Introduction to Next Generation Sequencing and Analysis for Microbes

Egon A. Ozer, MD PhD

Director, Center for Pathogen Genomics and Microbial Evolution

Northwestern University Feinberg School of Medicine

Department of Medicine, Division of Infectious Diseases

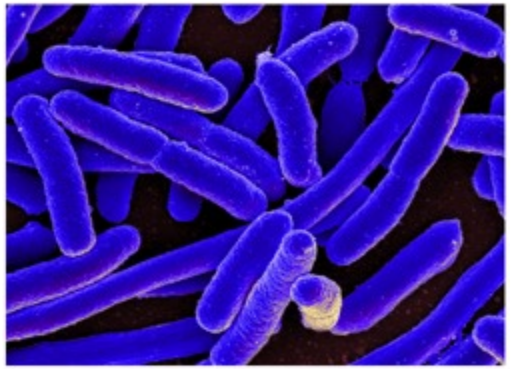
e-ozer@northwestern.edu

Outline

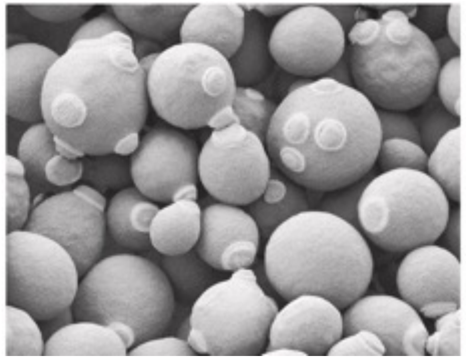
- Microbial genome sequencing
- Whole-genome assembly and alignment
- Genome annotation
- Reference-based read alignment
- Phylogenetic analysis

Microbes

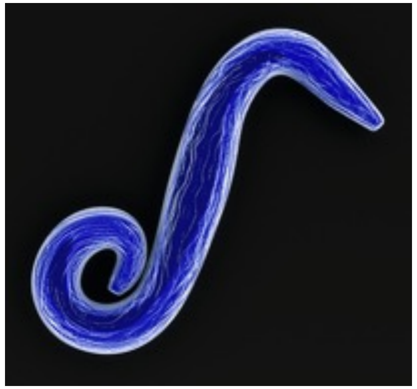
- Single-celled microscopic organisms
  - Prokaryotes:
    - Bacteria
    - Archaea
  - Eukaryotes:
    - Fungi (e.g. *Candida, Cryptococcus*, etc.)
    - Parasites (e.g. *Plasmodium falciparum, Toxoplasma gondii,* etc.)
    - Plants (e.g. Algae)
- Viruses



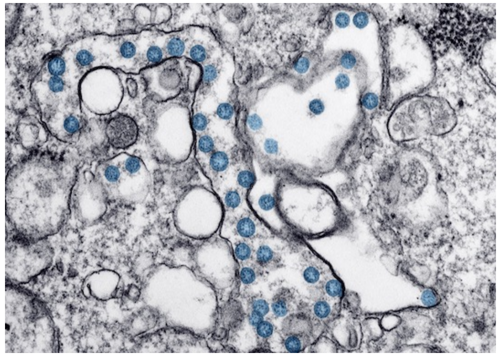
E. coli



Candida albicans



Plasmodium falciparum



SARS-CoV-2

Microbial Genomics vs Non-microbial

	Species and Common Name	Estimated Total Size of Genome (bp)	Estimated Number of Protein-Encoding Genes
Microbial	Staphylococcus aureus	2.8 million	2,700
	Escherichia coli	5.1 million	4,800
	Saccharomyces cerevisiae (baker’s yeast)	12 million	6,000
	Plasmodium falciparum (malaria)	23 million	5,000
	Trichomonas vaginalis	160 million	60,000
Non-Microbial	Caenorhabditis elegans (nematode)	95.5 million	18,000
	Drosophila melanogaster (fruit fly)	170 million	14,000
	Oryza sativa (rice)	470 million	51,000
	Canis familiaris (domestic dog)	2.4 billion	19,000
	Mus musculus (laboratory mouse)	2.5 billion	30,000
	Homo sapiens (human)	2.9 billion	20,000-25,000

Adapted from: Pray, L. (2008) *Nature Education* 1(1):96

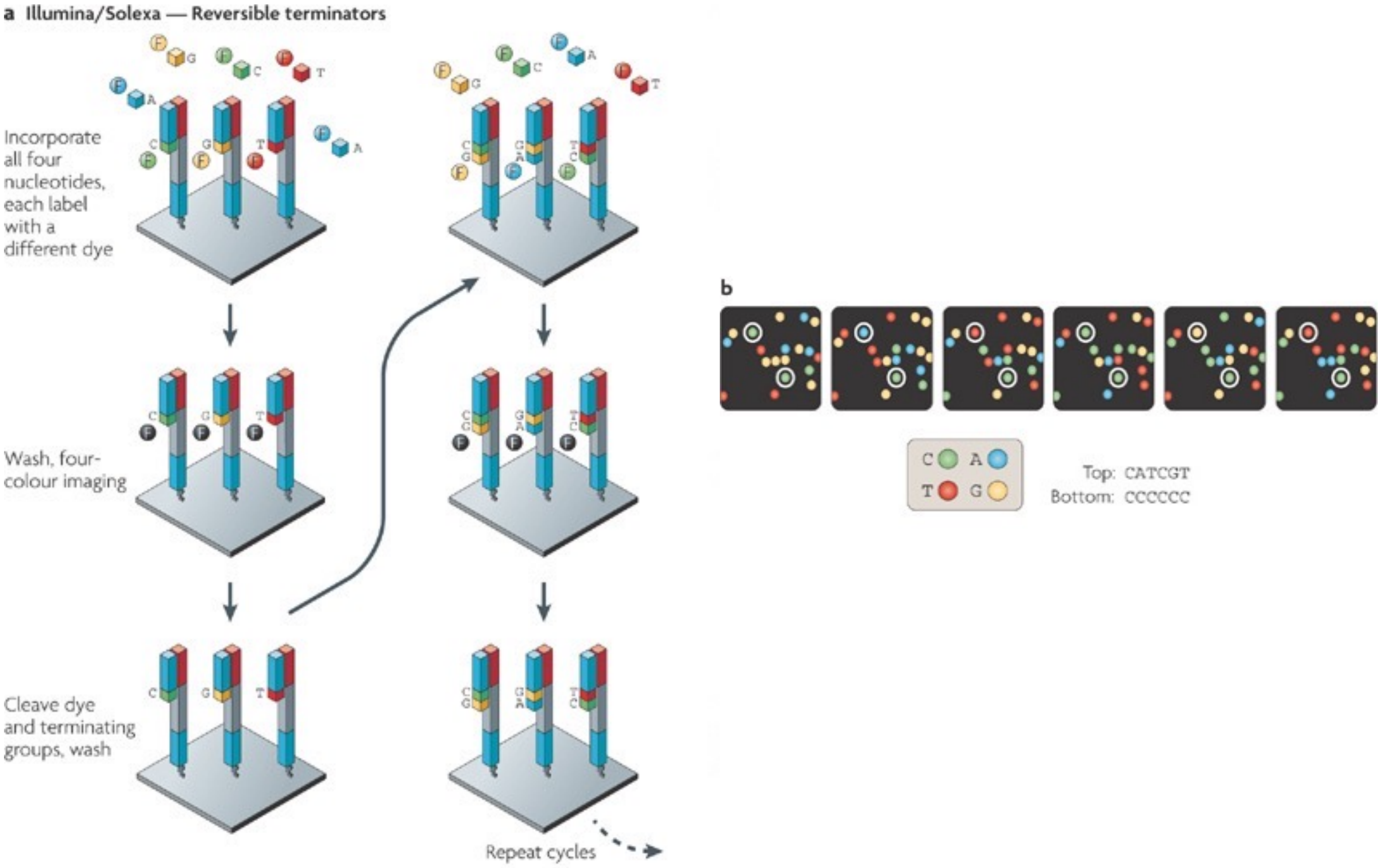
Microbial Genomics vs Non-microbial

	Non-microbial	Microbial
Sequence Size	Large (100’s or 1000’s of Mb)	Small (< 20 - 25 Mb)
Ploidy	Polyploid (mostly)	Haploid (mostly)
Chromosomes	Multiple	One (most bacteria) or more

Smaller doesn’t mean easier!

Whole-genome sequencing platforms

- Illumina
  - HiSeq
  - MiSeq
  - NextSeq
  - NovaSeq
  - MiniSeq


a Illumina/Solexa — Reversible terminators
Incorporate all four nucleotides, each label with a different dye
Wash, four-colour imaging
Cleave dye and terminating groups, wash
Repeat cycles
b
C A
T G
Top: CATCGT
Bottom: CCCCCC

Metzger ML. *Nat Rev Gen.* 2010. 11:31-46

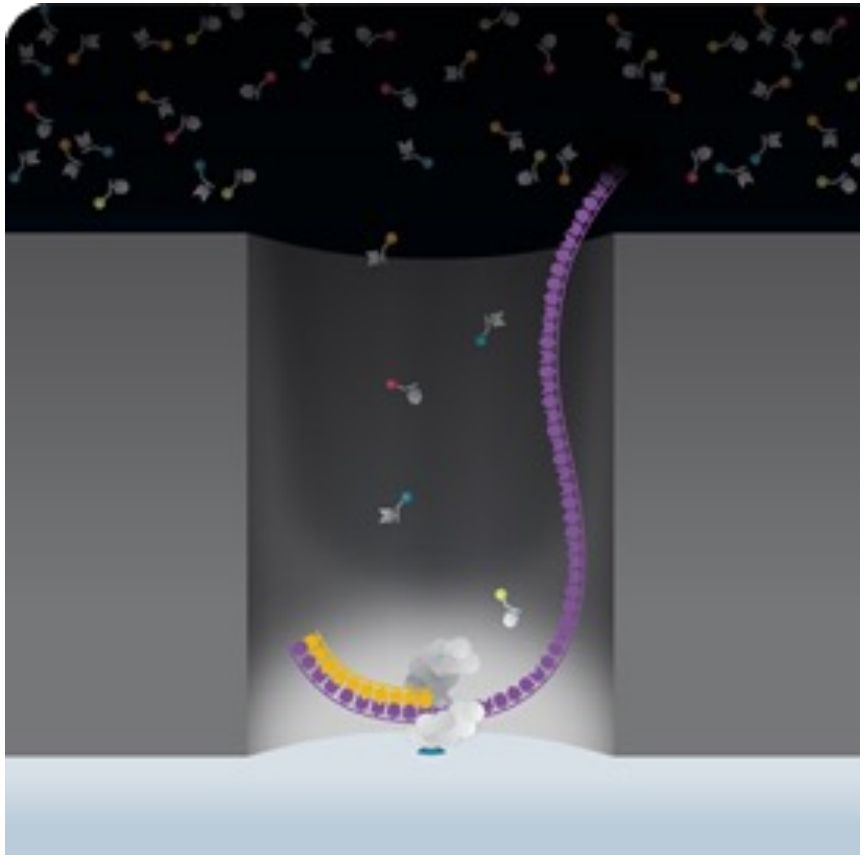
Whole-genome sequencing platforms

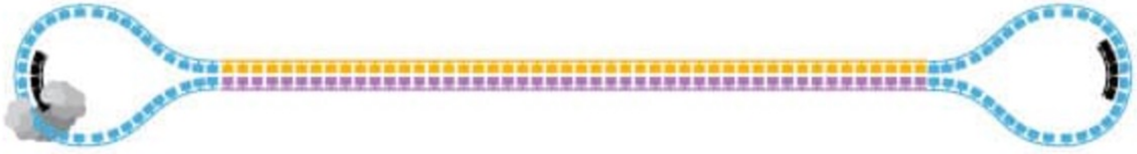
- Illumina (HiSeq, MiSeq, NextSeq, NovaSeq)
  - Benefits:
    - High-throughput
      - MiSeq: max 15 Gb per run
      - NextSeq: max 120 Gb per run
      - HiSeq: max 1,500 Gb per run
      - NovaSeq: max 6,000 Gb per run
    - Low error rate (~ 0.1%) – substitution errors more common than indel
    - Relatively low cost-per-base
  - Drawbacks:
    - PCR amplification required for sequencing
    - Short reads (max 150 - 300 bp)
    - Relatively slow (1 – 3 days)

https://www.illumina.com/systems/sequencing-platforms.html

Whole-genome sequencing platforms

- PacBio (Sequel, Sequel II)
  - SMRT = “Single Molecule, Real-Time”
  - Flow-cells contain millions of zero-mode waveguides (ZMWs)
  - Anchored polymerases at bases incorporate labeled bases → light emitted
  - Nucleotide incorporates read in real-time to generate sequence





SMRTbell library → ”HiFi” reads

http://www.pacificbiosciences.com

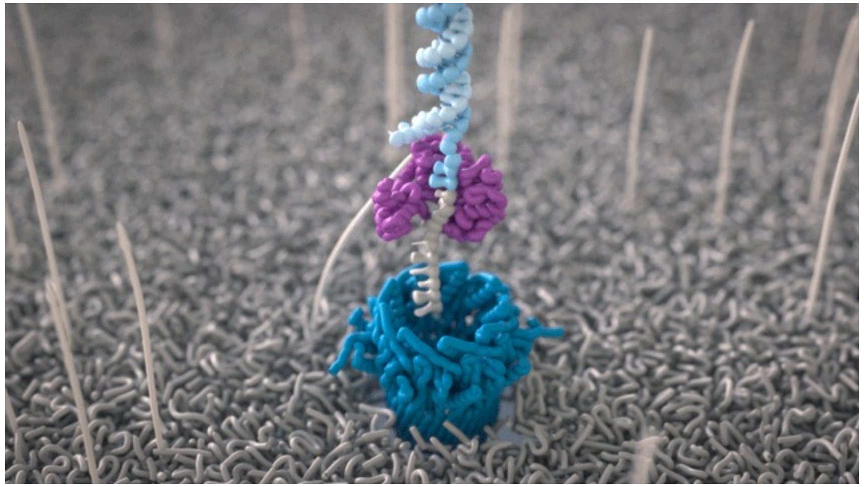
Whole-genome sequencing platforms

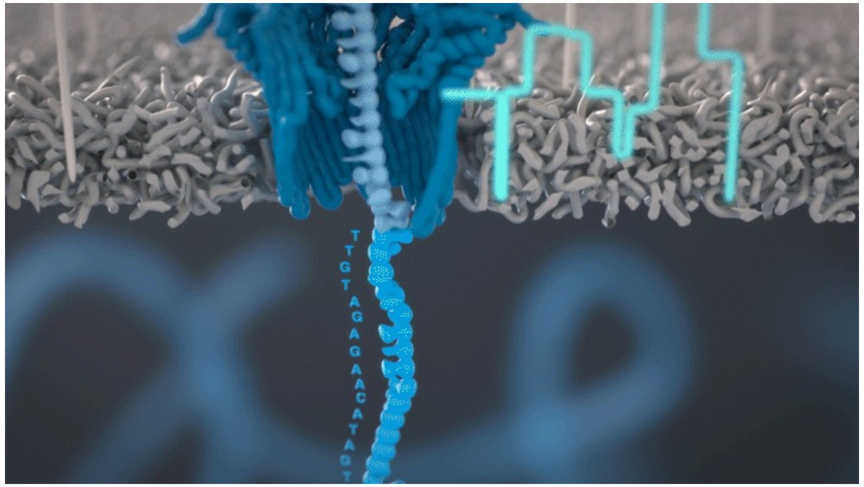
- PacBio (Sequel, Sequel II)
  - Benefits:
    - Long reads (max read length ~175 kb)
    - Intermediate - high throughput (20 Gb - 160 Gb)
    - Fast: run time 4 - 30 hours
    - No PCR amplification necessary
  - Drawbacks:
    - Higher error rates than Illumina - substitution and indel
      - Error rates can be much lower with circular consensus libraries (CCL), but homopolymers can still be a problem
    - Higher cost-per-base than Illumina platforms

https://www.pacb.com/products-and-services/sequel-system/
http://allseq.com/knowledge-bank/sequencing-platforms/pacific-biosciences/

Whole-genome sequencing platforms

- Oxford Nanopore (MinION, GridION)
  - Engineered protein pore α-hemolysin transports DNA molecules through a polymer membrane
  - Ionic current is passed through the nanopore
  - As nucleotides pass through pore, current is disrupted
  - Degree of current disruption is specific to individual nucleotides (A, C, T, or G)





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

Whole-genome sequencing platforms

- Oxford Nanopore (MinION, GridION)
  - Benefits:
    - Long reads (up to 900 kb)
    - Intermediate throughput (15 - 30 Gb per flow cell)
    - Fast: real-time results, run length depends on desired read depth
    - Affordable equipment costs (~ $1000 for instrument, $900 per flow cell)
    - No PCR amplification necessary
  - Drawbacks:
    - High error rates (5 – 15%) - substitution and indel
      - Newer generation flow cells → 1% error
    - Higher cost-per-base than (most) Illumina platforms

https://nanoporetech.com/products/comparison

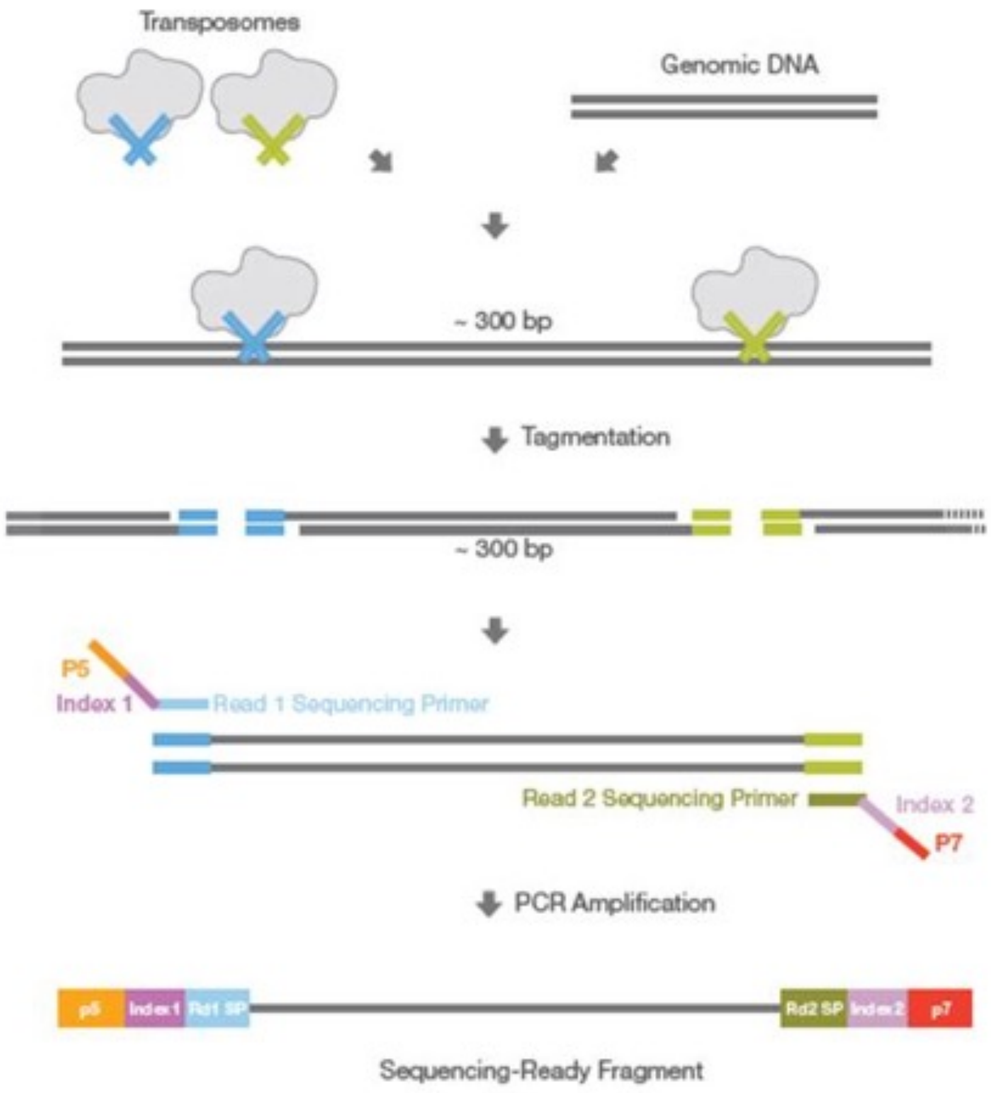
Genome multiplexing

Platform:	*S. aureus* (2.8 Mb)	*E. coli* (4.6 Mb)	*P. aeruginosa* (6.6 Mb)
MiSeq (300 bp)	80 genomes	50 genomes	30 genomes
NextSeq Mid	230 genomes	140 genomes	90 genomes
NextSeq High	710 genomes	430 genomes	300 genomes
HiSeq (one lane)	1,070 genomes	650 genomes	450 genomes
PacBio Sequel	110 genomes	70 genomes	50 genomes
ONT MinION	110 genomes	70 genomes	50 genomes

Goal of 60x coverage, i.e. each base in the genome sequenced ~ 60 times
Caveats: Ideal throughput, ideal library pooling, equal genome sizes, available indexes...

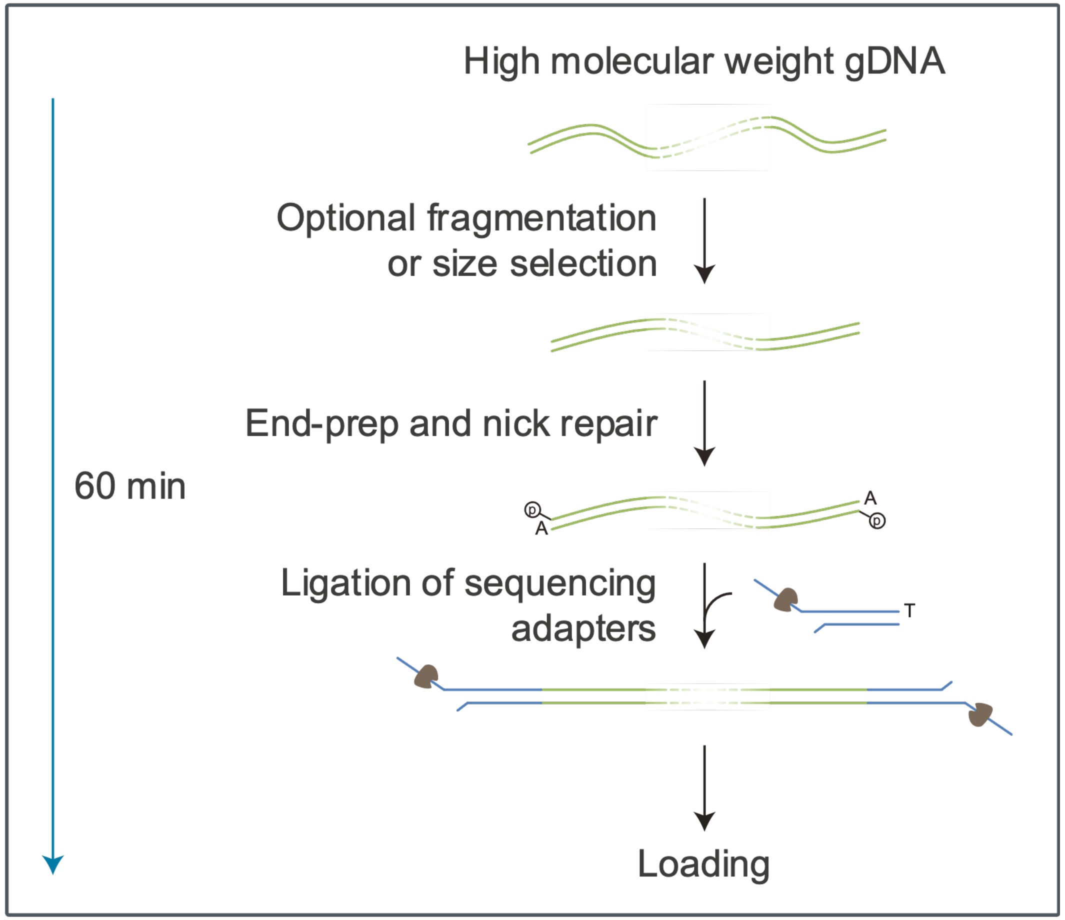
Library Preparation

- Genomic sequence (chromosome + plasmids) fragmented into smaller pieces
    - 500 bp up to 50 kb, depending on application
- Adapter sequences added
    - Adhere sequence to flowcell (Illumina)
    - Generate circularized single-stranded sequence (PacBio)
    - Ligation of sequencing adapters (Nanopore)



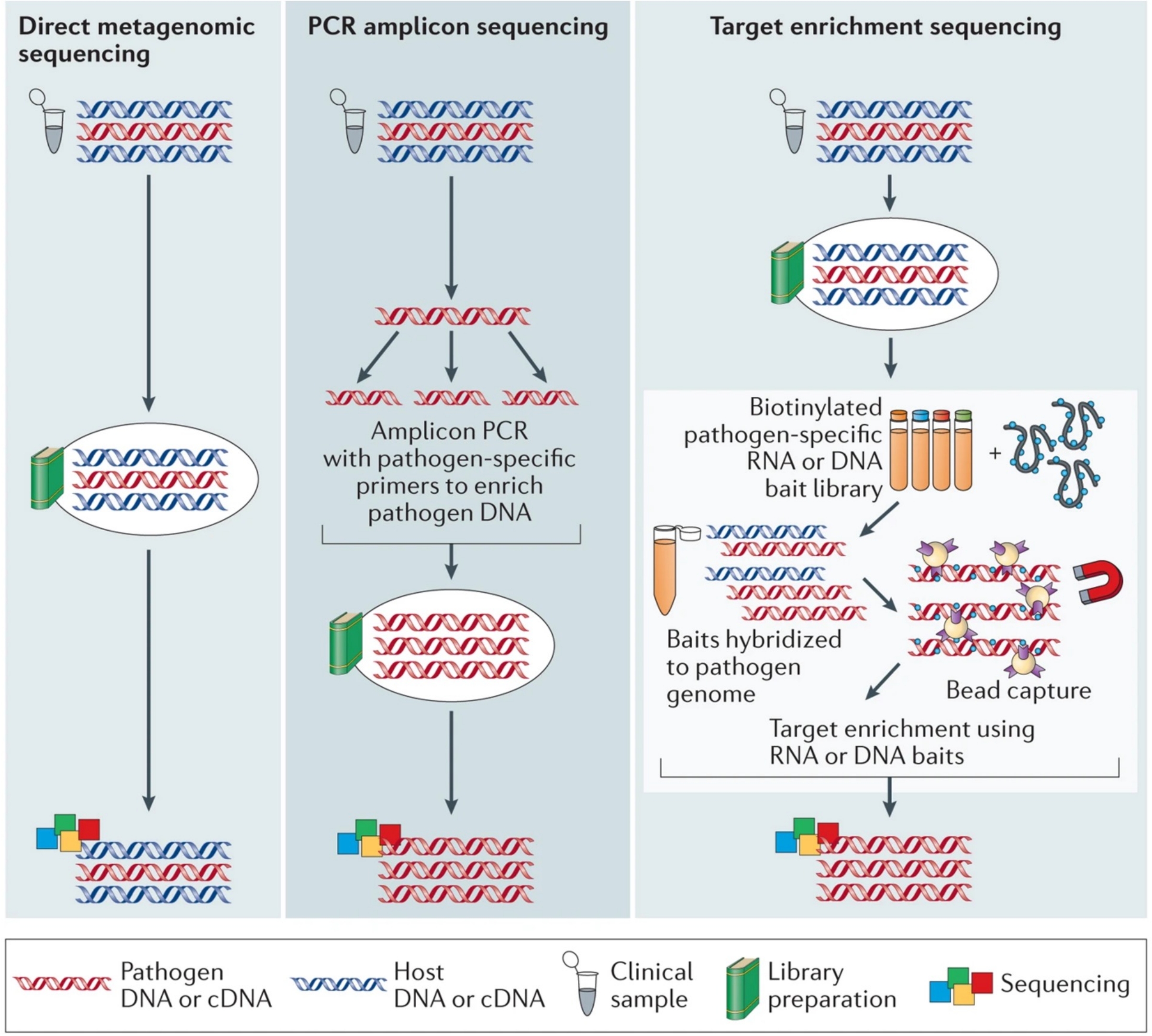
Library Preparation

- Genomic sequence (chromosome + plasmids) fragmented into smaller pieces
  - 500 bp up to 50 kb, depending on application
- Adapter sequences added
  - Adhere sequence to flowcell (Illumina)
  - Generate circularized single-stranded sequence (PacBio)
  - Ligation of sequencing adapters (Nanopore)



High molecular weight gDNA
Optional fragmentation or size selection
End-prep and nick repair
60 min
P
A
A
P
Ligation of sequencing adapters
T
Loading

Sequencing from Non-Cultured Specimens


Direct metagenomic sequencing
PCR amplicon sequencing
Target enrichment sequencing
Amplicon PCR with pathogen-specific primers to enrich pathogen DNA
Biotinylated pathogen-specific RNA or DNA bait library
+
Baits hybridized to pathogen genome
Bead capture
Target enrichment using RNA or DNA baits
Pathogen DNA or cDNA
Host DNA or cDNA
Clinical sample
Library preparation
Sequencing

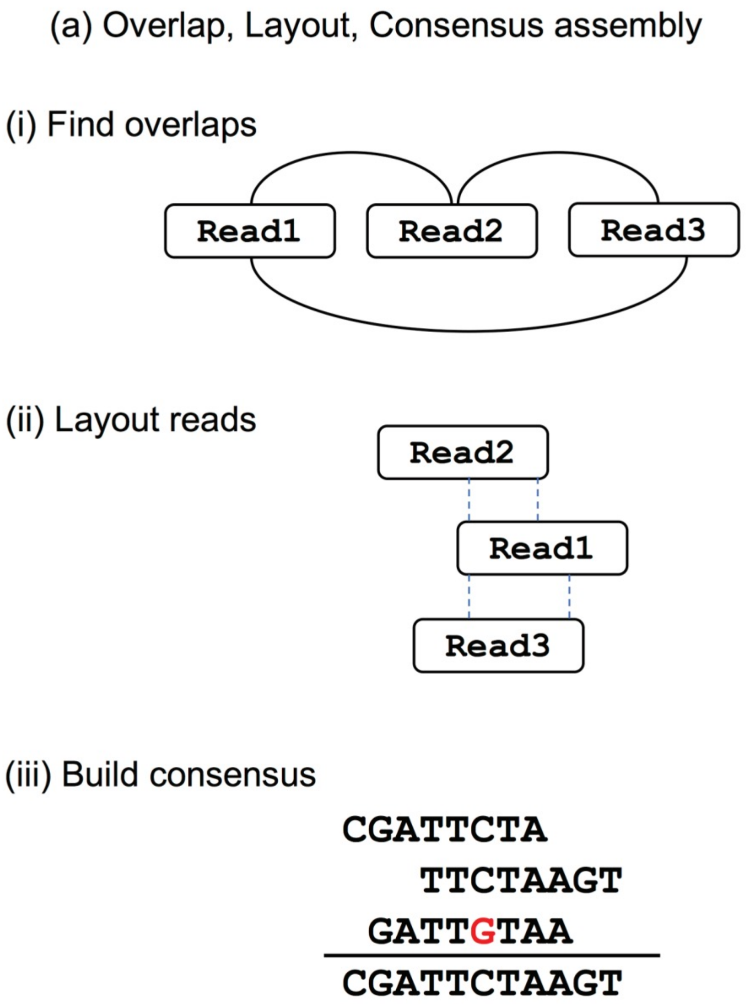
Houldcroft, Beale & Breuer. *Nat Rev Microbiol* **15,** 183–192 (2017).

Assembly vs. Alignment

- Sequencer produces reads. What’s next?
- Assembly
    - Recreate genome sequence by joining sequence reads with each other
    - “Putting together a puzzle”
- Alignment
    - Compare reads to a reference genome sequence
    - Identify single nucleotide variants, small indels

Assembly

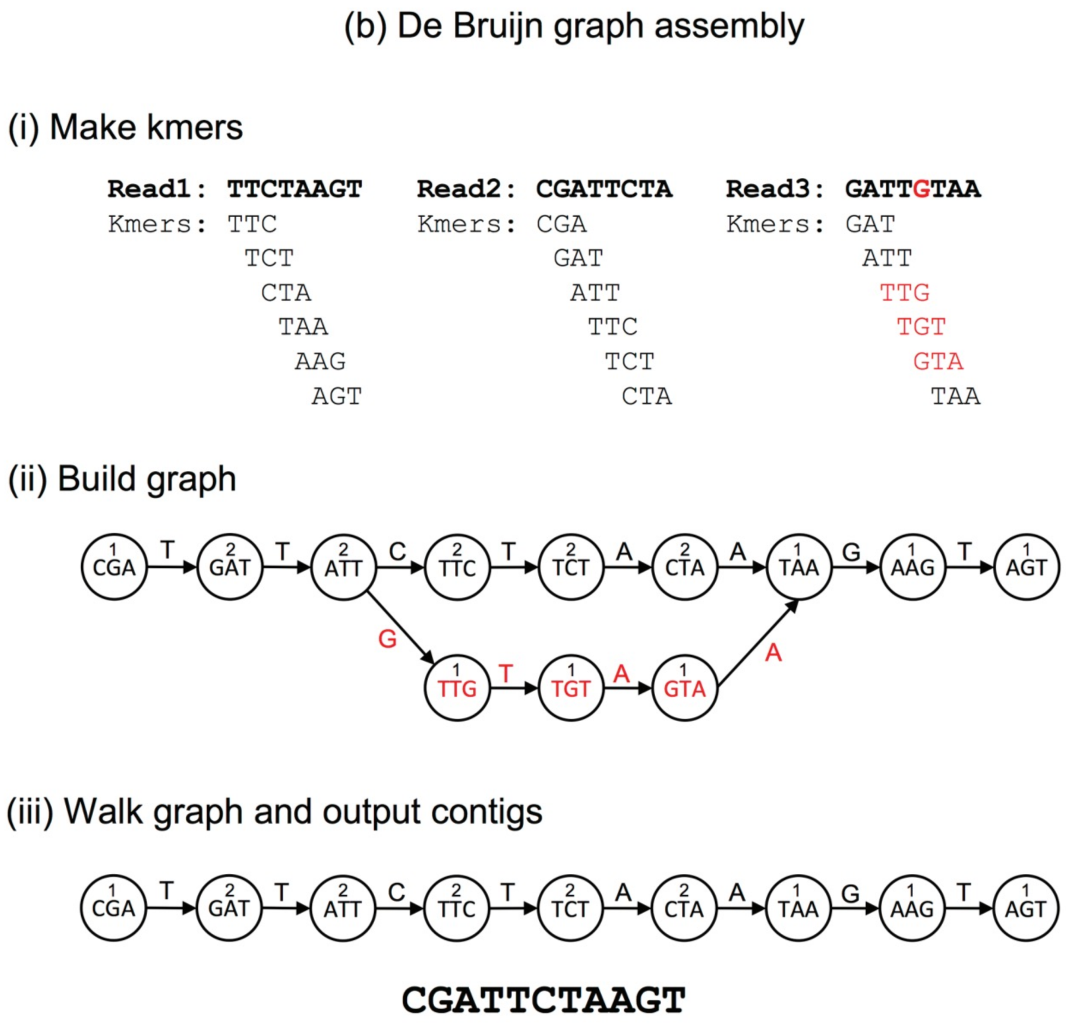
- Overlap layout consensus (OLC)
  - 1) Find overlaps among the reads, 2) create layout of all reads, 3) infer consensus sequence
  - Can be memory & computationally intensive
  - Best for lower numbers of long reads (PacBio or Nanopore)
  - Example software: Celera, miniasm



Ayling M, Clark MD, and Leggett RM. Brief in Bioinf. Feb 2019

Assembly

- De bruijn graph (DBG)
  - Chop reads into shorter k-mers, create graph of consecutive k-mers overlapping by k-1 bases. Recreate sequence by moving through the graph
  - More memory-efficient
  - Short reads or long reads
  - K-mer choice:
    - Short: more connections, less repeat resolution
    - Long: less connections, more repeat resolution
  - Example software: SPAdes, Velvet



(b) De Bruijn graph assembly

(i) Make kmers

```
Read1: TTCTAAGT     Read2: CGATTCTA     Read3: GATTGTAA
Kmers: TTC          Kmers: CGA          Kmers: GAT
        TCT                  GAT                  ATT
         CTA                  ATT                  TTG
          TAA                  TTC                  TGT
           AAG                  TCT                  GTA
            AGT                  CTA                  TAA
```

(ii) Build graph

(iii) Walk graph and output contigs

CGATTCTAAGT

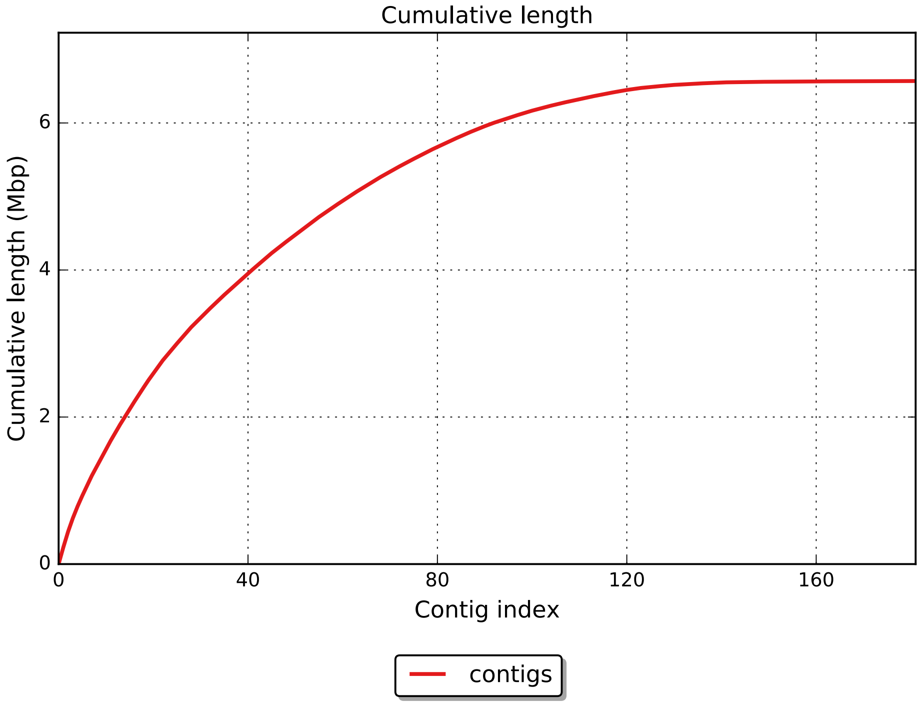
Ayling M, Clark MD, and Leggett RM. Brief in Bioinf. Feb 2019

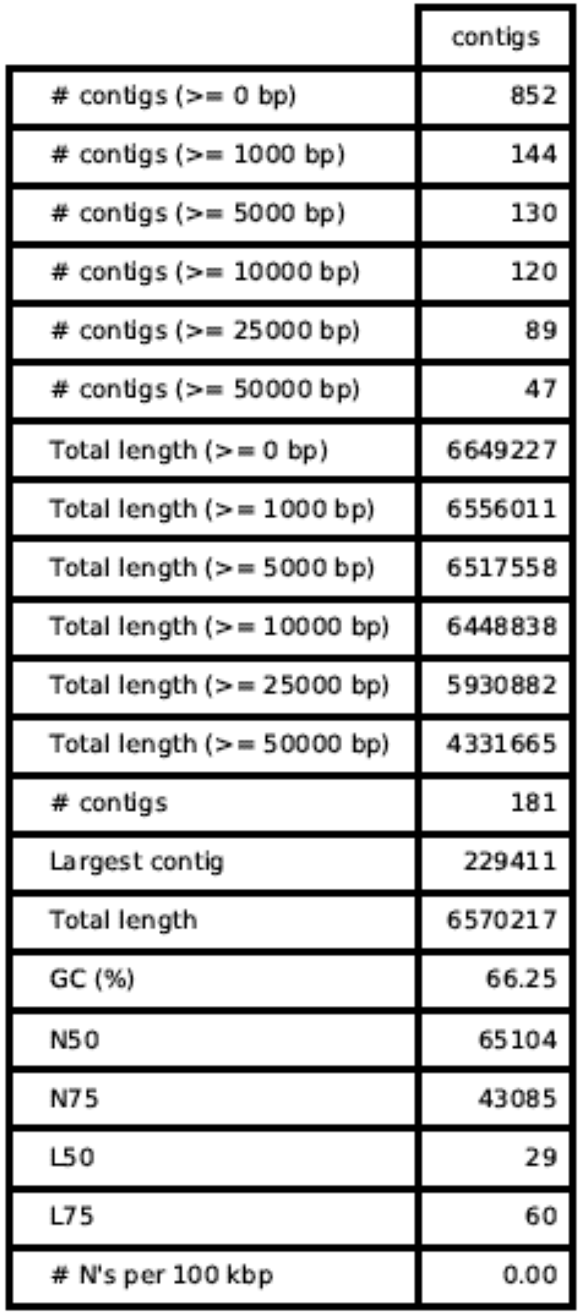
Assembly

- SPAdes Assembler
- De bruijn graph assembler
- Optimized for Illumina reads or hybrid short/long read assemblies
- Algorithm
    1. Read error correction
    2. Iterative repeats with multiple k-mer sizes to optimize assembly
    3. Aligns reads to assembly to correct mismatches & indels
- Output
    - Contigs: Contiguous assembled sequences
    - Scaffolds: Contigs linked and oriented using paired-end Illumina reads

Assembly

- Assessing results
  - Quast
    - Web: http://quast.bioinf.spbau.ru/





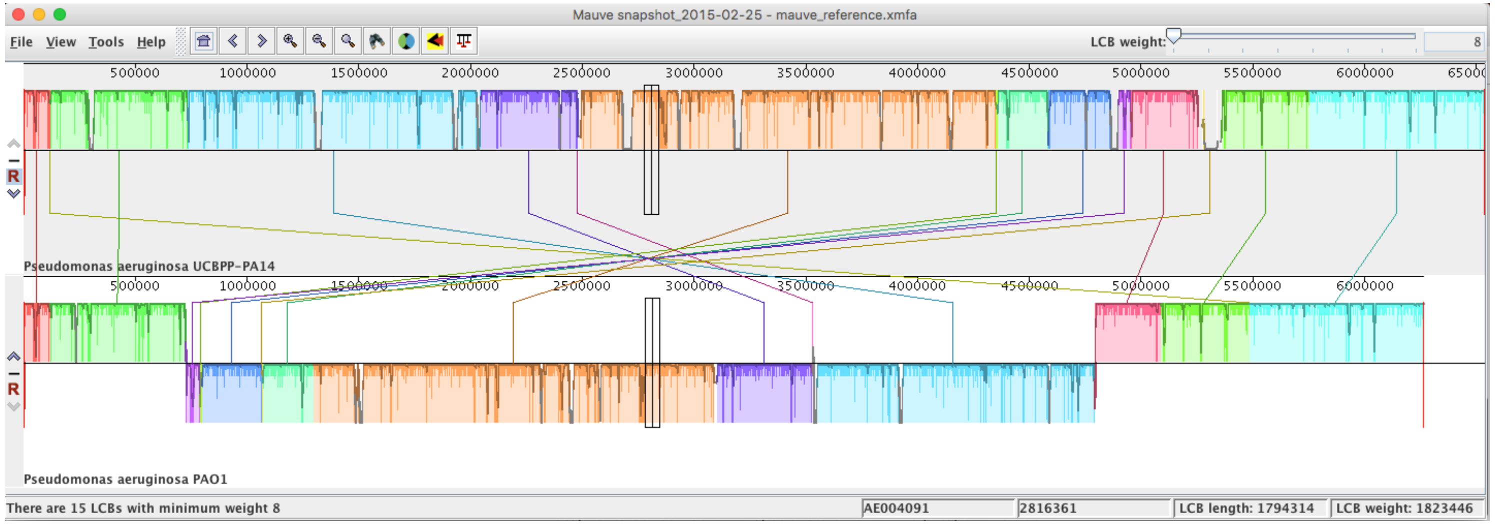
	contigs
# contigs (>= 0 bp)	852
# contigs (>= 1000 bp)	144
# contigs (>= 5000 bp)	130
# contigs (>= 10000 bp)	120
# contigs (>= 25000 bp)	89
# contigs (>= 50000 bp)	47
Total length (>= 0 bp)	6649227
Total length (>= 1000 bp)	6556011
Total length (>= 5000 bp)	6517558
Total length (>= 10000 bp)	6448838
Total length (>= 25000 bp)	5930882
Total length (>= 50000 bp)	4331665
# contigs	181
Largest contig	229411
Total length	6570217
GC (%)	66.25
N50	65104
N75	43085
L50	29
L75	60
# N's per 100 kbp	0.00

Aligning and Ordering Assemblies

- Mauve (http://darlinglab.org/mauve/mauve.html)
- Multiple genome aligner (up to about 10, max)
  - Newer program SibeliaZ (https://github.com/medvedevgroup/SibeliaZ) has higher capacity
- Contig reorder relative to reference
  - Alternative program: Nucmer (http://mummer.sourceforge.net/)
- Visualizer
- All functions available in GUI as well as command line

Mauve alignment

Complete genomes

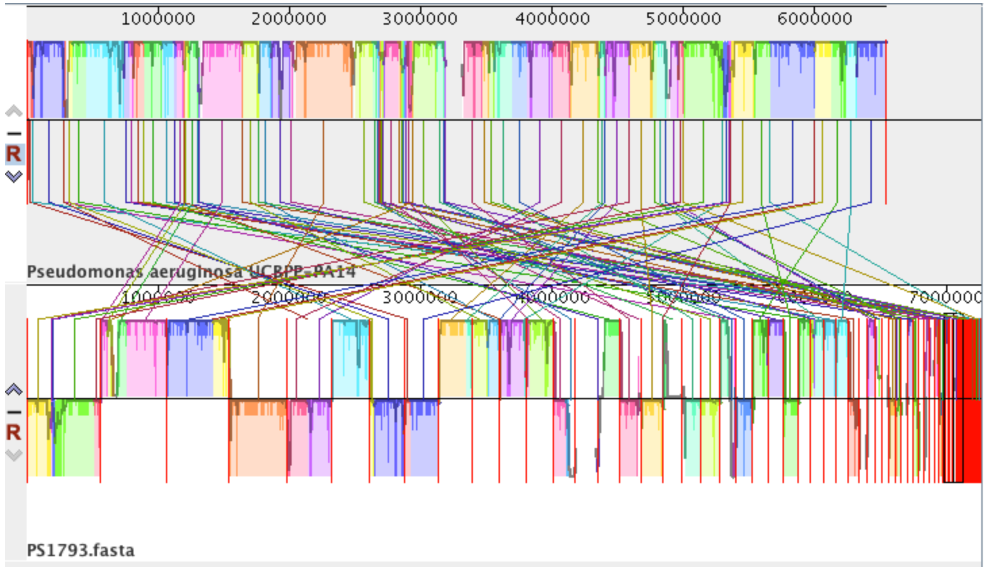


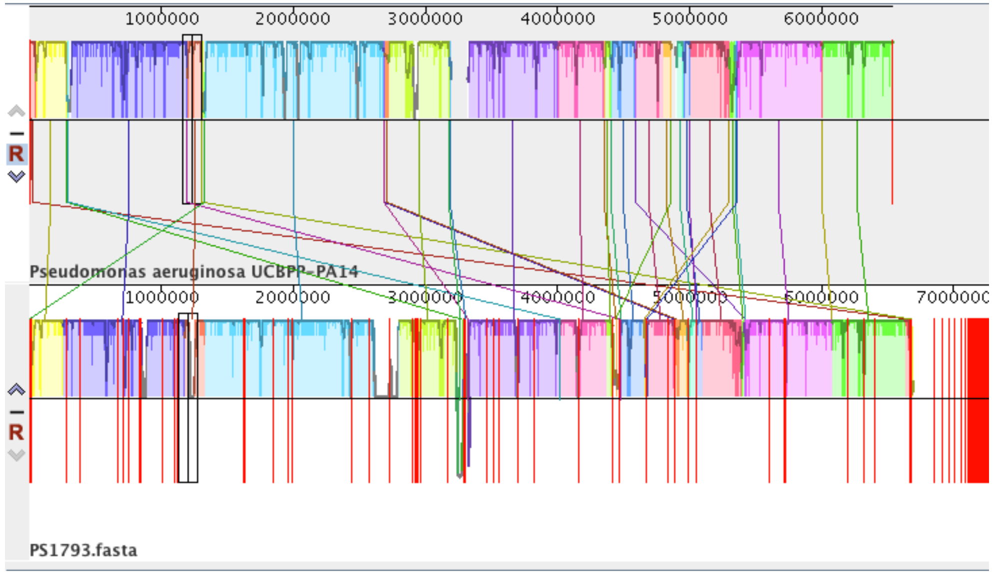
```
There are 15 LCBs with minimum weight 8    AE004091    2816361    LCB length: 1794314    LCB weight: 1823446
```

Contig reorderer

Ref

Draft





Hybrid Assembly

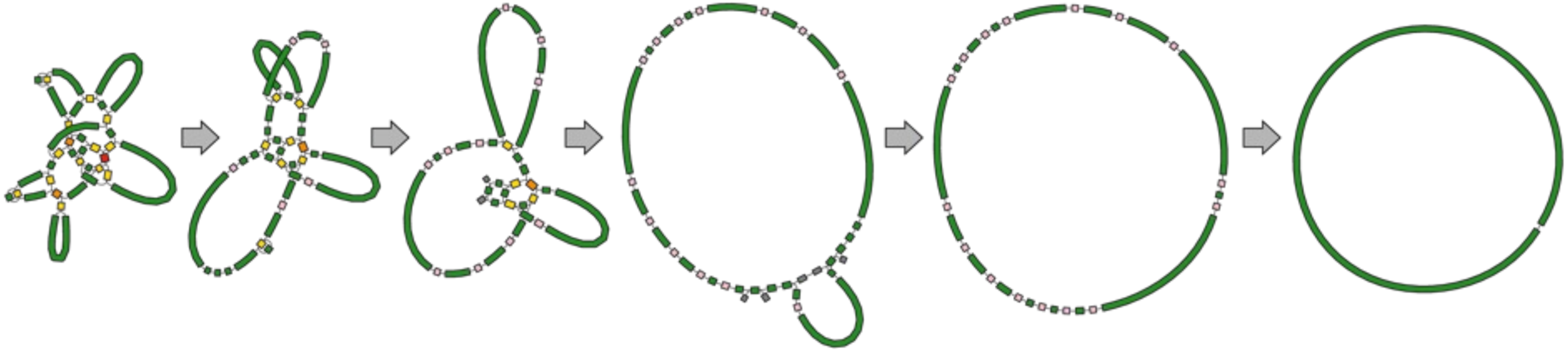
- Short read assemblies:
  - High accuracy, low error rate
  - Short reads can’t resolve long repeats → fragmentary assemblies, multiple contigs
  - Plasmid resolution difficult
- Long read assemblies:
  - Span long repeats → long assemblies, few contigs / complete chromosomes or plasmids
  - Higher error rates and homopolymer errors → lower accuracy
- Hybrid assembly methods combine strengths of both approaches

Hybrid Assembly Approaches

- Hybrid assembly
  - De-bruijn graph assembly with long and short reads or use long reads to order contigs generated with short reads and fill gaps
  - Example: SPAdes
- Sequential assembly
  - 1) Assemble using long reads alone
    - Example Software: Canu (https://github.com/marbl/canu/releases), SMRT-Analysis (https://www.pacb.com/support/software-downloads/), SMARTdenovo (https://github.com/ruanjue/smartdenovo), minimap/miniasm/racon (https://yiweiniu.github.io/blog/2018/03/Genome-assembly-pipeline-miniasm-Racon/)
  - 2) Correct assembly errors using short reads
    - Software: Pilon (https://github.com/broadinstitute/pilon/wiki)
  - 3) Join and circularize chromosomes and plasmids
    - Software: Circlator (https://sanger-pathogens.github.io/circlator/)

Hybrid Assembly Approaches

- All-In-One package
  - Unicycler (https://github.com/rrwick/Unicycler)
  - Uses SPAdes, miniasm, and Racon
  - First, generates assembly from short reads (SPAdes)
  - Second, assembly with long reads and contigs from first step (miniasm/Racon)
  - Third, tries to bridge ambiguous connections using long reads
  - Last, attempts to circularize contigs

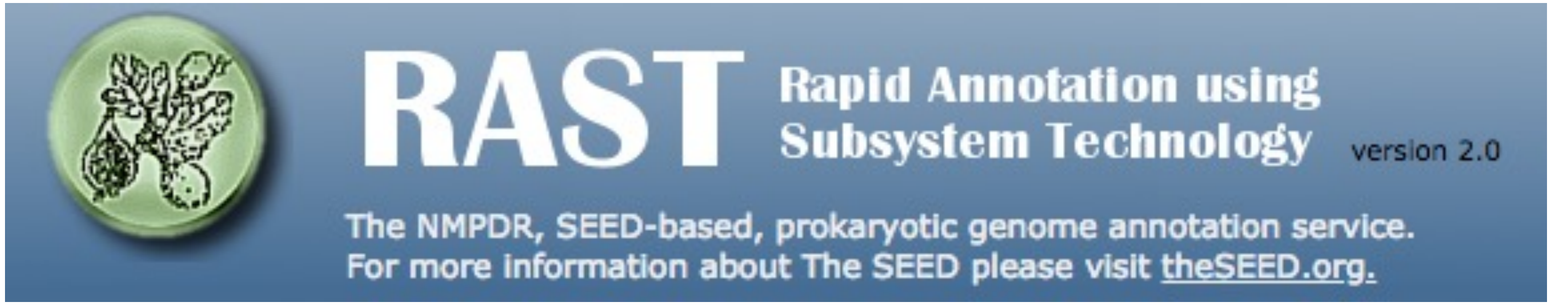


Whole-Genome Assembly

- Multiple approaches available
- No one-size-fits-all
- May need to try several techniques / software packages

Annotation

- Identification of genomic features (protein-coding sequences, RNA-encoding sequencings, others [CRISPRs, signal peptides, etc.])
- Online option: RAST
  - http://rast.nmpdr.org/ (includes written and video tutorials)
  - Requires registration (free)
  - Depending on server load, can take hours or days for results
  - Input: Fasta contig file
  - Output: Annotated genbank file



RAST Rapid Annotation using Subsystem Technology version 2.0

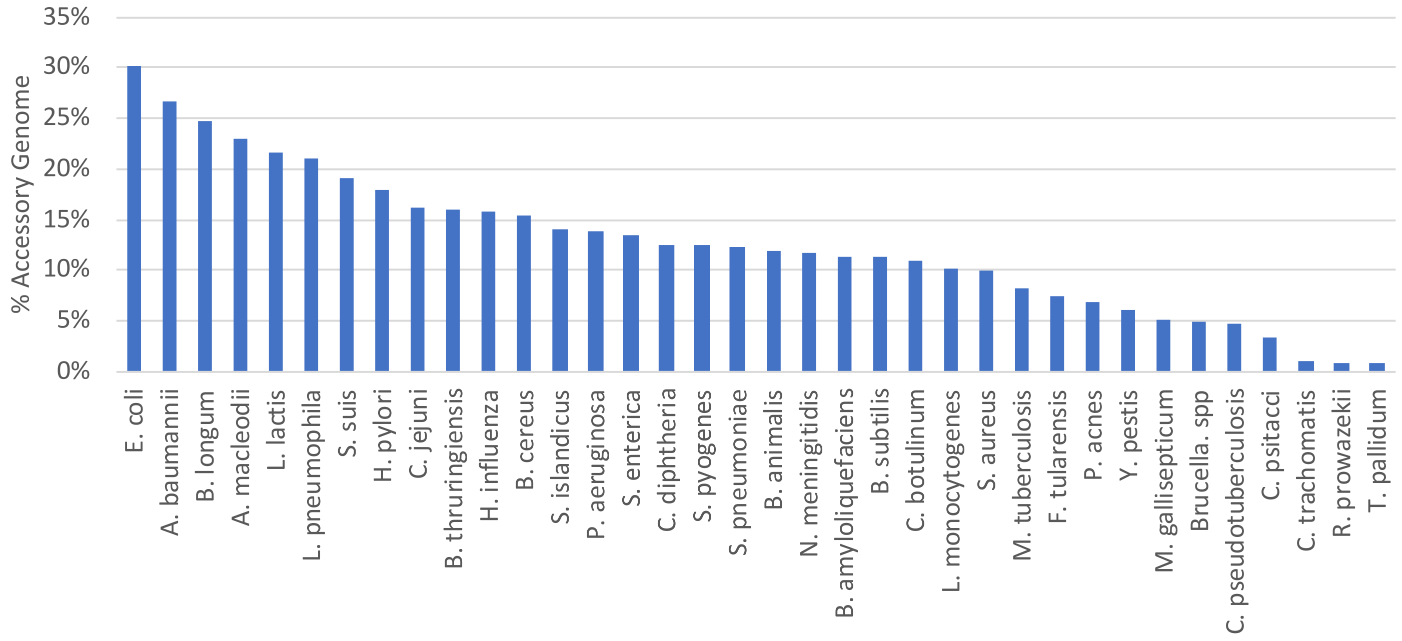
The NMPDR, SEED-based, prokaryotic genome annotation service.
For more information about The SEED please visit theSEED.org.

Annotation

- Command line option: Prokka
  http://www.vicbioinformatics.com/software.prokka.shtml
  - Advantages:
    - Local; no waiting on server load
    - Fast; less than 30 minutes per genome, usually
    - Output formatted for direct deposit to NCBI database
  - Disadvantages:
    - Requires installation of several support programs
    - Limited database, but customizable to your organism of interest

Core and Accessory Genome Analysis

- Core genome: Sequence shared by all or most representatives of a species
- Accessory genome: genetic elements present in some strains, absent in others
  - Plasmids, integrative and conjugative elements (ICEs), replacement islands, prophages and phage-like elements, transposons, insertion sequences (ISs), integrons



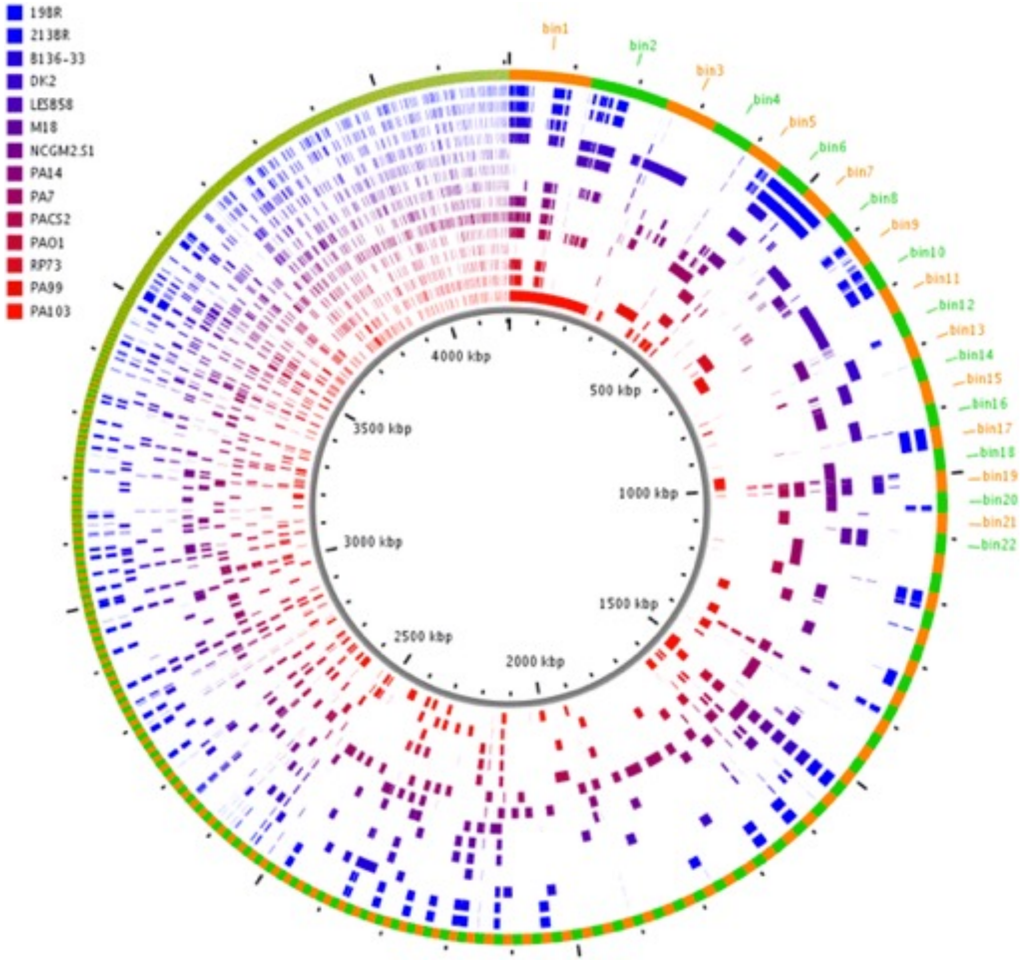
Adapted from: Bohlin J et al. BMC Genomics. 2017; 18: 151.

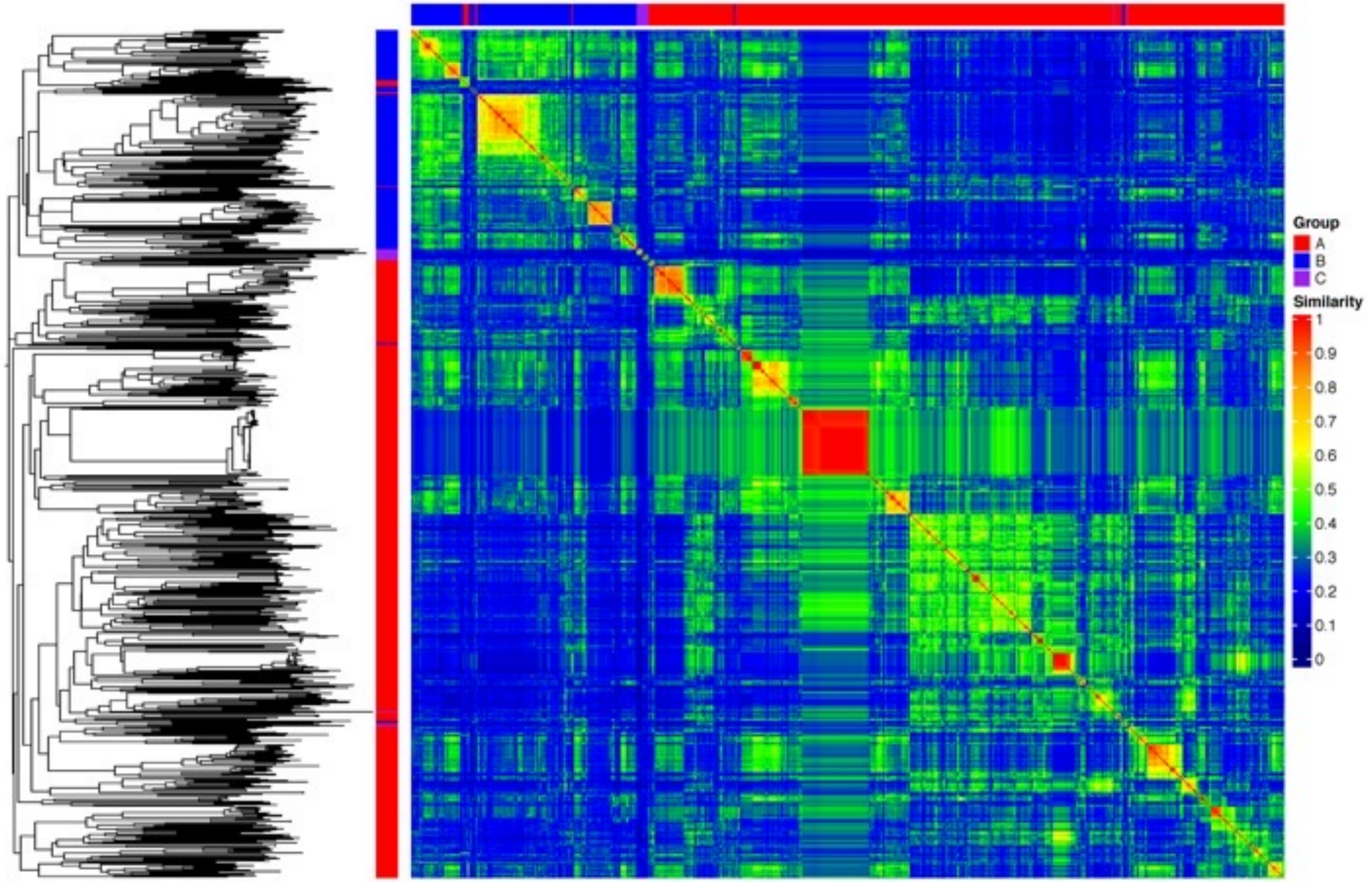
Core and Accessory Genome Analysis

- Spine, AGEnt, ClustAGE http://vfsmspineagent.fsm.northwestern.edu/index_age.html
- Spine: Identifies conserved core genome sequence using complete and/or draft sequences as input. Outputs representative core genome sequence
- AGEnt: Performs *in silico* subtractive hybridization to identify accessory genome
- ClustAGE: Determine the set of unique accessory sequences in a group of genomes and distribution of accessory elements among the isolates

Ozer EA, Allen JP, Hauser AR. BMC Genomics. 2014 Aug 29;15:737
Ozer EA.. BMC Bioinformatics 2018 19:150

Core and Accessory Genome Analysis

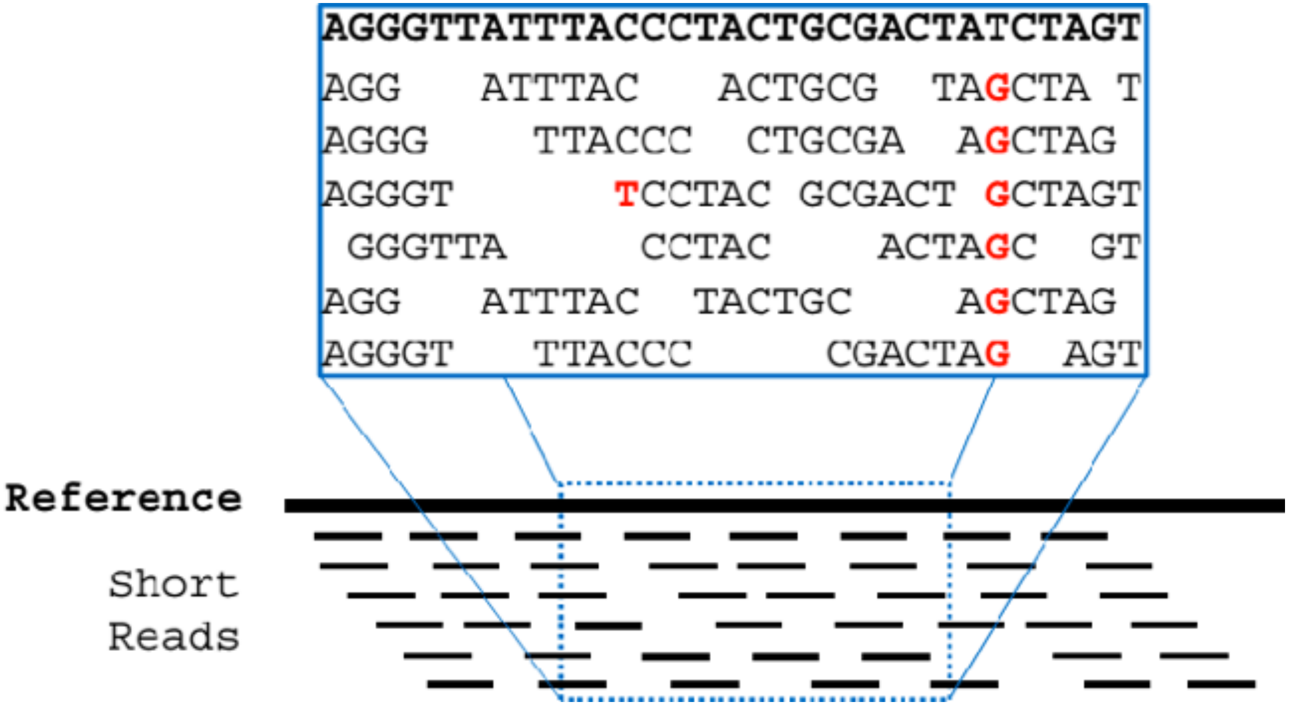




Ozer EA.. BMC Bioinformatics 2018 19:150

Alignment

- Align reads directly to a reference genome sequence (no assembly)
- Identify variants relative to reference


AGGGTTATTTACCCTACTGCGACTATCTAGT
AGG ATTTAC ACTGCG TAGCTA T
AGGG TTACCC CTGCGA AGCTAG
AGGGT TCCTAC GCGACT GCTAGT
GGGTTA CCTAC ACTAGC GT
AGG ATTTAC TACTGC AGCTAG
AGGGT TTACCC CGACTAG AGT
Reference
Short Reads

McVicar, Nathaniel et al. *ArXiv* abs/1805.00106 (2018)

Alignment

- Alignment programs:
  - bwa (Burrows-Wheeler aligner) http://bio-bwa.sourceforge.net/
  - Others: Stampy, Bowtie2, NovoAlign, Smalt

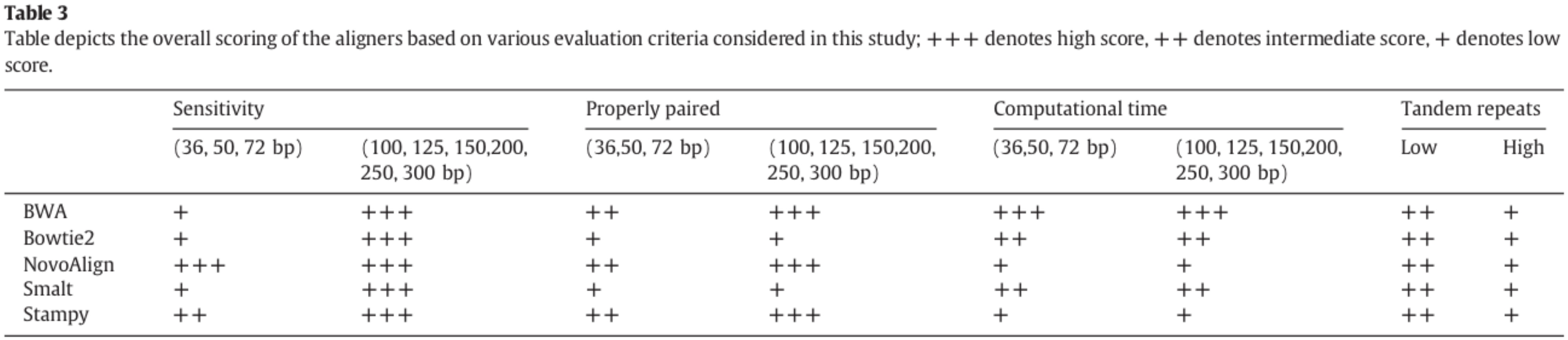


Table 3
Table depicts the overall scoring of the aligners based on various evaluation criteria considered in this study; +++ denotes high score, ++ denotes intermediate score, + denotes low score.

	Sensitivity		Properly paired		Computational time		Tandem repeats	
	(36, 50, 72 bp)	(100, 125, 150,200, 250, 300 bp)	(36,50, 72 bp)	(100, 125, 150,200, 250, 300 bp)	(36,50, 72 bp)	(100, 125, 150,200, 250, 300 bp)	Low	High
BWA	+	+++	++	+++	+++	+++	++	+
Bowtie2	+	+++	+	+	++	++	++	+
NovoAlign	+++	+++	++	+++	+	+	++	+
Smalt	+	+++	+	+	++	++	++	+
Stampy	++	+++	++	+++	+	+	++	+

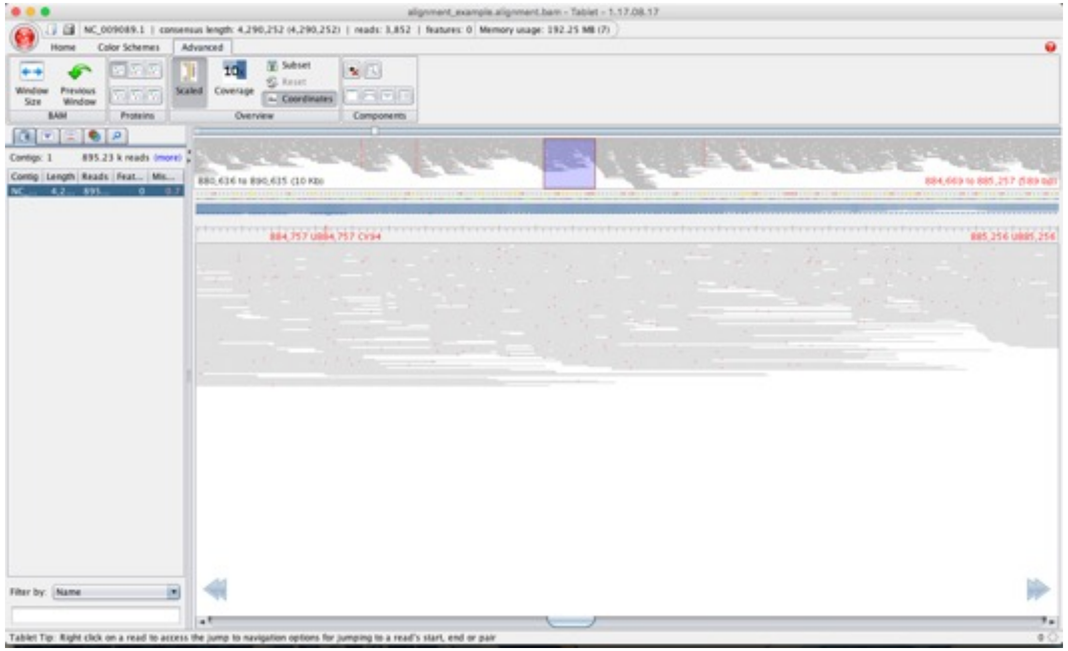
S. Thankaswamy-Kosalai et al. Genomics 109 (2017) 186–191

Alignment

- Inputs:
  - Reference genome sequence
  - Sequencing read files
- Output:
  - Alignment file, usually in SAM format
    - BAM is a binary-encoded SAM file
  - SAM file often post-processed using samtools program http://samtools.sourceforge.net/
  - Typical steps: filtering of non-aligned reads, sorting, indexing

Visualizing read alignments

- Tablet https://ics.hutton.ac.uk/tablet/
- Requires reference sequence file and sorted alignment file
  - Sam file = “flat” text file
  - Bam file = binary version of sam file. Tablet requires index file (.bai) produced by samtools to be in the same directory
- EXAMPLE:
  - Alignment: alignment_example.alignment.bam
  - Reference: alignment_example.reference.fasta



alignment_example.alignment.bam - Tablet - 1.17.08.17

NC_009089.1 | consensus length: 4,290,252 (4,290,252) | reads: 3,852 | features: 0 | Memory usage: 192.25 MB (7)

Contigs: 1 895.23 k reads (more)

880,636 to 890,635 (10 Kb)

884,669 to 885,257 (589 bp)

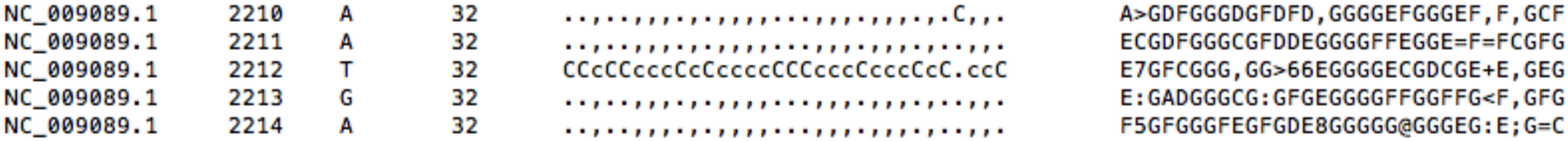
Filter by: Name

Tablet Tip: Right click on a read to access the jump to navigation options for jumping to a read's start, end or pair

Variant identification from alignments

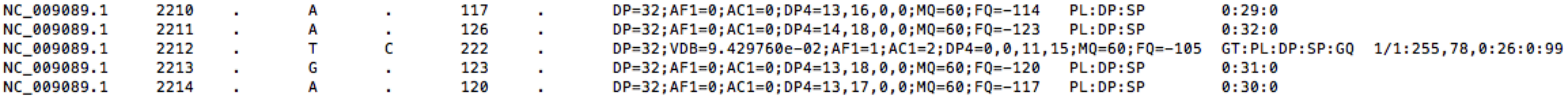
- Use alignment to identify variants (SNPs, indels) relative to the reference
- Programs:
  - Samtools / bcftools
    - http://www.htslib.org/
    - ‘bcftools mpileup’ to generate list of per-position alignments → ‘bcftools call’ to calculate SNP/indel calls in VCF format

Pileup:



```
NC_009089.1    2210    A    32    ..,..,,,.,.,,,,...,,,.,,,.,.C,,.    A>GDFGGGDGFDFD,GGGGEFGGGEF,F,GCF
NC_009089.1    2211    A    32    ..,..,,,.,.,,,,...,,,.,,,.,..,,.    ECGDFGGGCGFDDEGGGGFFEGGE=F=FCGFG
NC_009089.1    2212    T    32    CCcCCcccCcCccccCCCcccCcccCcC.ccC    E7GFCGGG,GG>66EGGGGECGDCGE+E,GEG
NC_009089.1    2213    G    32    ..,..,,,.,.,,,,...,,,.,,,.,..,,.    E:GADGGGCG:GFGEGGGGFFGGFFG<F,GFG
NC_009089.1    2214    A    32    ..,..,,,.,.,,,,...,,,.,,,.,..,,.    F5GFGGGFEGFGDE8GGGGG@GGGEG:E;G=C
```

VCF:

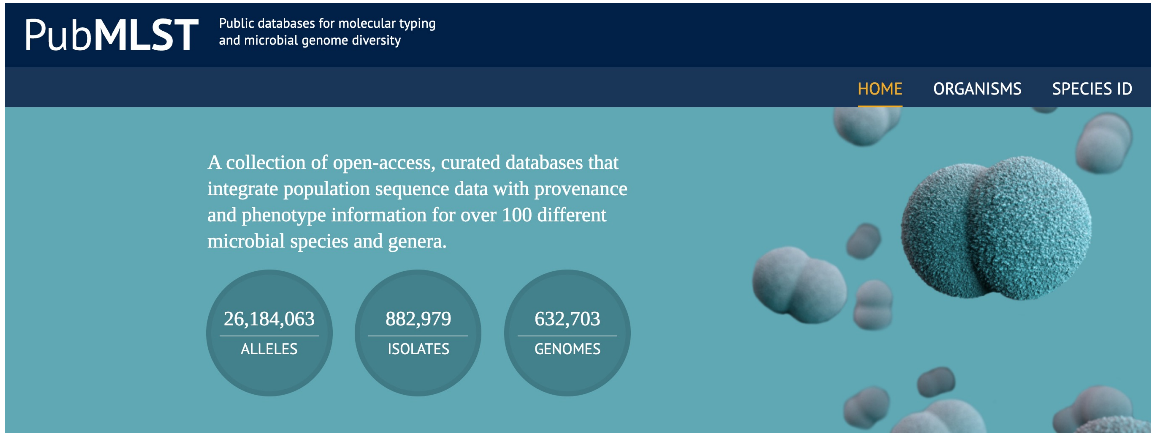


```
NC_009089.1   2210   .   A   .   117   .   DP=32;AF1=0;AC1=0;DP4=13,16,0,0;MQ=60;FQ=-114   PL:DP:SP        0:29:0
NC_009089.1   2211   .   A   .   126   .   DP=32;AF1=0;AC1=0;DP4=14,18,0,0;MQ=60;FQ=-123   PL:DP:SP        0:32:0
NC_009089.1   2212   .   T   C   222   .   DP=32;VDB=9.429760e-02;AF1=1;AC1=2;DP4=0,0,11,15;MQ=60;FQ=-105  GT:PL:DP:SP:GQ  1/1:255,78,0:26:0:99
NC_009089.1   2213   .   G   .   123   .   DP=32;AF1=0;AC1=0;DP4=13,18,0,0;MQ=60;FQ=-120   PL:DP:SP        0:31:0
NC_009089.1   2214   .   A   .   120   .   DP=32;AF1=0;AC1=0;DP4=13,17,0,0;MQ=60;FQ=-117   PL:DP:SP        0:30:0
```

Variant identification from alignments

- Other software:
  - FreeBayes https://github.com/ekg/freebayes
  - Also outputs in VCF format
  - Nice tutorial: http://clavius.bc.edu/~erik/CSHL-advanced-sequencing/freebayes-tutorial.html
- All-in-one solution
  - Snippy: https://github.com/tseemann/snippy
  - Pipeline for performing alignment (using bwa), variant calling (using FreeBayes), and multi-genome alignment for phylogenetics in microbial genomes

Phylogenetic Analysis



PubMLST Public databases for molecular typing and microbial genome diversity

A collection of open-access, curated databases that integrate population sequence data with provenance and phenotype information for over 100 different microbial species and genera.

26,184,063 ALLELES

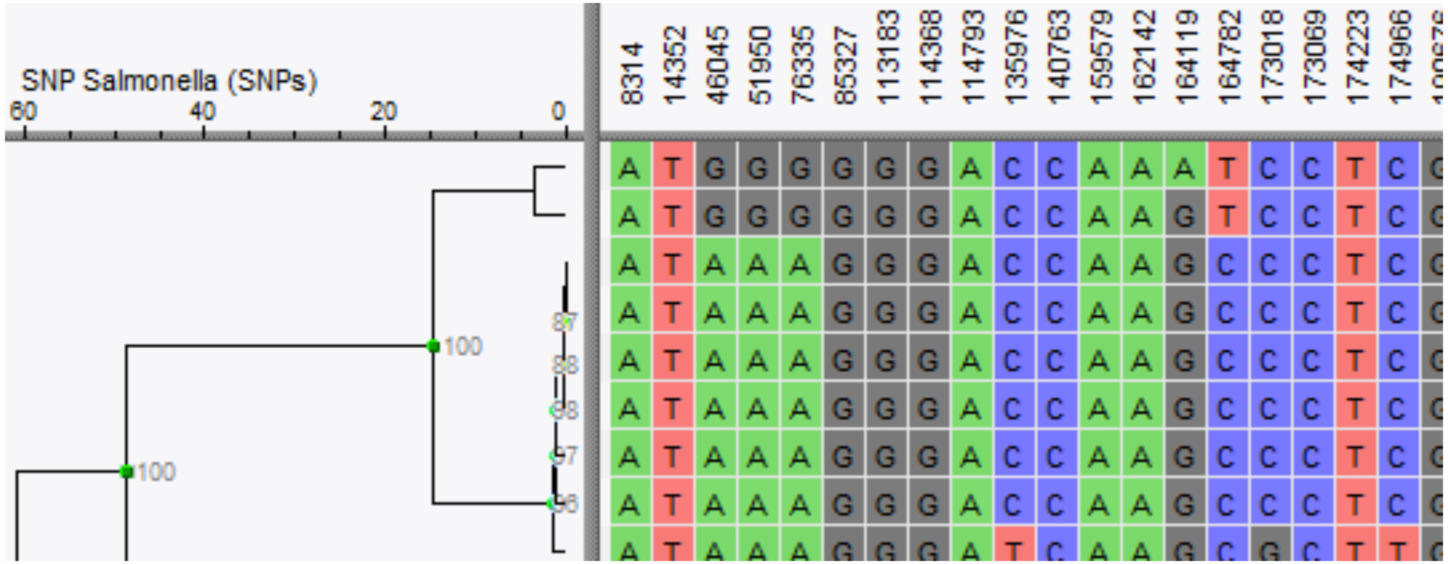
882,979 ISOLATES

632,703 GENOMES

- Isolate relatedness → epidemiologic or evolutionary inferences
- Multi-locus sequencing typing (**MLST**) or similar typing scheme
  - Usually 7 housekeeping genes
  - PCR amplification → Sanger sequencing
  - Gene sequences compared to database to assign allele numbers
  - Allele pattern associated with sequence type (ST) designation
- PubMLST https://pubmlst.org/
- Allele sequences from multiple strains could be concatenated and aligned to generate phylogenetic tree
- MLST trees are low resolution, can overestimate strain relatedness

Phylogenetics

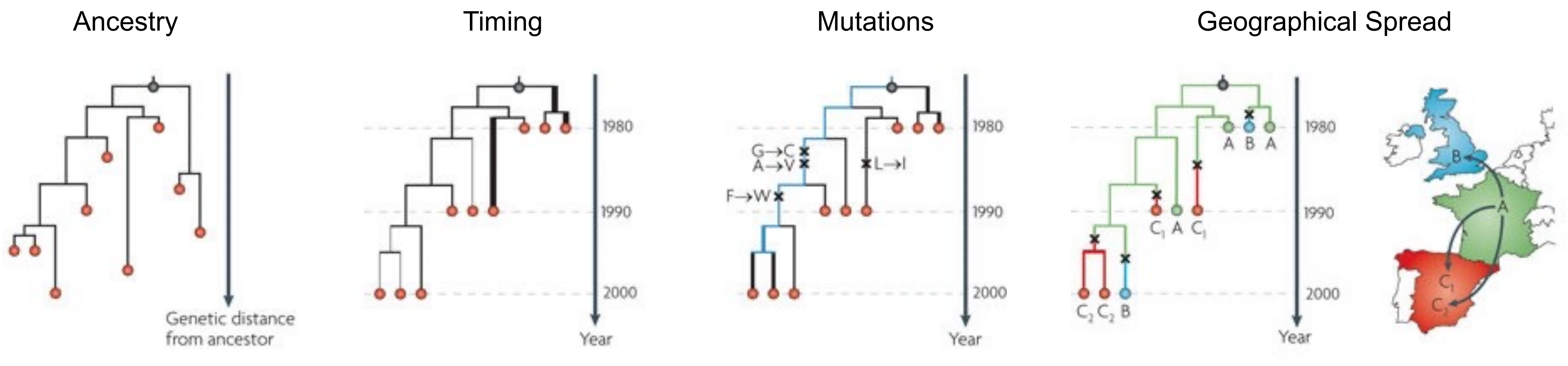
- From alignments, can use phylogenetic analysis to determine degrees of similarity and differences between genomes
  - Make inferences about transmission, evolution etc.
  - Can add also time of isolation or location to analysis



8314	14352	46045	51950	76335	85327	113183	114368	114793	135976	140763	159579	162142	164119	164782	173018	173069	174223	174966	
A	T	G	G	G	G	G	G	A	C	C	A	A	A	T	C	C	T	C	G
A	T	G	G	G	G	G	G	A	C	C	A	A	G	T	C	C	T	C	G
A	T	A	A	A	G	G	G	A	C	C	A	A	G	C	C	C	T	C	G
A	T	A	A	A	G	G	G	A	C	C	A	A	G	C	C	C	T	C	G
A	T	A	A	A	G	G	G	A	C	C	A	A	G	C	C	C	T	C	G
A	T	A	A	A	G	G	G	A	C	C	A	A	G	C	C	C	T	C	G
A	T	A	A	A	G	G	G	A	C	C	A	A	G	C	C	C	T	C	G
A	T	A	A	A	G	G	G	A	C	C	A	A	G	C	C	C	T	C	G
A	T	A	A	A	G	G	G	A	T	C	A	A	G	C	G	C	T	T	G

https://www.applied-maths.com/applications/whole-genome-snp-analysis

Whole Genome Sequencing can provide insights into the molecular epidemiology of pathogens



Ancestry

Timing

Mutations

Geographical Spread

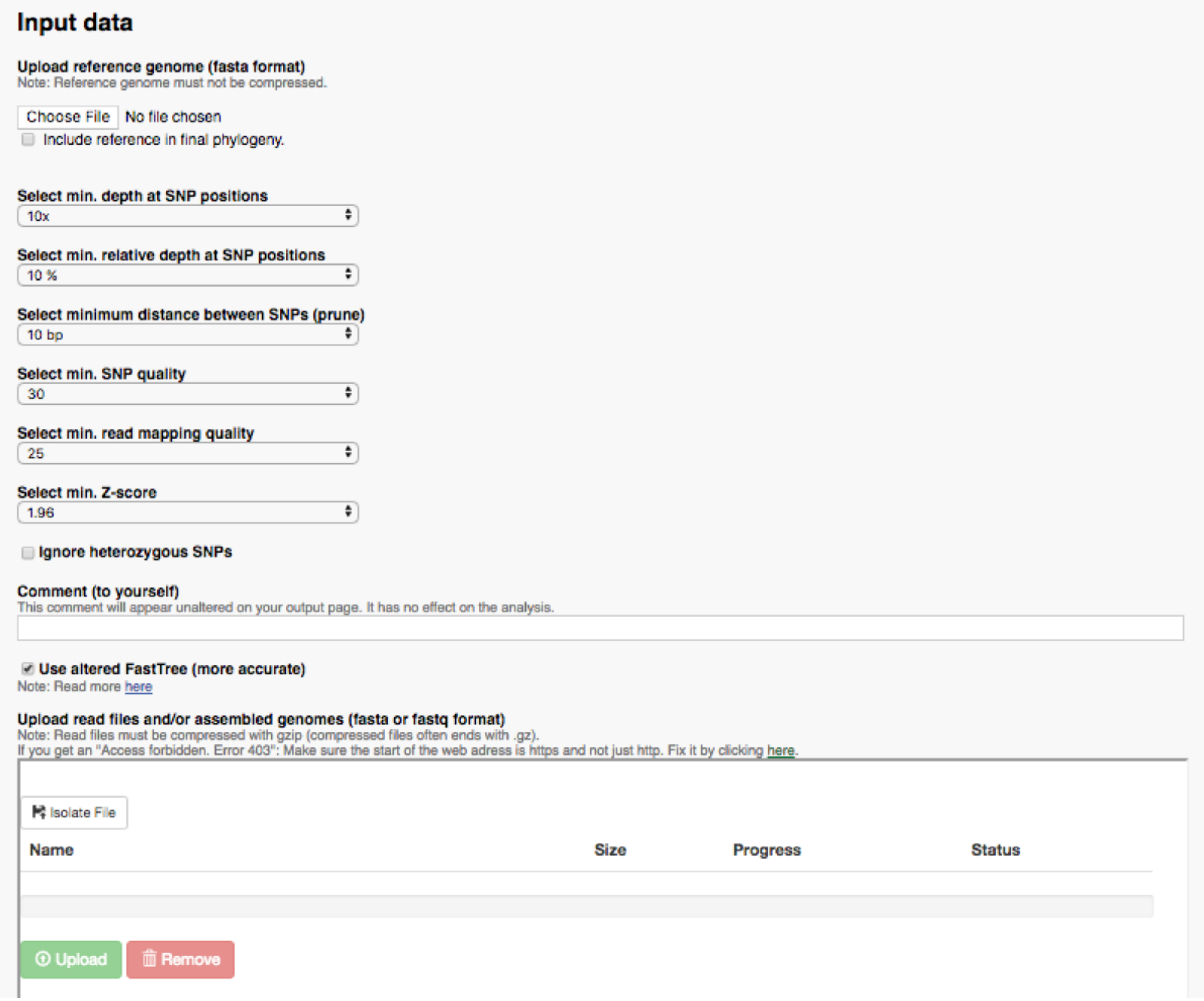
Pybus, O., Rambaut, A. 2009 *Nat. Rev. Genet.*

Whole-genome phylogenetics

- Reference-based alignment: Sequences aligned to a reference genome, variant positions identified relative to the reference
  - Can be slow, possible misalignments, can’t evaluate sequence not present in reference
  - CSI Phylogeny https://cge.cbs.dtu.dk/services/CSIPhylogeny/
  - REALPHY https://realphy.unibas.ch/fcgi/realphy

Whole-genome phylogenetics

- CSI Phylogeny



Input data

Upload reference genome (fasta format)
Note: Reference genome must not be compressed.

No file chosen

□ Include reference in final phylogeny.

Select min. depth at SNP positions
10x

Select min. relative depth at SNP positions
10 %

Select minimum distance between SNPs (prune)
10 bp

Select min. SNP quality
30

Select min. read mapping quality
25

Select min. Z-score
1.96

□ **Ignore heterozygous SNPs**

Comment (to yourself)
This comment will appear unaltered on your output page. It has no effect on the analysis.

☑ **Use altered FastTree (more accurate)**
Note: Read more here

Upload read files and/or assembled genomes (fasta or fastq format)
Note: Read files must be compressed with gzip (compressed files often ends with .gz).
If you get an "Access forbidden. Error 403": Make sure the start of the web adress is https and not just http. Fix it by clicking here.

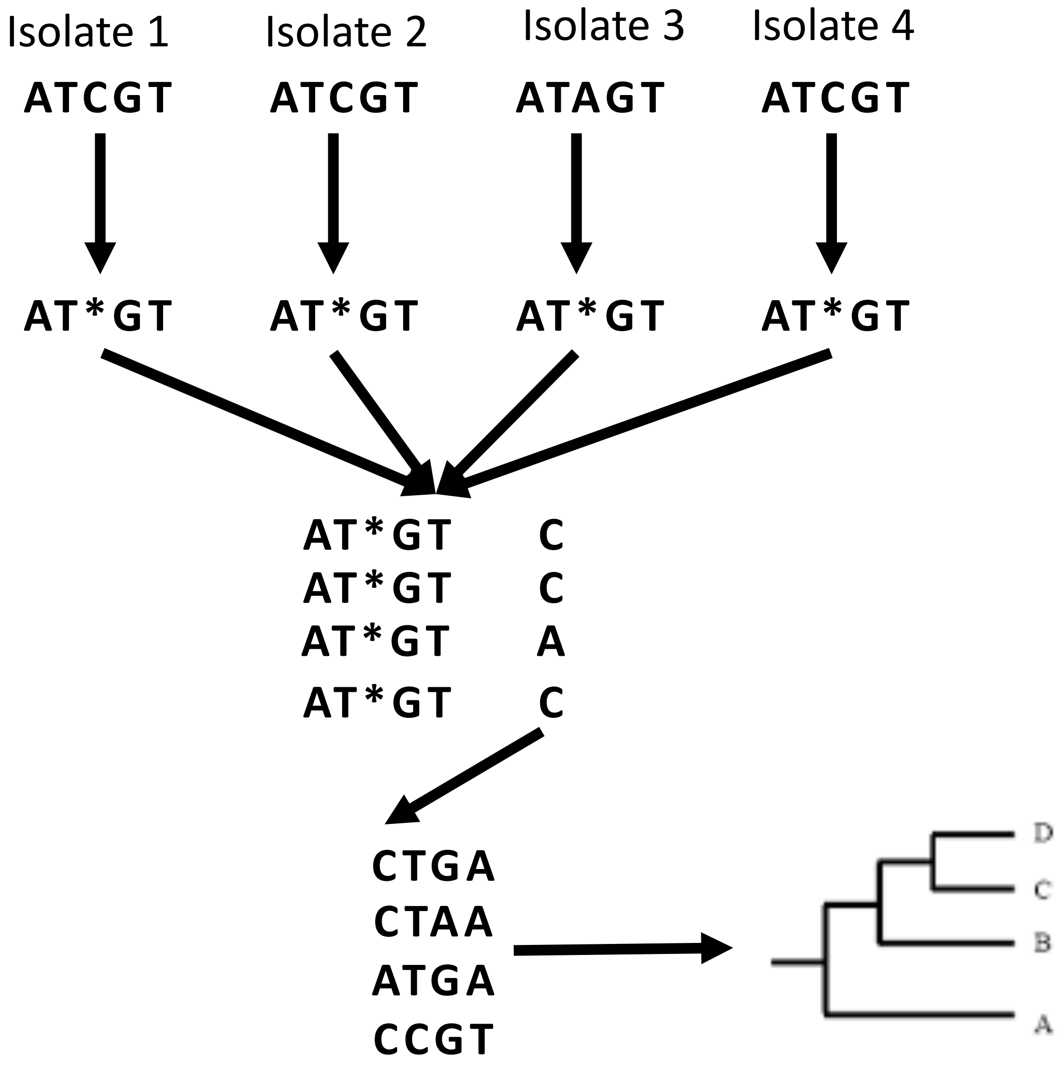
Name	Size	Progress	Status

Whole-genome phylogenetics

- Reference-free alignment: Sequences compared directly to each other to identify differences
  - Feature reduction improves computational performance, but may lose some resolution
  - kSNP https://sourceforge.net/projects/ksnp/
  - Mashtree: https://github.com/lskatz/mashtree

Whole-genome phylogenetics

- kSNP
  - Command line
  - Fragment each genome sequence into k-mers
  - Group all k-mers with wildcard as middle base
  - If middle base differs in any included isolate, count as SNP locus
  - Create matrix of all k-mers with variant middle bases
  - Generate phylogenetic tree



Isolate 1 Isolate 2 Isolate 3 Isolate 4

ATCGT ATCGT ATAGT ATCGT

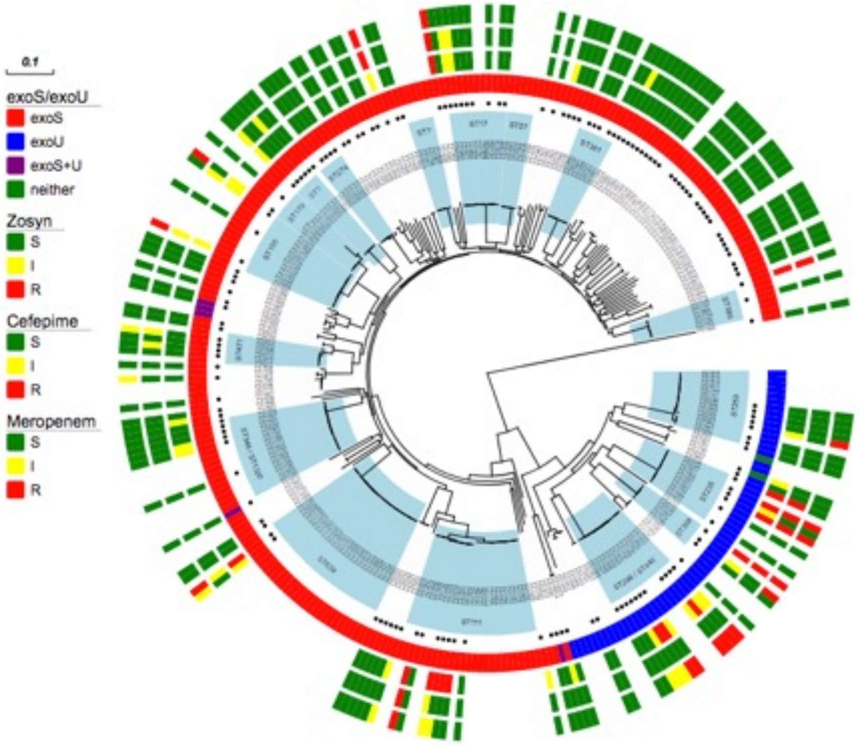
AT*GT AT*GT AT*GT AT*GT

AT*GT C
AT*GT C
AT*GT A
AT*GT C

CTGA
CTAA
ATGA
CCGT

Whole-genome phylogenetics

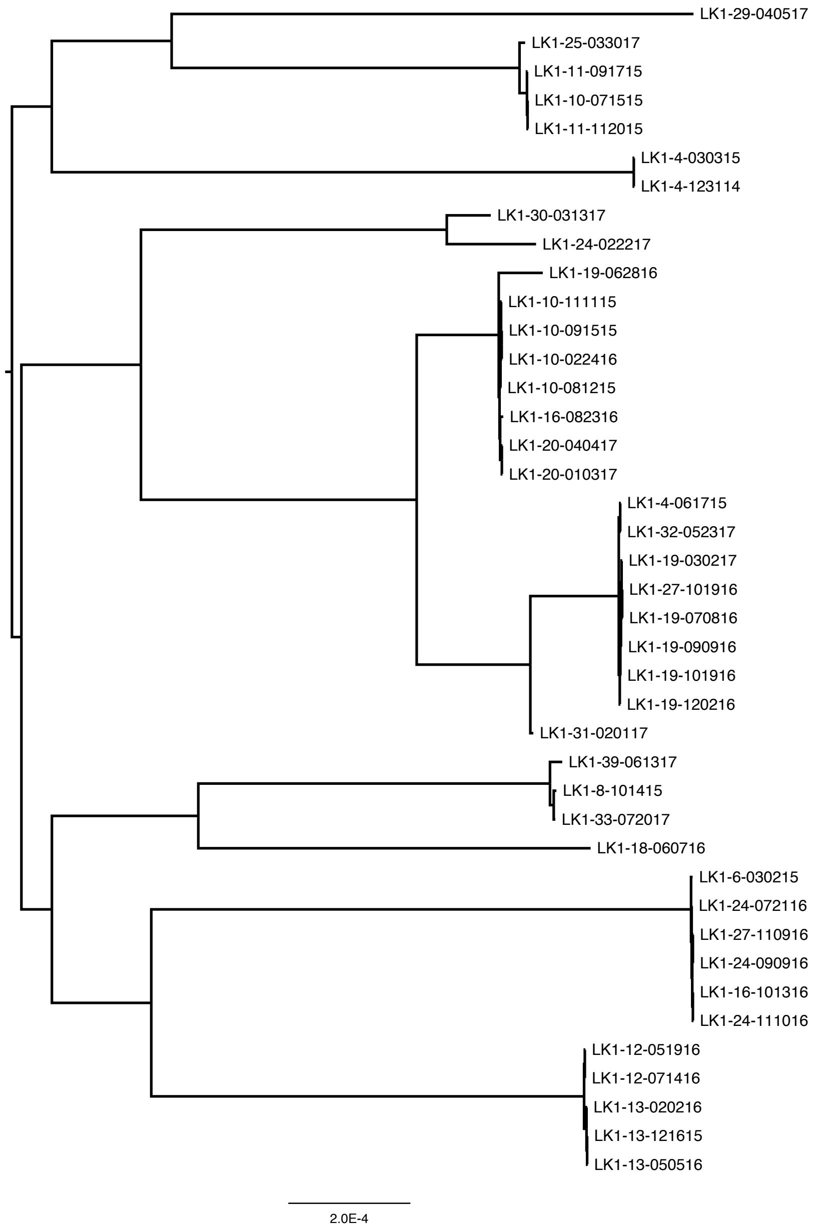
- Visualizing and annotating trees
- Common tree file format is Newick (.tre, .nwk)
- FigTree https://github.com/rambaut/figtree/releases
  - Local, nice for rapid visualization
  - EXAMPLE: tree_example.tre
- EvolView http://www.evolgenius.info/evolview/
  - Web-based, powerful annotation tools. Account required
- IToL https://itol.embl.de/
  - Web-based, account required. Heatmaps
- ggtree https://bioconductor.org/packages/release/bioc/html/ggtree.html
  - R-based package for tree annotation



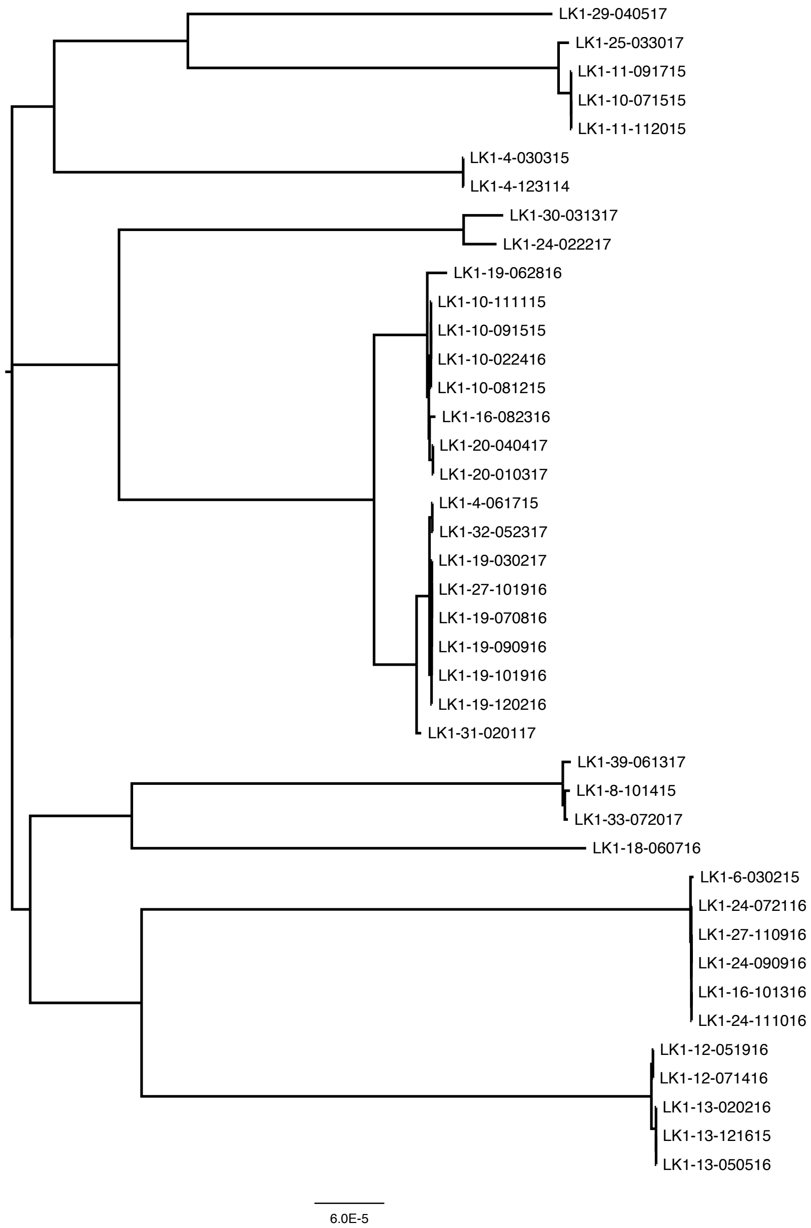
Recombination

- Most phylogenies assume a single evolutionary history
  - Acquisition of variants over time and propagation to descendants
- Recombination of microbial genomes can result in acquisition of multiple variants simultaneously independent of ancestral sequence
  - Can lead to false inferences about relative strain relatedness → increased branch lengths
  - Topology of trees not affected by presence of recombination
- Software for detecting and filtering regions subject to recombination in genome alignments:
  - ClonalFrameML https://github.com/xavierdidelot/ClonalFrameML
  - Gubbins https://sanger-pathogens.github.io/gubbins/

Didelot & Wilson. PLoS Comput Biol 2015 11(2): e1004041
Croucher N. J. et al. Nucleic Acids Research, 2014



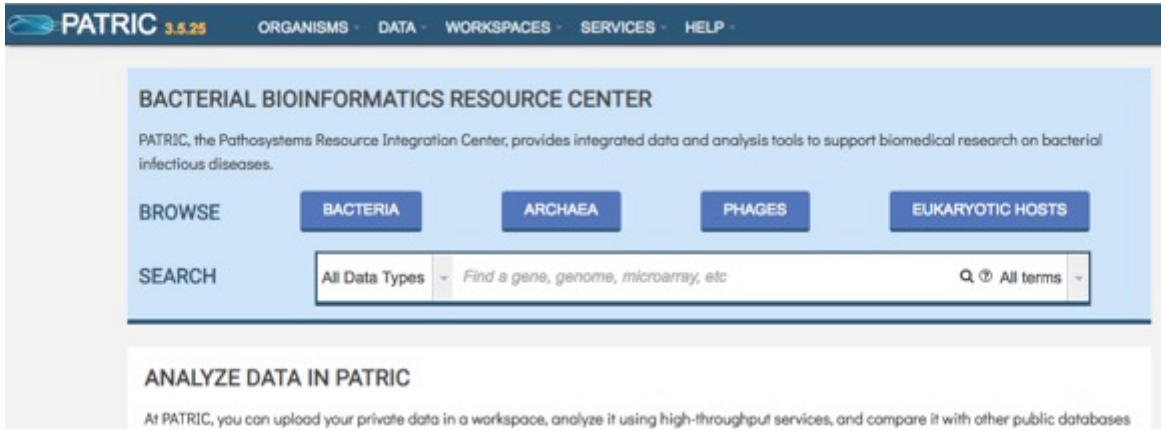
Core genome maximum likelihood tree



Following removal of recombination effects by ClonalFrameML

PATRIC (Pathosystems Resource Integration Center)

- https://www.patricbrc.org/
- Web-based service
- Services offered:
  - Assembly, alignment, annotation, phylogenetics, metagenomics, and much more
- Integration with NCBI



BACTERIAL BIOINFORMATICS RESOURCE CENTER

PATRIC, the Pathosystems Resource Integration Center, provides integrated data and analysis tools to support biomedical research on bacterial infectious diseases.

ANALYZE DATA IN PATRIC

At PATRIC, you can upload your private data in a workspace, analyze it using high-throughput services, and compare it with other public databases

Other resources

- Phylogenetic Trees Made Easy by Barry G. Hall
- Comparative Genomics Tutorial: https://holtlab.net/2017/07/01/update-to-comparative-bacterial-genomics-tutorial/