Bioinformatics: An introduction

Isheng Jason Tsai

GSB
Computation Class: Lecture 1

Welcome!

Lecture outline

- Bioinformatics and Computational biology – Origin and History

#Why is the above important?

- Case study: myself

#What should you gain in the first semester of GSB?

- NGS Analytics: usage and history (emphasis on **genomics**)

What about your own field?

- More case studies

What about your research?

Bioinformatics and Computational biology – Origin and History

- the very beginnings of bioinformatics occurred more than 50 years ago, when desktop computers were still a hypothesis and DNA could not yet be sequenced.”

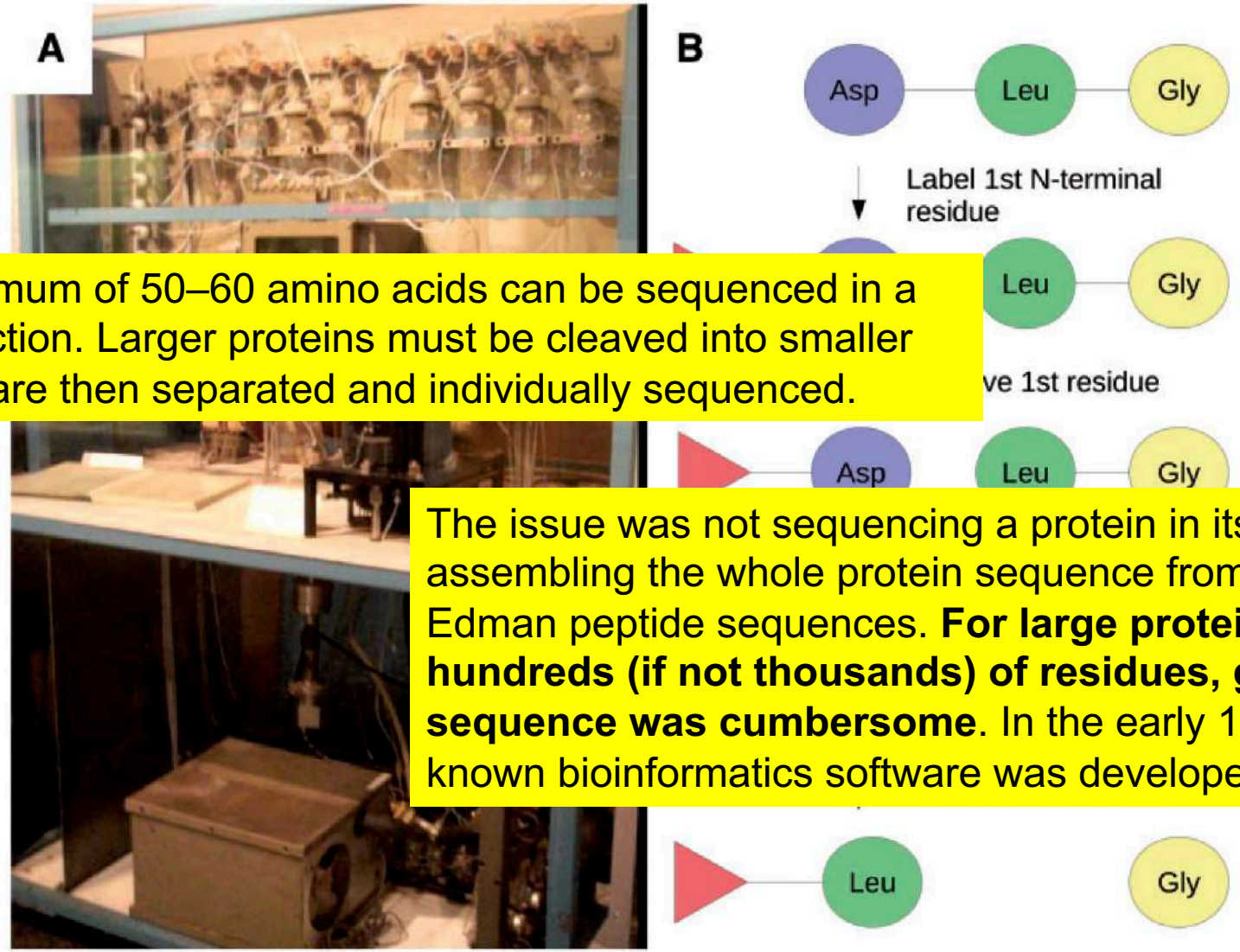
- The foundations of bioinformatics were laid in the early 1960s with the application of computational methods to protein sequence analysis (notably, *de novo* sequence assembly, biological sequence databases and substitution models).

- Later on, DNA analysis also emerged due to parallel advances in (i) molecular biology methods, which allowed easier manipulation of DNA, as well as its sequencing, and (ii) computer science, which saw the rise of increasingly miniaturized and more powerful computers, as well as novel software better suited to handle bioinformatics tasks. In the 1990s through the 2000s, major improvements in sequencing technology, along with reduced costs, gave rise to an exponential increase of data.

- The arrival of ‘Big Data’ has laid out new challenges in terms of data mining and management, calling for more expertise from computer science into the field.

- Universities are now fully integrating this discipline into the curriculum of biology students. Recent subdisciplines such as synthetic biology, systems biology and whole-cell modeling have emerged from the ever-increasing complementarity between computer science and biology.

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



A theoretical maximum of 50–60 amino acids can be sequenced in a single Edman reaction. Larger proteins must be cleaved into smaller fragments, which are then separated and individually sequenced.

The issue was not sequencing a protein in itself but rather assembling the whole protein sequence from hundreds of small Edman peptide sequences. **For large proteins made of several hundreds (if not thousands) of residues, getting back the final sequence was cumbersome**. In the early 1960s, one of the first known bioinformatics software was developed to solve this problem.

Figure 1. Automated Edman peptide sequencing. (**A**) One of the first automated peptide sequencers, designed by William J. Dreyer. (**B**) Edman sequencing: the first N-terminal amino acid of a peptide chain is labeled with phenylisothiocyanate (PITC, red triangle), and then cleaved by lowering the pH. By repeating this process, one can determine a peptide sequence, one N-terminal amino acid at a time.

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

Dayhoff: the first bioinformatician

Margaret Dayhoff (1925-1983)

- Designed one letter amino acid code
- Trained in quantum chemistry and mathematics, she became interested in proteins and molecular evolution around 1960.
- to explore mathematical approaches for analysing amino-acid sequence data
- Her initial project was writing a series of FORTRAN programs to determine the amino-acid sequences of protein molecules.

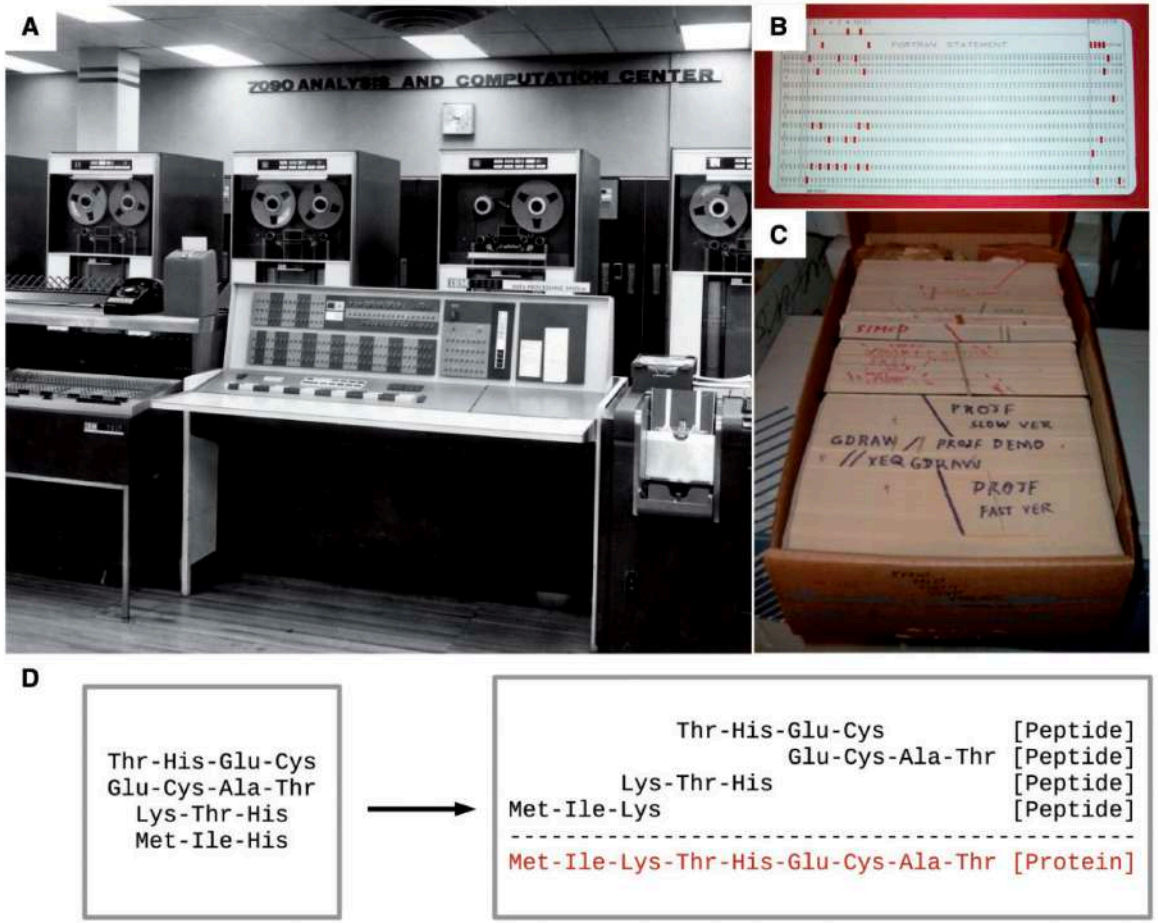


Figure 2. COMPROTEIN, the first bioinformatics software. (A) An IBM 7090 mainframe, for which COMPROTEIN was made to run. (B) A punch card containing one line of FORTRAN code (the language COMPROTEIN was written with). (C) An entire program's source code in punch cards. (D) A simplified overview of COMPROTEIN's input (i.e. Edman peptide sequences) and output (a consensus protein sequence).

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

Hagen (2000) Nat Rev Genetics doi: 10.1038/35042090

Ancestral sequences and Molecular clock (Emile Zuckerkandl and Linus Pauling)

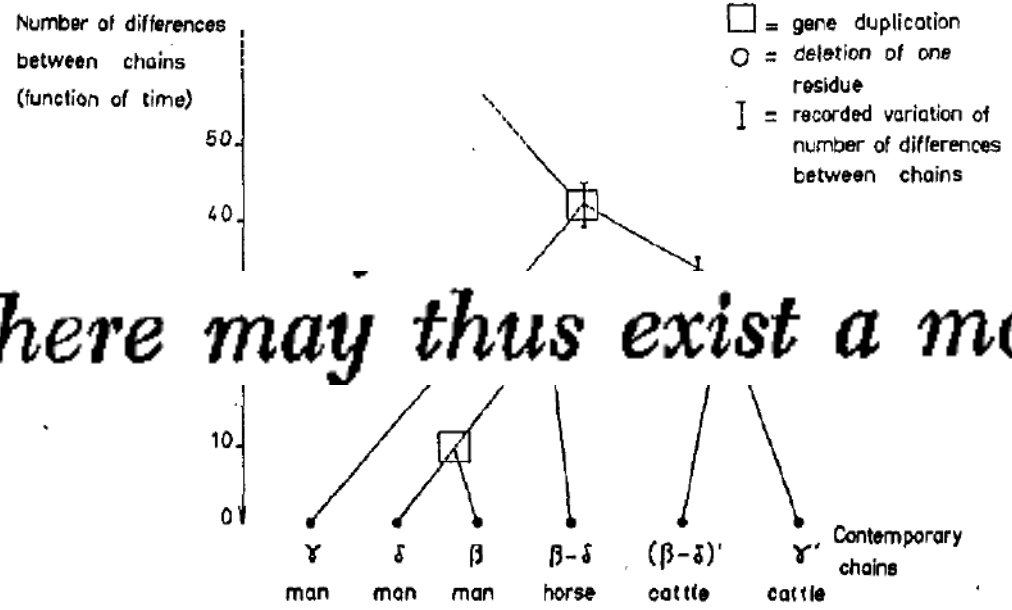
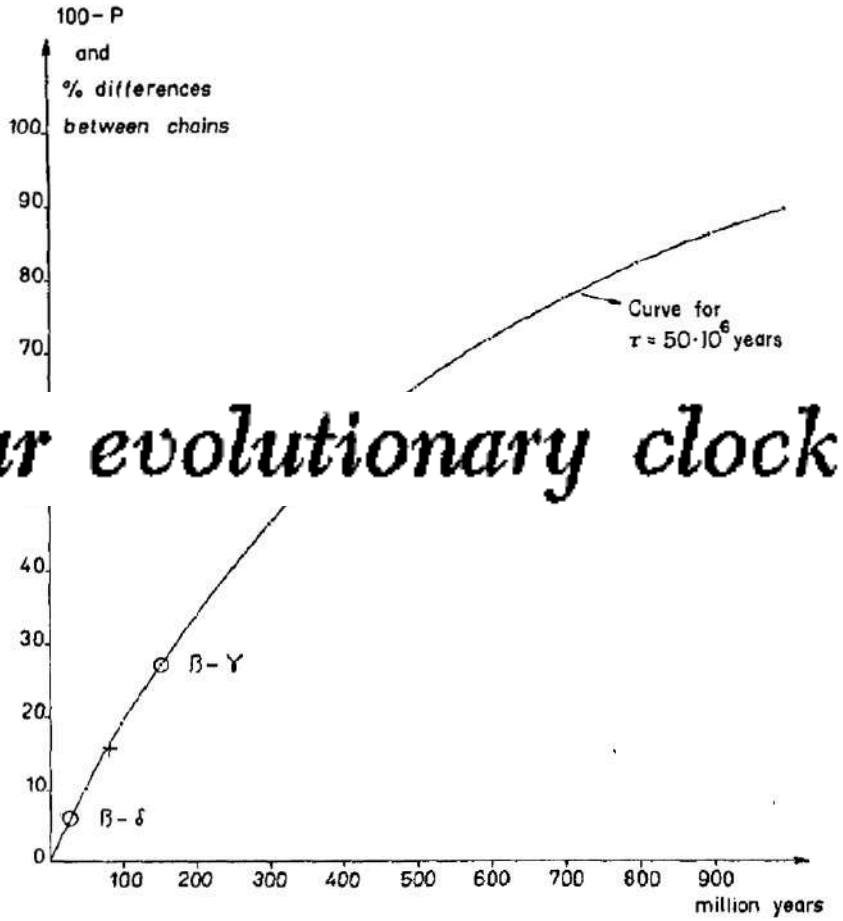


FIG. 4. Probable evolutionary relationship of some mammalian hemoglobin chains.

There may thus exist a molecular evolutionary clock.[1]

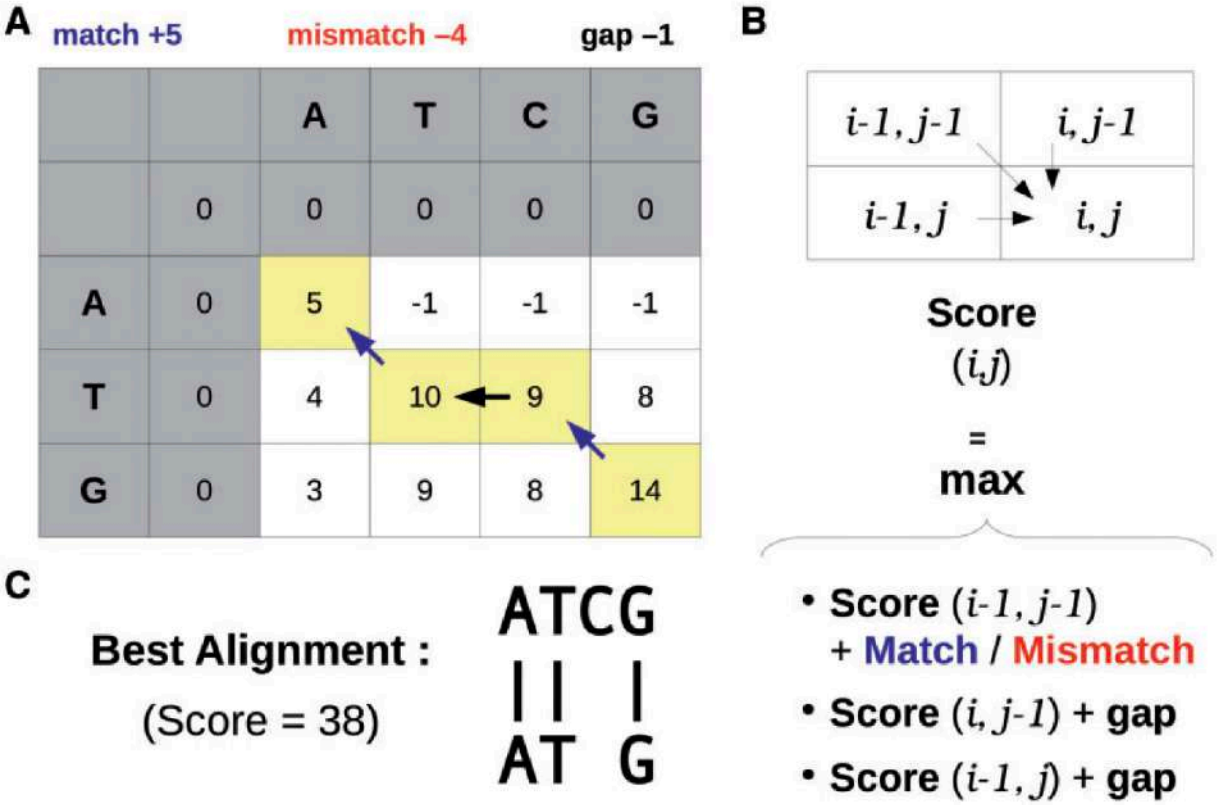


"Zuckerkandl and Pauling hypothesized that orthologous proteins evolved through divergence from a common ancestor. Consequently, by comparing the sequence of hemoglobin in currently extant organisms, it became possible to predict the 'ancestral sequences' of hemoglobin and, in the process, its evolutionary history up to its current forms"

Evolutionary divergence and convergence in proteins
Zuckerkandl, E. and Pauling, L (1965)

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

A mathematical framework for sequence alignments



		A	T	C	G
	0	0	0	0	0
A	0	5	-1	-1	-1
T	0	4	10	9	8
G	0	3	9	8	14

Table 1. An excerpt of the PAM1 amino acid substitution matrix

10^4 P[a]		Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	... ...	Val V
Ala	A	9867	2	9	10	3	8	...	18
Arg	R	1	9913	1	0	1	10	...	1
Asn	N	4	1	9822	36	0	4	...	1
Asp	D	6	0	42	9859	0	6	...	1
Cys	C	1	1	0	0	9973	0	...	2
Gln	Q	3	9	4	5	0	9876	...	1
...	...	...	...	...	...	...	...	...	...
Val	V	13	2	1	1	3	2	...	9901

[a]Each numeric value represents the probability that an amino acid from the *i*-th column be substituted by an amino acid in the *j*-th row (multiplied by 10 000).

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

1970-2000s – Paradigm shifts and parallel advances in biology and computer science

- Protein sequencing to DNA sequencing (faster / cheaper)
- Use DNA sequences to infer phylogenetic trees
- Sequence of marker genes and genomes
- Beyond sequences (structural bioinformatics)

- Faster computers
- GPUs
- Free software movement
- New Programming languages (Perl created by Larry Wall in 1987)

- Internet
- Online databases (NCBIs)

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

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

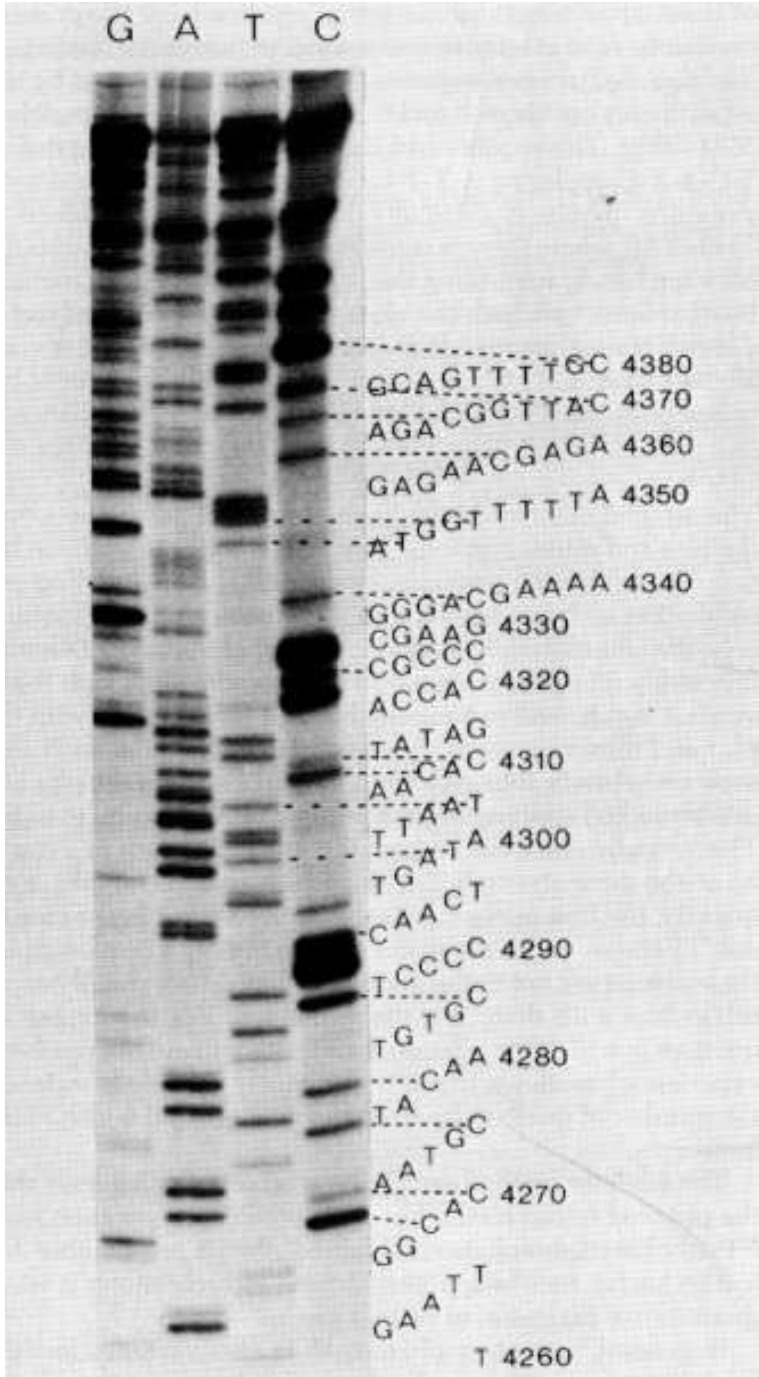
DNA sequencing with chain-terminating inhibitors

(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

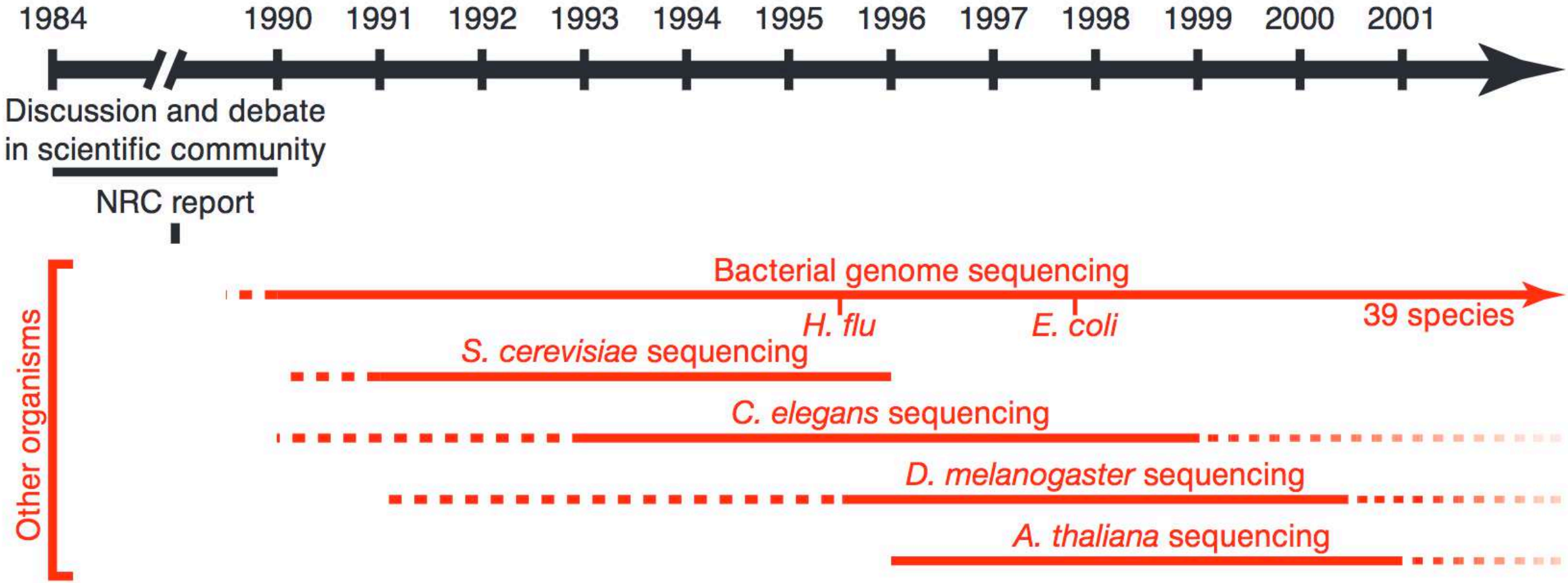
Contributed by F. Sanger, October 3, 1977



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

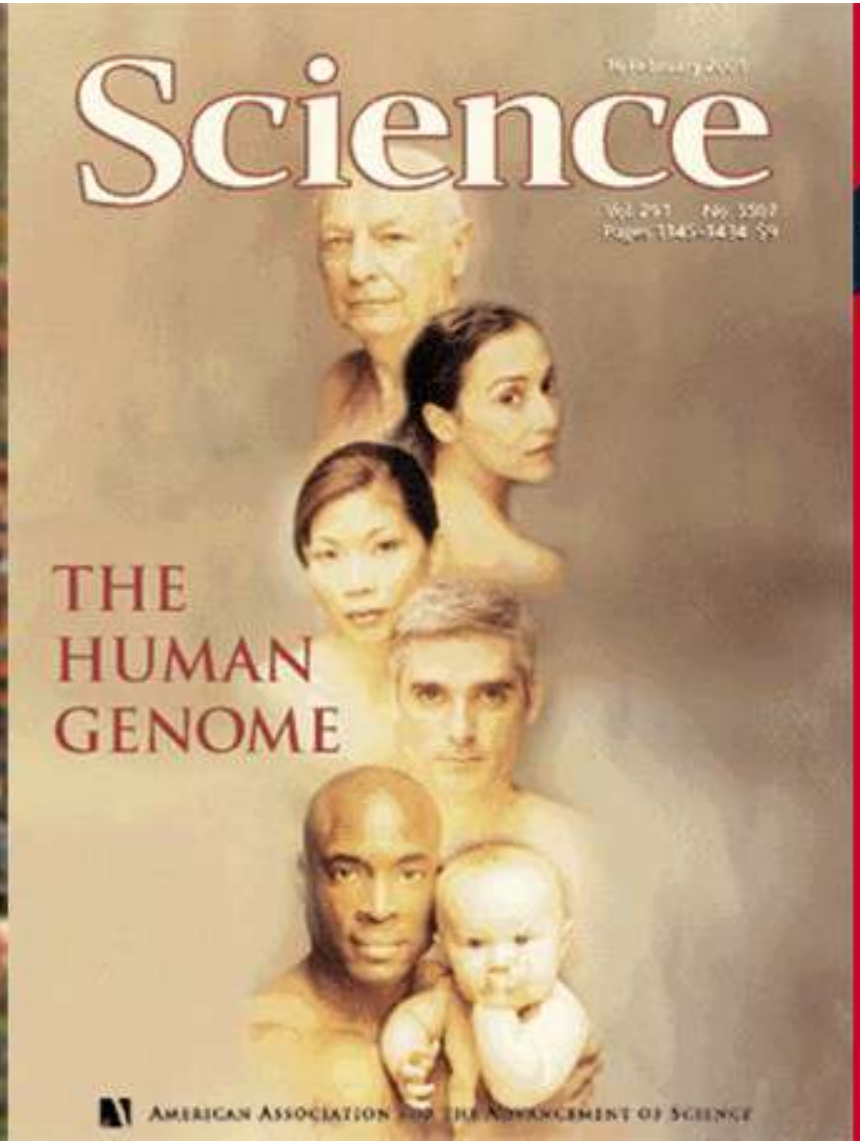
ABI 3730xi at TIGR

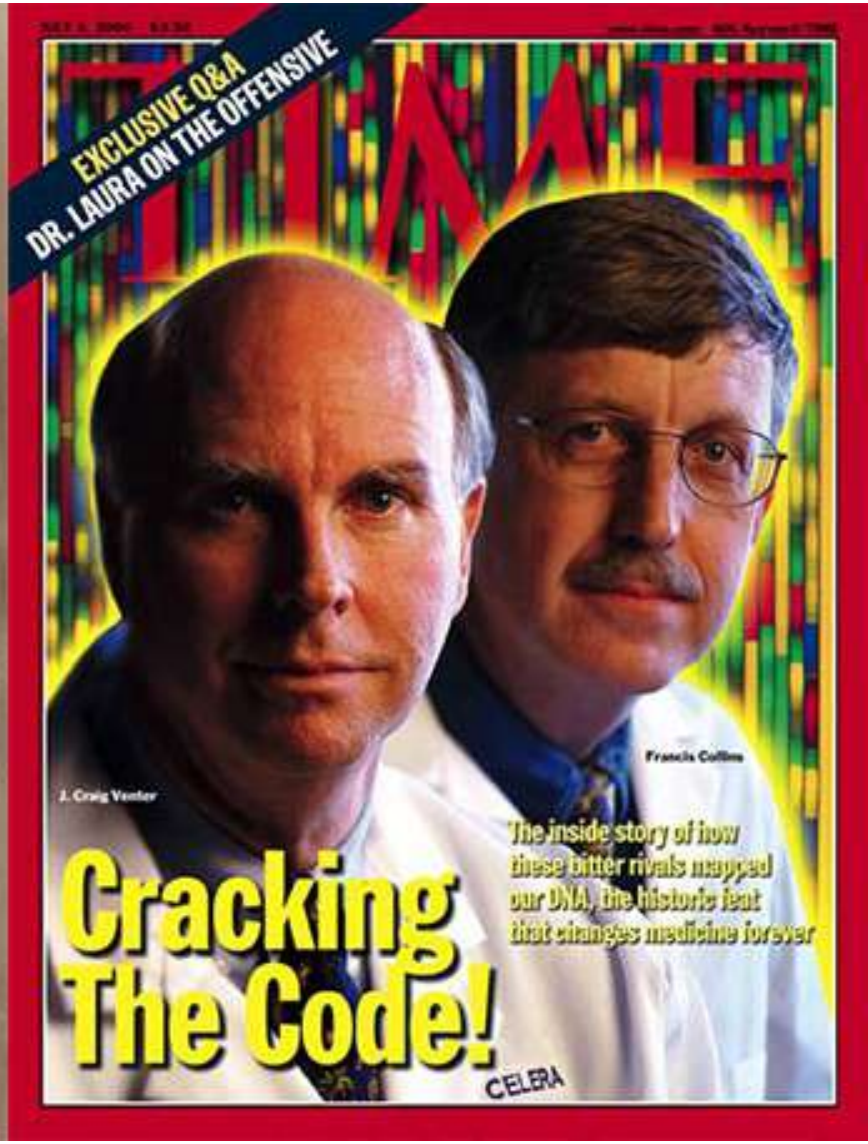
https://www.flickr.com/photos/jurvetson/57080968



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


nature
the human genome


Science
THE HUMAN GENOME
AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE

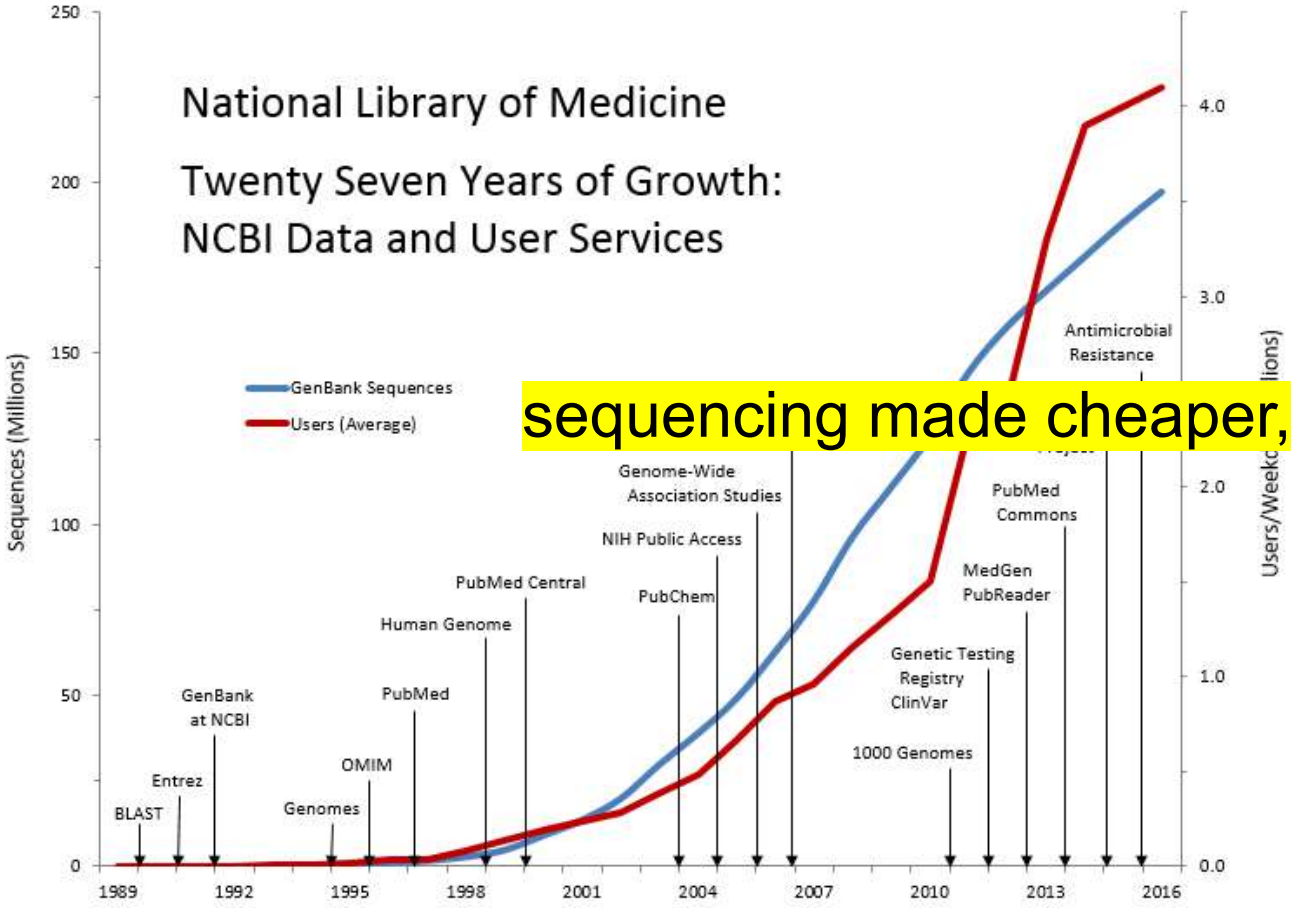

EXCLUSIVE Q&A
DR. LAURA ON THE OFFENSIVE
J. Craig Venter
Francis Collins
The inside story of how these bitter rivals mapped our DNA, the historic feat that changes medicine forever
Cracking The Code!
CELERA

Calculating the economic impact of the Human Genome Project

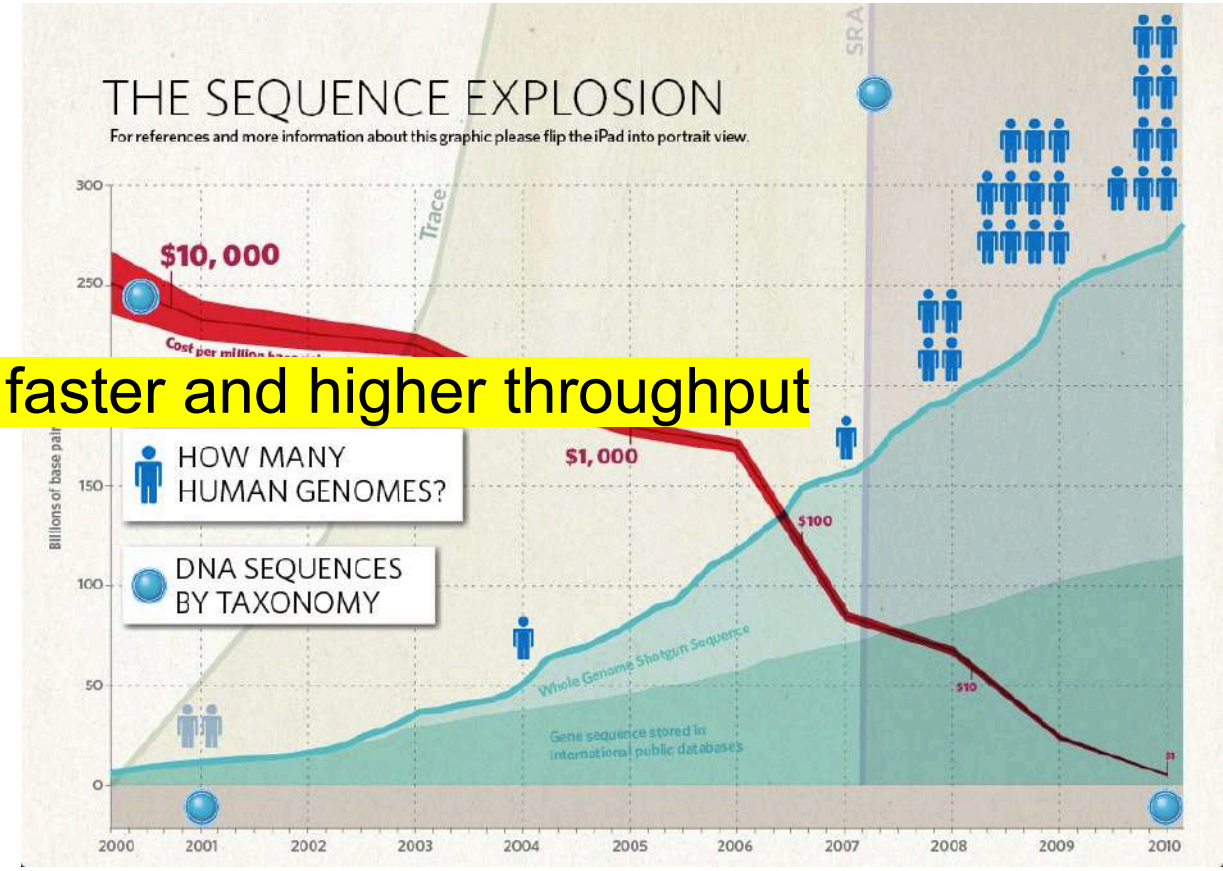
Public funding of scientific R&D has a significant positive impact on the wider economy, but quantifying the exact impact of research can be difficult to assess. A new report by research firm Battelle Technology Partnership Practice estimates that **between 1988 and 2010, federal investment in genomic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

FOCUS ON GENOMES OF ICELANDERS

ARTICLES

nature genetics

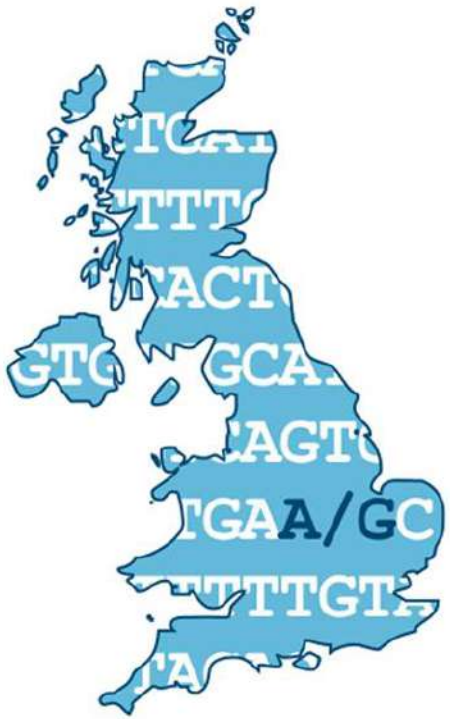
Large-scale whole-genome sequencing of the Icelandic population

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

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

The blood of a thousand Icelanders.
Photo: Chris Lund

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



UK 10K

RARE GENETIC VARIANTS IN HEALTH AND DISEASE

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

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

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

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

https://www.uk10k.org/

Cold Spring Harbor Laboratory
bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

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

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

The UK Biobank is a prospective study of 502,543 individuals, combining extensive phenotypic and genotypic data with streamlined access for researchers around the world.

https://www.biorxiv.org/content/10.1101/572347v1

	Variants in WES, n=49,960 Participants		Median Per Participant (IQR)	
	# Variants	# Variants MAF<1%	# Variants	# Variants MAF<1%
Total	9,693,526	9,547,730	48,982 (627)	1,626 (133)
Targeted Regions[1]	4,735,722	4,665,684	24,332 (283)	780 (63)
Variant Type[1]				
SNVs	4,520,754	4,453,941	23,529 (276)	739 (61)
Indels	214,968	211,743	803 (29)	42 (10)
Multi-Allelic	591,340	580,728	3,388 (63)	117 (18)
Functional Prediction				
Synonymous	1,229,303	1,203,043	9,619 (128)	228 (28)
Missense	2,498,947	2,472,384	8,781 (137)	380 (39)
LOF (any transcript)	231,631	230,790	219 (16)	24 (8)
LOF (all transcripts)	153,903	153,441	111 (12)	15 (6)

Table 2 | Summary statistics for variants in sequenced exomes of 49,960 UKB participants

https://www.biorxiv.org/content/10.1101/572347v1

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

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

Smithsonian

Institute for

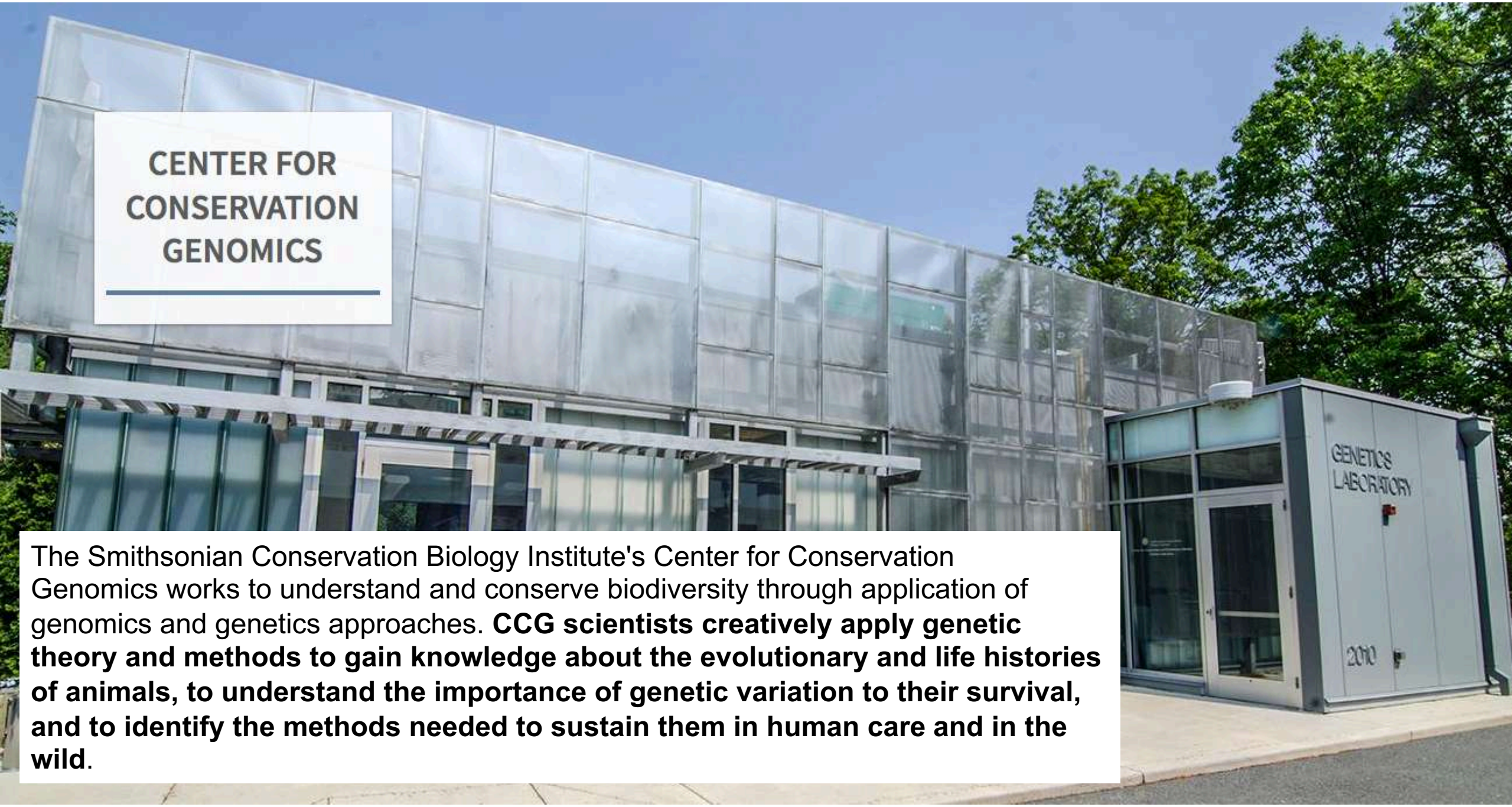
Biodiversity Genomics

How do we sustain life on our changing planet?

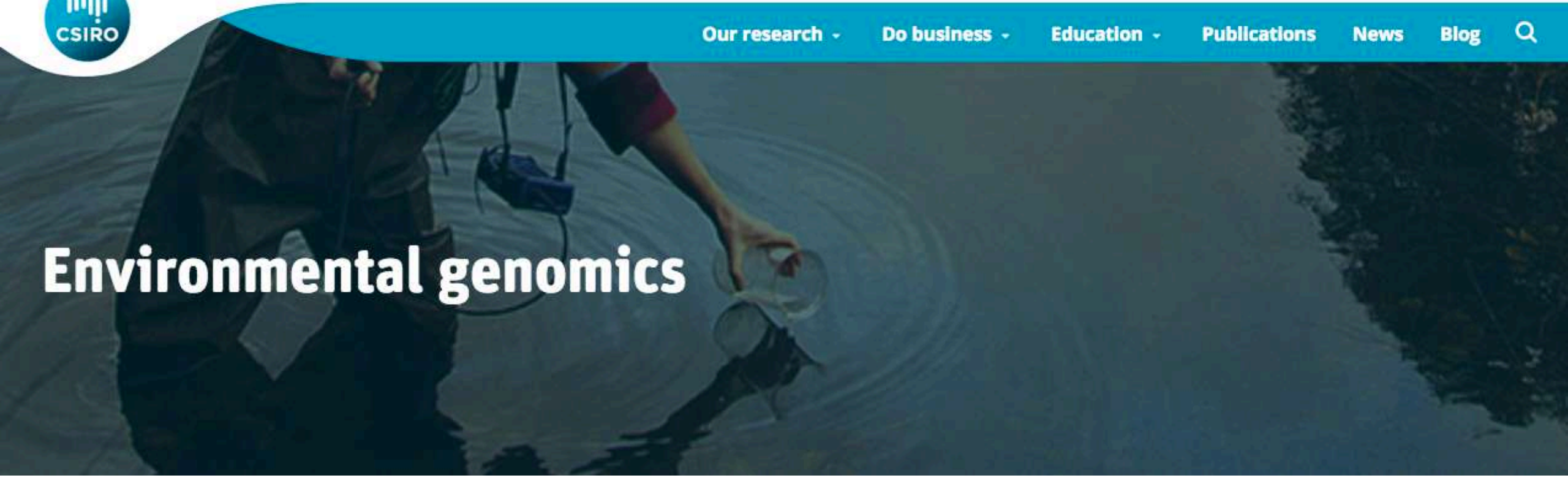
Biodiversity—our planet's complex web of interdependent species and ecosystems—is critical to our survival and includes the water we drink, the air we breathe, the food we eat, the medicines that heal, and the soils that nurture.

But our biodiversity faces serious challenges.

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



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



Environmental genomics

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

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/

Problems and Challenges

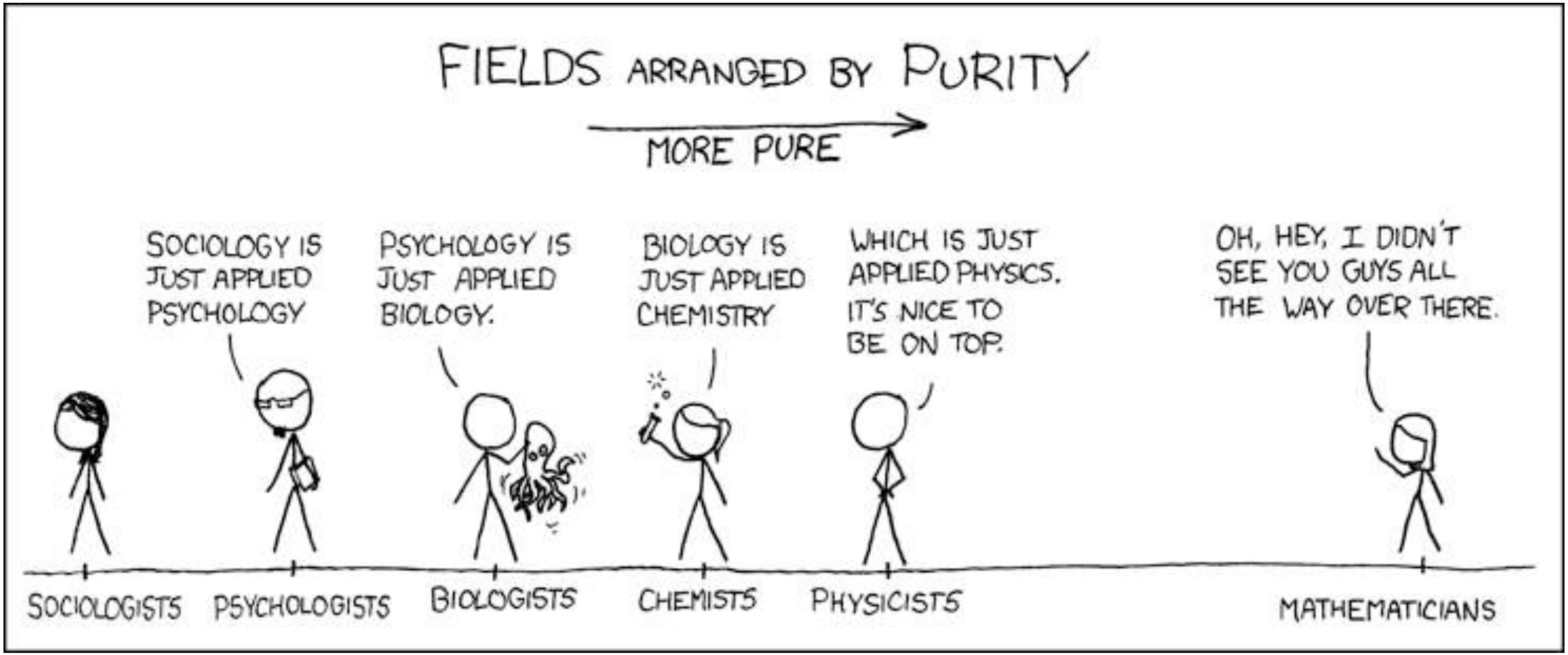
- Do it because you can (lots of $$, want to jump in)
- Don’t hate it because you don’t know how to do it
  - Typical scenario: “We should focus on more traditional methods because NGS is expensive”
  - Typical scenario 2: “These people who do mathematics (?) don’t know what ecology/biology/conservation are”

- Biological Big data or too much data
- Integrating different kinds of data
- 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 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

Purity?



https://xkcd.com/435/

Current perspective and challenges

Garcia-Milian et al., (2018) PeerJ

Data challenges of biomedical researchers in the age of omics

Rolando Garcia-Milian[1], Denise Hersey[2], Milica Vukmirovic[3] and Fanny Duprilot[4]

[1] Bioinformatics Support Program, Research and Education Services, Cushing/Whitney Medical Library, Yale University, New Haven, CT, United States of America
[2] Science Libraries, Lewis Science Library, Princeton University, Princeton, NJ, United States of America
[3] Pulmonary Critical Care & Sleep Medicine, Yale School of Medicine, Yale University, New Haven, CT, United States of America
[4] Service commun de la documentation, Université Denis Diderot (Paris VII), Paris, France

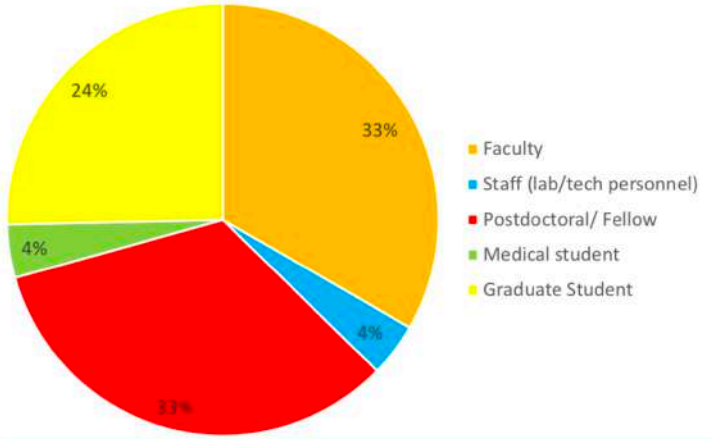


Figure 1 Response to the question: which of the following best describes your role? Total responses: 157.

Full-size DOI: 10.7717/peerj.5553/fig-1

Data analysis	Not important	Important	Very important	Total responses
Analysis of high-throughput data (e.g., microarray data, RNA-seq)	16 (11.9%)	21 (15.7%)	97 (72.4%)	134 (100%)
Signaling, network, and pathway analysis	13 (10%)	33 (25.4%)	84 (64.6%)	130 (100%)
Functional analysis of high-throughput data	20 (15.4%)	36 (27.7%)	74 (56.9%)	130 (100%)
Transcription factor and gene regulatory sequence analysis	25 (19.1%)	38 (29.0%)	68 (51.9%)	131 (100%)
Integrated searches of literature and high-throughput data	15 (11.6%)	50 (38.8%)	64 (49.6%)	129 (100%)
DNA/protein sequence manipulation and analysis	17 (13.3%)	50 (39.1%)	61 (47.7%)	128 (100%)
SNP, genetic variation, Genome wide association data analysis	42 (31.8%)	42 (31.8%)	48 (36.4%)	132 (100%)
Other data analysis needs	11 (43.4%)	4 (12.5%)	17 (53.1)	32 (100%)

Garcia-Milian et al., (2018) PeerJ

A global perspective on evolving bioinformatics and data science training needs

Teresa K. Attwood, Sarah Blackford, Michelle D. Brazas, Angela Davies and Maria Victoria Schneider

Corresponding author: Teresa K. Attwood, School of Computer Science, The University of Manchester, Manchester, M13 9PL, UK. Tel.: +44 (0)161-275-6259; E-mail: teresa.k.attwood@manchester.ac.uk

Table 3. Summary of some of the most important training needs reported in recent surveys

Survey	Training needs						
	Data analysis (including visualization/ interpretation)	Data mining/ manipulation/ management	Data integration	Scaling to cloud/HPC	Basic program-ming/scripting	Statistics	Bioinformatics tools/resources
SEB 2013	■					■	
GOBLET 2014	■					■	
ABPI 2014	■	■				■	■
ELIXIR 2014	■	■				■	■
NSF 2016		■	■	■			
EMBL-ABR 2016	■	■	■		■		■

Attwood et al., (2019) Briefings in Bioinformatics

Challenges

Personnel challenges

Among the 15 interviewees, 13 commented on challenges related to personnel. Experts in data analysis are in high demand and many labs, particularly smaller ones, have difficulty getting access to staff with those skill sets. As a result, researchers feel that valuable time is wasted either waiting to work with an expert or attempting to do the analysis with their own staff.

Training challenges

Training is also a significant challenge to successfully analyze data. Among the 15 interviewees, 10 commented on problems related to the lack of training that would allow them to conduct their own data analysis.

In order to efficiently use the preponderance of new tools and resources, a certain level of expertise in data analysis is required. Unfortunately, researchers recognize that they do not always have a strong background in this type of work.

Garcia-Milian et al., (2018) PeerJ

Skills needed as a bioinformatician (Defining training in bioinformatics)

PLOS COMPUTATIONAL BIOLOGY

MESSAGE FROM ISCB

The development and application of bioinformatics core competencies to improve bioinformatics training and education

Nicola Mulder[1‡*], Russell Schwartz[2‡], Michelle D. Brazas[3], Cath Brooksbank[4], Bruno Gaeta[5], Sarah L. Morgan[4], Mark A. Pauley[6], Anne Rosenwald[7], Gabriella Rustici[8], Michael Sierk[9], Tandy Warnow[10], Lonnie Welch[11]

EDUCATION

Fostering bioinformatics education through skill development of professors: *Big Genomic Data Skills Training for Professors*

Yingqian Ada Zhan[1], Charles Gregory Wray[2*], Sandeep Namburi[1], Spencer T. Glantz[1], Reinhard Laubenbacher[1,3], Jeffrey H. Chuang[1,4*]

Zhan et al., (2019) PLOS Computational Biology

Table 1. Program structure of *Big Genomic Data Skills Training for Professors.*

	Teaching Focus	Specific Topics
JAX BD2K Program Structure	Basics	High-throughput sequencing technologies
		Statistics
		Scripting in R/UNIX
	Modules	RNA-Seq
		Cancer variant
		Exome
		Microbiome
		ChIP-Seq
	Miscellaneous	Setting up educational cloud and grants
		Curriculum discussion
		Slack discussion community

Table 2. Description of data analysis modules in *Big Genomic Data Skills Training for Professors.*

Modules	Skills	Biology Question	Platform	Data Source
RNA-Seq	Differential gene expression analysis and gene set enrichment	What are genes affected by Pax6 knockout in male mice?	Galaxy	Mitchell and colleagues 2017
Cancer Variant	Data manipulation—grouping and sorting	What is the common driver mutation in three melanoma tumors?	Excel/R	Berger and colleagues 2012
Exome	Variant calling and filtering	Identify the exonic variant and gene responsible for the phenotype of "Leg dragger"* in mice.	Galaxy	Fairfield and colleagues 2011
Microbiome	16S analysis and bacterial taxon cataloging	What is the role of the microbiome in the development of type 1 diabetes in infants?	R/UNIX	https://pubs.broadinstitute.org/diabimmune
ChIP-Seq	Peak calling and motif analysis	Identify CTCF binding motif.	UNIX (Cloud)	ENCFF000ARV, ENCFF000ARP, ENCFF000ARK

You?

SKILLS SPECTRUM

There are **three** essential skill sets bioinformaticians need. Here's where to start.

1. COMMAND

Understand how **Unix** commands work.

2. PROGRAM

Learn **Python**, a basic language. Then consider **R**, a useful language for handling statistics.

3. DATA

Understanding what type of data is in different kinds of **databases**, and how to mine it, is essential. Learning relational database techniques is another **plus**.

NatureJobs (2016) doi:10.1038/nj0478

Personal journey

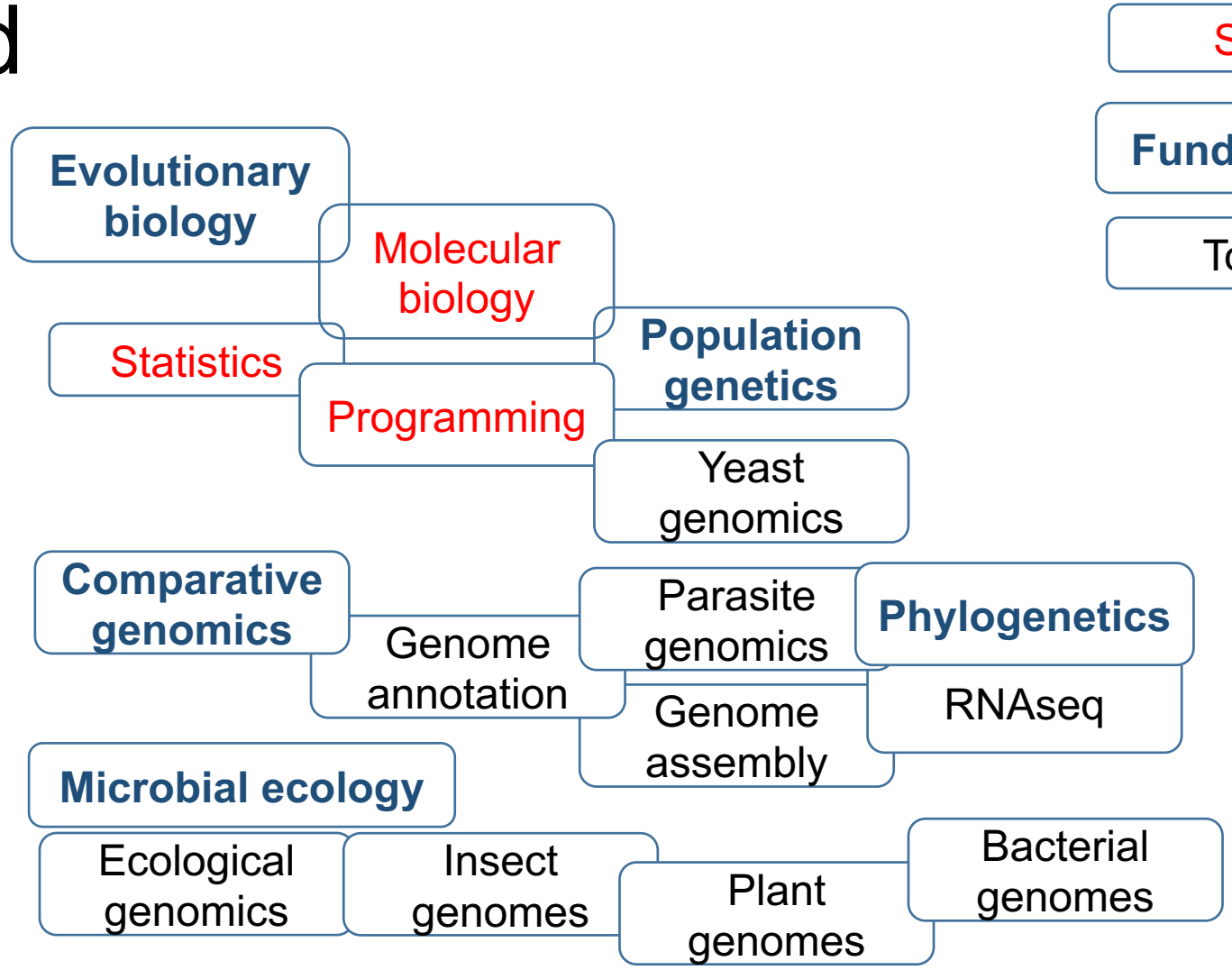
My background

Undergraduate:
Biochemistry and Genetics

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

2009-14 ; Postdoc:
Genomics & parasitology

2015 - ; Academia Sinica:
Microbial diversity & Bioinformatics

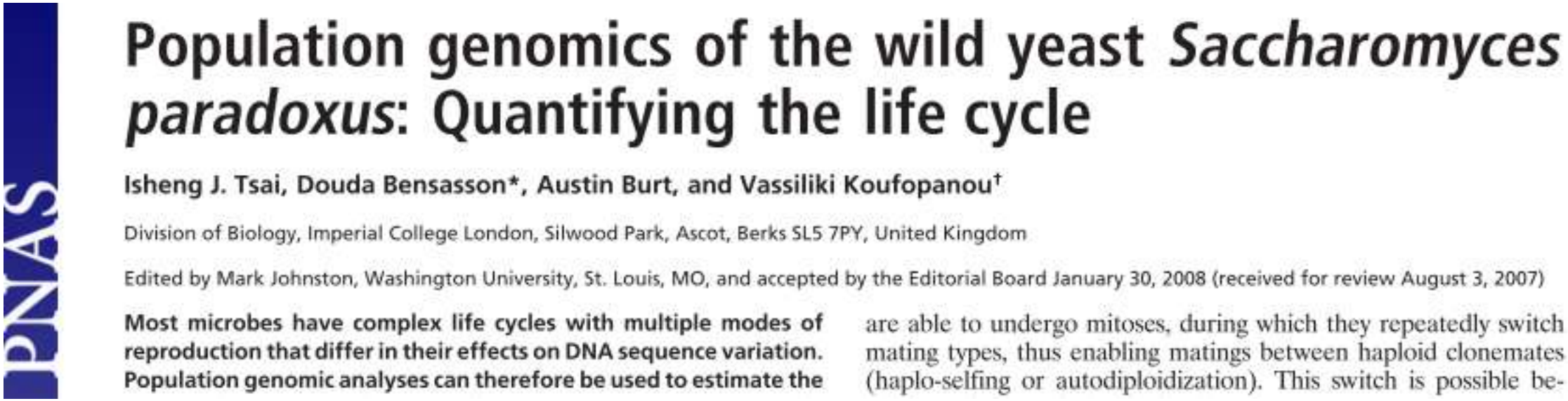


- **39 publications**
  (2 Nature, 1 Science, 2 Nature Genetics, 2 PNAS,
  3 Genome Biology, 1 Molecular Ecology)

2005 – *Saccharomyces paradoxus*

- Capillary read sequenced full Chromosome III **(~315kb**) of 20 isolates
- Costed £**750k**
- One of the first scale re-sequencing projects

- Took me 3 years to sequence, align, annotate and analyse (= PhD)



PNAS

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

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

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

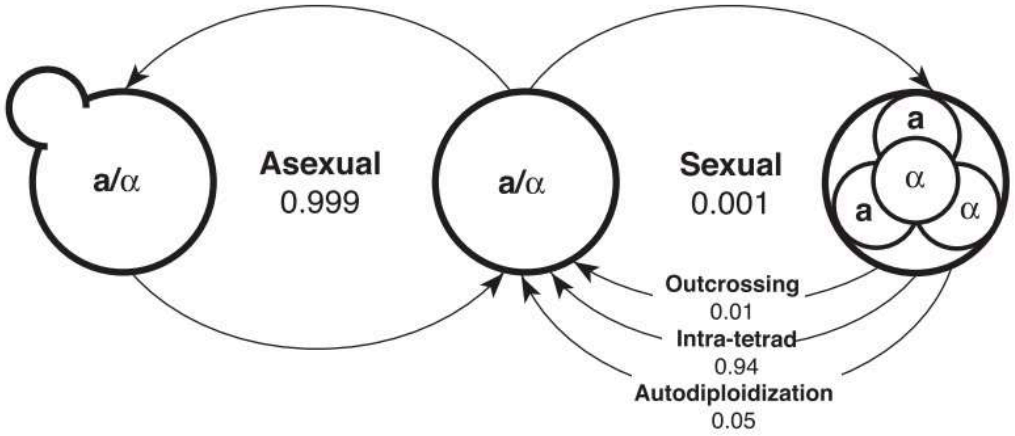
Edited by Mark Johnston, Washington University, St. Louis, MO, and accepted by the Editorial Board January 30, 2008 (received for review August 3, 2007)

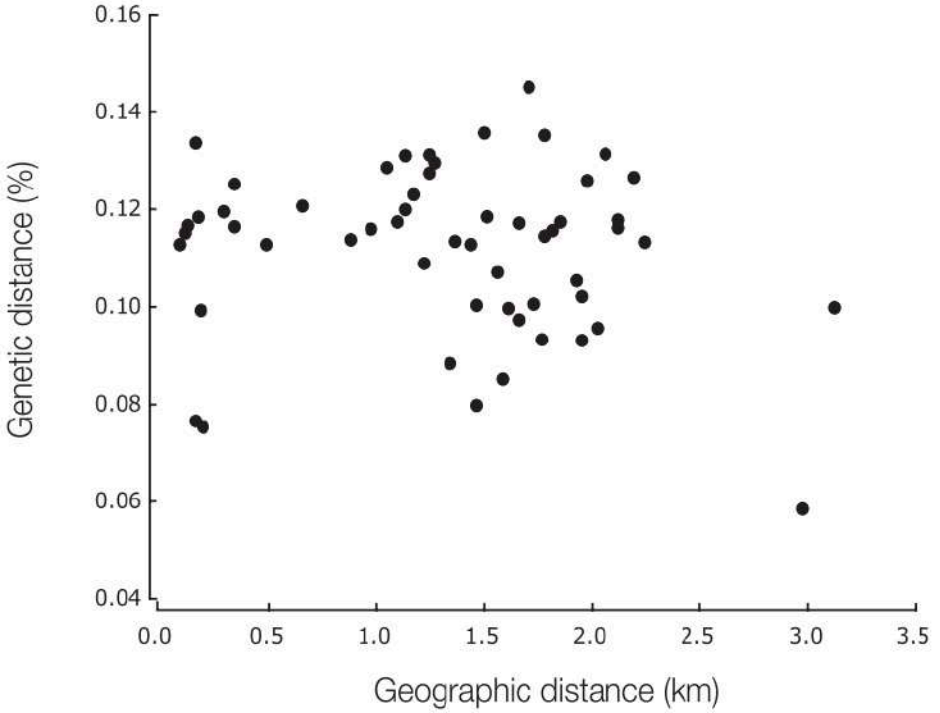
Most microbes have complex life cycles with multiple modes of reproduction that differ in their effects on DNA sequence variation. Population genomic analyses can therefore be used to estimate the

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

2005 – *Saccharomyces paradoxus*

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





2009 – *Saccharomyces* resequencing genome project

- 70 isolates at 1X-10X coverage
- ~2 years project with 26 authors
- At the start of NGS period (36bp Solexa reads)

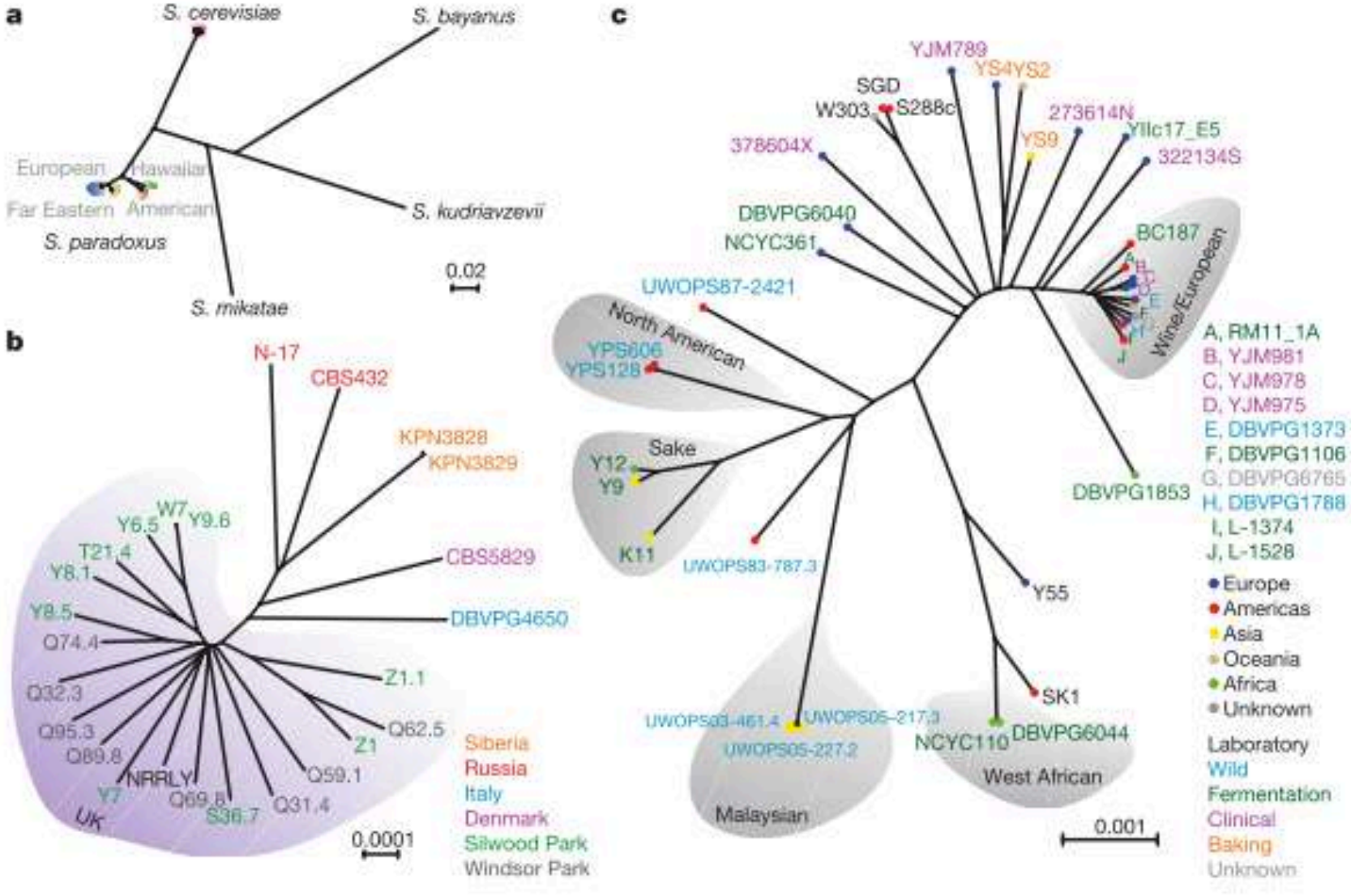
- **Now= We are collecting and sequencing hundreds of isolates in Taiwan**

Population genomics of domestic and wild yeasts

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

2009 – *Saccharomyces* resequencing genome project

Phylogeny of ~70 isolates



2013 – Tapeworm genome project

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

ARTICLE

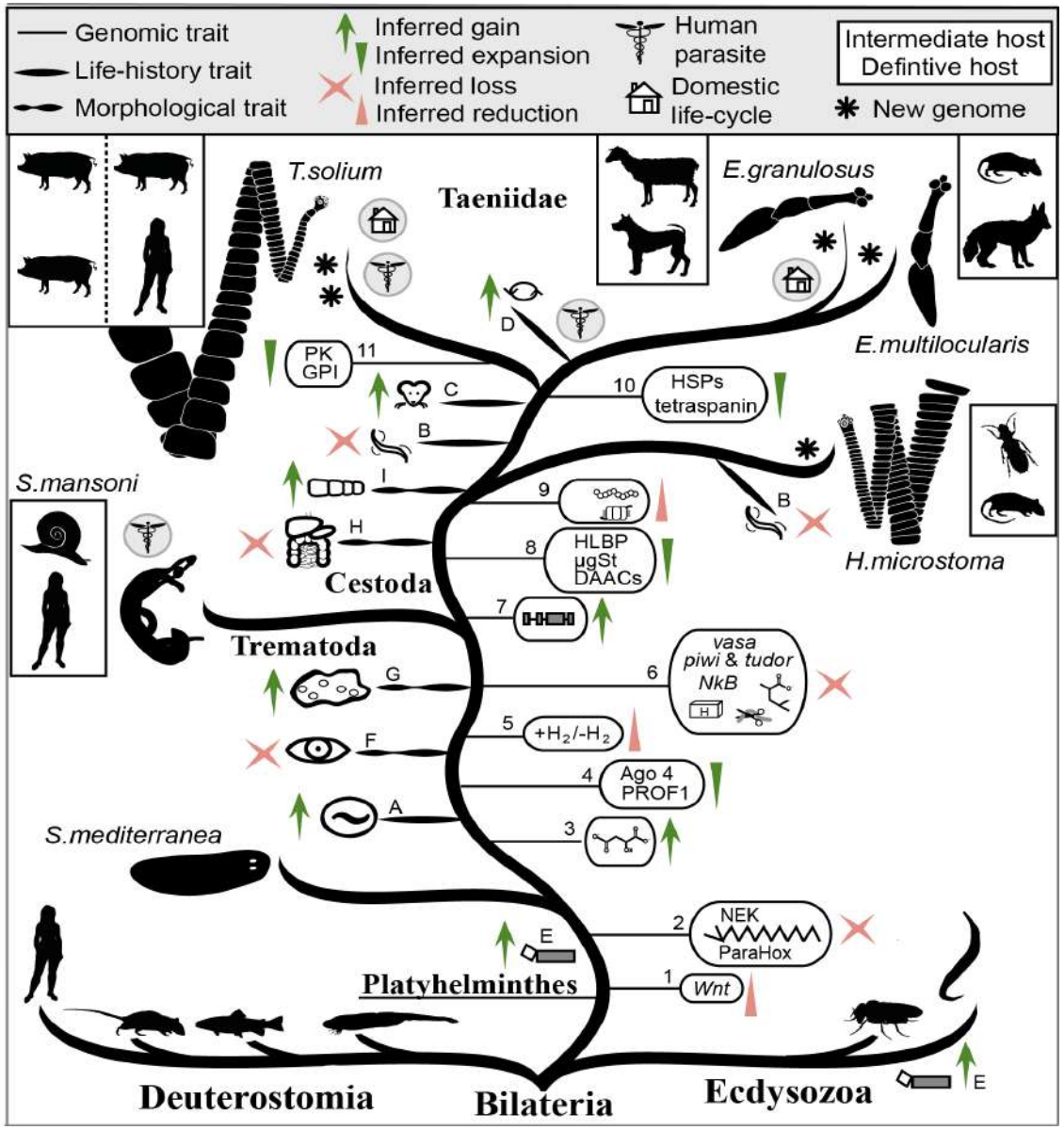
OPEN

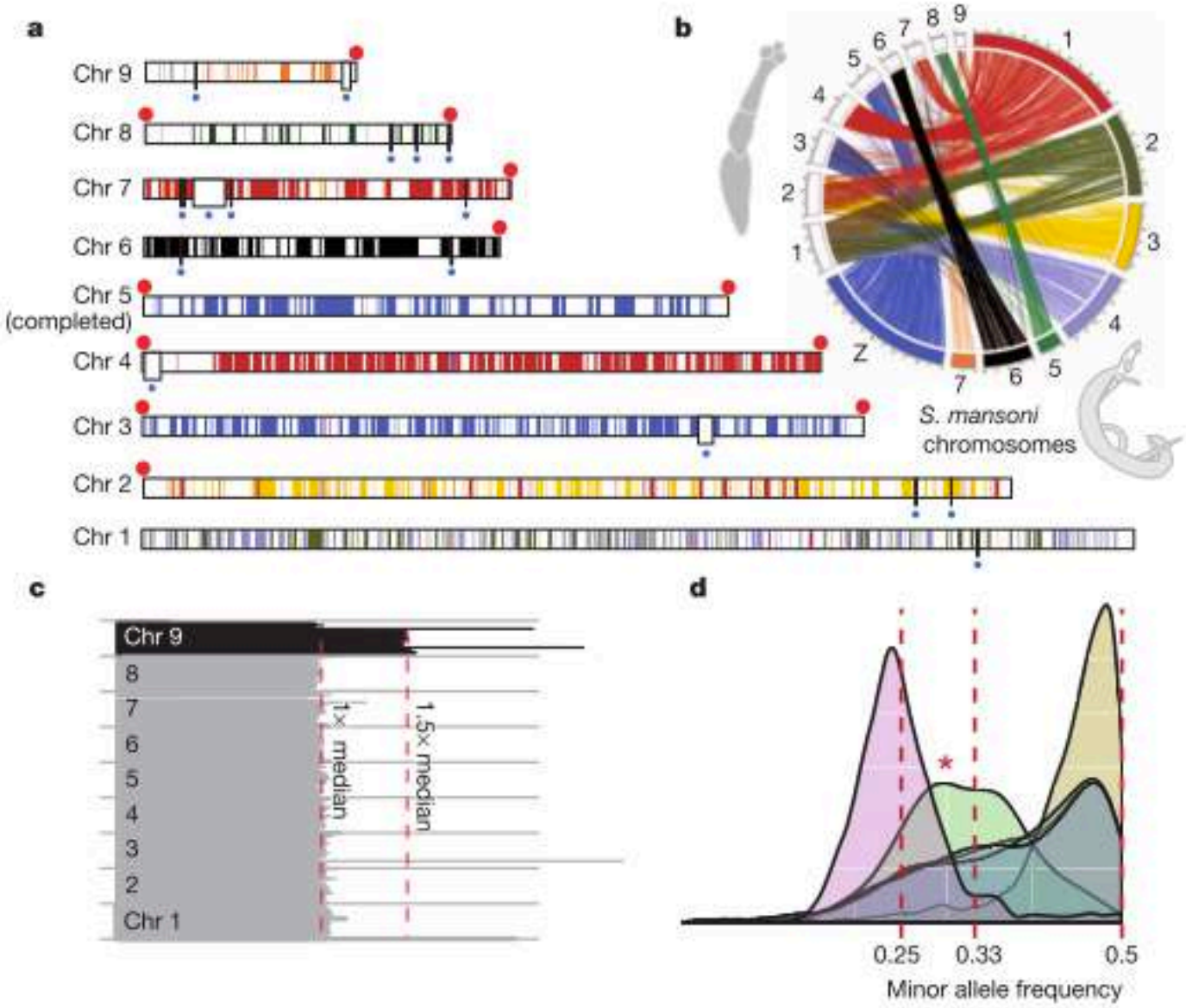
doi:10.1038/nature12031

The genomes of four tapeworm species reveal adaptations to parasitism

Isheng J. Tsai[1,2]*, Magdalena Zarowiecki[1]*, Nancy Holroyd[1]*, Alejandro Garciarrubio[3]*, Alejandro Sanchez-Flores[1,3], Karen L. Brooks[1], Alan Tracey[1], Raúl J. Bobes[4], Gladis Fragoso[4], Edda Sciutto[4], Martin Aslett[1], Helen Beasley[1], Hayley M. Bennett[1], Jianping Cai[5], Federico Camicia[6], Richard Clark[1], Marcela Cucher[6], Nishadi De Silva[1], Tim A. Day[7], Peter Deplazes[8], Karel Estrada[3], Cecilia Fernández[9], Peter W. H. Holland[10], Junling Hou[5], Songnian Hu[11], Thomas Huckvale[1], Stacy S. Hung[12], Laura Kamenetzky[6], Jacqueline A. Keane[1], Ferenc Kiss[13], Uriel Koziol[13], Olivia Lambert[1], Kan Liu[11], Xuenong Luo[5], Yingfeng Luo[11], Natalia Macchiaroli[6], Sarah Nichol[1], Jordi Paps[10], John Parkinson[12], Natasha Pouchkina-Stantcheva[14], Nick Riddiford[14,15], Mara Rosenzvit[6], Gustavo Salinas[9], James D. Wasmuth[16], Mostafa Zamanian[17], Yadong Zheng[5], The *Taenia solium* Genome Consortium†, Xuepeng Cai[5], Xavier Soberón[3,18], Peter D. Olson[14], Juan P. Laclette[4], Klaus Brehm[13] & Matthew Berriman[1]

2013 – Tapeworm genome project





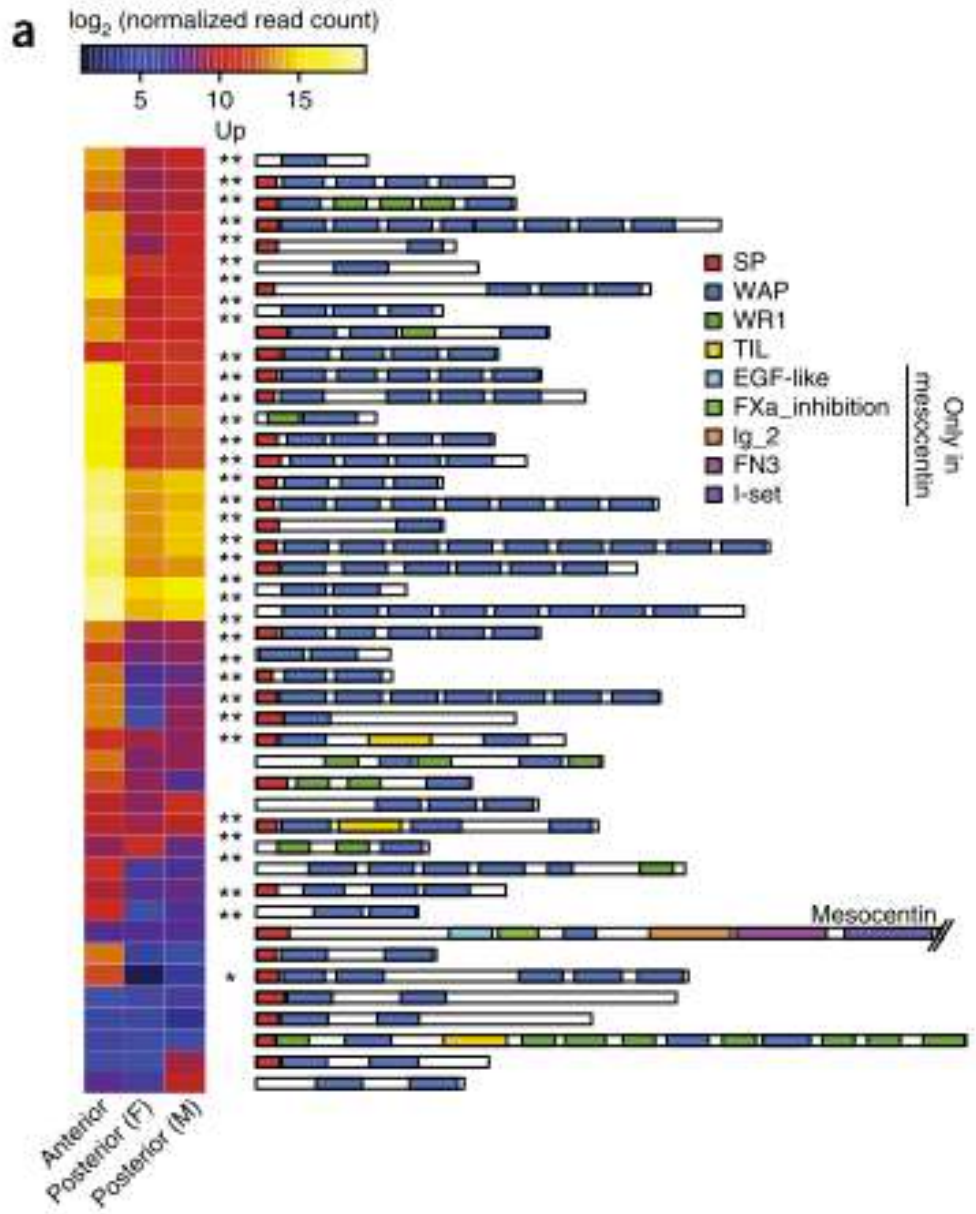
2014 – *Trichuris* genome project

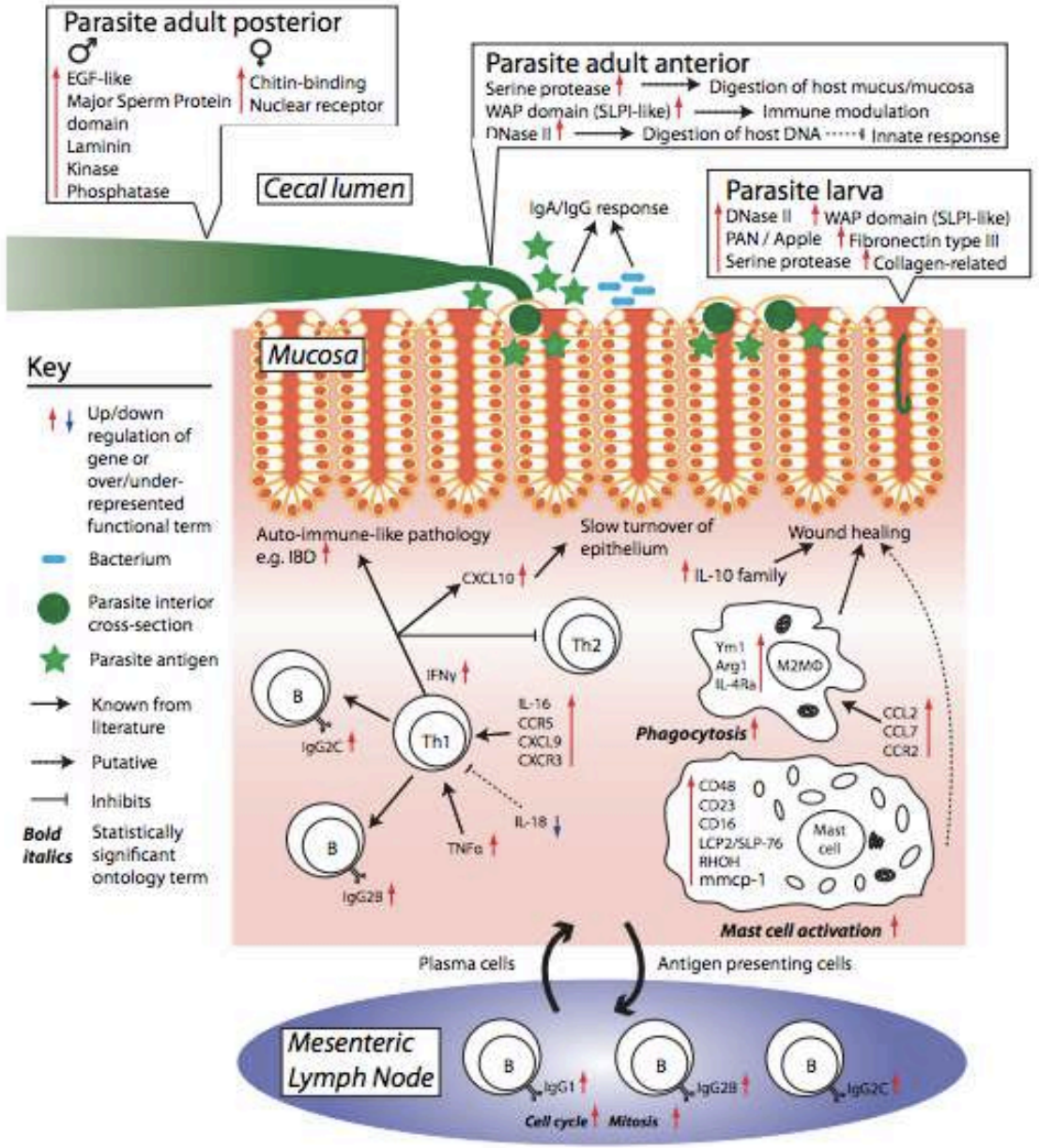
- 2 genomes probably costs less than £10,000k
- About **40 RNAseq** libraries of different life cycle stages, host infecting stages
- Paradigm shifts to RNAseq

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

Bernardo J Foth[1,7], Isheng J Tsai[1,2,7], Adam J Reid[1,7], Allison J Bancroft[3,7], Sarah Nichol[1], Alan Tracey[1], Nancy Holroyd[1], James A Cotton[1], Eleanor J Stanley[1], Magdalena Zarowiecki[1], Jimmy Z Liu[4], Thomas Huckvale[1], Philip J Cooper[5,6], Richard K Grencis[3] & Matthew Berriman[1]

2014 – *Trichuris* genome project





2014 – *Taphrina* genome project

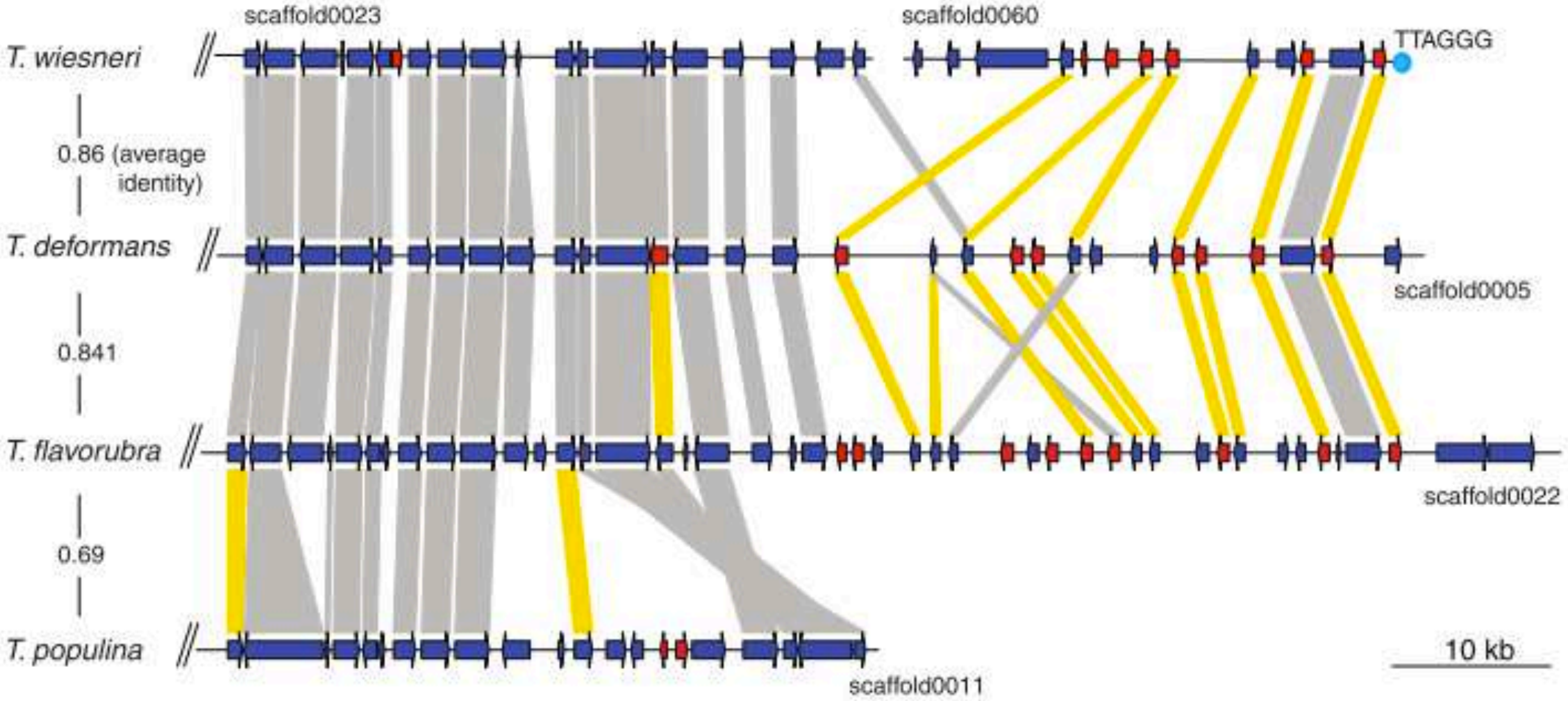
- 3 fungi genomes (~18Mb) of Illumina PE
- RNAseq for annotation purpose
- Costs probably less than 200,000 NT
- 2 months to analyse

GBE

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

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

2014 – *Taphrina* genome project



2017 – *Phellinus* genome project

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

Received: 3 July 2017 | Revised: 8 September 2017 | Accepted: 11 September 2017

DOI: 10.1111/mec.14359

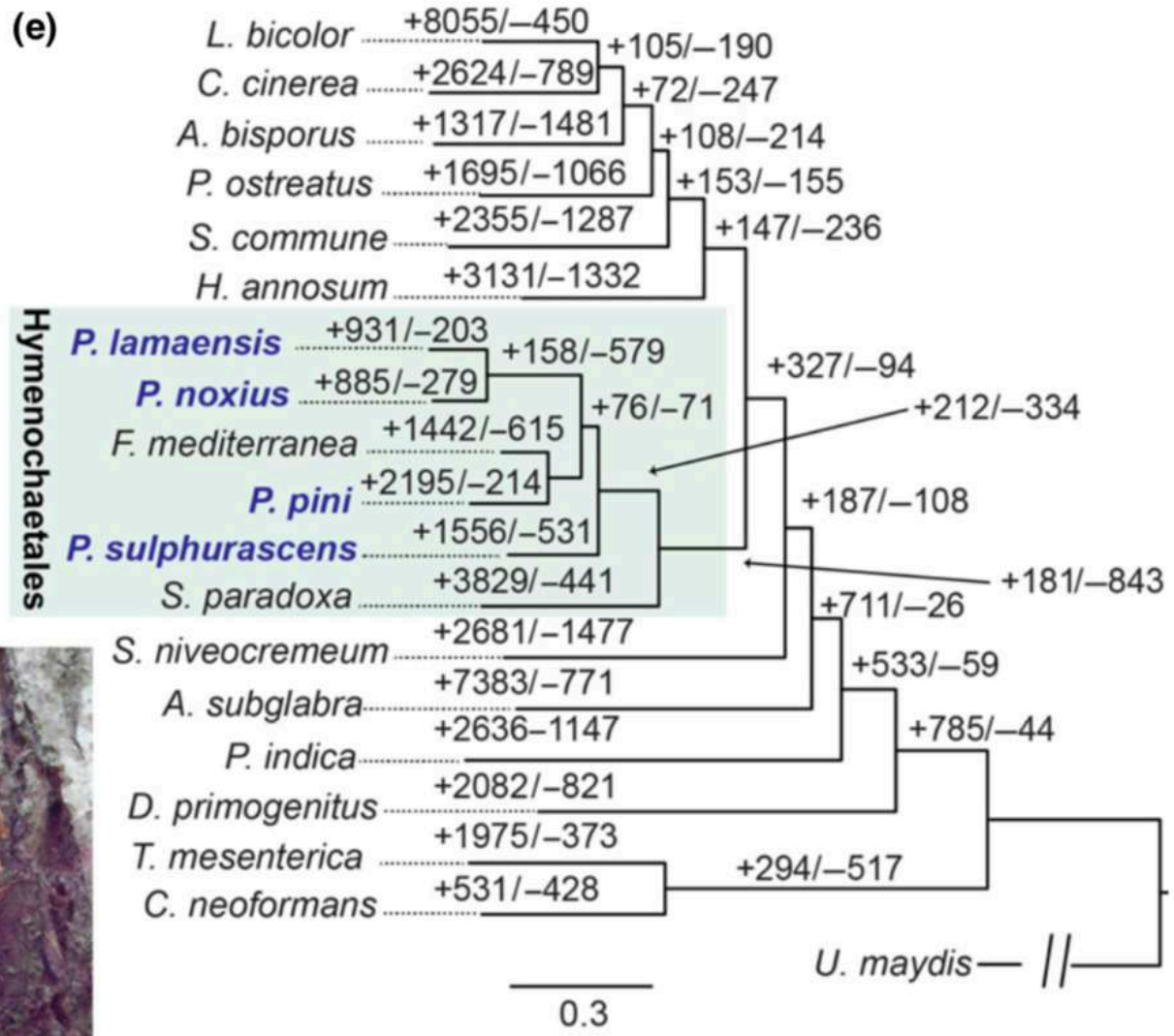
ORIGINAL ARTICLE

WILEY MOLECULAR ECOLOGY

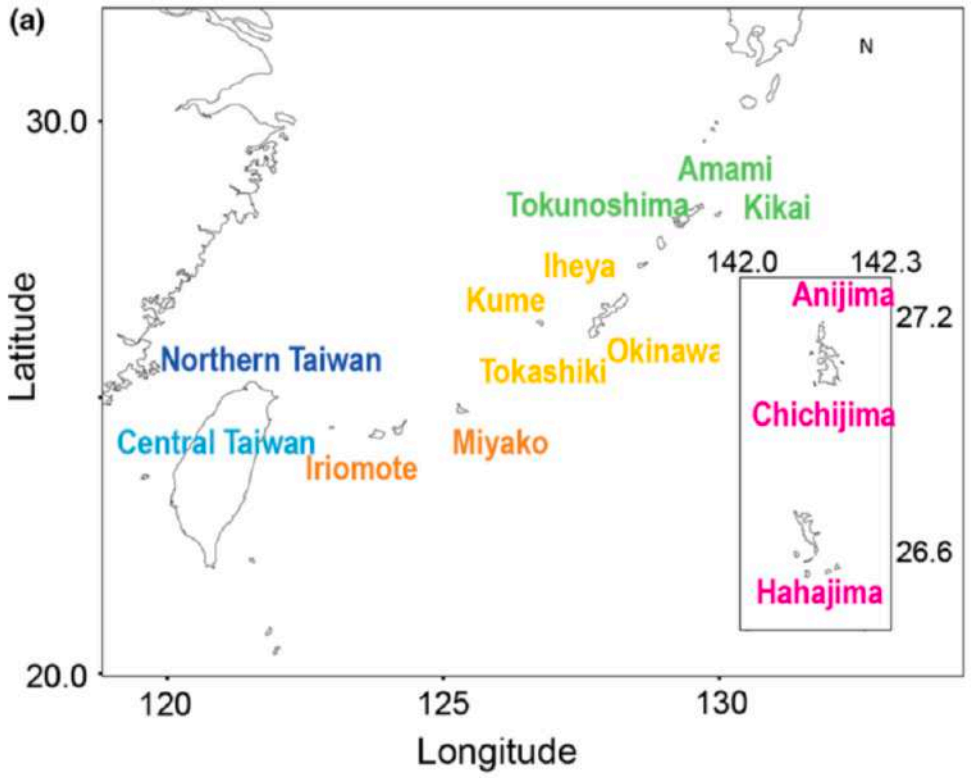
Comparative and population genomic landscape of *Phellinus noxius*: A hypervariable fungus causing root rot in trees

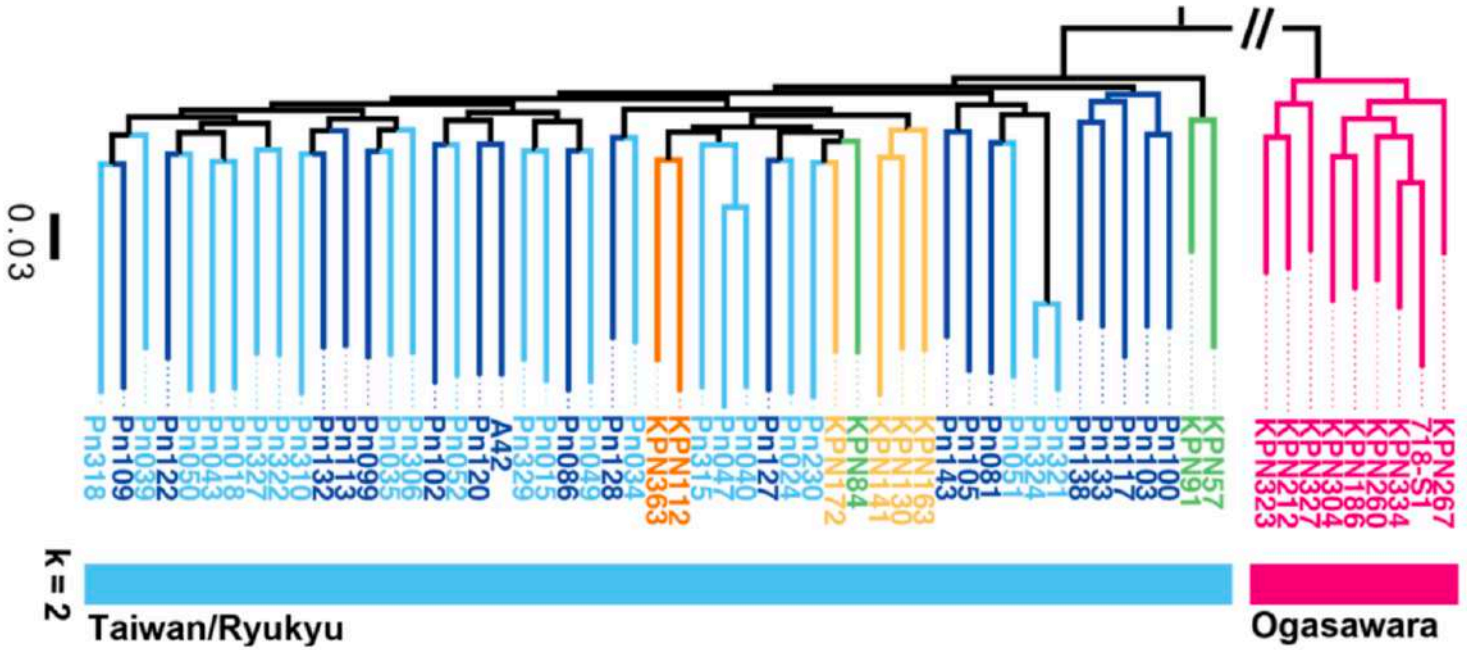
Chia-Lin Chung[1,2*] | Tracy J. Lee[3,4,5] | Mitsuteru Akiba[6] | Hsin-Han Lee[1] | Tzu-Hao Kuo[3] | Dang Liu[3,7] | Huei-Mien Ke[3] | Toshiro Yokoi[6] | Marylette B. Roa[3,8] | Mei-Yeh J. Lu[3] | Ya-Yun Chang[1] | Pao-Jen Ann[9] | Jyh-Nong Tsai[9] | Chien-Yu Chen[10] | Shean-Shong Tzean[1] | Yuko Ota[6,11]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

2017 – *Phellinus* genome project



2017 – *Phellinus* genome project





[In comparison]

Article

Cell

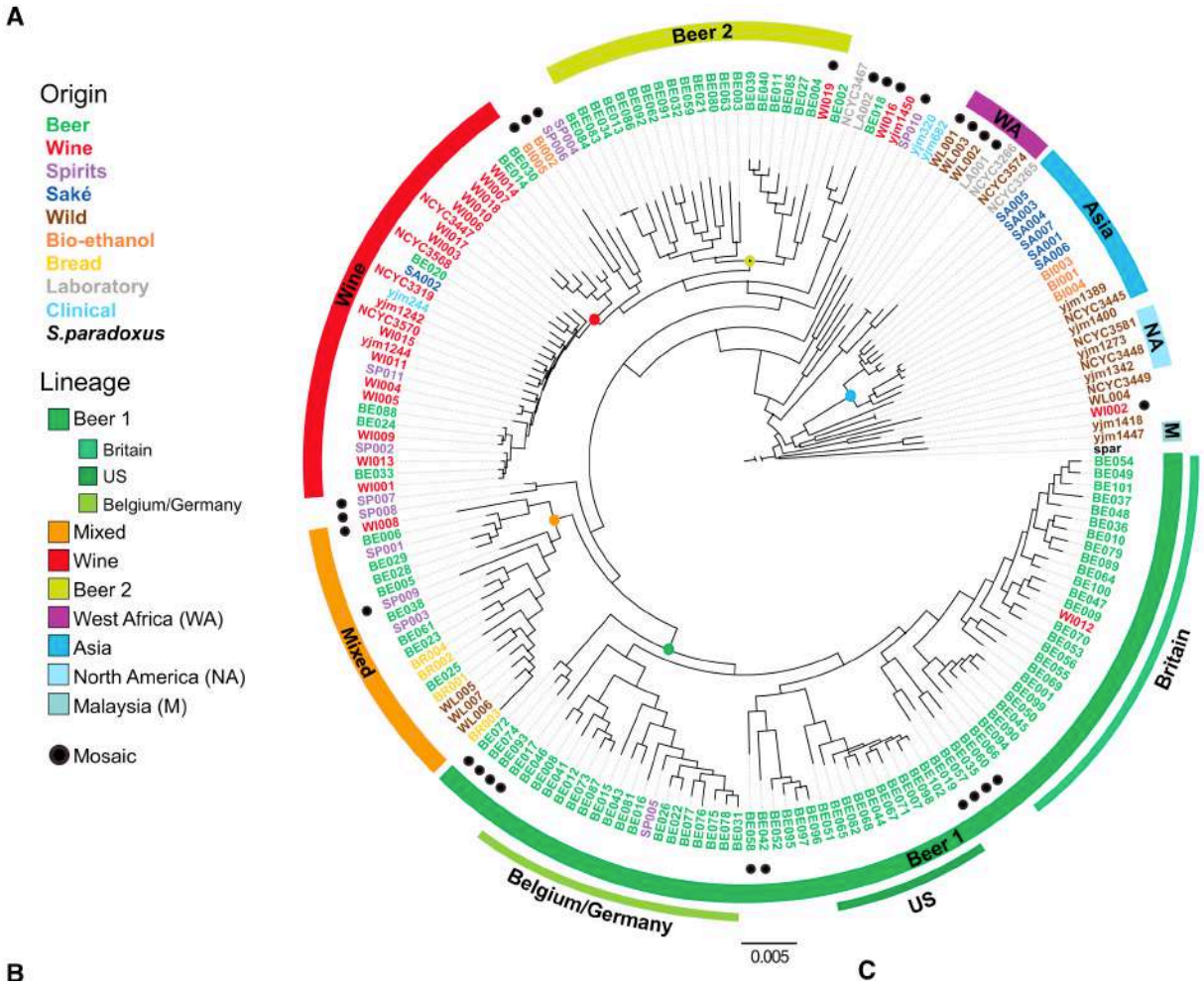
Domestication and Divergence of *Saccharomyces cerevisiae* Beer Yeasts

••Sequenced and phenotyped **157** *S. cerevisiae* yeasts

••Present-day industrial yeasts originate from only a few domesticated ancestors

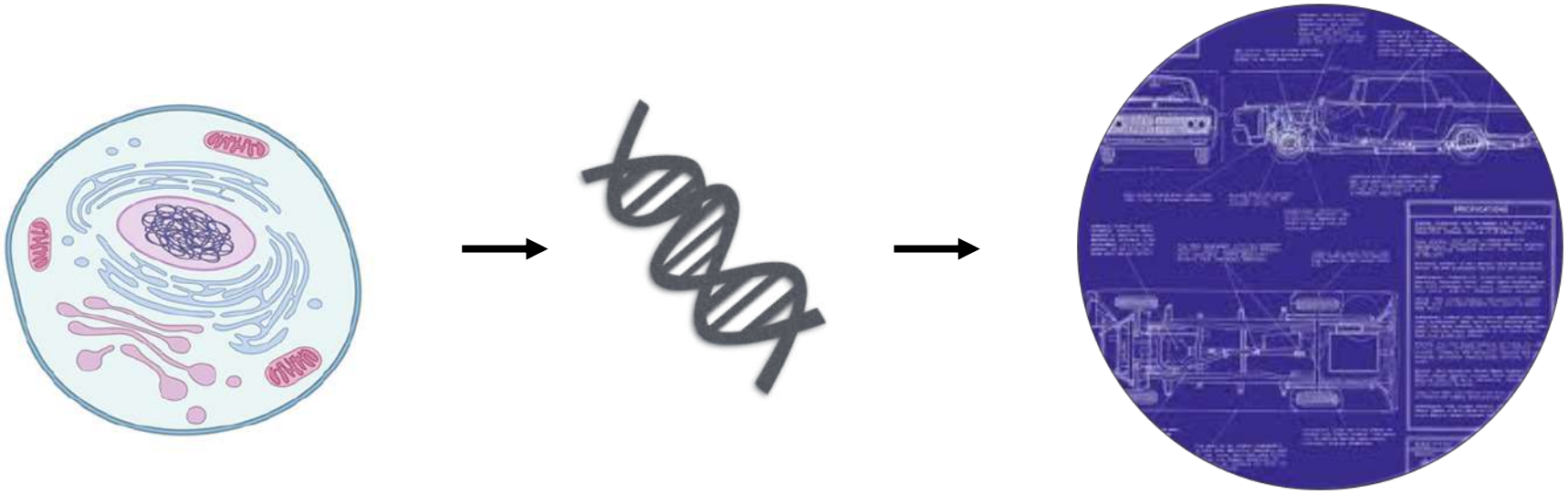
••Beer yeasts show strong genetic and phenotypic hallmarks of domestication

••Domestication of industrial yeasts predates microbe discovery



Break

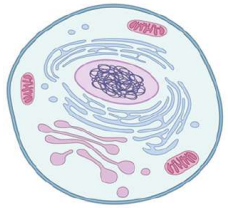
Genome



Genome = Parts list of a single genome

A genome project

Wet lab work



Extract DNA
(or RNA)

Shear DNA
(or RNA)

Sequence

Bioinformatics

DNA or RNA Reads
50-500 bp

Mapping
RNAseq

Annotation

ATCG
ATGG
ATCG

Variant calling

Data QC

Reads
50-500 bp

Assembly

Contigs
1kb – 100 kbp

Scaffolding

Scaffolds
Hopefully Mbp

Three situations you are most likely to encounter

Genome reference is available (for example, humans):

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

Genome reference is NOT available

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

Counting:

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

Project examples

- Sequencing a species (Comparative genomics)
  - Map, assemble

- Sequencing multiple individuals of a species (Population genomics)
  - Map, count

- Combination of (1) and (2)

- **Further analysis**
  # Integrating multiple data types
  # Correlation
  # Dimension reduction

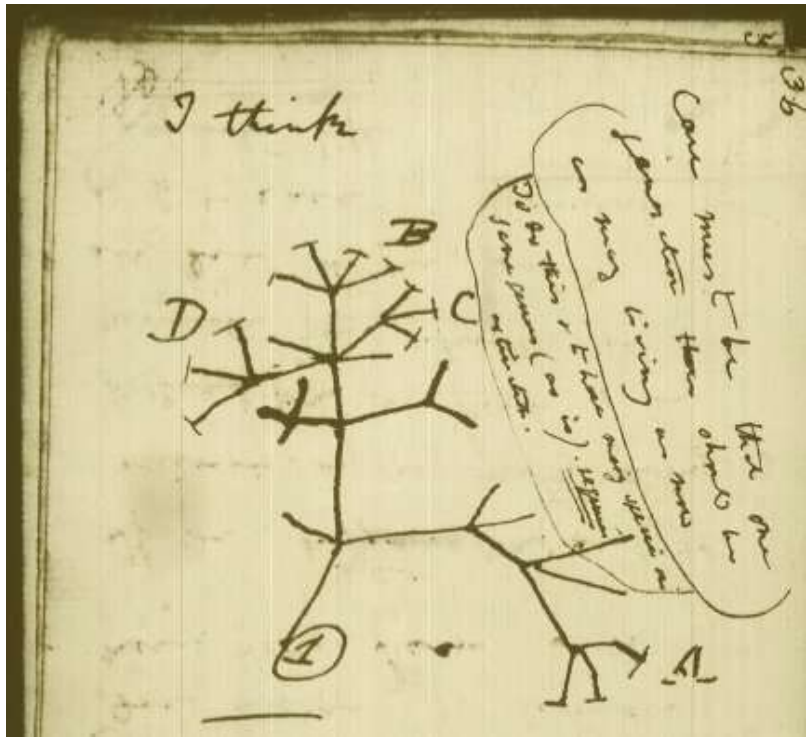
Why sequence a genome?

- 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**

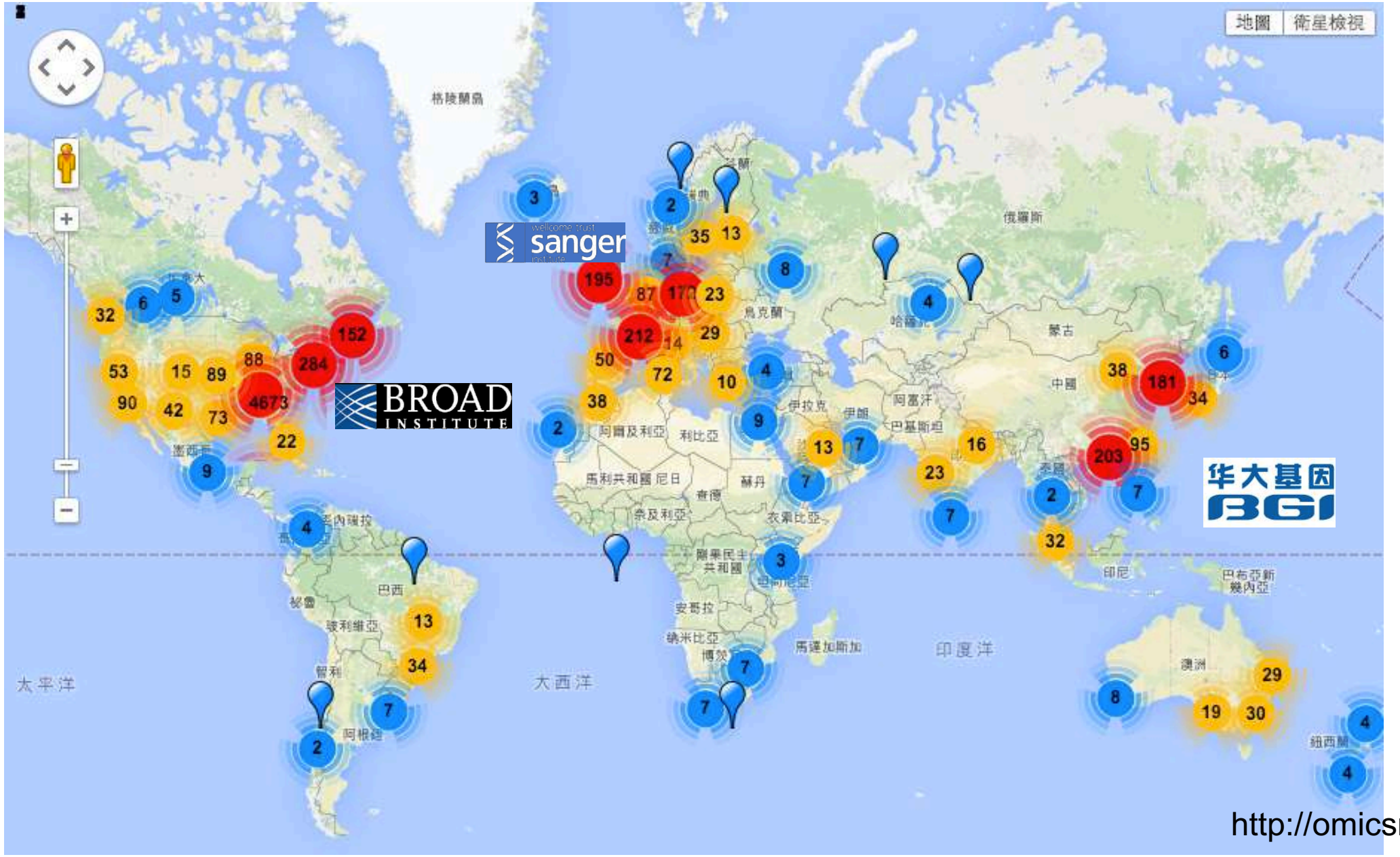
Nothing makes sense in the light of evolution

Theodosius Dobzhansky 1973



Different sequencing platforms / History of sequencing

World competing for sequencing power

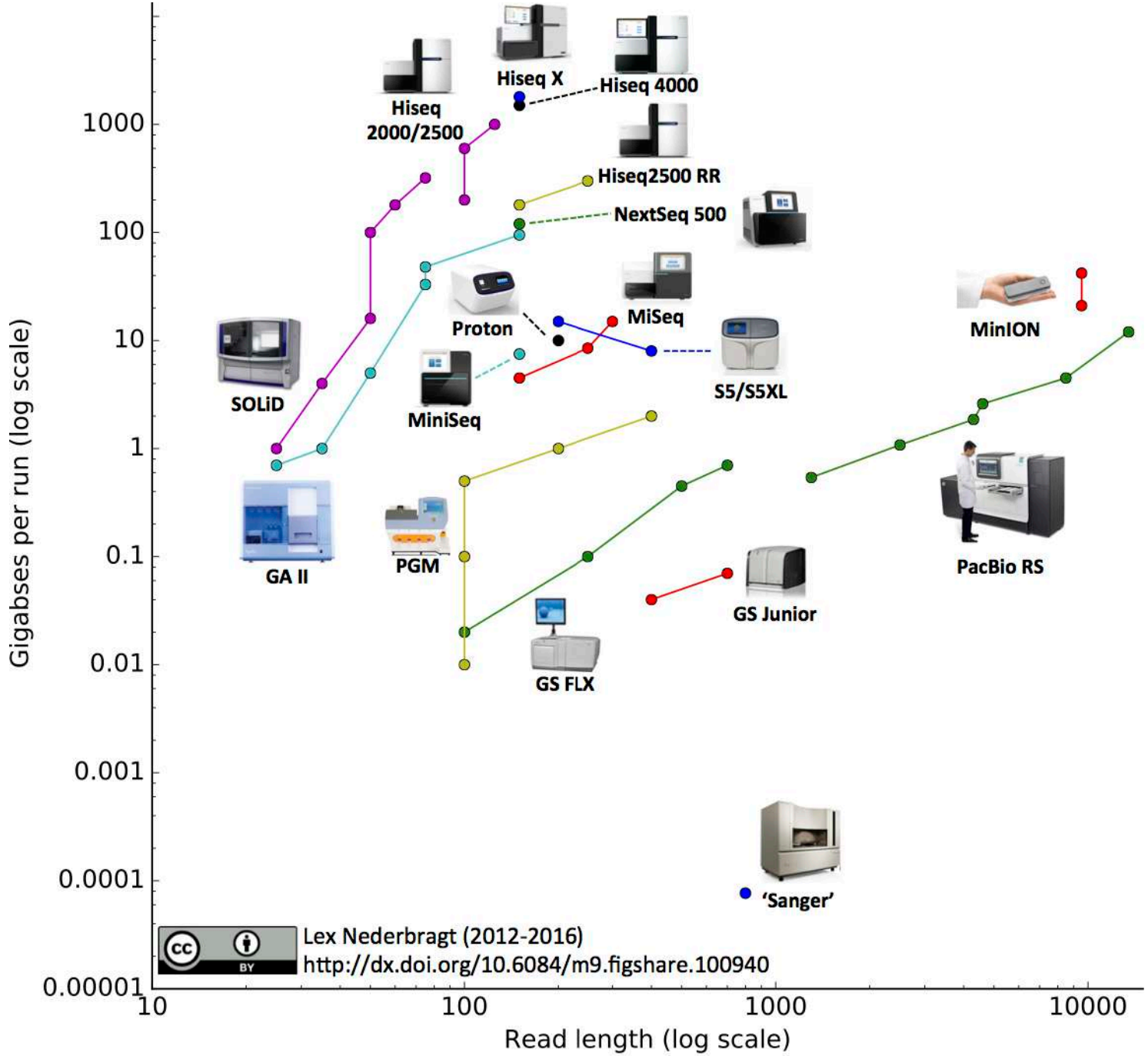


http://omicsmaps.com/

Sequencing Platforms

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

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


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

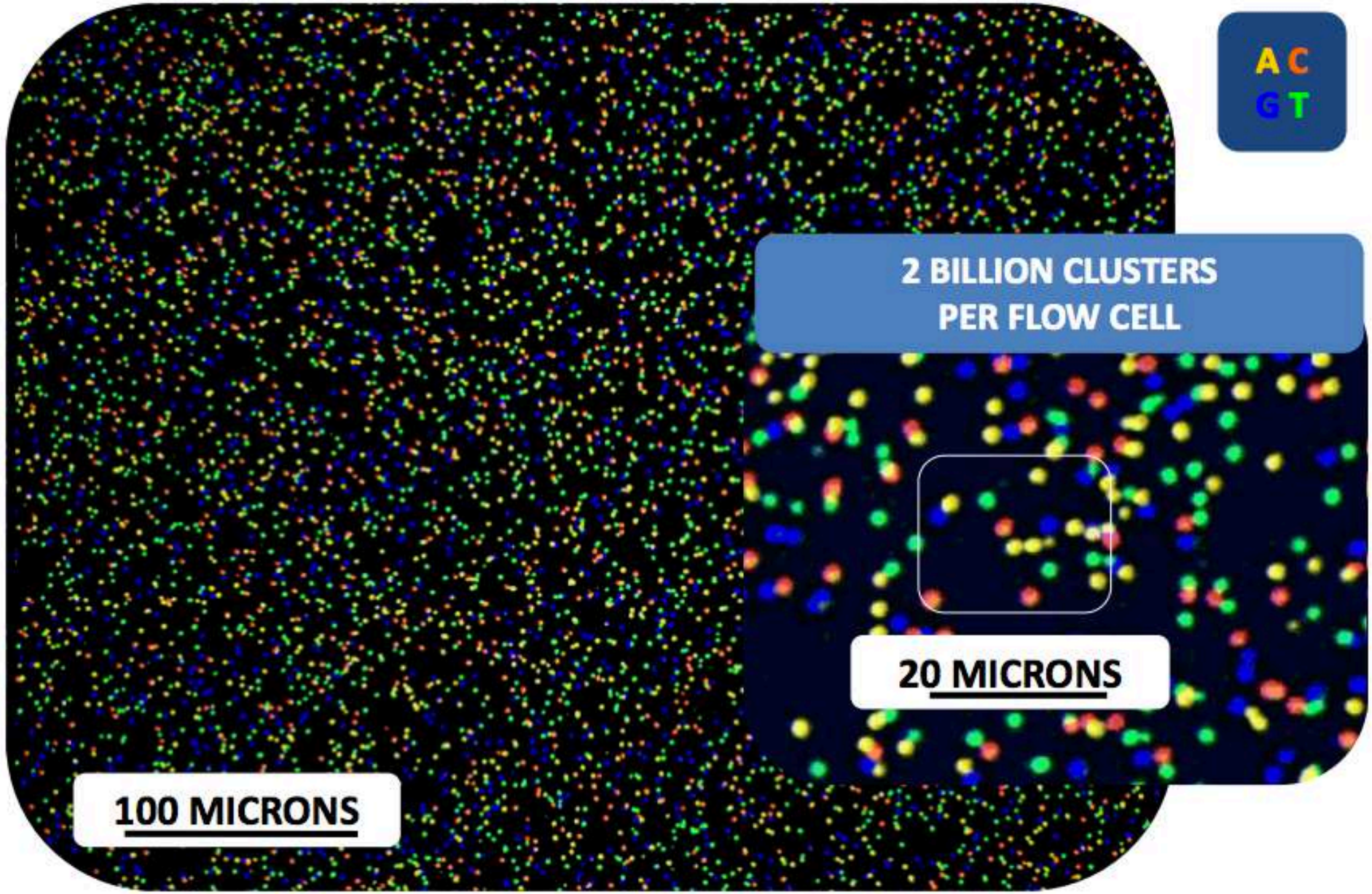
Platform	Reads x run: (M)	Read length: (paired-end*, Half of data in reads**)	Run time: (d)	Yield: (Gb)	Rate: (Gb/d)	Reagents: ($K)	per-Gb: ($)	hg-30x: ($)	Machine: ($)
iSeq 100 1fcell	4	150*	0.77	1.2	1.56	0.625	521	62500	19.9K
MiniSeq 1fcell	25	150*	1	7.5	7.5	1.75	233	28000	49.5K
MiSeq 1fcell	25	300*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

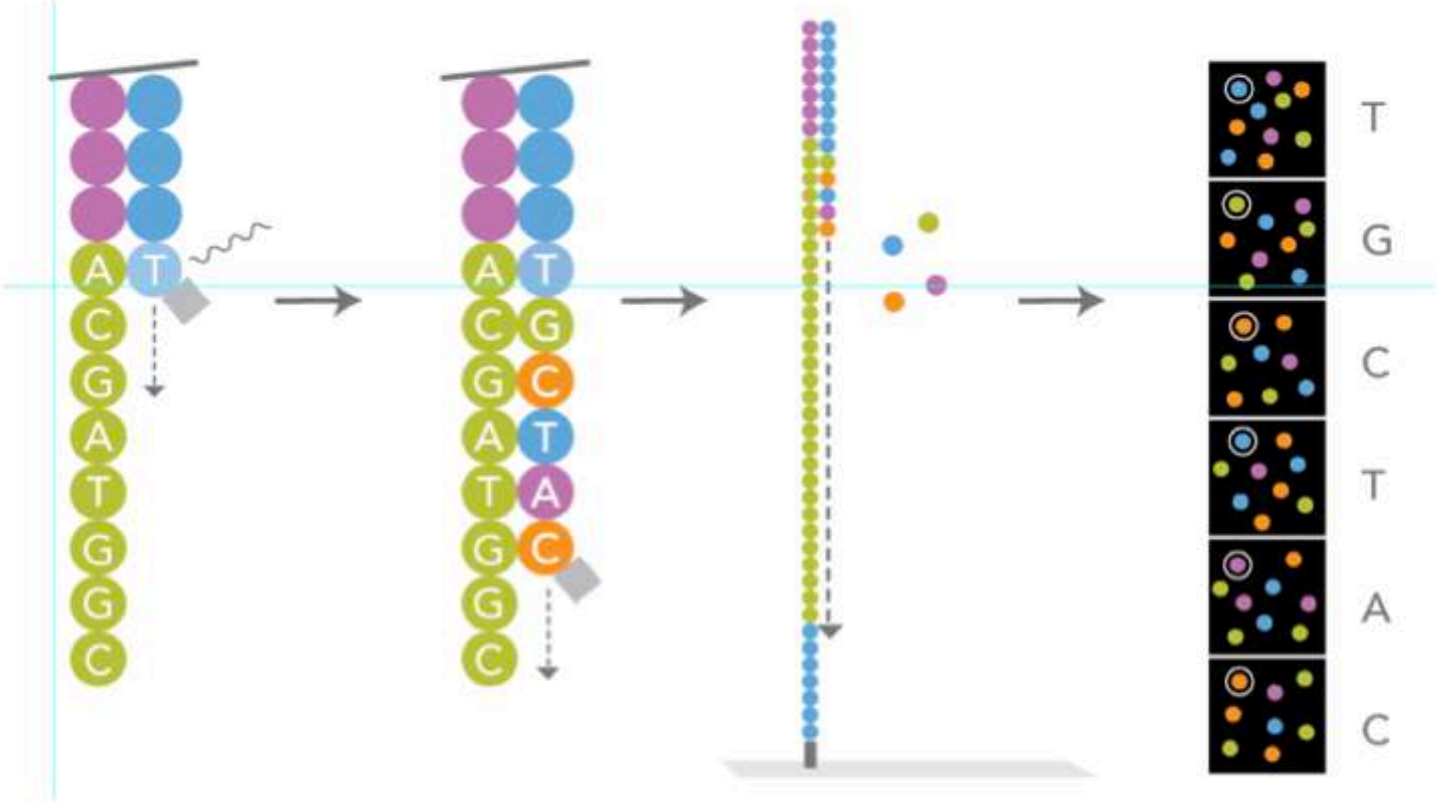
http://twitter.com/albertvilella ; Last updated: 2018.02.20

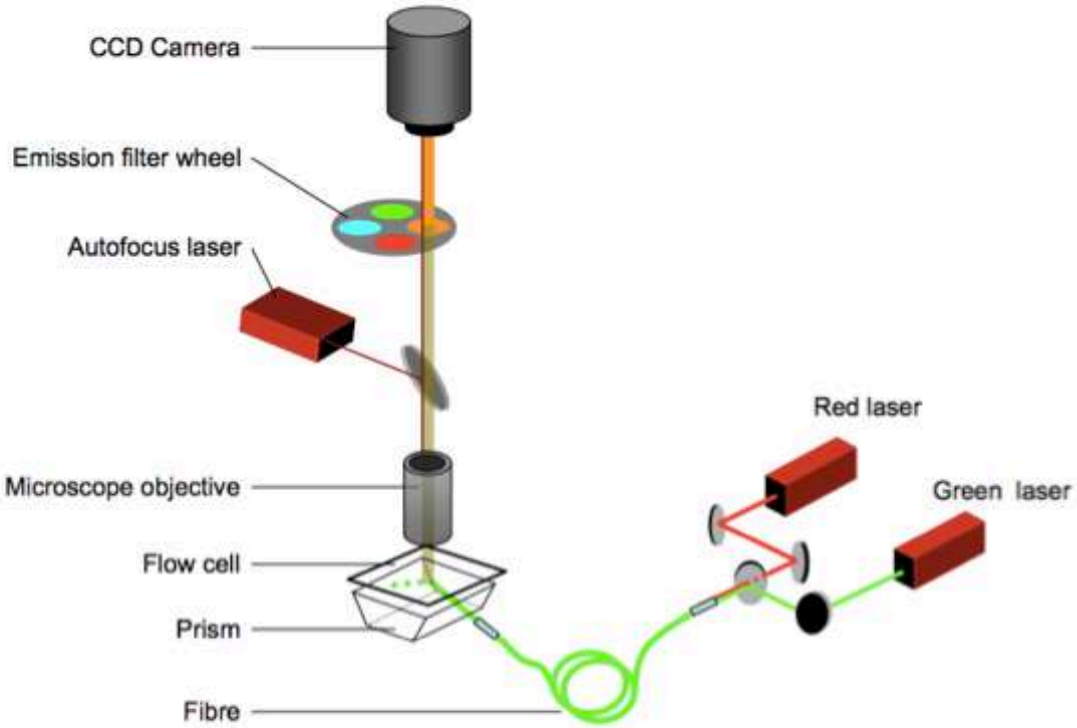
Illumina NovaSeq

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

PacBio (Pacific Biosciences)

Sequel

System Performance

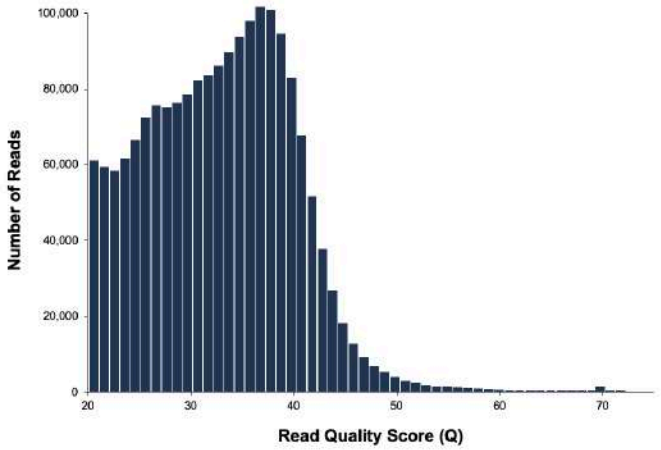
Example data from genomic libraries generated using the continuous long read (CLR) and HiFi read modes of sequencing on the Sequel II System.

Highly Accurate Long Reads

HiFi Sequencing

Number of >99% (Q20) 9-13 kb Reads:

Up to 2 million



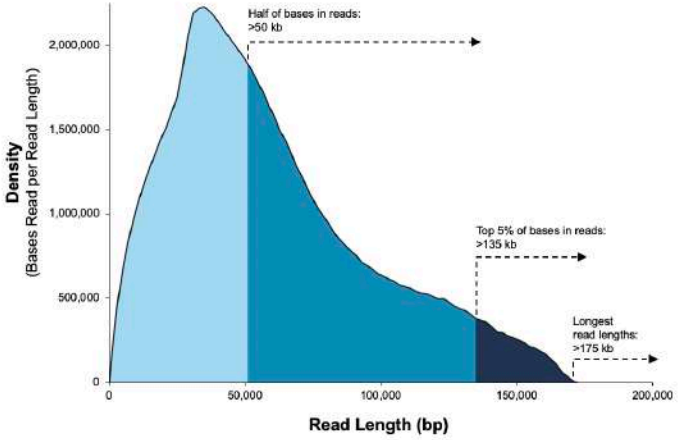
Data from a 11 kb size-selected human library using the SMRTbell Template Prep Kit 1.0 on a Sequel II System (1.0 Chemistry, Sequel II System Software v7.0, 30-hour movie)*.

Long Read Lengths

CLR Sequencing

Half the Data in Reads: >50 kb

Data per SMRT Cell: Up to 160 Gb



Data from a 35 kb size-selected *E. coli* library using the SMRTbell Express Template Prep Kit 2.0 on a Sequel II System (1.0 Chemistry, Sequel II System Software v7.0, 15-hour movie)*.

https://www.pacb.com/products-and-services/sequel-system/latest-system-release/

Oxford Nanopore

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

Oxford Nanopore – how it works

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

Read length go beyond

John Tyson @DrT1973 · Aug 21

Achievement unlocked ..... half the data in 100Kb+ (Ultra) reads. New Rapid-Bead Free Ultra-Library prep for @nanopore DNA sequencing 😄. Coming your way via @longreadclub

https://twitter.com/longreadclub

Come and go of technologies

Programming languages

Perl

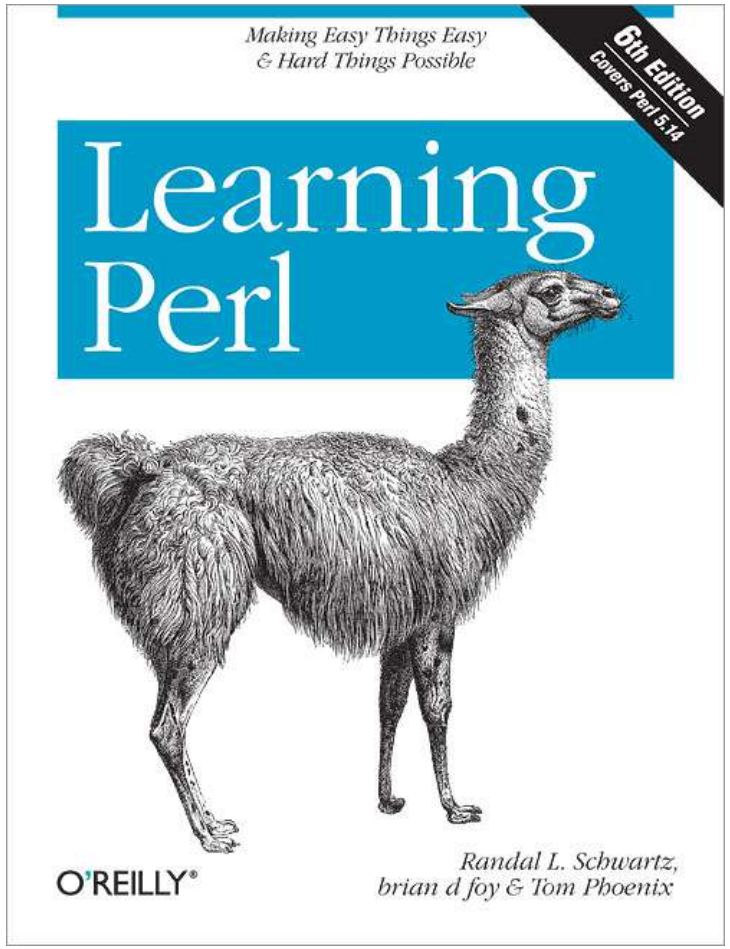


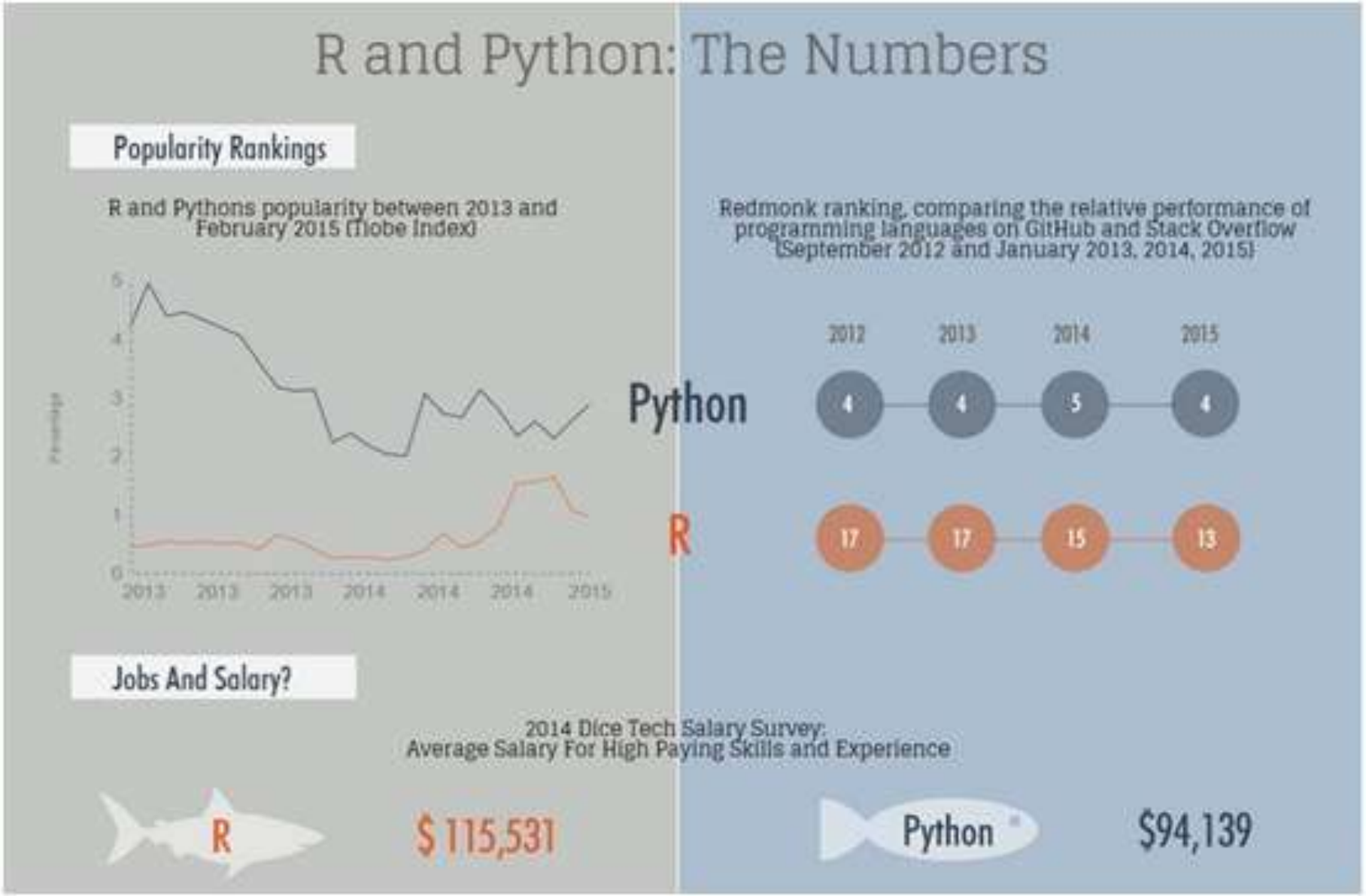
Table 2. Selected early bioinformatics software written in Perl

Software	Year released	Use	Reference
GeneQuiz	1994 (oldest)	Workbench for protein sequence analysis	[65]
LabBase	1998	Making relational databases of sequence data	[66]
Phred-Phrap-Consed	1998	Genome assembly and finishing	[67]
Swissknife	1999	Parsing of SWISS-PROT data	[68]
MUMmer	1999	Whole genome alignment	[69]

PubMed Key: (perl bioinformatics) AND ("1987"[Date-Publication]:"2000"[Date-Publication]).

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

Python and R


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

Data type

EARTH BIOGENOME PROJECT

A GRAND CHALLENGE

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

A GRAND VISION

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

A lot of data

- We biologists generate a lot of data
  - Experiments, sequencing
  - Everything is more high throughput, but not necessarily less noisy

- Different data types
  - Images, Sequences, Signals, Locations, Linkage, Frequencies…

- How do we
  - analyse them?
  - store them?
  - publish them?
  - reuse them?

A small project's typical output

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

8 exome samples ;
2 Illumina Hiseq lanes with **184GB** of data

~100X of human exome to detect disease causing SNP

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

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

More data but less people with informatics skills

- Sequencing is the result of many types of experiment

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

- **Bottleneck is bioinformatics analysis**

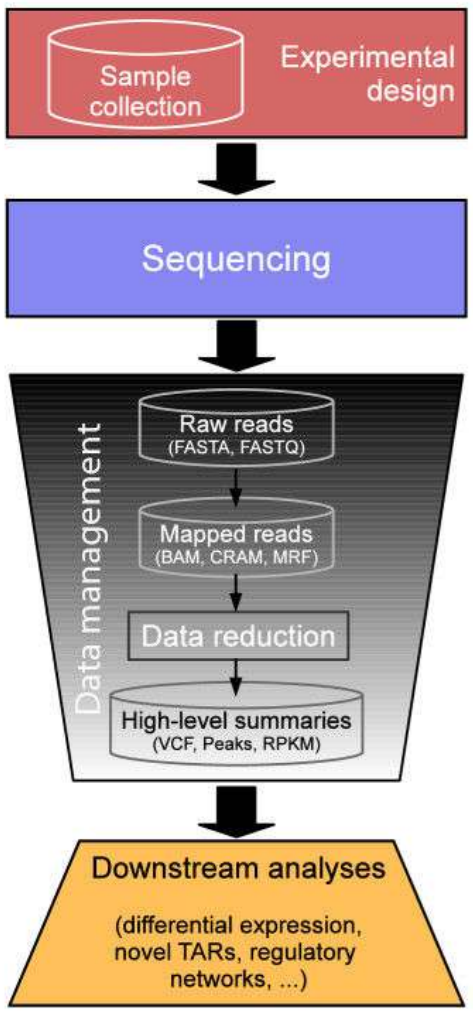
Sboner *et al. Genome Biology* 2011, **12**:125
http://genomebiology.com/2011/12/8/125

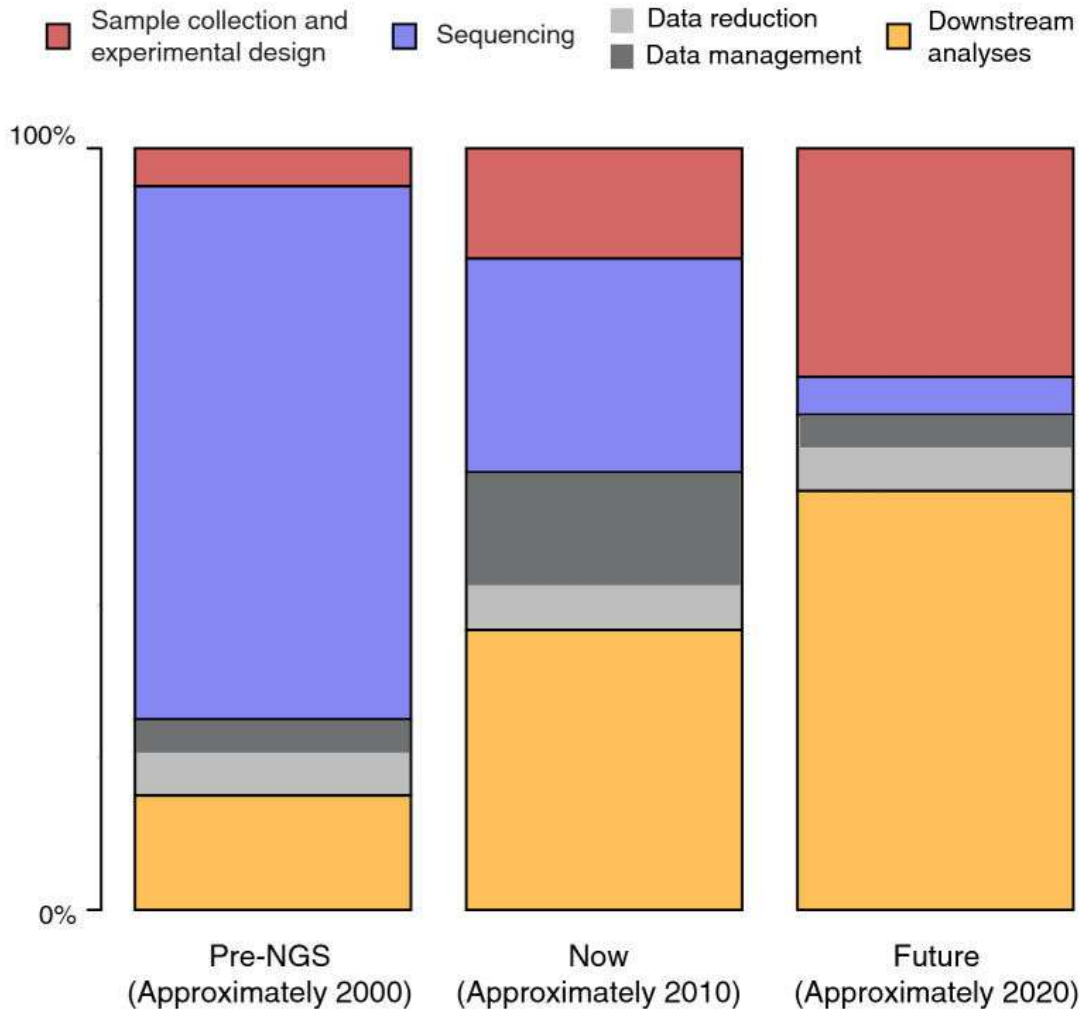
Genome Biology

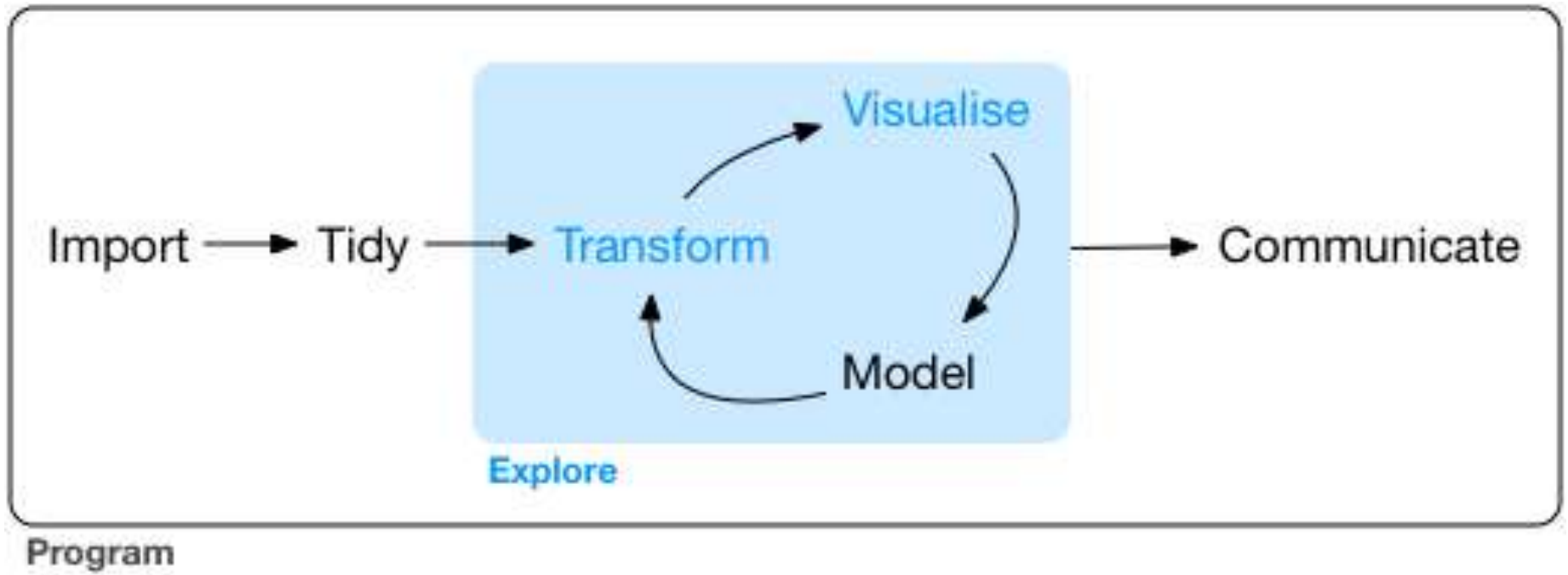
OPINION

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]







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

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

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

Important problems:

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

No ’perfect’ pipeline – learn through experience

Always understand your data / programs

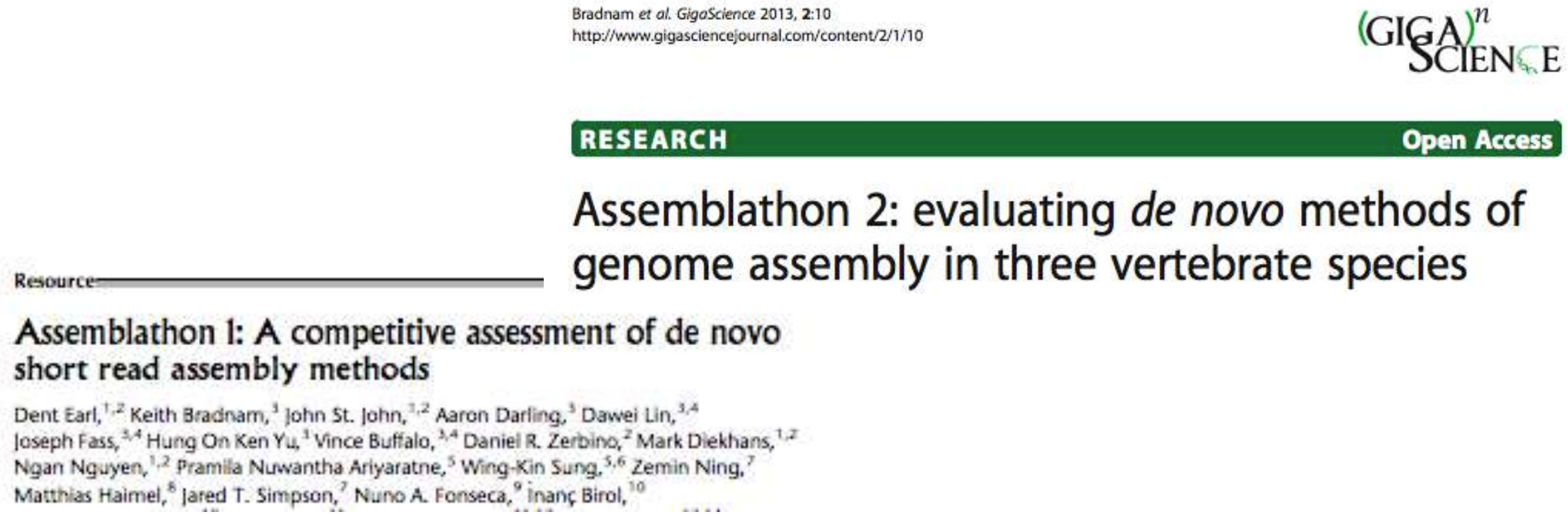
- Understand:
  - Data format
  - The nature of your data

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

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

If unsure – always check **benchmark** studies

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



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

(GIGA)^n SCIENCE

RESEARCH Open Access

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

Resource

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

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

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.**

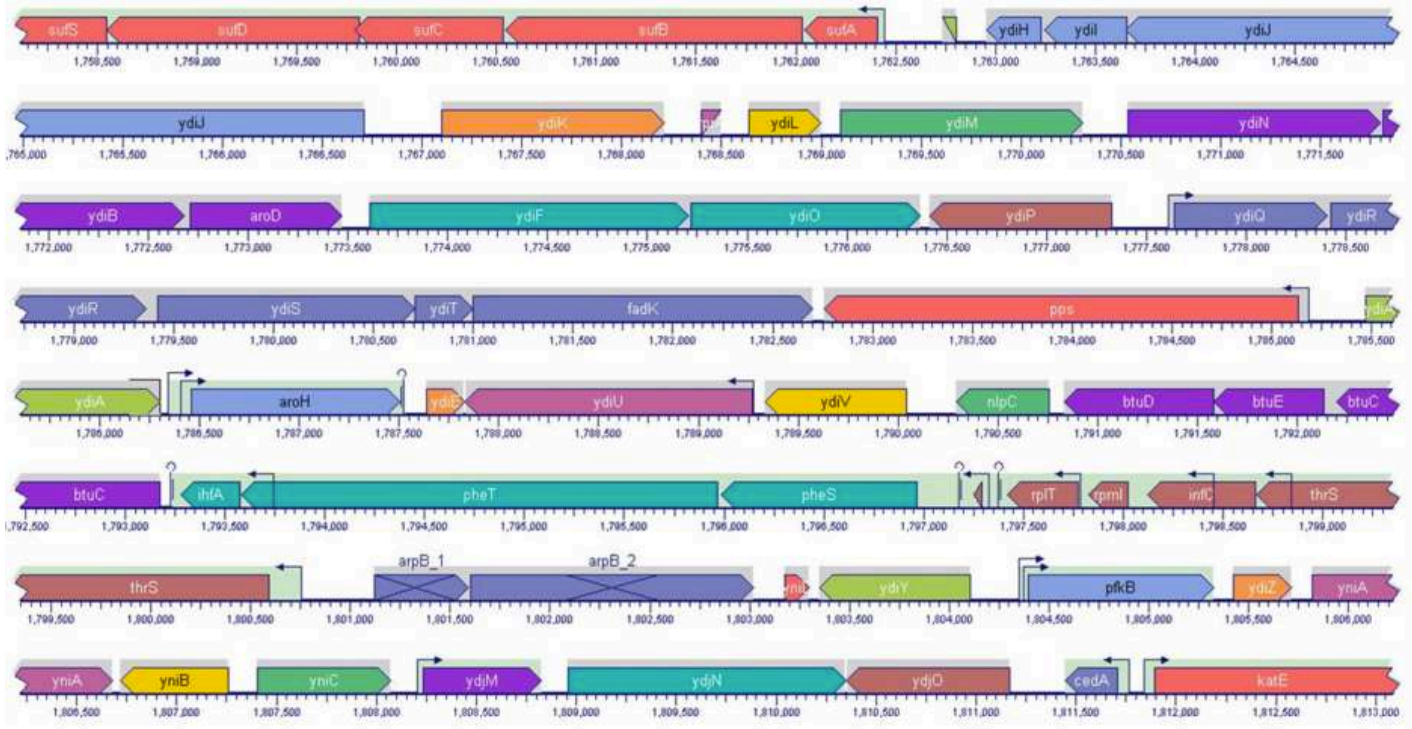
-Wikipedia

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

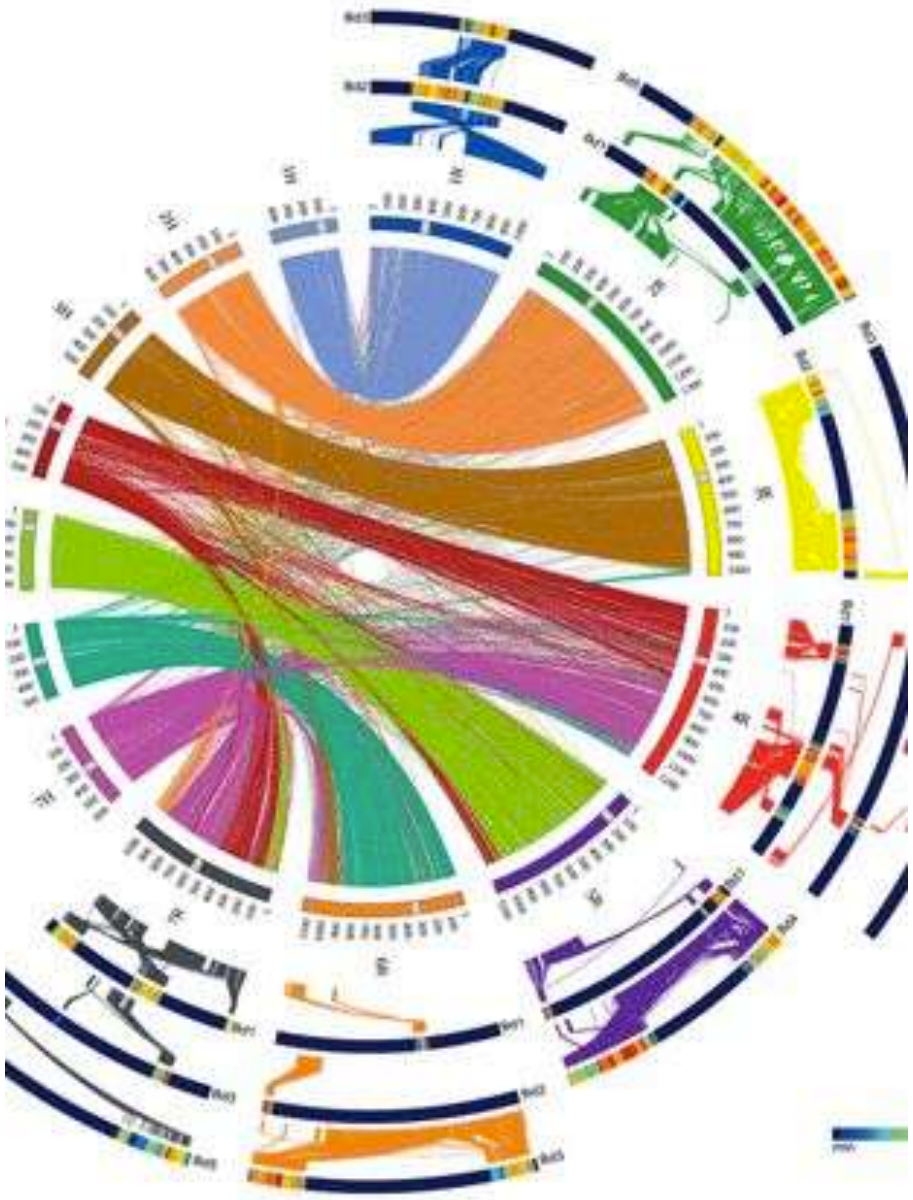
1. **Single line ID** with at symbol (“@”) in the first column.
2. There should be not space between “@” symbol and the first letter of the identifier.
3. Sequences are in multiple lines after the ID line
4. Single line with plus symbol (“+”) in the first column to represent the quality line.
5. Quality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

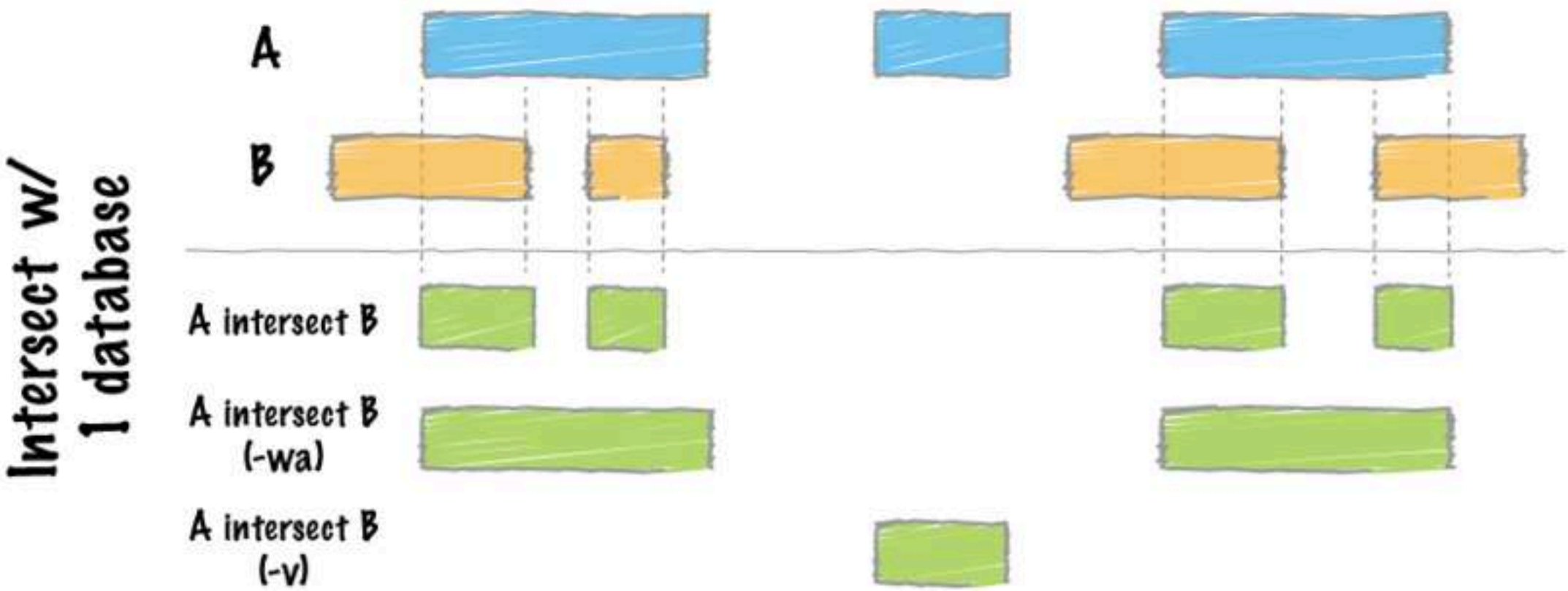
```
chr7    127471196  127472363
chr7    127472363  127473530
chr7    127473530  127474697
chr7    127474697  127475864
chr7    127475864  127477031
chr7    127477031  127478198
chr7    127478198  127479365
chr7    127479365  127480532
chr7    127480532  127481699
```

```
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

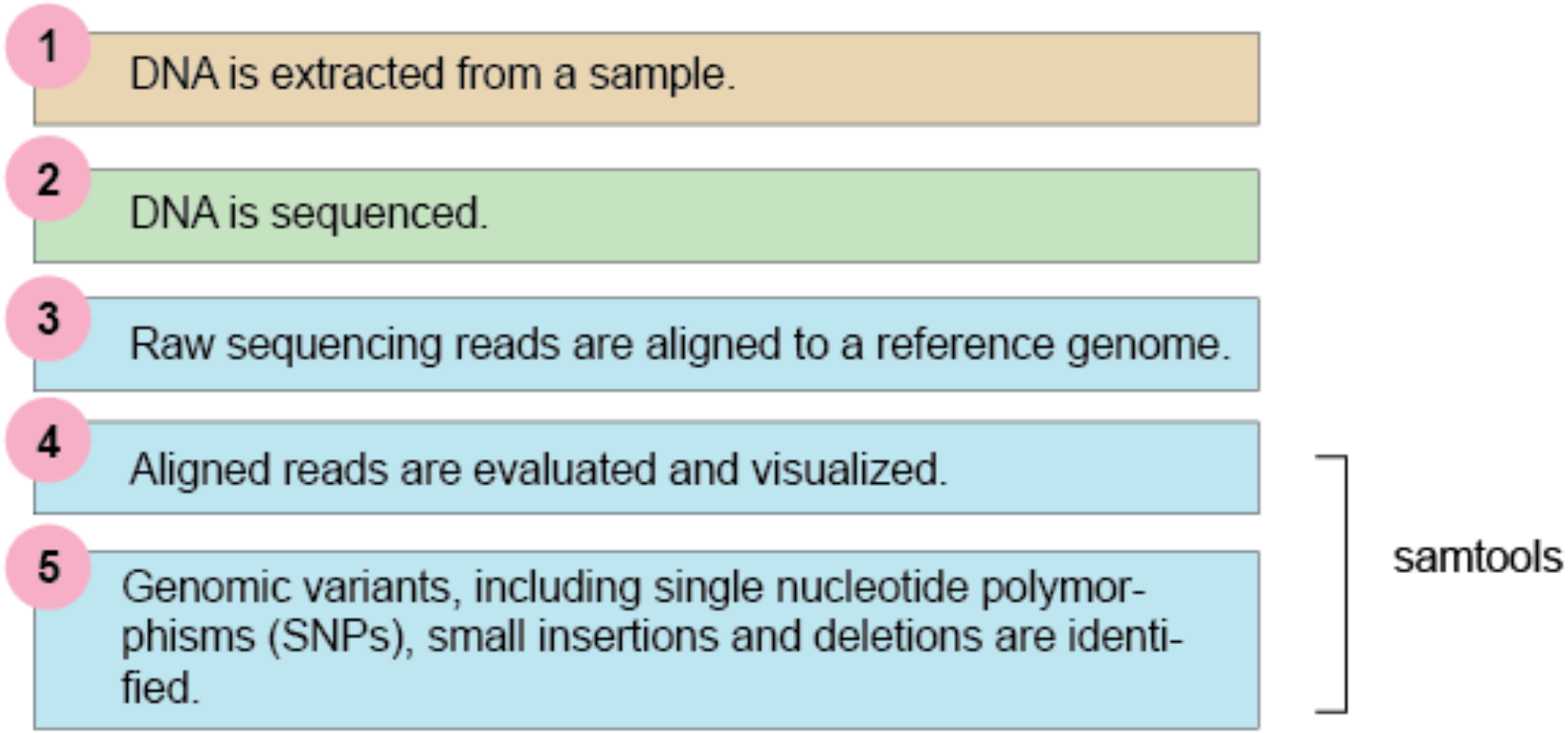
http://gmod.org/wiki/GFF2

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

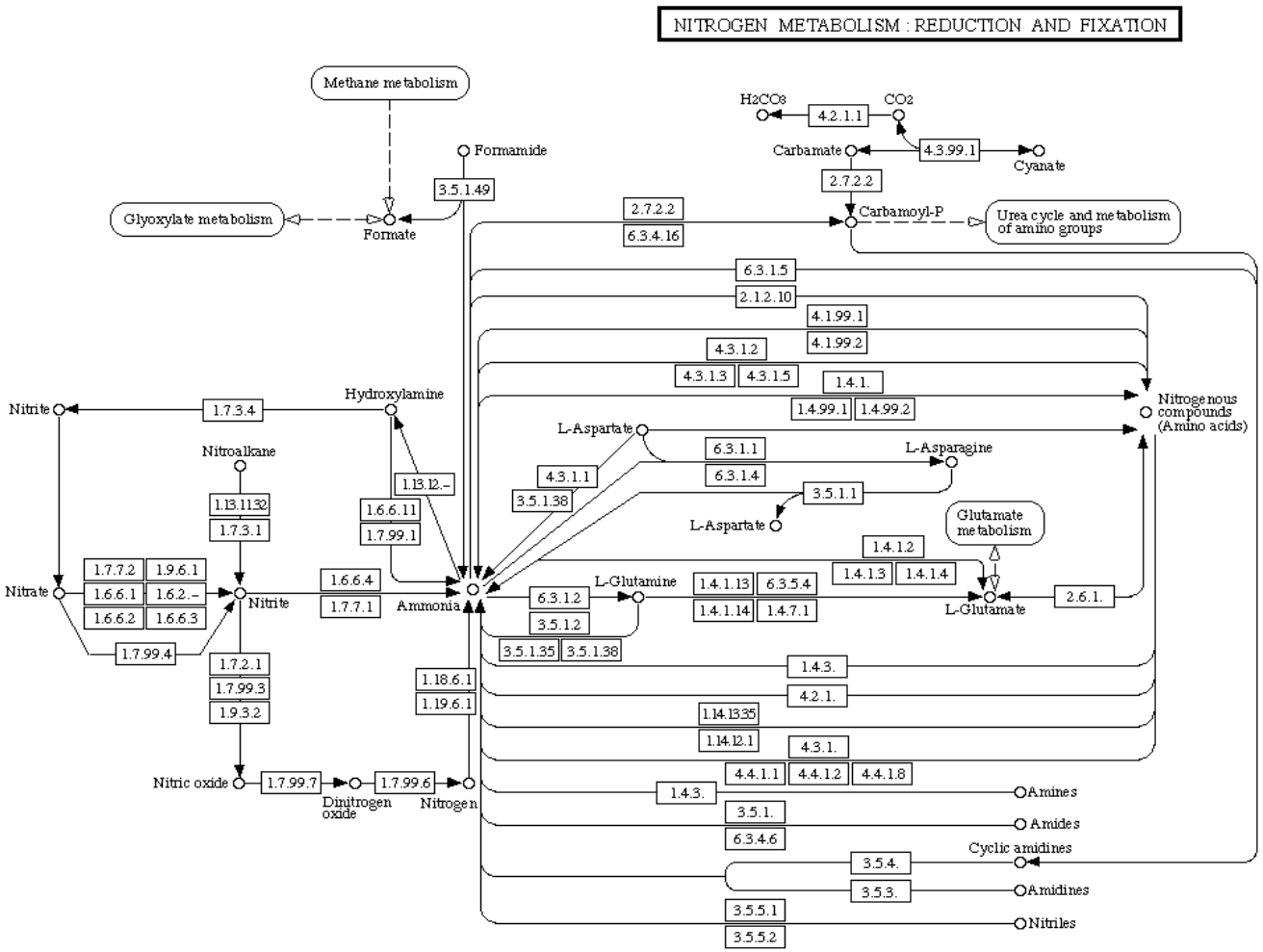
http://www.htslib.org/doc/samtools.html

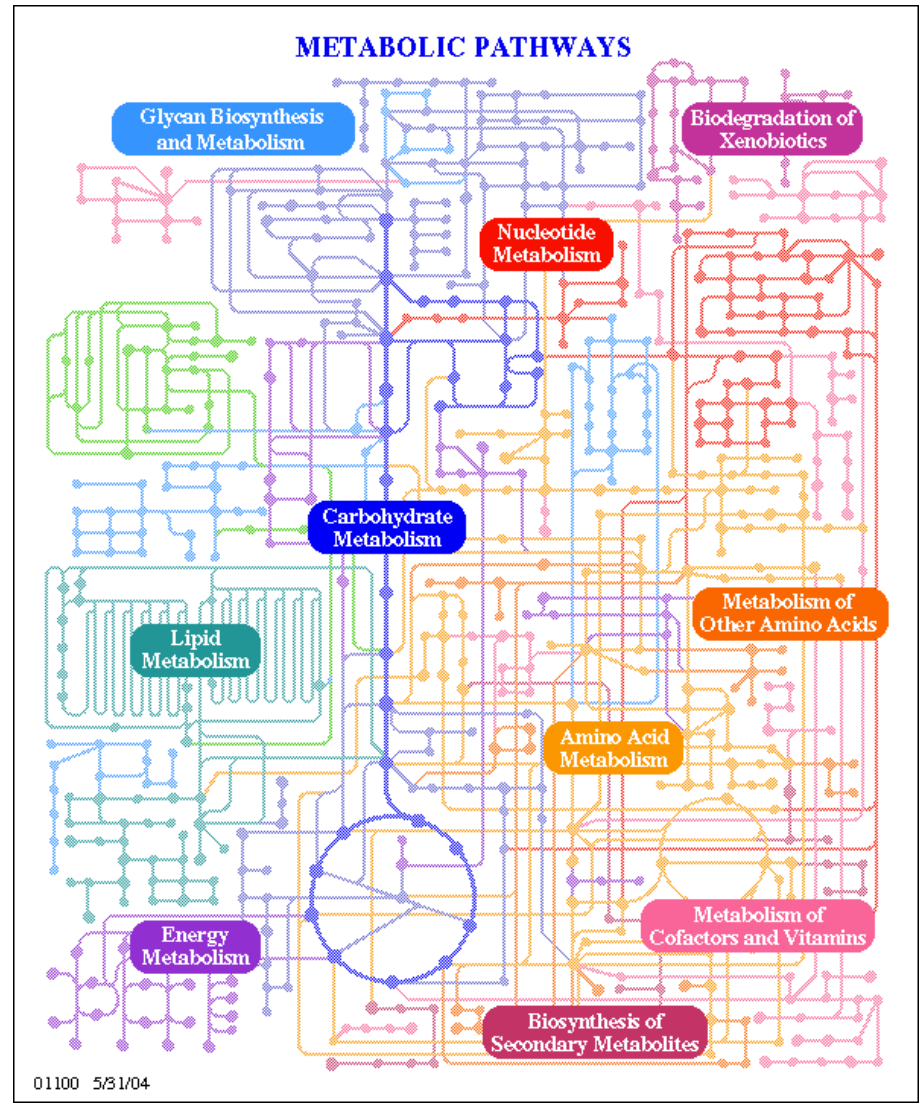
Public databases

- 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

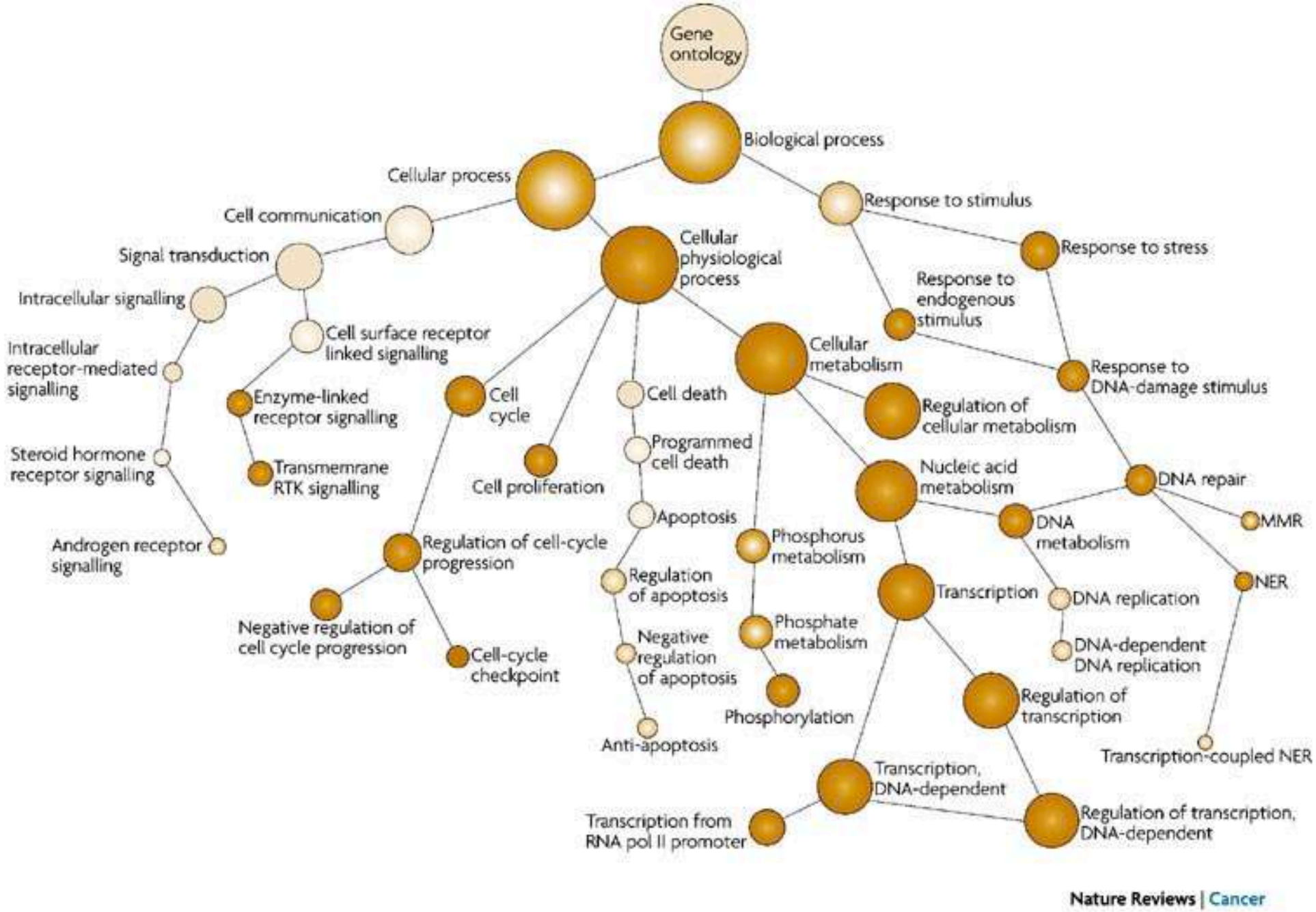
KEGG: Kyoto Encylopedia of Genes and Genomes





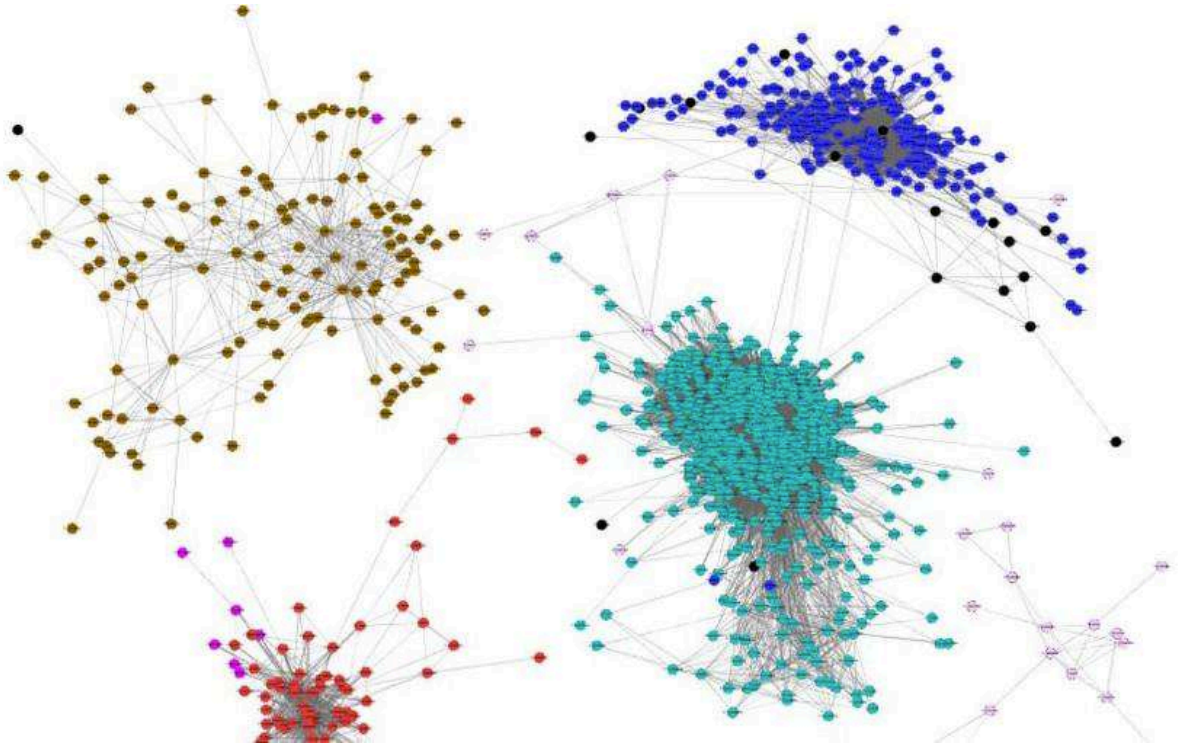
Genome Ontology

- 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)



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

Importance of networks in biology



Gene interaction networks

Protein interaction network

Interactions in a multi-omics world

TABLE 1 | Overview on selected resources for molecular interactions and omics datasets.

Resource	Data type	Organisms	References
STRING	P-P[a]	> 5000	Szklarczyk et al., 2015
BioGrid	P-P	> 60	Stark et al., 2006
inBio map	P-P	HS	Li et al., 2017
GWAS catalog	D-PH	HS	MacArthur et al., 2017
KEGG	multiple	> 5000	Kanehisa and Goto, 2000
APID	P-P	> 400	Alonso-Lopez et al., 2016
doRINA	P-R, miR-R	HS, MM, DM, CE	Blin et al., 2015
REMAP	P-D	HS	Chèneby et al., 2018
IntAct	P-P[b]	multiple	Orchard et al., 2014
Pathway Commons	multiple	multiple	Cerami et al., 2011
AGRIS	P-D	AT	Yilmaz et al., 2011
ENCODE	G, T, E	HS	The ENCODE Project Consortium, 2012
modENCODE	G, T, E	DM, CE	Celniker et al., 2009
GTEx	G, T	HS	Carithers et al., 2015
ROADMAP	E, T	HS	Roadmap Epigenomics Consortium, 2015
GEO	G, T, E	multiple	Edgar et al., 2002; Barrett et al., 2013
ARCHS4	T	HS, MM	Lachmann et al., 2018
The Human Protein Atlas	T, P	HS	Thul et al., 2017
MetaboLights	M	multiple	Haug et al., 2013
TCGA	G, T, E	HS	Weinstein et al., 2013

Data type column depicts either the type of interactions (e.g., protein-protein interaction, P-P) or the type of omics data available in the data collection. Interactions: M, metabolite; P, protein; D, DNA; R, RNA; PH, phenotype; Organisms: HS, H. sapiens; AT, A. thaliana; MM, M. musculus; DM, D. melanogaster; CE, C. elegans; Omics: G, genomic; E,epigenomic; T,transcriptomic.

[a] *includes functional interactions.*

[b] *focus on P-P, but arbitrary interactions possible.*

Hawe *et al* (2019) Front. Genet. 10:535.
doi: 10.3389/fgene.2019.00535

Interactions in a multi-omics world

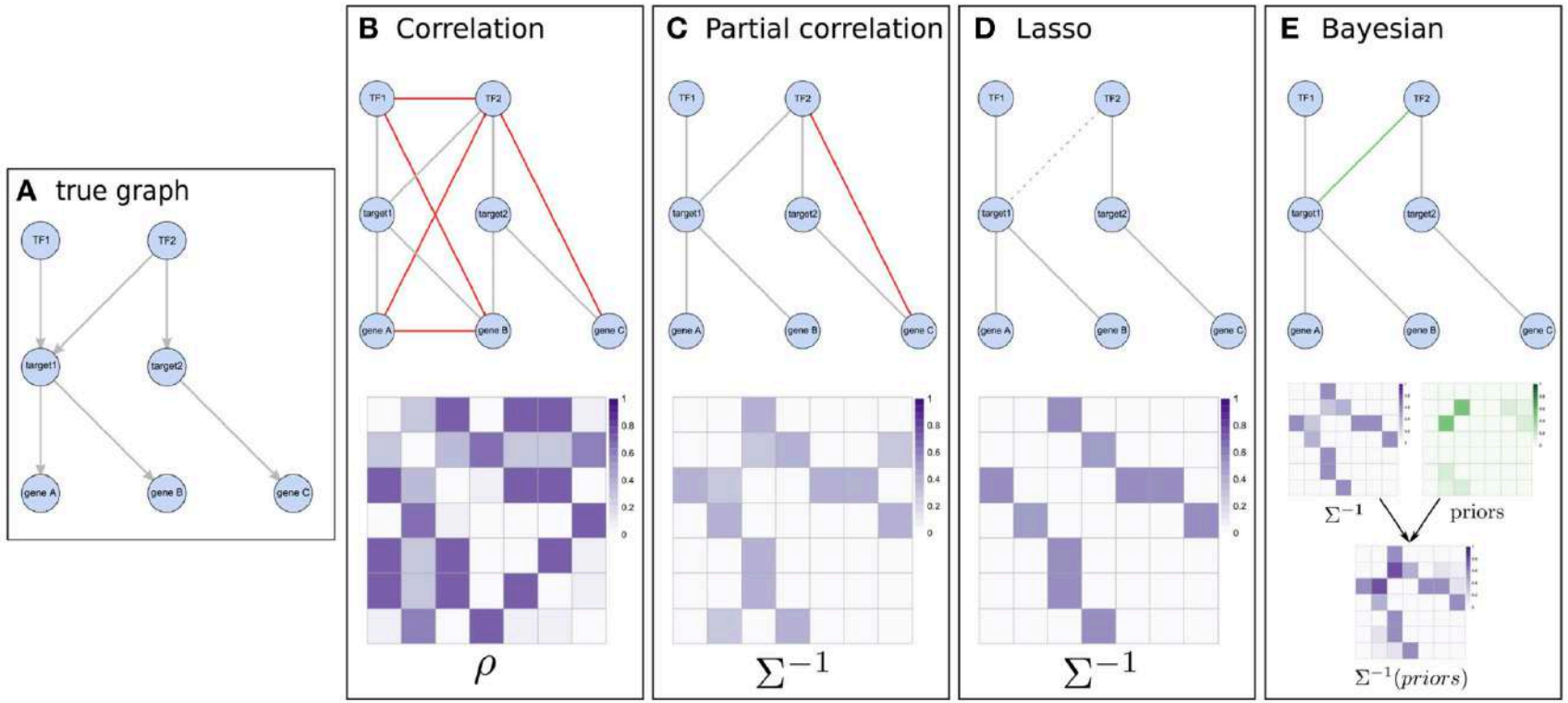
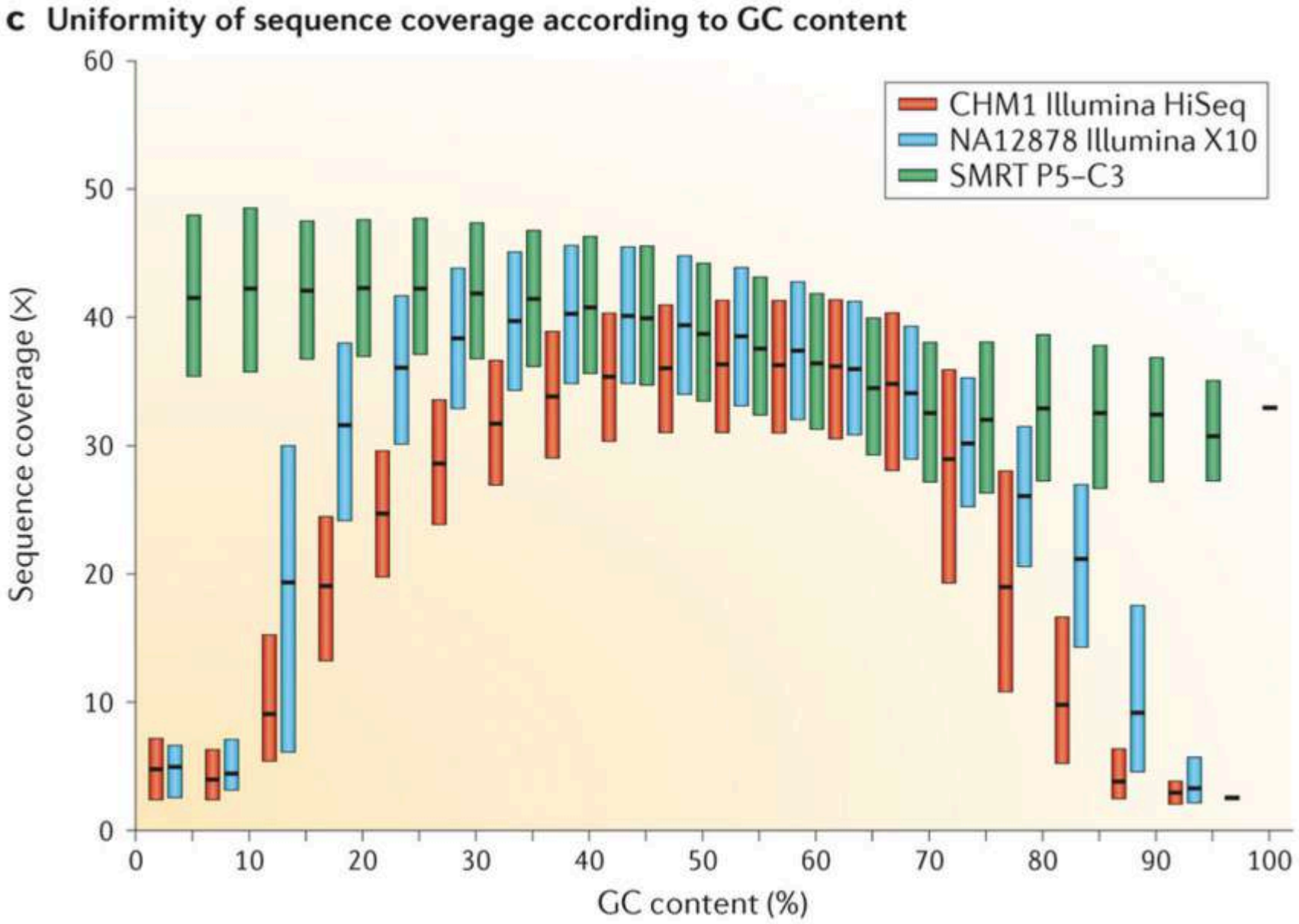


FIGURE 3 | Illustration of the concept of different network inference methods. **(A)** represents a known pathway structure which should be recovered from functional data using the different approaches: two transcription factors influencing expression of two target genes which in turn affect the expression of other downstream genes. **(B,C)** show correlation based results and their estimated matrices (correlation and partial-correlation, respectively). While using Pearson correlation results in many indirect associations (shown in red), this is largely amended by using partial correlations. **(D)** The graphical lasso pushes weaker associations (e.g., between *TF1* and *gene C*) toward zero in the precision matrix and might do so even for real edges which have relatively low evidence in the data (like the edge between *TF2* and *target1*). **(E)** When considering prior information, weak associations still have a chance of getting selected if their respective prior (shown in green) supports them.

Hawe *et al* (2019) Front. Genet. 10:535.
doi: 10.3389/fgene.2019.00535

Analysis and interpretation

Is your data good enough? - Sequencing Biases

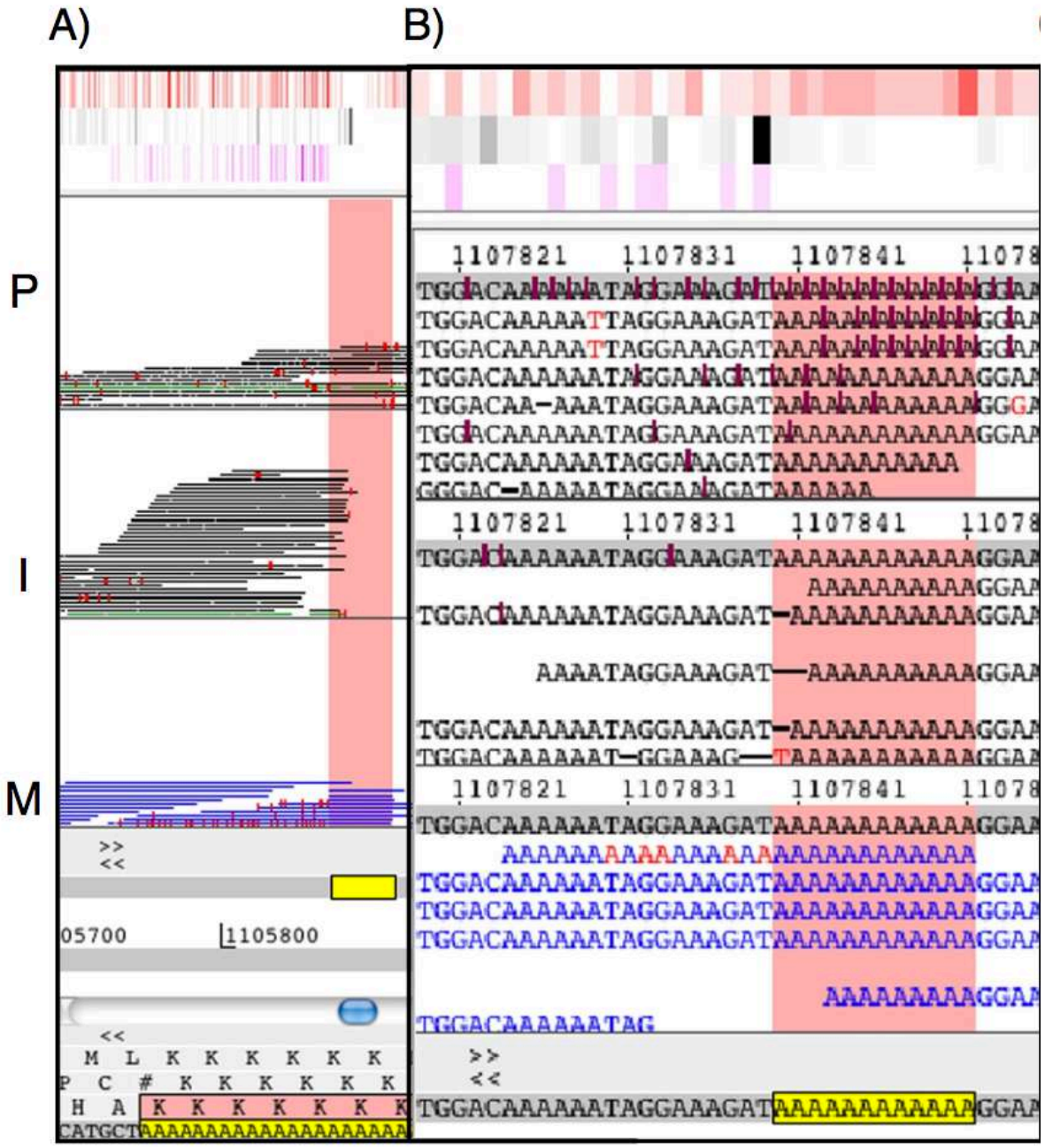


http://www.nature.com/nrg/journal/v16/n11/fig_tab/nrg3933_F2.html

Sequencing Errors

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

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



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

Three situations you are most likely to encounter

Genome reference is available (for example, humans):

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

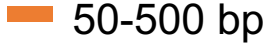
Genome reference is NOT available

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

Counting:

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

More Definition

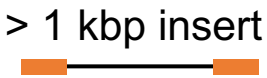


Read

A sequenced piece of DNA

Paired-end read

Sequencing both ends of a short DNA fragment



Mate-pair read

Sequencing both ends of a long DNA fragment

Insert size

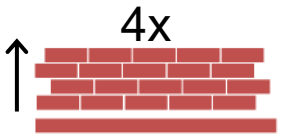
The length of the DNA fragment

Contig

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

Scaffold

Contigs separated by gaps of known length



Coverage

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

What is an alignment?

Align the following two sequences:

```
ATTGAAAGCTA
GAAATGAAAAGG
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

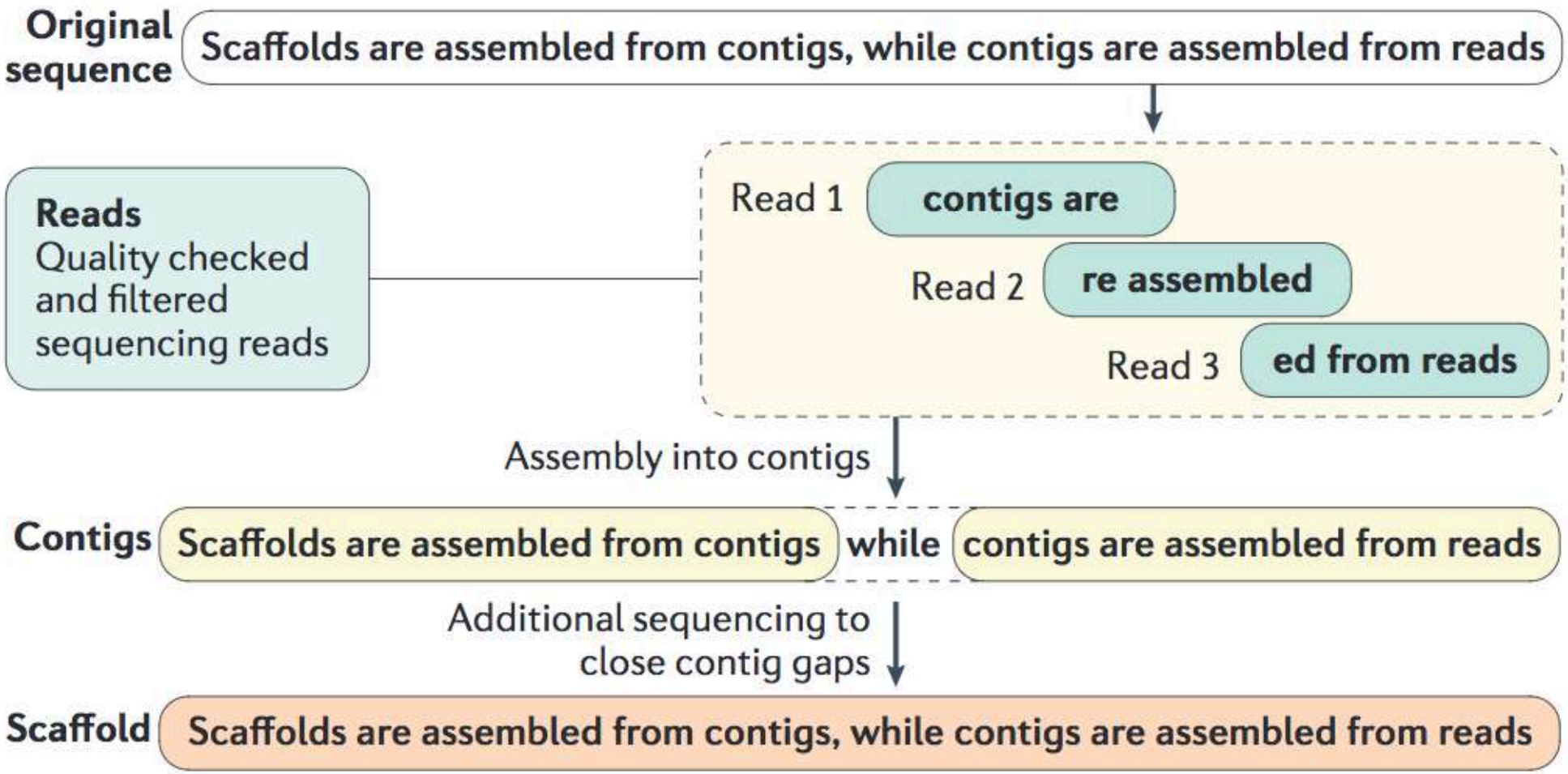
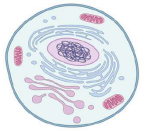


Figure 2 | **Sequence read assembly.** A mock example explaining bioinformatic sequence assembly along with the terms sequence, reads, contigs and scaffolds.

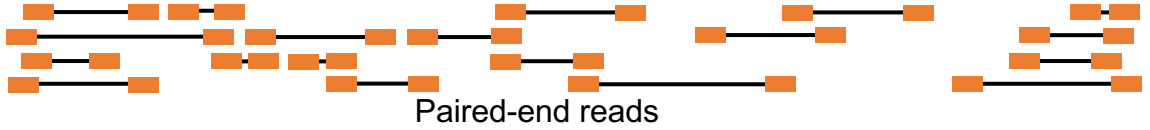
Assembly



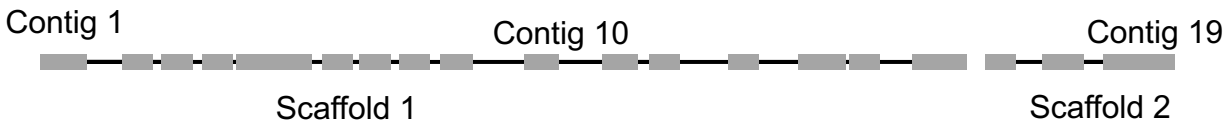
Genome

Fragment

Sequence



Assemble

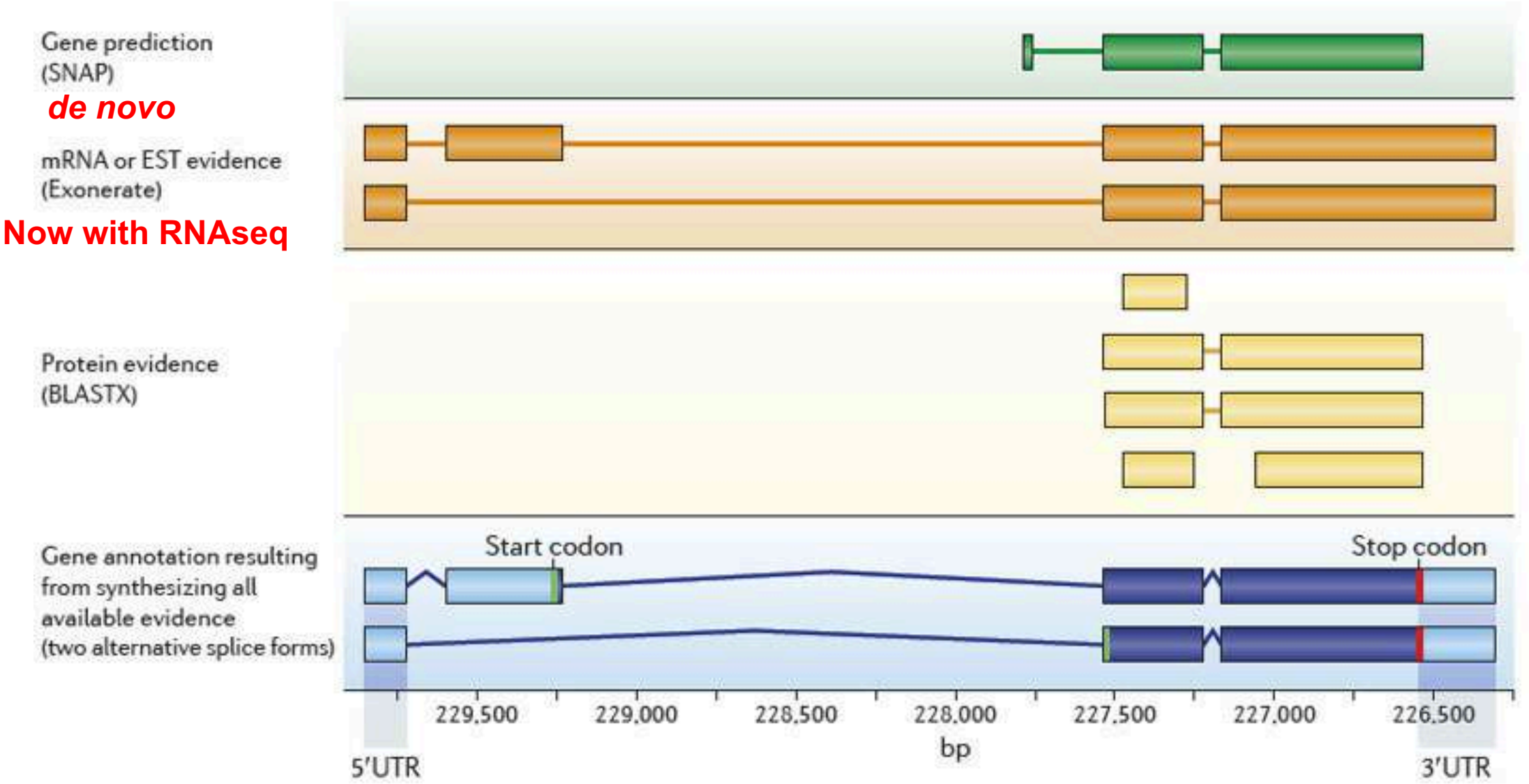


After assembly

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

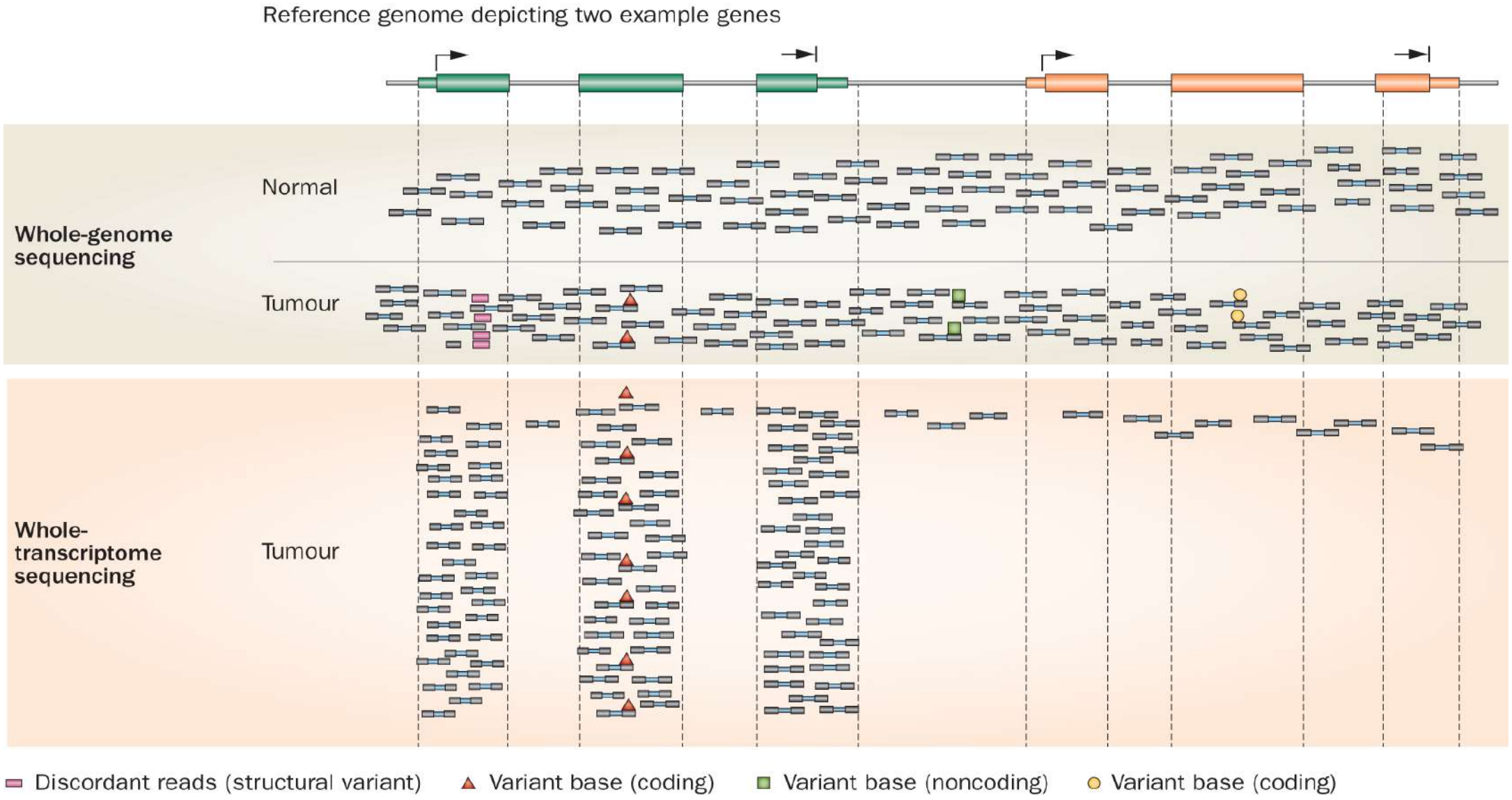
- How accurate is it?
- Have you tried different assemblers?
- Can you improve with additional data or diminishing returns?
- Is there contamination?
- How does it compare to other species?

Annotation



Yandell and Ence Nature Genetics Review (2012)

Mapping



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.

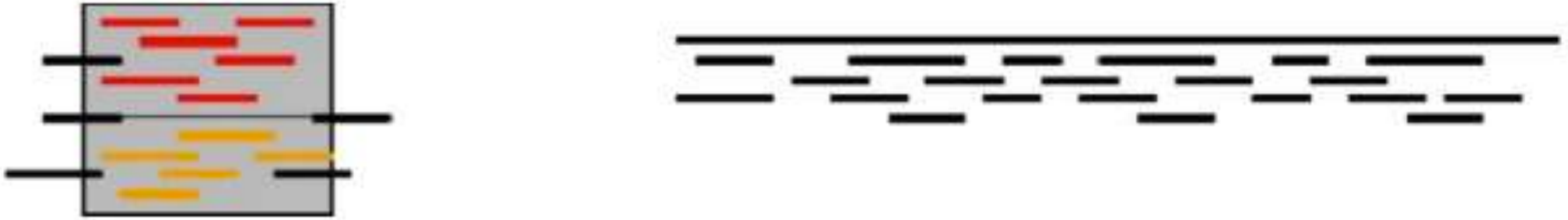
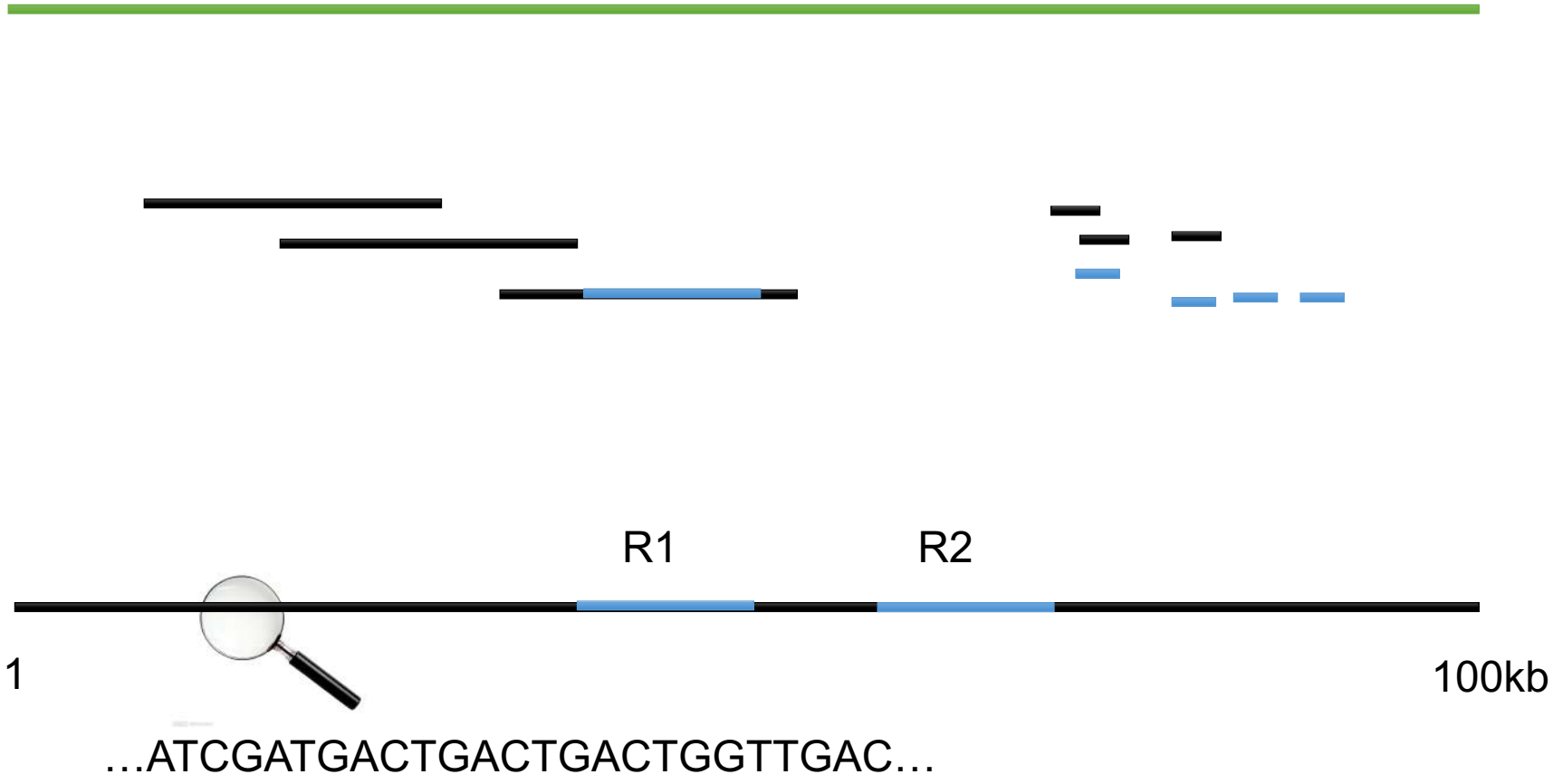


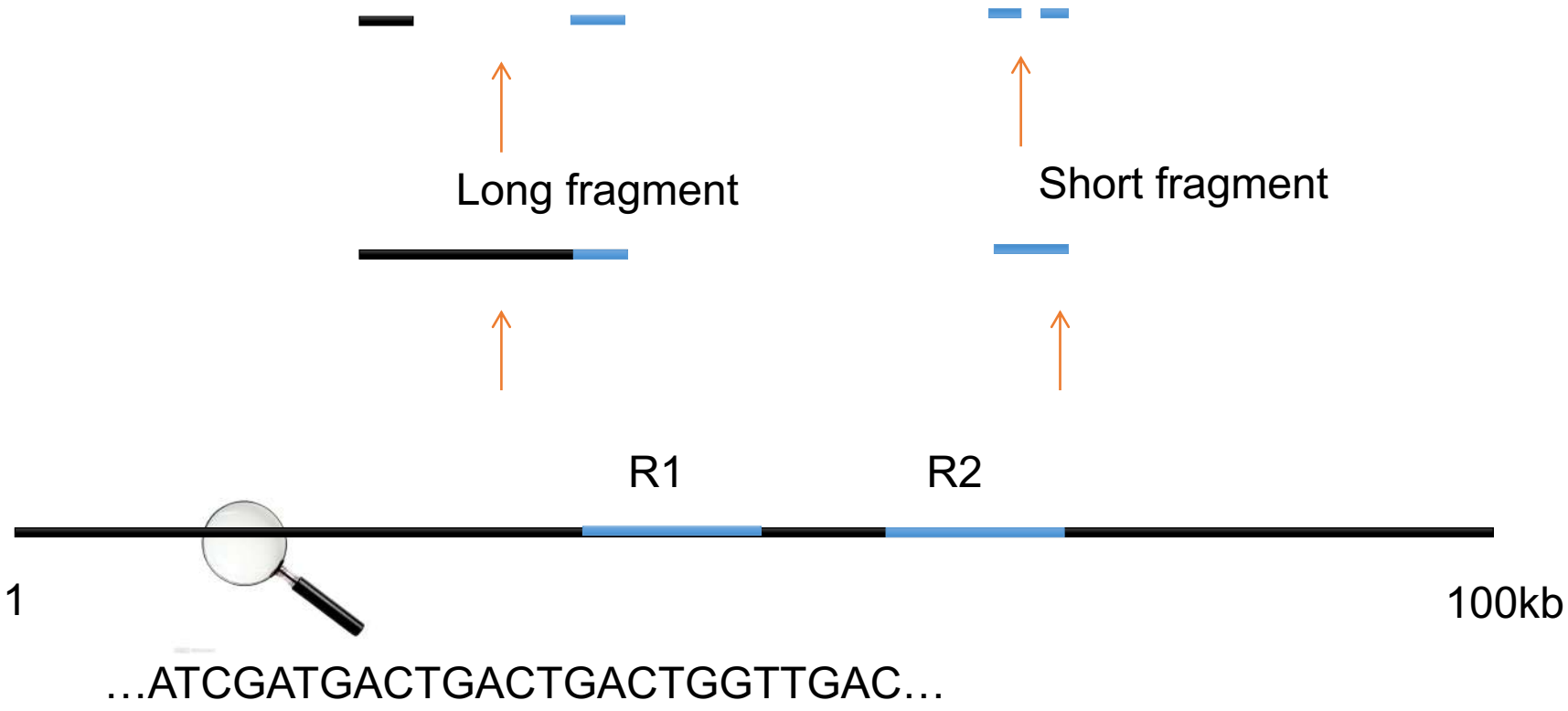
Figure 6. Genome mis-assembled due to a repeat. The assembly program incorrectly combined the reads from the two copies of the repeat leading to the creation of two separate contigs

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

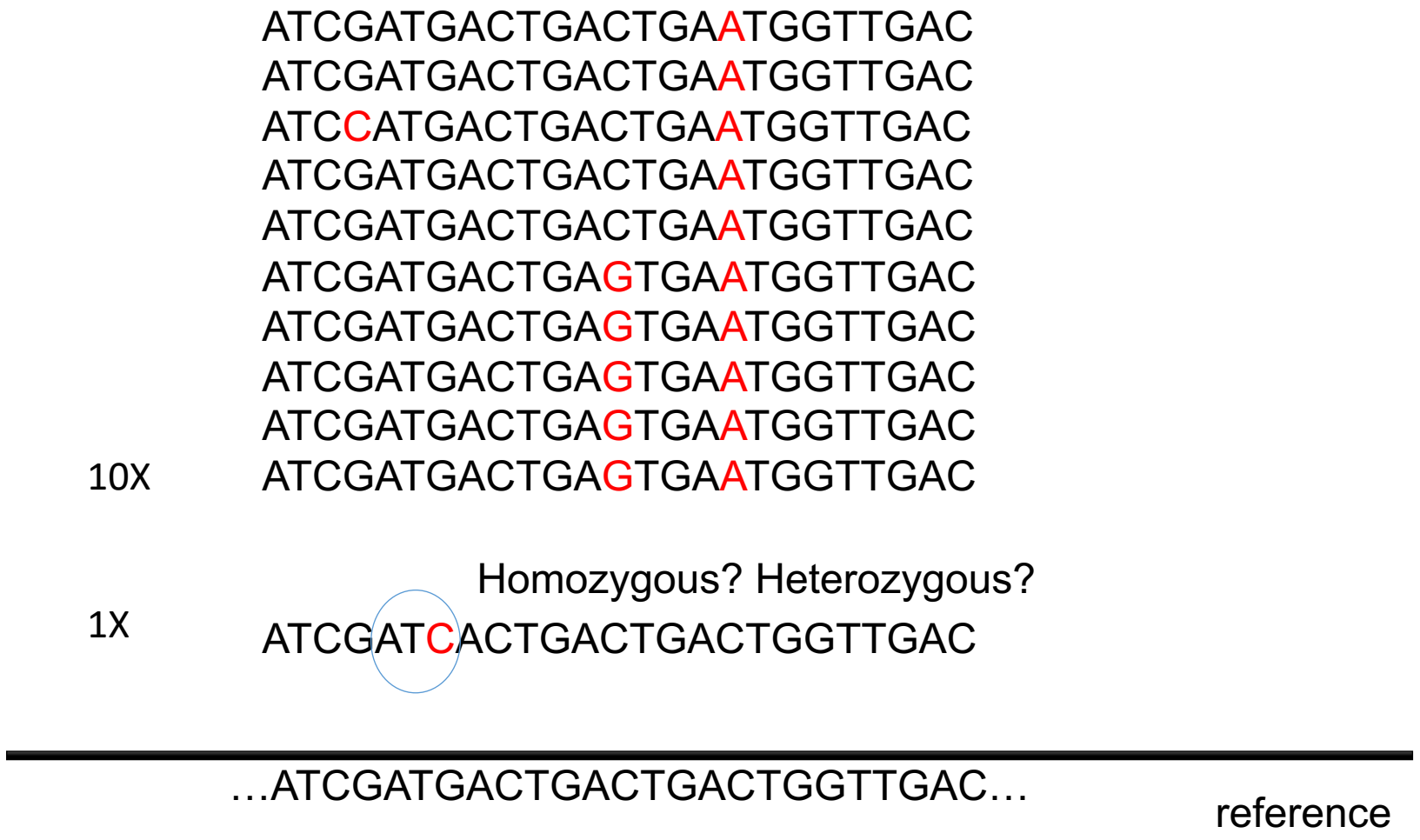
Read length matters in sequencing



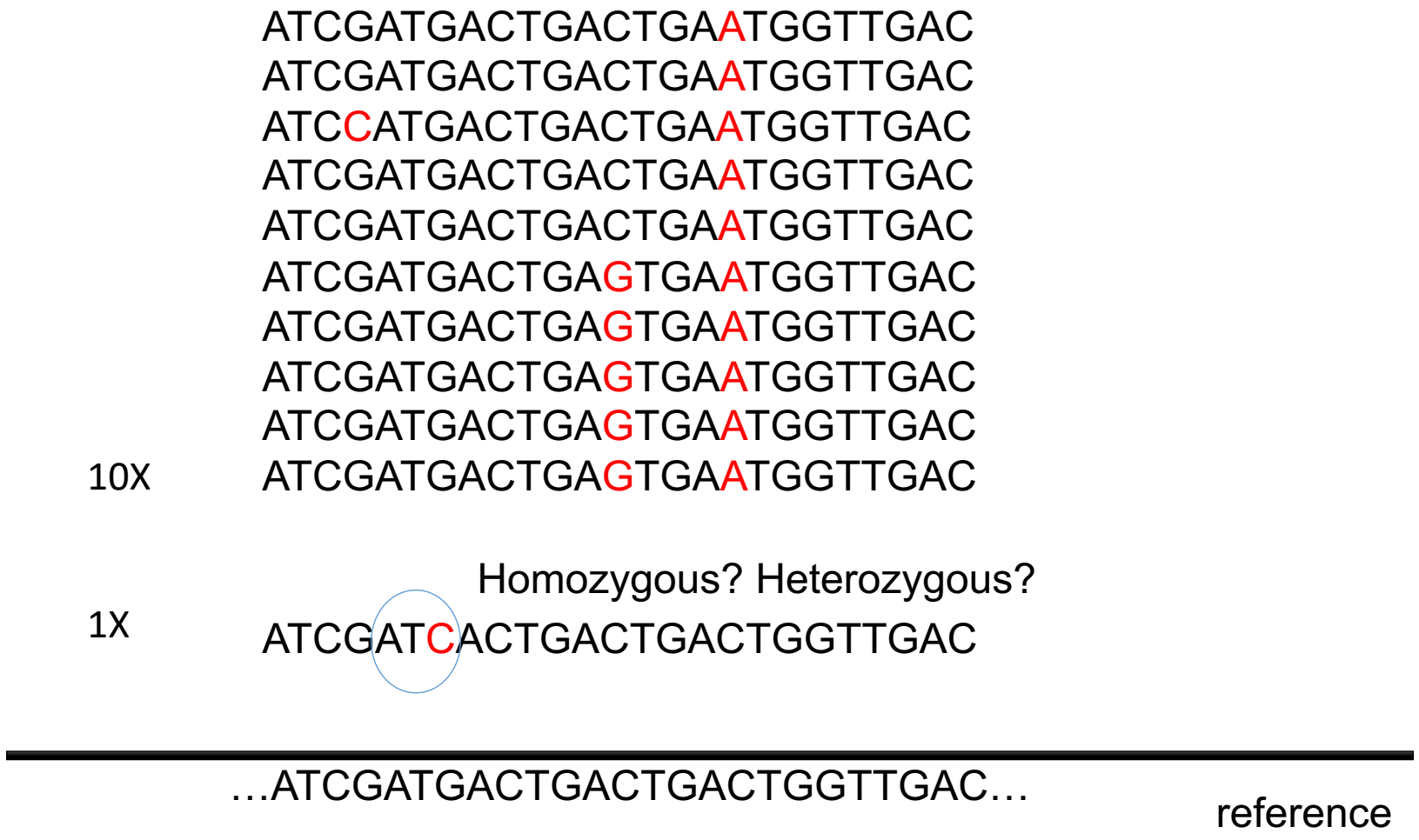
Paired end and insert size matter in sequencing



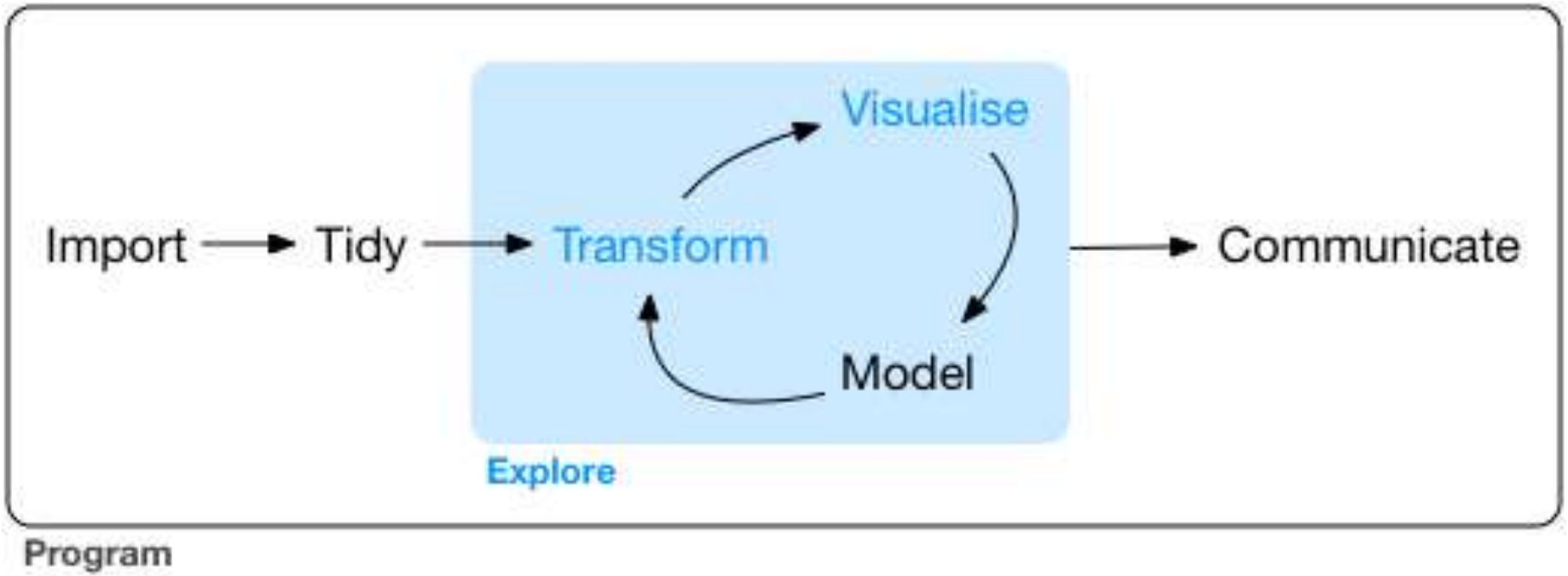
Depth matters in sequencing



Interpretation

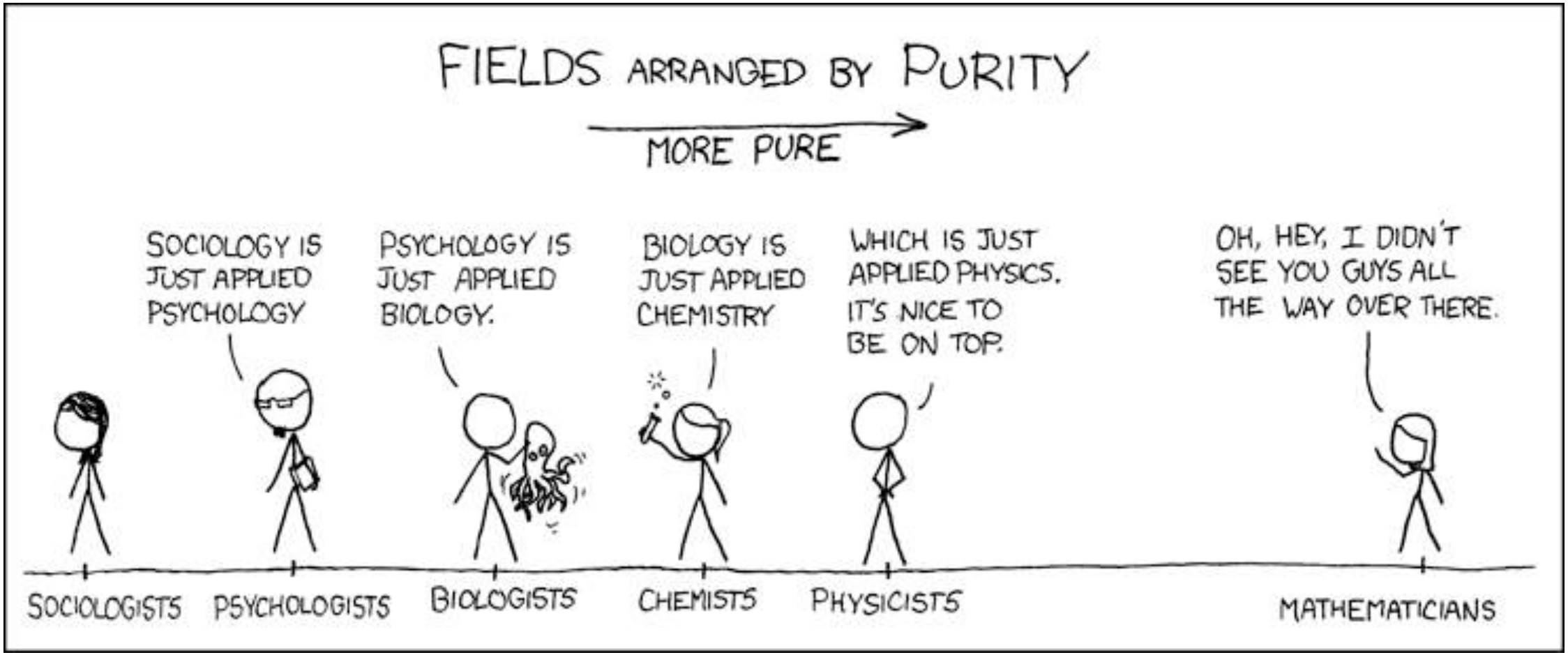


Modeling and Visualisation



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

Purity?



https://xkcd.com/435/

Visual Vocabulary

Designing with data

Correlation

Show the relationship between two or more variables. Be mindful that, unless you tell them otherwise, many readers will assume the relationships you show them to be causal (i.e. one causes the other)

Examples of use

Inflation & unemployment, income & life expectancy

Chart types

scatterplot

The standard way to show the relationship between two variables, each of which has its own axis

line-column

A good way of showing the relationship between an amount (columns) and a rate (line)

scatterplot-connected

Usually used to show how the relationship between 2 variables has changed over time

Bubble

Like a scatterplot, but adds additional detail by sizing the circles according to a third variable

XY-heatmap

A good way of showing the patterns between 2 categories of data, less good at showing fine differences in amounts

Ranking

Use where an item's position in an ordered list is more important than its absolute or relative value. Don't be afraid to highlight the points of interest.

Examples of use

Wealth, deprivation, league tables, constituency election results

Chart types

bar-ordered

Standard bar charts display the ranks of values much more easily when sorted into order

column-ordered

Standard column charts display the ranks of values much more easily when sorted into order

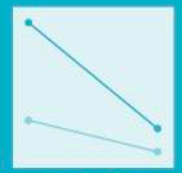
symbol-proportional-ordered

Use when there are big variations between values and/or seeing fine differences between data is not so important.

dot-plot-strip

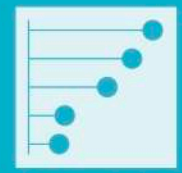
Dots placed in order on a strip are a space-efficient method of laying out ranks across multiple categories.

slope



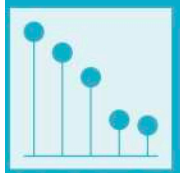
Perfect for showing how ranks have changed over time or vary between categories.

lollipop-h



Lollipop charts draw more attention to the data value than standard bar/column and can also show rank effectively

lollipop-v



Lollipop charts draw more attention to the data value than standard bar/column and can also show rank effectively

bump

https://ft-interactive.github.io/visual-vocabulary/

Case studies

Classical genetics

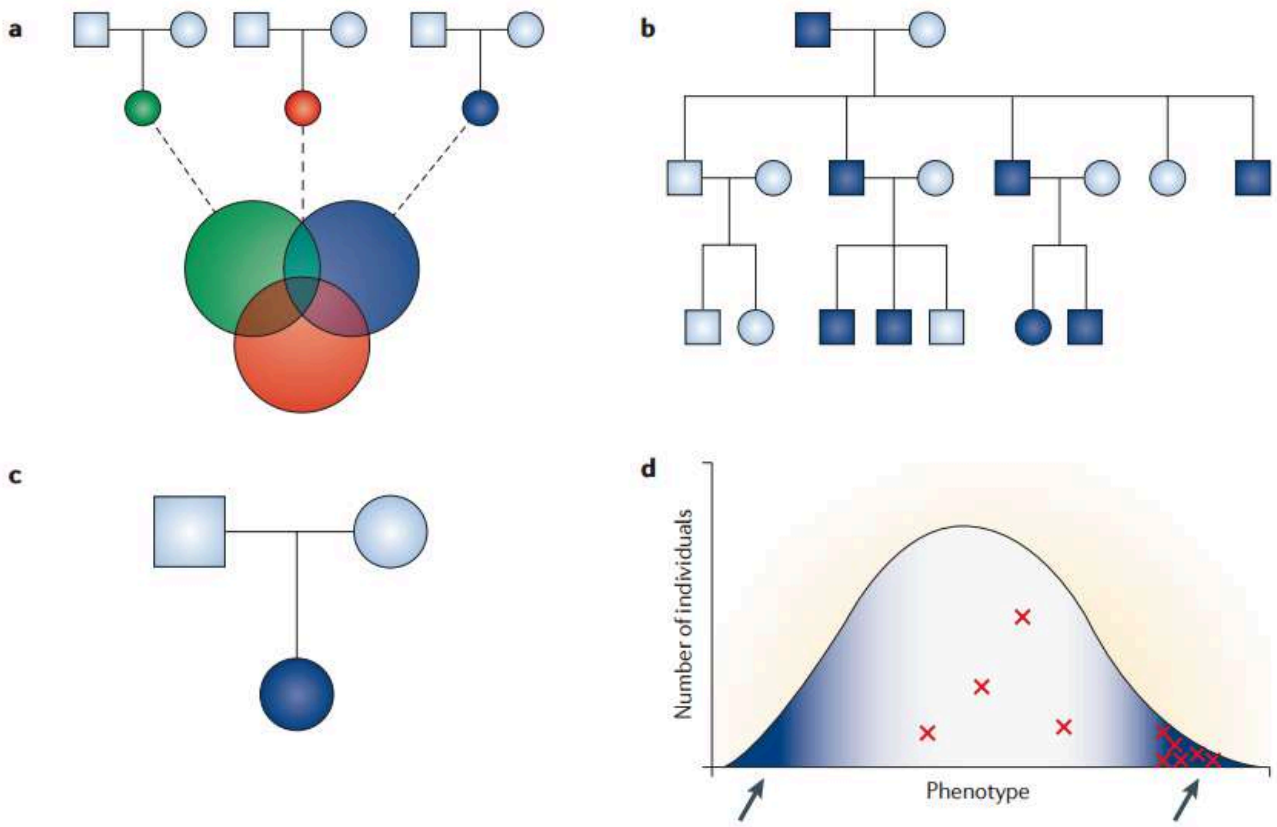
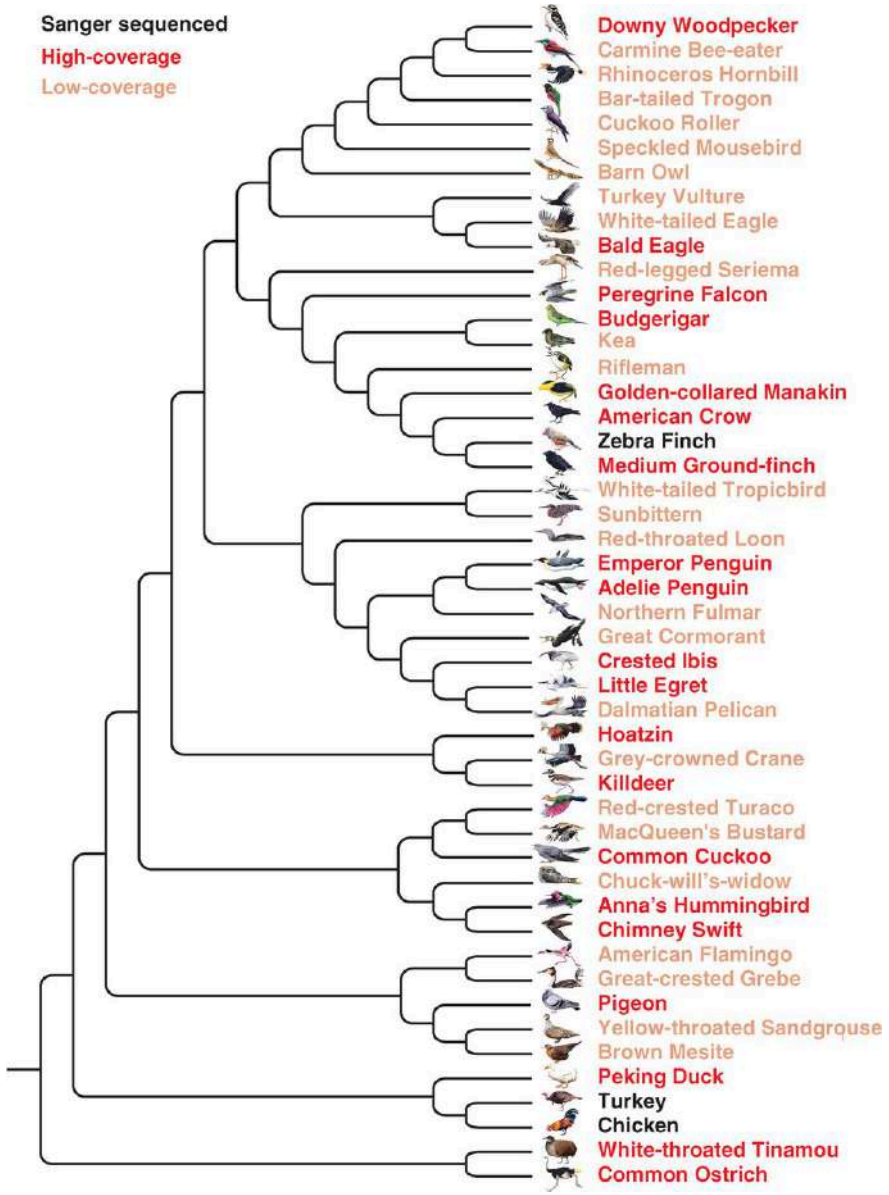


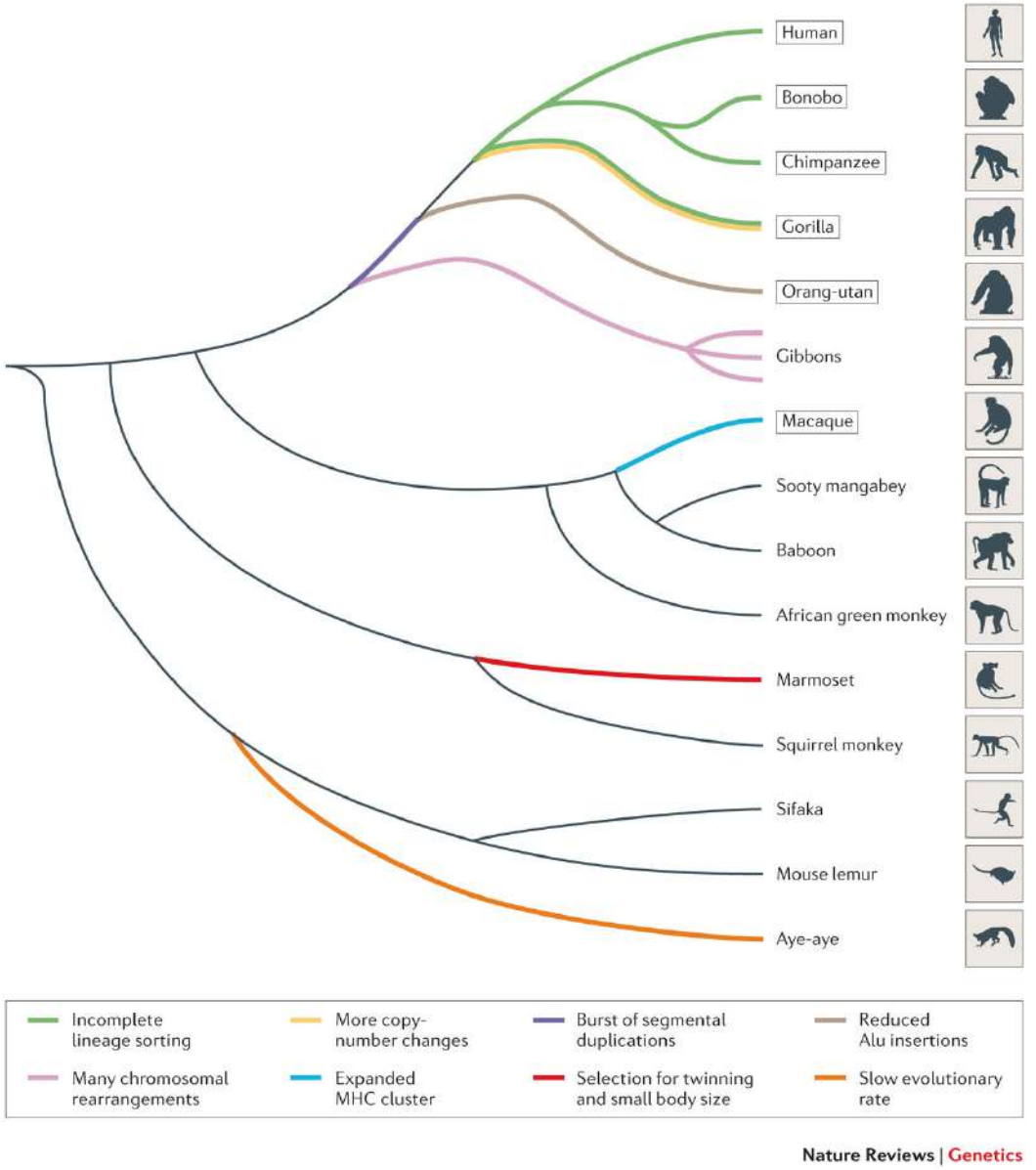
Figure 2 | **Strategies for finding disease-causing rare variants using exome sequencing.** Four main strategies are illustrated. **a** | Sequencing and filtering across multiple unrelated, affected individuals (indicated by the three coloured circles). This approach is used to identify novel variants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

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

Comparative genomics



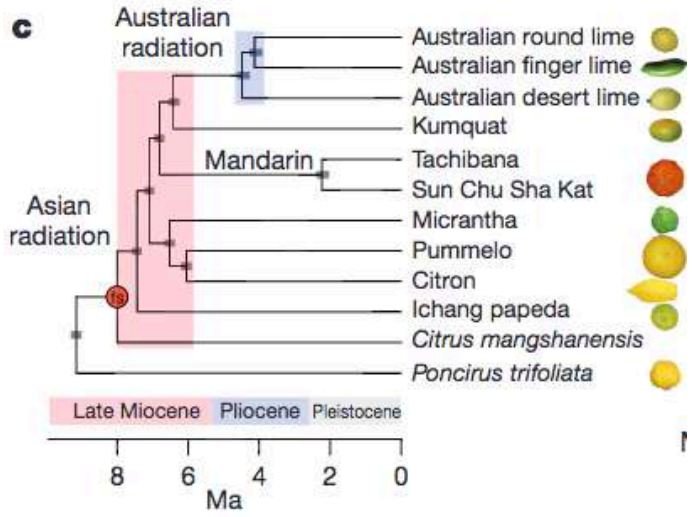
Guojie Zhang et al. Science (2014)

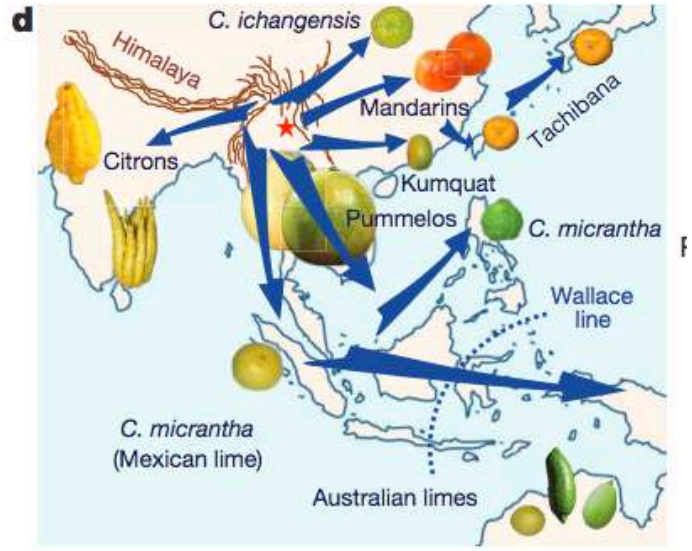


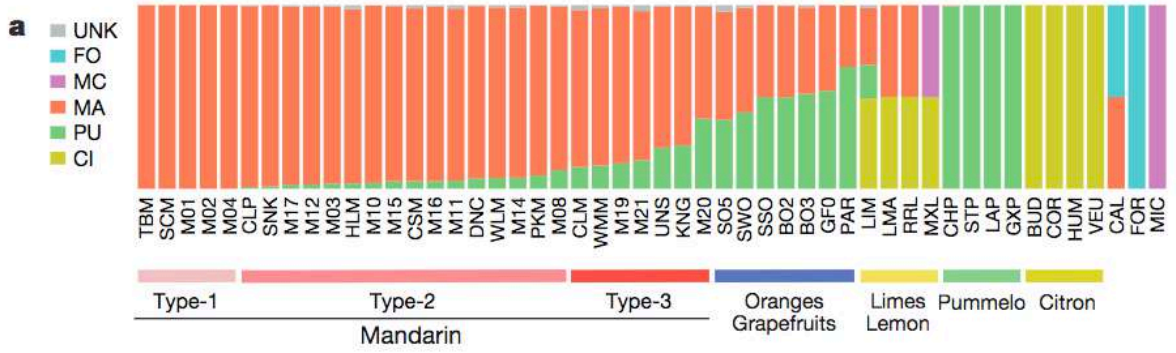
Roger & Gibbs Nature Reviews Genetics (2014)

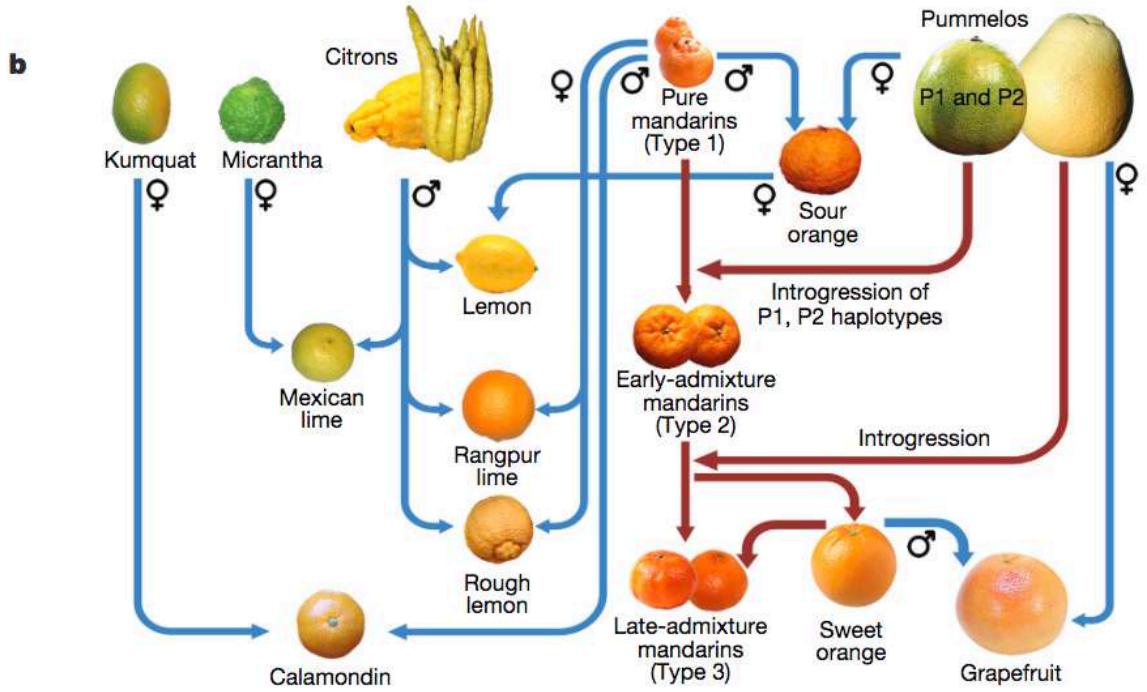
Comparative genomics

Genomics of the origin and evolution of Citrus









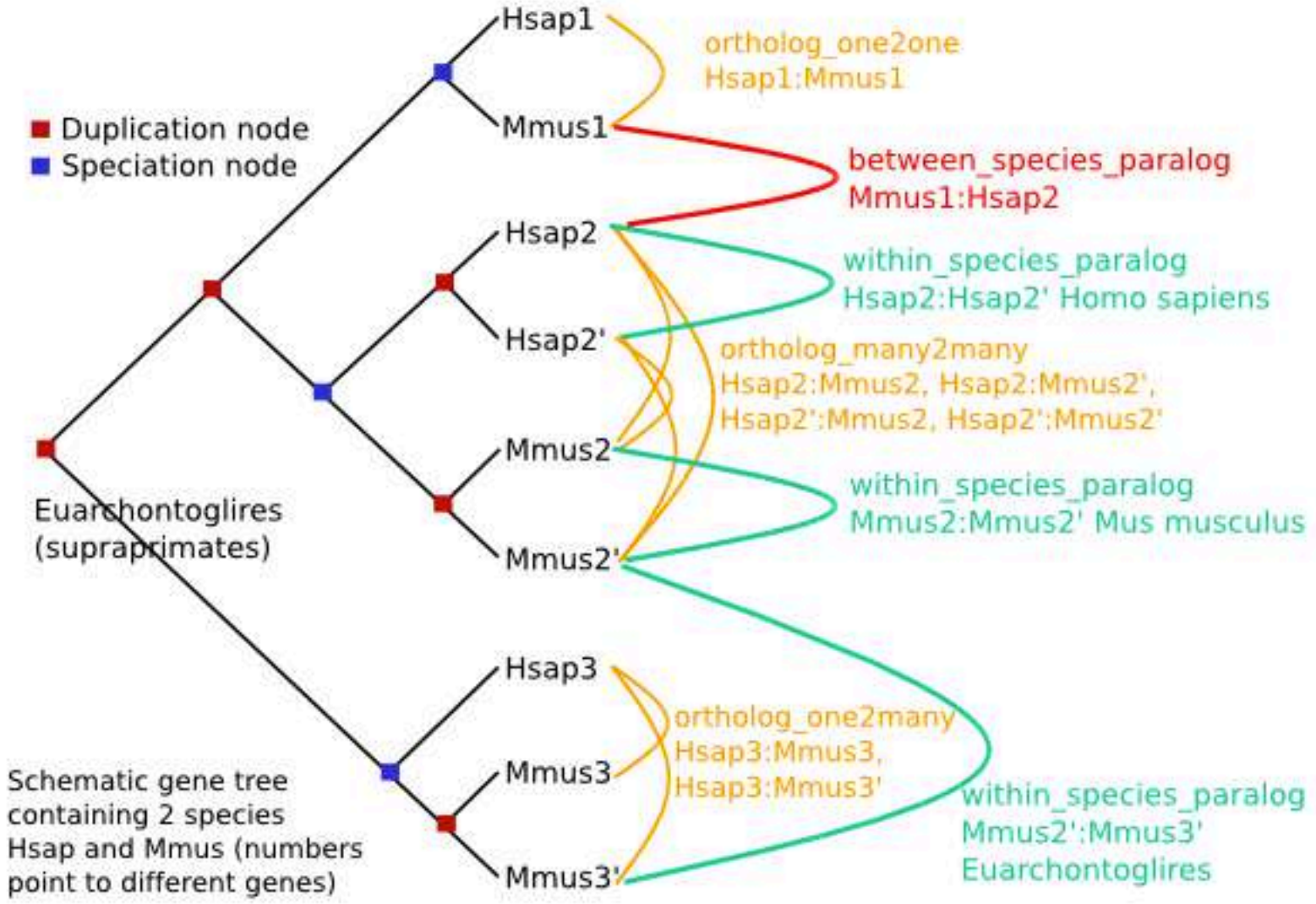
Wu *et al.*, Nature (2018)

Homologs: Orthologs and paralogs

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

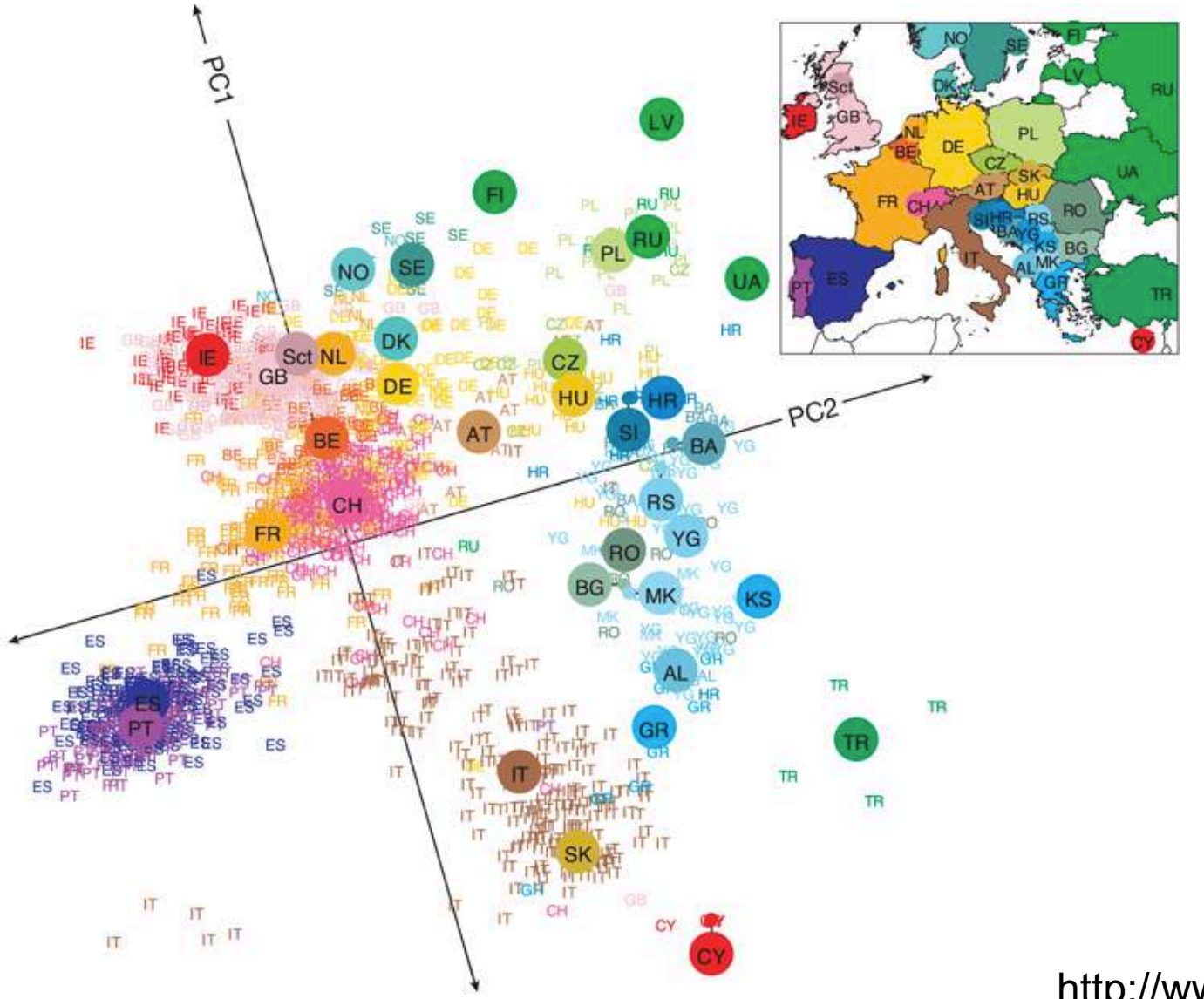
Depending on the number of genes found in each species, we differentiate among 1:1, 1:many and many:many relationships.

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

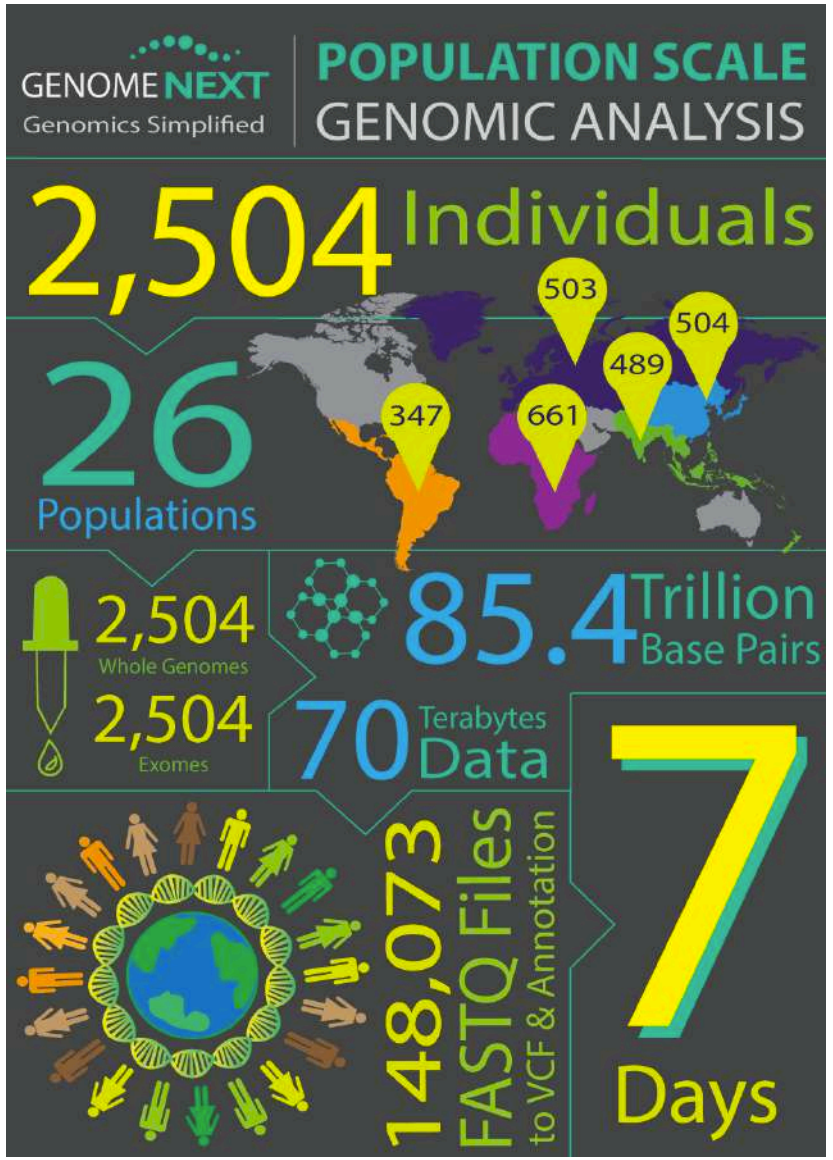


http://asia.ensembl.org/info/genome/compara/homology_method.html?redirect=no

Population genomics

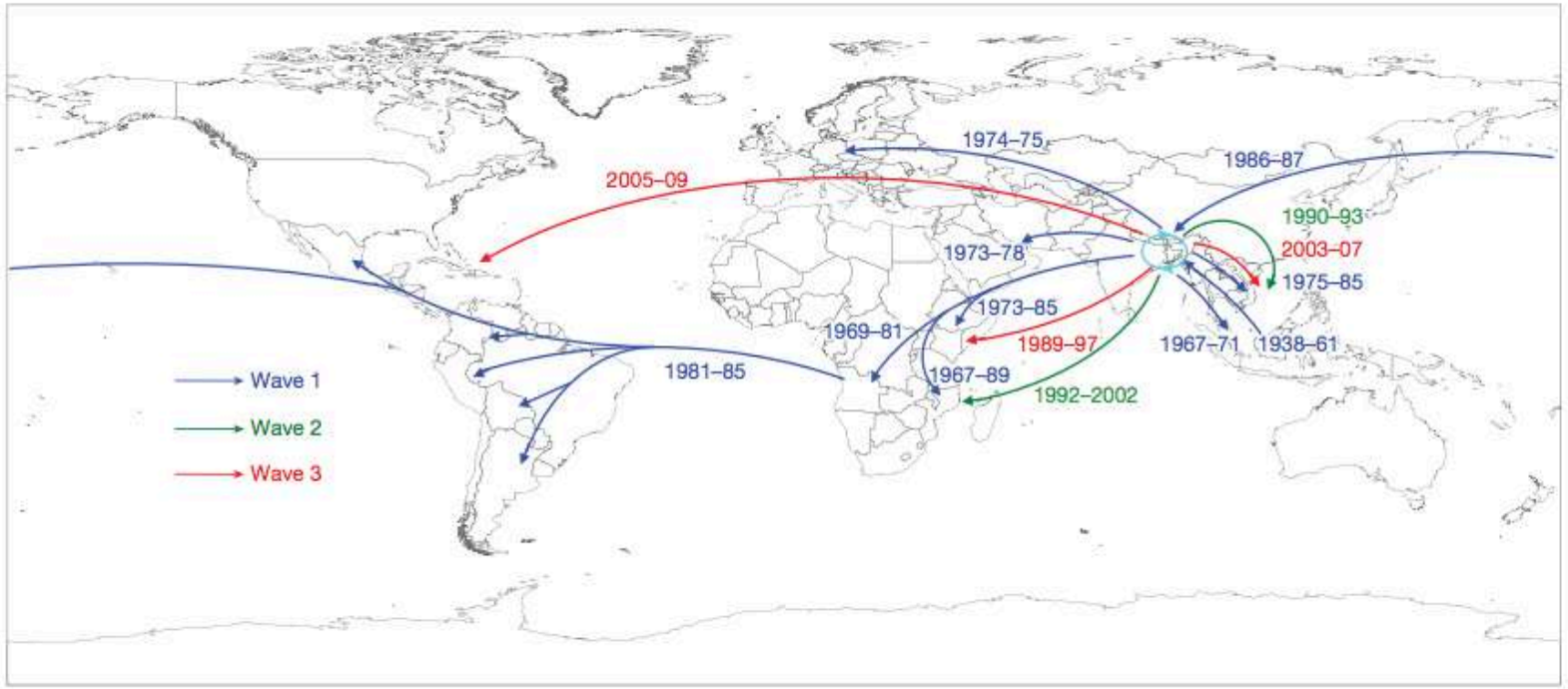


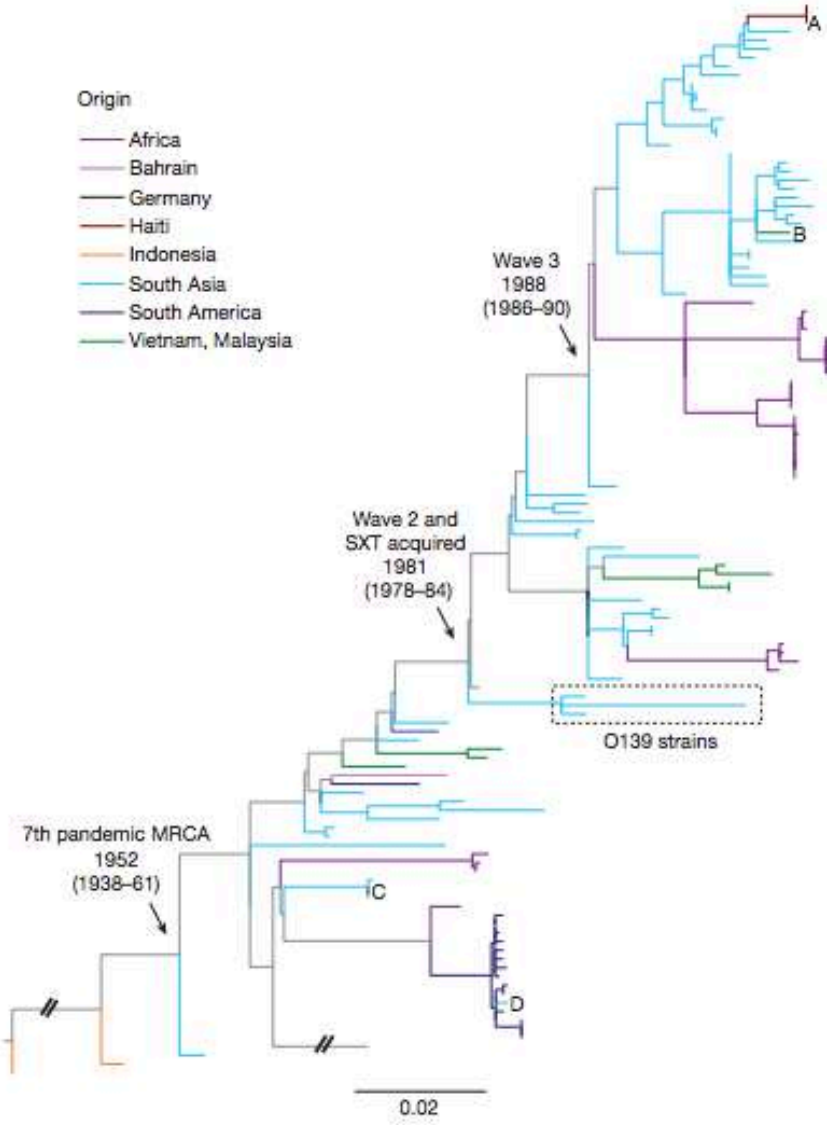
Novembre et al Nature (2008)



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

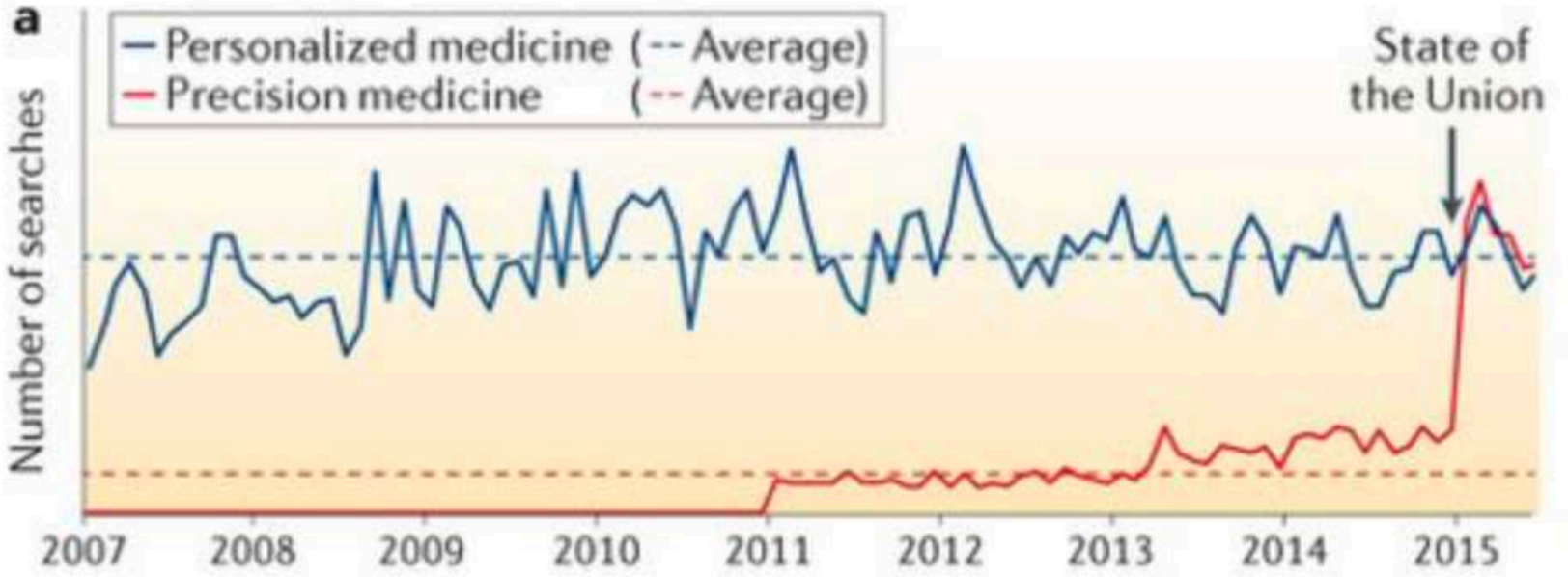
Population genomics





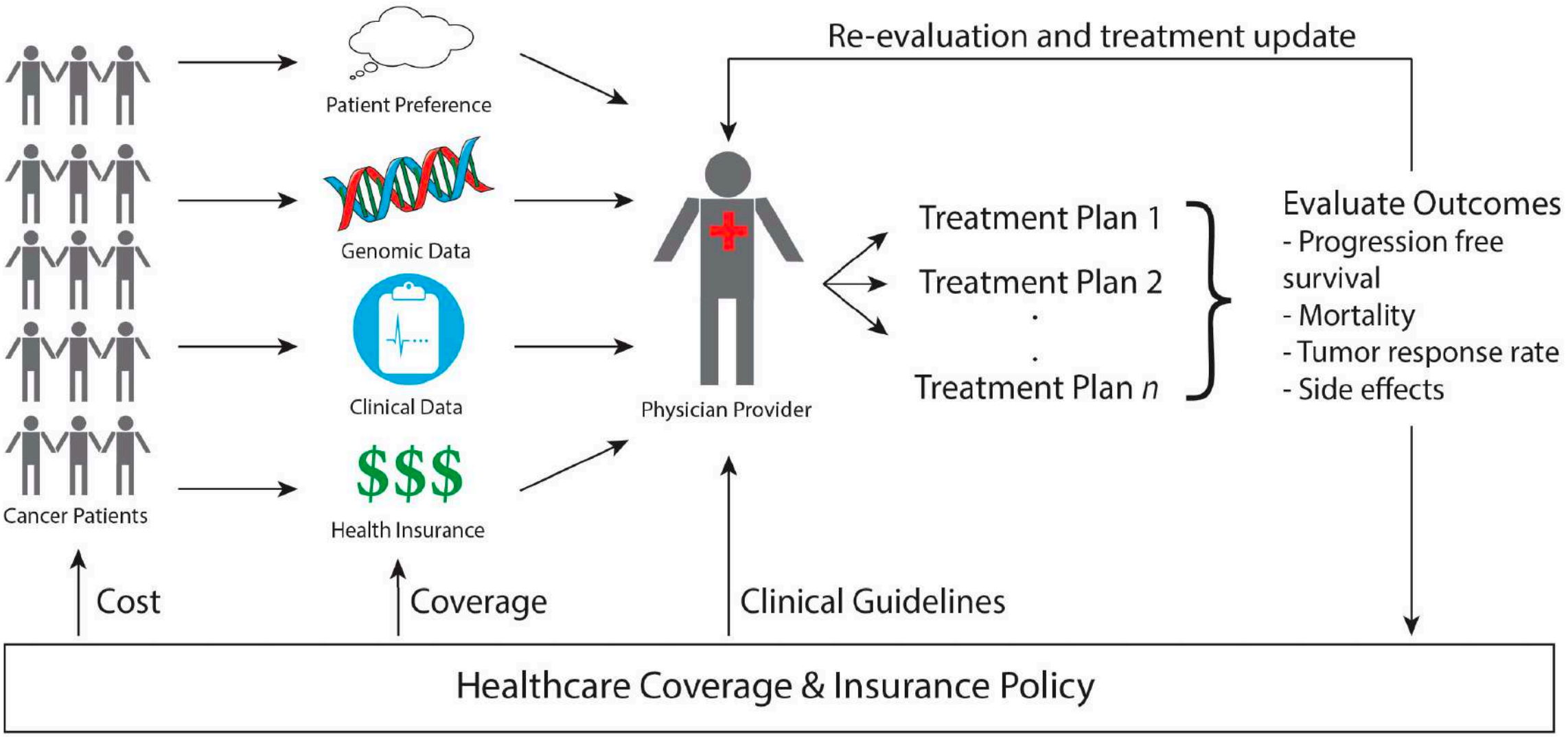
Mutreja *et al.*, Nature Genetics (2011)

Precision medicine



Ashley (2016) Nature Review Genetics

Outline of precision medicine



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

PM examples

Table 1 | **Examples of precision medicine**

Condition	Gene	Action
Mendelian disease		
Cystic fibrosis	*CFTR*	Specific therapies such as ivacaftor and a combination of lumacaftor and ivacaftor
Long QT syndrome	*KCNQ1*, *KCNH2* and *SCN5A*	Specific therapy for patients with *SCN5A* mutations
Duchenne muscular dystrophy	*DMD*	Ongoing phase III clinical trials of exon-skipping therapies
Malignant hyperthermia susceptibility	*RYR1*	Avoid volatile anaesthetic agents; avoid extremes of heat
Familial hypercholesterolaemia (FH)	*PCSK9*, *APOB* and *LDLR*	• Heterozygous FH (HeFH): eligible for PCSK9 inhibitor drugs • Homozygous FH (HoFH): eligible for PCSK9 inhibitor drugs in addition to lomitapide and mipomersen
Dopa-responsive dystonia	*SPR*	Therapy with dopamine precursor L-dopa and the serotonin precursor 5-hydroxytryptophan
Thoracic aortic aneurysm	*SMAD3*, *ACTA2*, *TGFBR1*, *TGFBR2* and *FBN1*	Customization of surgical thresholds based on patient genotype
Left ventricular hypertrophy	*MYH7*, *MYBPC3*, *GLA* and *TTR*	Sarcomeric cardiomyopathy, Fabry disease and transthyretin cardiac amyloid disease have specific therapies
Precision oncology		
Lung adenocarcinoma	*EGFR* and *ALK*	Targeted kinase inhibitors, such as gefitinib and crizotinib
Breast cancer	*HER2*	HER2 (also known as ERBB2)-targeted treatment, such as trastuzumab and pertuzumab
Gastrointestinal stromal tumour	*KIT*	Targeted KIT kinase activity inhibitors, such as imatinib
Melanoma	*BRAF*	BRAF inhibitors, such as vemurafenib and dabrafenib
Pharmacogenomics		
Warfarin sensitivity	*CYP2C9* and *VKORC1*	Adjust dosage of warfarin or consider alternative anticoagulant
Clopidogrel sensitivity, post-stent procedure	*CYP2C19*	Consider alternative antiplatelet therapy (for example, prasugrel or ticagrelor)
Thiopurine sensitivity	*TPMT*	Reduce thiopurine dosage or consider alternative agent
Codeine sensitivity	*CYP2D6*	Avoid use of codeine; consider alternatives such as morphine and non-opioid analgesics
Simvastatin sensitivity	*SLCO1B1*	Reduce dose of simvastatin or consider an alternative statin; consider routine creatine kinase surveillance

Ashley (2016) Nature Review Genetics

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.

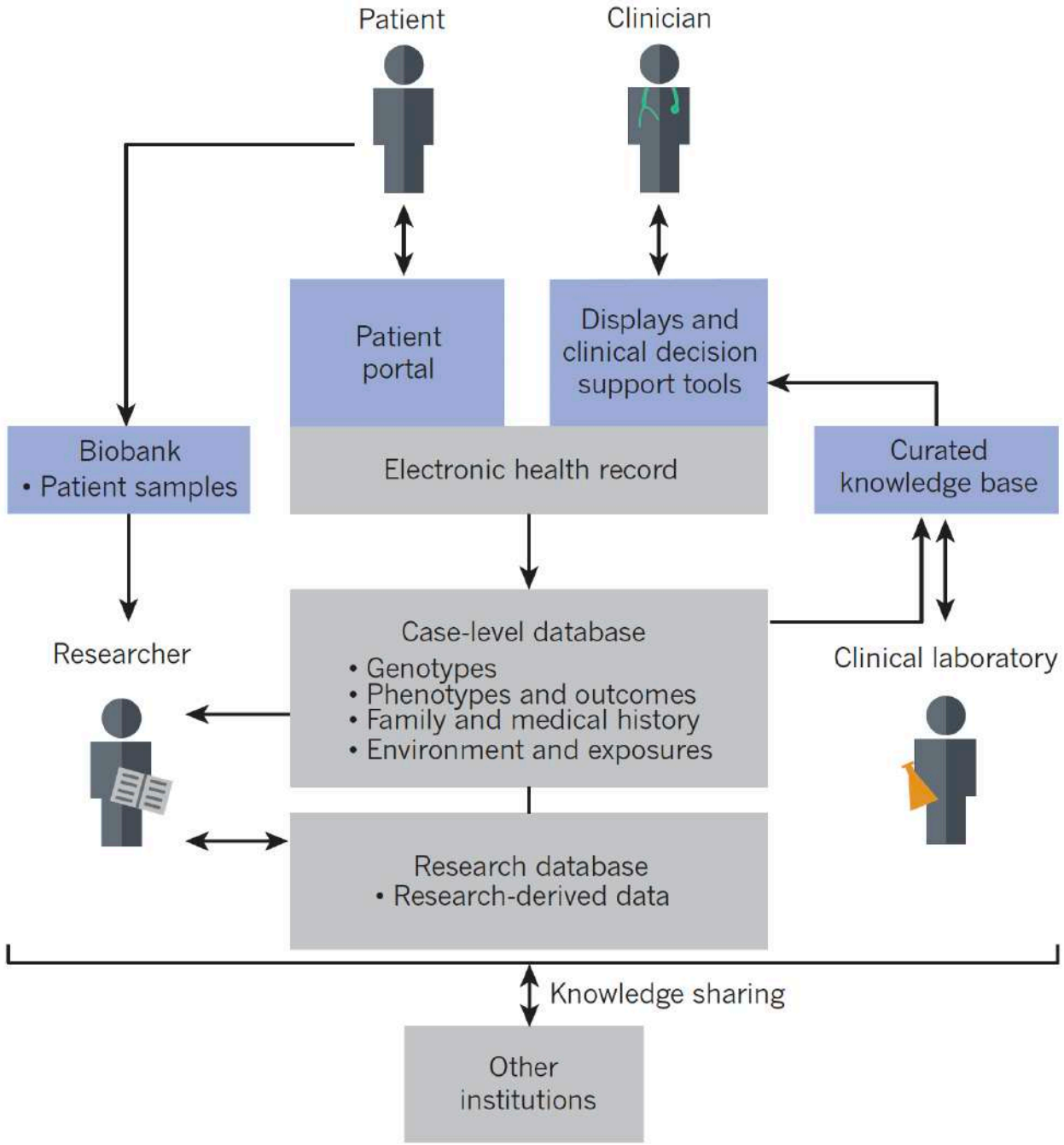
Morash *et al* (2018) Journal of Personalized Medicine

REVIEW

doi

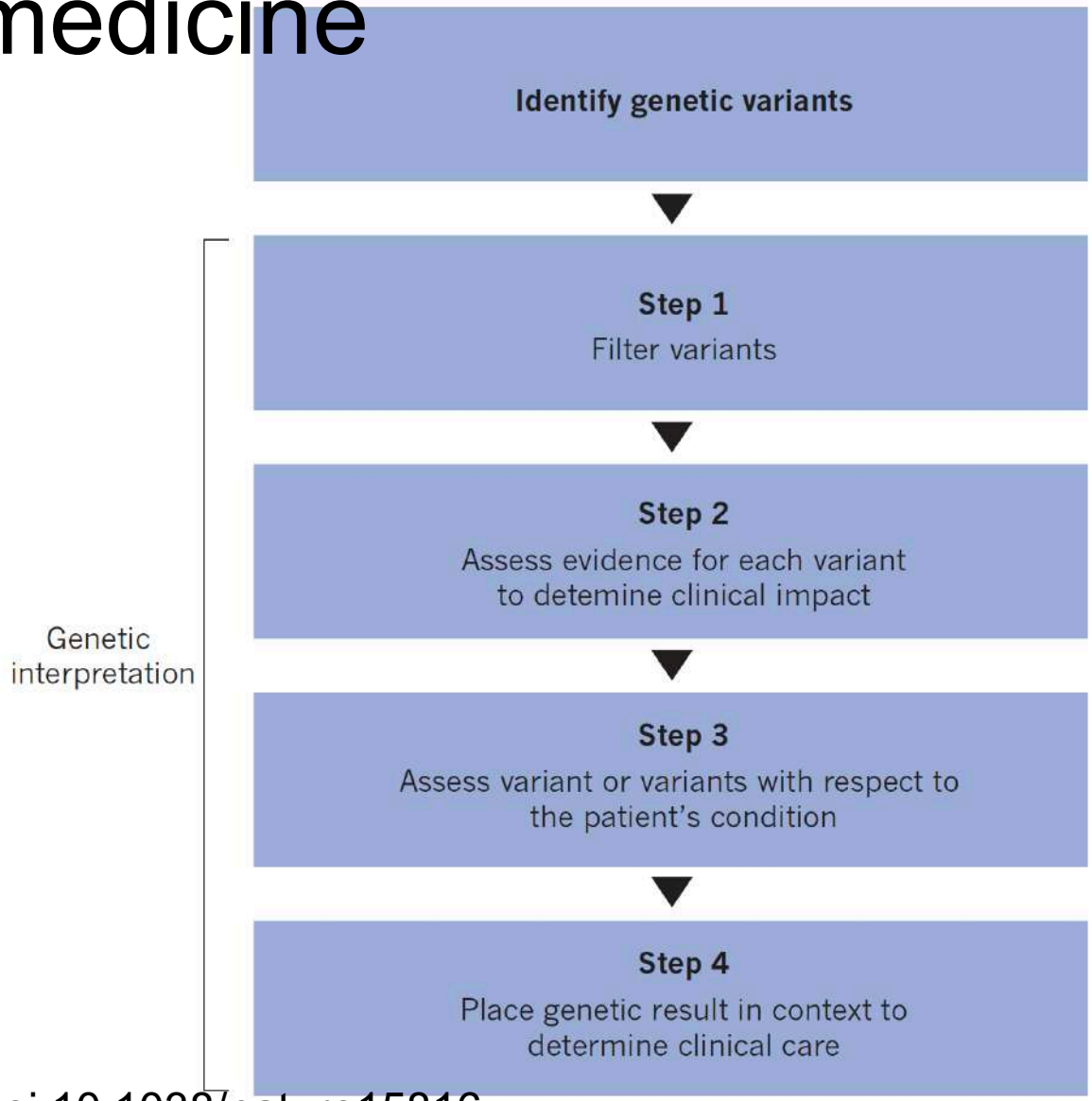
Building the foundation for genomics in precision medicine

Samuel J. Aronson[1,2] & Heidi L. Rehm[1,3,4,5]



doi:10.1038/nature15816

Building the foundation for genomics in precision medicine



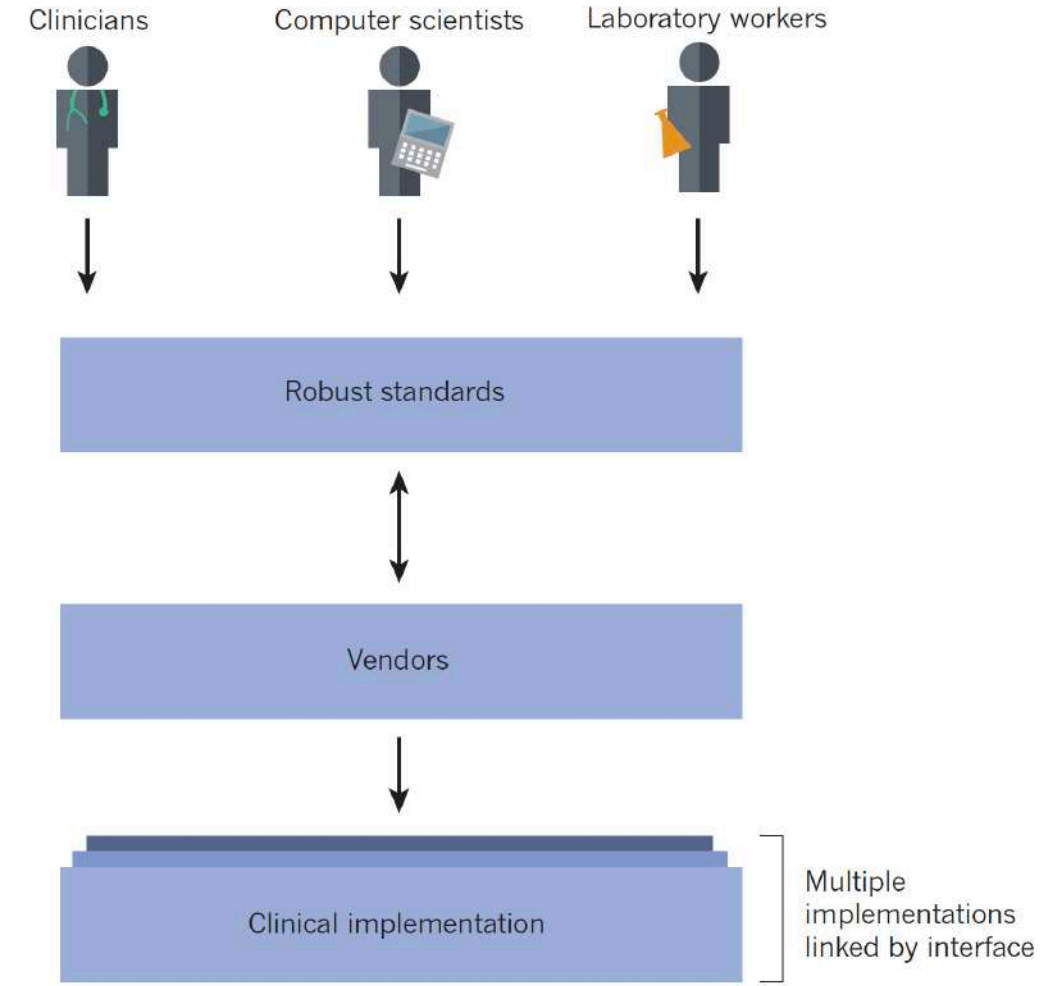
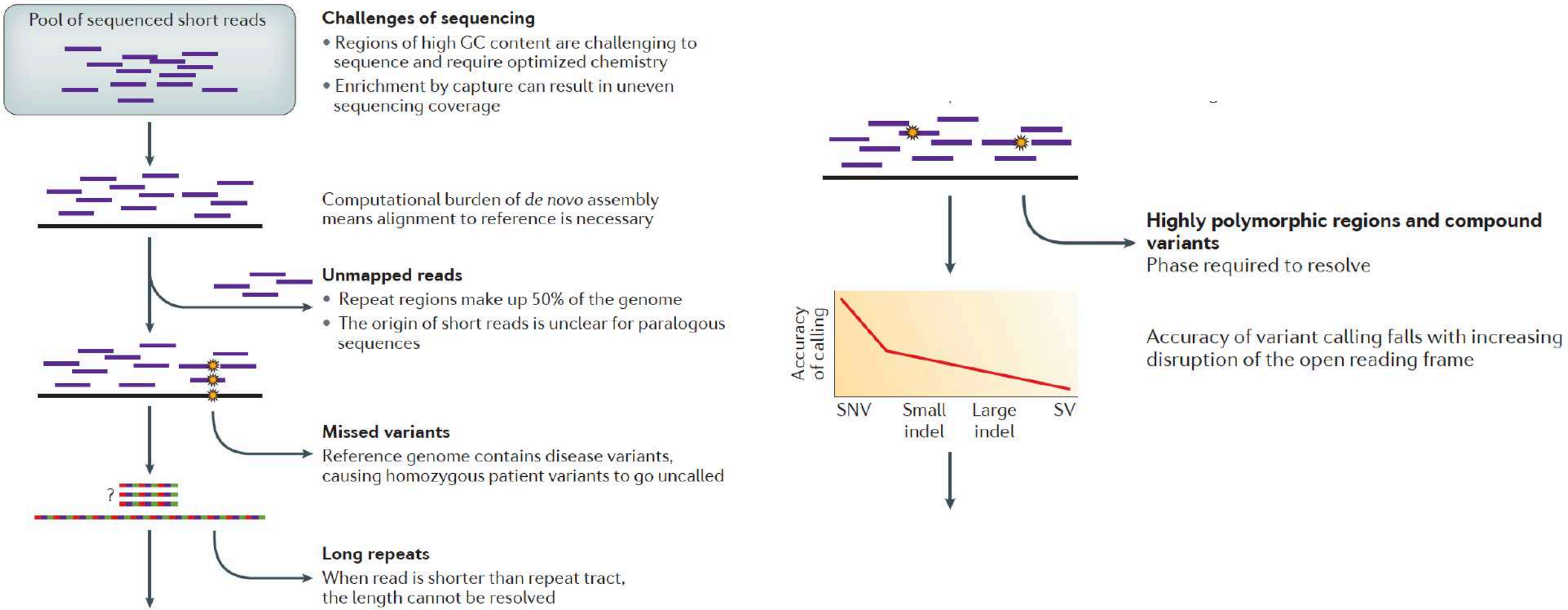


Figure 3 | Creating and implementing robust standards for the description and structuring of data in laboratory processing and patient-care systems.

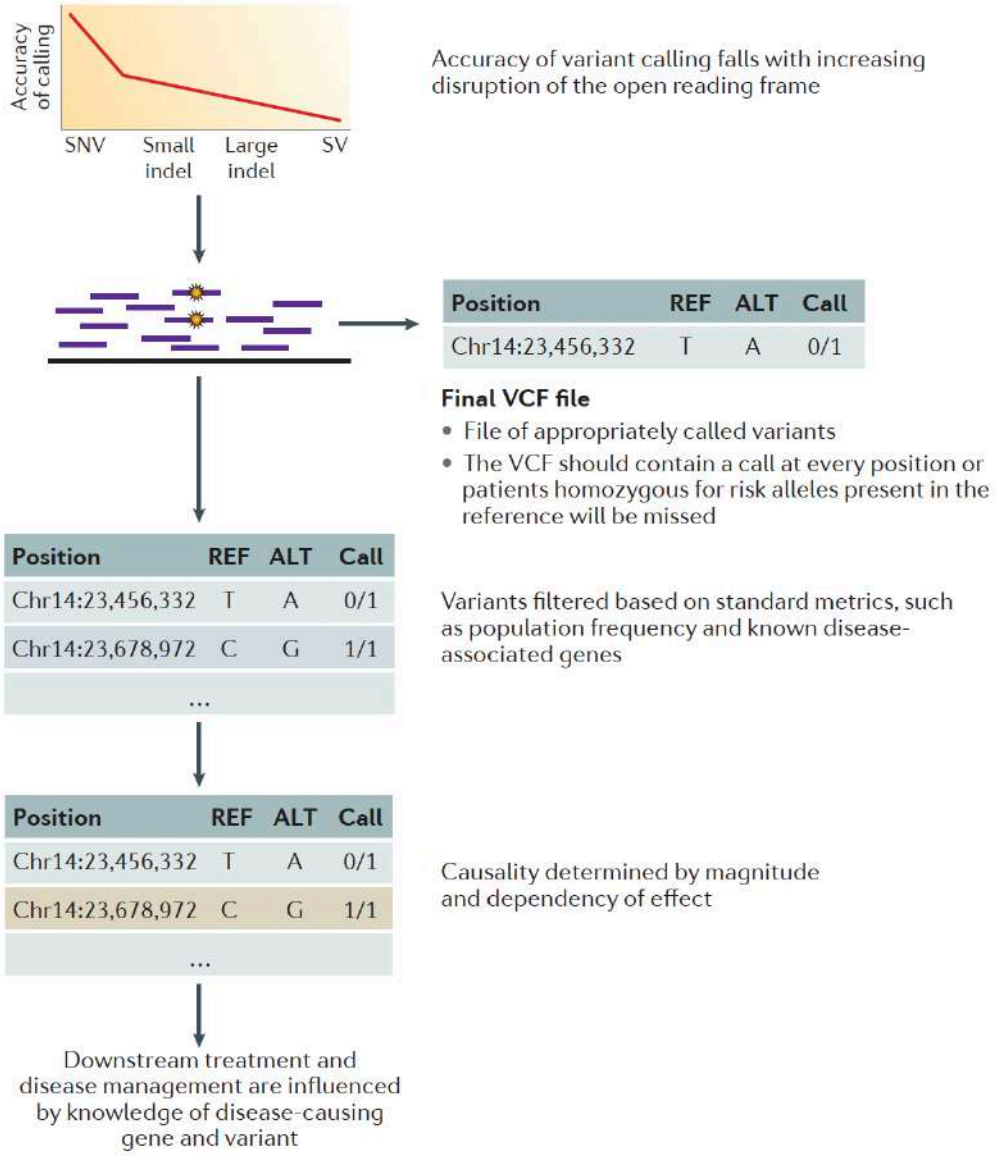
doi:10.1038/nature15816

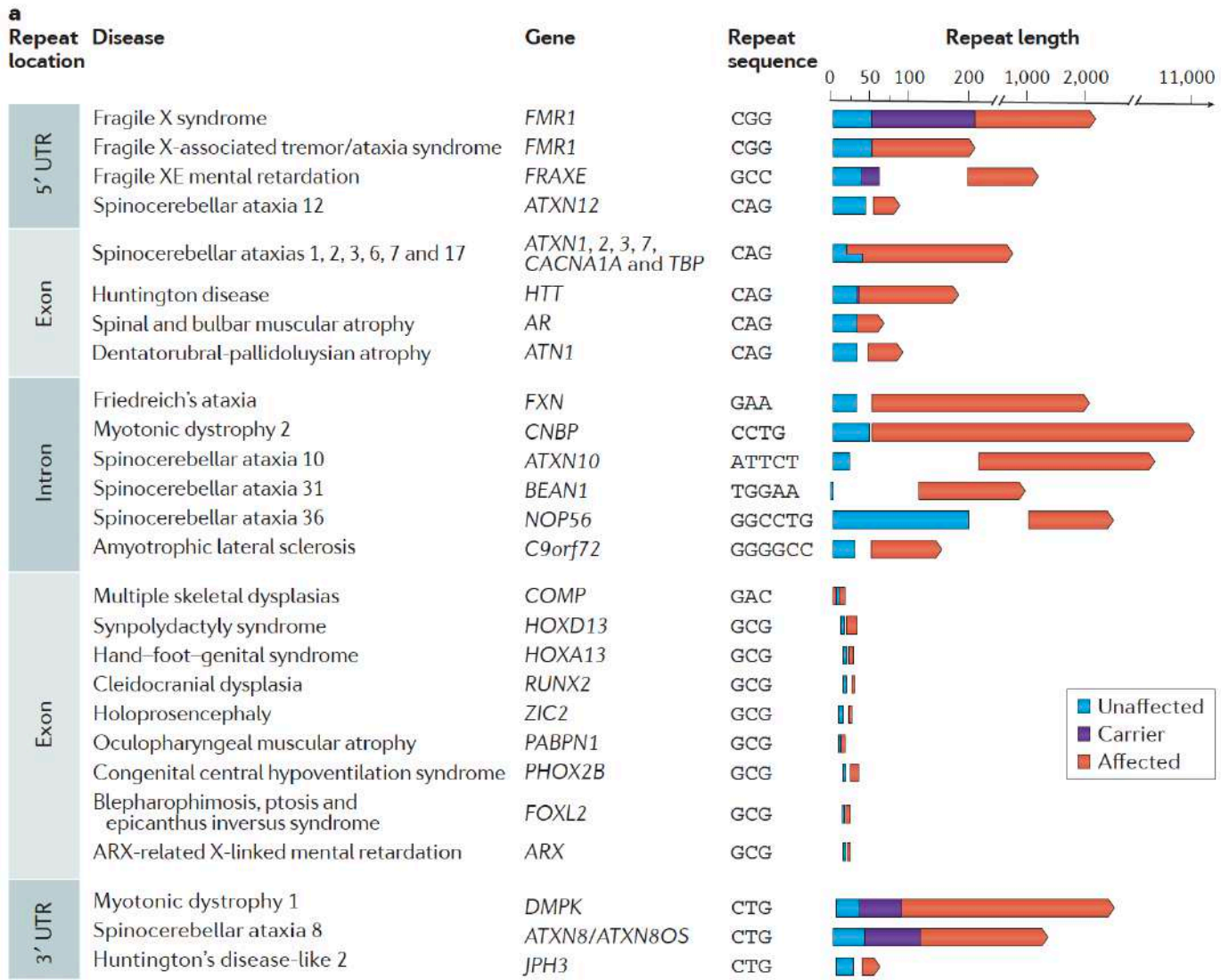
Challenges and reduced accuracies



Ashley (2016) Nature Review Genetics

Challenges and reduced accuracies

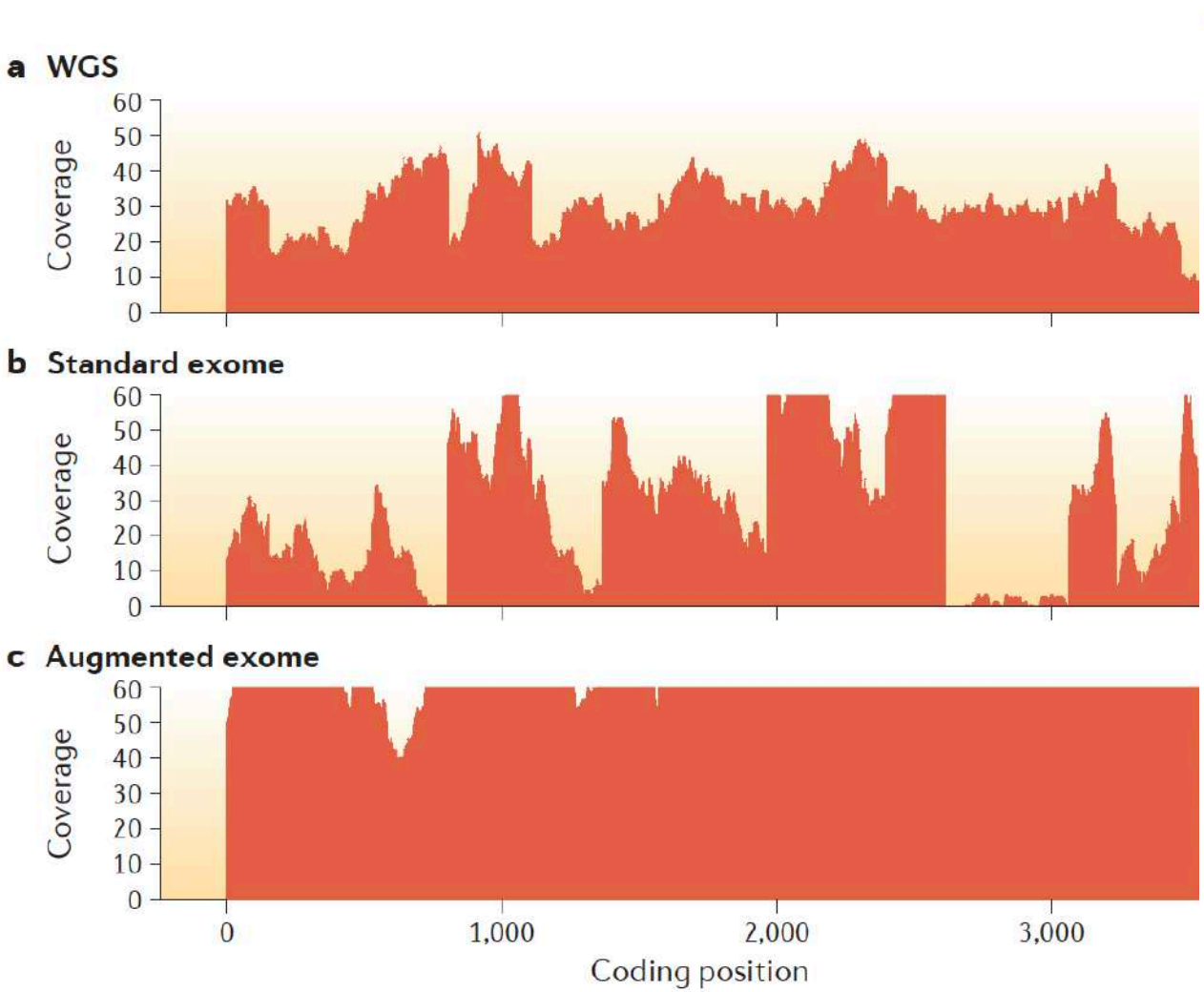


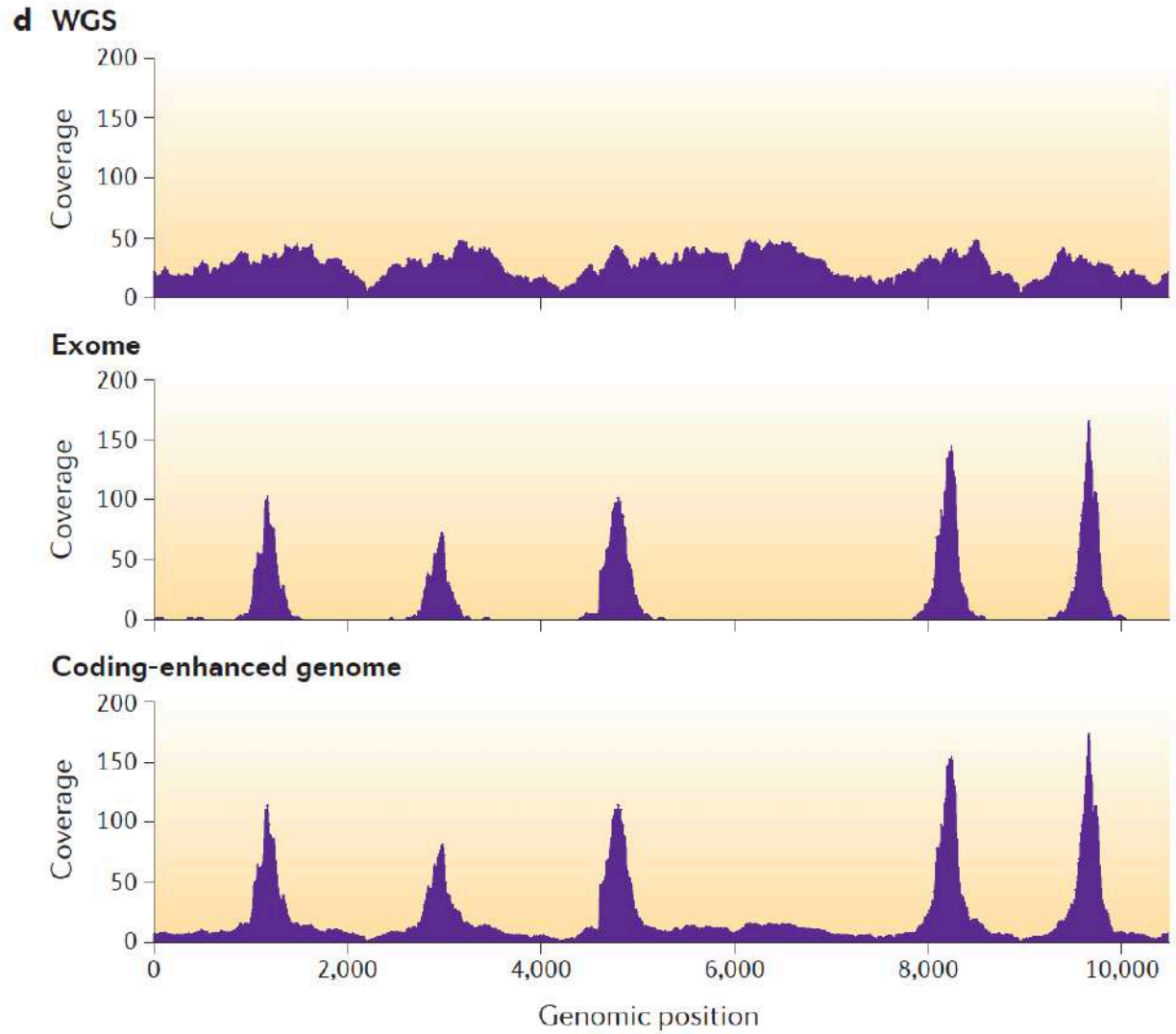


Ashley (2016) Nature Review Genetics

Resolution

Zoomed out





Ashley (2016) Nature Review Genetics

Summary:

- A genome, a few genomes are no longer “enough”
  - ~since everybody can do it reasonably well
- Genome sequencing projects are
  - being done on a per-lab basis and no longer exclusive to sequencing centers
  - moving away from exploration to question orientated.
- Data being produced on a **much faster speed** at a **much higher throughput**, and a much **cheaper scale**

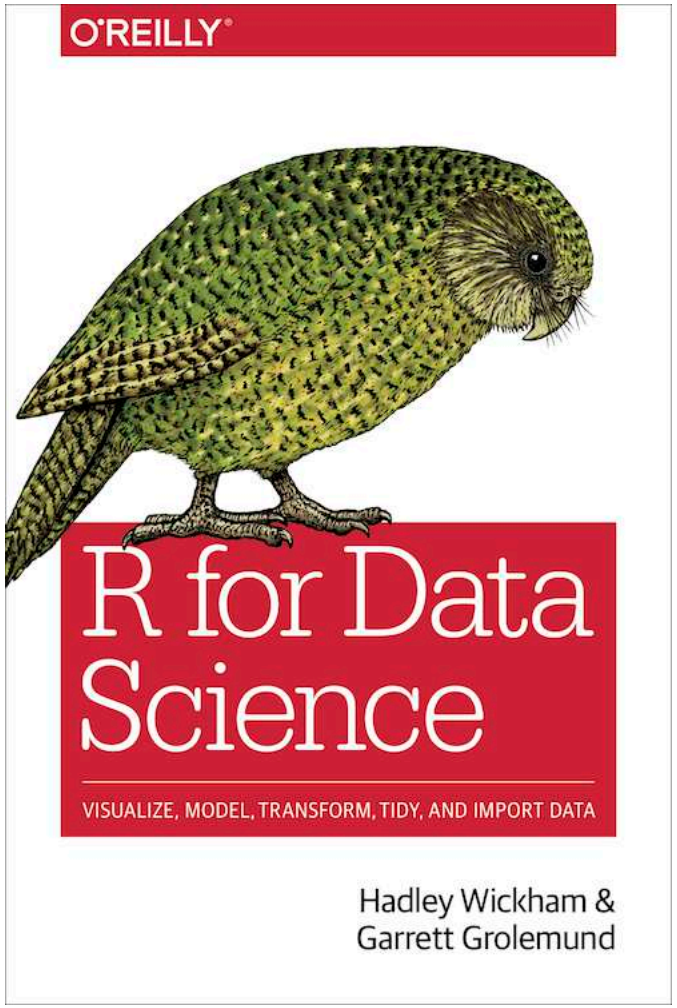
- More methods, analysis, tools, experiments…
  - Not always better

It is an exciting time to be in

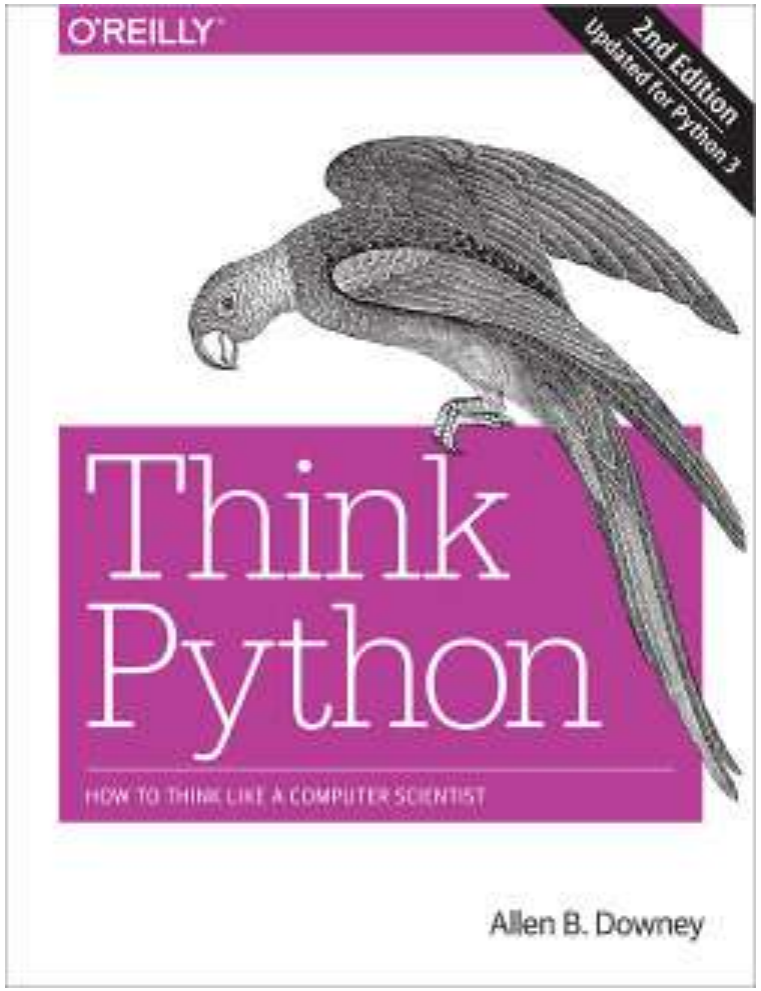
Next two lectures:

- Metagenomics
- Transcriptomics / RNAseq

Resources



https://r4ds.had.co.nz/



https://greenteapress.com/wp/think-python-2e/

Resources

Some very useful websites:

陳倩瑜老師 BioDataMining
https://www.youtube.com/channel/UCBIt6qJh3XA8wshVaw6-hVg

Next-Gen Sequence Analysis Workshop [UC Davis]
https://angus.readthedocs.io/en/2019/

Workshop
http://evomics.org/

Lecures: Computational Genomics: Applied Comparative Genomics
https://github.com/schatzlab/appliedgenomics2019

Written assignment

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