Sequencing Technologies and Microbial Genomics Overview

Egon A. Ozer, MD PhD

Director, Center for Pathogen Genomics and Microbial Evolution

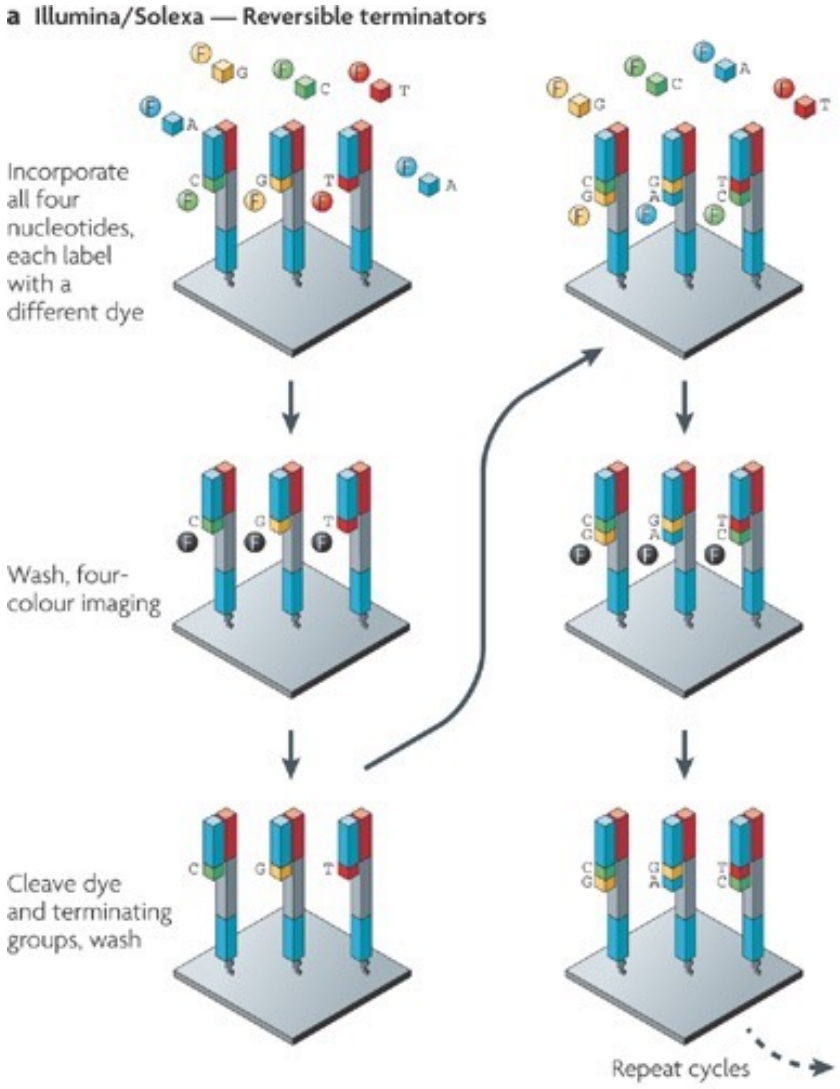
Department of Medicine, Division of Infectious Diseases

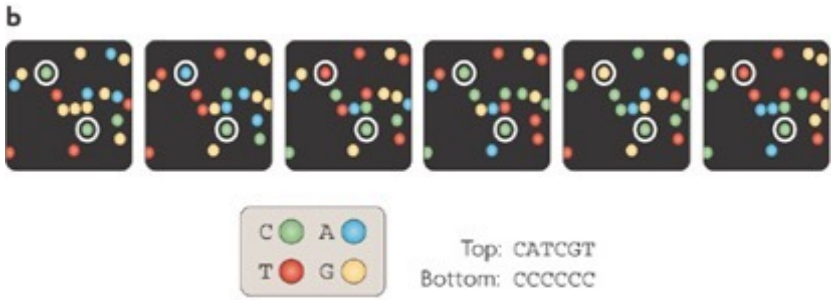
Northwestern University Feinberg School of Medicine

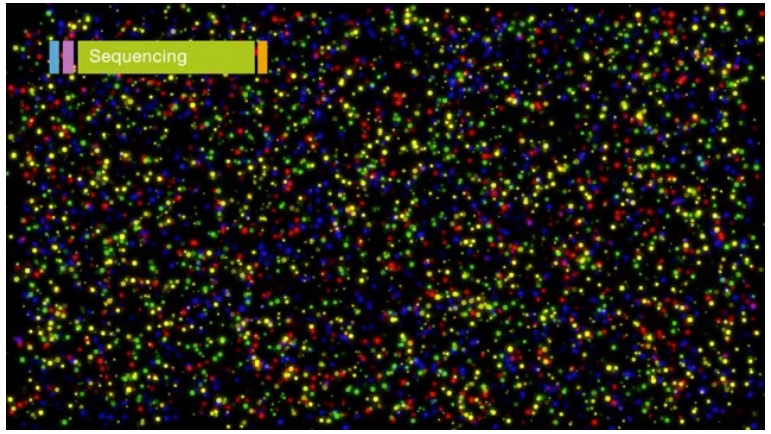
e-ozer@northwestern.edu

Next generation sequencing platforms

- Illumina
  - MiSeq
  - NextSeq
  - NovaSeq







Metzger ML. *Nat Rev Gen.* 2010. 11:31-46
https://www.illumina.com/systems/sequencing-platforms.html

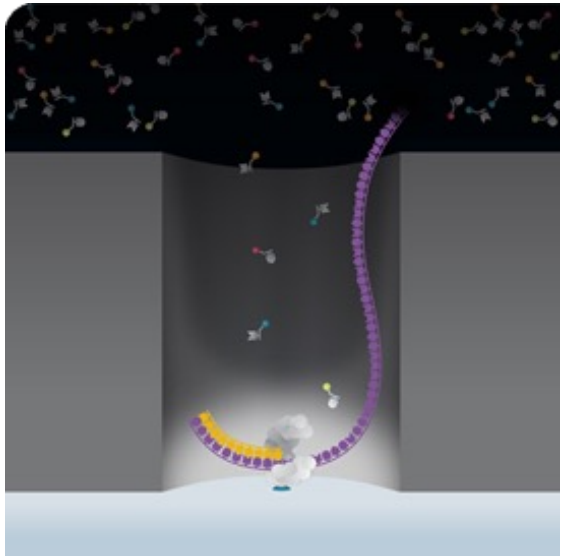
Whole-genome sequencing platforms

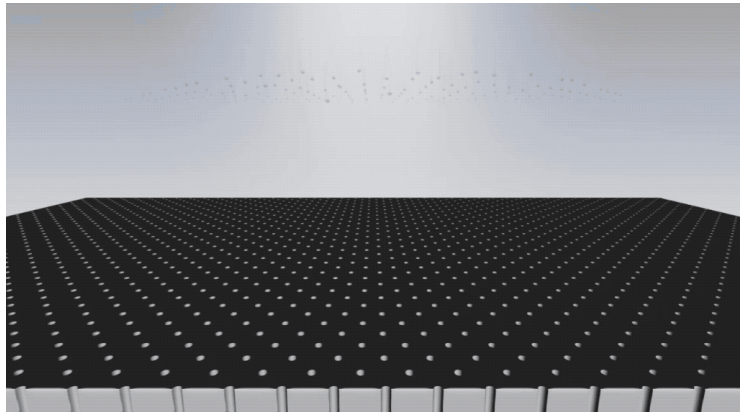
- Illumina (HiSeq, MiSeq, NextSeq, NovaSeq)
  - Benefits:
    - High-throughput
      - MiSeq: 15 Gb per flow cell
      - NextSeq 2000: 540 Gb per flow cell
      - NovaSeq X: 8,000 Gb (8 Tb) per flow cell
    - Low error rate (~ 0.1%) – substitution errors more common than indel
    - Relatively low cost-per-base
      - ¢0.02 – ¢0.6 – ¢13.5 / Mb (flow cell only)
  - Drawbacks:
    - PCR amplification required for sequencing
    - Short reads (50 - 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

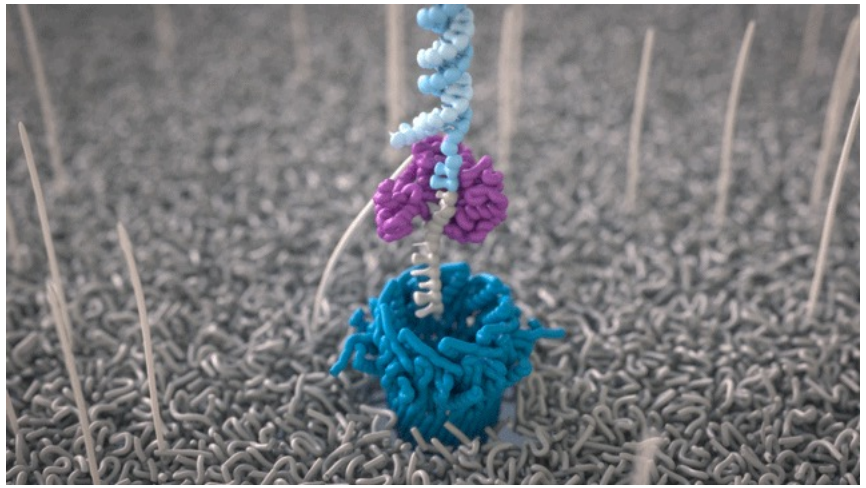
Whole-genome sequencing platforms

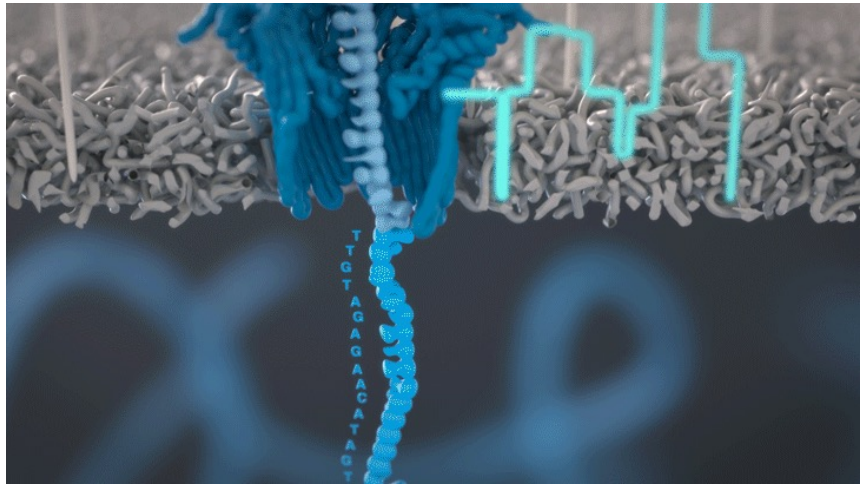
- PacBio (Sequel, Sequel II, Revio)
  - Benefits:
    - Long reads (15 – 20 kb)
    - Intermediate - high throughput (30 Gb - 90 Gb)
    - Fast: run time 4 - 30 hours
    - No PCR amplification necessary
  - Drawbacks:
    - Higher error rates than Illumina (5 – 15%) - substitution and indel
      - Error rates can be much lower (<1%) with circular consensus libraries (CCL), but homopolymers can still be a problem
    - Higher cost-per-base than Illumina platforms
      - ¢1.4 - ¢6.3 / Mb (flow cell only)

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

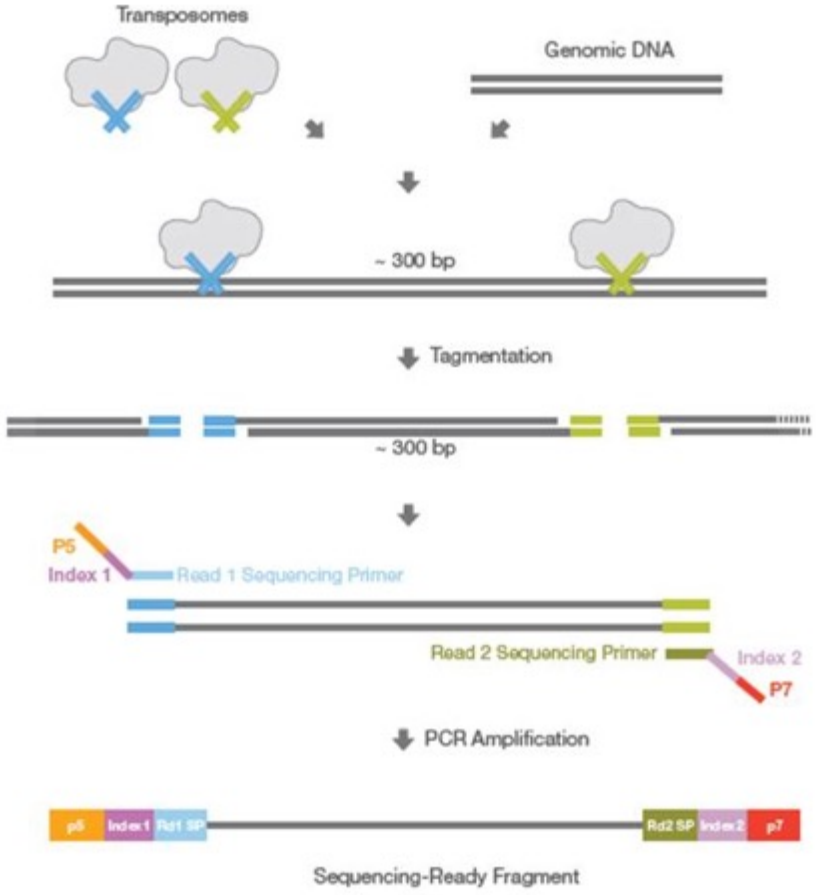
Whole-genome sequencing platforms

- Oxford Nanopore (MinION, GridION, PromethION)
  - Benefits:
    - Long reads (up to 900 kb)
    - Intermediate to high throughput
      - MinION / GridION: 35 Gb per flow cell
      - PromethION: 200 Gb per flow cell
    - Fast: real-time results, run length depends on desired read depth
    - Affordable equipment costs (~ $2000 for instrument, $700 per flow cell)
    - No PCR amplification necessary
  - Drawbacks:
    - Error rates higher than Illumina (~ 1%) - substitution and indel
    - Higher cost-per-base than (most) Illumina platforms
      - ¢0.9 - ¢2.7 / Mb (flow cell only)

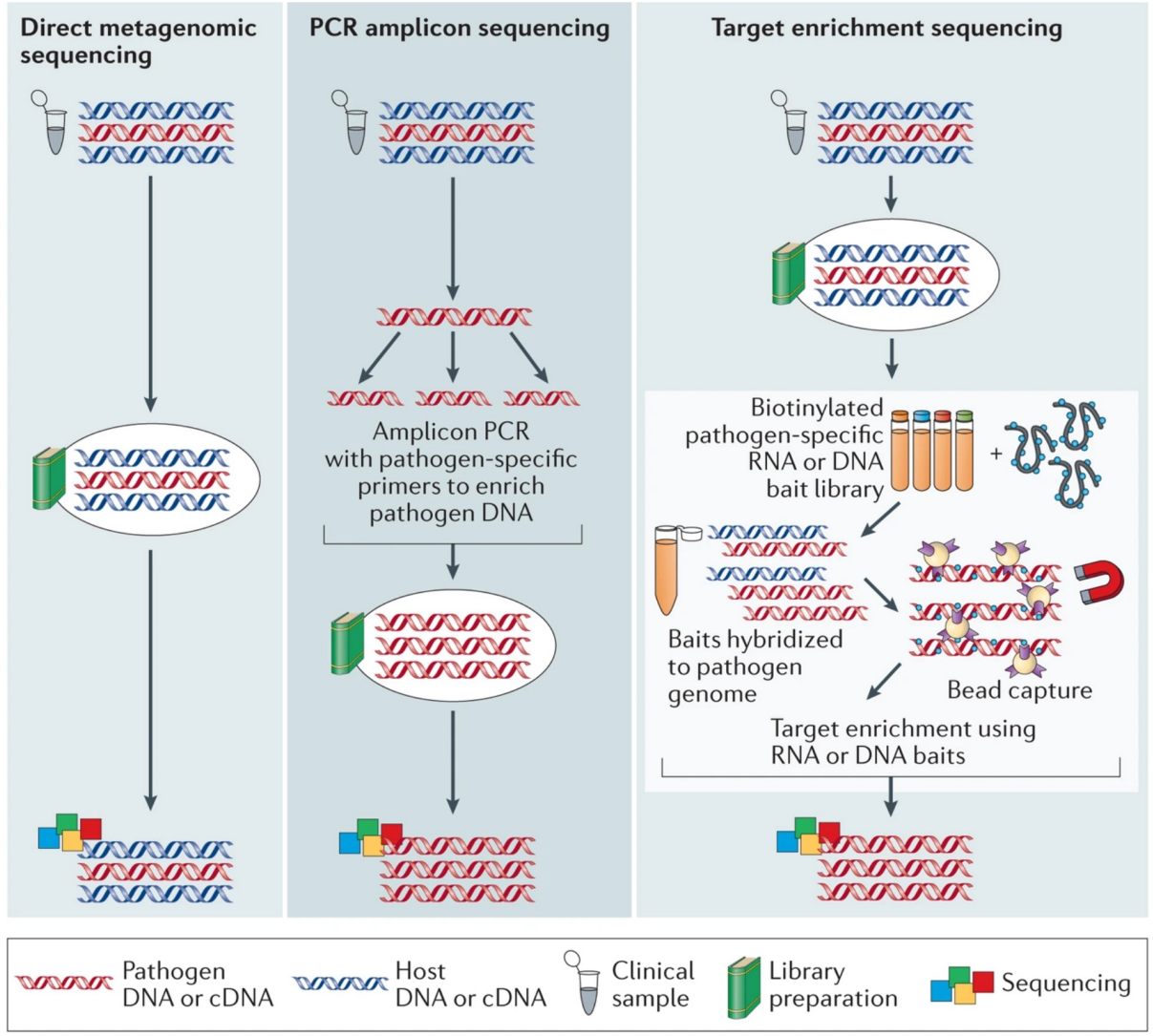
https://nanoporetech.com/products/comparison

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)



Sequencing from Non-Cultured Specimens



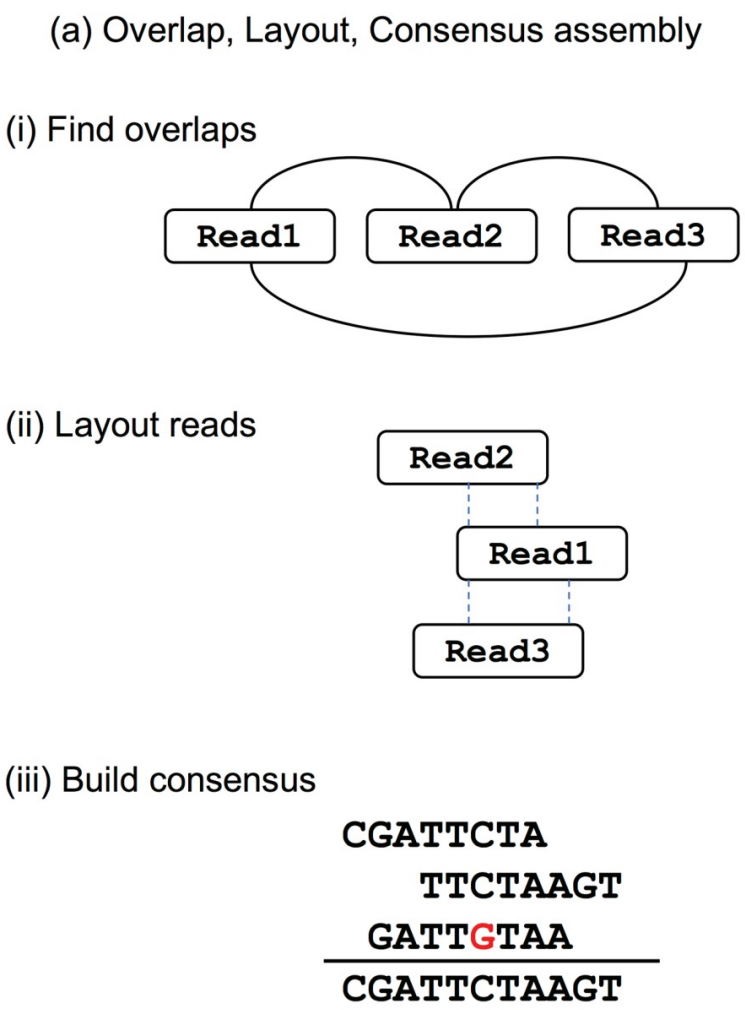
Houldcroft, Beale & Breuer. *Nat Rev Microbiol* **15,** 183–192 (2017).

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

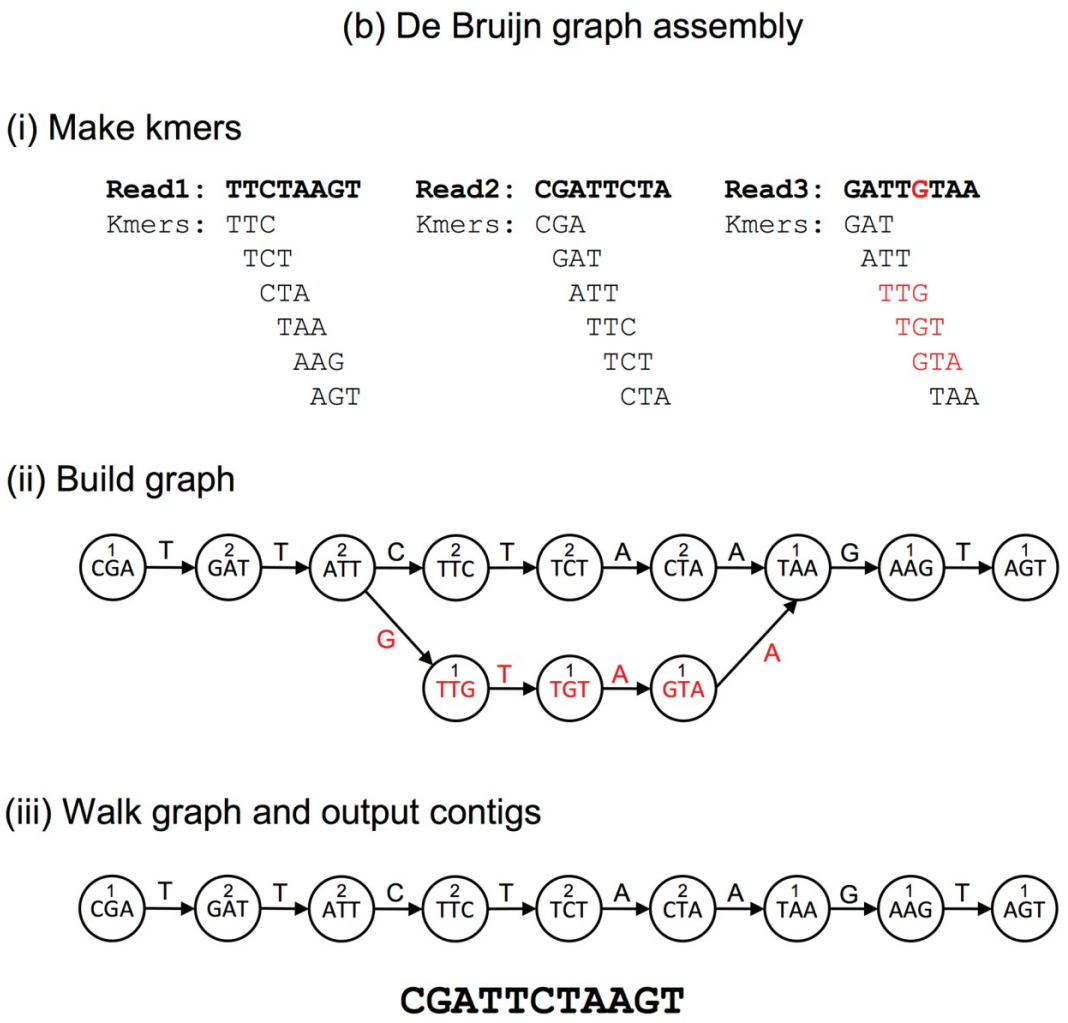
- Overlap layout consensus (OLC)
  - 1) Find overlaps among the reads, 2) create layout of all reads, 3) infer consensus sequence
  - Can be memory & computationally intensive
  - Best for lower numbers of long reads (PacBio or Nanopore)
  - Example software: Celera, miniasm



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



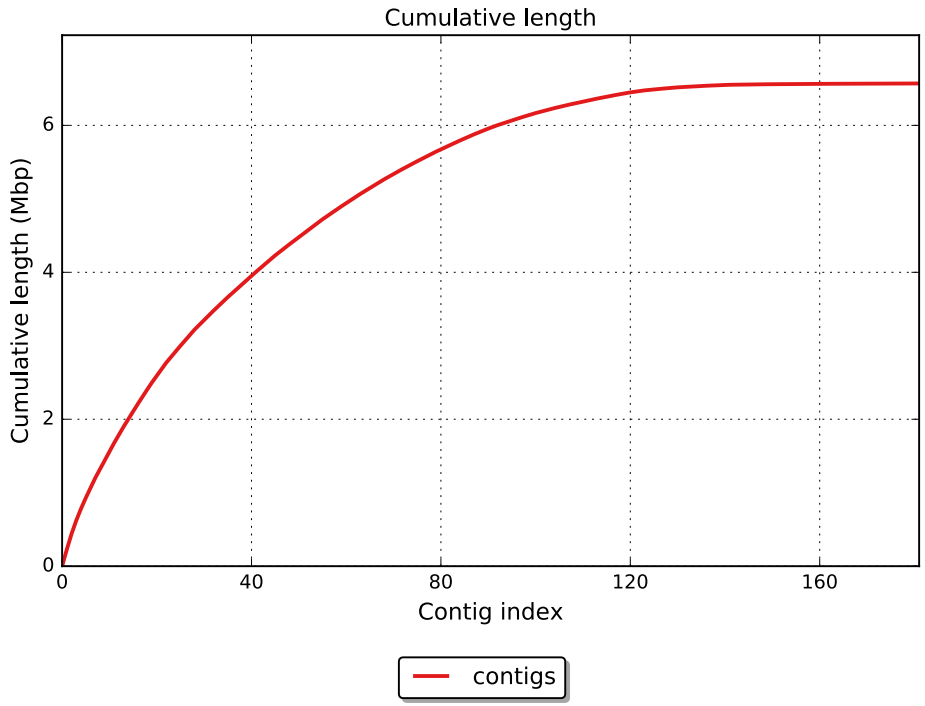
Ayling M, Clark MD, and Leggett RM. Brief in Bioinf. Feb 2019

Assembly

- SPAdes Assembler
  - http://cab.spbu.ru/software/spades/
  - Manual: http://cab.spbu.ru/files/release3.13.0/manual.html
- 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

Assembly

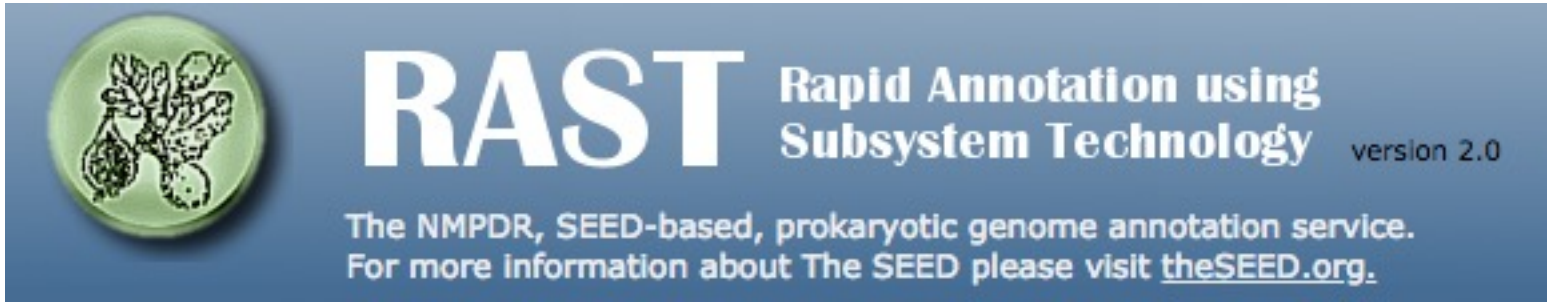
- Assessing results
  - Quast
    - Web: http://cab.cc.spbu.ru/quast/
    - Command Line: http://quast.sourceforge.net/quast



	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

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

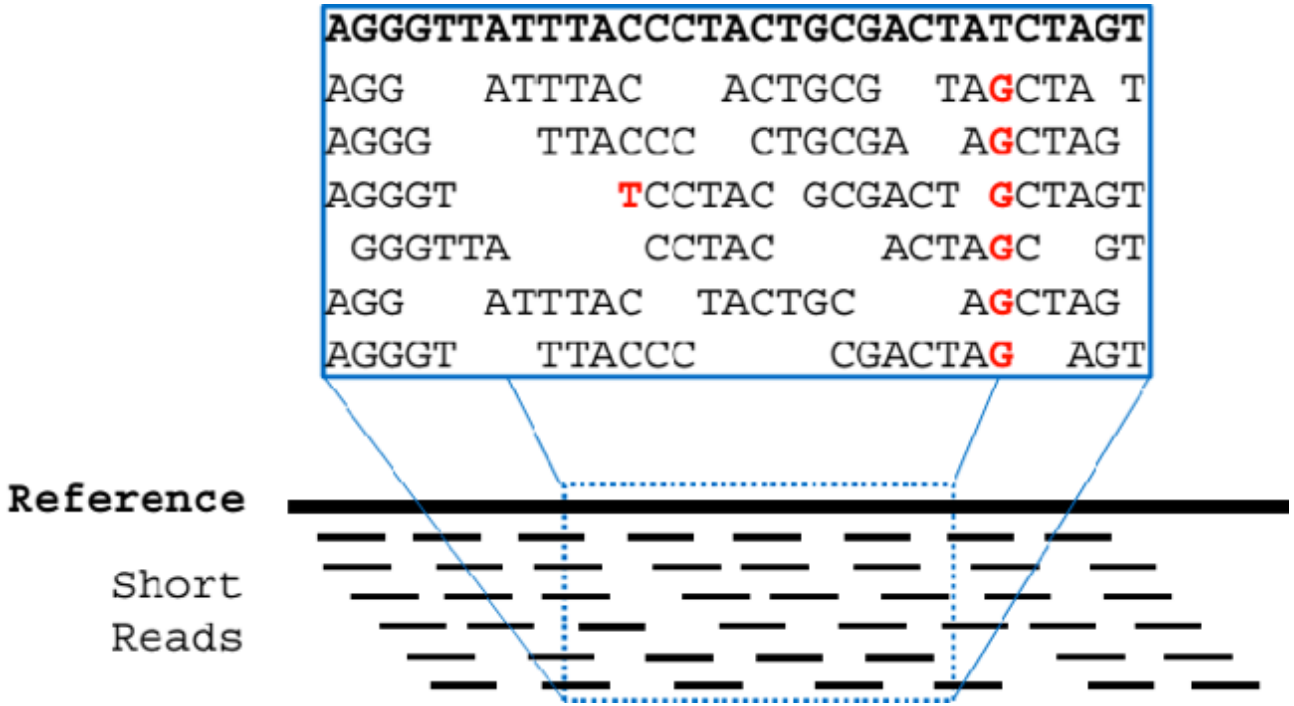


Annotation

- Command line option: PGAP https://github.com/ncbi/pgap
  - Official prokaryote annotation pipeline from NCBI
  - *Ab initio* gene prediction algorithms with homology-based methods
  - Curated protein profile hidden Markov models (HMMs), curated complex domain architectures for functional annotation of proteins
  - Advantages: Comprehensive, well-supported
  - Disadvantages: Big database, slow (hours), resource intensive
- 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:
    - Limited database, but customizable to your organism of interest

Alignment

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



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	++	+++	++	+++	+	+	++	+

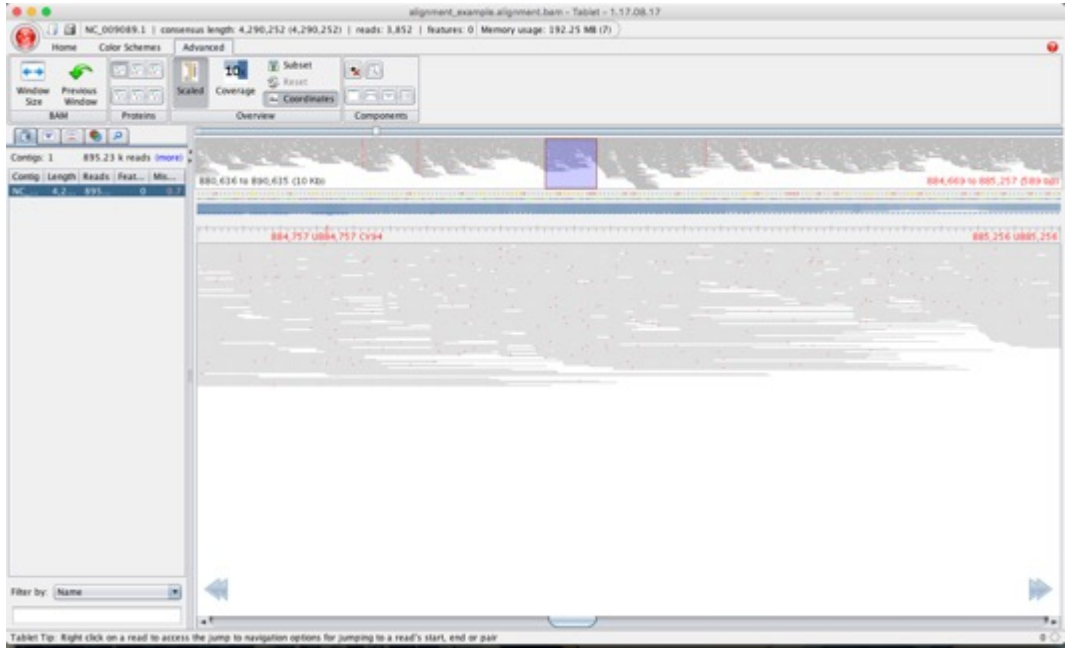
S. Thankaswamy-Kosalai et al. Genomics 109 (2017) 186–191

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



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
  - FreeBayes https://github.com/ekg/freebayes
    - 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

BV-BRC (Bacterial and Viral Bioinformatics Resource Center)

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