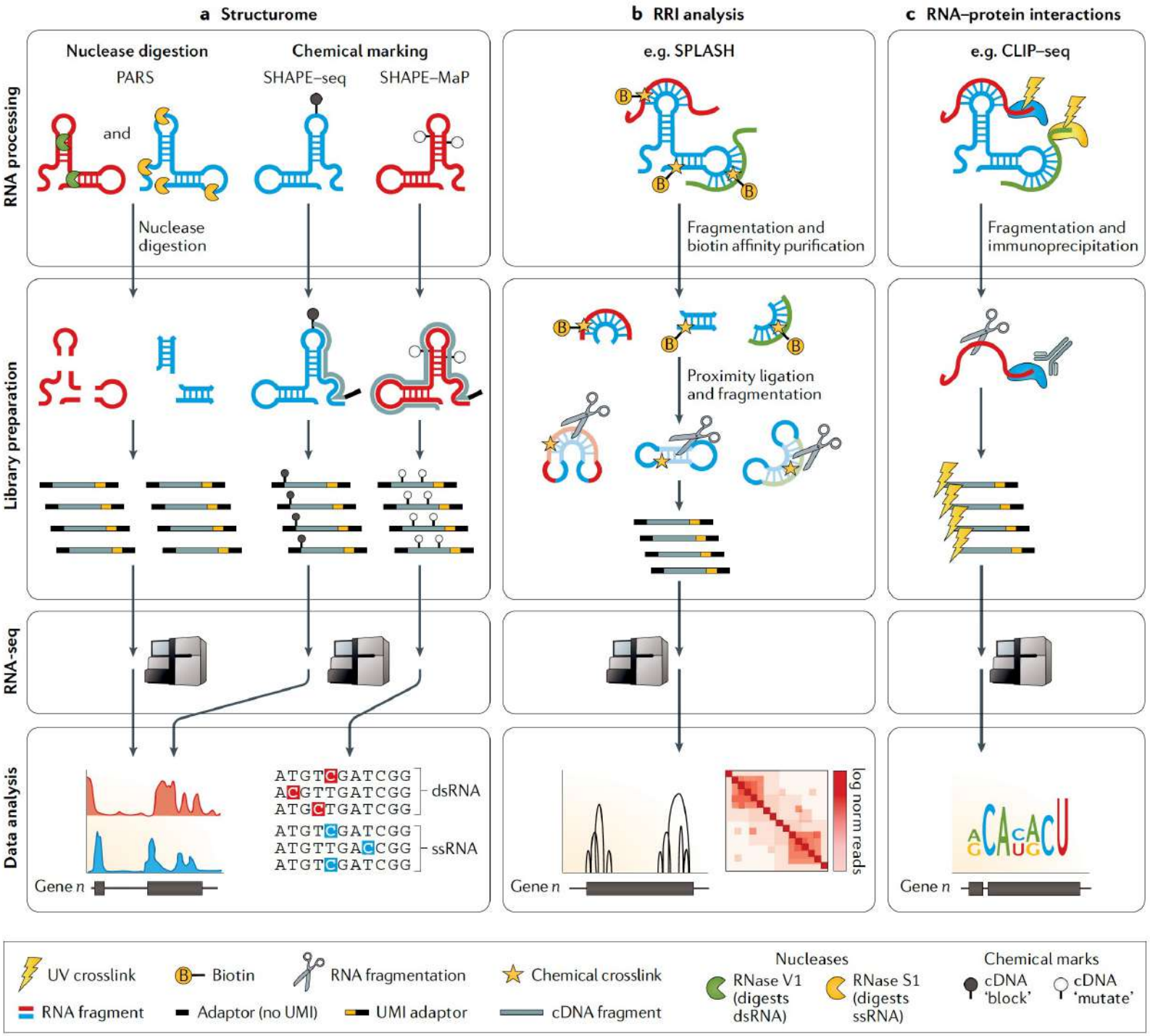
- RNA-sequencing from ribsomally bound RNA
- mRNA ribosome density correlates with the protein synthesis level



Stark *et al.,* Nature Review Genetics (2019)

RNA-RNA interaction
RNA-protein interaction

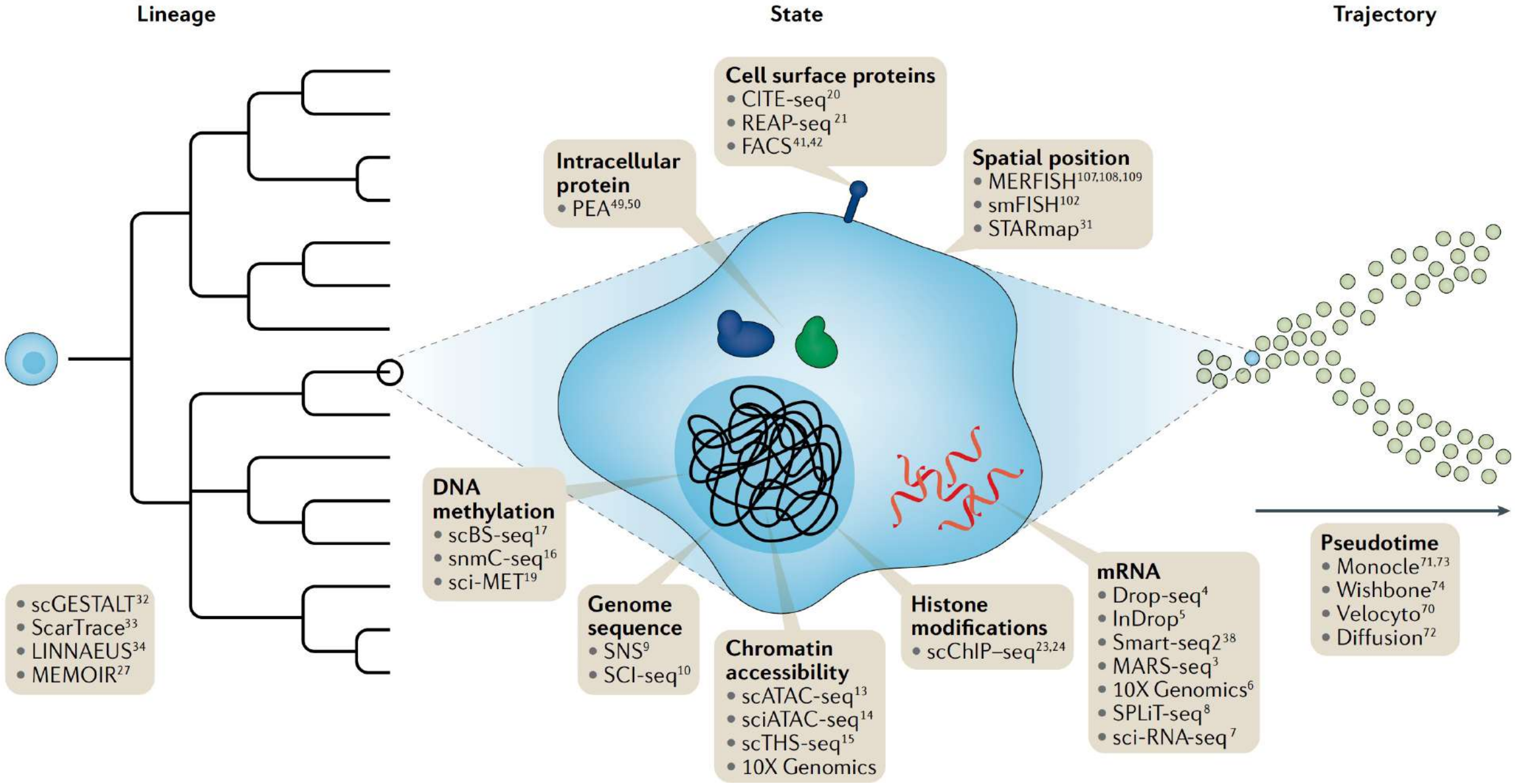
- A) Probe structured (ddRNA) or unstructured (ssRNA) RNA in transcriptome level
- B) Crosslinking interacting RNA with biotinylated psoralen
- C) Crosslinking immunoprecipitation of RNA followed by sequencing



Stark *et al.*, Nature Review Genetics (2019)

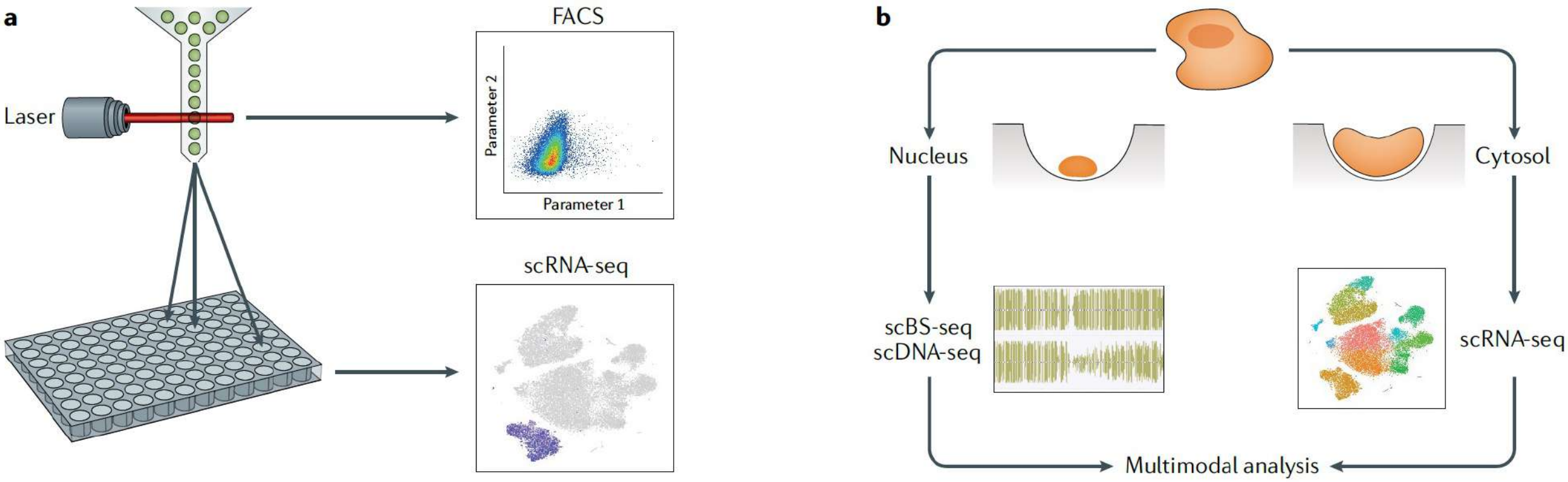
Single cell sequencing + multi omic perspective

Overview of current methods for single cell data integration



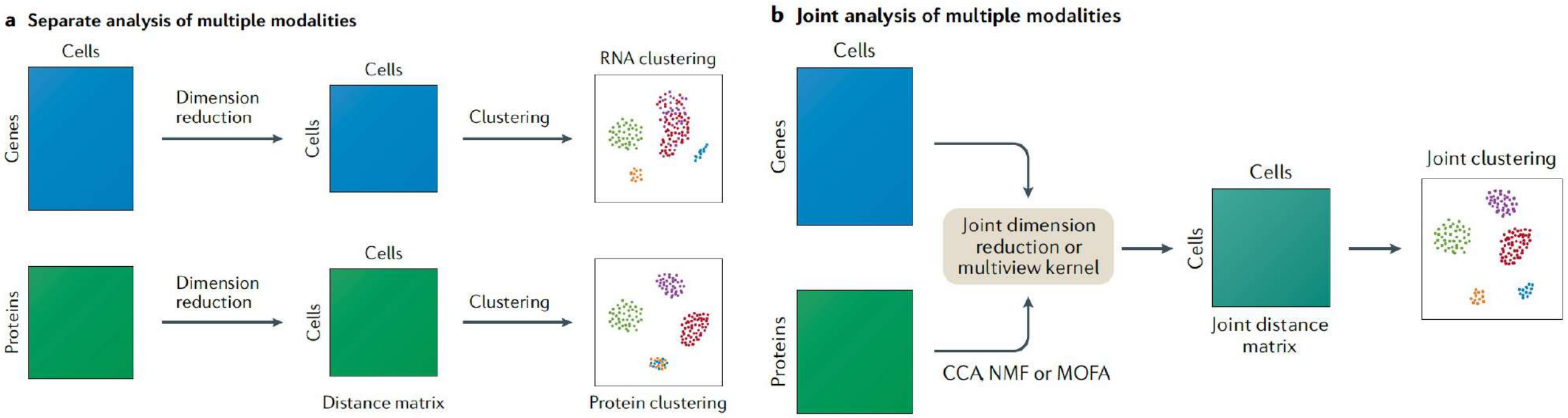
Stuart and Satija (2019) Nature Review Genetics

Example of experimental methods for performing single-cell multimodal measurements



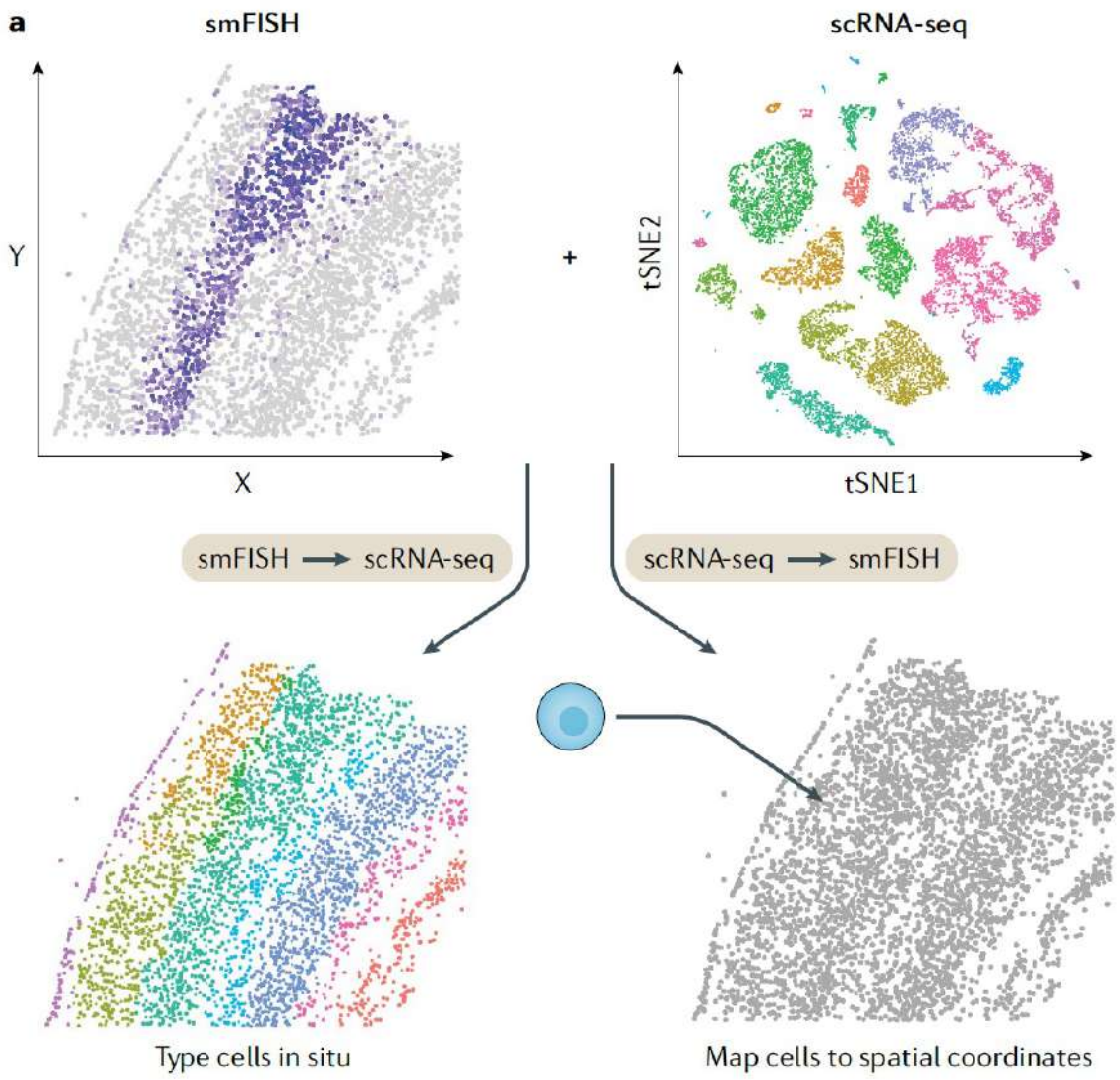
Stuart and Satija (2019) Nature Review Genetics

Multi-modal data can lead to better power at identifying cell states



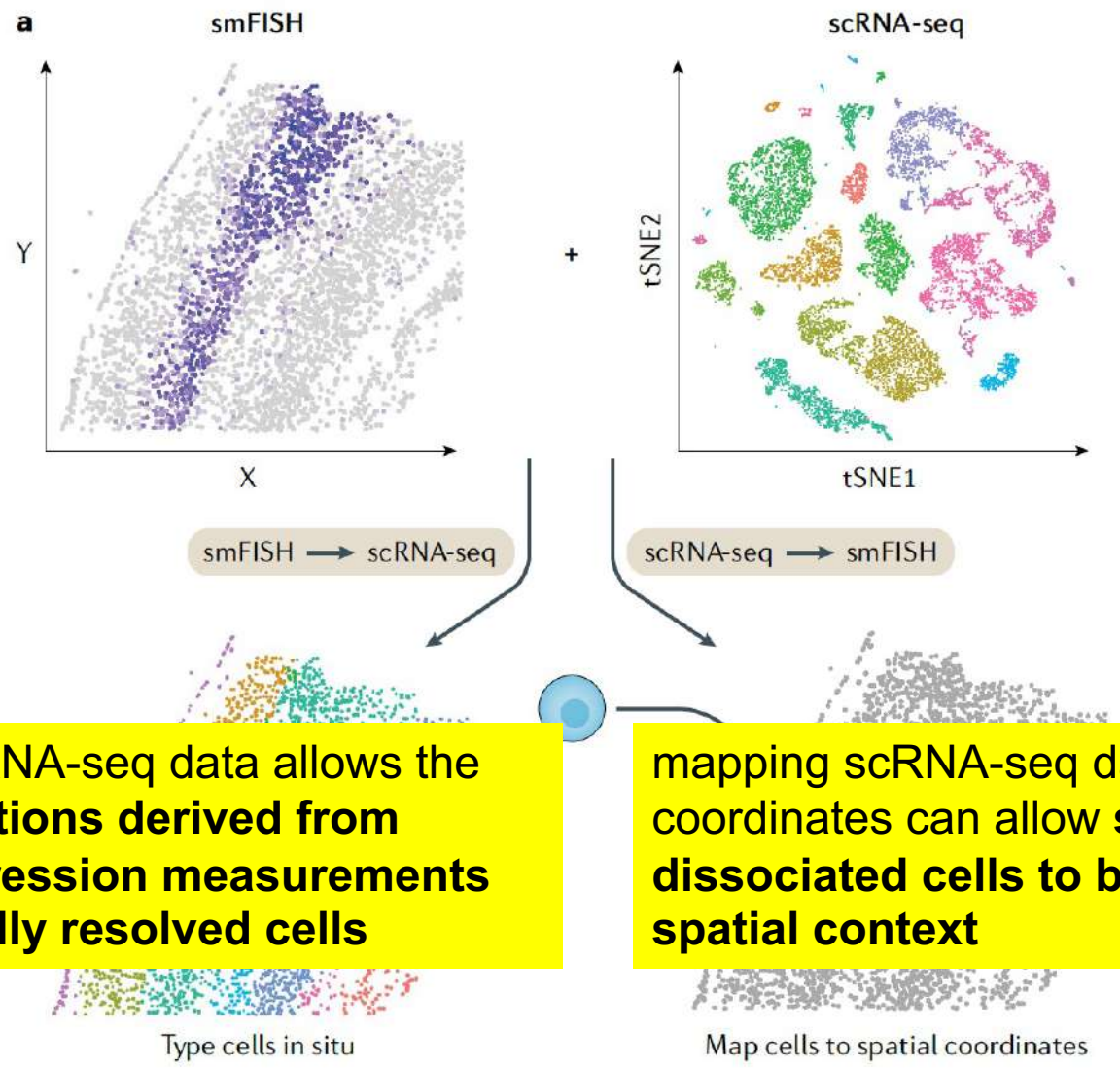
Stuart and Satija (2019) Nature Review Genetics

Spatial omics + scRNA-seq



Stuart and Satija (2019) Nature Review Genetics

Spatial omics + scRNA-seq

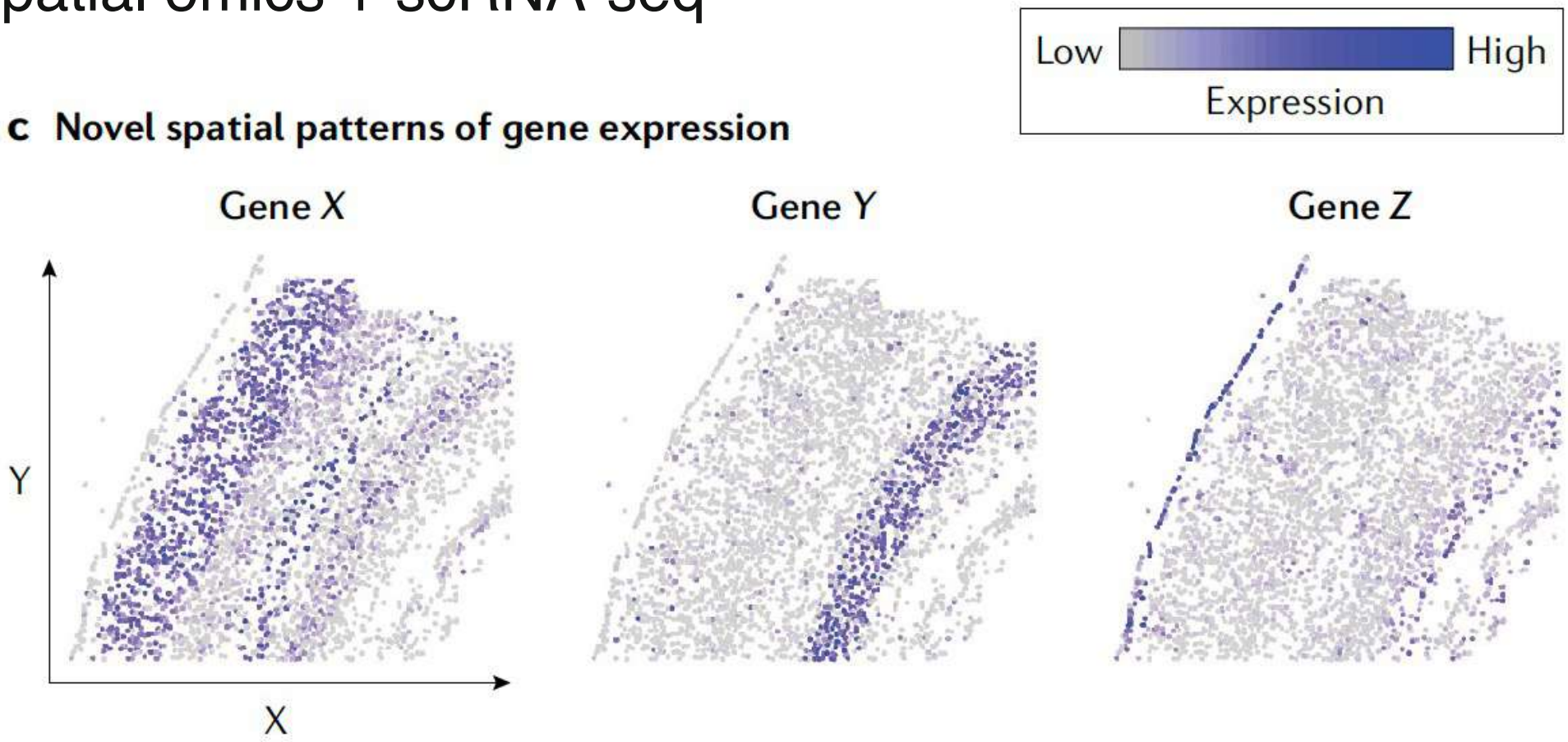


Mapping smFISH cells onto scRNA-seq data allows the **transfer of cell-type classifications derived from transcriptome-wide gene expression measurements to be transferred to the spatially resolved cells**

mapping scRNA-seq data onto smFISH-profiled spatial coordinates can allow **scRNA-seq data from dissociated cells to be placed back into their spatial context**

Stuart and Satija (2019) Nature Review Genetics

Spatial omics + scRNA-seq

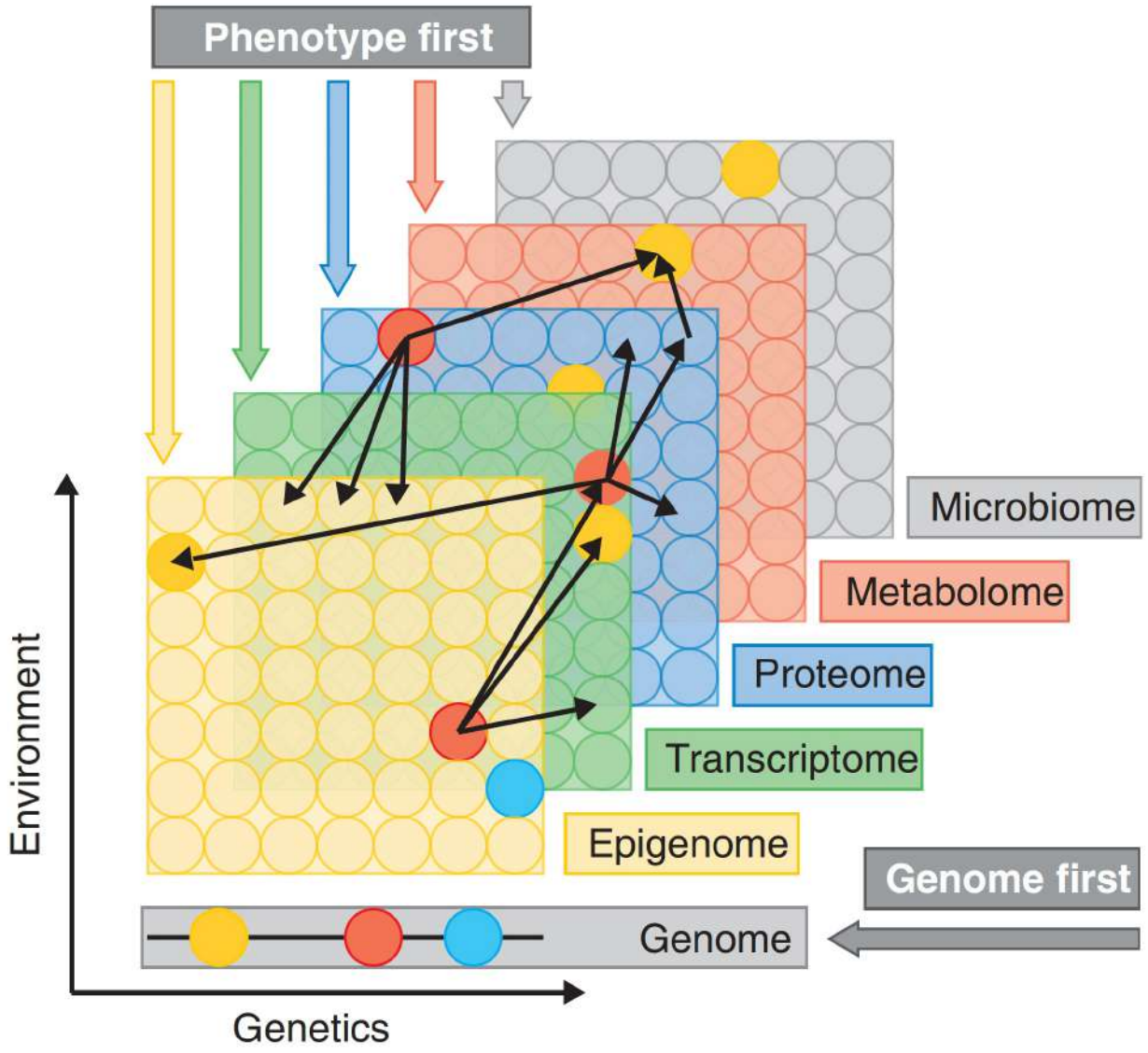


By mapping scRNA-seq-profiled cells onto spatially resolved coordinates through the integration with smFISH data, **spatial patterns of gene expression can be predicted for any gene measured in the scRNA-seq data set**. Through these predictions, novel spatial patterns of gene expression may be identified through the analysis of genes that were not profiled by smFISH

Stuart and Satija (2019) Nature Review Genetics

a multi-omic perspective

Multiple omics data types



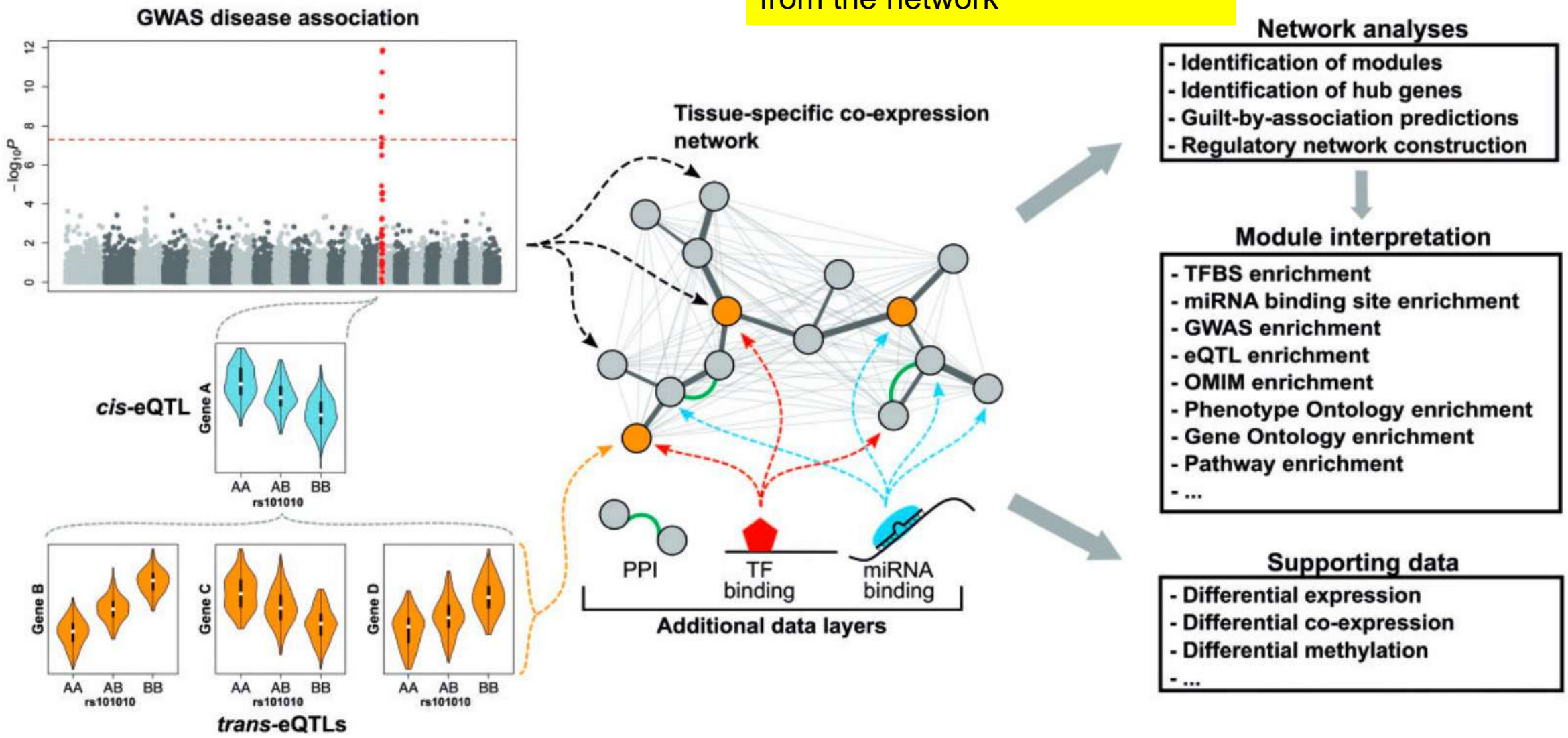
- Genome first or Phenotype first or environment first?
- Genome first -> GWAS

- “Locus-centered integration of additional omics layers can help to identify causal single nucleotide polymorphisms(SNPs) and genes at GWAS loci and then to examine how these perturb pathways leading to disease”

Hain *et al* (2017) Genome Biology

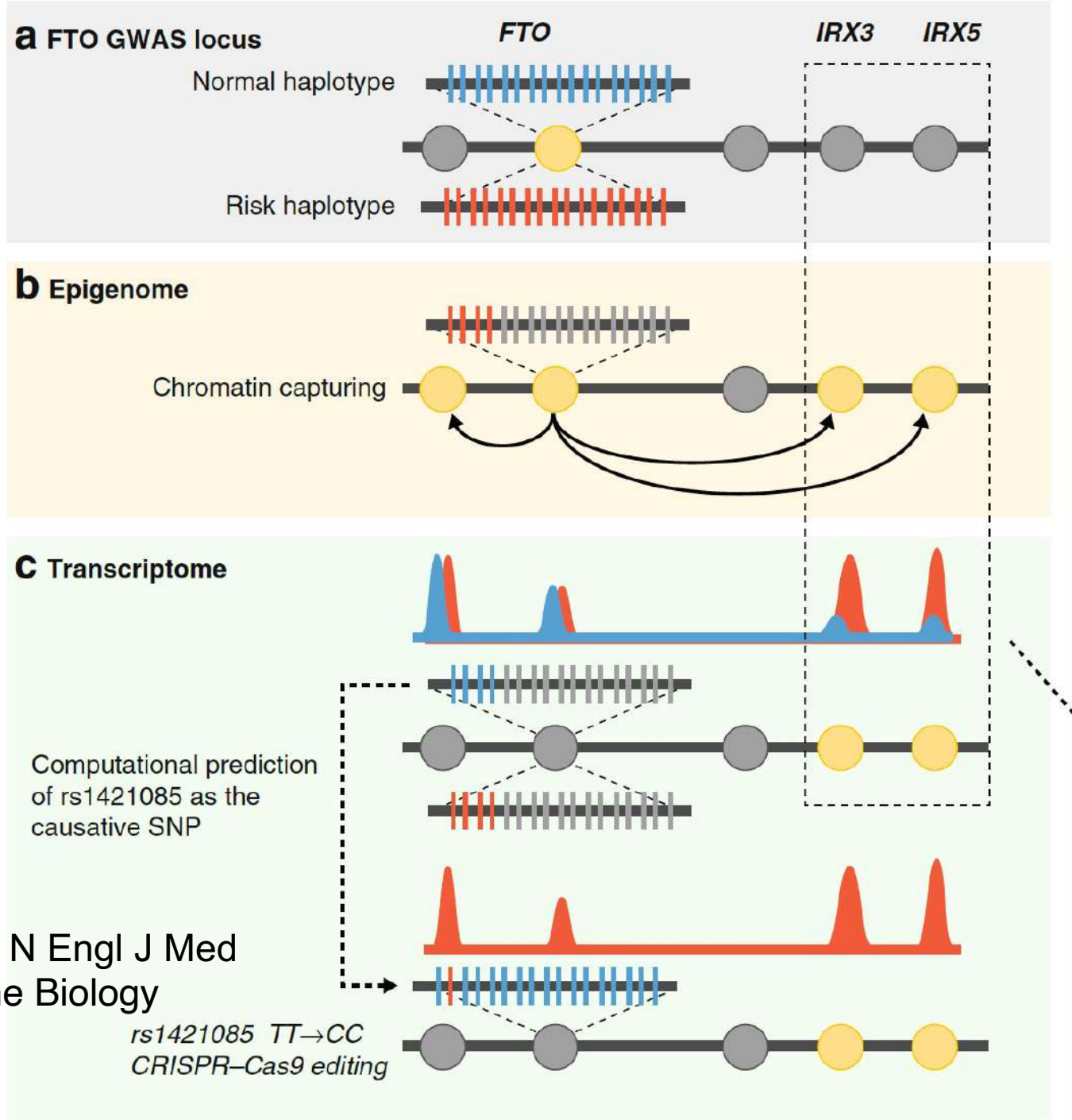
Integrating multi-omics to network

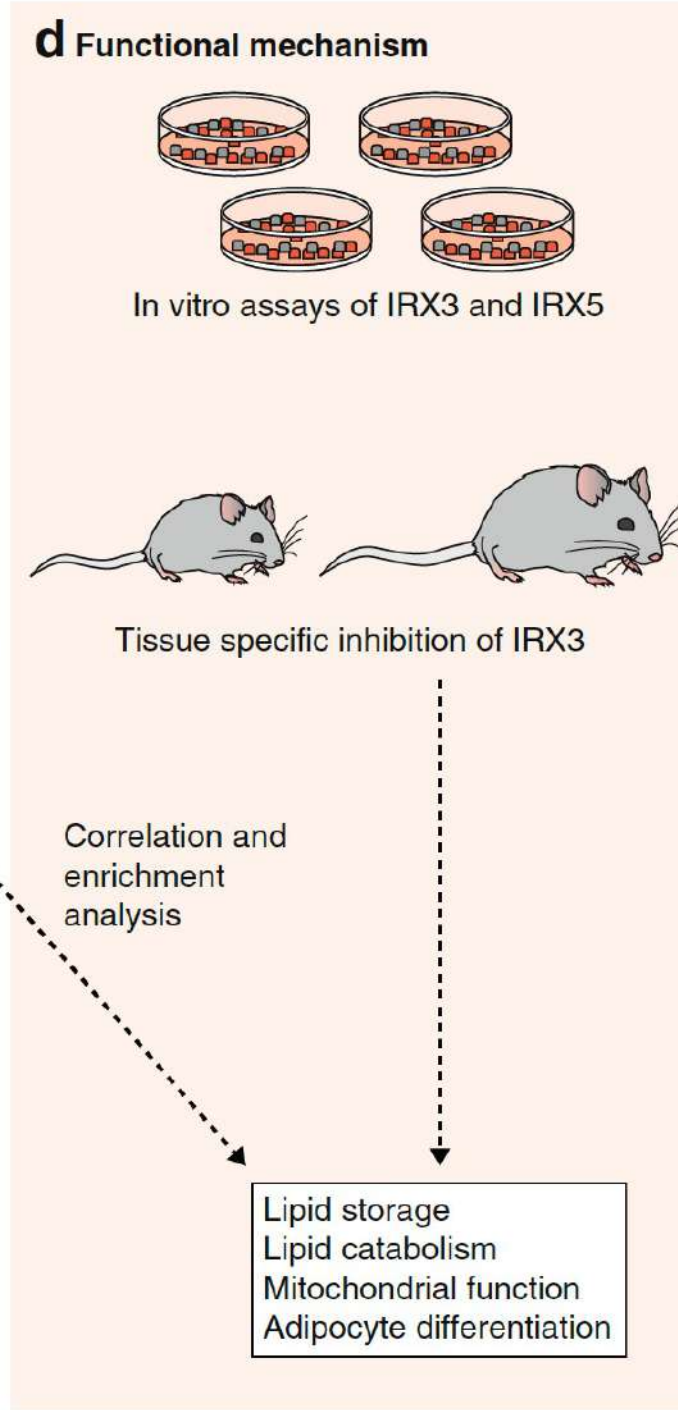
Various additional data can then be used to enrich and extract biological relevant information from the network



Dam *et al* (2016) Briefings in Bioinformatics

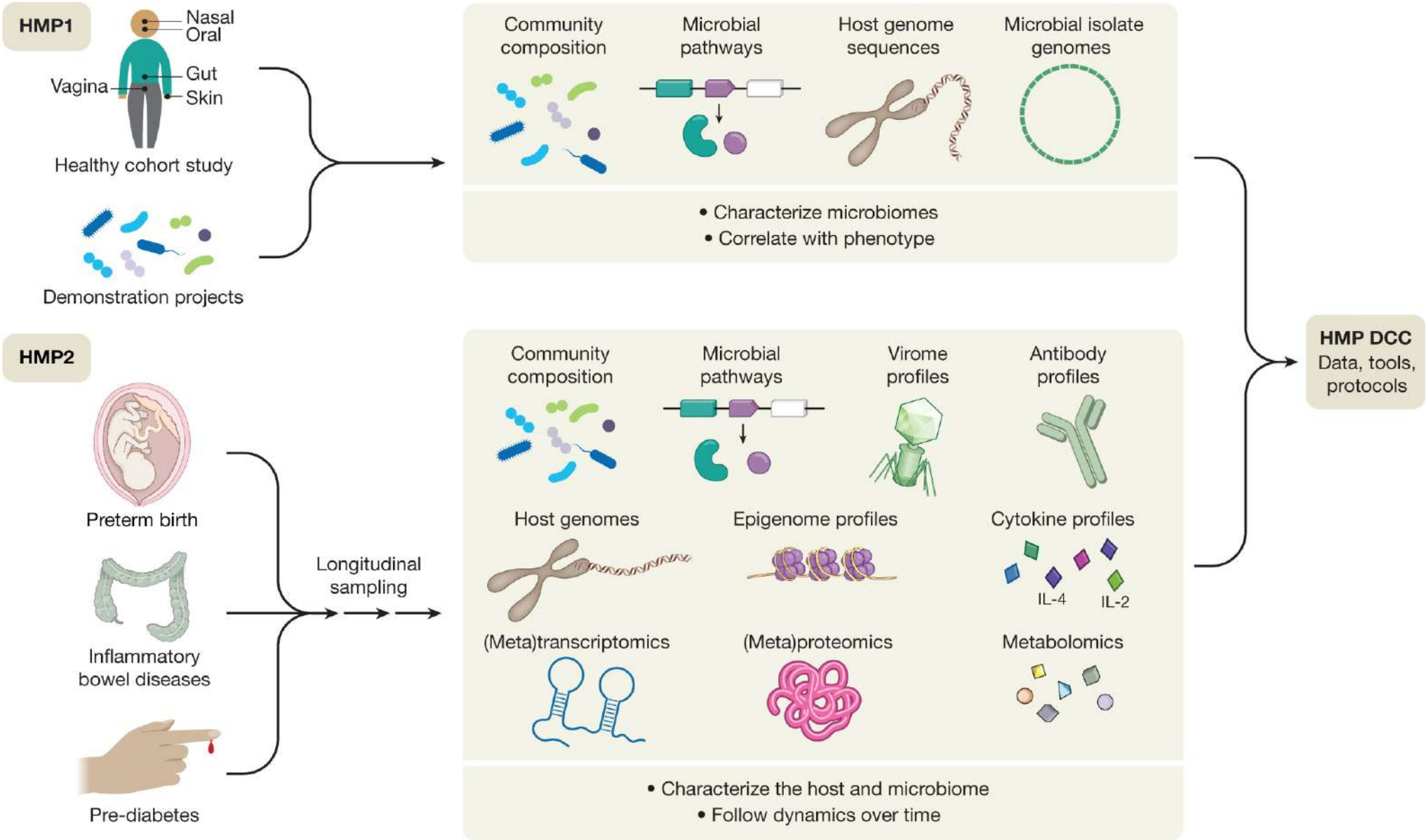
Example: FTO GWAS locus





Claussnitzer *et al* (2015) N Engl J Med
Hain *et al* (2017) Genome Biology

“a paradigm for future multi-omic studies of the human microbiome”



Proctor, L.M., Creasy, H.H., Fettweis, J.M. *et al.* The Integrative Human Microbiome Project. *Nature* **569,** 641–648 (2019). https://doi.org/10.1038/s41586-019-1238-8

Summaries

VIEWPOINT

The road ahead in genetics and genomics

Amy L. McGuire, Stacey Gabriel, Sarah A. Tishkoff, Ambroise Wonkam, Aravinda Chakravarti, Eileen E. M. Furlong, Barbara Treutlein, Alexander Meissner, Howard Y. Chang, Núria López-Bigas, Eran Segal and Jin-Soo Kim

Abstract | In celebration of the 20th anniversary of *Nature Reviews Genetics*, we asked 12 leading researchers to reflect on the key challenges and opportunities faced by the field of genetics and genomics. Keeping their particular research area in mind, they take stock of the current state of play and emphasize the work that remains to be done over the next few years so that, ultimately, the benefits of genetic and genomic research can be felt by everyone.

1. Making genomics truly equitable
2. Genome sequencing at population scale
3. A global view of human evolution
4. African genomics — the next frontier
5. Decoding multifactorial phenotypes
6. Enhancers and embryonic development
7. Spatial multi- omics in single cells
8. Unravelling the layers of the epigenome
9. Long non- coding RNAs: a time to build
10. FAIR genomics to track tumorigenesis
11. Integrating genomics into medicine
12. CRISPR genome editing enters the clinic

https://www.nature.com/articles/s41576-020-0272-6

New challenges

- So much data
- Technology advancement
- **Integrating different kinds of data (multi-omic)**
- High performance
- Reproducibility crisis

- Bioinformaticians as a profession
- Only biology has a specific term to refer to the use of computers in this discipline (‘bioinformatics’)
- Proper integration into academic curriculums

A personal take on science and society

World view

By Paul Nurse

Biology must generate ideas as well as data

> "It would have been a pity if Darwin had stopped thinking after describing the shapes and sizes of finch beaks."

Accepting a Nobel prize nearly two decades ago, my old friend Sydney Brenner had a warning for biology. **"We are drowning in a sea of data and starving for knowledge,"** he said. That admonishment, from one of the founders of molecular biology, who established the nematode worm *Caenorhabditis elegans* as a model organism, is even more relevant to biology today.

13 September 2021

https://www.nature.com/articles/d41586-021-02480-z

Yanai and Lercher *Genome Biology* (2020) 21:231
https://doi.org/10.1186/s13059-020-02133-w

Genome Biology

EDITORIAL Open Access

A hypothesis is a liability

Itai Yanai[1*] and Martin Lercher[2*]

" 'When someone seeks,' said Siddhartha, 'then it easily happens that his eyes see only the thing that he seeks, and he is able to find nothing, to take in nothing. [...] Seeking means: having a goal. But finding means: being free, being open, having no goal.' " Hermann Hesse

Shift in paradigm 2005-2021 (My personal take)

- Genome and transcriptome 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

Thank you