Lecture 23. Genomic Futures

Michael Schatz

April 20, 2020

JHU 600.749: Applied Comparative Genomics

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	ByOhinPho	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

appliedgenomics2020/finalre

github.com/schatzlab/appliedgenomics2020/blob/master/project/finalreport.md

21 lines (16 sloc) 1.45 KB

Raw Blame History

Final Project Report

Due Date: Wednesday, May 13, 2020 @ 11:59pm

Each team should submit a PDF of your final project proposal (6 to 9 pages) to GradeScope by 11:59pm on Wednesday May 13. No late days can be used as grades must be submitted to the registrar on Thursday.

The report should have at least:

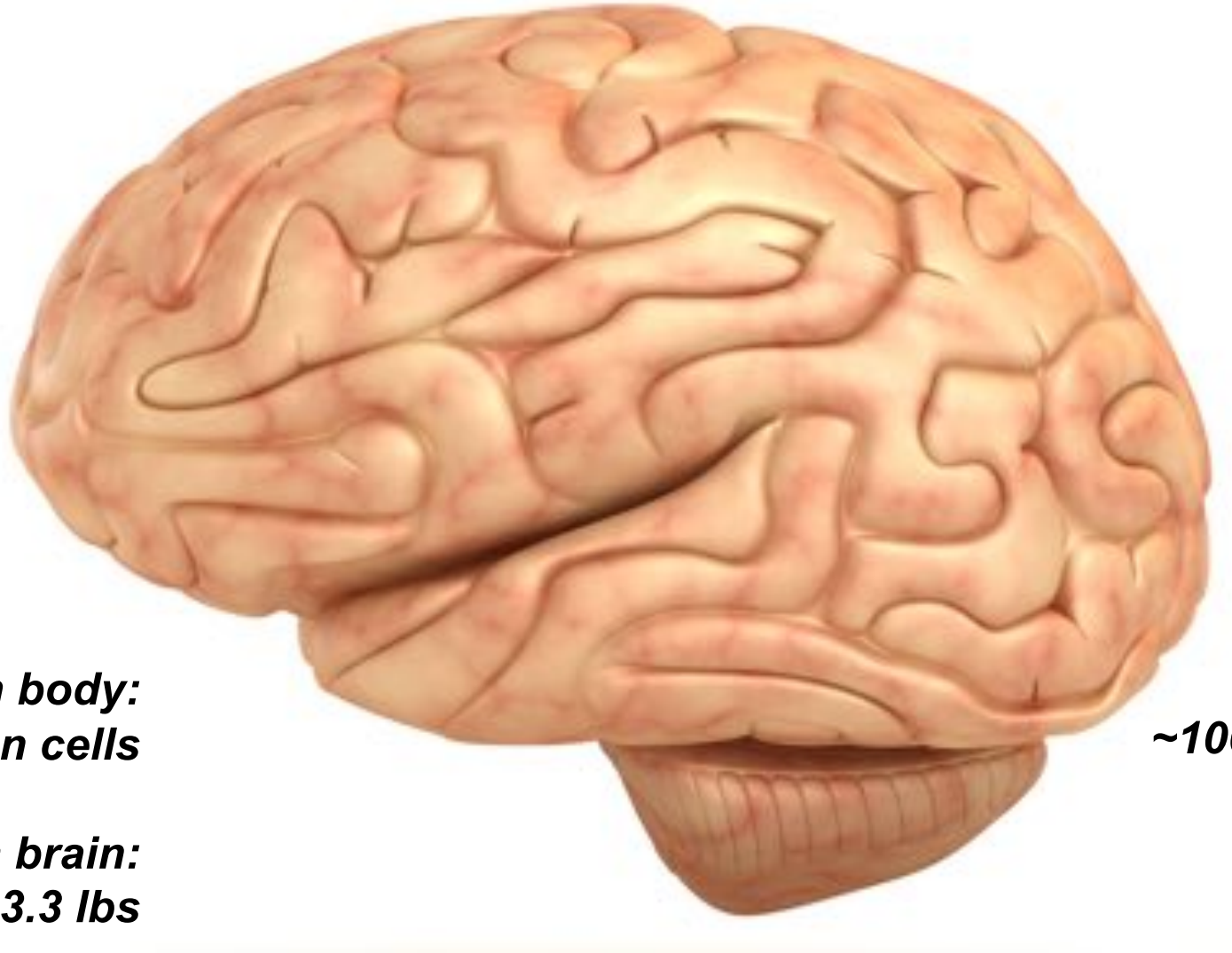
- Title of your project
- List of team members and email addresses
- 1-2 paragraph abstract summarizing the project
- 1-2 pages of Introduction: Background, what is the big problem/question you are addressing, overview of data used, summary of results
- 2-3 pages of Methods that you are using: if you are primarily using existing methods, please describe those methods
- 2-3 pages of Results: be sure to describe the data evaluated along with the results of your analysis. If computational time is measured, please list the machine specifications
- 1 page of Discussion: what you have seen or how that relates to other papers
- Please include 4-6 main figures showing your results. If you have more figures, please include them in a supplemental figures section at the end of the PDF.
- 1 paragraph of acknowledgements
- 1/2 to 1 page of references to relevant papers and data

The report should use the Bioinformatics style template. Word and LaTeX templates are available at https://academic.oup.com/bioinformatics/pages/submission_online. You can (and should) expand on your preliminary report into the full report.

Please use Piazza if you have any general questions!

Part I. Metagenomics

Your second genome?



Human body:
~10 trillion cells

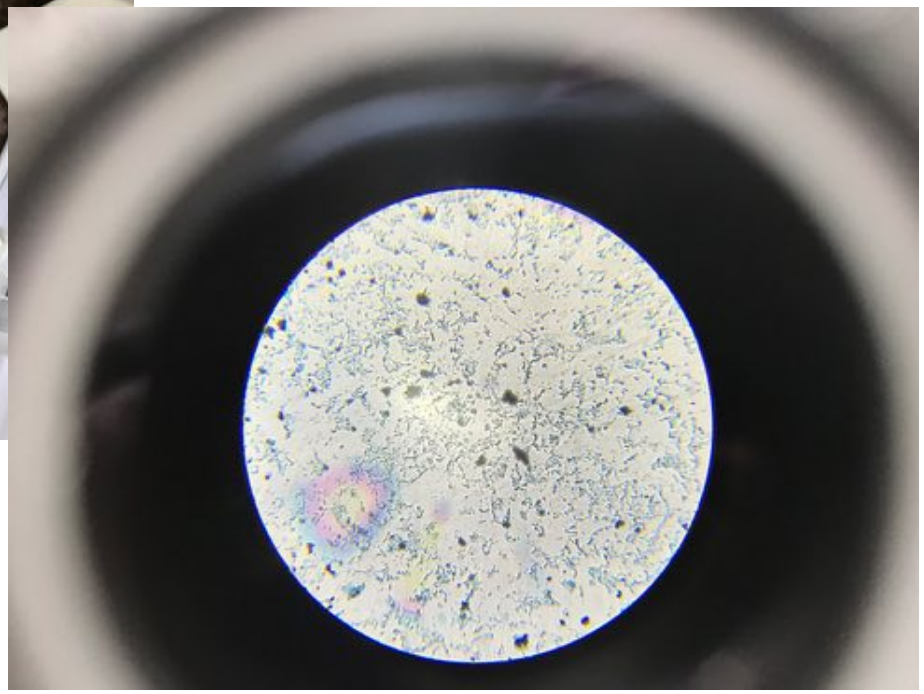
Human brain:
~3.3 lbs

Microbiome
~100 trillion cells

Total mass:
~3.3 lbs

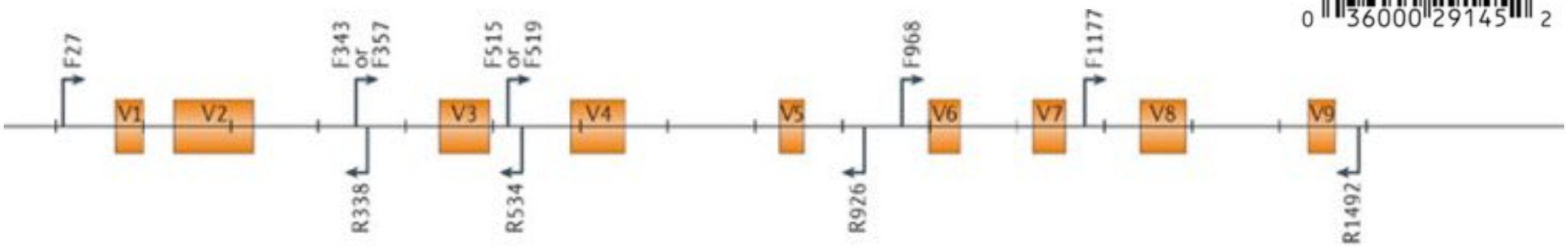
Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans
Sender et al (2016) Cell. http://doi.org/10.1016/j.cell.2016.01.013

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


PRIUS

16S versus shotgun NGS

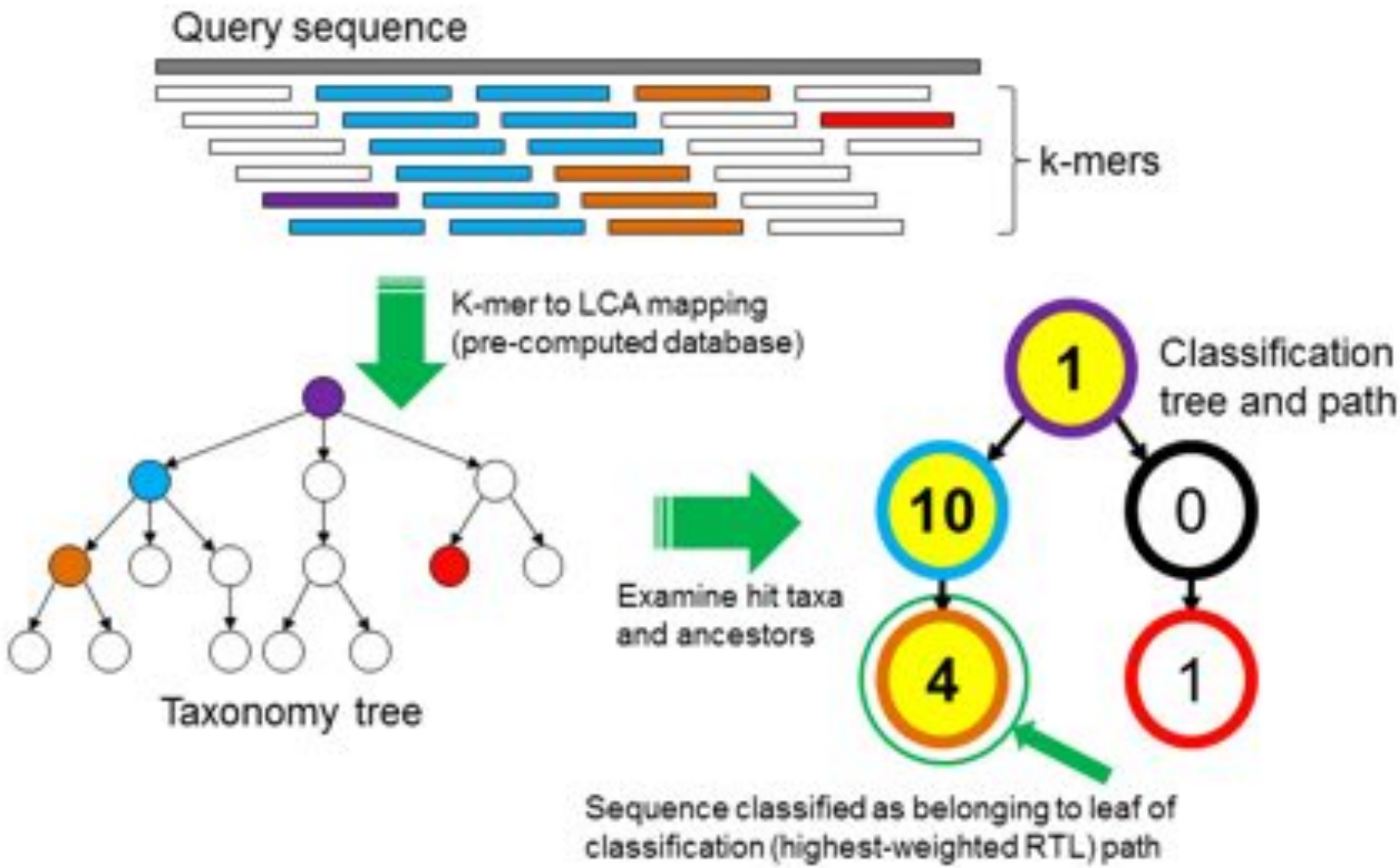
16S

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

NGS

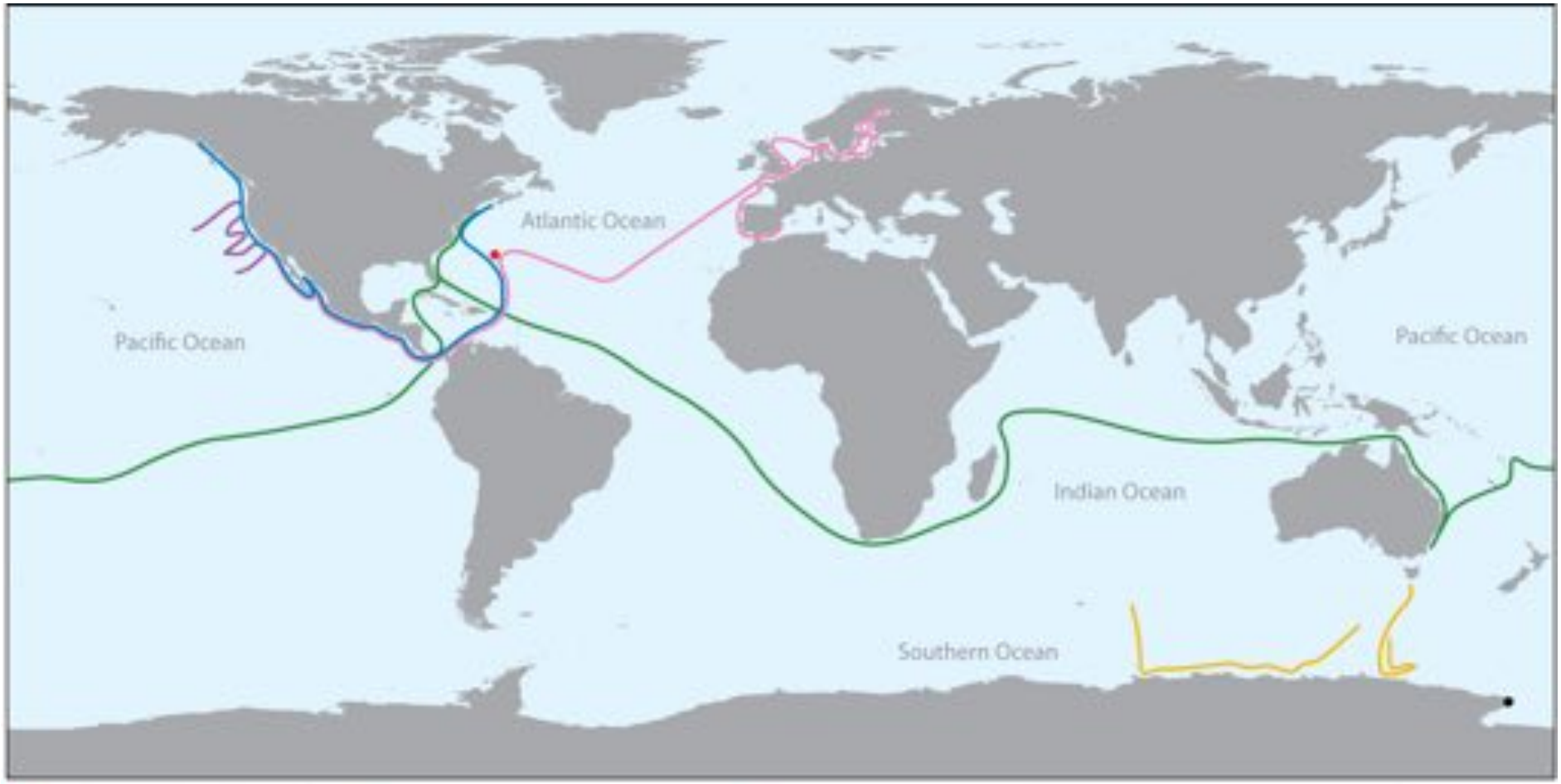
Slower (hours to days)
Whole Metagenome
More expensive per sample
Species/Strain Identification
Genes presence/absence
Variant analysis
Eukaryotic hosts
Can ID fungi, viruses, etc.

Kraken



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

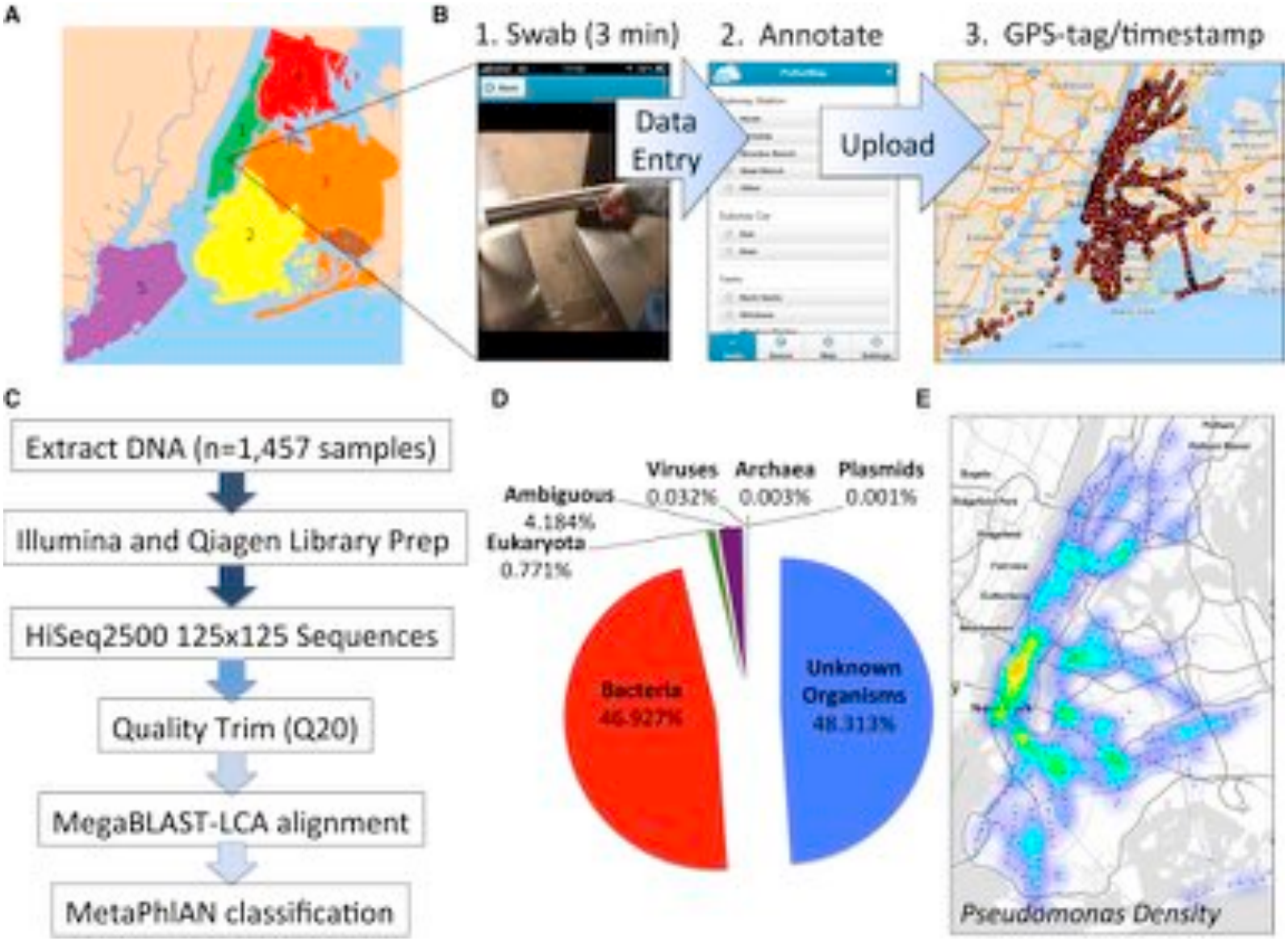
Global Ocean Survey



• 2003 Sargasso Sea pilot study
— 2003–2006 circumnavigation
— 2006–2007 Antarctica cruises
— 2007 east-to-west coast USA
— 2007 collaborative cruises
• 2009 Antarctica sea ice and water samples
— 2009–2010 Europe expedition

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

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

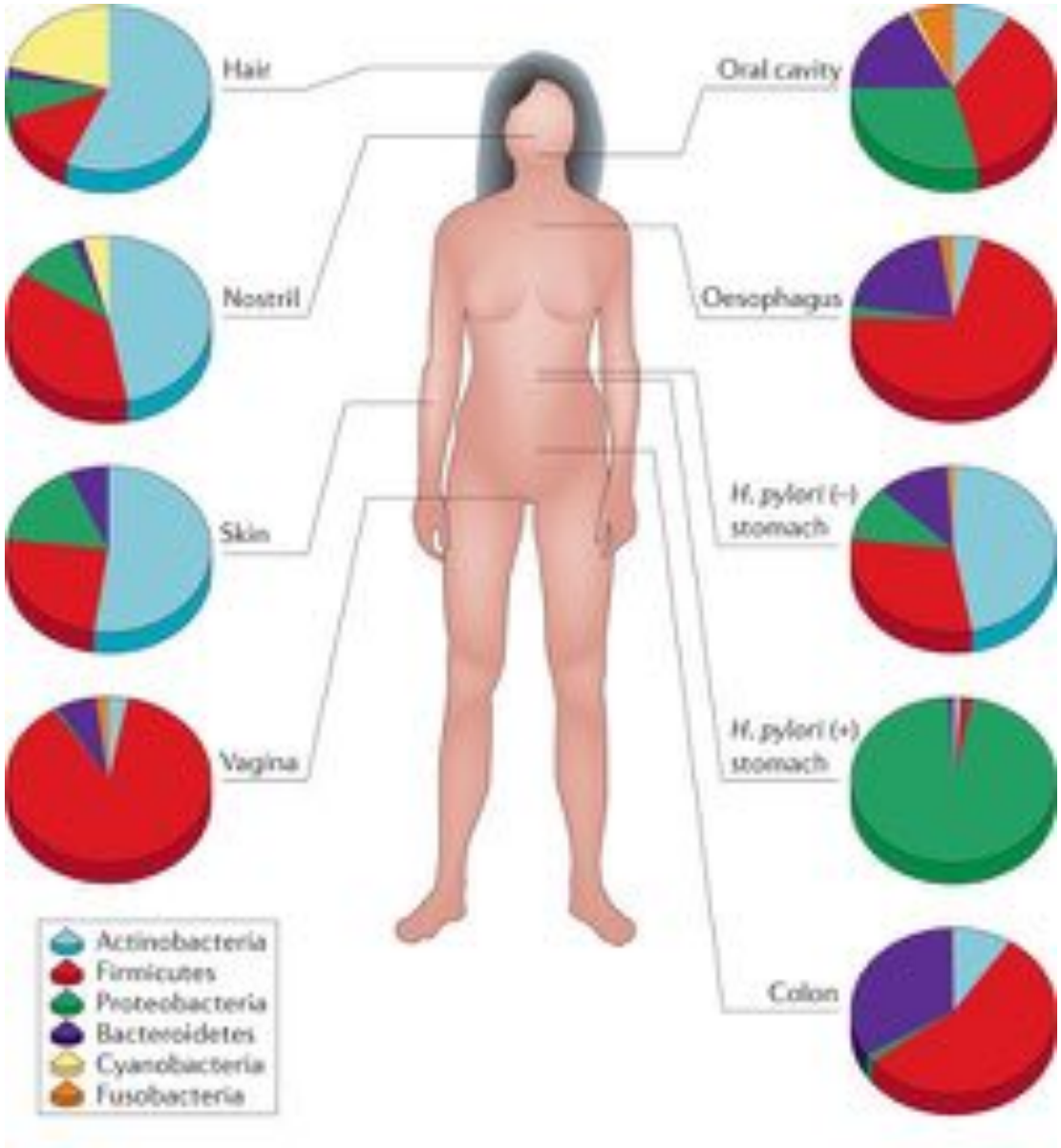
In October, riders were not deterred after reports that an Ebola-infected man had ridden the subway just before he fell ill. Robert Stolarik for The New York Times

Microbes and Human Health

"MICROBE DIET Mice fed microbes from obese people tend to gain fat. Microbes from lean people protect mice from excessive weight gain, even when animals eat a high-fat, low-fiber diet."

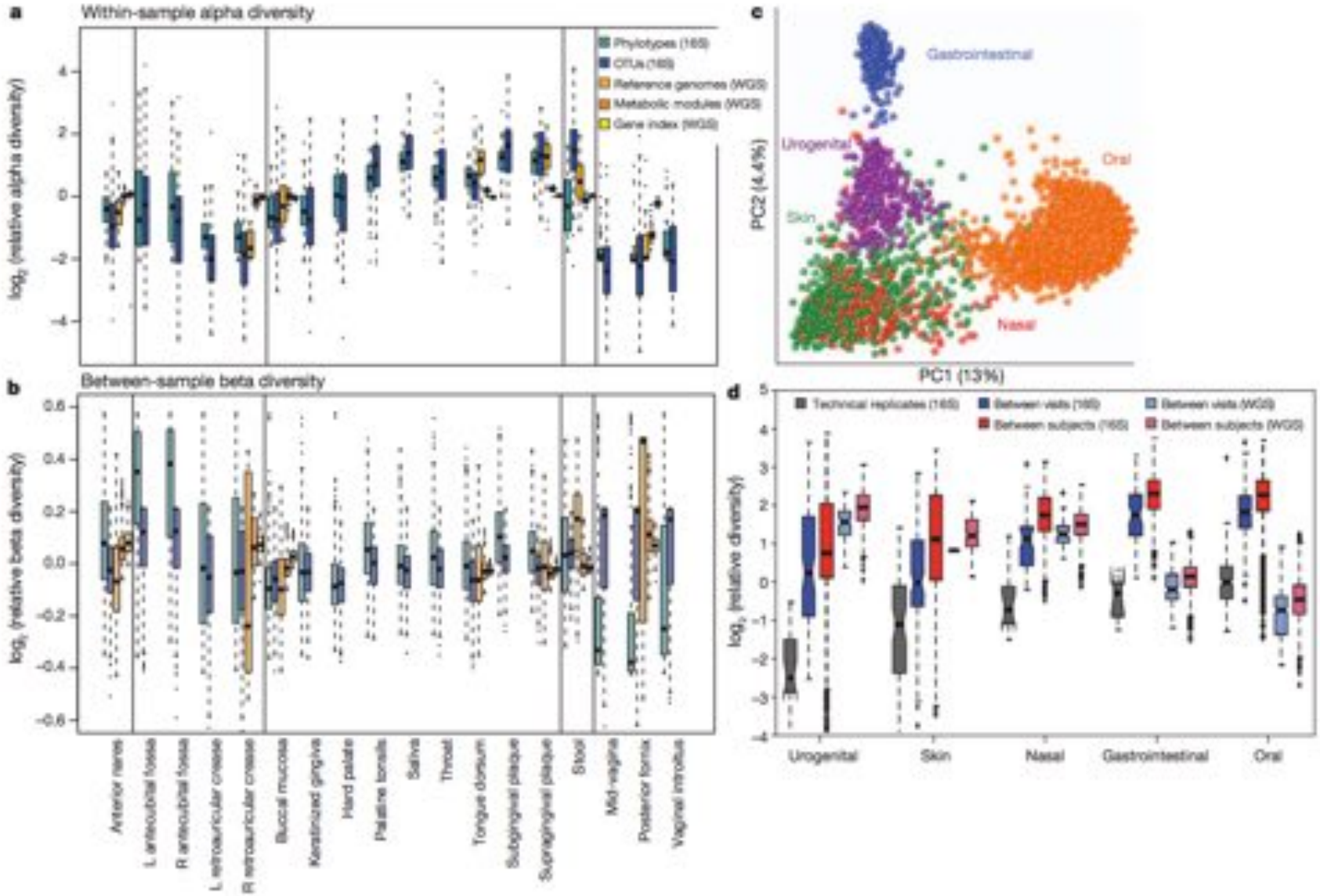
Gut Microbiota from Twins Discordant for Obesity Modulate Metabolism in Mice
Ridaura et al (2013) Science. doi: 10.1126/science.1241214

Microbes and Human Health



The human microbiome: at the interface of health and disease
Cho & Blaser (2012) Nature Reviews Genetics. doi:10.1038/nrg3182

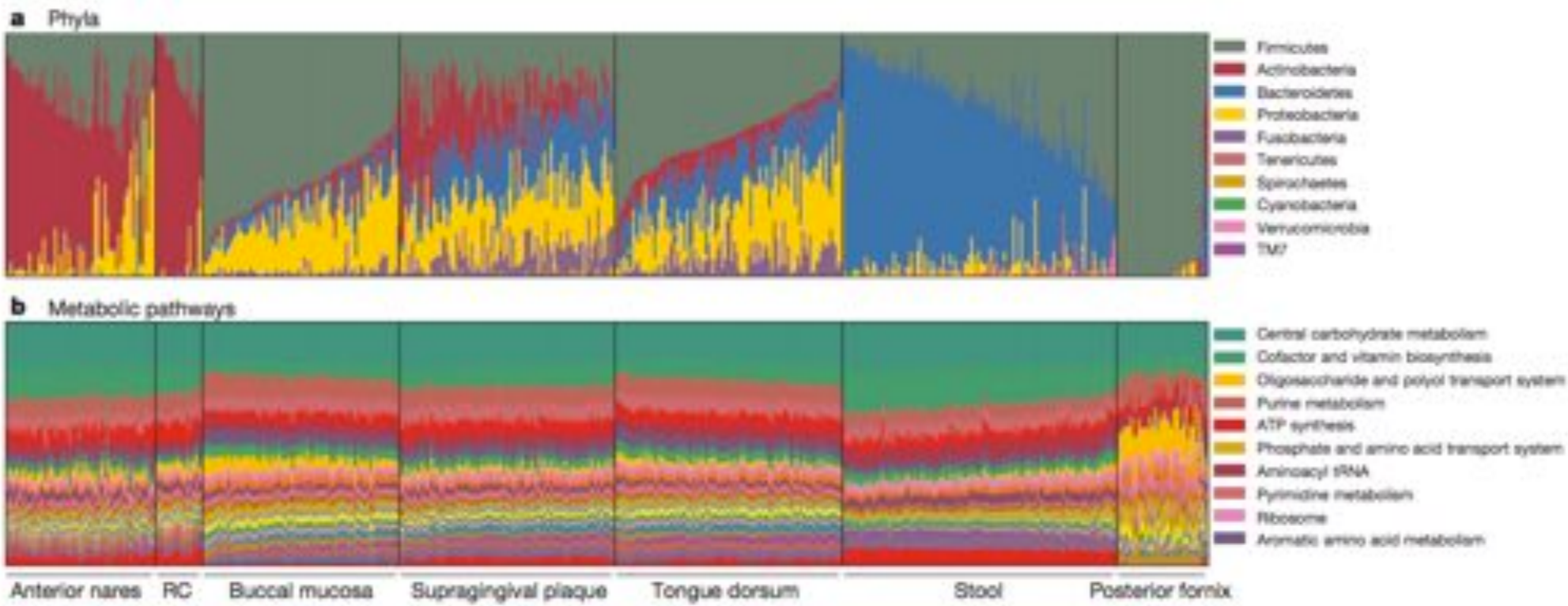
Human Microbiome Project



Structure, function and diversity of the healthy human microbiome

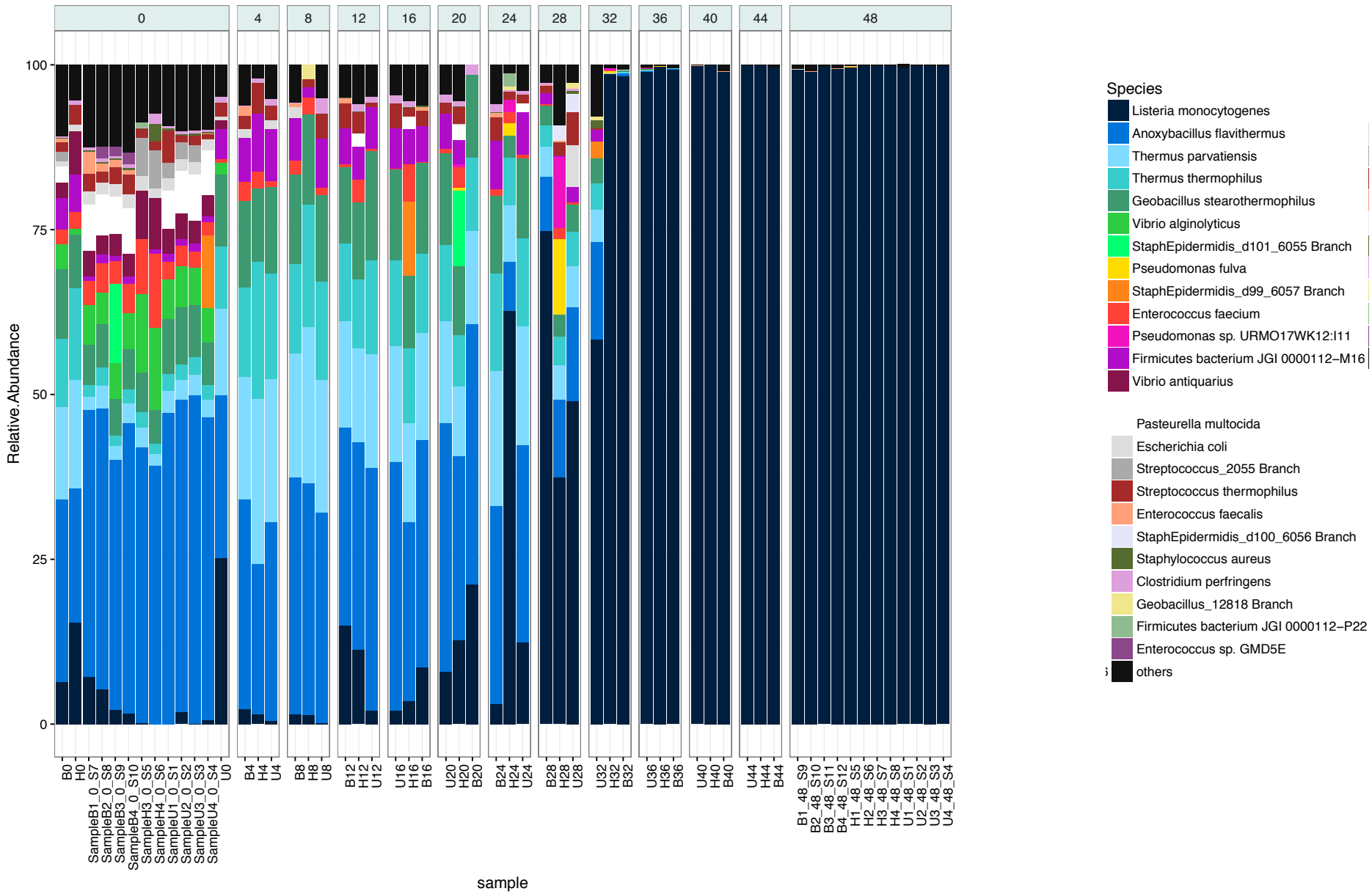
The Human Microbiome Project Consortium (2012) Nature. doi:10.1038/nature11234

Functional composition tends to be more stable than genome composition



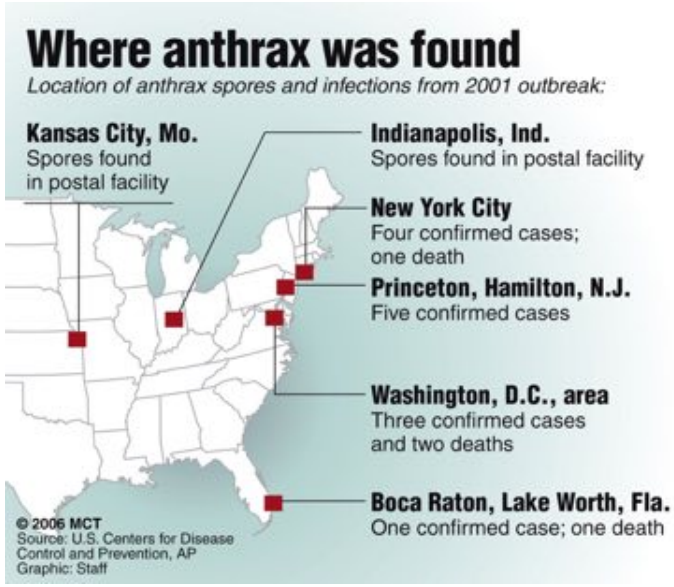
Structure, function and diversity of the healthy human microbiome
The Human Microbiome Project Consortium (2012) Nature. doi:10.1038/nature11234

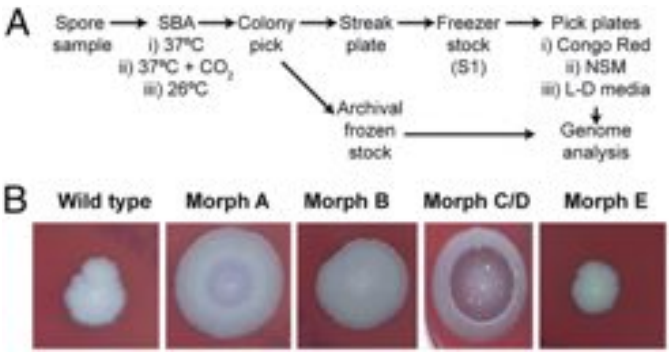
Listeria in ice cream

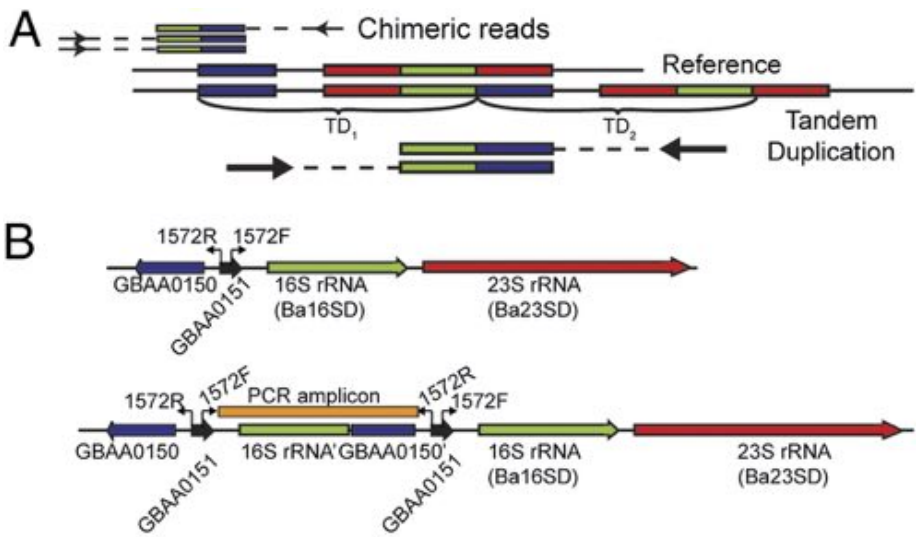


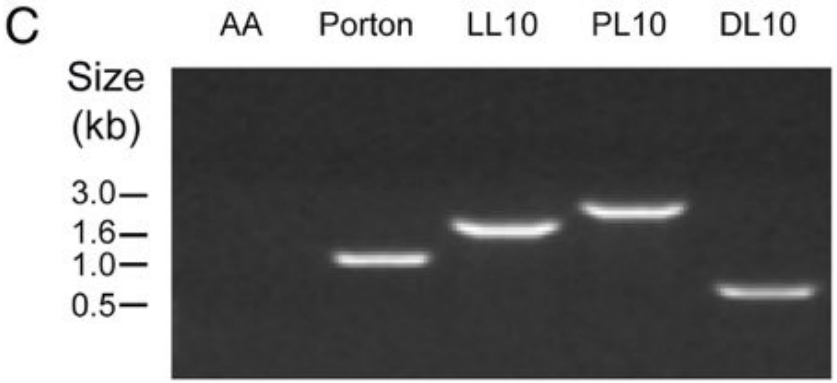
COSMOSID®

Amerithrax Analysis



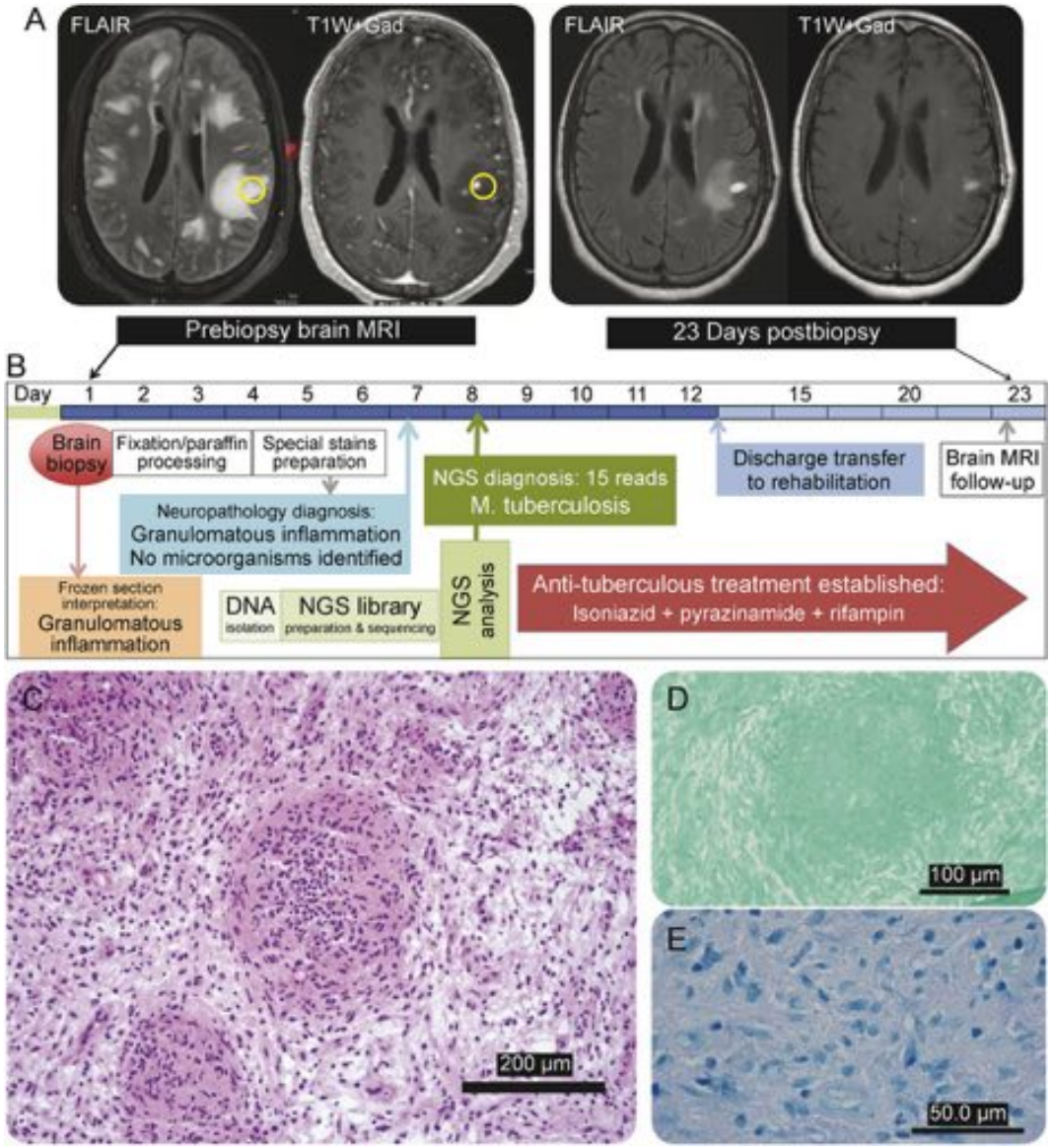




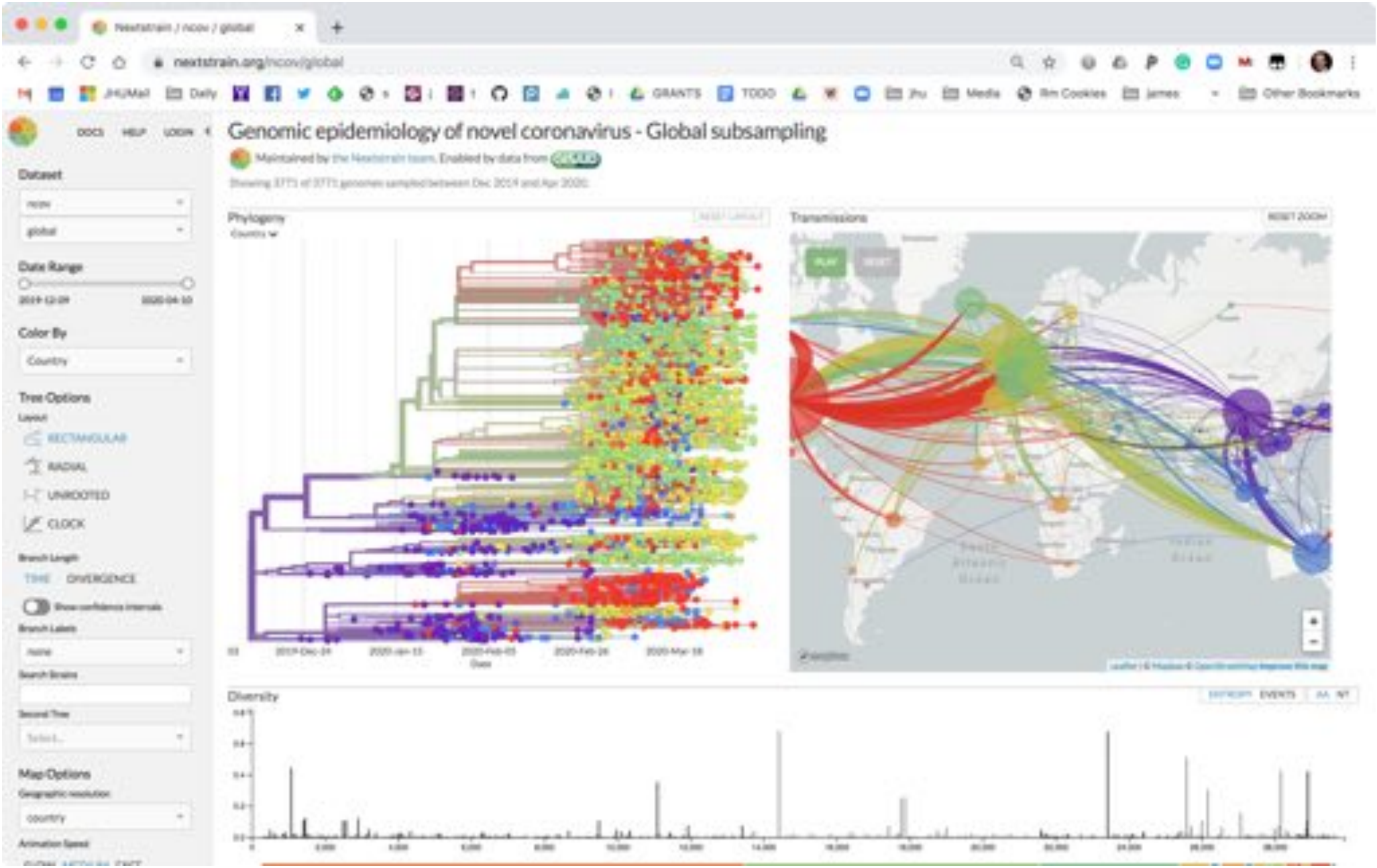


Bacillus anthracis comparative genome analysis in support of the Amerithrax investigation
Rasko et al (2011) PNAS. doi: 10.1073/pnas.1016657108

Diagnosing Brain Infections with NGS



Next-generation sequencing in neuropathologic diagnosis of infections of the nervous system
Salzberg et al (2016) Neurol Neuroimmunol Neuroinflamm dx.doi.org/10.1212/NXI.0000000000000251


nextstrain.org/ncov/global
Genomic epidemiology of novel coronavirus - Global subsampling
Dataset
Date Range
Color By
Tree Options
RECTANGULAR
RADIAL
UNROOTED
CLOCK
Phylogeny
Transmissions
Diversity
Map Options

The Future of Metagenomics

- Applications:
  - WGS metagenomics in the clinic for anaerobic infections and high risk patients (NICU etc.)
  - Surveillance: bioterror agents and epidemiology

- Methods:
  - Single cell, Hi-C, and long read sequencing
  - Computational challenges
    - Species level binning of large datasets
    - Plasmid analysis (antimicrobial resistance genes)
    - Going from associations to specific mechanisms
    - Functional analysis

Part II:

Genetic Privacy

Science
AAAS

Identifying Personal Genomes by Surname Inference

Melissa Gymrek *et al.*

Science **339**, 321 (2013);
DOI: 10.1126/science.1229566

What are microsatellites

- **Tandemly repeated sequence motifs**
  - Motifs are 1 – 6 nt long
  - So far, min. 8 nt length, min. 3 tandem repeats for our analyses
- **Ubiquitous in human genome**
  - >5.7 million uninterrupted microsatellites in hg19
- **Extremely unstable**
  - Mutation rate thought to be ~10^{-3} per generation in humans
- **Unique mutation mechanism**
  - Replication slippage during mitosis and meiosis
- **May be under neutral selection**

cCTCTCTCTCTCTCTCTCTCTCTCTCa ➔ $(CT)_{13}$ tCAACAACAACAACAACAACAACAAa ➔ $(CAA)_7$

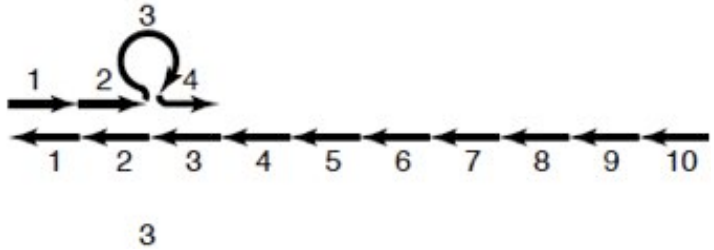
tTTGTCTTGTCTTGTCTTGTCTTGTCTTGTCc ➔ $(TTGTC)_6$ cCATTCATTCATTCATTa ➔ $(CATT)_4$

Microsatellites: Simple Sequences with Complex Evolution
Ellegren (2004) *Nature Reviews Genetics*. doi:10.1038/nrg1348

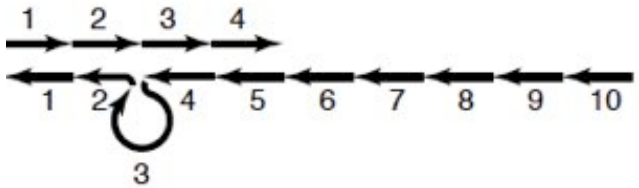
Replication slippage

- **Out-of-phase re-annealing**
  - Nascent and template strands dissociate and re-anneal out-of-phase
- **Loops repaired by mismatch repair machinery (MMR)**
  - Very efficient for small loops
  - Possible strand-specific repair
- **Stepwise process**
  - Nascent strand gains or loses full repeat units
  - Typically single unit mutations
- **Varies by motif length, motif composition, etc.**

Expansion:

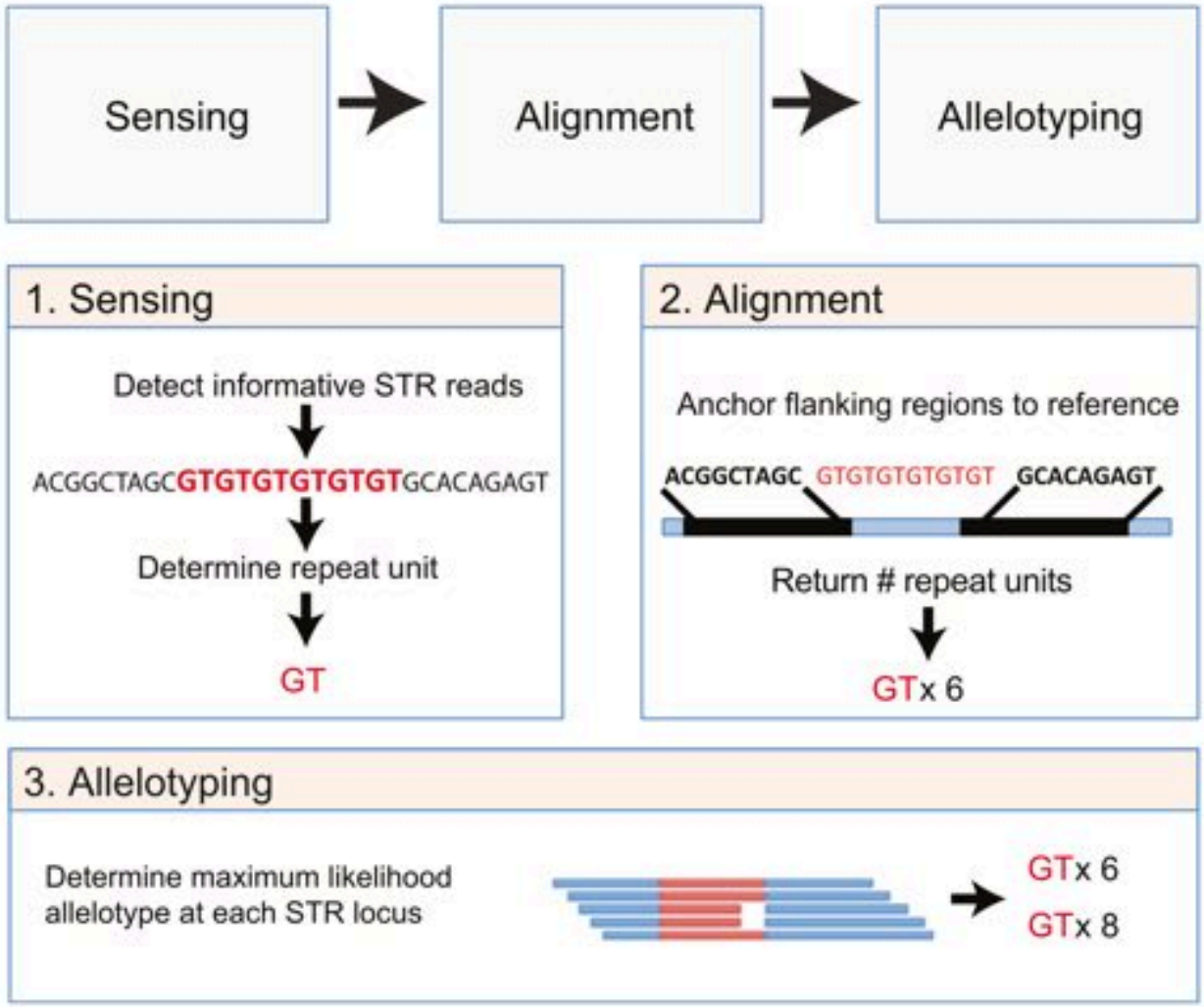


Contraction:



Microsatellites: Simple Sequences with Complex Evolution
Ellegren (2004) *Nature Reviews Genetics*. doi:10.1038/nrg1348

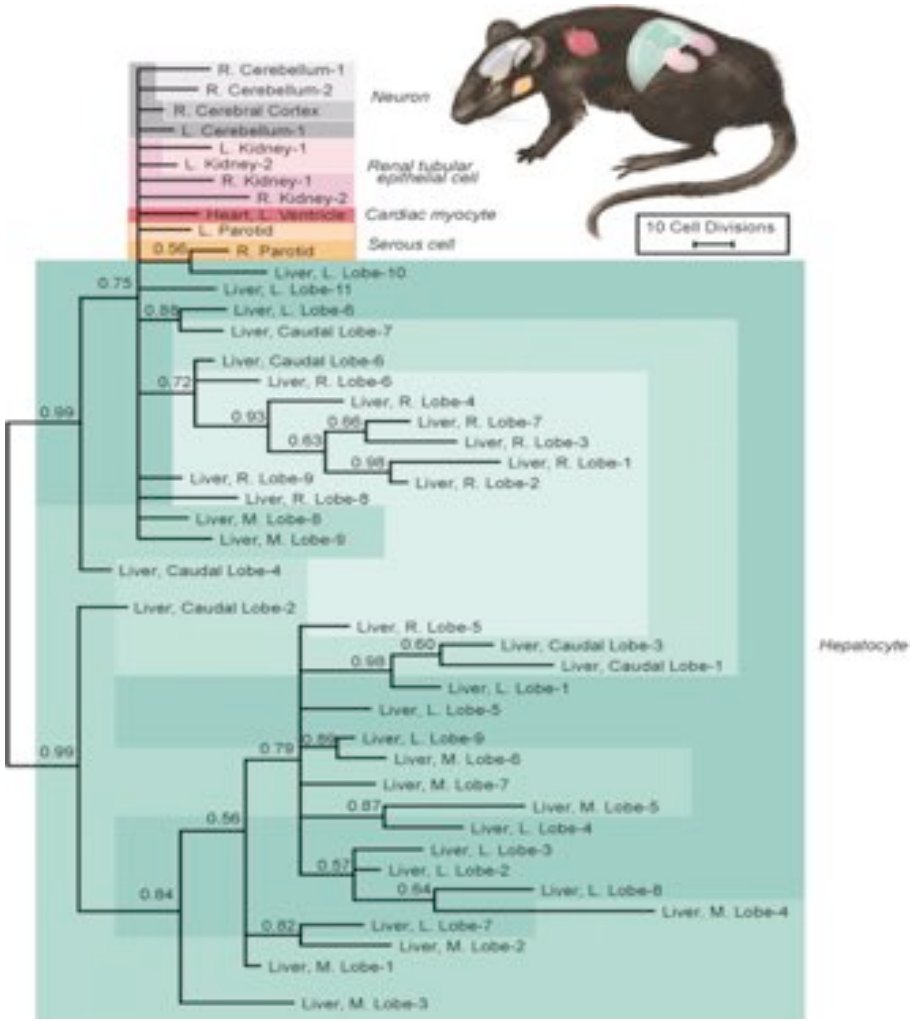
lobSTR Algorithm Overview



lobSTR: A short tandem repeat profiler for personal genomes
Gymrek et al. (2012) *Genome Research.* doi:10.1101/gr.135780.111

Why should we care about microsatellites?

- Polymorphism and mutation rate variation
- Disease
  - Huntington's Disease
  - Fragile X syndrome
  - Friedrich's ataxia
- Mutations as lineage
  - Organogenesis/embryonic development
  - Tumor development



Phylogenetic fate mapping
Salipante (2006) *PNAS*. doi: 10.1073/pnas.0601265103

Combined DNA Index System (CODIS)

The Combined DNA Index System, or CODIS, blends forensic science and computer technology into a tool for linking violent crimes. It enables federal, state, and local forensic laboratories to exchange and compare DNA profiles electronically, thereby linking serial violent crimes to each other and to known offenders. Using the National DNA Index System of CODIS, the National Missing Persons DNA Database also helps identify missing and unidentified individuals.

Overview

CODIS generates investigative leads in cases where biological evidence is recovered from the crime scene. Matches made among profiles in the Forensic Index can link crime scenes together, possibly identifying serial offenders. Based upon a match, police from multiple jurisdictions can coordinate their respective investigations and share the leads they developed independently. Matches made between the Forensic and Offender Indexes provide investigators with the identity of suspected perpetrators. Since names and other personally identifiable information are not stored at NDIS, qualified DNA analysts in the laboratories sharing matching profiles contact each other to confirm the candidate match.

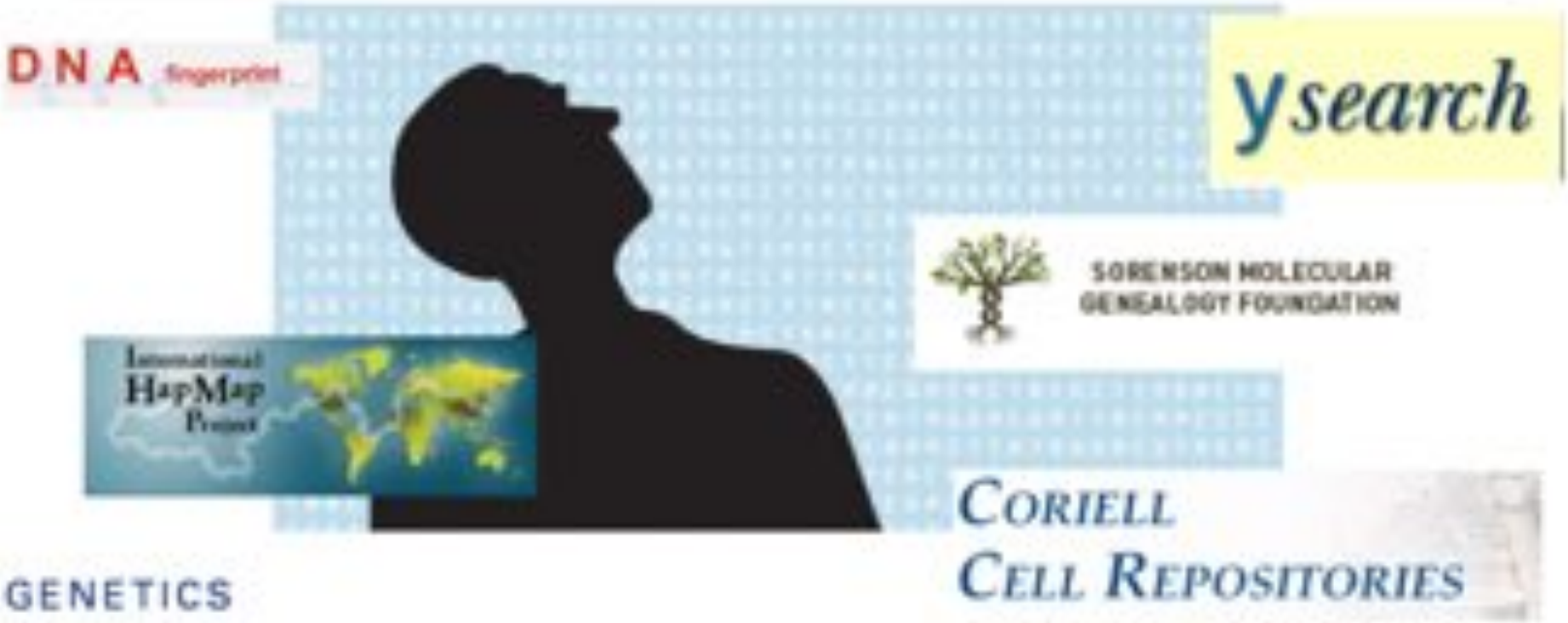
History

The FBI Laboratory's CODIS began as a pilot software project in 1990, serving 14 state and local laboratories. The DNA Identification Act of 1994 formalized the FBI's authority to establish a National DNA Index System (NDIS) for law enforcement purposes. Today, over 190 public law enforcement laboratories participate in NDIS across the United States. Internationally, more than 90 law enforcement laboratories in over 50 countries use the CODIS software for their own database initiatives.

Mission

The CODIS Unit manages CODIS and NDIS. It is responsible for developing, providing, and supporting the CODIS program to federal, state, and local crime laboratories in the United States and selected international law enforcement crime laboratories to foster the exchange and comparison of forensic DNA evidence from violent crime investigations. The CODIS Unit

Genealogy Databases



GENETICS

Genealogy Databases Enable Naming Of Anonymous DNA Donors

Surname Inference

Whose sequence reads are these?

Identifying Personal Genomes by Surname Inference
Gymrek et al (2013) *Science*. doi: 10.1126/science.1229566

Step 1. Profile Y-STRs from the individual's genome.

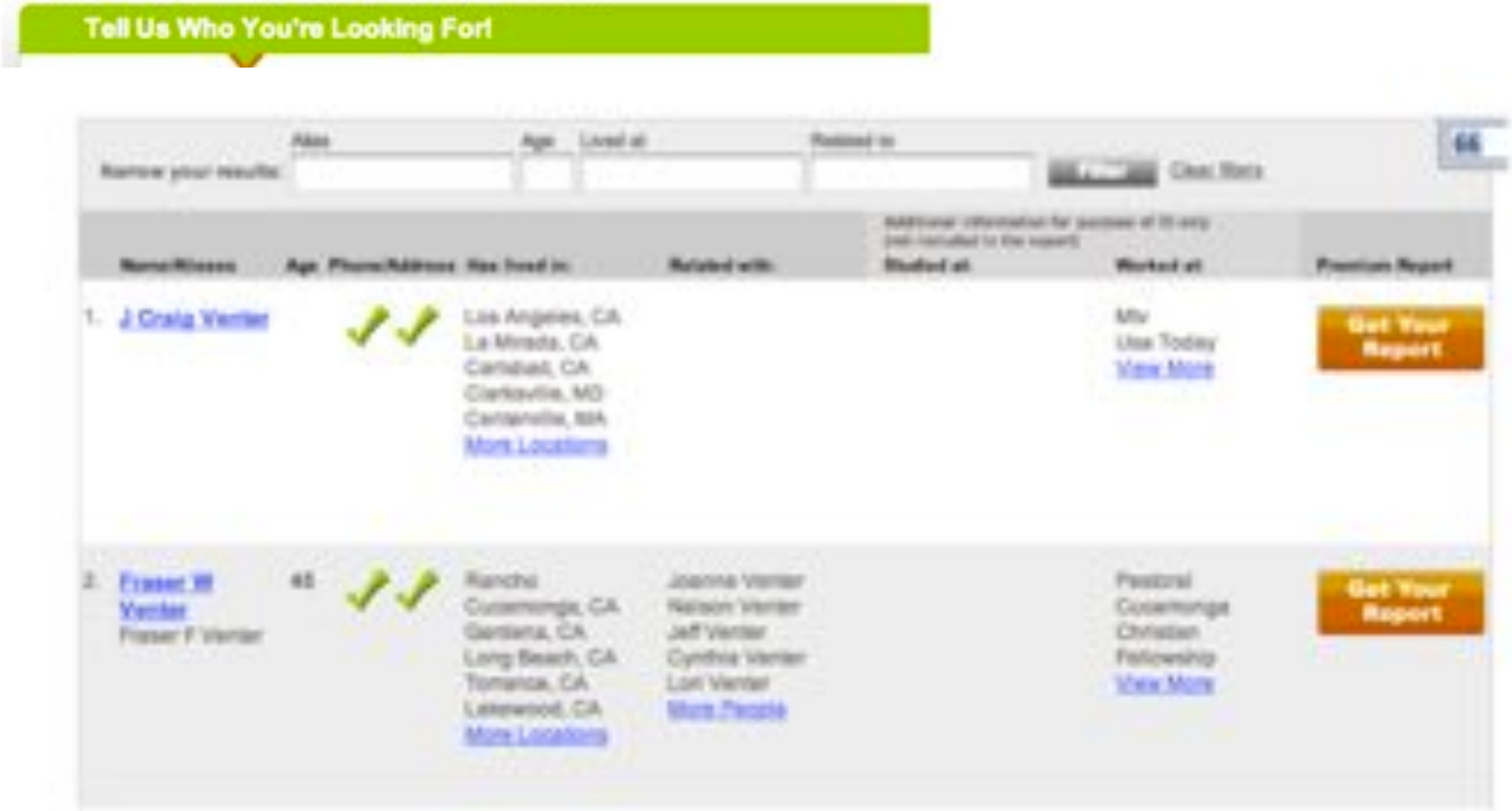
The human reference genome contains 16 copies of "TTTC". Venter has an extra copy of "TTTC", giving him a genotype of "17" at this marker. In a similar way, we can profile all other genealogical STR markers on the Y-chromosome where we know Venter's genome sequence to get the value of a whole panel of these markers.

Step 2. Search for a surname hit in online genetic genealogy databases.

http://www.ysearch.org

Step 3. Search with additional metadata to narrow down the individual.

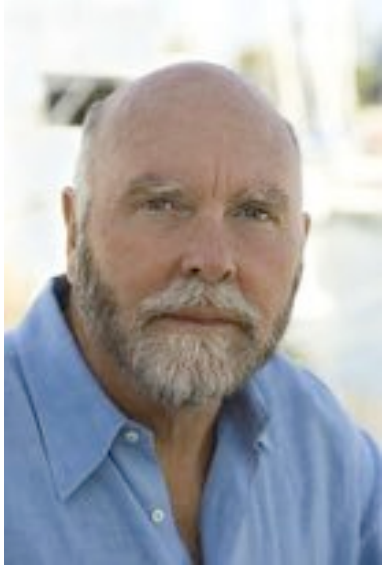
We enter the search information: Venter, CA, and 66:



http://www.ussearch.com

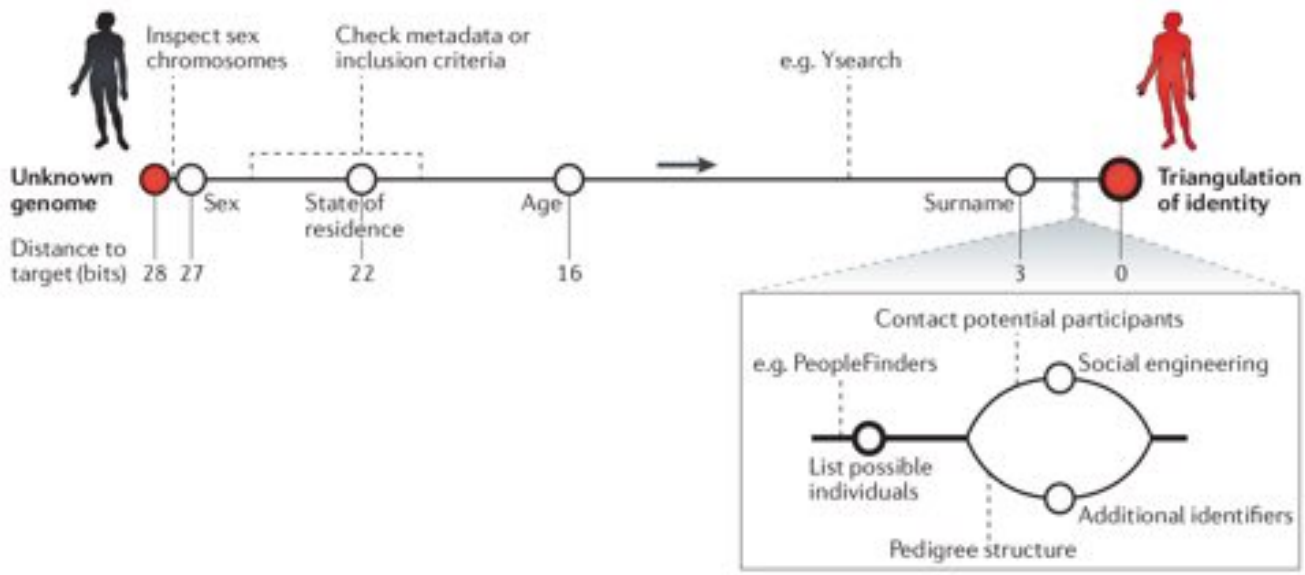
Surname Inference

It's Craig Venter!



Identifying Personal Genomes by Surname Inference
Gymrek et al (2013) *Science*. doi: 10.1126/science.1229566

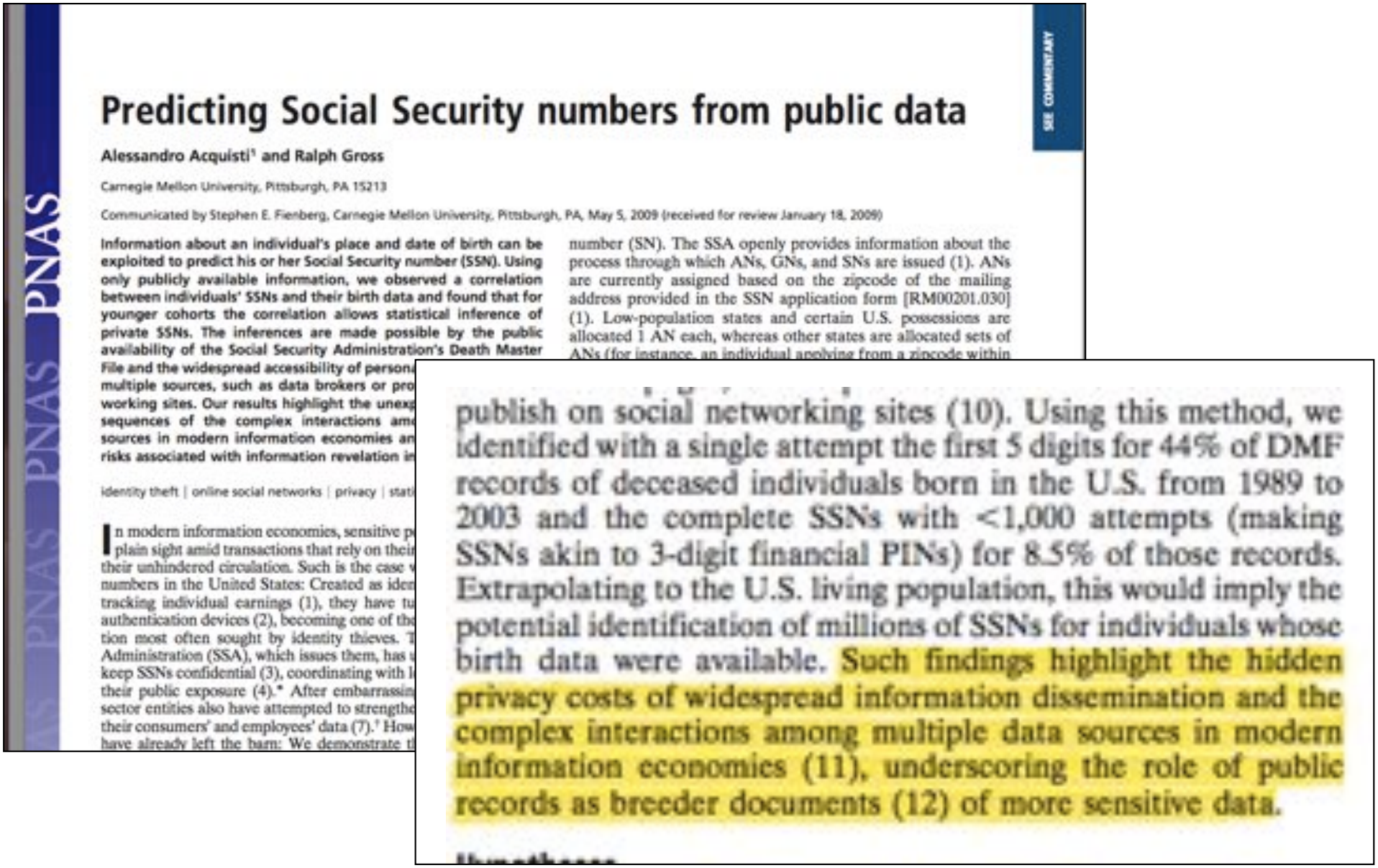
Possible route for identity tracing



- *US population: ~313.9 million individuals*
- *$\log_2$ 313,900,000 = 28.226 bits*
- *Sex ~ 1.0 information bits*
- *$\log_2$ 156,950,000 = 27.226 bits*

- Tracing attacks combine metadata and surname inference to triangulate the identity of an unknown individual.
- With no information, there are roughly 300 million matching individuals in the US, equating to 28.0 bits of entropy.
- Sex reduces entropy by 1 bit, state of residence and age reduces to 16, successful surname inference reduces to ~3 bits.

The risks of big data?



PNAS

SEE COMMENTARY

Predicting Social Security numbers from public data

Alessandro Acquisti[1] and Ralph Gross

Carnegie Mellon University, Pittsburgh, PA 15213

Communicated by Stephen E. Fienberg, Carnegie Mellon University, Pittsburgh, PA, May 5, 2009 (received for review January 18, 2009)

Information about an individual's place and date of birth can be exploited to predict his or her Social Security number (SSN). Using only publicly available information, we observed a correlation between individuals' SSNs and their birth data and found that for younger cohorts the correlation allows statistical inference of private SSNs. The inferences are made possible by the public availability of the Social Security Administration's Death Master File and the widespread accessibility of persona multiple sources, such as data brokers or pro working sites. Our results highlight the unexp sequences of the complex interactions am sources in modern information economies an risks associated with information revelation i

identity theft | online social networks | privacy | stati

In modern information economies, sensitive p plain sight amid transactions that rely on thei their unhindered circulation. Such is the case numbers in the United States: Created as ide tracking individual earnings (1), they have tu authentication devices (2), becoming one of th tion most often sought by identity thieves. Administration (SSA), which issues them, has keep SSNs confidential (3), coordinating with their public exposure (4).* After embarrassin sector entities also have attempted to strength their consumers' and employees' data (7).† Ho have already left the barn: We demonstrate t

number (SN). The SSA openly provides information about the process through which ANs, GNs, and SNs are issued (1). ANs are currently assigned based on the zipcode of the mailing address provided in the SSN application form [RM00201.030] (1). Low-population states and certain U.S. possessions are allocated 1 AN each, whereas other states are allocated sets of ANs (for instance, an individual applying from a zipcode within

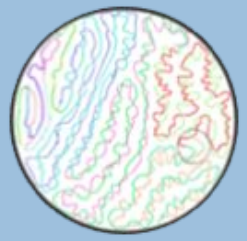
publish on social networking sites (10). Using this method, we identified with a single attempt the first 5 digits for 44% of DMF records of deceased individuals born in the U.S. from 1989 to 2003 and the complete SSNs with <1,000 attempts (making SSNs akin to 3-digit financial PINs) for 8.5% of those records. Extrapolating to the U.S. living population, this would imply the potential identification of millions of SSNs for individuals whose birth data were available. Such findings highlight the hidden privacy costs of widespread information dissemination and the complex interactions among multiple data sources in modern information economies (11), underscoring the role of public records as breeder documents (12) of more sensitive data.

Genomic Futures?

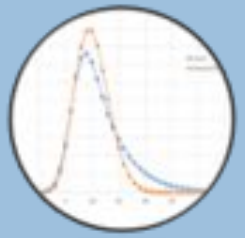
The rise of a digital immune system
Schatz & Phillippy (2012) GigaScience 1:4

Computational Research Landscape

- ***Avoid***
  - New Illumina/PacBio base callers
  - Entirely new genome assembler from scratch
- ***Good***
  - Alignment/Assembly/Analysis methods robust to errors, polyploidy, aneuploidy
  - Use insights from long-reads to improve analysis of short-reads
- ***Best***
  - Synthesis of large numbers of samples ("pan-genome assembly") and/or multiple data types ("multi-omics")
  - Prioritization and interpretation of variations



FALCON

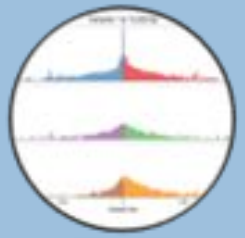


LRSim

SURVIVOR

NGM+Sniffles

Ribbon



Assemblytics

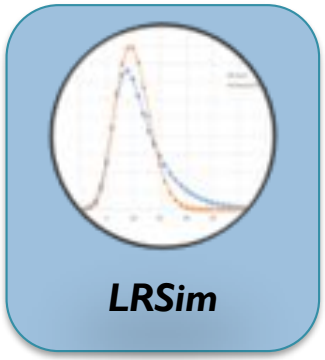
http://schatz-lab.org

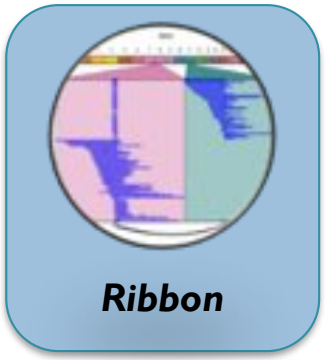
Computational Research Landscape

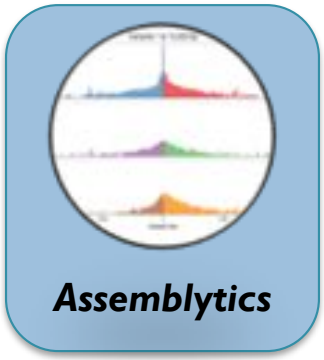
- ***Avo***
  -
  -
- ***Goo***
  -
  -
- ***Bes***
  -
  -

Also consider starting a company!

neuploidy







http://schatz-lab.org