Lecture 22. Metagenomics

Michael Schatz

April 15, 2020

JHU 600.749: Applied Comparative Genomics

Preliminary Project Report

Assignment Date: March 30, 2019
Due Date: Monday, April 13, 2019 @ 11:59pm

Each team should submit a PDF of your preliminary project proposal (2 to 3 pages) to GradeScope by 11:59pm on Monday April 13.

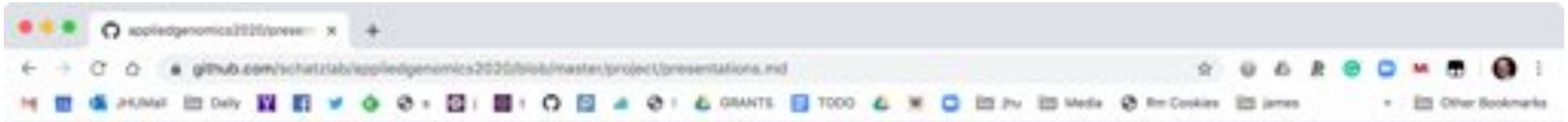
The preliminary report should have at least:

- Title of your project
- List of team members and email addresses
- 1 paragraph abstract summarizing the project
- 1+ paragraph of Introduction
- 1+ paragraph of Methods that you are using
- 1+ paragraph of Results, describing the data evaluated and any any preliminary results
- 1+ paragraph of Dicsussion (what you have seen or expect to see)
- 1+ figure showing a preliminary result
- 5+ References to relevant papers and data

The preliminary report should use the Bioinformatics style template. Word and LaTeX templates are available at https://academic.oup.com/bioinformatics/pages/submission_online. Overleaf is recommended for LaTex submissions. Google Docs is recommended for non-latex submissions, especially group projects. Paperpile is recommended for citation management.

Later, you will present your project in class starting the week of April 22. You will also submit your final written report (5-7 pages) of your project by May 13

Please use Piazza if you have any general questions!



JHU EN.600.749: Computational Genomics: Applied Comparative Genomics

Project Presentations

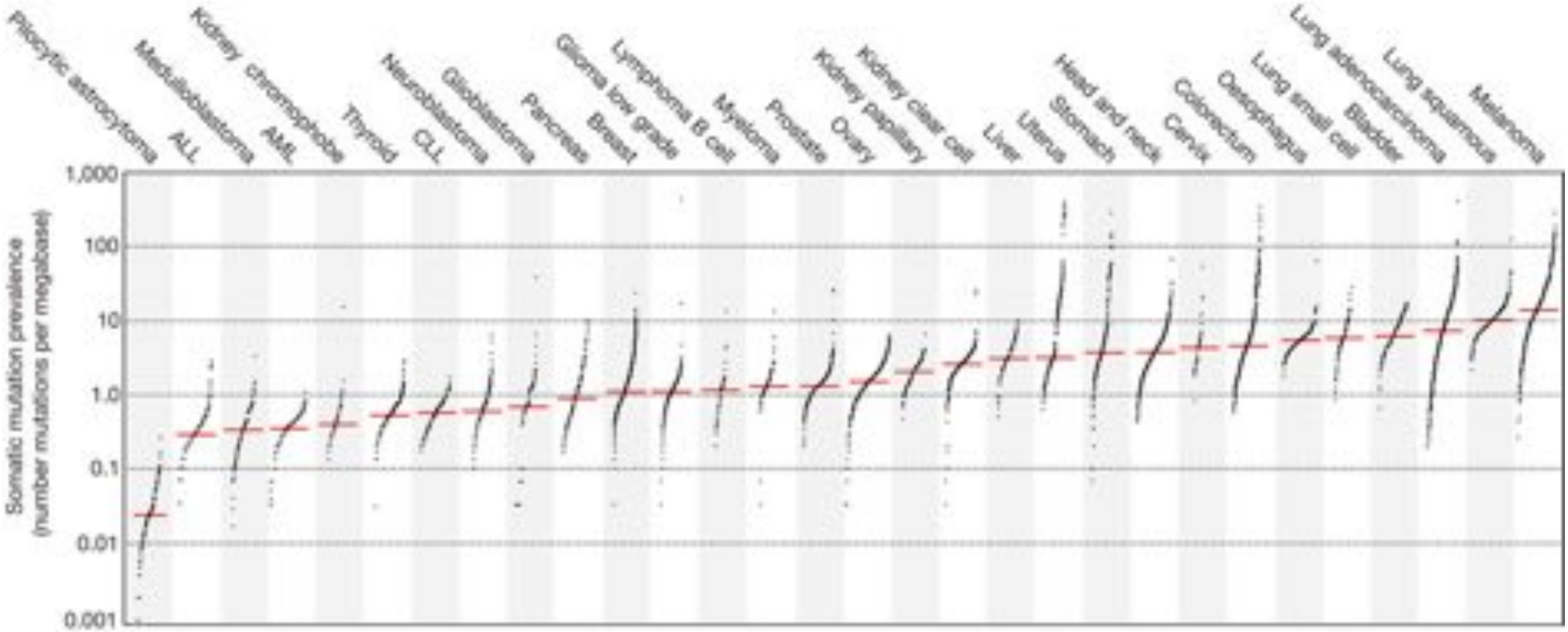
Presentations will be a total of 20 minutes: 15 minutes for the presentation, followed by 5 minutes for questions. We will strictly keep to the schedule to ensure that all groups can present in class!

Schedule of Presentations

Day	Time	Team Name	Students	Title
Wed 4/22	1:30 - 1:50	Predict enhancer-promoter interactions	Sandeep Kambhampati, Kevin Zhan, Tatiana Gelaf	Using deep learning approaches on DNA sequence and DNA methylation data to predict enhancer-promoter interactions
Wed 4/22	1:50 - 2:10	Team Cao	Hongyu Cao	Benchmarking variant calling algorithms and performance
Wed 4/22	2:10 - 2:30	SAMtools	Samantha Zarate, April Kim, Michelle Shu	Phylogenetic and Comparative Analysis of SARS-CoV-2
Mon 4/27	1:30 - 1:50	Two-Step Project	Lukas Voortman	Determining the generality of the two-step mechanism in the Drosophila genome
Mon 4/27	1:50 - 2:10	Gviz	Ebenezer Armah	Genomic Data Visualization
Mon 4/27	2:10 - 2:30	ByOhmPho	Louis (Jinrui) Liu, Yijun Li	Assess the performance of Monocle Algorithm
Wed 4/29	1:30 - 1:50	Metagenomics Team	Harrison Huh, Qing Dai, Victor Wang	CNN approach to metagenomics

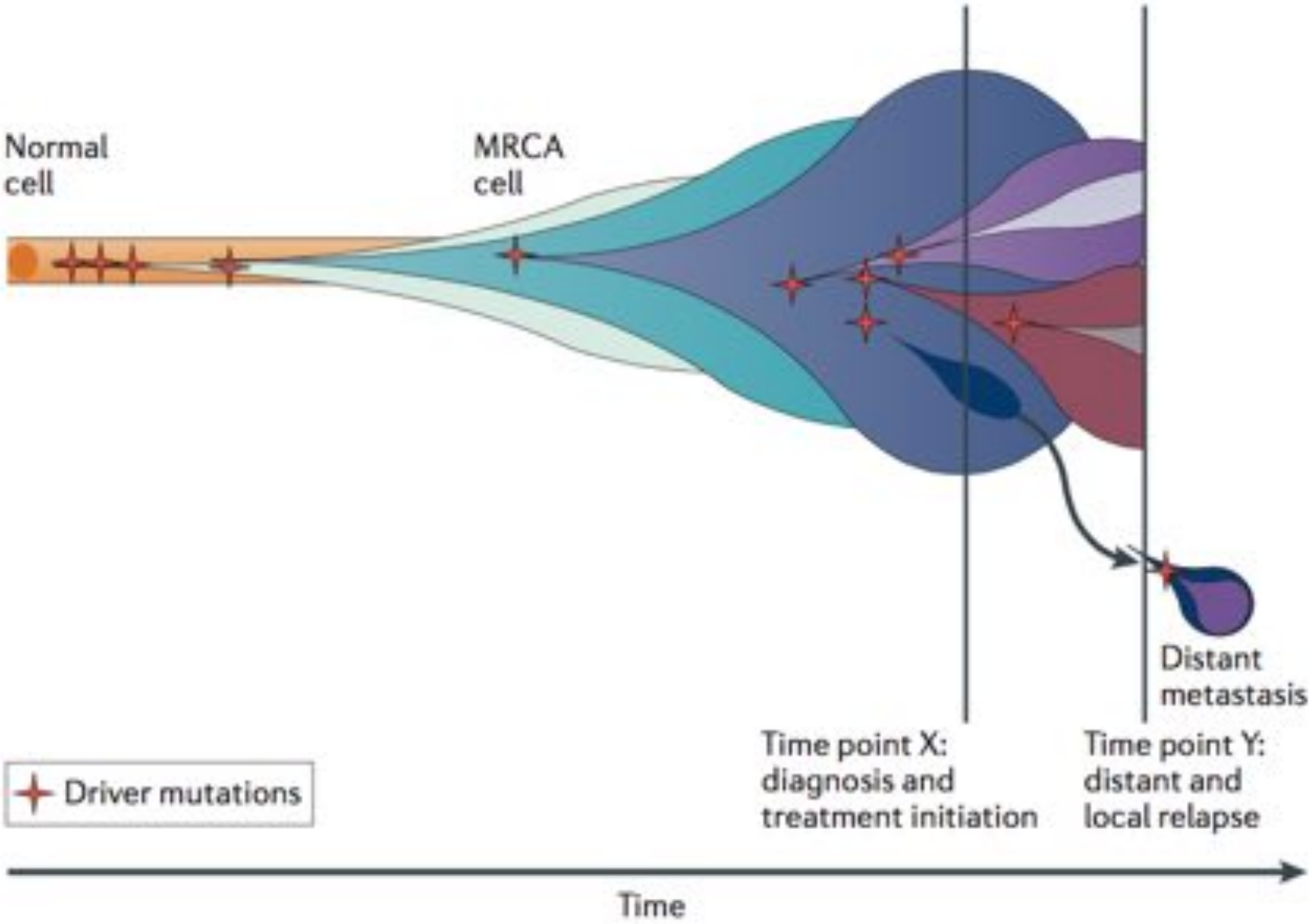
Part 1:
Cancer Genetics

Somatic Mutations In Cancer



Signatures of mutational processes in human cancer
Alexandrov et al (2013) *Nature*. doi:10.1038/nature12477

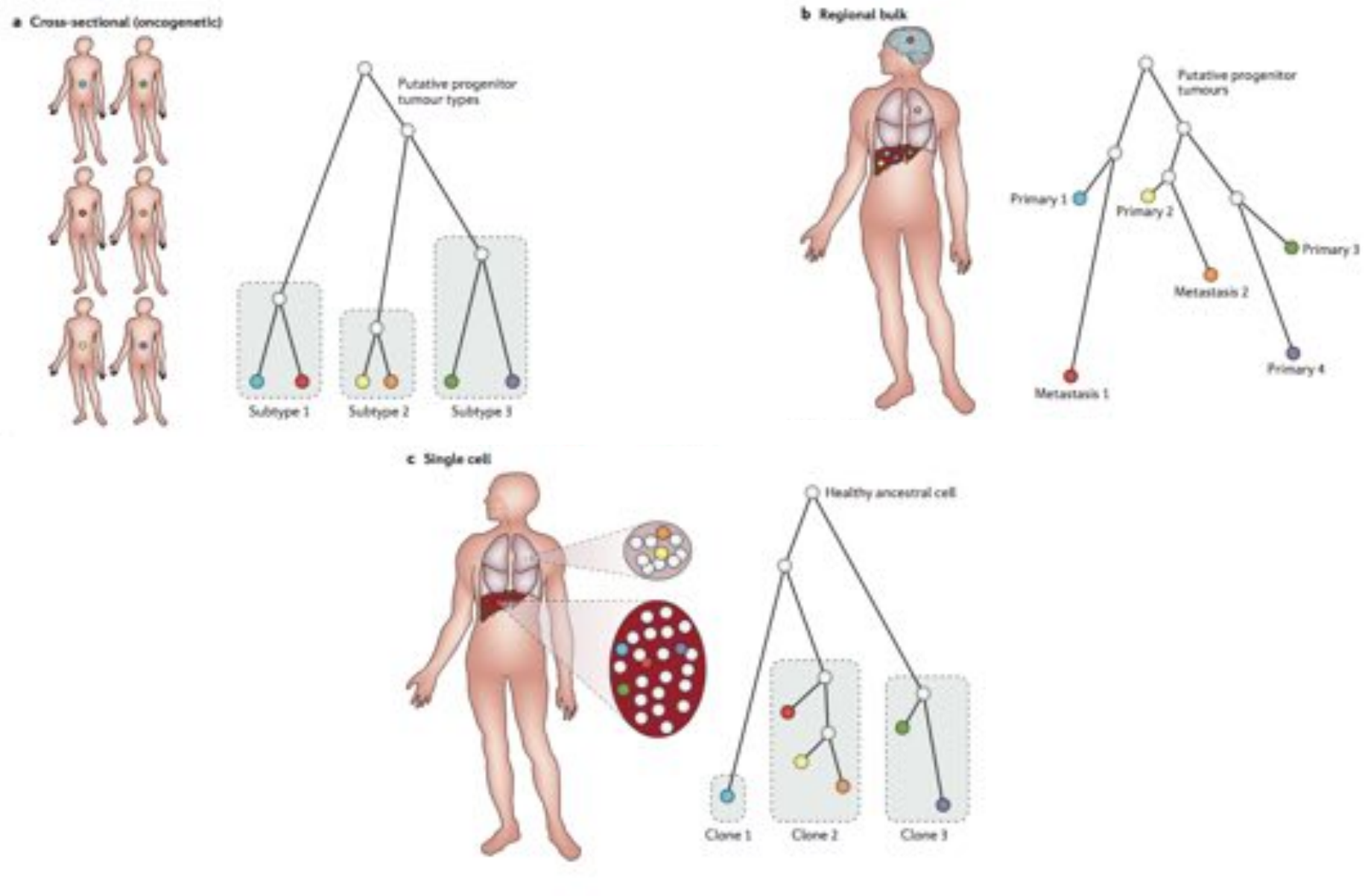
Tumor Evolution



Evolution of the cancer genome

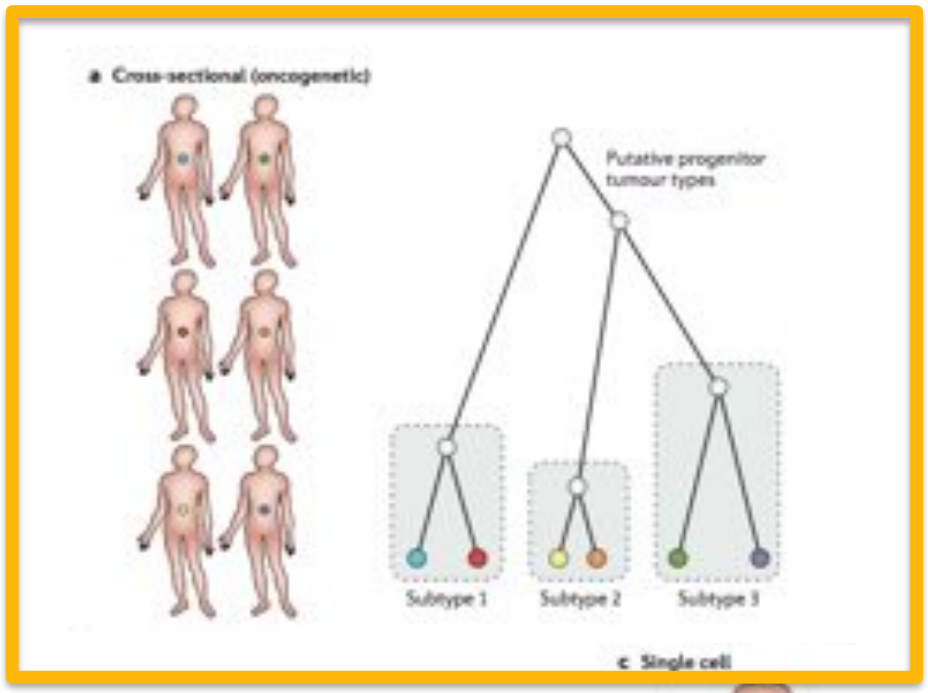
Yates & Campbell (2012) Nature Review Genetics. doi:10.1038/nrg3317

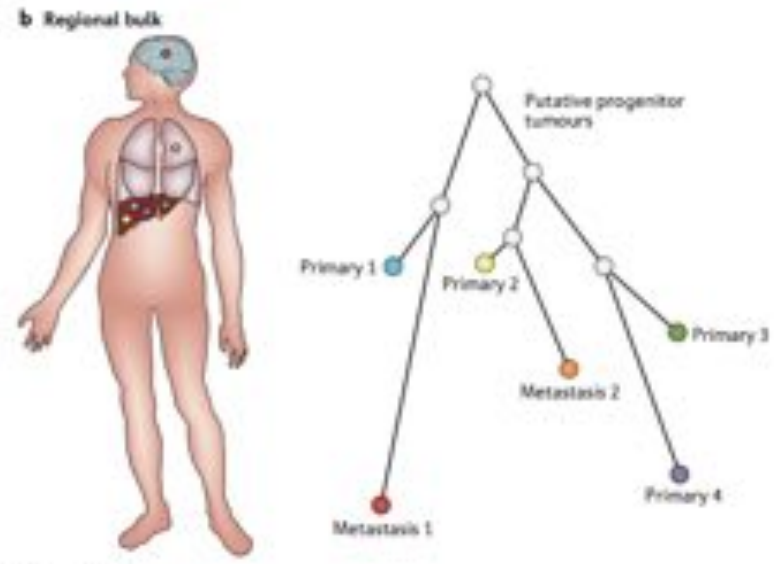
Tumor Heterogeneity

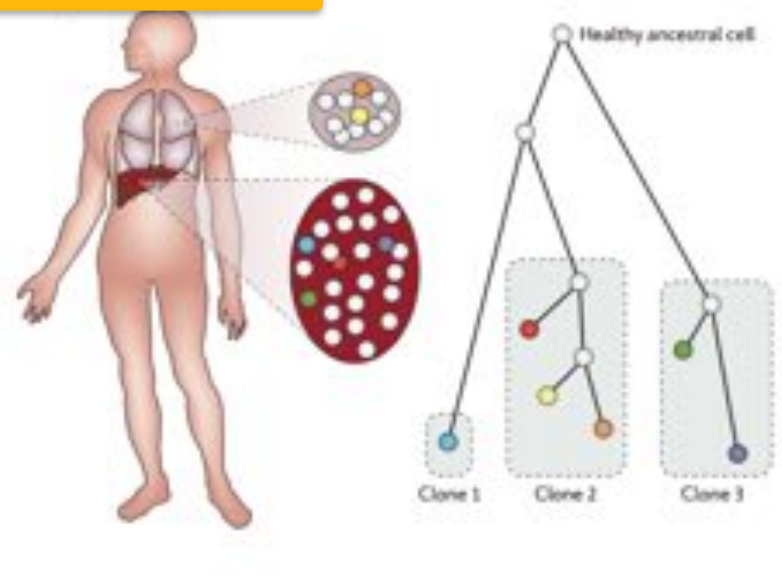


The evolution of tumour phylogenetics: principles and practice
Schwarz and Schaffer (2017) *Nature Reviews Genetics. doi:10.1038/nrg.2016.170*

Tumor Heterogeneity

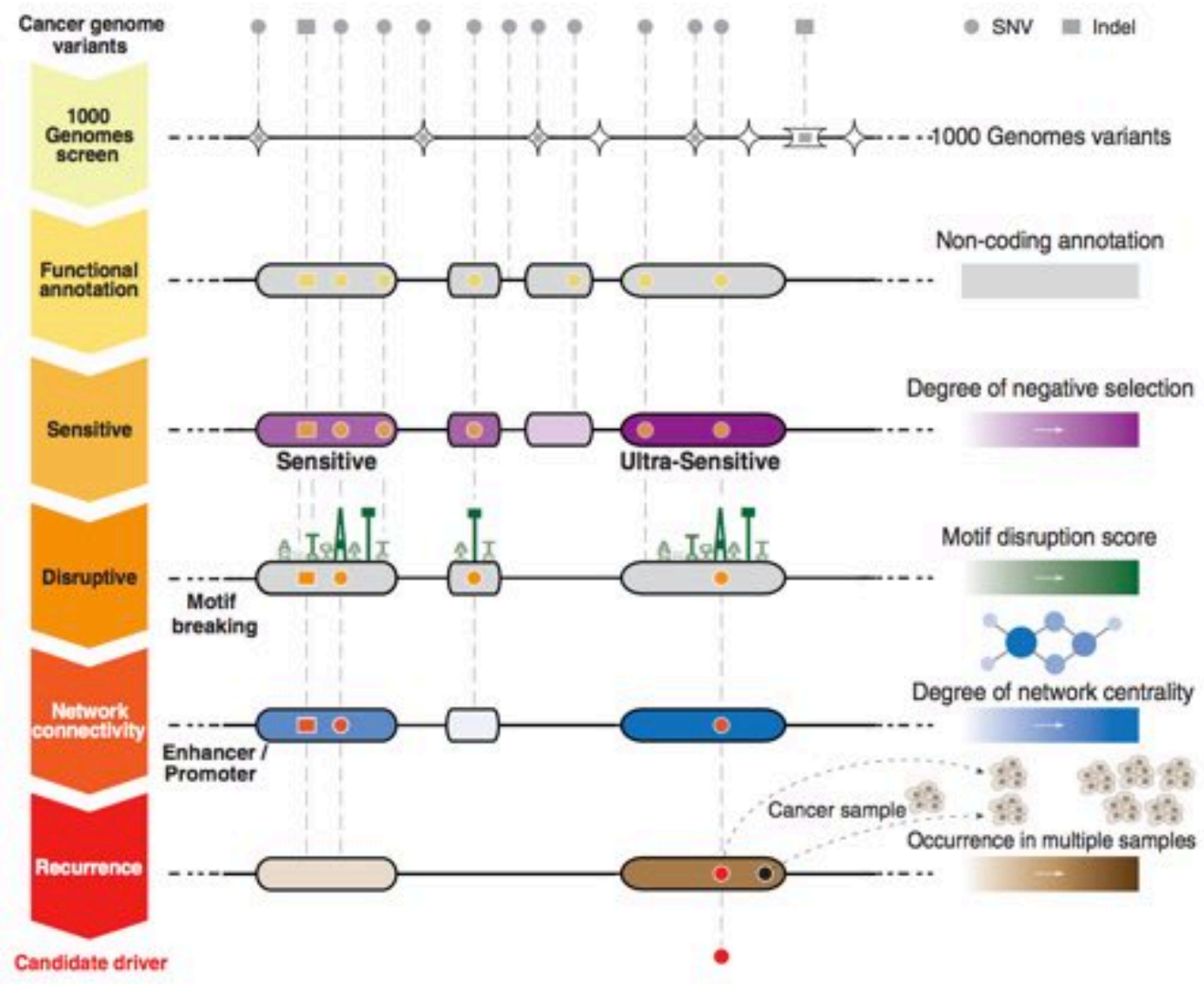






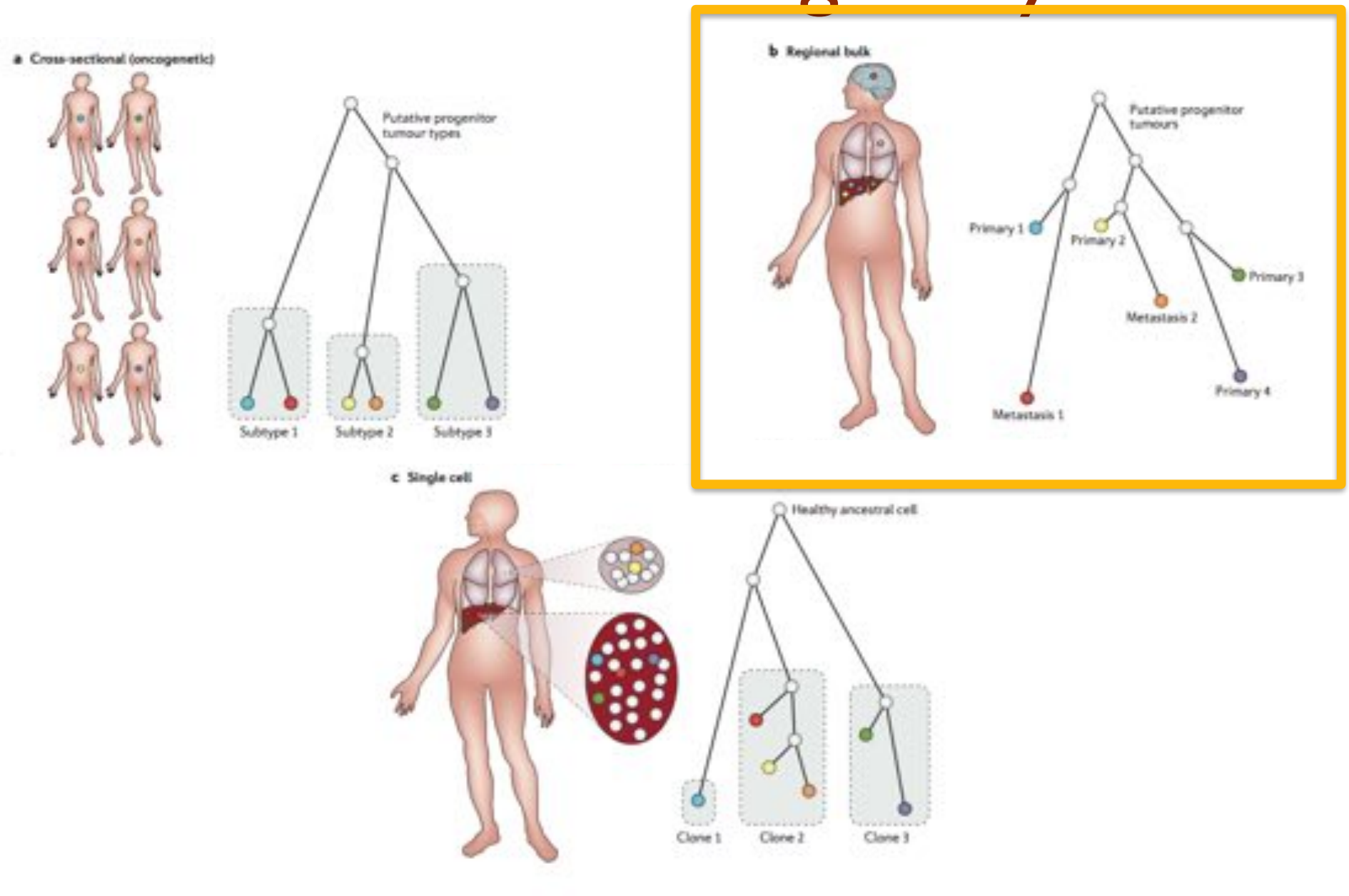
The evolution of tumour phylogenetics: principles and practice
Schwarz and Schaffer (2017) *Nature Reviews Genetics. doi:10.1038/nrg.2016.170*

Finding Driving Mutations



Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics
Khurana et al (2013) Science. DOI: 10.1126/science.1235587

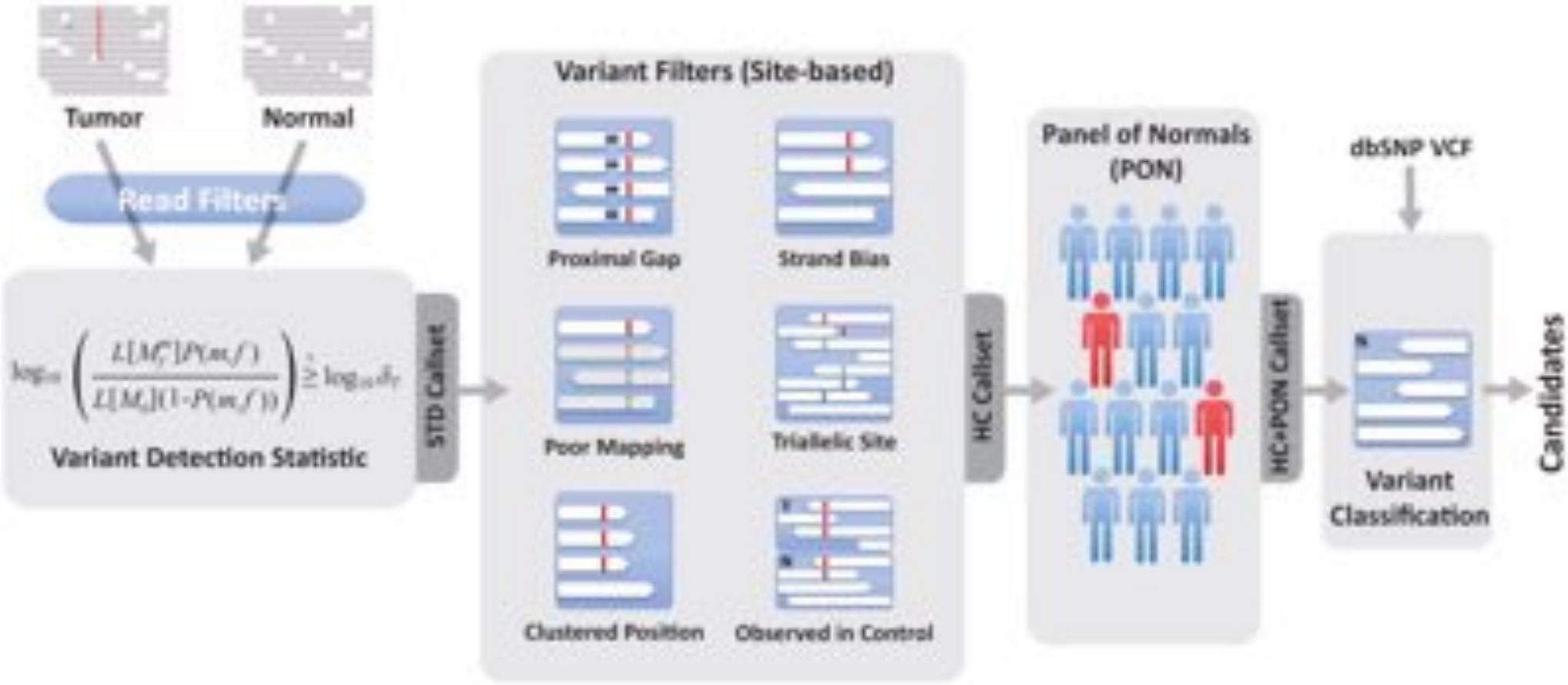
Tumor Heterogeneity



The evolution of tumour phylogenetics: principles and practice

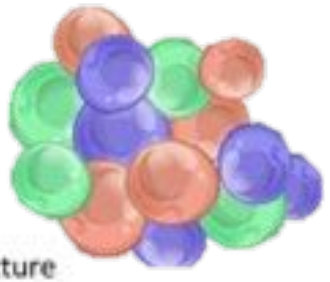
Schwarz and Schaffer (2017) *Nature Reviews Genetics. doi:10.1038/nrg.2016.170*

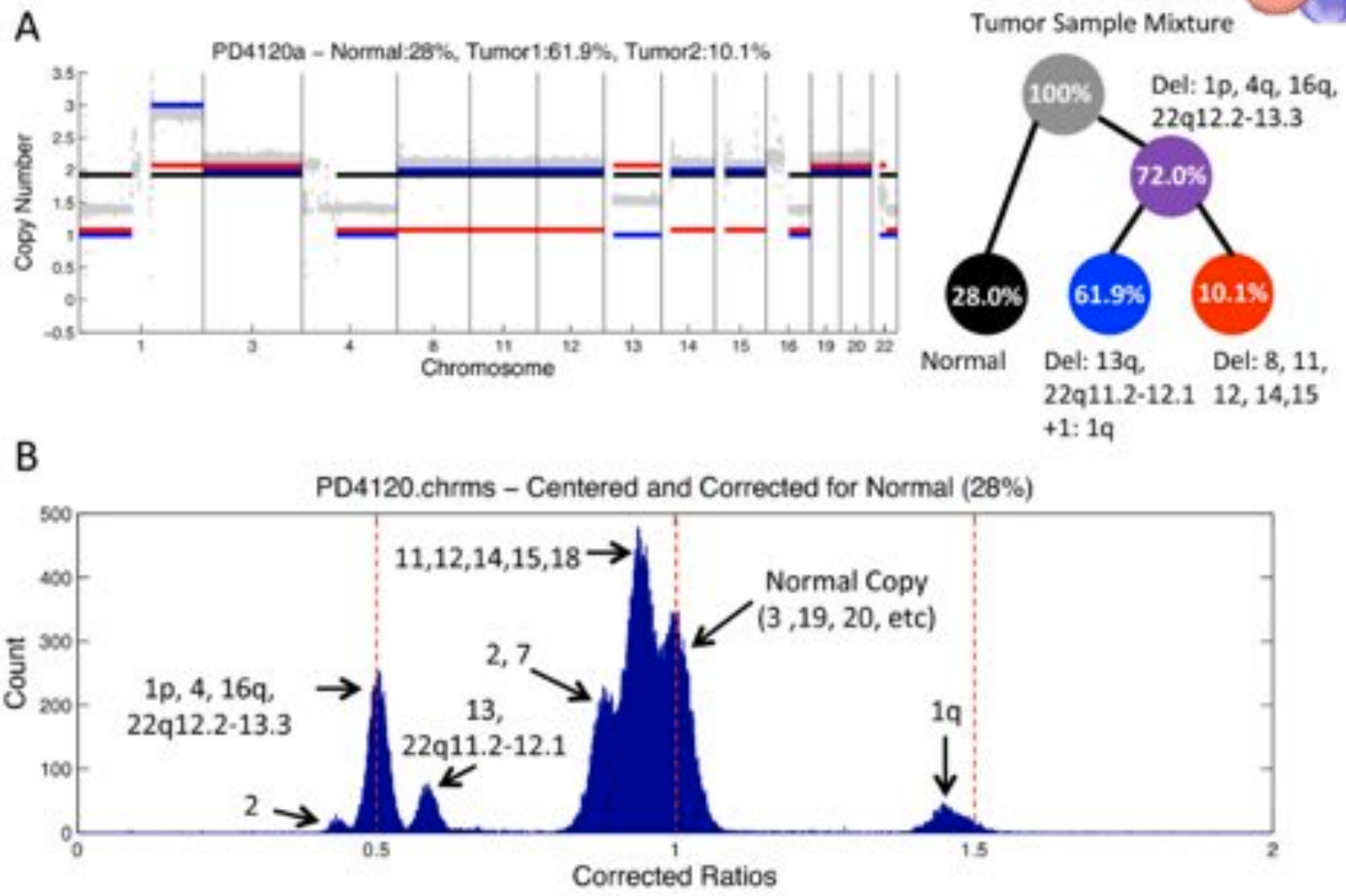
Tumor-Normal Pairs



Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples
Cibulskis et al (2013) Nature Biotech. doi:10.1038/nbt.2514

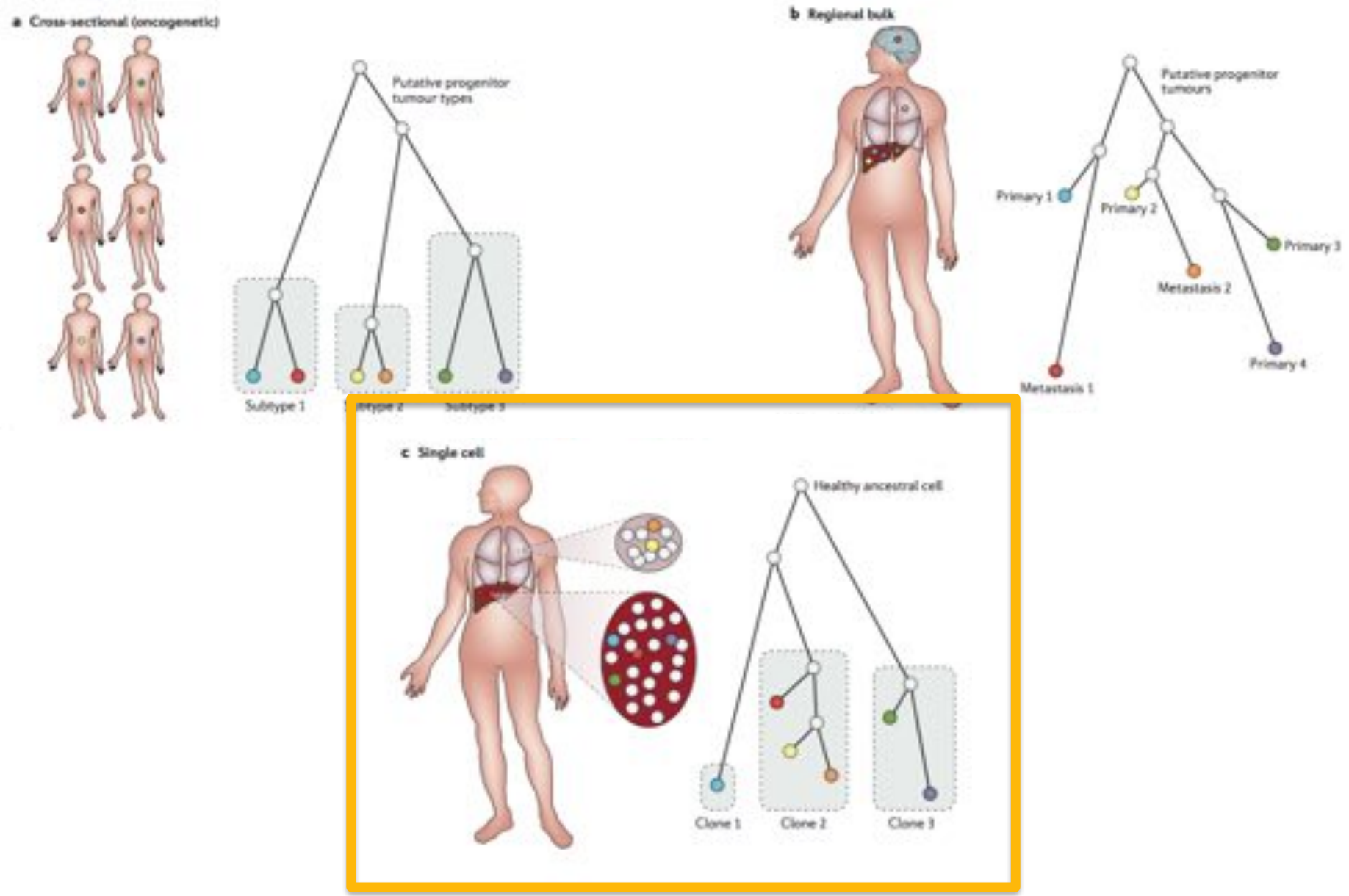
Bulk Heterogeneity





THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data
Oesperet al (2013) Genome Biology. DOI: 10.1186/gb-2013-14-7-r80

Tumor Heterogeneity



The evolution of tumour phylogenetics: principles and practice

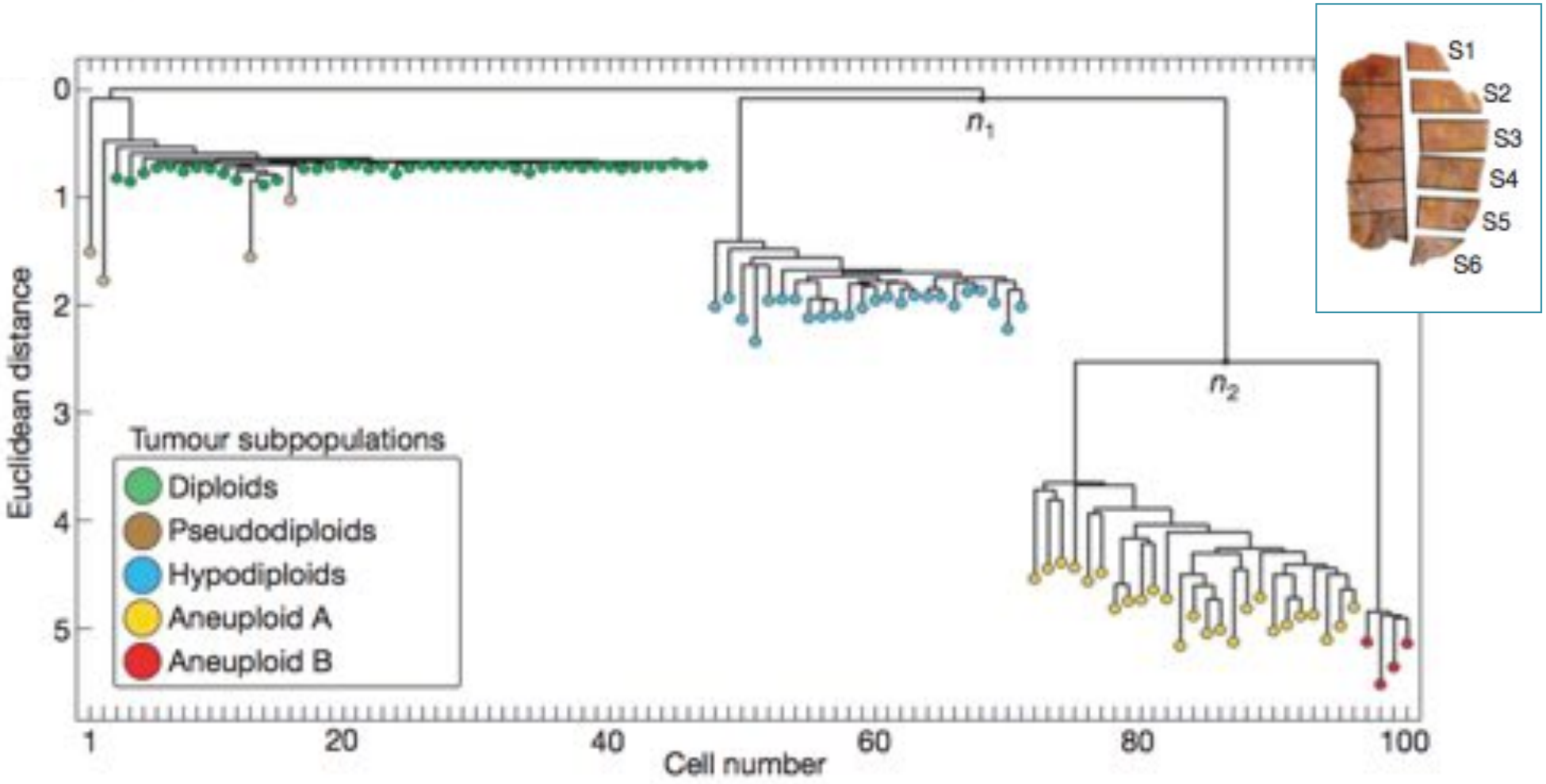
Schwarz and Schaffer (2017) *Nature Reviews Genetics. doi:10.1038/nrg.2016.170*

LETTER

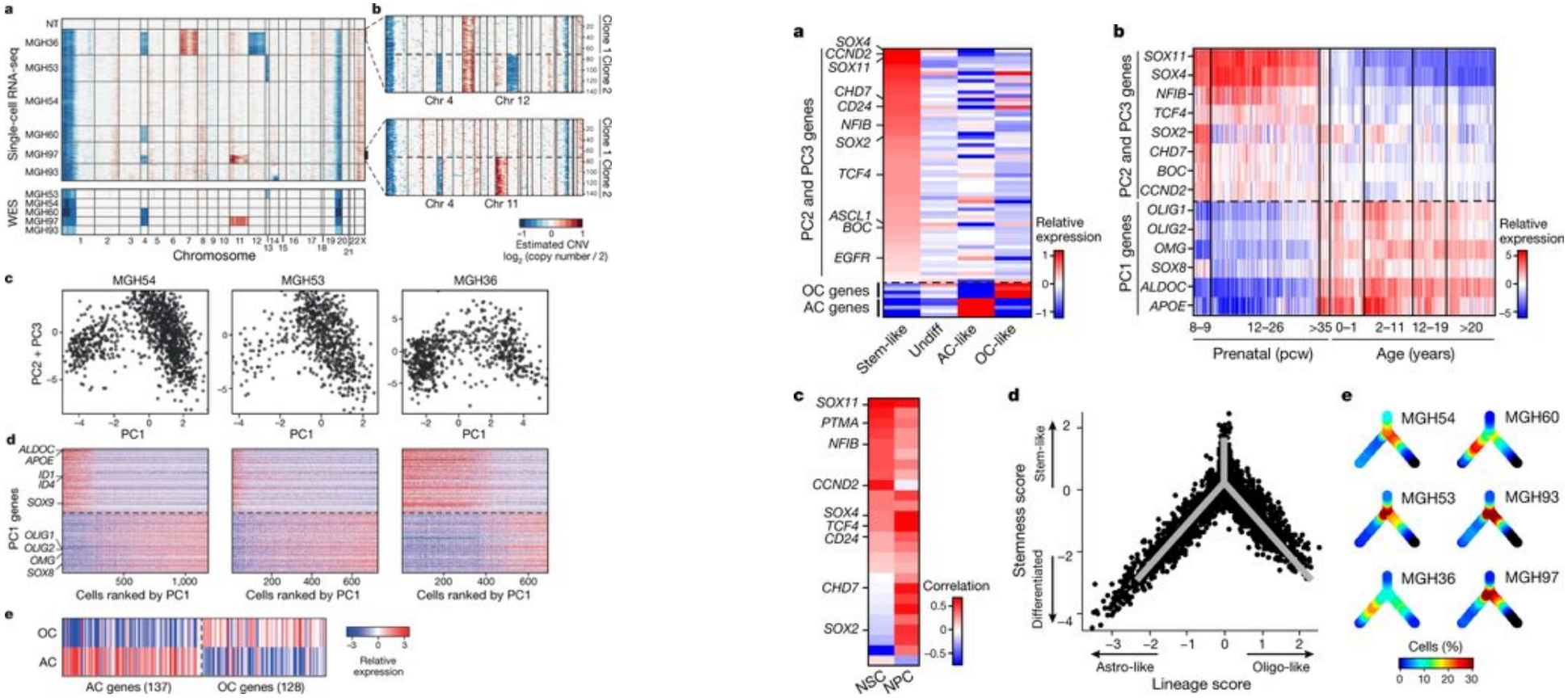
doi:10.1038/nature09807

Tumour evolution inferred by single-cell sequencing

Nicholas Navin[1,2], Jude Kendall[1], Jennifer Troge[1], Peter Andrews[1], Linda Rodgers[1], Jeanne McIndoo[1], Kerry Cook[1], Asya Stepansky[1], Dan Levy[1], Diane Esposito[1], Lakshmi Muthuswamy[3], Alex Krasnitz[1], W. Richard McCombie[1], James Hicks[1] & Michael Wigler[1]

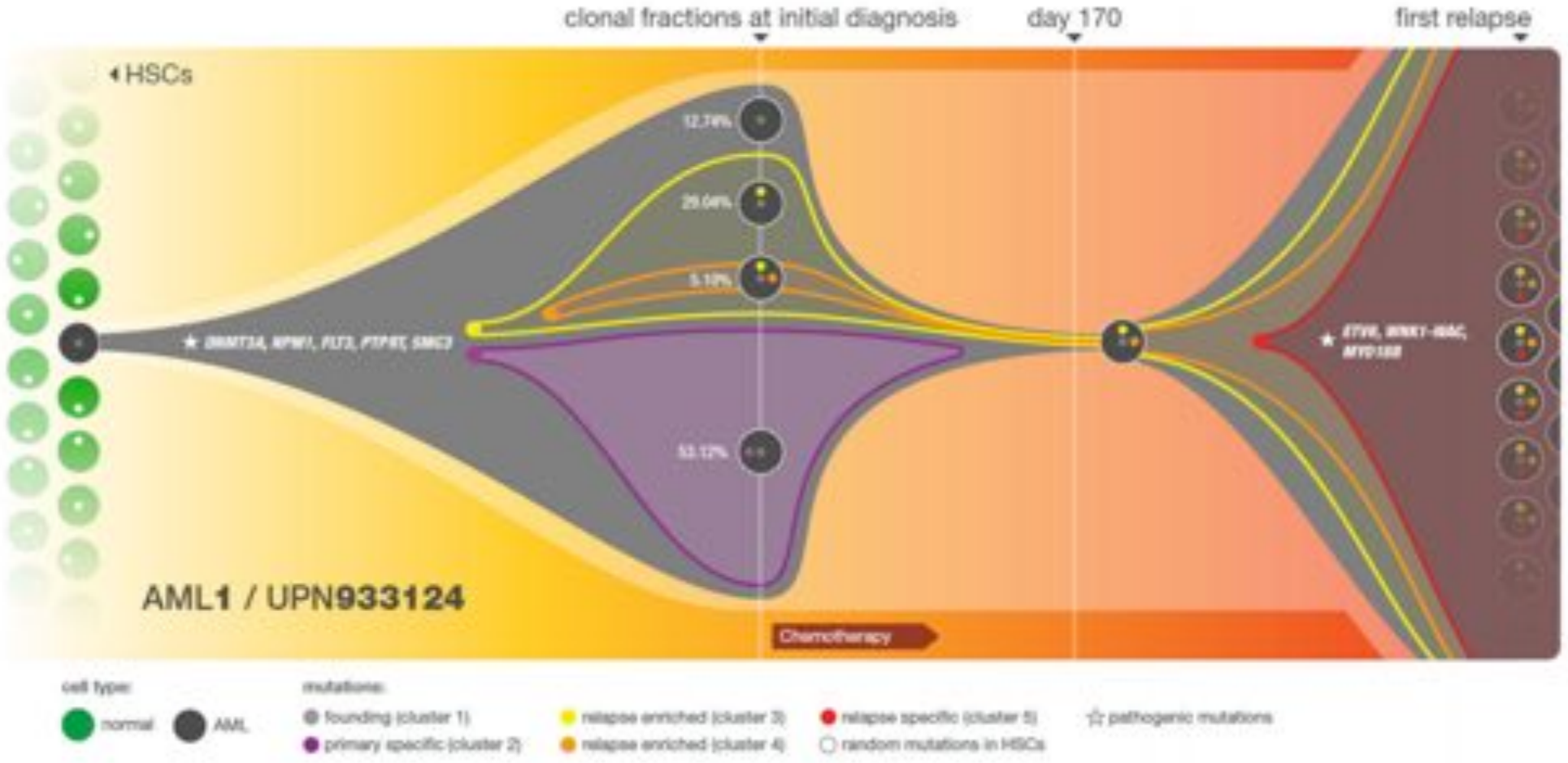


Single Cell RNA-seq of Cancer



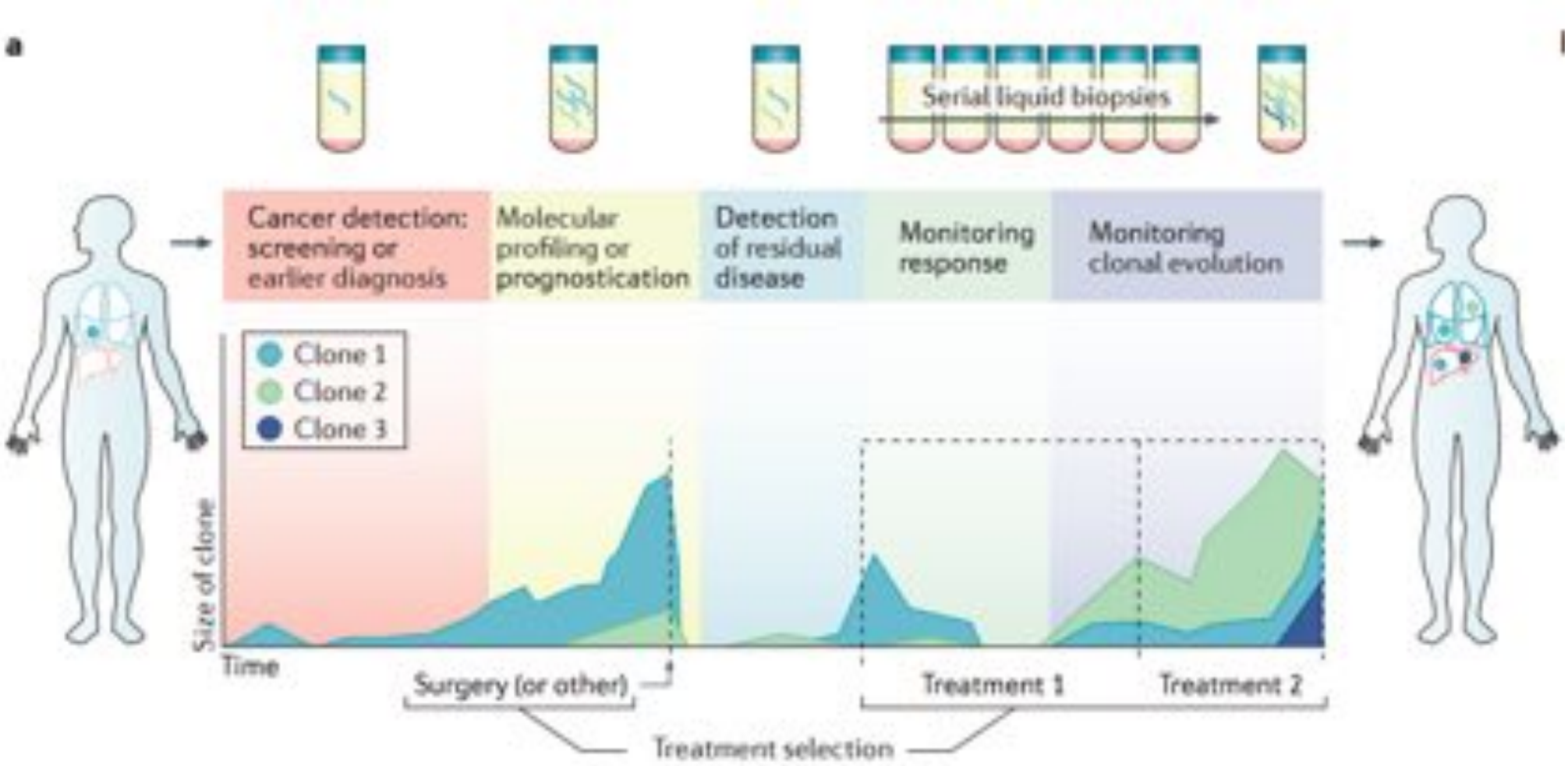
Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma
Tirosh et al (2016) Nature. doi:10.1038/nature20123

Tumor Heterogeneity and Treatment



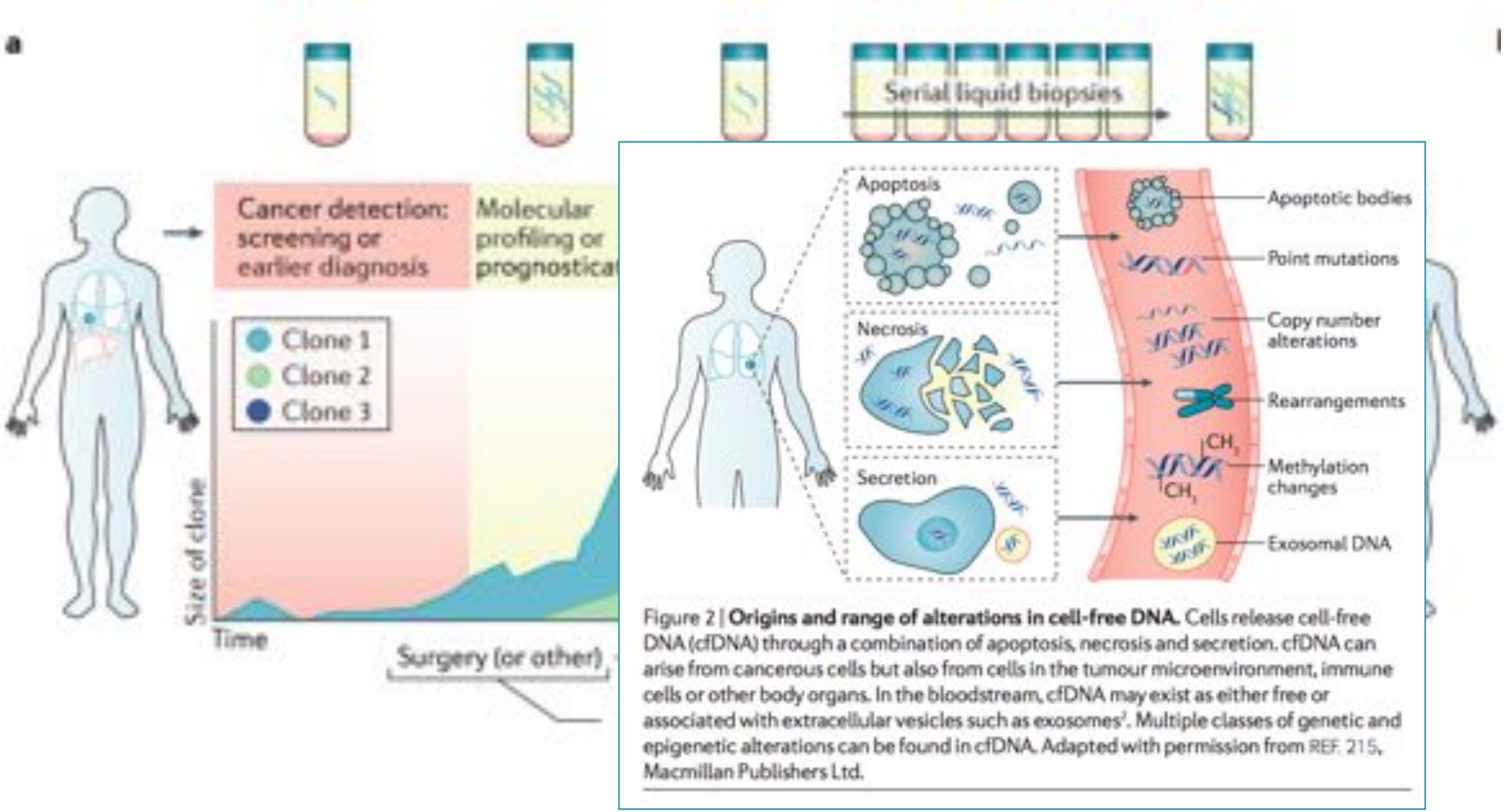
Clonal evolution in relapsed acute myeloid leukemia revealed by whole genome sequencing
Ding et al (2012) Nature. doi:10.1038/nature10738

Liquid Biopsies



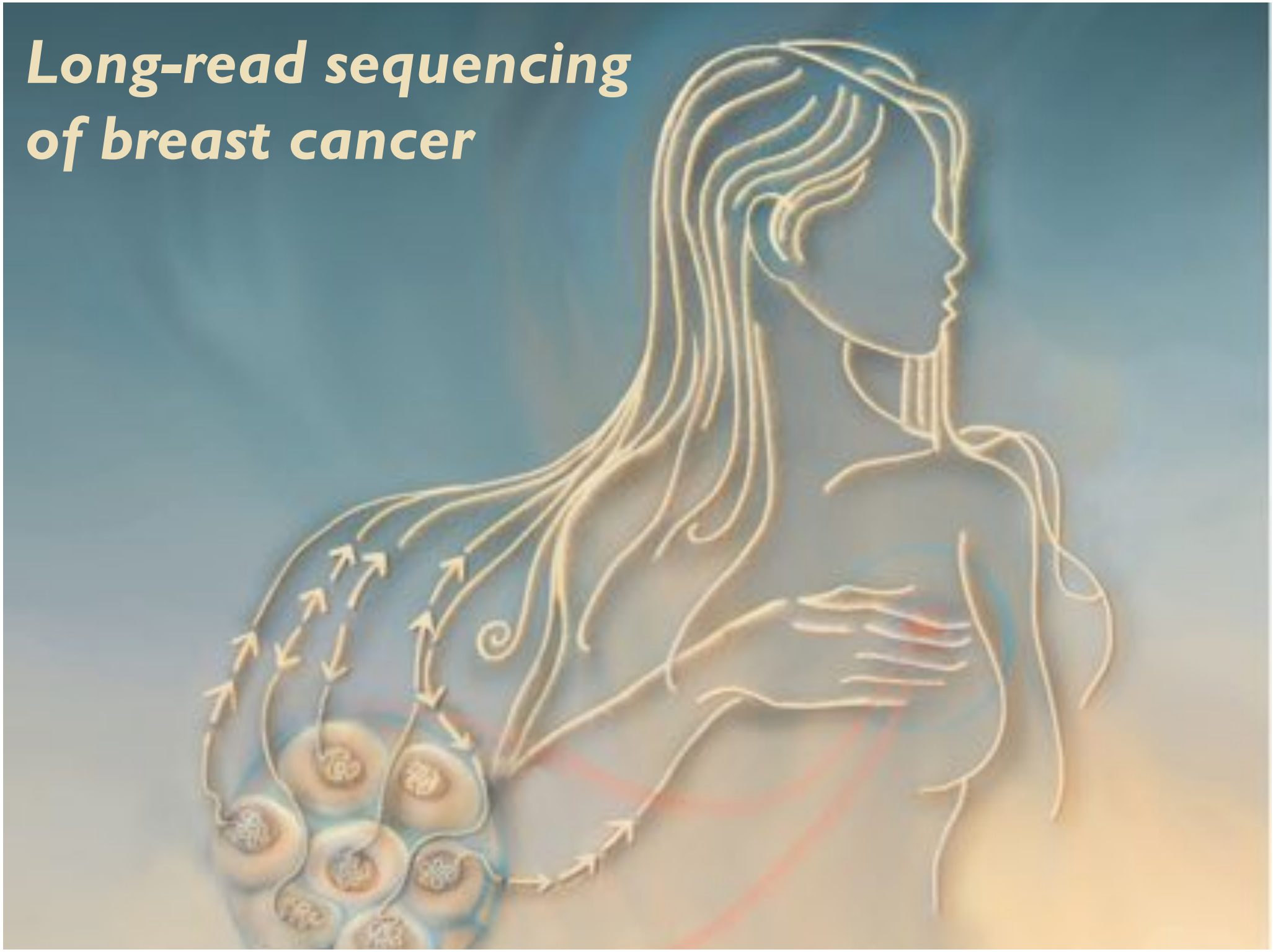
Liquid biopsies come of age: towards implementation of circulating tumour DNA
Wan et al (2017) Nature Review Cancer. doi:10.1038/nrc.2017.7

Liquid Biopsies



Liquid biopsies come of age: towards implementation of circulating tumour DNA
Wan et al (2017) Nature Review Cancer. doi:10.1038/nrc.2017.7

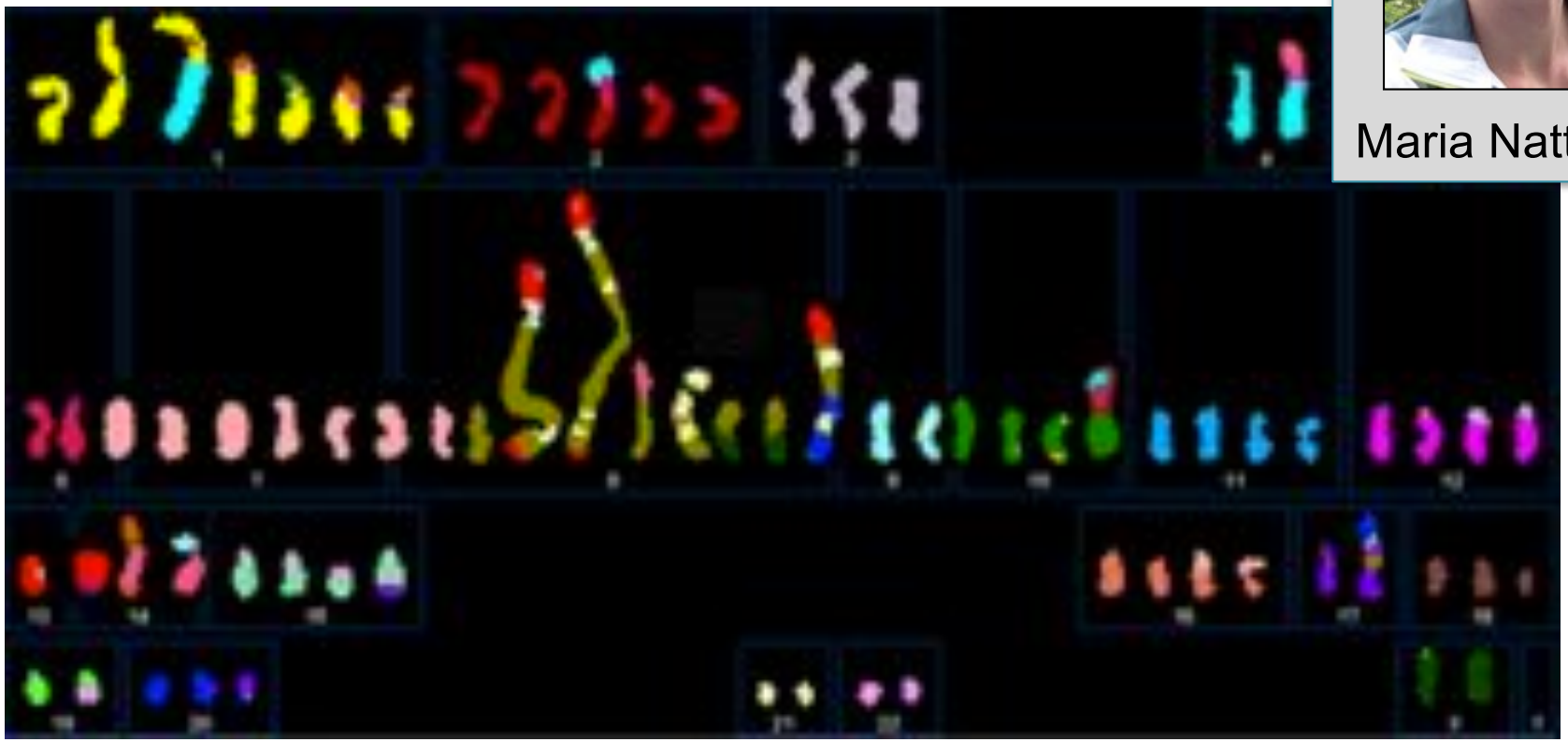
Long-read sequencing of breast cancer



SK-BR-3

Maria Nattestad

Most commonly used Her2-amplified breast cancer



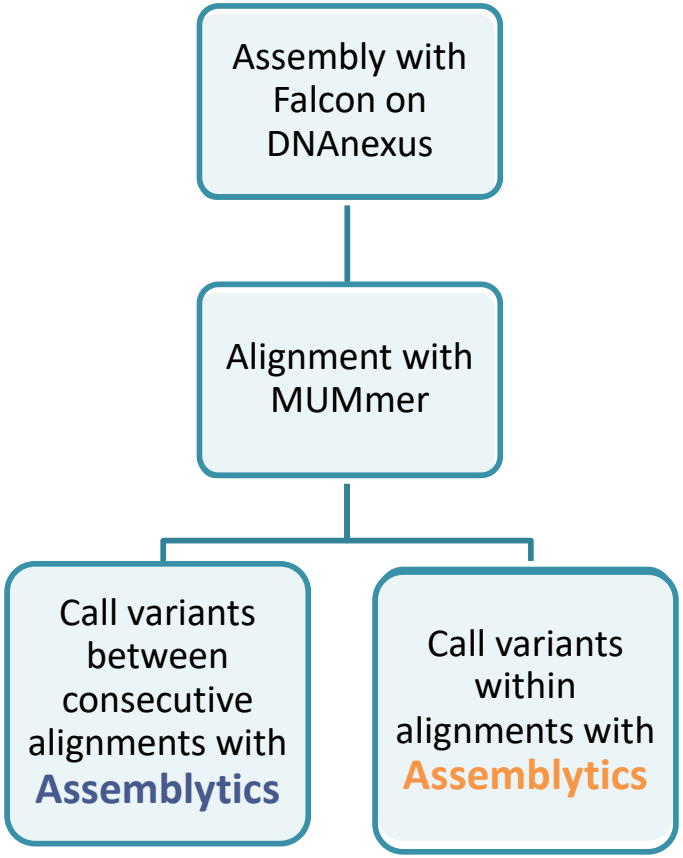
(Davidson et al, 2000)

Can we resolve the complex structural variations, especially around Her2?

Recent collaboration between JHU, CSHL and OICR to *de novo* assemble and analyze the complete cell line genome with PacBio long reads

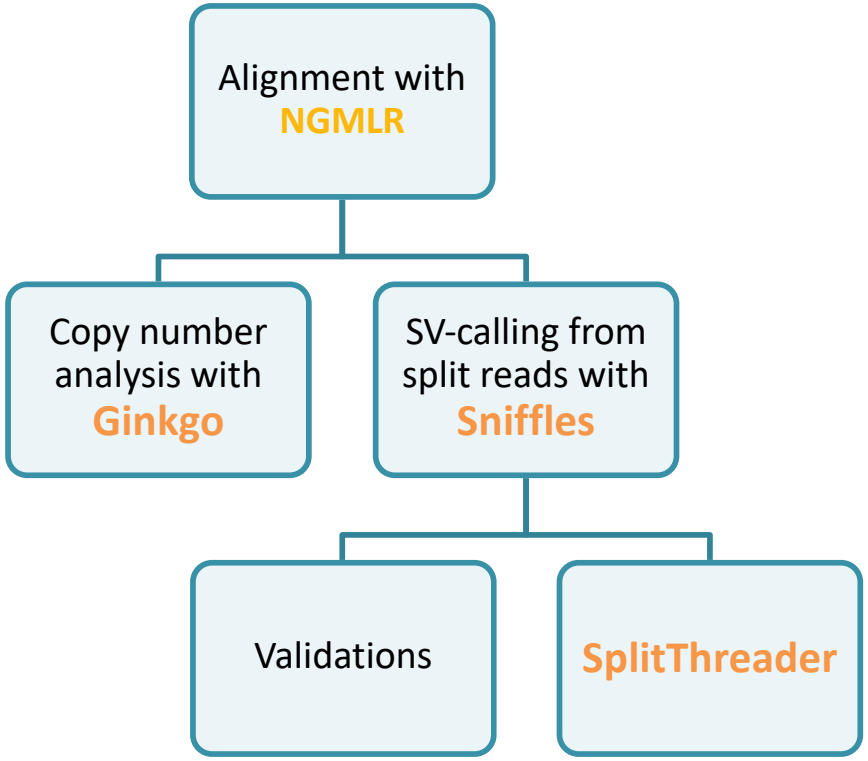
Structural Variation Analysis

Assembly-based



~ 11,000 structural variants
50 bp to 10 kbp

Split-Read based

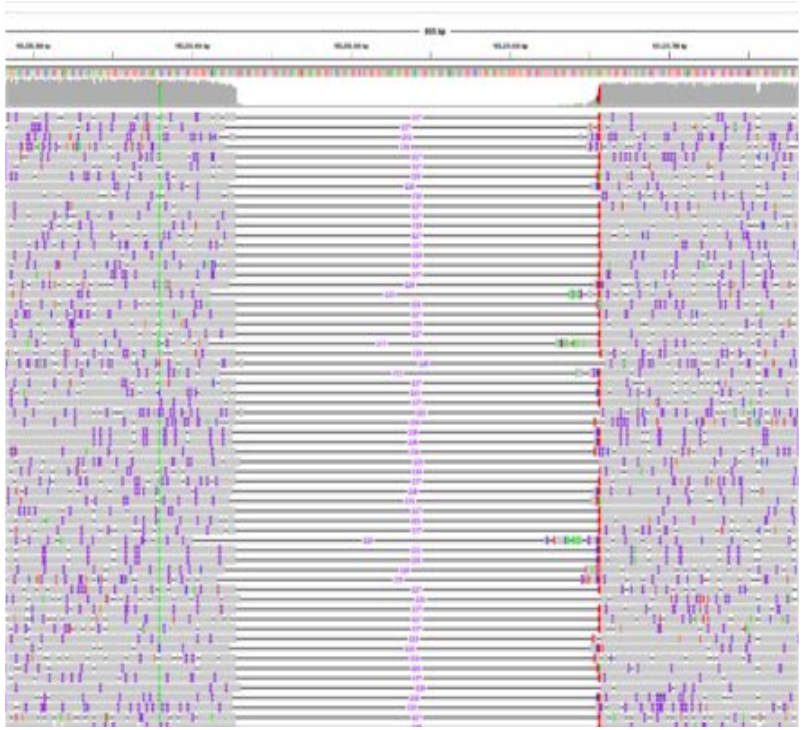


~ 20,000 structural variants
Including many inter-chromosomal rearrangements

NGMLR + Sniffles

BWA-MEM:

NGMLR:



NGMLR: Convex scoring model to accommodate many small gaps from sequencing errors along with less frequent but larger SVs

Accurate detection of complex structural variations using single molecule sequencing
Sedlazeck, Rescheneder et al (2018) *Nature Methods. doi:10.1038/s41592-018-0001-7*

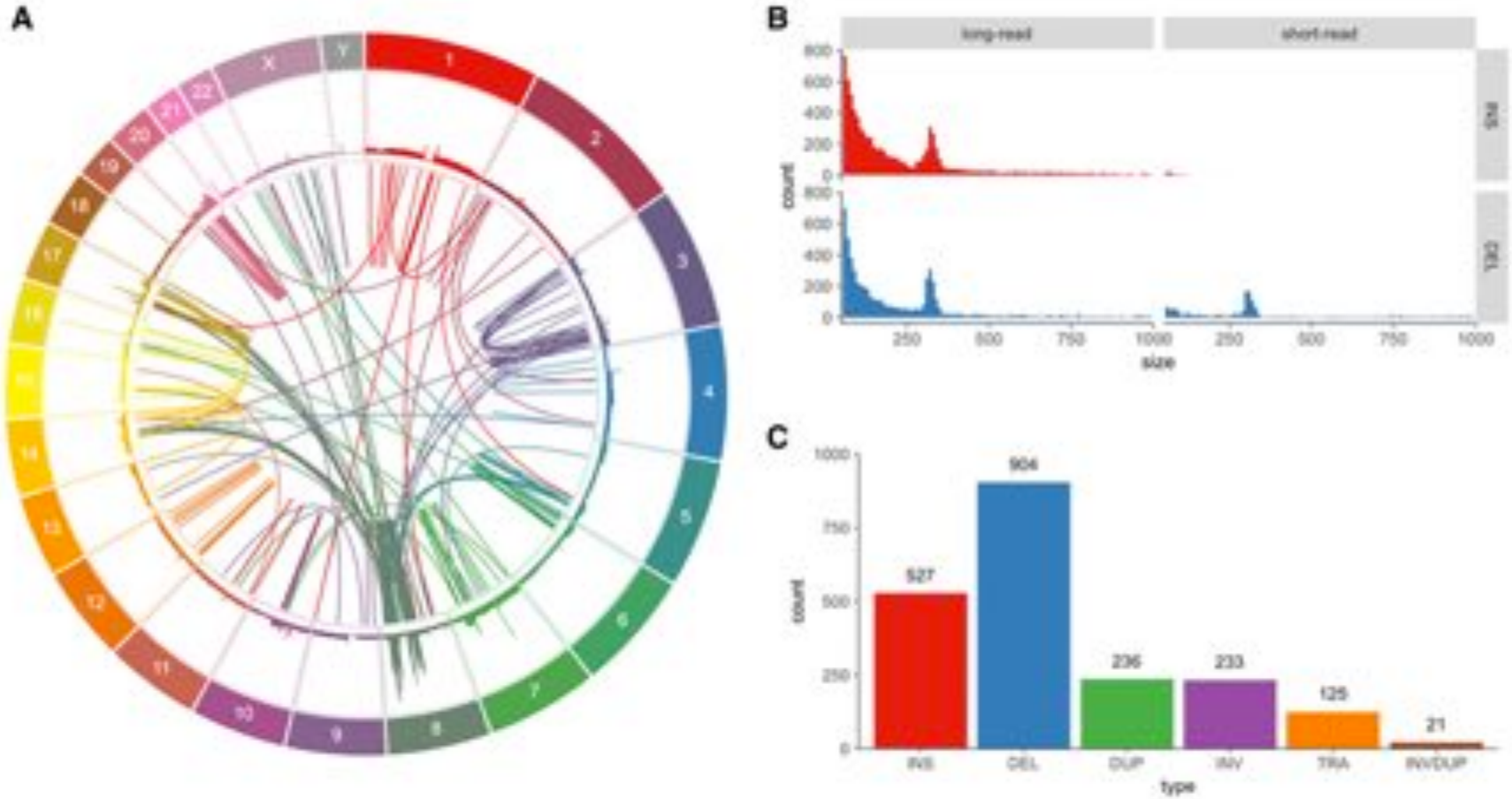


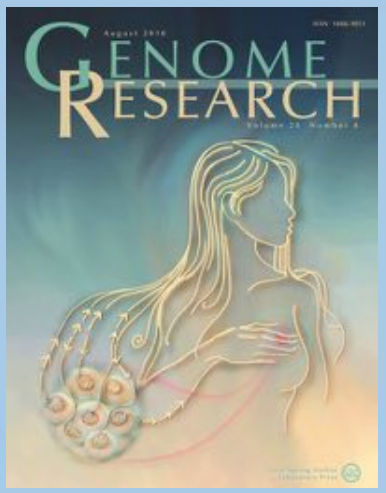
Figure 1. Variants found in SK-BR-3 with PacBio long-read sequencing. (*A*) Circos (Krzywinski et al. 2009) plot showing long-range (larger than 10 kbp or inter-chromosomal) variants found by Sniffles from split-read alignments, with read coverage shown in the *outer* track. (*B*) Variant size histogram of deletions and insertions from size 50 bp up to 1 kbp found by long-read (Sniffles) and short-read (SURVIVOR 2-caller consensus) variant calling, showing similar size distributions for insertions and deletions from long reads but not for short reads, where insertions are greatly underrepresented. (*C*) Sniffles variant counts by type for variants above 1 kbp in size, including translocations and inverted duplications.

Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line

Nattestad et al. (2018) *Genome Research.* doi: 10.1101/gr.231100.117

Highlights

- Finding 10s of thousands of additional variants
- PCR validation confirms high accuracy of long reads
- Detect many novel gene fusions
- Identify early vs late mutations in the cancer



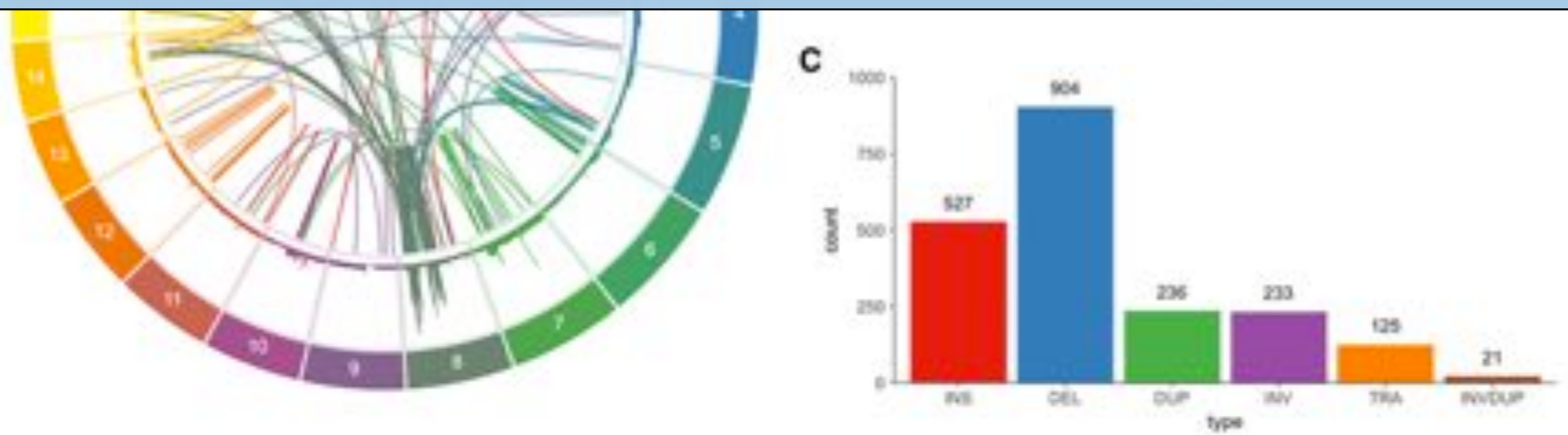
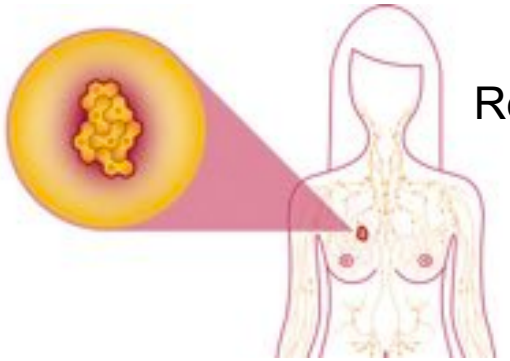


Figure 1. Variants found in SK-BR-3 with PacBio long-read sequencing. (*A*) Circos (Krzywinski et al. 2009) plot showing long-range (larger than 10 kbp or inter-chromosomal) variants found by Sniffles from split-read alignments, with read coverage shown in the *outer* track. (*B*) Variant size histogram of deletions and insertions from size 50 bp up to 1 kbp found by long-read (Sniffles) and short-read (SURVIVOR 2-caller consensus) variant calling, showing similar size distributions for insertions and deletions from long reads but not for short reads, where insertions are greatly underrepresented. (*C*) Sniffles variant counts by type for variants above 1 kbp in size, including translocations and inverted duplications.

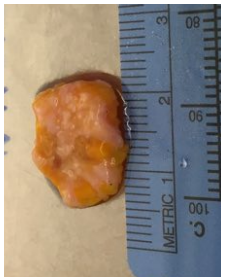
Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line

Nattestad et al. (2018) *Genome Research.* doi: 10.1101/gr.231100.117

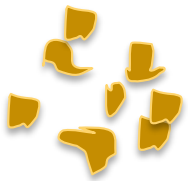
Taking Long Read Sequencing into the Clinic



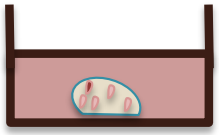
Resected primary tumor



Mechanical/ enzymatic



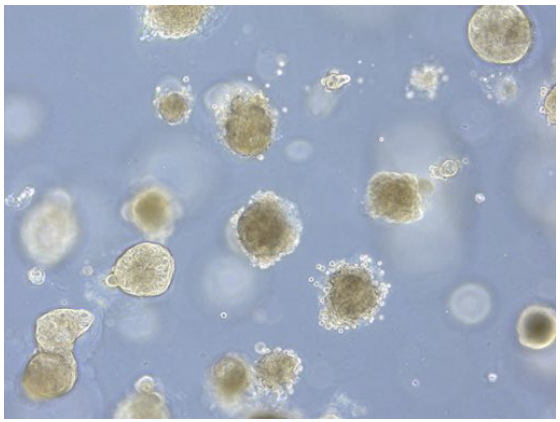
Purification



Plating on Matrigel

Add growth factors

Tumor organoids in culture



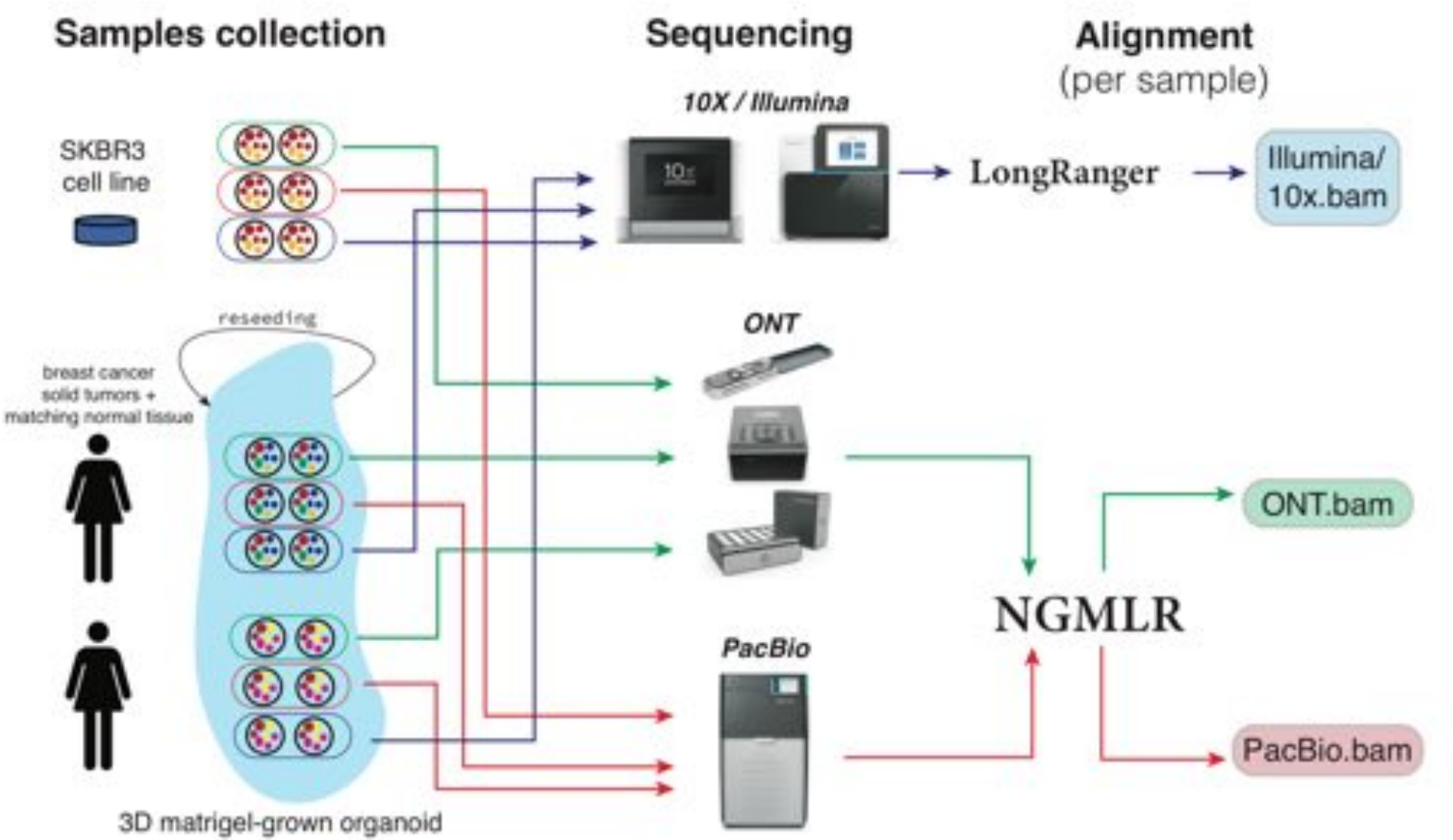
- ✓ Stable Growth in 3D
- ✓ Recapitulate tumor pathology & treatment response
- ✓ Maintenance of tissue/tumor heterogeneity
- ✓ "2017 Method of the Year" - Nature Methods

CSH

Northwell Health

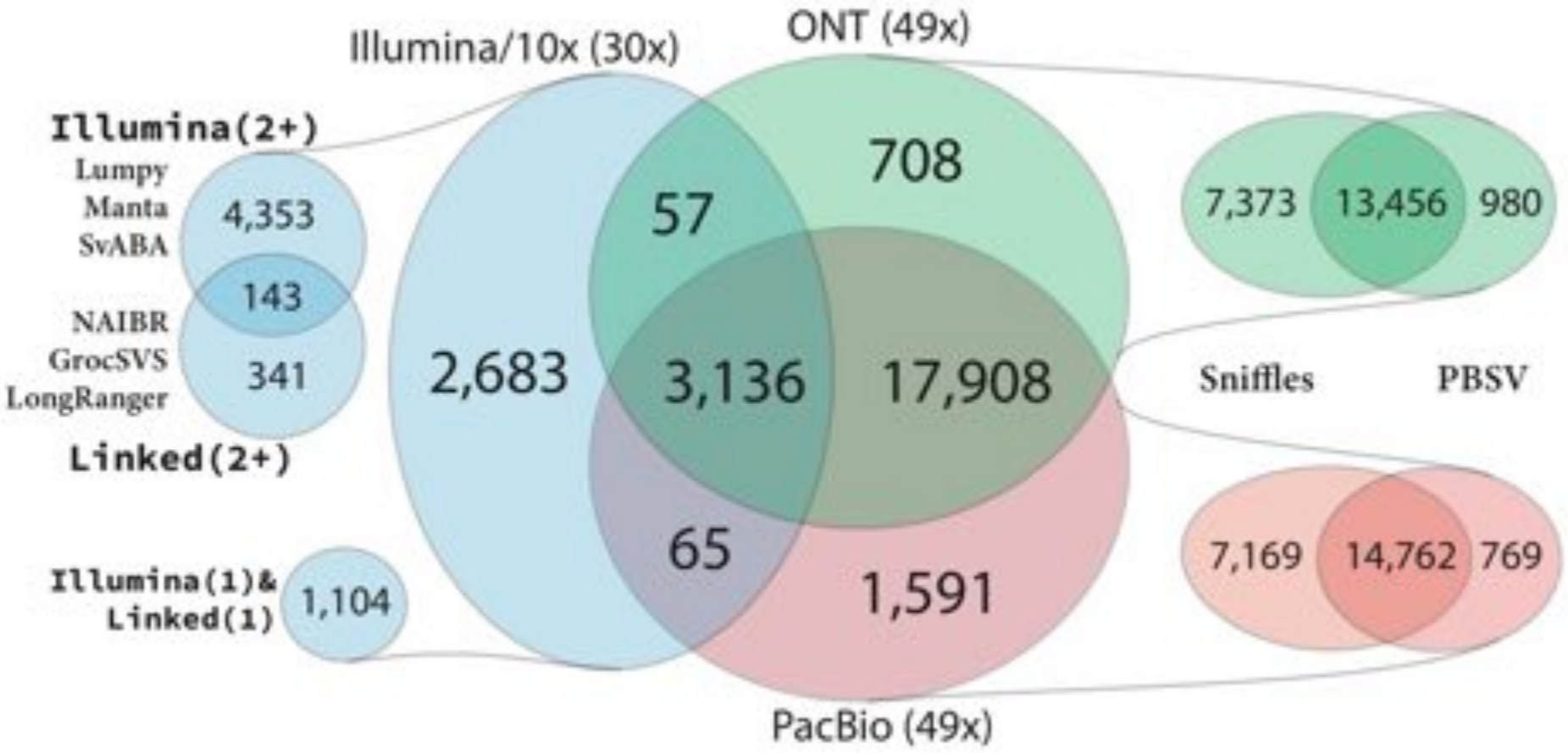
David Spector Karen Kostroff

Data Production

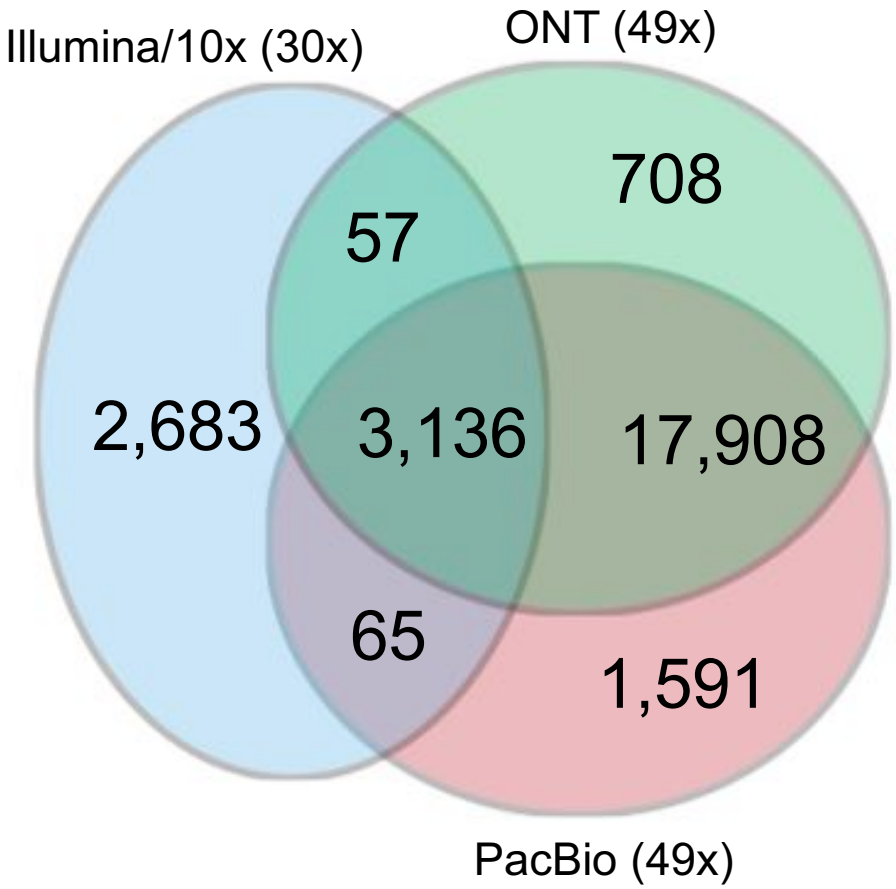


Comprehensive analysis of structural variants in breast cancer genomes using single molecule sequencing
Aganezov, S et al. (2019) *bioRxiv* doi: https://doi.org/10.1101/847855

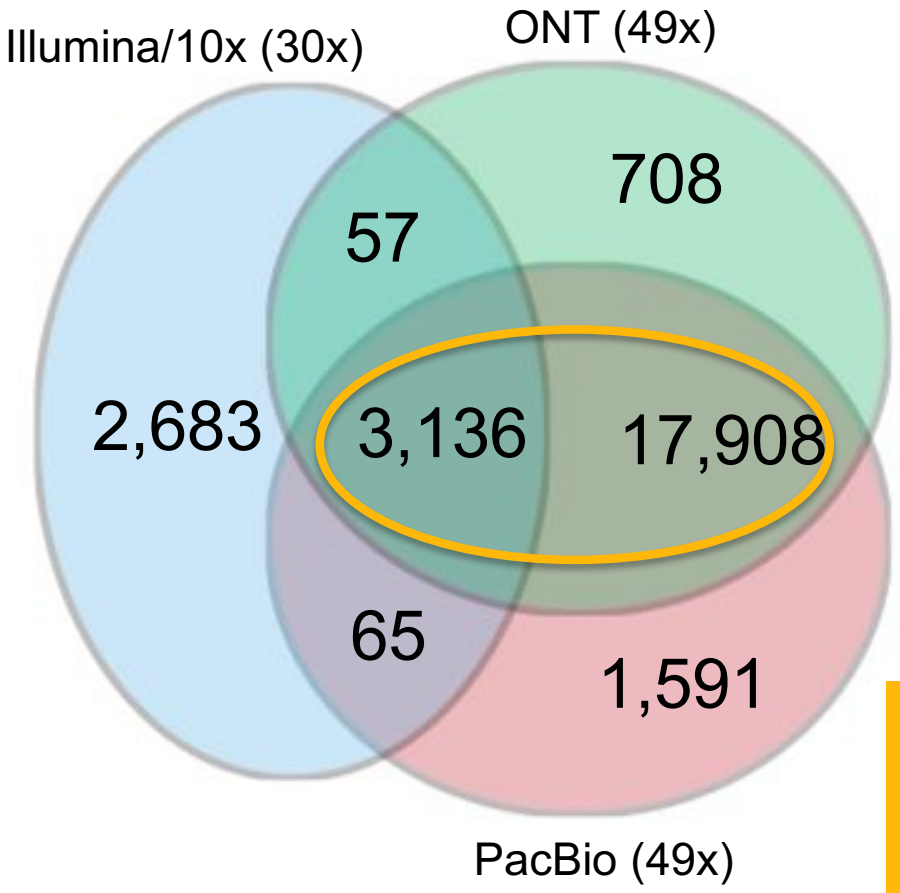
Structural Variation Consistency



Structural Variation Consistency

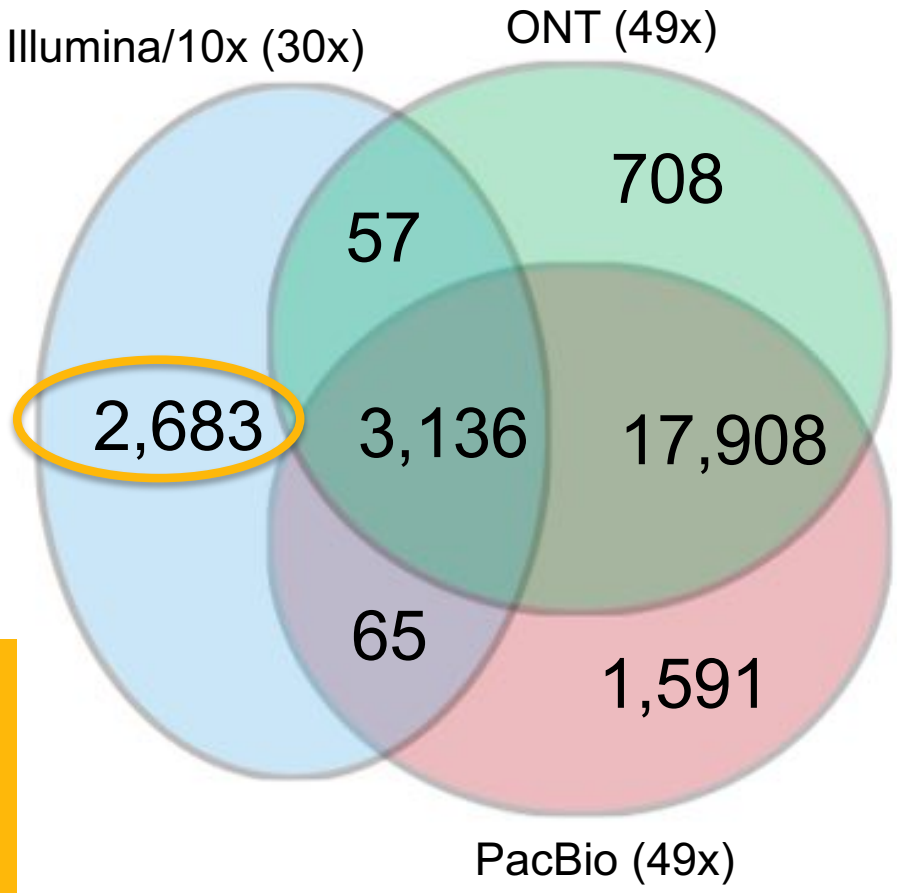


Structural Variation Consistency



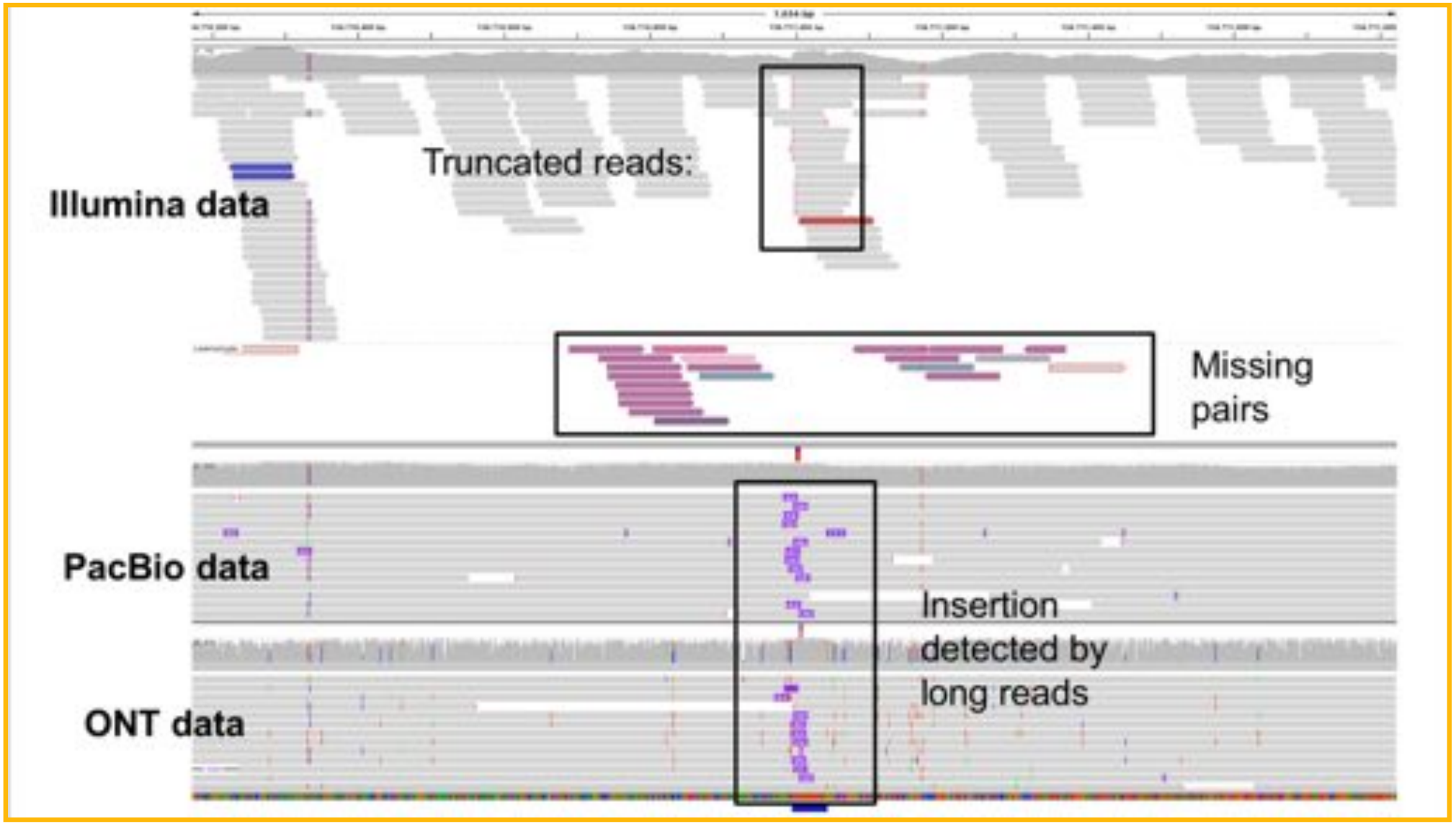
- Very strong concordance between long read platforms
- Substantially more variants than detected by short reads

Structural Variation Consistency

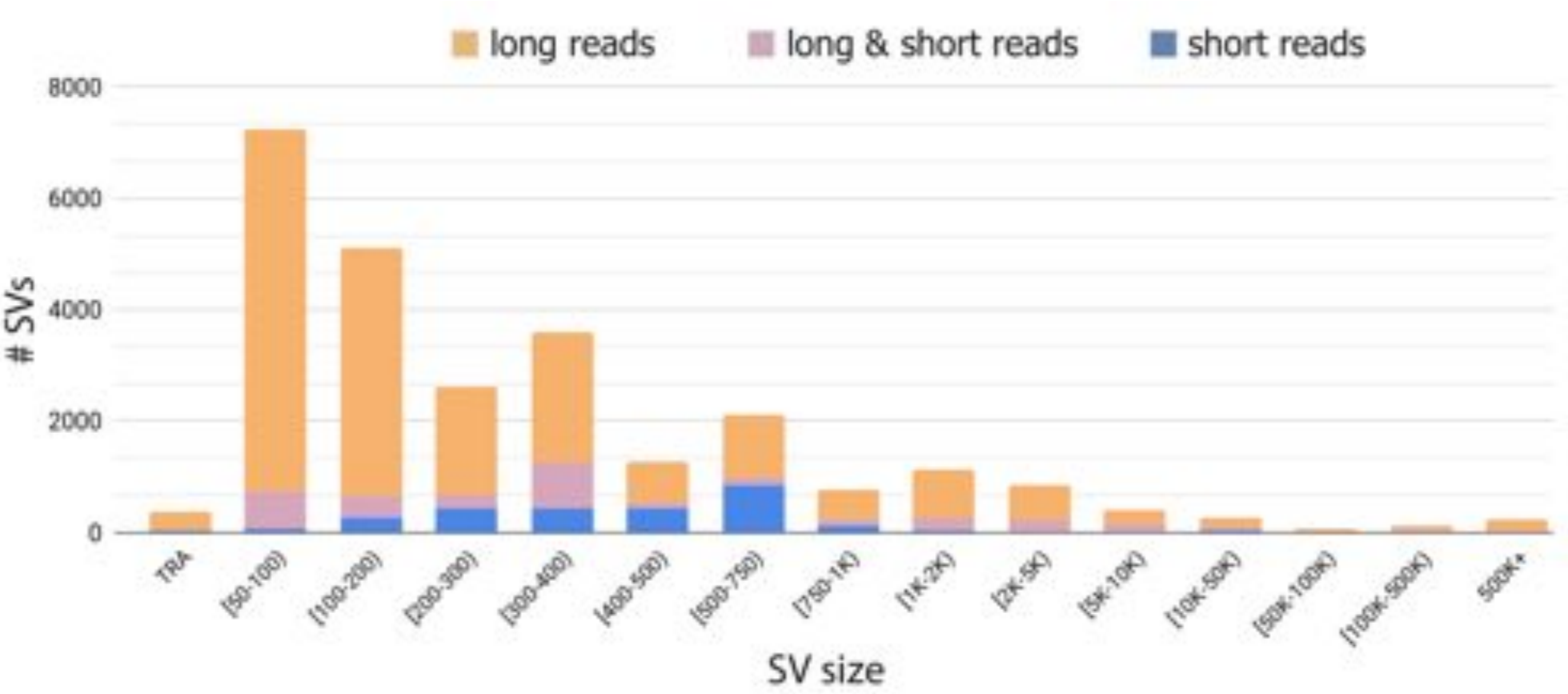


- PCR validation shows most Illumina-only calls are false positives

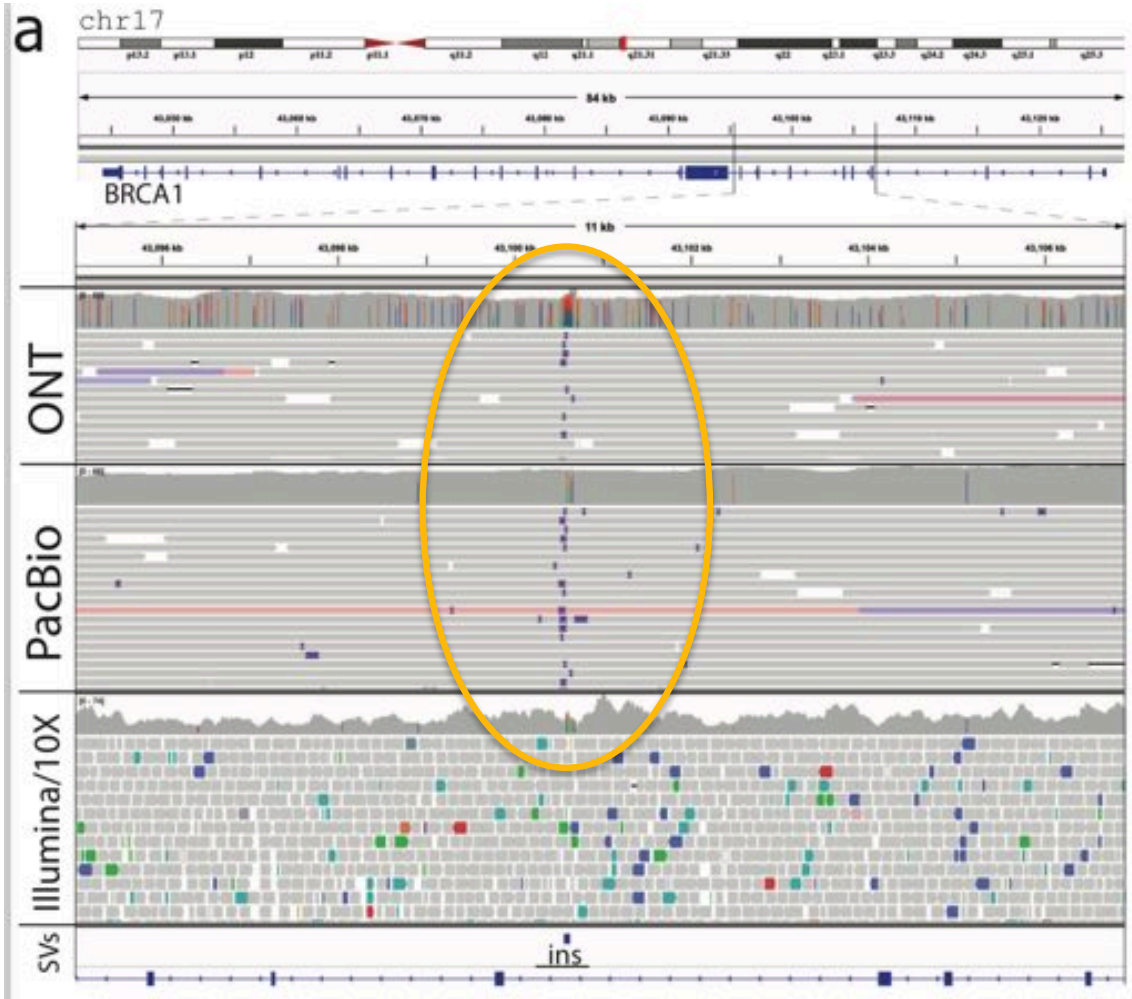
Structural Variation Consistency



Structural Variation Identification

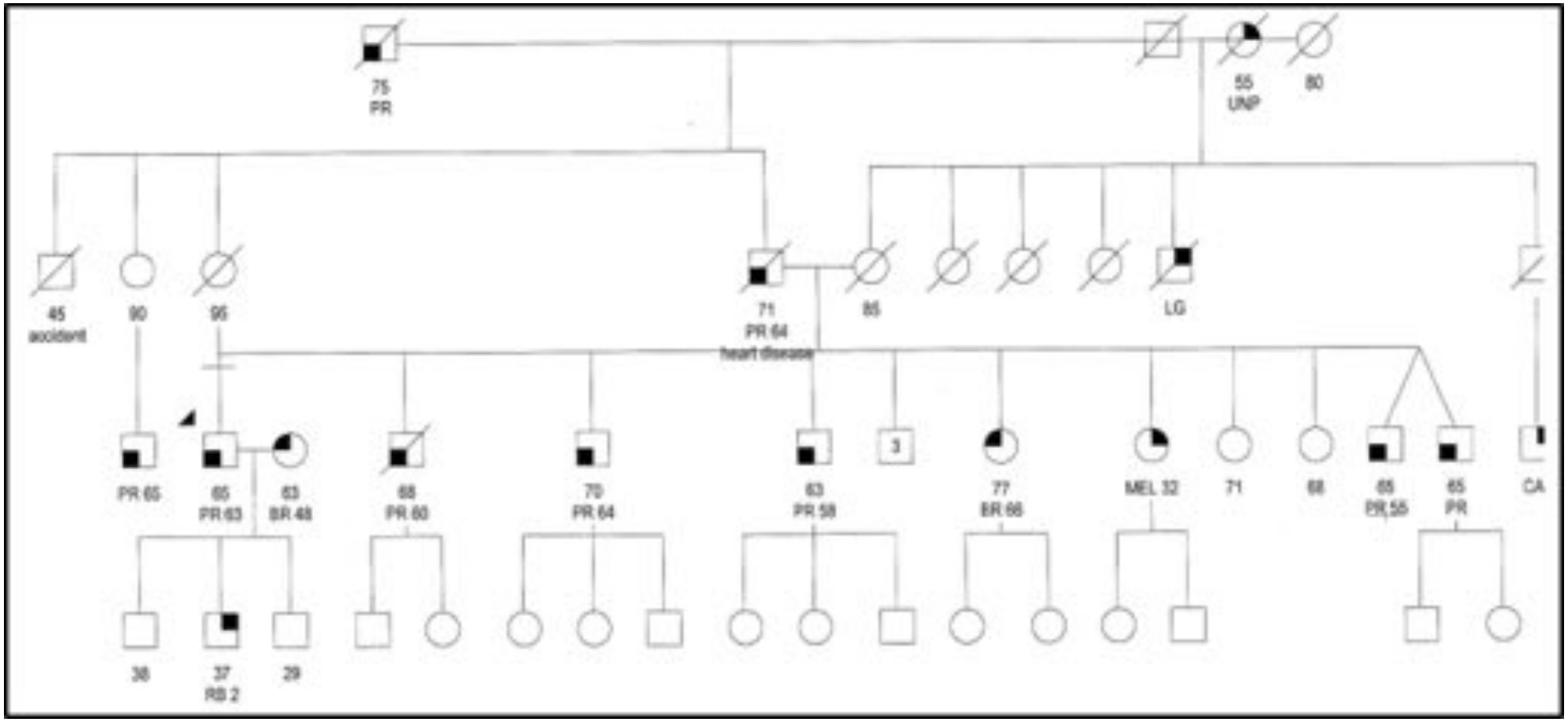


Hidden Variants in Breast Cancer Genes



62bp repeat expansion in BRCA1 detected in normal tissue that is undetectable using a panel or short read sequencing

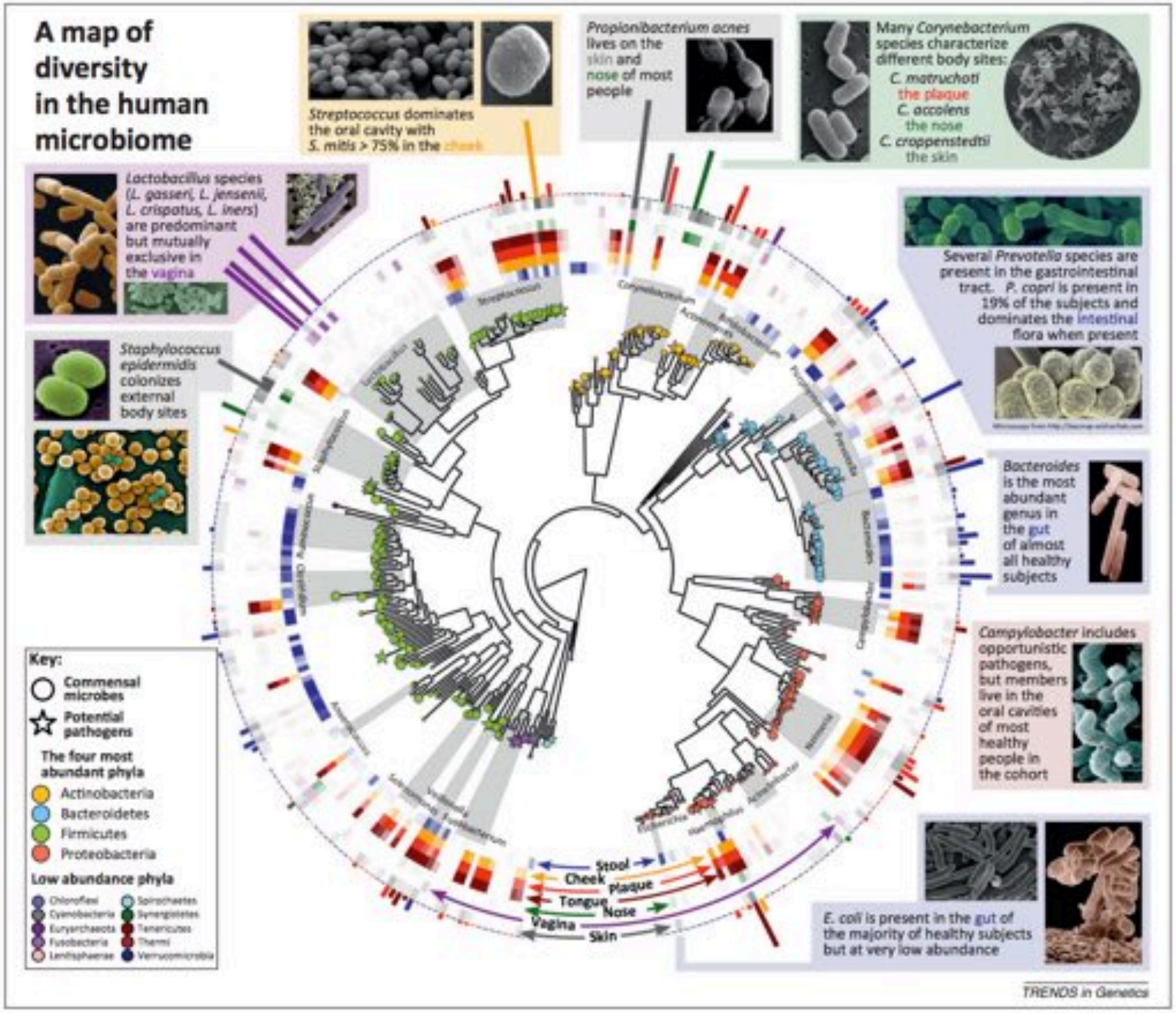
What causes "outlier" families?



The Mark Foundation® for Cancer Research

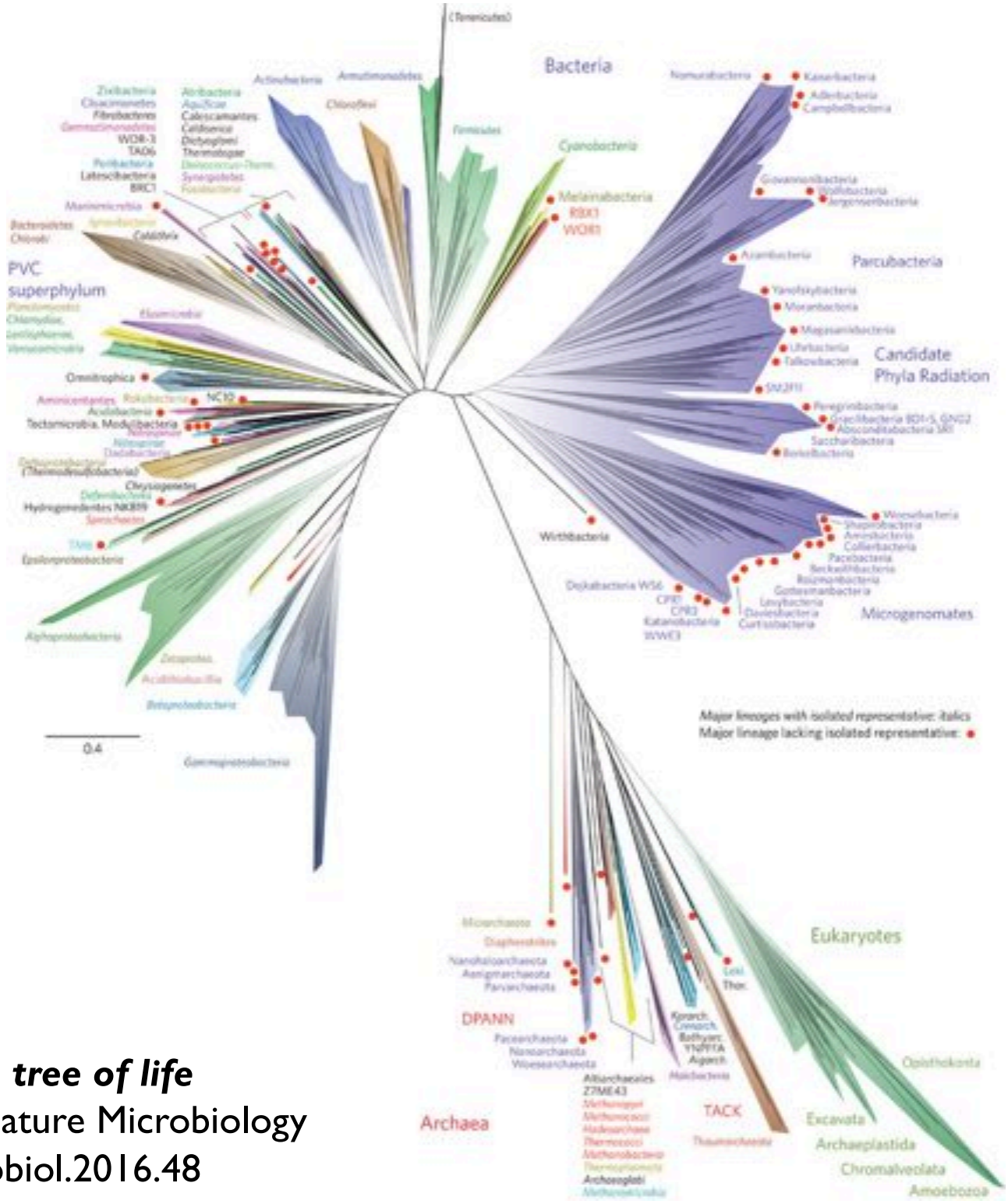
Eli Van Allen,
Dana-Farber Cancer Institute

Part 2: Metagenomics



Biodiversity and functional genomics in the human microbiome

Morgan et al (2013) Trends in Genetics. http://doi.org/10.1016/j.tig.2012.09.005



A new view of the tree of life
Hug et al. (2016) Nature Microbiology
doi:10.1038/nmicrobiol.2016.48

Your second genome?

Human body:
~10 trillion cells

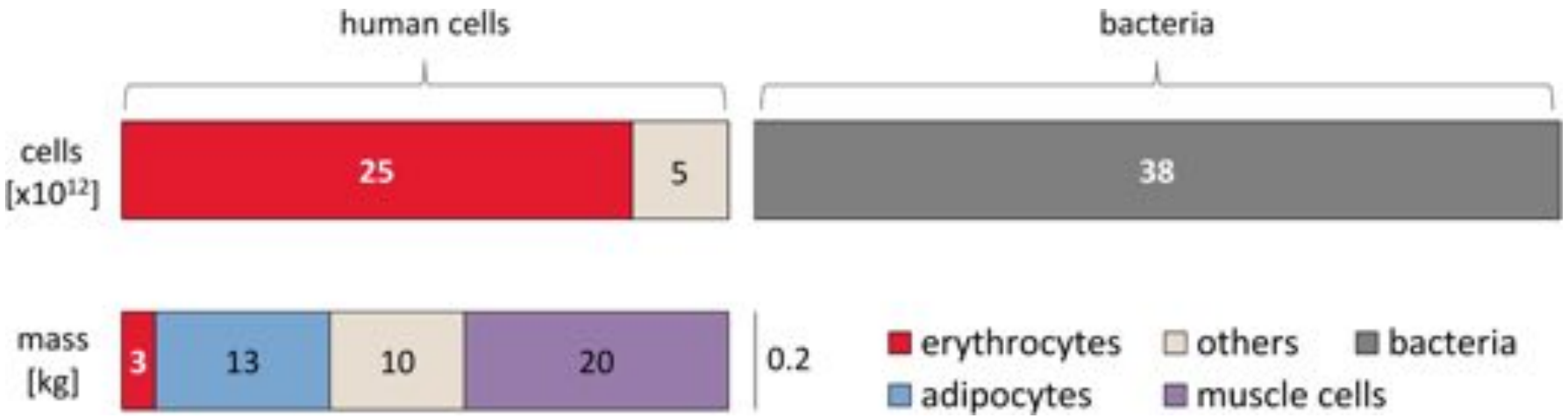
Human brain:
~3.3 lbs

Microbiome
~100 trillion cells

Total mass:
~3.3 lbs

Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans
Sender et al (2016) Cell. http://doi.org/10.1016/j.cell.2016.01.013

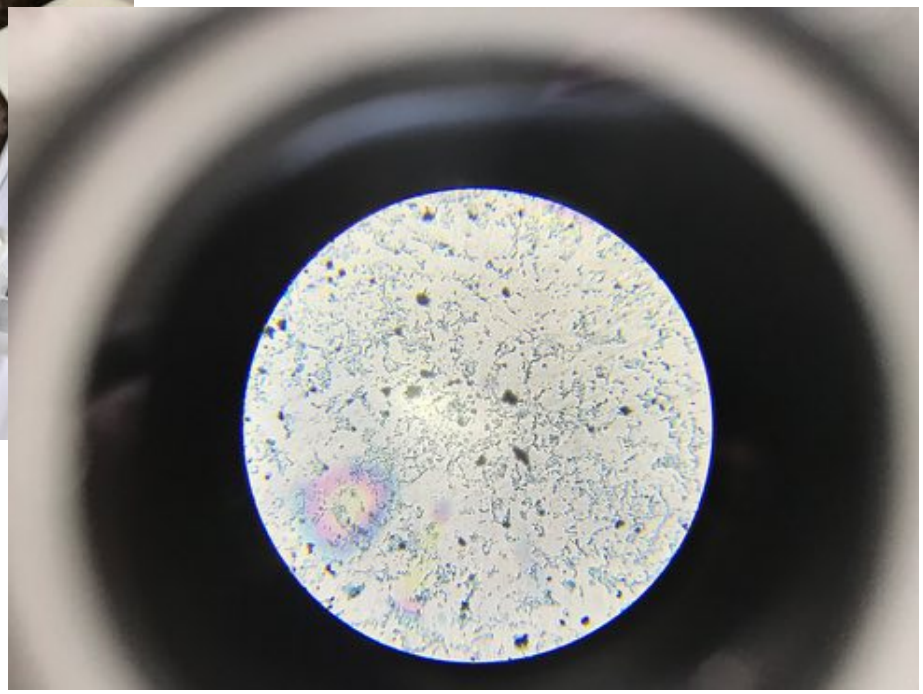
Okay, maybe not 10x more cells but still a lot! ☺



population segment	body weight [kg]	age [y]	blood volume [L]	RBC count [10^{12}/L]	colon content [g]	bac. conc. [10^{11}/ g wet] (1)	total human cells [10^{12}] (2)	total bacteria [10^{12}]	B:H
ref. man	70	20–30	4.9	5.0	420	0.92	30	38	**1.3**
ref. woman	63		3.9	4.5	480	0.92	21	44	**2.2**
young infant	4.4	4 weeks	0.4	3.8	48	0.92	1.9	4.4	**2.3**
infant	9.6	1	0.8	4.5	80	0.92	4	7	**1.7**
elder	70	66	3.8 (3)	4.8	420	0.92	22	38	**1.8**
obese	140		6.7	5.0 (4)	610 (5)	0.92	40	56	**1.4**

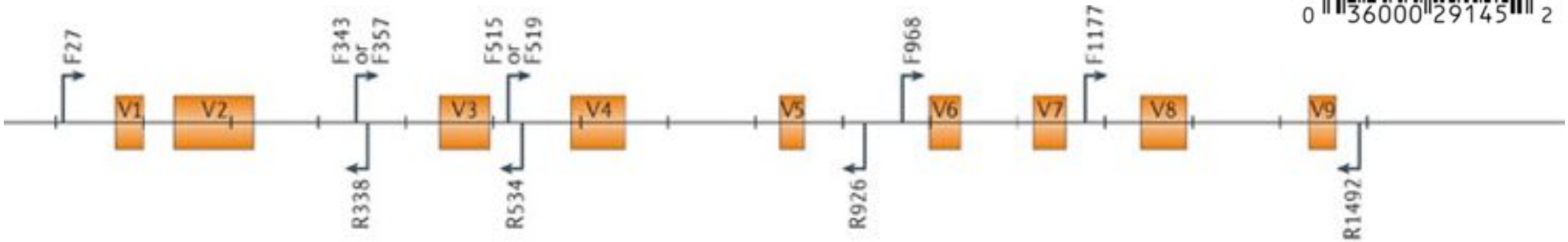
Revised Estimates for the Number of Human and Bacteria Cells in the Body
Sender et al (2016) PLOS Biology. https://doi.org/10.1371/journal.pbio.1002533

Pre-PCR: Gram-Staining



Gram staining differentiates bacteria by the chemical and physical properties of their cell walls by detecting peptidoglycan, which is present in the cell wall of Gram-positive bacteria

16S rRNA



The 16S rRNA gene is a section of prokaryotic DNA found in all bacteria and archaea. This gene codes for an rRNA, and this rRNA in turn makes up part of the ribosome.

The 16S rRNA gene is a commonly used tool for identifying bacteria for several reasons. First, traditional characterization depended upon phenotypic traits like gram positive or gram negative, bacillus or coccus, etc. Taxonomists today consider analysis of an organism's DNA more reliable than classification based solely on phenotypes. Secondly, researchers may, for a number of reasons, want to identify or classify only the bacteria within a given environmental or medical sample. Thirdly, the 16S rRNA gene is relatively short at 1.5 kb, making it faster and cheaper to sequence than many other unique bacterial genes.

http://greengenes.lbl.gov/cgi-bin/JD_Tutorial/nph-16S.cgi

Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 6955–6959, October 1985
Evolution

0 36000 29145 2

Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses

(reverse transcriptase/dideoxynucleotide)

DAVID J. LANE*, BERNADETTE PACE*, GARY J. OLSEN*, DAVID A. STAHL†‡, MITCHELL L. SOGIN†, AND NORMAN R. PACE*§

*Department of Biology and Institute for Molecular and Cellular Biology, Indiana University, Bloomington, IN 47405; and †Department of Molecular and Cellular Biology, National Jewish Hospital and Research Center, Denver, CO 80206

Communicated by Ralph S. Wolfe, June 26, 1985

ABSTRACT **Although the applicability of small subunit ribosomal RNA (16S rRNA) sequences for bacterial classification is now well accepted, the general use of these molecules has been hindered by the technical difficulty of obtaining their sequences. A protocol is described for rapidly generating large blocks of 16S rRNA sequence data without isolation of the 16S rRNA or cloning of its gene. The 16S rRNA in bulk cellular RNA preparations is selectively targeted for dideoxynucleotide-terminated sequencing by using reverse transcriptase and synthetic oligodeoxynucleotide primers complementary to universally conserved 16S rRNA sequences. Three particularly useful priming sites, which provide access to the three major 16S rRNA structural domains, routinely yield 800–1000 nucleotides of 16S rRNA sequence. The method is evaluated with respect to accuracy, sensitivity to modified nucleotides in the template RNA, and phylogenetic usefulness, by examination of several 16S rRNAs whose gene sequences are known. The relative simplicity of this approach should facilitate a rapid expansion of the 16S rRNA sequence collection available for phylogenetic analyses.**

described here rapidly provides partial sequences of 16S rRNA that are useful for phylogenetic analysis.

MATERIALS AND METHODS

Purification of RNA Templates. Bulk, cellular RNA was purified by phenol extraction of French pressure cell lysates as detailed by Pace *et al.* (6), except that ribosomes were not pelleted before extraction. High molecular weight RNA was then prepared by precipitation with 2 M NaCl (6). Although not essential, NaCl precipitation of the RNA generally increased the amount of legible sequence data and reduced backgrounds on gels, presumably by eliminating fragmented DNA from the reactions. RNA was stored at 2 mg/ml in 10 mM Tris·HCl (pH 7.4) at −20°C.

Oligodeoxynucleotide Primers. Oligodeoxynucleotide primers were synthesized manually by using the appropriate blocked and protected nucleoside diisopropylphosphoramidites and established coupling protocols (7). Deblocked products were purified by polyacrylamide gel electrophore-

Box 1 | **Species definitions and concepts in microbiology**

Definitions

Microbes are currently assigned to a common species if their reciprocal, pairwise DNA re-association values are ≥70% in DNA–DNA hybridization experiments under standardized conditions and their ΔT_m (melting temperature) is ≤5°C[79]. In addition, all strains within a species must possess a certain degree of phenotypic consistency, and species descriptions should be based on more than one type strain[11]. A species name is only assigned if its members can be distinguished from other species by at least one diagnostic phenotypic trait[79]. Microbes with 16S ribosomal RNAs (rRNAs) that are ≤98.7% identical are always members of different species, because such strong differences in rRNA correlate with <70% DNA–DNA similarity[80]. However, the opposite is not necessarily true, and distinct species have been occasionally described with 16S rRNAs that are >98.7% identical. Most uncultured microbes cannot be assigned to a classical species because we do not know their phenotype. In some cases, uncultured microbes can be assigned a provisional '*Candidatus*' designation if their 16S rRNA sequences are sufficiently different from those of recognized species, if experimental *in situ* hybridization can be used to specifically detect them and if a basic description of their morphology and biology has been provided[81].

Microbial diversity and the genetic nature of microbial species
Achtman & Wagner (2008) Nature Reviews Microbiology. doi:10.1038/nrmicro1872

Box 1 | **Species definitions and concepts in microbiology**

Definitions

Microbes are currently assigned to a common species if their reciprocal, pairwise DNA re-association values are ≥70% in DNA–DNA hybridization experiments under standardized conditions and their ΔT_m (melting temperature) is ≤5°C[79]. In addition, all strains within a species must possess a certain degree of phenotypic consistency, and species descriptions should be based on more than one type strain[11]. A species name is only assigned
diagnostic ph
≤98.7% ident
differences in
is not necessa
rRNAs that ar
classical spec
microbes can
sequences ar
in situ hybridi
their morpho

Concepts

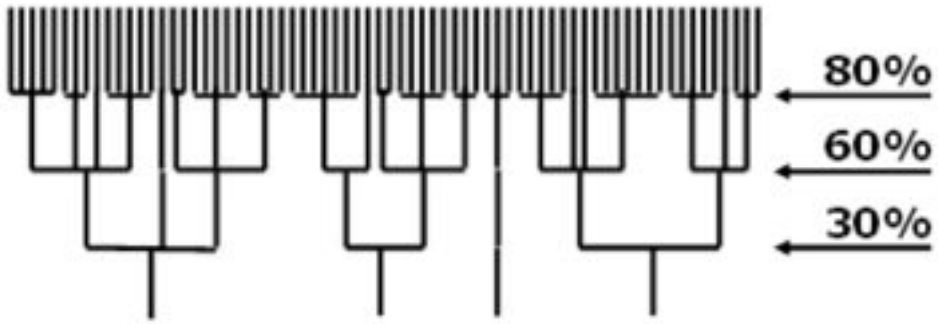
Various concepts have been suggested for microbial species, but none have been generally accepted[9]. The following quotes represent several published concepts that were chosen to illustrate the lack of consensus:

- "A species could be described as a monophyletic and genomically coherent cluster of individual organisms that show a high degree of overall similarity in many independent characteristics, and is diagnosable by a discriminative phenotypic property." (REF. 9)
- "Species are considered to be an irreducible cluster of organisms diagnosably different from other such clusters and within which there is a parental pattern of ancestry and descent." (REF. 82)
- "A species is a group of individuals where the observed lateral gene transfer within the group is much greater than the transfer between groups." (REF. 83)
- "Microbes ... do not form natural clusters to which the term "species" can be universally and sensibly applied." (REF. 84)
- "Species are (segments of) metapopulation lineages." (REF. 7)

Microbial diversity and the genetic nature of microbial species
Achtman & Wagner (2008) Nature Reviews Microbiology. doi:10.1038/nrmicro1872

Operational Taxonomic Units (OTUs)

OTUs take the place of "species" in many microbiome diversity analyses because named species genomes are often unavailable for particular marker sequences.



- Although much of the 16S rRNA gene is highly conserved, several of the sequenced regions are variable or hypervariable, so small numbers of base pairs can change in a very short period of evolutionary time.
- Because 16S regions are typically sequenced using only a single pass, there is a fair chance that they will thus contain at least one sequencing error. This means that requiring tags to be 100% identical will be extremely conservative and treat essentially clonal genomes as different organisms.
- Some degree of sequence divergence is typically allowed - 95%, 97%, or 99% are sequence similarity cutoffs often used in practice [18] - and the resulting cluster of nearly-identical tags (and thus assumedly identical genomes) is referred to as an Operational Taxonomic Unit (OTU) or sometimes phylotype.

Chapter 12: Human Microbiome Analysis
Morgan & Huttenhower (2012) PLOS Comp Bio.https://doi.org/10.1371/journal.pcbi.1002808

16S versus shotgun NGS

16S

Fast (minutes – hours)
Directed analysis
Cheap per sample
Family/Genus Identification

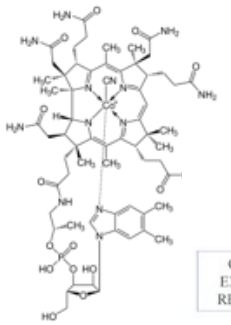
NGS

Slower (hours to days)
Whole Metagenome
More expensive per sample
Species/Strain Identification
Genes presence/absence
Variant analysis
Eukaryotic hosts
Can ID fungi, viruses, etc.

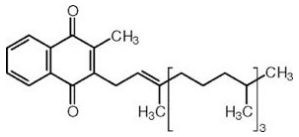
E. coli

?

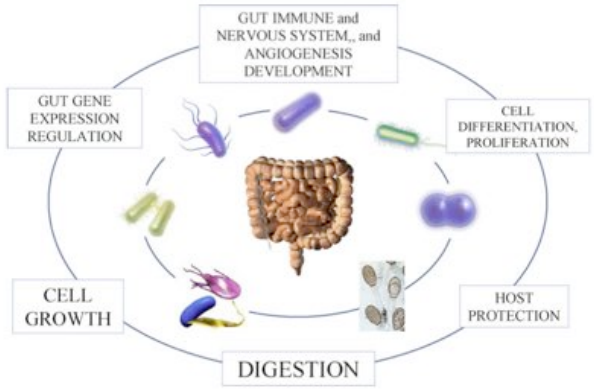
Commensal?



Vitamin B12



Vitamin K



Colonization Barrier Effect Against Pathogens

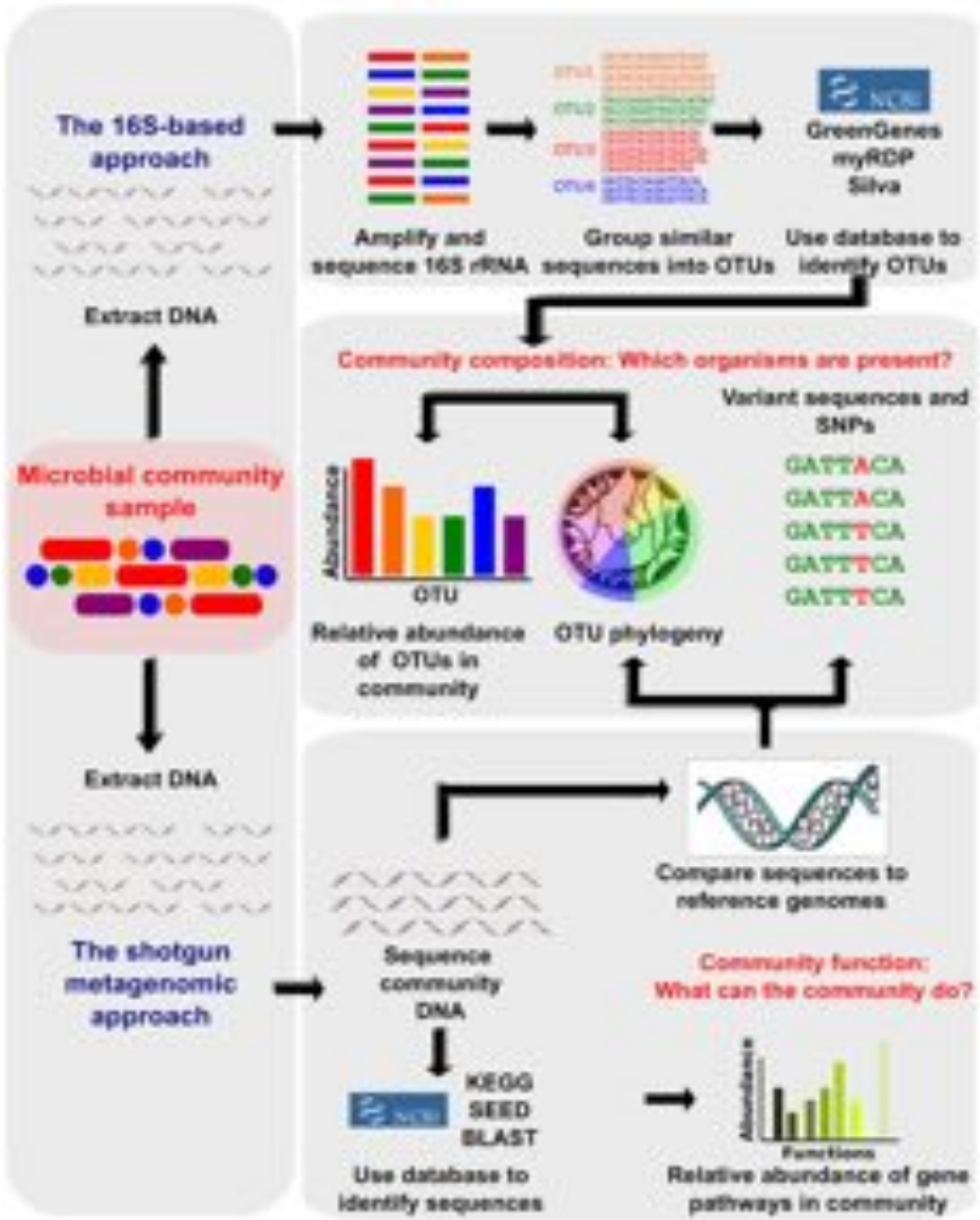
Pathogenic?

Strain O157:H7

Hemorrhagi c enteritis


PRIUS

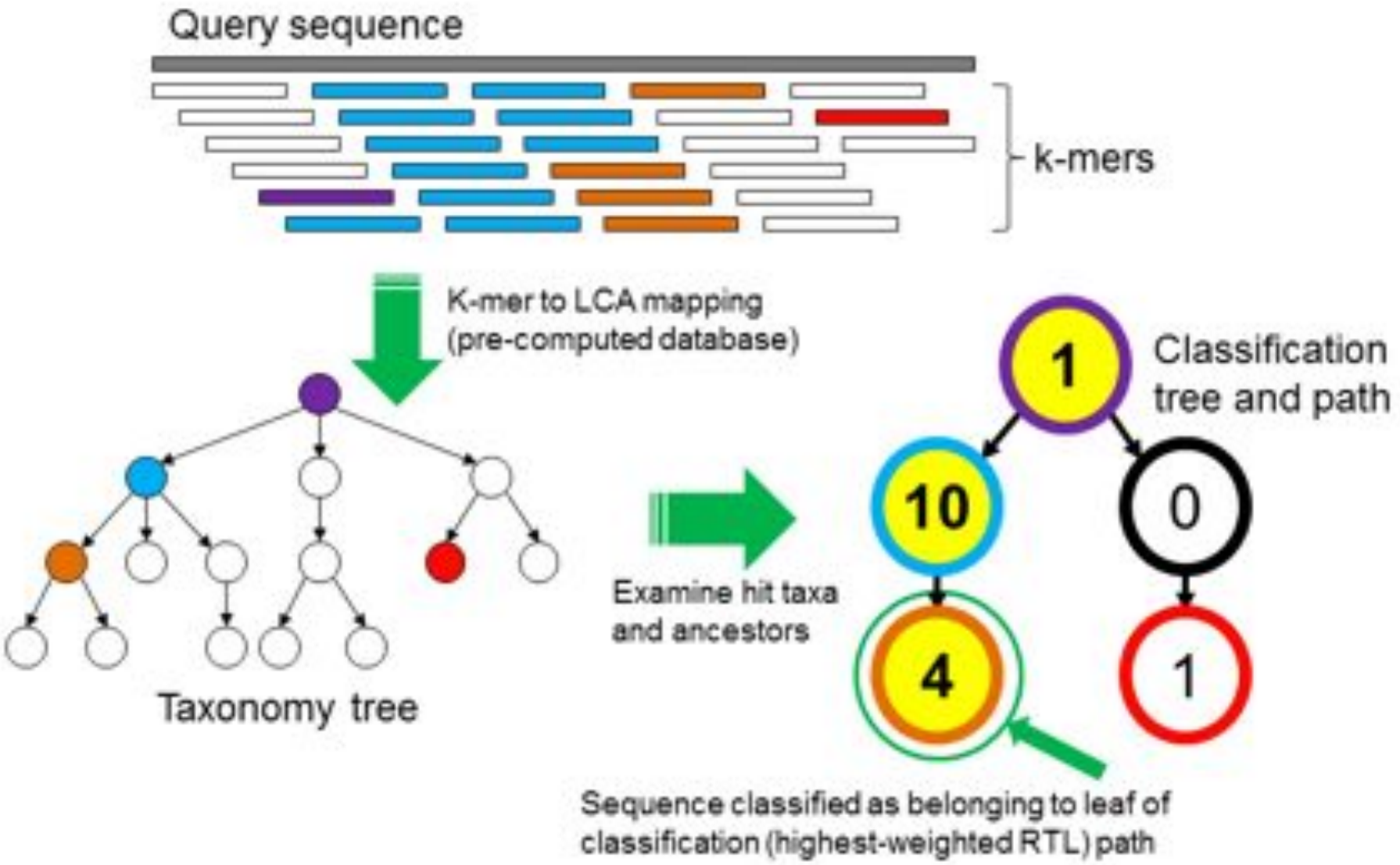
Part III:
Metagenomics Methods



Chapter 12: Human Microbiome Analysis

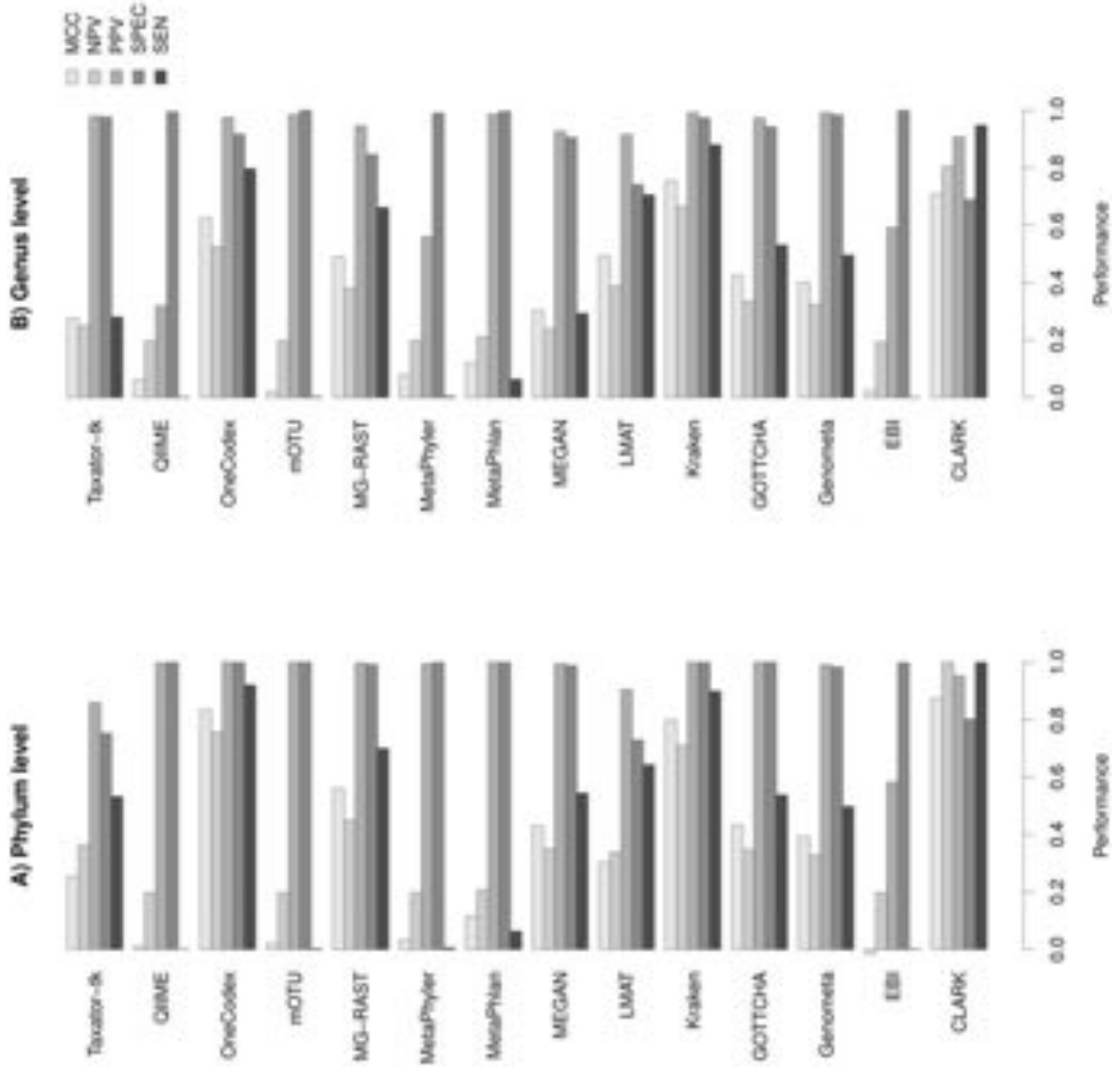
Morgan & Huttenhower (2012) PLOS Comp Bio.https://doi.org/10.1371/journal.pcbi.1002808

Kraken



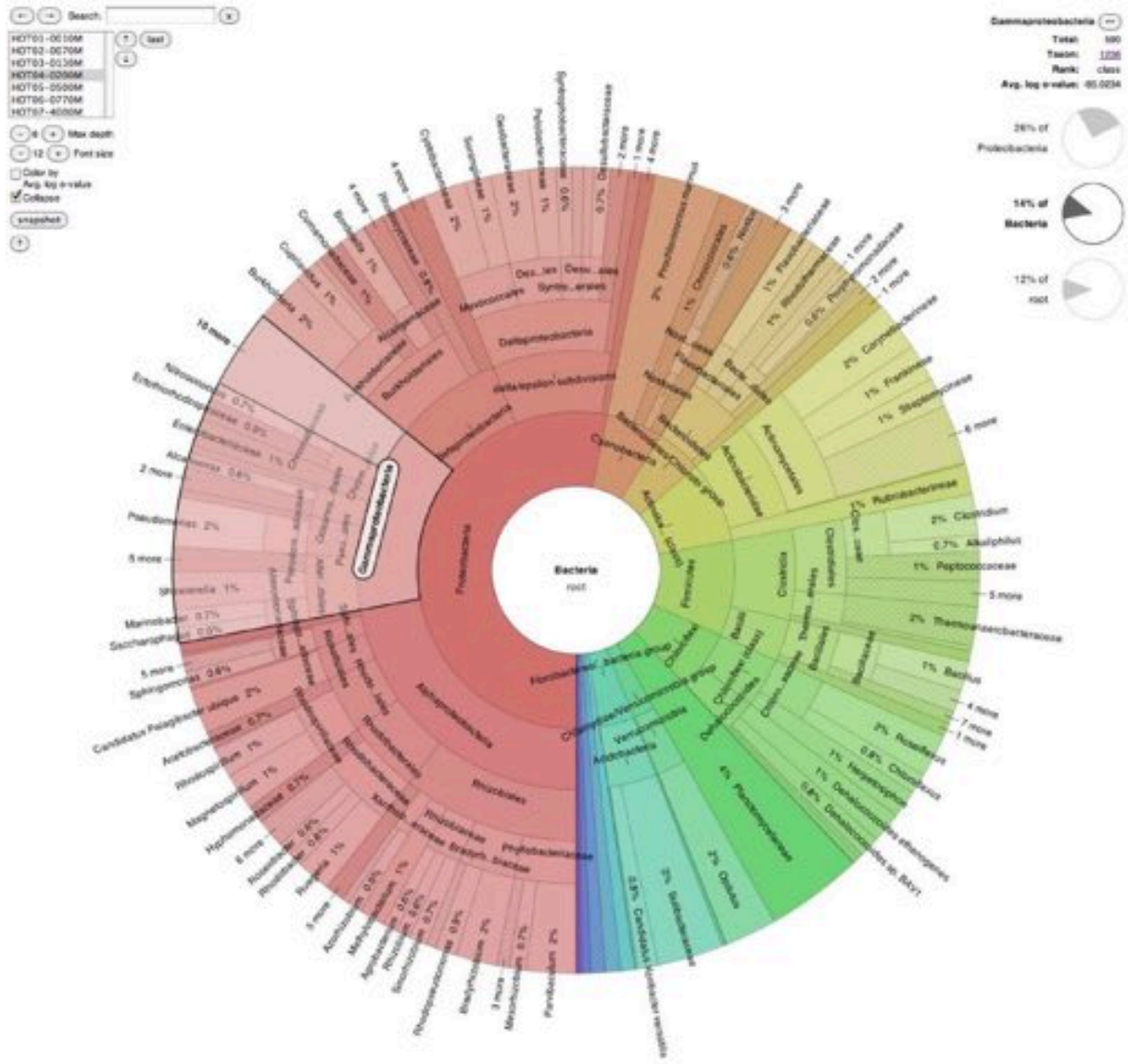
Kraken: ultrafast metagenomic sequence classification using exact alignments
Wood and Salzberg (2014) Genome Biology. DOI: 10.1186/gb-2014-15-3-r46

Metagenomics Benchmarking



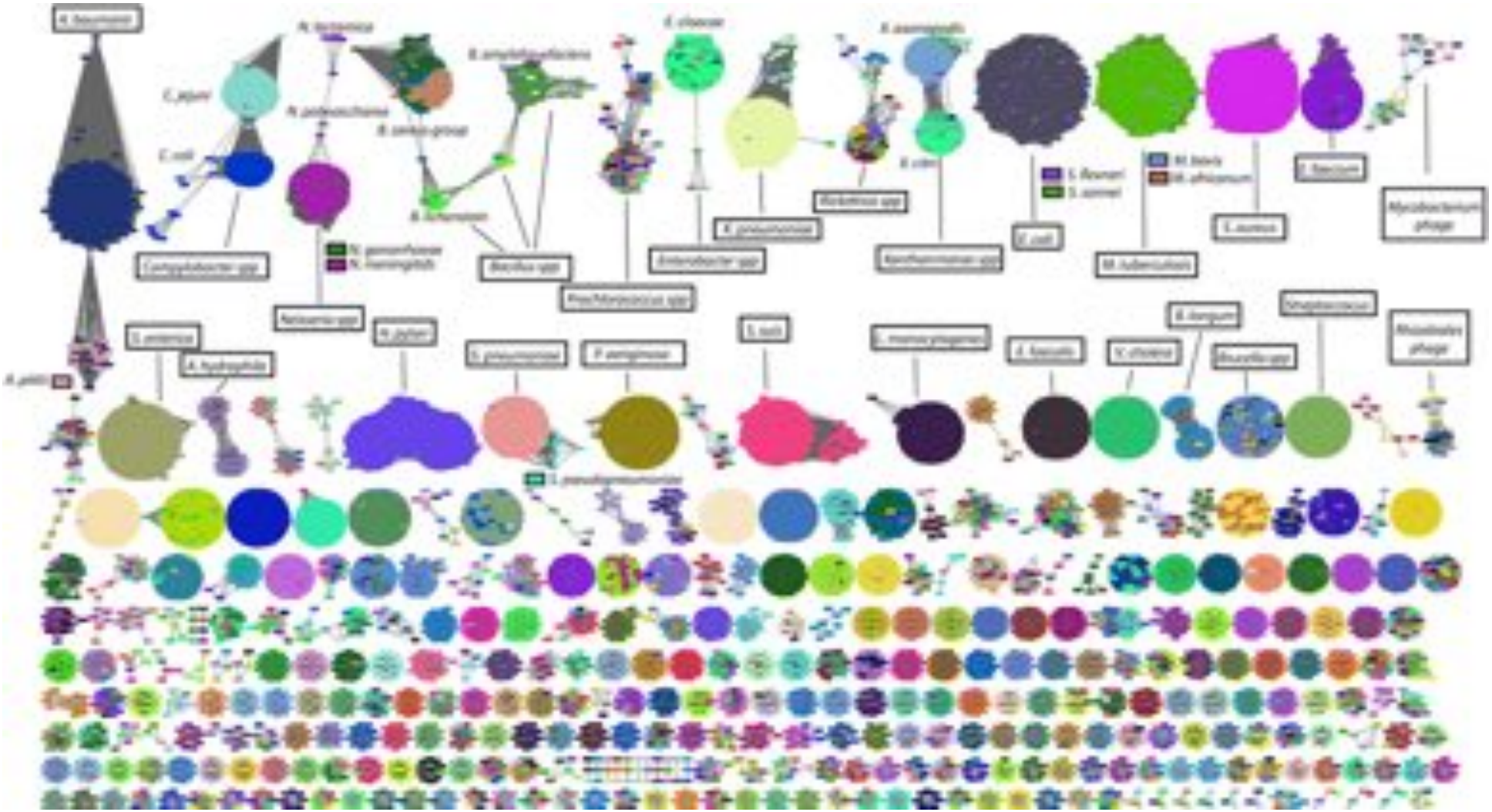
An evaluation of the accuracy and speed of metagenome analysis tools
Lindgreen et al (2016) Scientific Reports. doi:10.1038/srep19233

Krona Plots



Interactive metagenomic visualization in a Web browser
Ondov et al (2011) BMC Bioinformatics. DOI: 10.1186/1471-2105-12-385

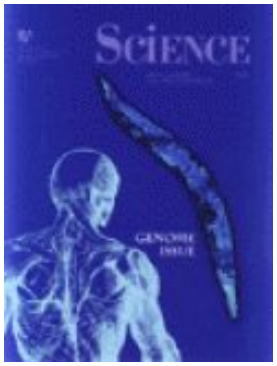
Min-Hash: Comparing all 54,118 RefSeq genomes in 1 day on a laptop



Mash: fast genome and metagenome distance estimation using MinHash
Ondov et al. (2016) Genome Biology. DOI: 10.1186/s13059-016-0997-x

Part IV: Results

The first microbial genomes



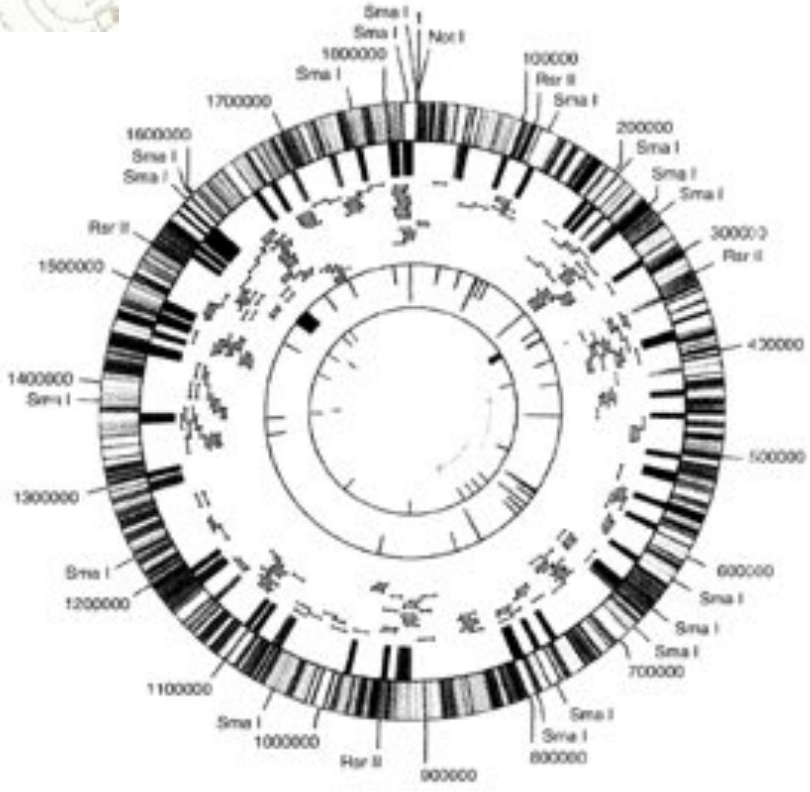
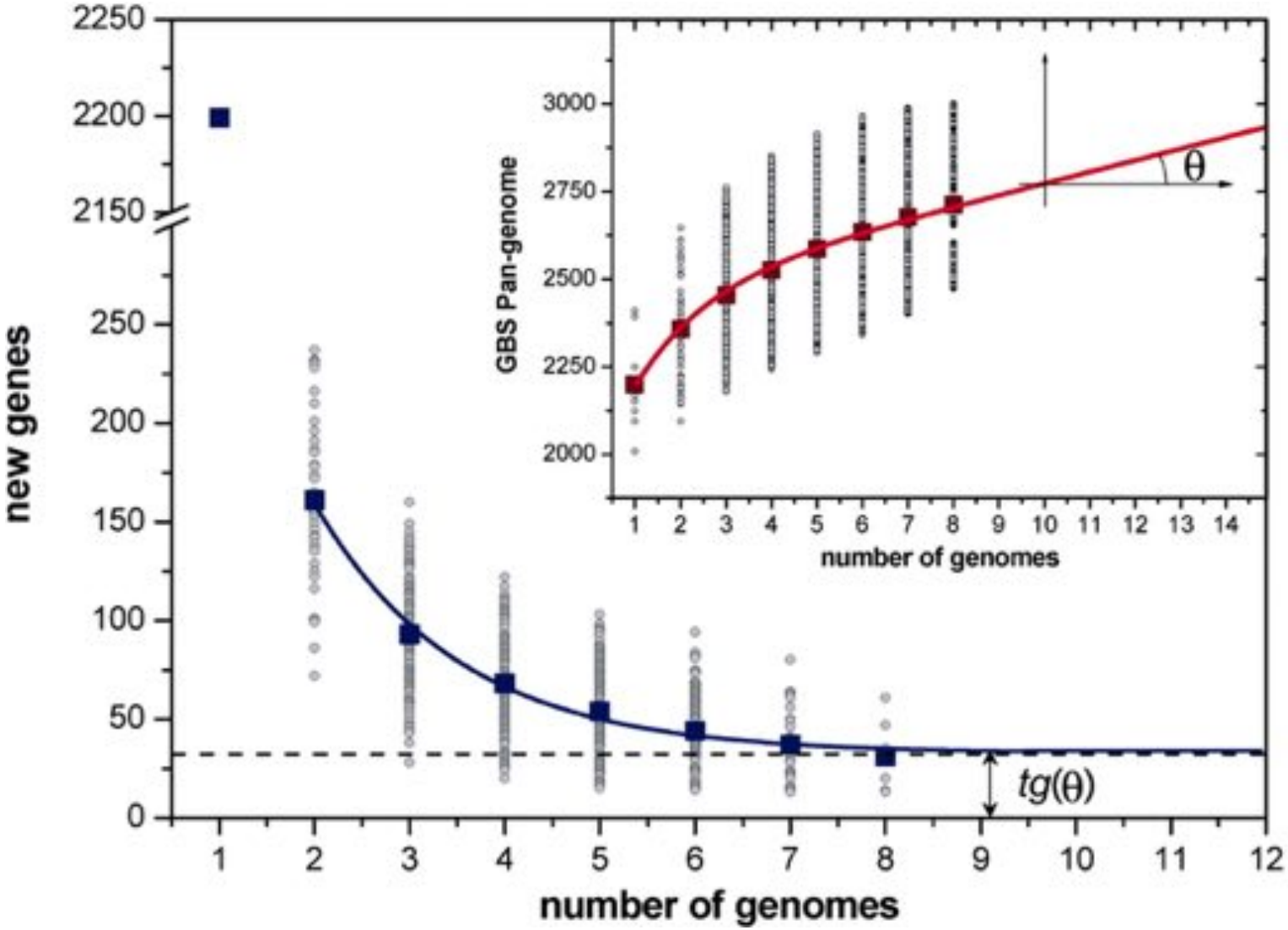


Fig. 1. Gene map of the *M. genitalium* genome. Predicted coding regions are shown, and the direction of transcription is indicated by arrows. Each line in the figure represents 24,000 bp of sequence in the *M. genitalium* genome. Genes are color-coded by role category as described in the key. Gene identification numbers correspond to those in Table 1. The rRNA operon, tRNA genes, and adhesin protein (MgPa) operon repeats are labeled.

Whole-genome random sequencing and assembly of Haemophilus influenzae Rd
Fleischmann et al (1995) Science. doi: 10.1126/science.7542800

The Minimal Gene Complement of *Mycoplasma genitalium*
Fraiser et al (1995) Science. doi: 10.1126/science.270.5235.397

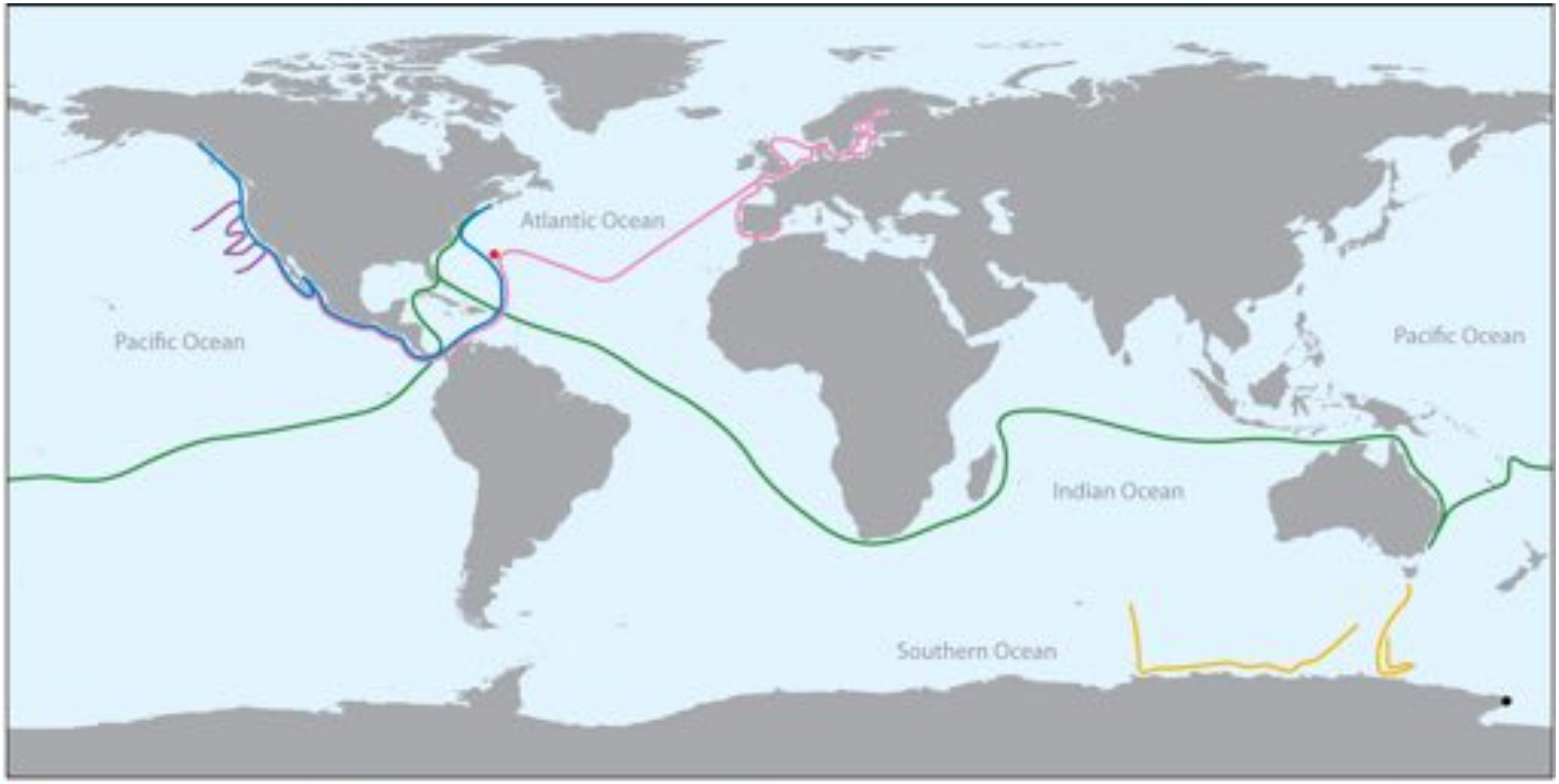
The first pan genome: *Streptococcus agalactiae*



Hervé Tettelin et al. PNAS 2005;102:13950-13955

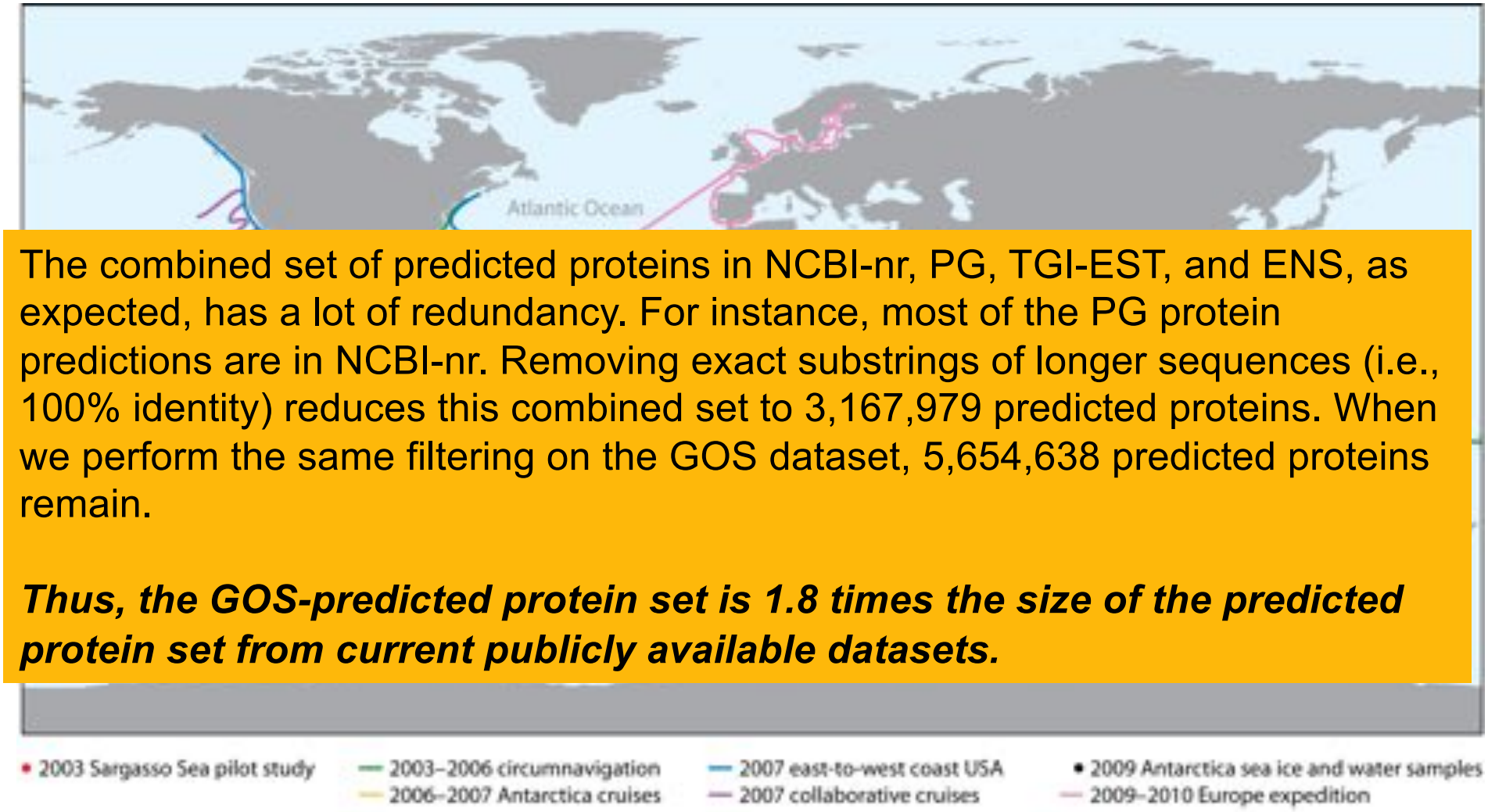
©2005 by National Academy of Sciences

Global Ocean Survey



Gilbert JA, Dupont CL. 2011.
Annu. Rev. Mar. Sci. 3:347–71

Global Ocean Survey

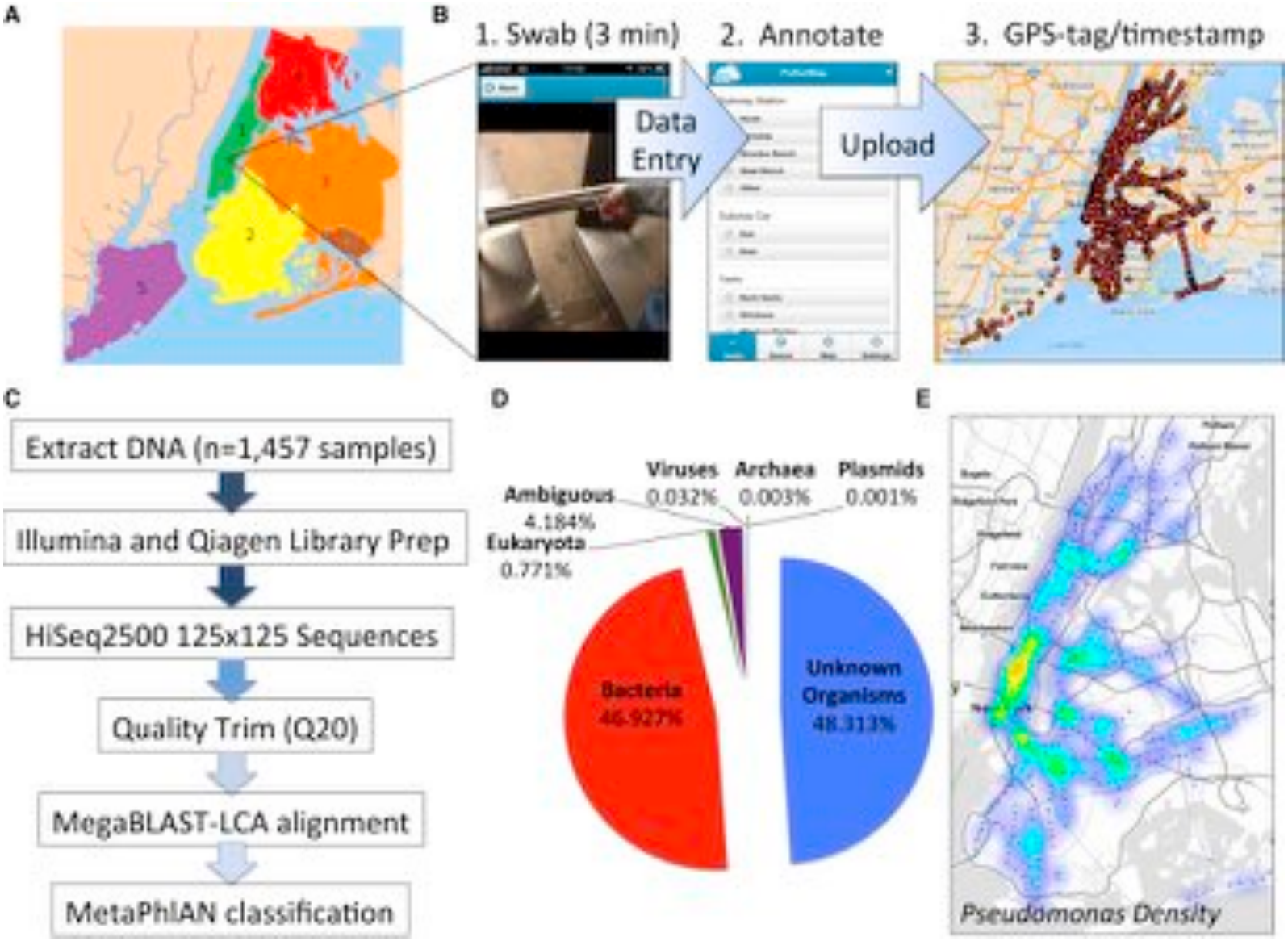


The combined set of predicted proteins in NCBI-nr, PG, TGI-EST, and ENS, as expected, has a lot of redundancy. For instance, most of the PG protein predictions are in NCBI-nr. Removing exact substrings of longer sequences (i.e., 100% identity) reduces this combined set to 3,167,979 predicted proteins. When we perform the same filtering on the GOS dataset, 5,654,638 predicted proteins remain.

Thus, the GOS-predicted protein set is 1.8 times the size of the predicted protein set from current publicly available datasets.

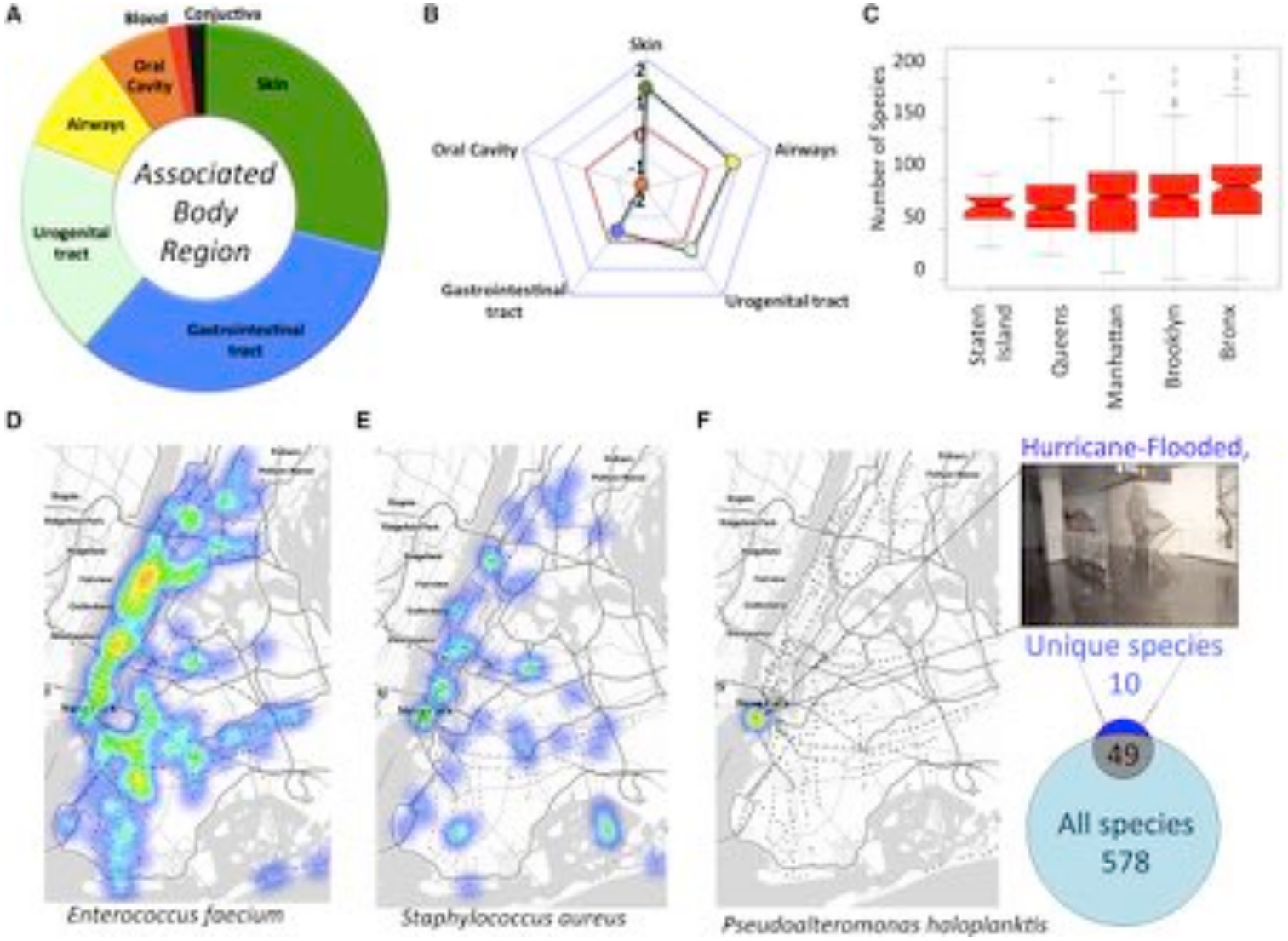
Gilbert JA, Dupont CL. 2011.
Annu. Rev. Mar. Sci. 3:347–71

Metasub



Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics
Afshinnekoo et al (2016) Cell Systems. http://dx.doi.org/10.1016/j.cels.2015.01.001

Different subway stations resembled different body sites



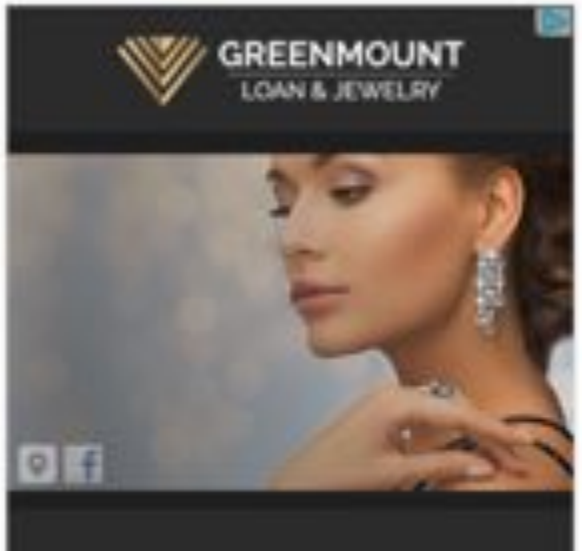
Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics
Afshinnekoo et al (2016) Cell Systems. http://dx.doi.org/10.1016/j.cels.2015.01.001

Dangerous pathogens and mystery microbes ride the subway

FEBRUARY 6, 2015 / 10:42 AM / CBS NEWS

New York City's subway system has never been known for its cleanliness, but even the most jaded city dweller may be shocked and disgusted to learn just what types of microorganisms are lurking on the average subway pole.

A group of researchers led by Christopher Mason of the department of physiology and biophysics at Weill Cornell Medical College swabbed surfaces and collected specimens from the subway system to develop a map of what they called an "urban microbiome." The result, seen above, is called the PathoMap and it illustrates



Bubonic Plague in the Subway System? Don't Worry About It

In October, riders were not deterred after reports that an Ebola-infected man had ridden the subway just before he fell ill. Robert Stolarik for The New York Times