Pan-genomics: theory & practice

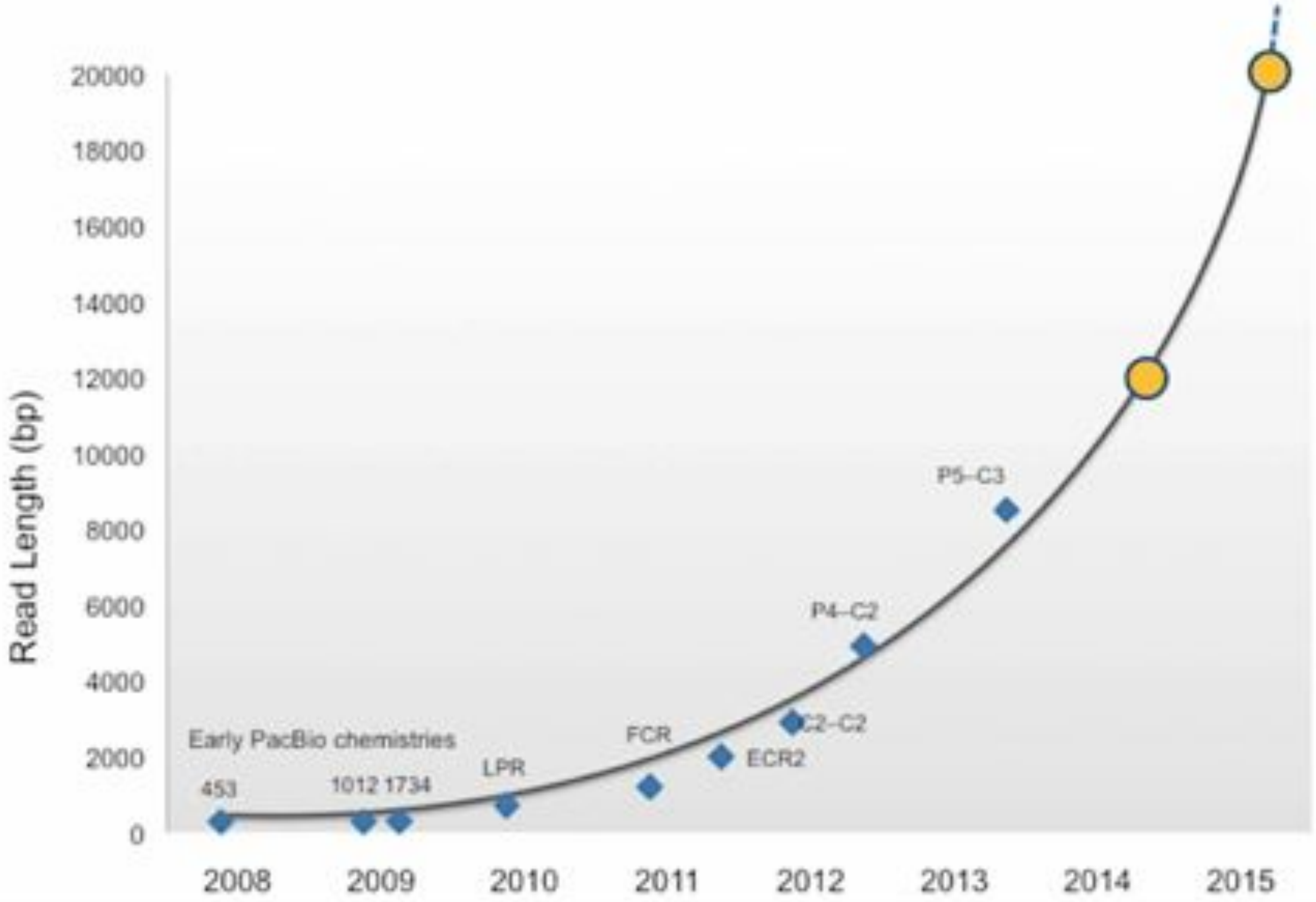
Michael Schatz

Sept 20, 2014
GRC Assembly Workshop

#gi2014 / @mike_schatz

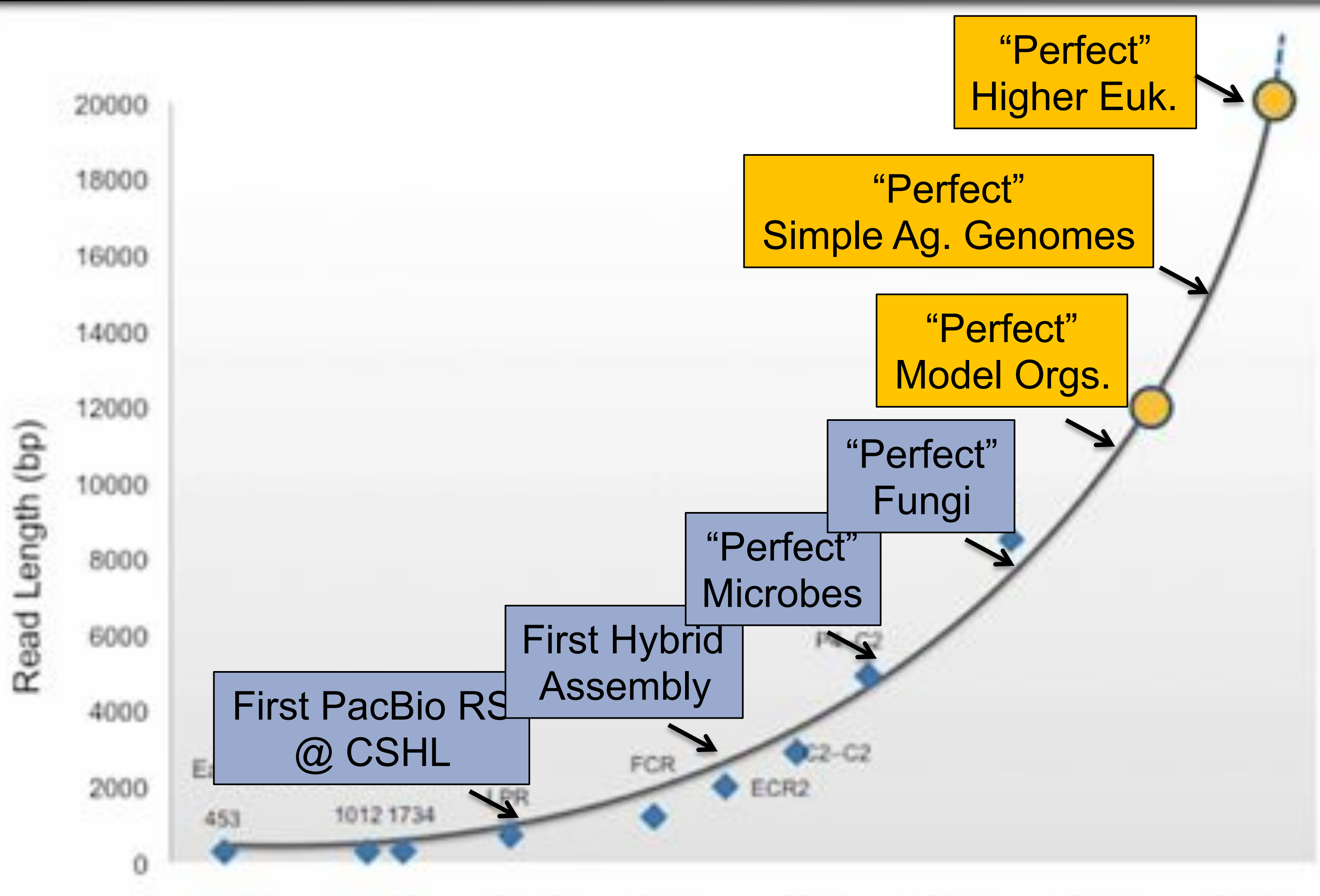
Part 1: Theory

PacBio® Advances in Read Length



Advances in Assembly

“Perfect” Human Assembly



“Perfect” Higher Euk.

“Perfect” Simple Ag. Genomes

“Perfect” Model Orgs.

“Perfect” Fungi

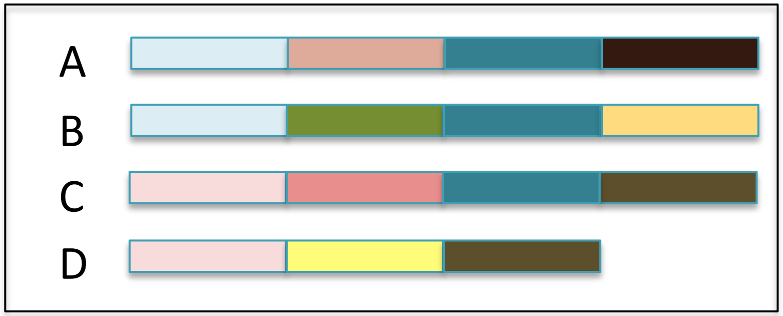
“Perfect” Microbes

First Hybrid Assembly

First PacBio RS @ CSHL

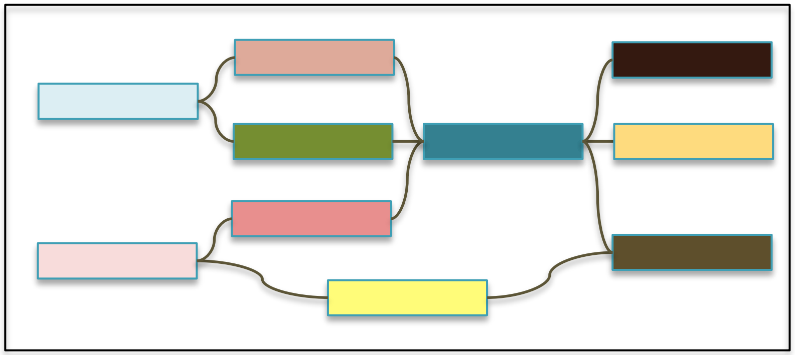
Error correction and assembly complexity of single molecule sequencing reads.
Lee, H*, Gurtowski, J*, Yoo, S, Marcus, S, McCombie, WR, Schatz, MC
http://www.biorxiv.org/content/early/2014/06/18/006395

Pan-Genome Alignment & Assembly



Time to start considering problems for which *N* complete genomes are the input to study the "pan-genome"

- Available today for many microbial species, near future for higher eukaryotes



Pan-genome colored de Bruijn graph

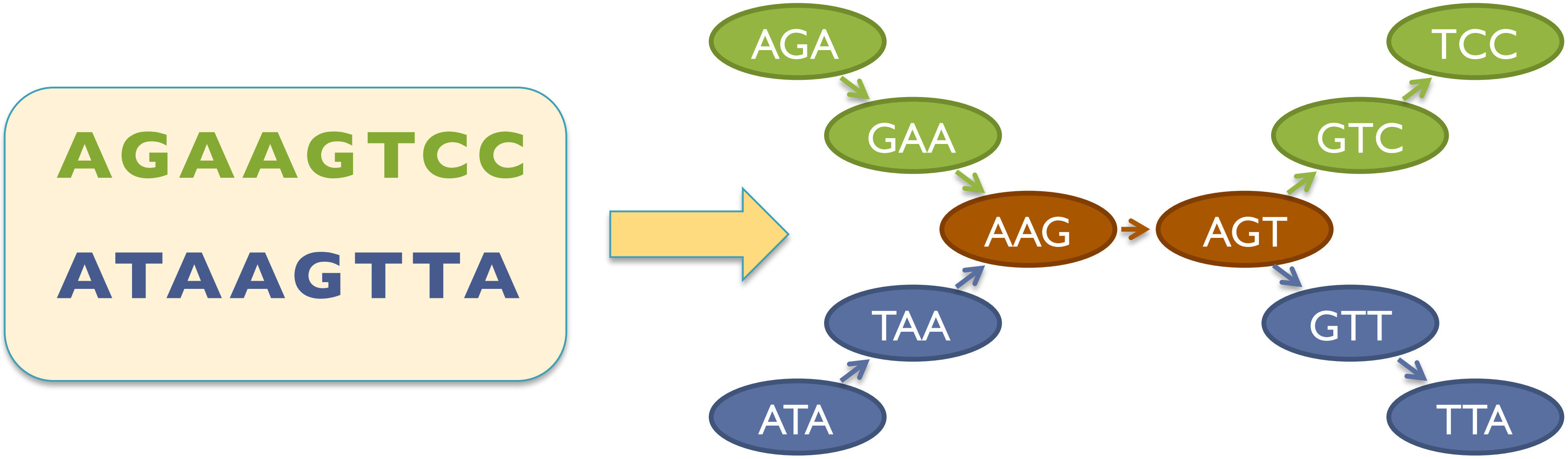
- Encodes all the sequence relationships between the genomes
- How well conserved is a given sequence?
- What are the pan-genome network properties?

SplitMEM: Graphical pan-genome analysis with suffix skips
Marcus, S, Lee, H, Schatz, MC
http://biorxiv.org/content/early/2014/04/06/003954

Graphical pan-genome analysis

Colored de Bruijn graph

- Node for each distinct kmer
- Directed edge connects consecutive kmers
- Nodes overlap by k-1 bp



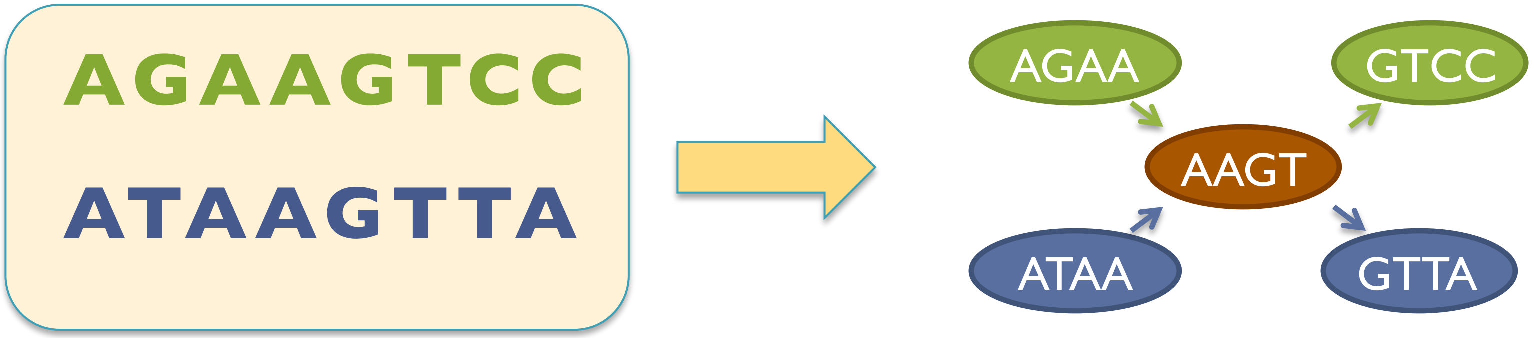
AGAAGTCC

ATAAGTTA

Graphical pan-genome analysis

Colored de Bruijn graph

- Node for each distinct kmer
- Directed edge connects consecutive kmers
- Nodes overlap by k-1 bp



AGAAGTCC

ATAAGTTA

AGAA

GTCC

AAGT

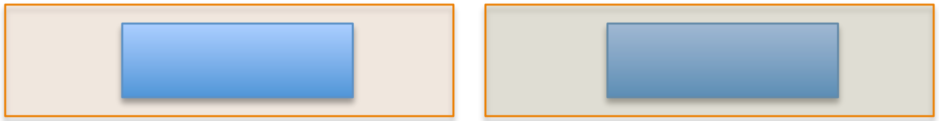
ATAA

GTTA

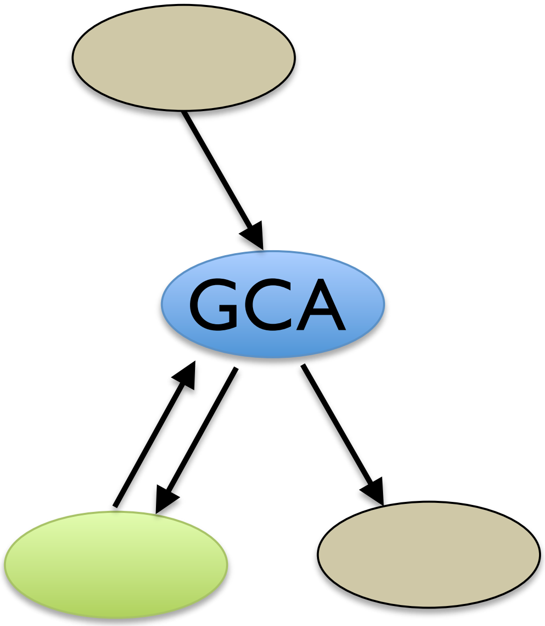
More specifically:

- We aim to build the compressed de Bruijn graph as quickly as possible without considering every distinct kmer

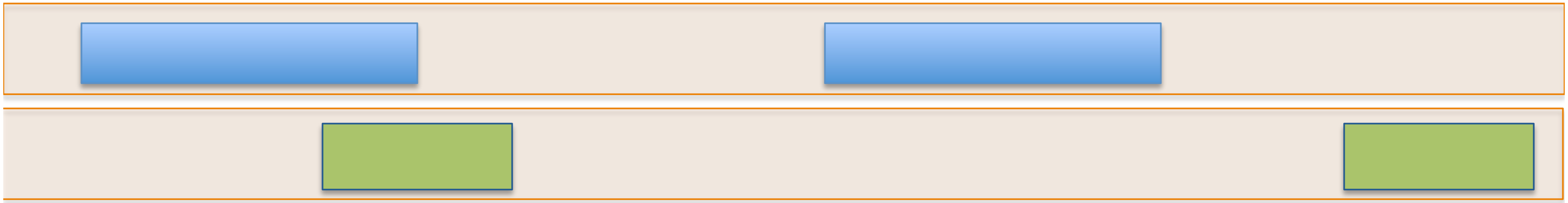
Maximal Exact Matches (MEMs) to de Bruijn Graphs



TGCAC…GGCAA

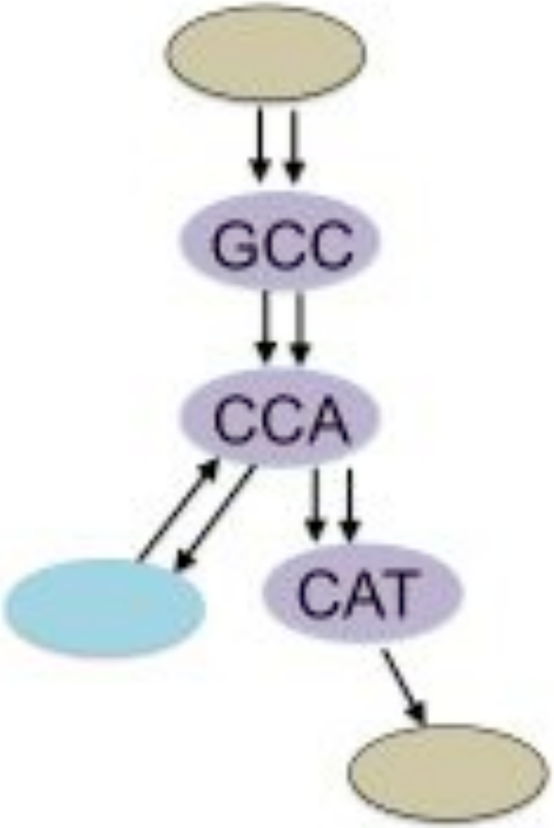


Overlapping MEMs

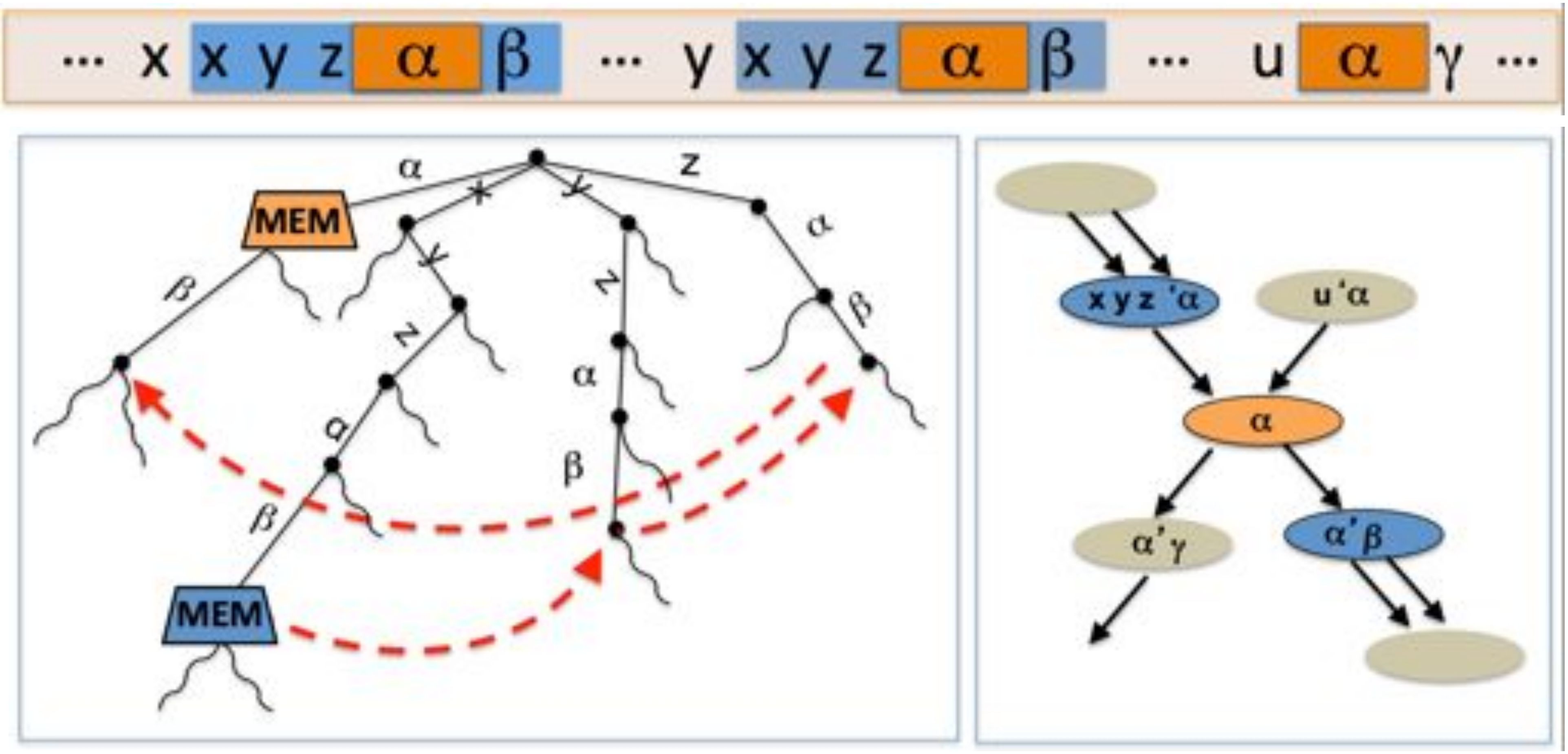


TGCCATCGCCAACCAT

TGCCATCGCCAACCAT



Suffix Trees & de Bruijn Graphs

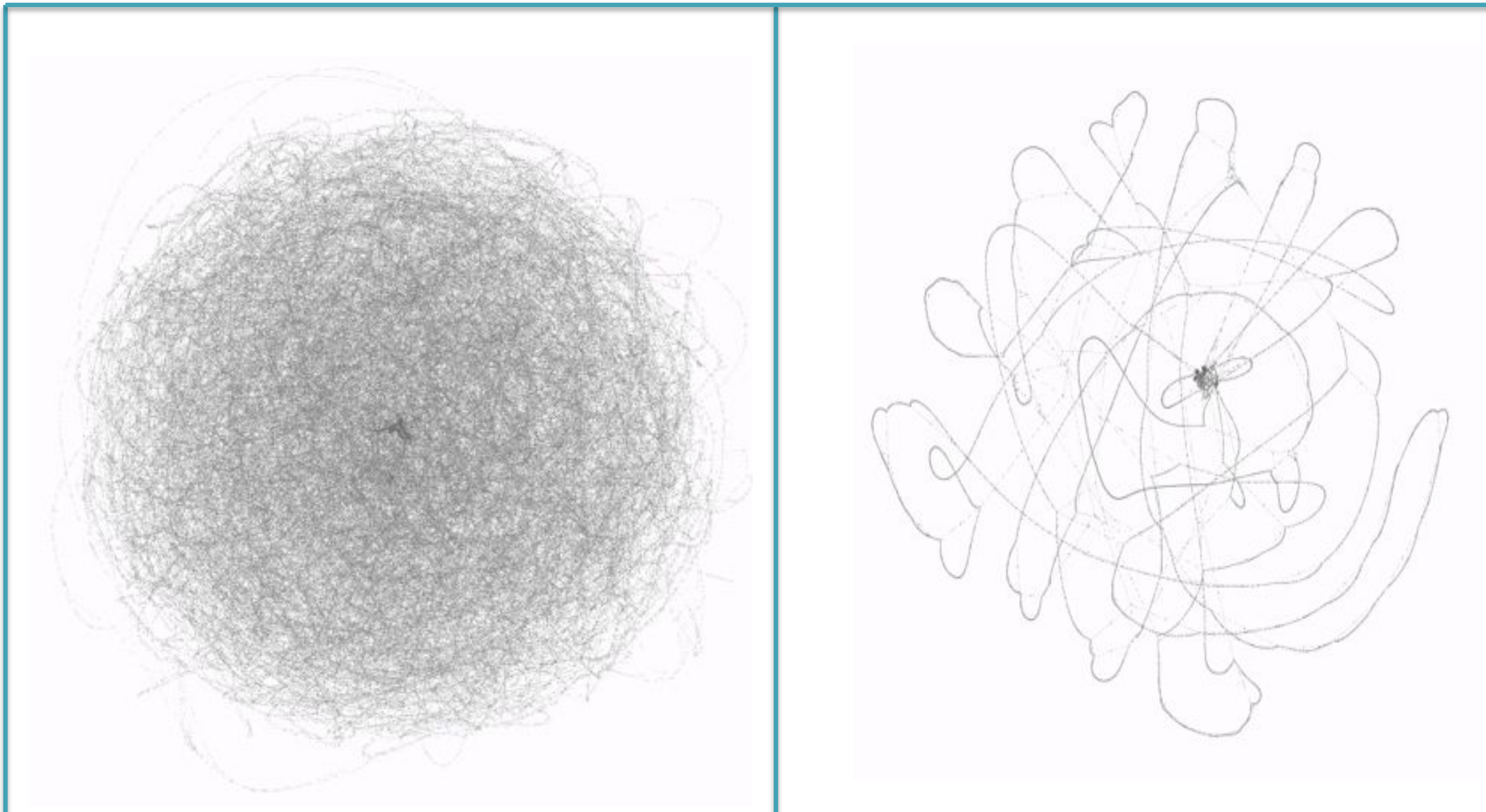


Key concepts:

- Shared sequences form repeats called "maximal exact matches" (MEM)
- Easy to identify MEMs in a suffix tree, but may be nested within other MEMs
- Use "suffix skips" to quickly decompose MEMs, add in the missing nodes and edges

B. anthracis pan-genome

(9 strains)



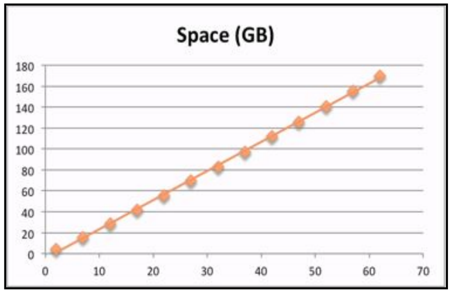
k=25

k=1000

Microbial Pan-Genomes

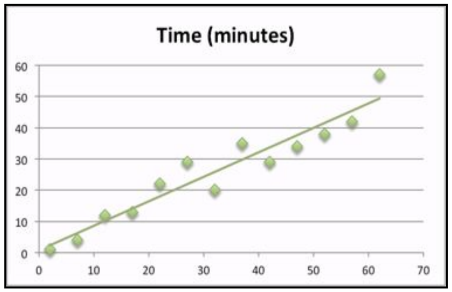
E. coli (62) and B. anthracis (9) pan-genome analysis

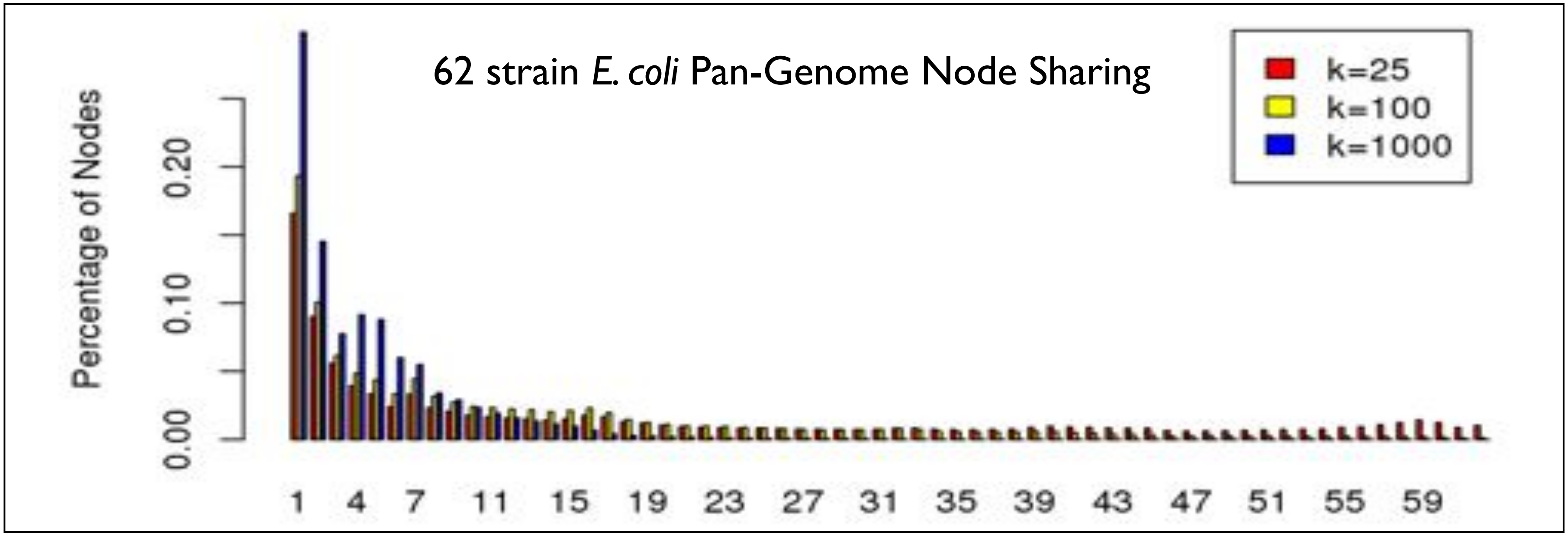
- Analyzed all available strains in Genbank
- Space and time are effectively linear in the number of genomes
    - O(n log g) where g is the length of the longest genome


Space (GB)

Many possible applications:

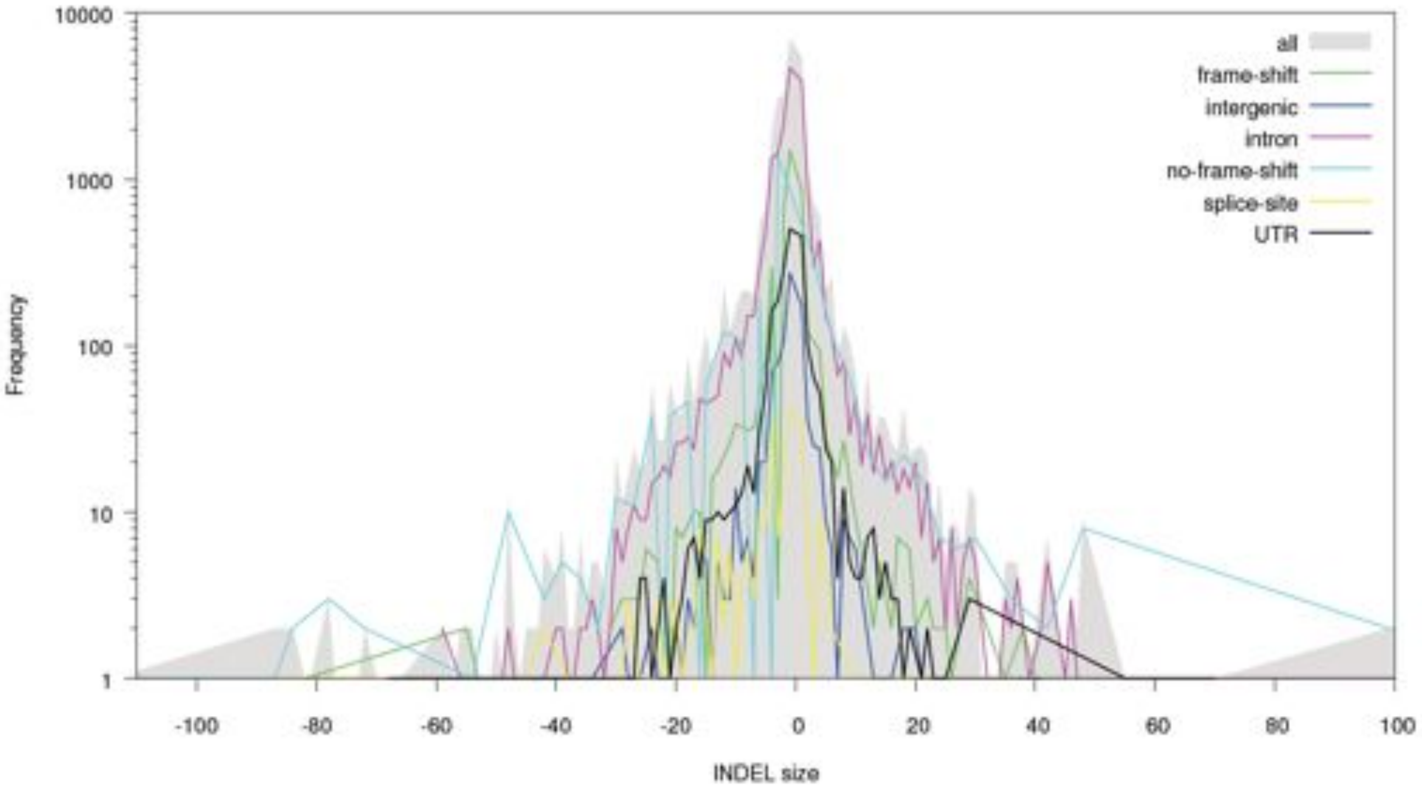
- Identifying "core" genes present in all strains
- Characterizing highly variable regions (+ flanking shared)
- Cataloging sequences shared by pathogenic varieties


Time (minutes)


62 strain *E. coli* Pan-Genome Node Sharing

Part 2: Practice

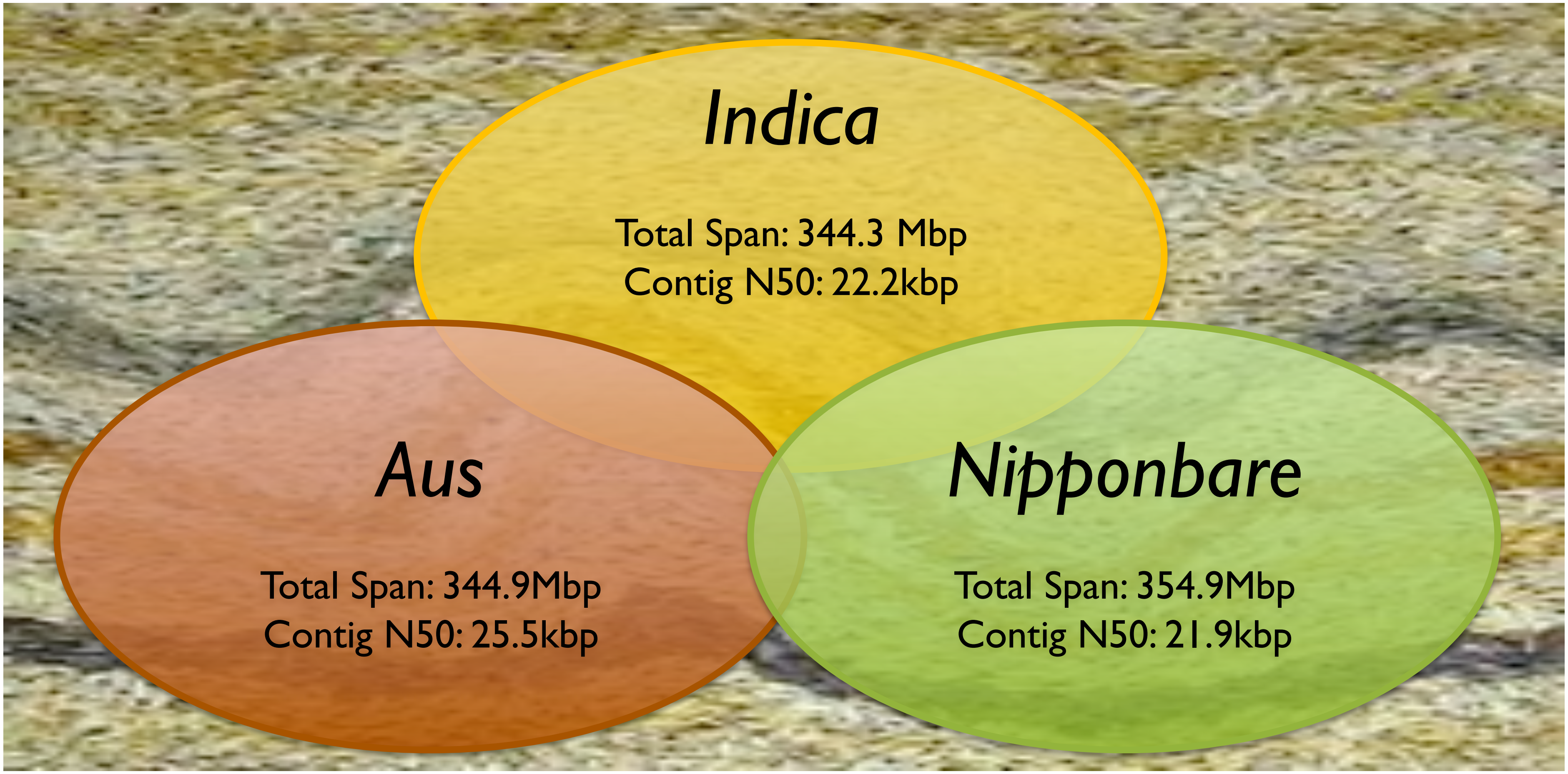
Genetics of Autism Spectrum Disorders



1. Constructed database of >1M transmitted and de novo indels in ~1000 families
2. For practical reasons, analysis is computed relative to the (unpatched) reference genome
   - We use population statistics to "clean" problematic regions
   - We believe we are missing and/or misinterpreting some interesting variants

Accurate de novo and transmitted indel detection in exome-capture data using microassembly.
Narzisi et al. (2014) *Nature Methods.* doi:10.1038/nmeth.3069

Population structure of *Oryza sativa*



Indica

Total Span: 344.3 Mbp
Contig N50: 22.2kbp

Aus

Total Span: 344.9Mbp
Contig N50: 25.5kbp

Nipponbare

Total Span: 354.9Mbp
Contig N50: 21.9kbp

Whole genome de novo assemblies of three divergent strains of rice (*O. sativa*) documents novel gene space of *aus* and *indica*
Schatz, Maron, Stein et a*l* (2014) *http://biorxiv.org/content/early/2014/04/02/003764*

Pan-genomics of draft assemblies

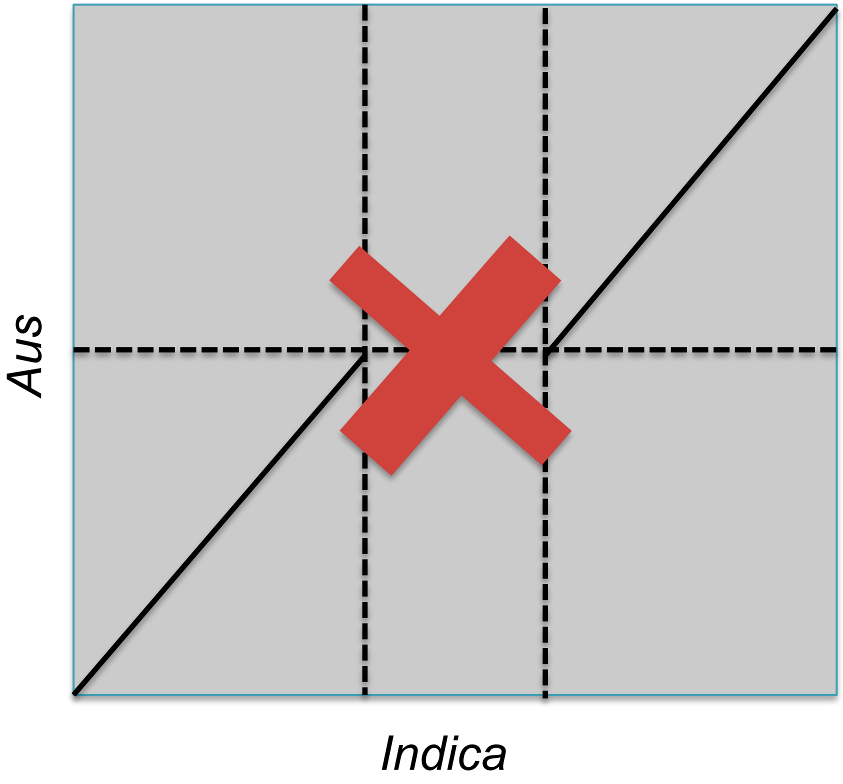
Strategy:

1. Align the genomes to each other (MUMmer)
2. Identify segments of genome A that do not align anywhere to genome B (BEDTools)

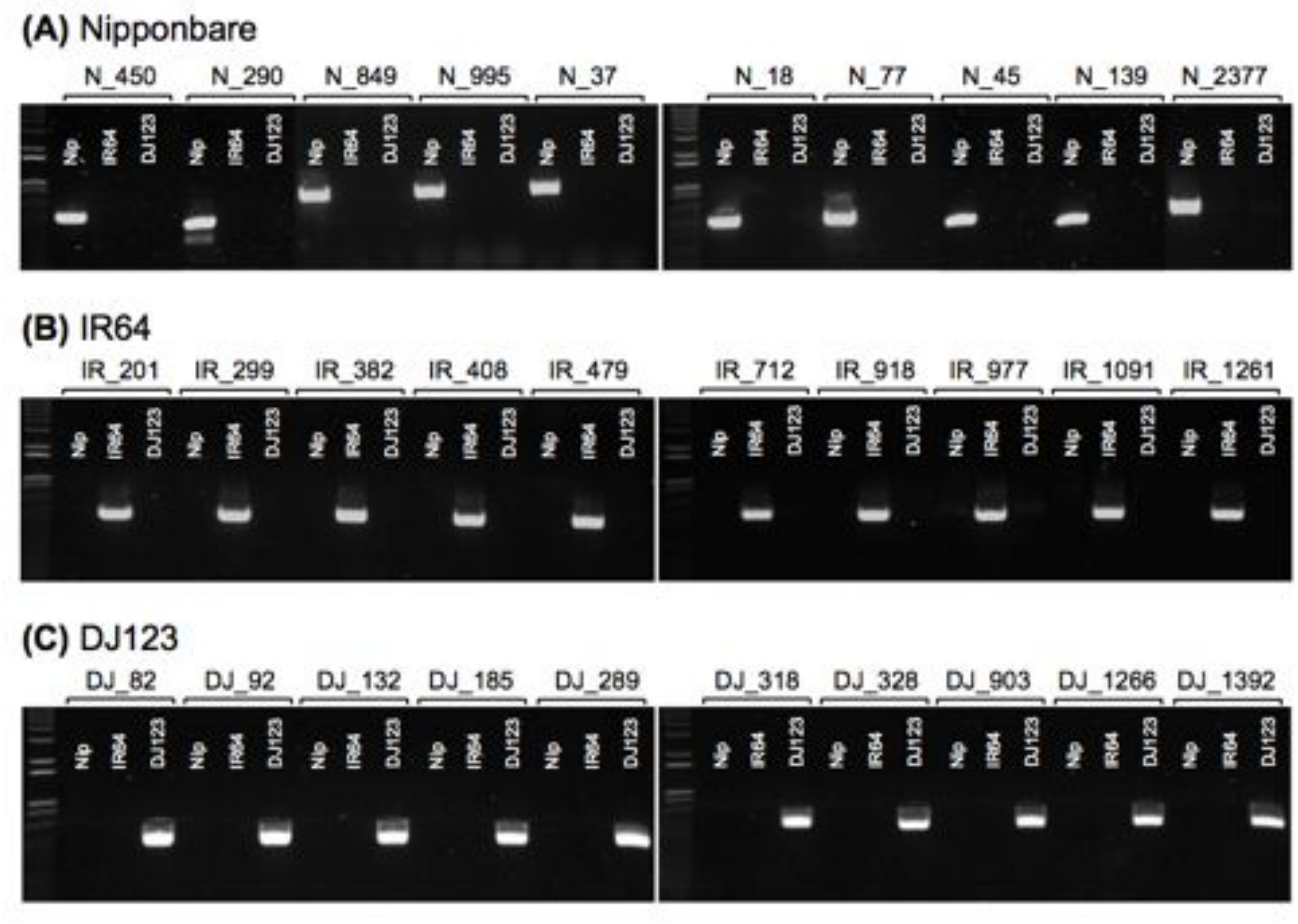
→ Megabases specific to each genome!!!!

3. Screen regions that fail to align with their k-mer frequencies (jellyfish)
   - "Genome specific regions" averaged over 10,000x kmer coverage while unique regions were ~50x

→ 100s of KB specific to each genome!!!

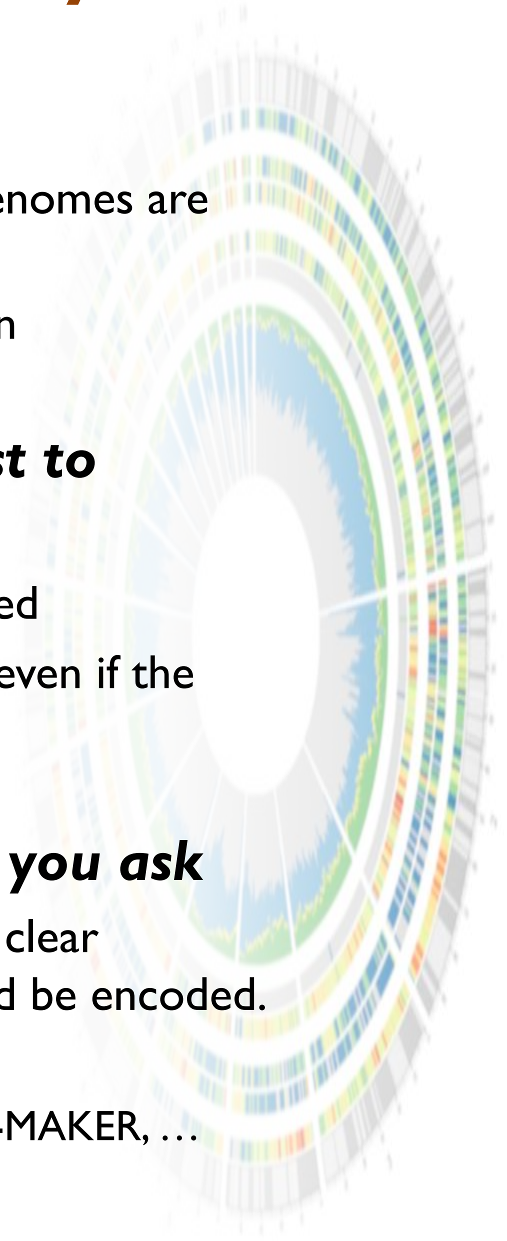


Genome-specific Regions



Successfully able to identify many regions specific to each genome (30/30 PCR validation)
Enriched for genes for disease resistance & other interesting phenotypes

Pan-genomics Summary



- ***Now is the time to study pan-genomes***
  - Perfect assemblies of microbes and many smaller eukaryotic genomes are now routine
  - Expect to rapidly scale up these results to larger genomes soon

- ***Algorithms must scale to large collections, be robust to errors, gaps, and ambiguity***
  - Large body of assembly and alignment theory can be repurposed
  - Simple refinements, like k-mer screening, can be very effective even if the sequence is lacking

- ***The "right approach" will depend on the questions you ask***
  - We all agree we need to work from a graph, but there is not a clear consensus of what the graph should represent or how it should be encoded.
  - Ultimately the needs will be driven by applications
    - graph-BLAST, -BWA, -SAMTools, -TopHat/Cufflinks, -IGV, -UCSC, -MAKER, …

Acknowledgements

Schatz Lab

Rahul Amin
Tyler Gavin
James Gurtowski
Han Fang
Hayan Lee
Maria Nattestad
Aspyn Palatnick
Srividya Ramakrishnan
Eric Biggers
Ke Jiang
Shoshana Marcus
Giuseppe Narzisi
Rachel Sherman
Greg Vurture
Alejandro Wences

CSHL

Hannon Lab
Gingeras Lab
Jackson Lab
Hicks Lab
Iossifov Lab
Levy Lab
Lippman Lab
Lyon Lab
Martienssen Lab
McCombie Lab
Tuveson Lab
Ware Lab
Wigler Lab

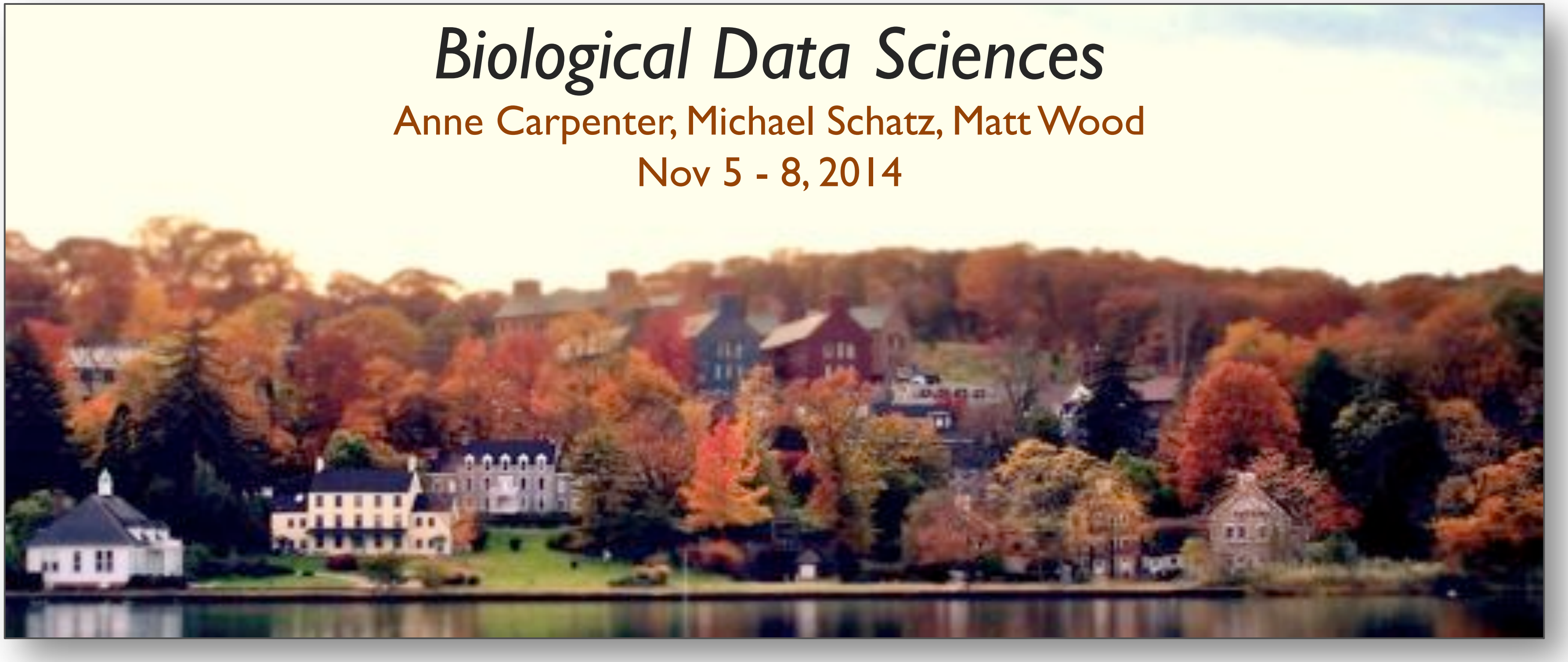
Pacific Biosciences
Oxford Nanopore

NSF

National Human Genome Research Institute

U.S. DEPARTMENT OF ENERGY

SFARI
SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE



Biological Data Sciences

Anne Carpenter, Michael Schatz, Matt Wood

Nov 5 - 8, 2014

Thank you

http://schatzlab.cshl.edu

@mike_schatz