Genome and Transcriptome Assembly

FAS Informatics

April 3, 2014

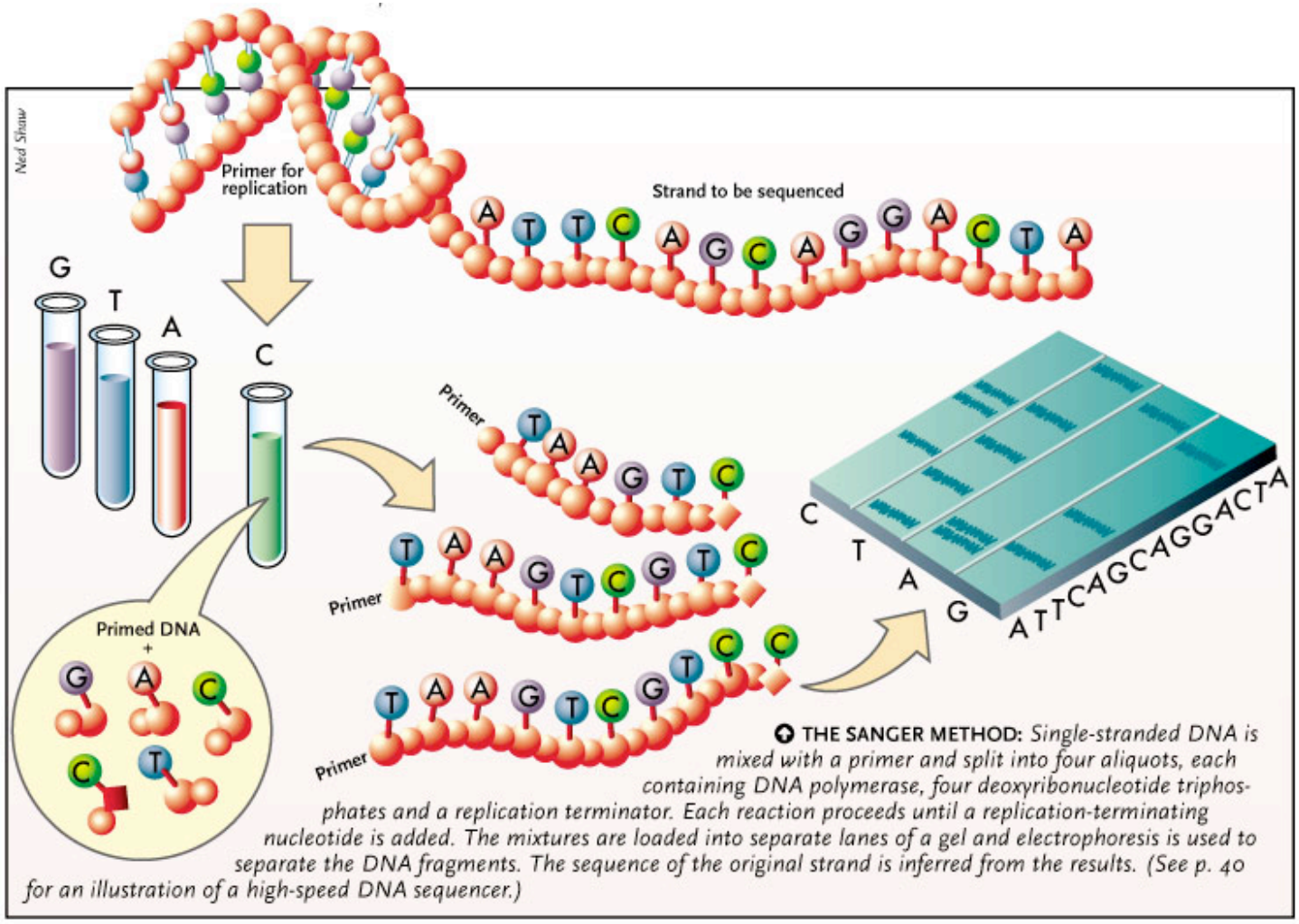
Agenda

- Genome assembly
  - Traditional (Sanger)
  - Next-Gen Sequencing
  - Hybrid approaches
  - Example: *Arthrobacter* bacteria
- Transcriptome assembly
  - Algorithms and assemblers
  - Example: Zebra finch
- New methods

Agenda

- Genome assembly
  - Traditional (Sanger)
  - Next-Gen Sequencing
  - Hybrid approaches
  - Example: *Arthrobacter* bacteria
- Transcriptome assembly
  - Algorithms and assemblers
  - Example: Zebra finch
- New methods

Sanger Sequencing



THE SANGER METHOD: *Single-stranded DNA is mixed with a primer and split into four aliquots, each containing DNA polymerase, four deoxyribonucleotide triphosphates and a replication terminator. Each reaction proceeds until a replication-terminating nucleotide is added. The mixtures are loaded into separate lanes of a gel and electrophoresis is used to separate the DNA fragments. The sequence of the original strand is inferred from the results. (See p. 40 for an illustration of a high-speed DNA sequencer.)*

http://www.the-scientist.com/?articles.view/articleNo/15939

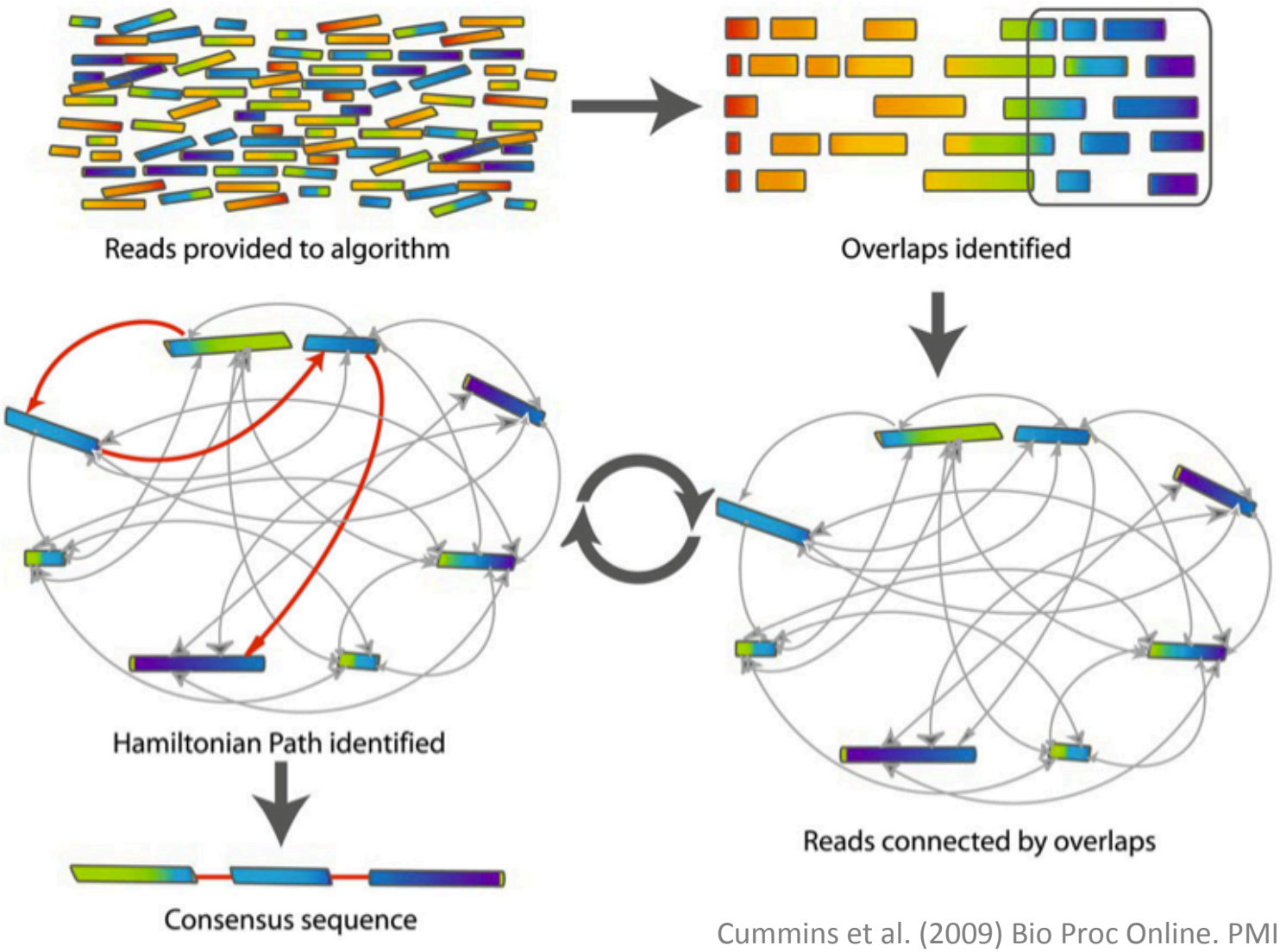
Sanger Sequencing

- 500–1000 nt reads
- Adjacent reads usually overlap by a couple of hundred base pairs
- Coverage needed for assembly: 6-8x
- Error rate < 2%
- “Gold standard” of genome sequencing

Assembly for Sanger Sequencing: Overlap-Layout-Consensus

- Overlap
  - Find pairs of reads sharing k-mers (k = ~24)
  - Try to extend the alignment with dynamic programming
- Layout
  - Create local multiple alignments from pair-wise overlaps
- Consensus
  - Derive consensus sequence via weighted voting of aligning segments

Assembly for Sanger Sequencing: Overlap-Layout-Consensus



Cummins et al. (2009) Bio Proc Online. PMID: 19495914

Assembly for Sanger Sequencing: Overlap-Layout-Consensus

- Requires computation of all pair-wise alignments $O(n^2)$
  - Ameliorated by kmer sorting (k = ~24)
- No efficient algorithm to find Hamiltonian path (all vertices) through overlap
- OLC is sensitive to kmer size and minimum percent identify required for an overlap

Assembly for Sanger Sequencing: Overlap-Layout-Consensus

OLC Assemblers:

- ARACHNE
  - Used to assemble Drosophila genome from Sanger reads in 2000
- CELERA
  - Used to assemble human genome from Sanger reads in 2001
- Phrap
- CAP
- TIGR

Agenda

- Genome assembly
  - Traditional (Sanger)
  - Next-Gen Sequencing
  - Hybrid approaches
  - Example: *Arthrobacter* bacteria
- Transcriptome assembly
  - Algorithms and assemblers
  - Example: Zebra finch
- New methods

Agenda

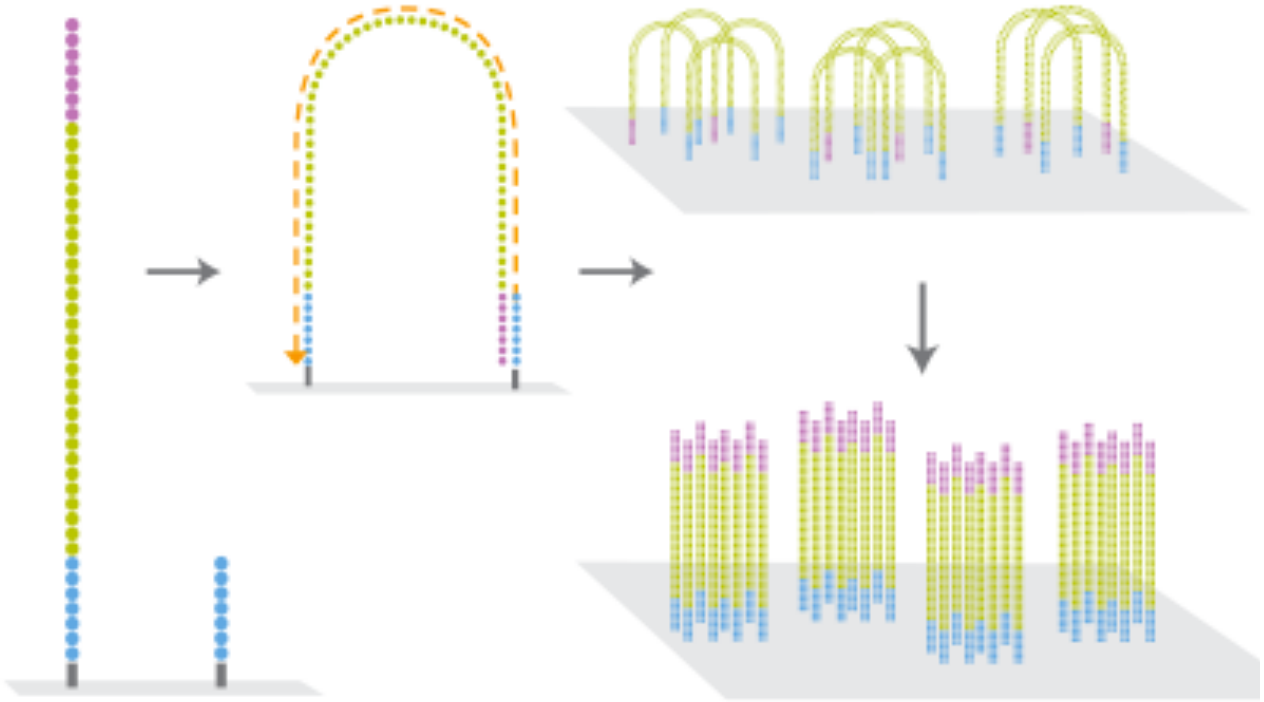
- Genome assembly
  - Traditional (Sanger)
  - Next-Gen Sequencing
  - Hybrid approaches
  - Example: *Arthrobacter* bacteria
- Transcriptome assembly
  - Algorithms and assemblers
  - Example: Zebra finch
- New methods

Next-Gen Sequencing

- Shorter reads: 35–150 nt from fragments 200-500 nt in length
- Per-base cost is 1/100th to 1/1000th of Sanger sequencing
- As many as 100M reads per sample
- Quantitative
- Error rate < 2%
- Capability to sequence both ends of library fragment

Next-Gen Sequencing (Illumina)

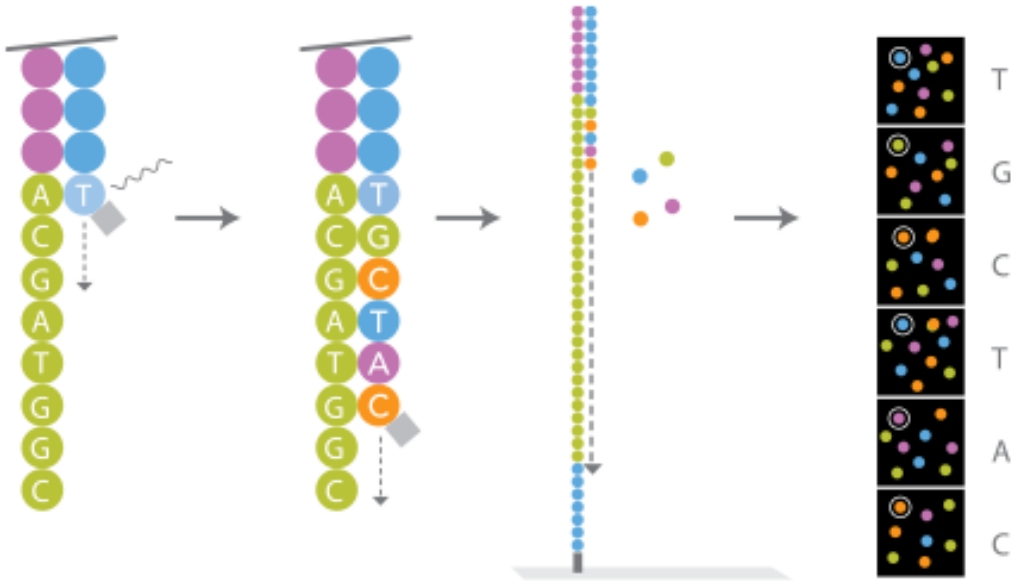
1. Adaptor-ligated fragments fixed to surface
2. Fragments amplified



http://illumina.com

Next-Gen Sequencing (Illumina)

3. Adaptor-ligated fragments fixed to surface
4. Fragments amplified
5. Fluorescent, reversibly terminated bases added
6. Surface imaged

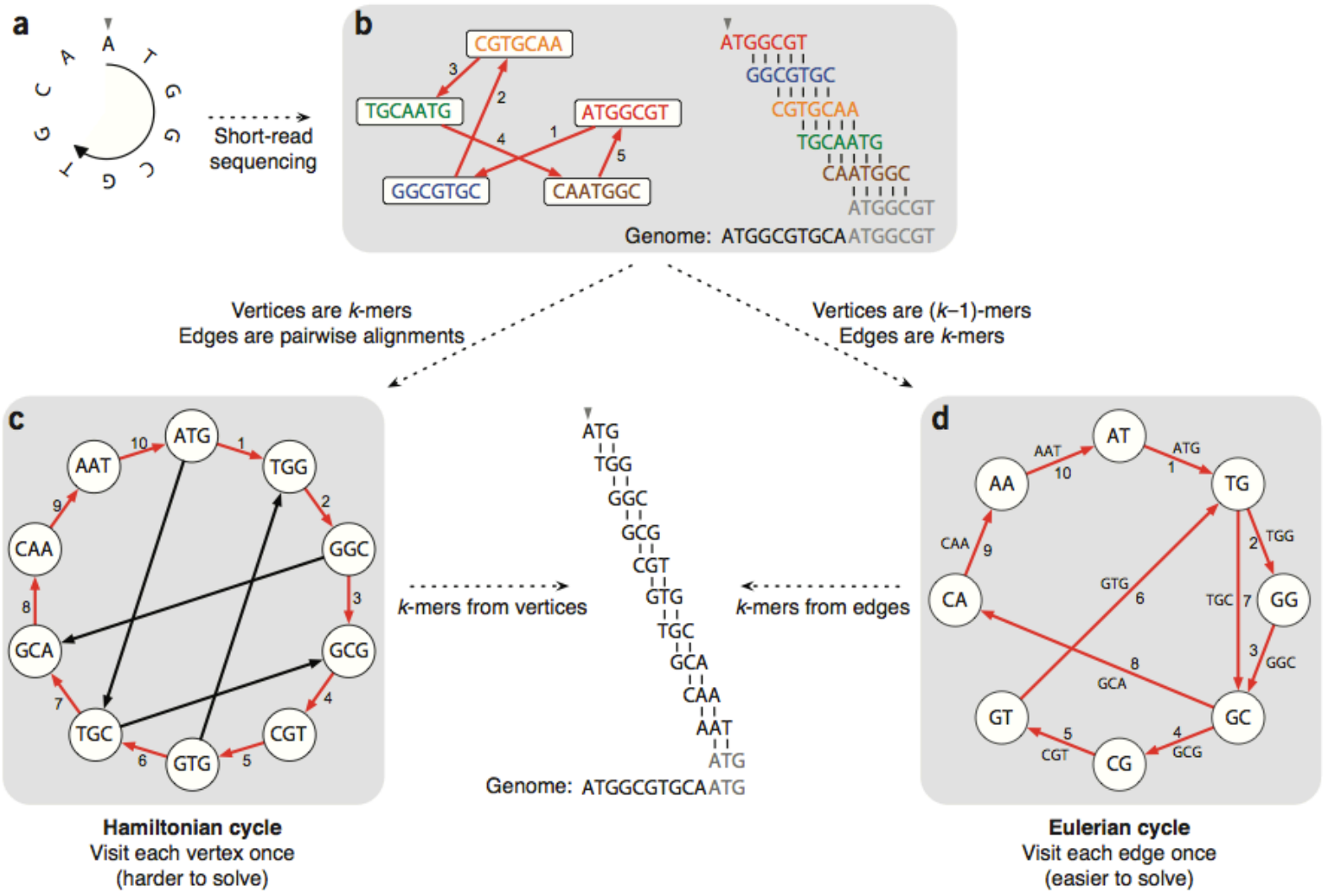


http://illumina.com

Genome Assembly for NGS

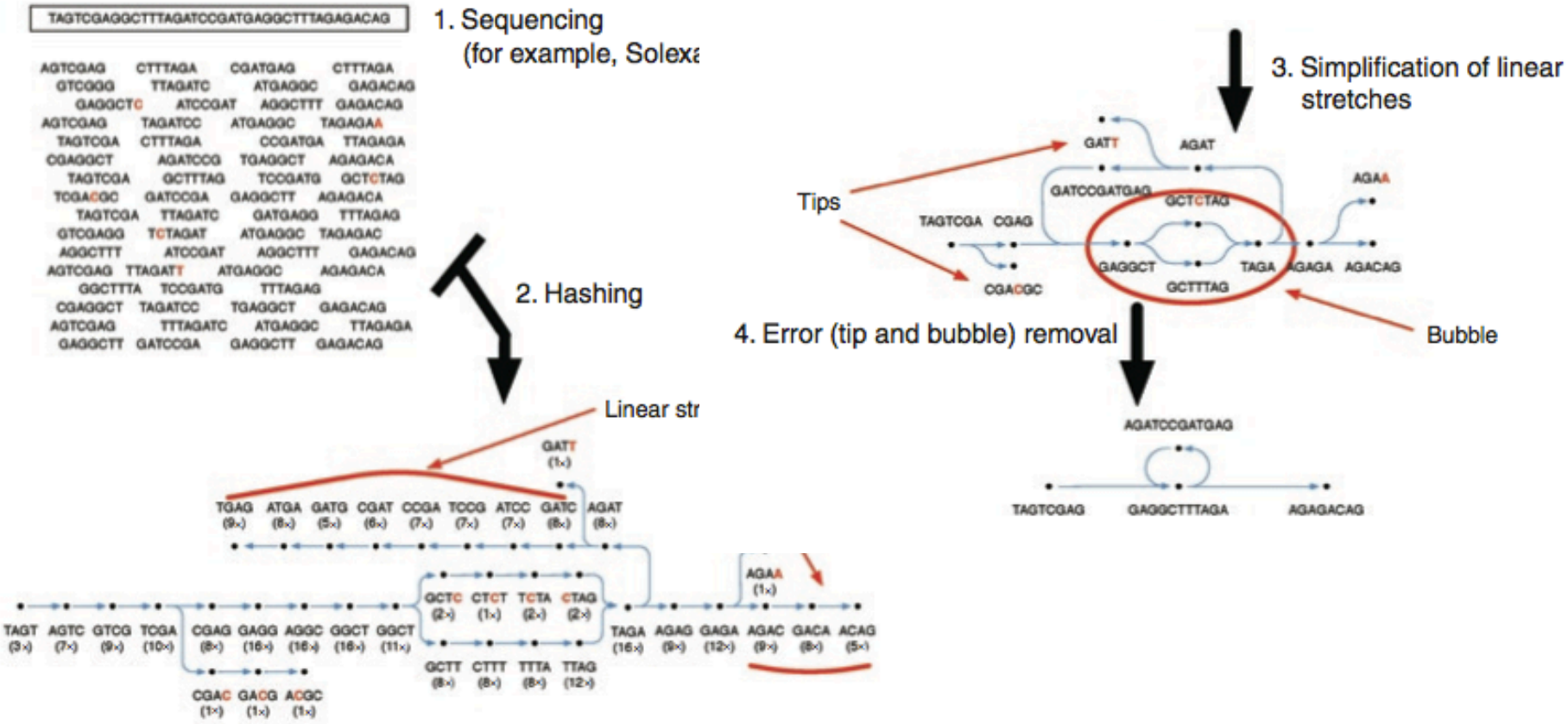
- Computing pair-wise overlaps for OLC assembly too computationally intensive for millions of reads
- Need alternative algorithm

Assembly for NGS: de Bruijn Graphs



Compeau et al (2011) Nat Biotech PMID: 22068540

Assembly for NGS: de Bruijn Graphs



Flicek et al (2009) Nat Methods PMID: 19844229

Genome Assembly for NGS: de Bruijn Graphs

- Advantages
  - No need for pairwise alignment calculation
  - Efficient algorithm exists to find Eulerian (all edges) path
- Disadvantages
  - Sequence information is lost when reads are split into kmers
    - Heuristic: thread reads through graph
  - Algorithm itself not robust to sequencing errors
    - Heuristic: preprocessing and error correction
  - Sensitive to value of k

Genome Assembly for NGS: de Bruijn Graphs

De Bruijn graph short-read assemblers:

- ABySS
  - Allows distributed computing across nodes
- Allpaths-LG
  - Requires mix of short and “jump” reads
- Velvet
  - User can specify range of k values
- SOAPdenovo
  - Built on SOAP OLC assembler

Genome Assembler Performance

- GAGE: 2012 study comparing assembly performance on 4 species
- 3 species (*S. aureus*, *R. sphaeroides* and human) had finished genomes for comparison

Species	*S. aureus*	*R. sphaeroides*	Human Chr14	*B. impatiens*
Size (Mb)	2.90	4.60	88.29	250 (est.)
Read length	101, 37	101	101	124
Fragment size, Library 1	180	180	155	400
Number of reads, Library 1	1,294,104	2,050,868	36,504,800	303,118,594
Fragment size, Library 2	3500	3500	2280–2800	3000–4000
Number of reads, Library 2	3,494,070	2,050,868	22,669,408	129,118,270
Fragment size, Library 3			35 kb	8 kb
Number of reads, Library 3			2,405,064	65,081,280

Salzberg et al. (2012) Genome Res. PMID: 22147368

Genome Assembler Performance

- Assembly with highest N50 (SOAPdenovo) also had highest number of errors
- N50's dropped considerably when contigs were broken at errors

Table 3. Assemblies of *R. sphaeroides* (genome size 4,603,060)

	Contigs				Scaffolds			
Assembler	**Num**	**N50 (kb)**	**Errors**	**N50 corr. (kb)**	**Num**	**N50 (kb)**	**Errors**	**N50 corr. (kb)**
ABySS	1915	5.9	76	4.2	1701	9	3	5
ALLPATHS-LG	204	42.5	49	**34.4**	**34**	**3192**	**0**	**3192**
Bambus2	**177**	93.2	373	12.8	92	2439	2	2419
CABOG	322	20.2	44	17.9	130	66	5	55
MSR-CA	395	22.1	52	19.1	43	2,976	5	2966
SGA	3067	4.5	**12**	2.9	2096	51	**0**	51
SOAPdenovo	204	**131.7**	422	14.3	166	660	3	658
Velvet	583	15.7	43	14.5	178	353	6	270

Salzberg et al. (2012) Genome Res. PMID: 22147368

Genome Assembler Performance

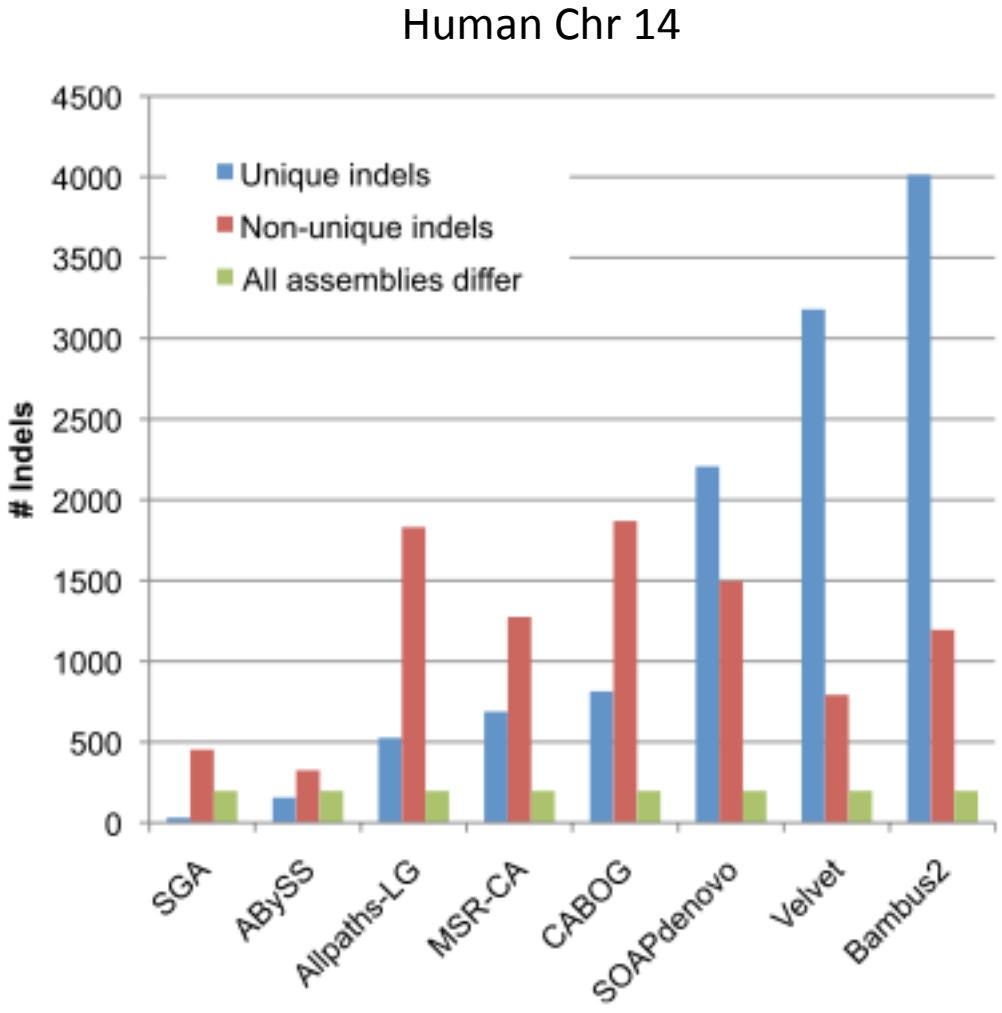
- In human, ALLPATHS-LG had the highest corrected N50, but ABySS had fewer errors
- ALLPATHS-LG outperformed other assemblers, but note that the sequencing read sizes and insert lengths were chosen to be optimal for this assembler

Table 4. Assemblies of human chromosome 14 (ungapped size 88,289,540)

	Contigs				Scaffolds			
Assembler	Num	N50 (kb)	Errors	N50 corr. (kb)	Num	N50 (kb)	Errors	N50 corr. (kb)
ABySS	51,924	2.0	**704**	2.0	51,301	2.1	**9**	2
ALLPATHS-LG	4529	36.5	2760	21.0	**225**	**81,647**	45	**4702**
Bambus2	13,592	5.9	11,943	4.3	1792	324	143	161
CABOG	**3361**	**45.3**	3181	**23.7**	479	393	597	26
MSR-CA	30,103	4.9	5550	4.3	1425	893	1068	94
SGA	56,939	2.7	981	2.7	30,975	83	19	79
SOAPdenovo	22,689	14.7	6424	7.4	13,502	455	268	214
Velvet	45,564	2.3	4910	2.1	3,565	1190	9156	27

Salzberg et al. (2012) Genome Res. PMID: 22147368

Genome Assembler Performance



Salzberg et al. (2012) Genome Res. PMID: 22147368

Other Assembler Comparisons

- Assemblathon 1 (2011)
  - Simulated genome and NGS reads
  - Teams submit assemblies using assembler and parameter settings of choice
- Assemblathon 2 (2012)
  - Real reads provided for three previously unsequenced species (a bird, fish and snake)
  - Assemblies assessed from optical data, fosmid sequences and statistical measures

Salzberg et al. (2012) Genome Res. PMID: 22147368

Agenda

- Genome assembly
  - Traditional (Sanger)
  - Next-Gen Sequencing
  - Hybrid approaches
  - Example: *Arthrobacter* bacteria
- Transcriptome assembly
  - Algorithms and assemblers
  - Example: Zebra finch
- New methods

Agenda

- Genome assembly
  - Traditional (Sanger)
  - Next-Gen Sequencing
  - Hybrid approaches
  - Example: *Arthrobacter* bacteria
- Transcriptome assembly
  - Algorithms and assemblers
  - Example: Zebra finch
- New methods

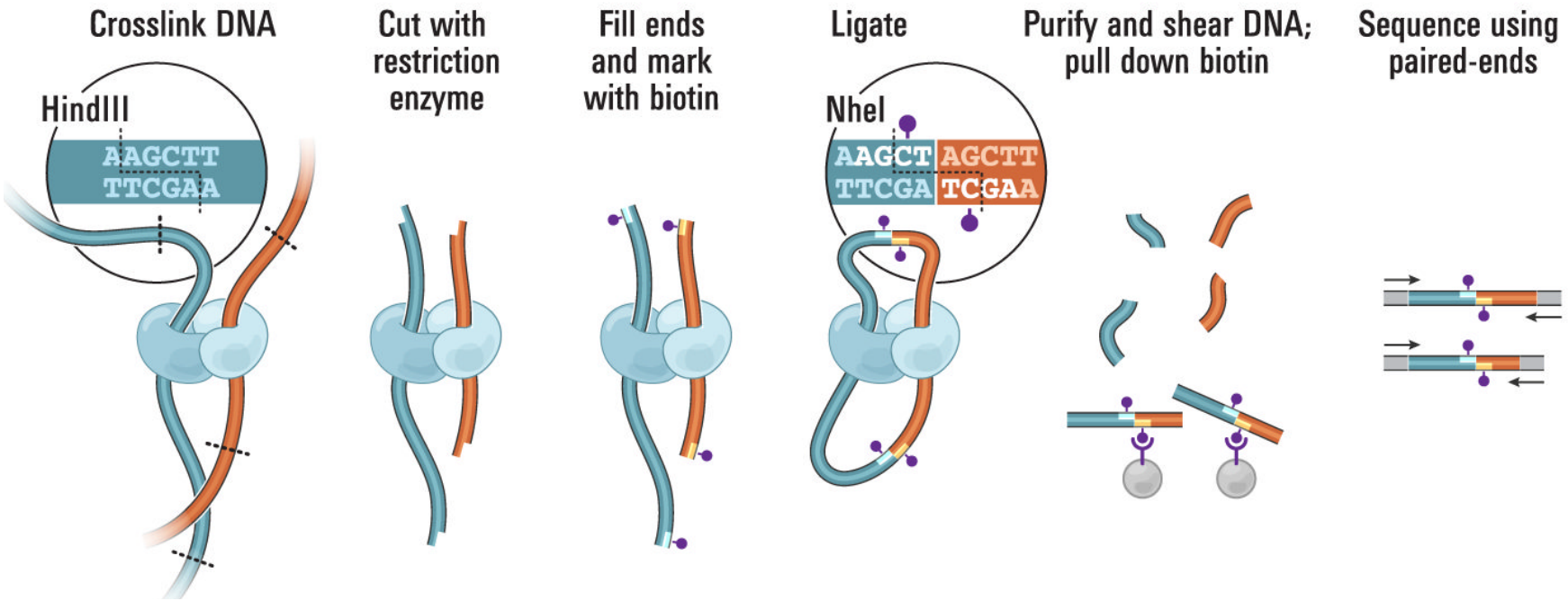
“Hybrid” Assembly

Use additional information in NGS assembly:

- Hi-C / interaction frequency
- Optical mapping
- Genome from closely-related species

Hybrid Assembly: Hi-C + NGS

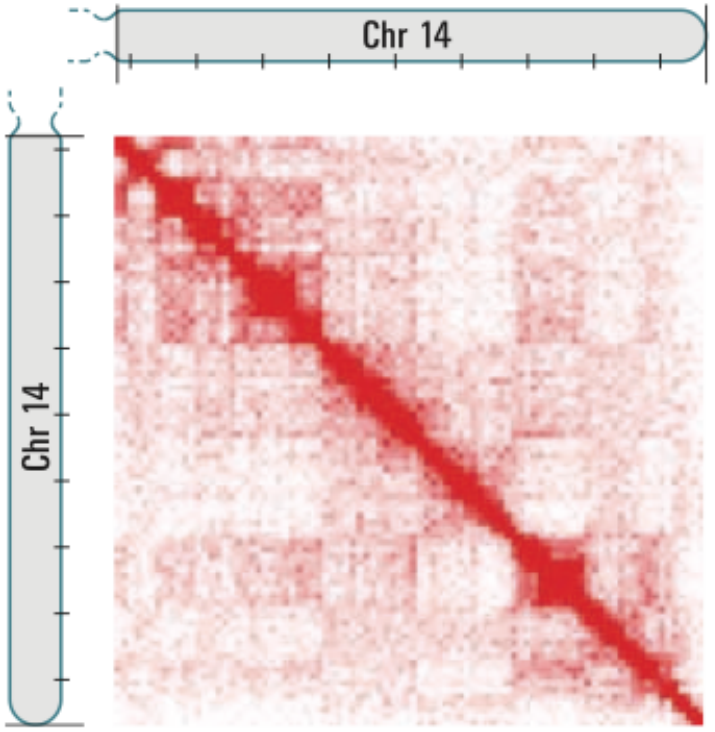
- Hi-C library preparation:



Lieberman-Aiden et al (2009) Science PMID: 19815776

Hybrid Assembly: Hi-C + NGS

- Hi-C read count reflective of physical proximity
- Segments from different chromosomes will have fewer reads than pairs from same chromosome
- Number of reads spanning two contigs defines proximity
- Contigs can then be hierarchically clustered into chromosomes
- Reads then used to order and orient contigs in chromosome

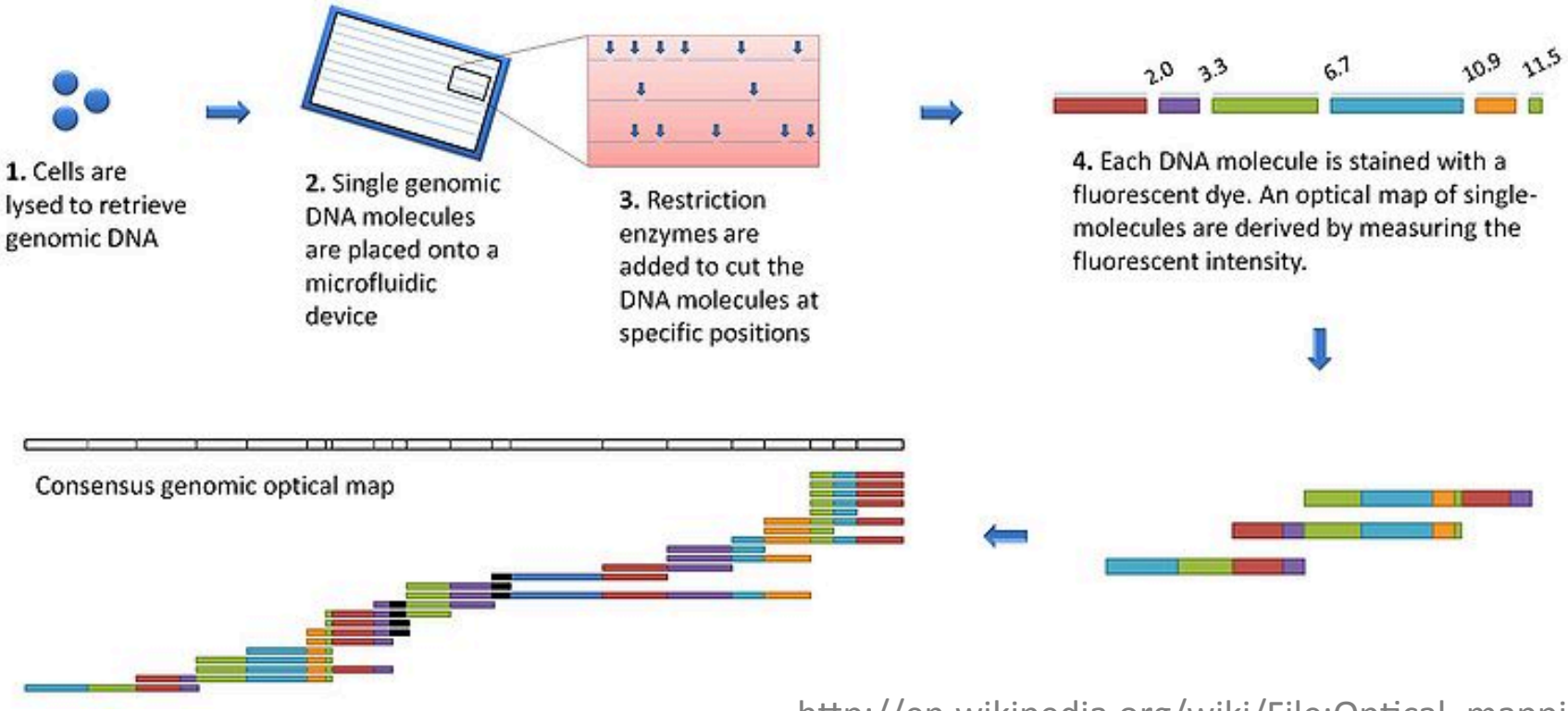


Hi-C read counts along Chr 14

Lieberman-Aiden et al (2009) Science PMID: 19815776

Hybrid Assembly: Optical map + NGS

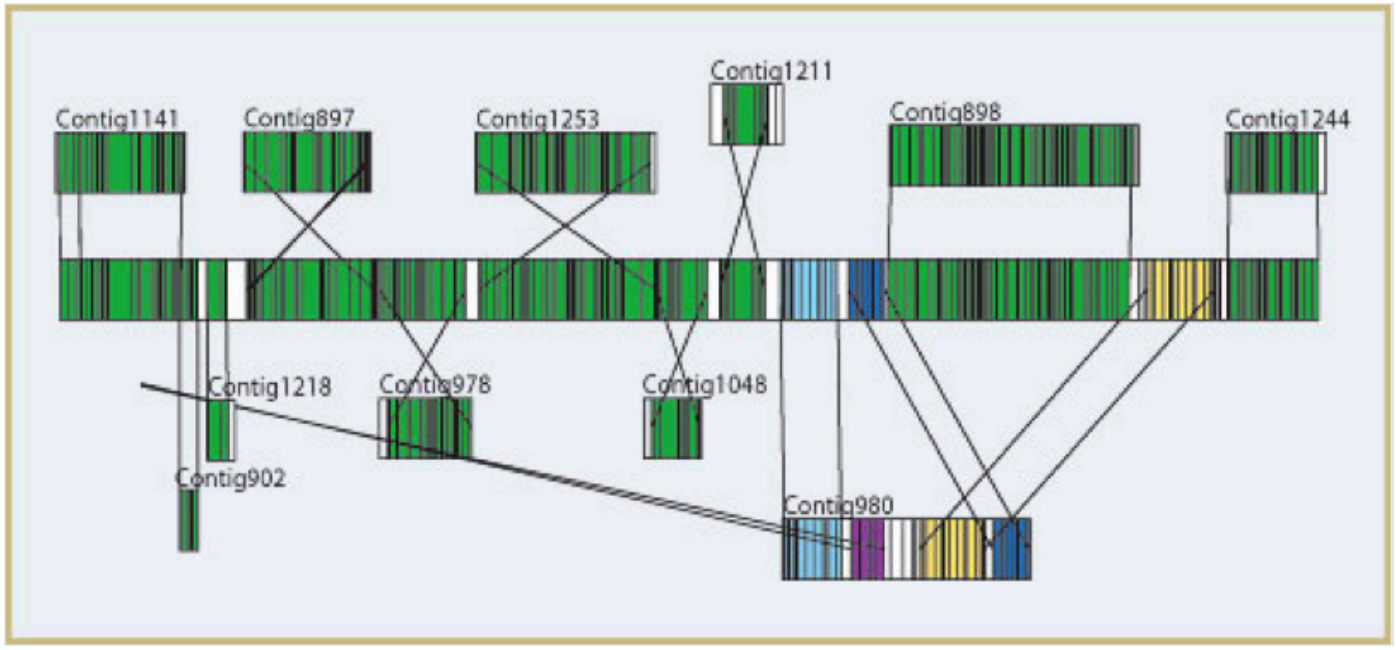
- DNA is stretched along a channel, cut by a restriction enzyme, and dyed
- Order and length of segments is measured



http://en.wikipedia.org/wiki/File:Optical_mapping.jpg

Hybrid Assembly: Optical map + NGS

- Restriction sites on contigs are identified by alignment
- Sites in contigs are matched to restriction sites in optical map:



http://www.amebioscience.com/nordic/whole-genome-seq-assembly

Hybrid Assembly: Use of Reference Genome

- "Comparative assembly"
- Reads are aligned to reference genome
- Given enough contradictory reads, assembly departs from reference and assembles new regions
- Fraction of runtime of *de novo* assemblers
- ABBA
  - Uses amino acid sequences as reference as they are more conserved than nucleotides
- AMOScmp
  - Alignment-layout-consensus

Agenda

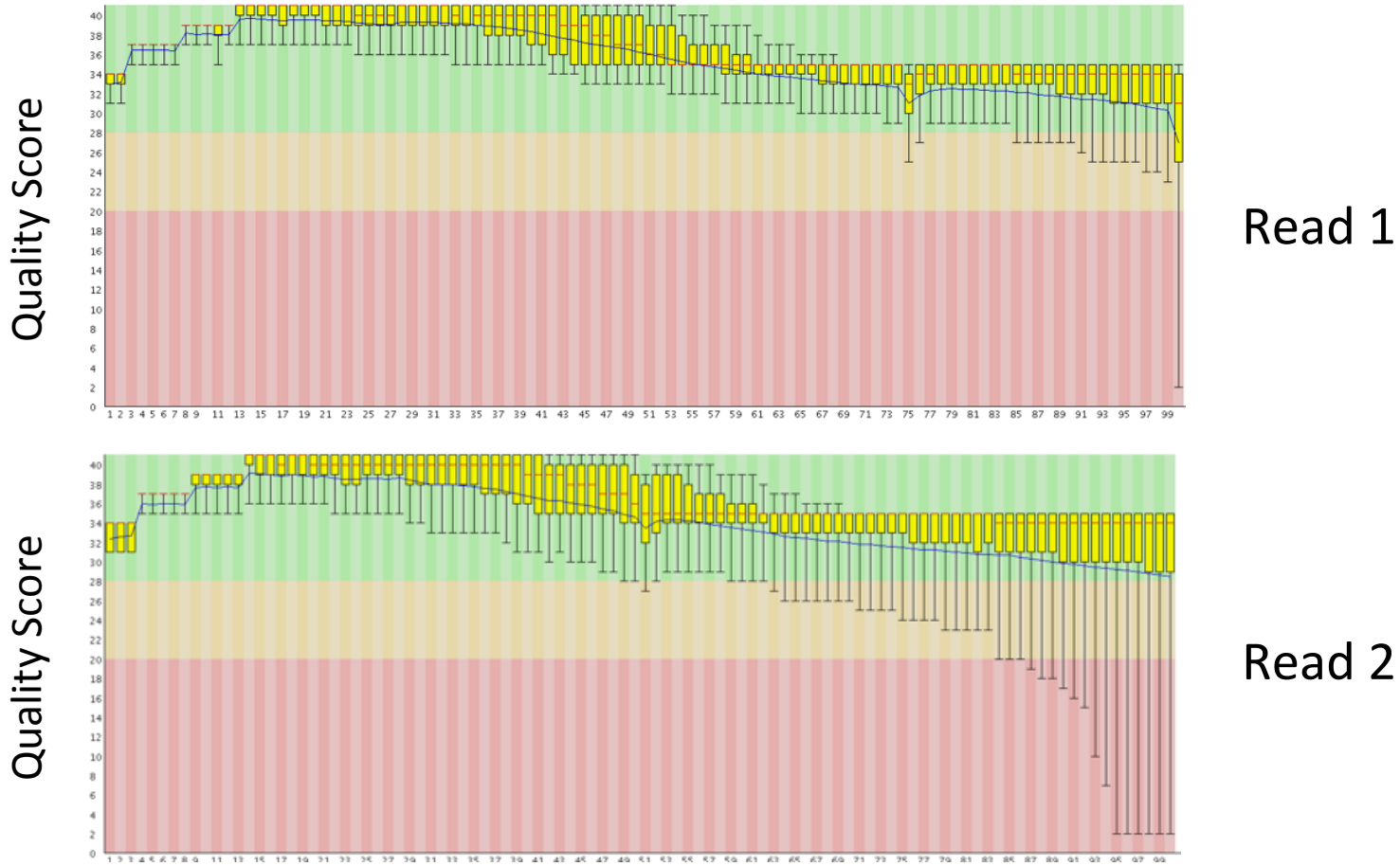
- Genome assembly
  - Traditional (Sanger)
  - Next-Gen Sequencing
  - Hybrid approaches
  - Example: *Arthrobacter* bacteria
- Transcriptome assembly
  - Algorithms and assemblers
  - Example: Zebra finch
- New methods

Agenda

- Genome assembly
  - Traditional (Sanger)
  - Next-Gen Sequencing
  - Hybrid approaches
  - Example: *Arthrobacter* bacteria
- Transcriptome assembly
  - Algorithms and assemblers
  - Example: Zebra finch
- New methods

Example Genome Assembly: *Arthrobacter* bacteria

- 8.5M paired-end, 100 nt reads from Illumina



Arthrobacter Genome Assembly

Preprocessing

Example script to preprocess reads with Trimmomatic:

```
#!/bin/bash
#SBATCH –N 1                    # Number of nodes
#SBATCH –n 6                    # Number of cores
#SBATCH -t 100                  # Runtime in minutes
#SBATCH -p serial_requeue       # Partition to submit to
#SBATCH --mem-per-cpu=1000      # Memory per cpu in MB

module load centos6/Trimmomatic-0.30
java -jar $TRIMMOMATIC/trimmomatic-0.30.jar PE \
        -phred33 –threads 6 R1.fq R2.fq \
        out.R1.paired.fq out.R1.unpaired.fq \
        out.R2.paired.fq out.R2.unpaired.fastq \
ILLUMINACLIP:$TRIMMOMATIC/adapters/TruSeq3-PE.fa:2:30:10 \
SLIDINGWINDOW:40:30 MINLEN:30 LEADING:3 TRAILING:3
```

Save to file `preprocess.sh`, and launch job with:

```
sbatch preprocess.sh
```

Arthrobacter Genome Assembly

Build contigs with Velvet

Run Velvet (try range of k values) to assemble contigs:

```
#!/bin/bash
#SBATCH --nodes=1                       # Number of nodes
#SBATCH --ntasks=1                      # Number of cores
#SBATCH --time=3:00:00                  # Runtime HH:MM:SS
#SBATCH --partition=serial_requeue      # Partition to submit to
#SBATCH –mem=10000                      # Memory in MB

module load centos6/velvet-1.2.10_gcc-4.8.0

velveth velvet 31 -fastq -short allOrphans.fastq \
   -shortPaired –separate R1.paired.fastq R2.paired.fastq
velvetg velvet_31 -cov_cutoff auto
```

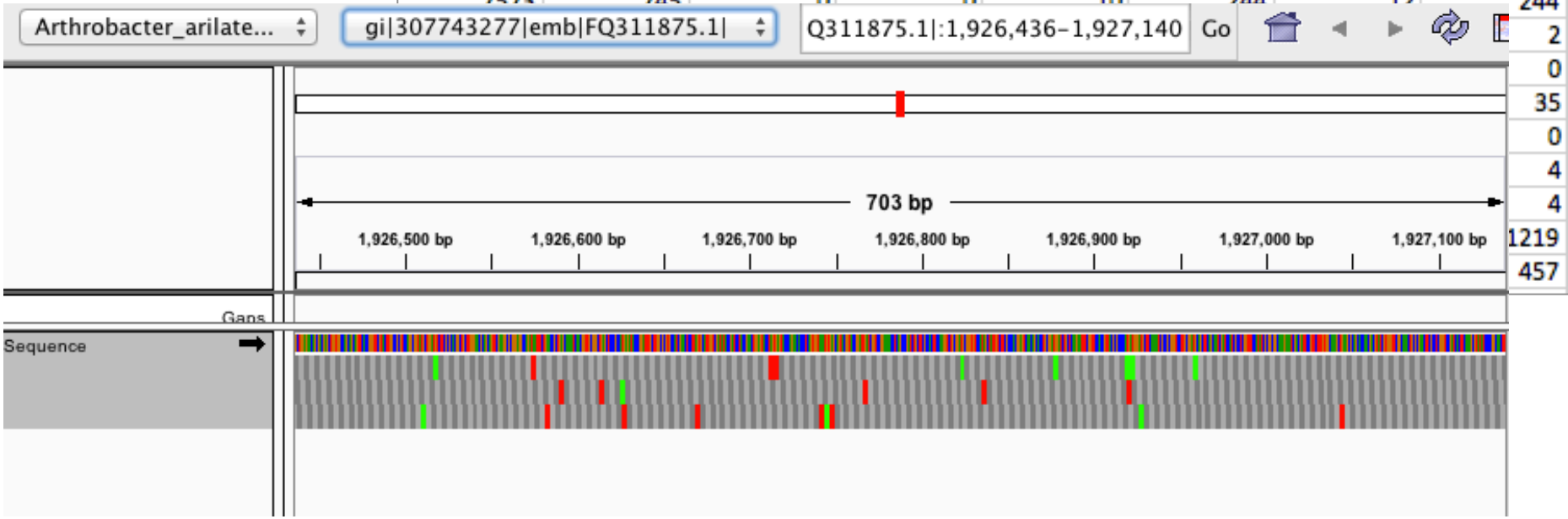
Save to file `runVelvet.sh`, and launch job with:

```
sbatch runVelvet.sh
```

Arthrobacter Genome Assembly

Compare results to assembled genome

match	mis-match	rep. match	N's	Q gap count	Q gap bases	T gap count	T gap bases	
52	2	0	0	1	1	1	1	
34	1	0	0	0	0	0	0	
34	1	0	0	0	0	0	0	
33	1	0	0	0	0	0	0	
41464	123	0	0	2	13	3	1516	
34	0	0	0	0	0	0	0	
34	1	0	0	0	0	0	0	
10858	125	0	0	1	55	2	12134	
7573	745	0	0	10	244	12	244	
							2	
							0	
							35	
							0	
							4	
							4	
							1219	
							457	



Agenda

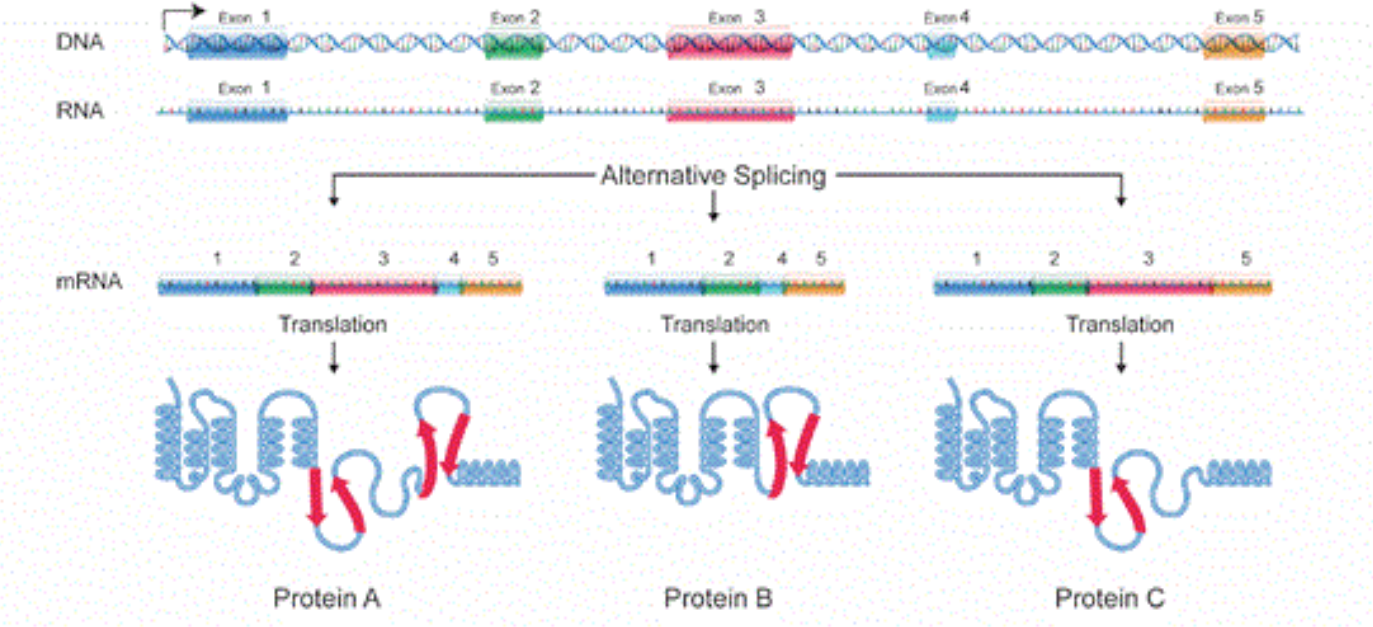
- Genome assembly
  - Traditional (Sanger)
  - Next-Gen Sequencing
  - Hybrid approaches
  - Example: *Arthrobacter* bacteria
- Transcriptome assembly
  - Algorithms and assemblers
  - Example: Zebra finch
- New methods

Agenda

- Genome assembly
  - Traditional (Sanger)
  - Next-Gen Sequencing
  - Hybrid approaches
  - Example: *Arthrobacter* bacteria
- Transcriptome assembly
  - Algorithms and assemblers
  - Example: Zebra finch
- New methods

Transcriptome Assembly

- Variation in read counts can be due to high expression or repeated sequences
- Need to differentiate between isoforms via differences in read counts across gene
- Lowly-expressed genes can be hard to detect

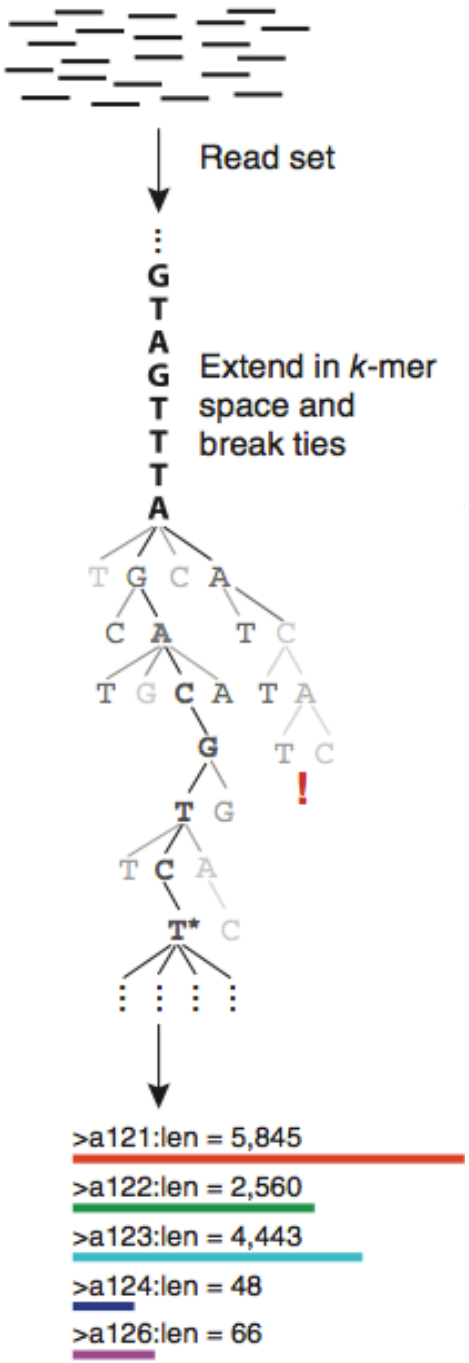


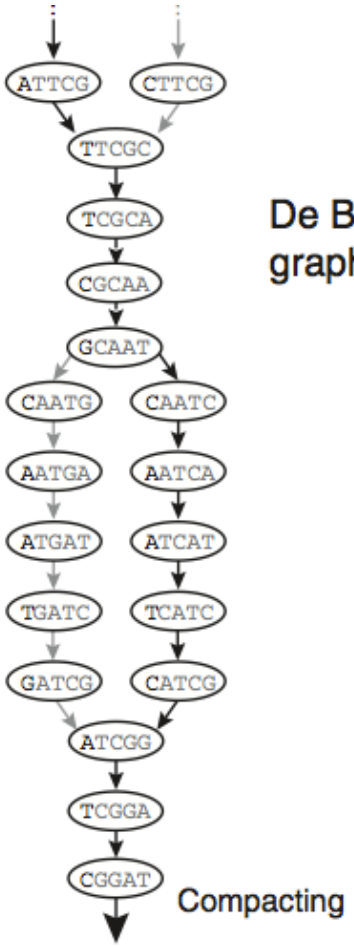
http://www.genomebc.ca/glossary/Alternative%20Splicing%20%28colourful%29_Image.gif

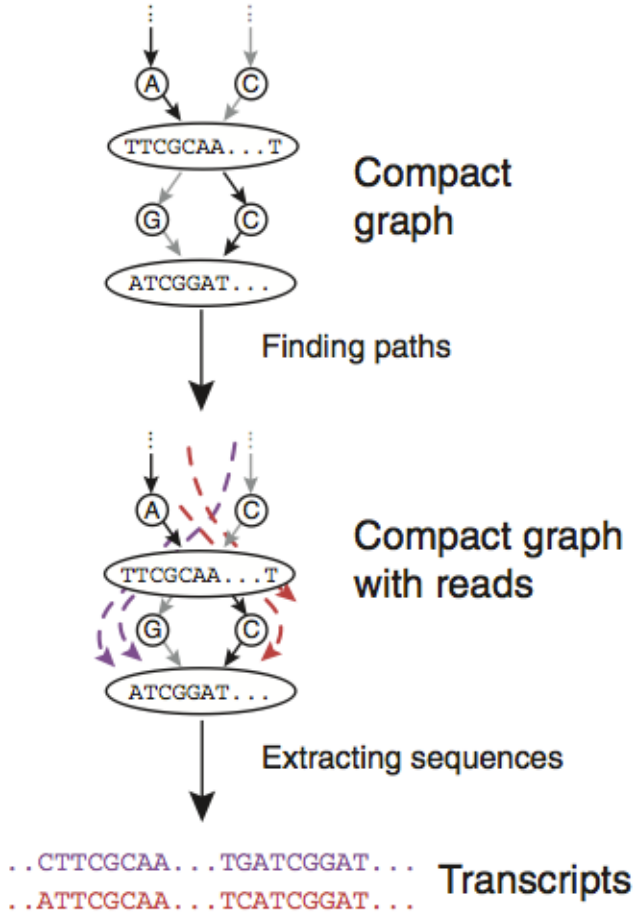
Transcriptome Assembly

- De Novo
  - Assemble transcripts with no knowledge of genome
- Reference-based
  - Assemble transcripts with no knowledge of genome
- Reference-guided
  - Aid assembly with a genome of a closely-related species

Assembly algorithm (Trinity)







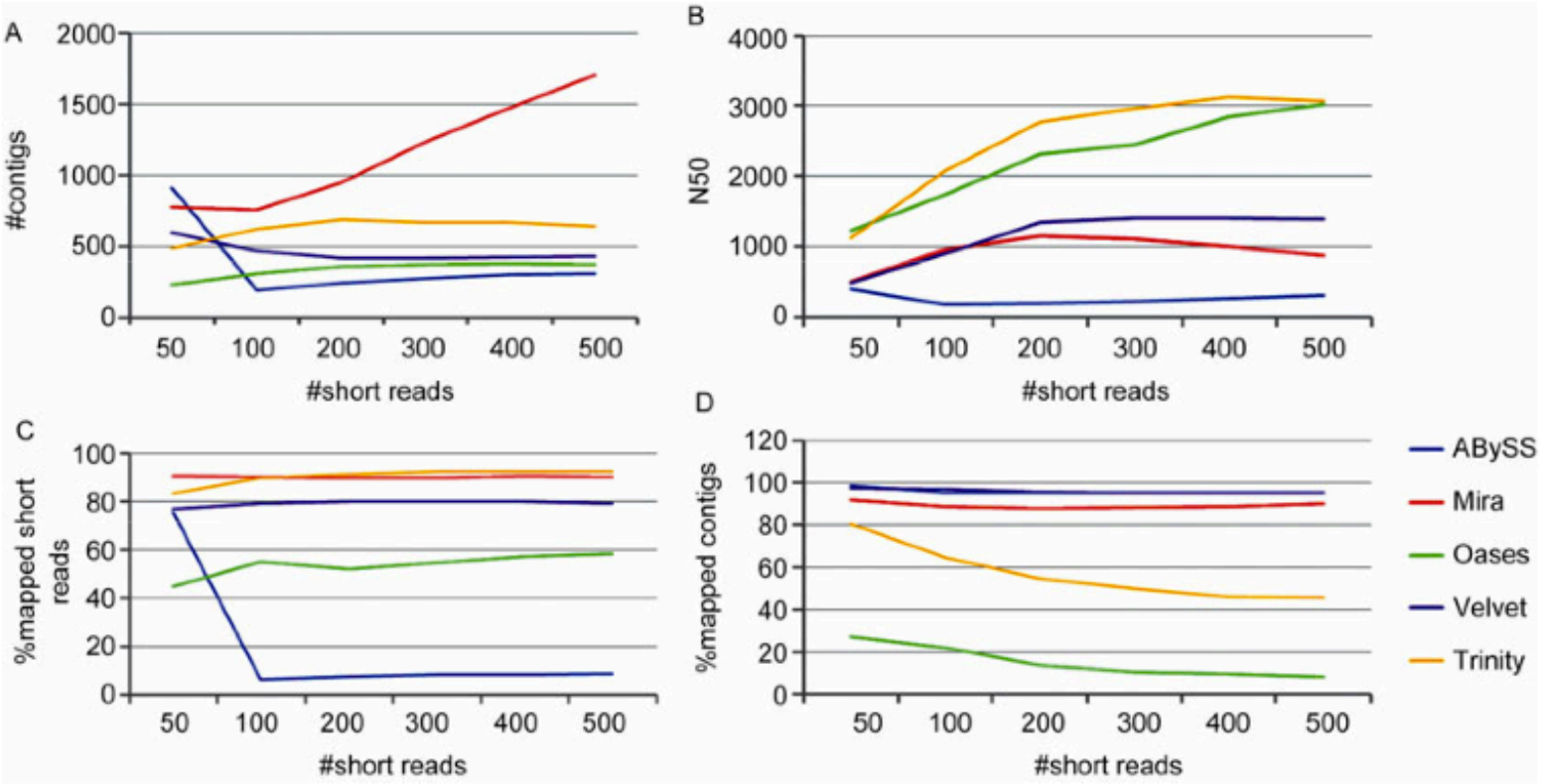
Grabher et al. (2011) Nat Biotech PMID: 21572440

Transcriptome Assembly

- Other assemblers:
  - Oases
    - Built on Velvet genome assembler
  - Trans-ABySS
    - Based on ABySS genome assembler
  - SOAPdenovo-Trans
    - Based on SOAPdenovo assembler

Transcriptome Assembly

- Assembler performance on simulated data:



Clarke et al (2013) Sci China PMID: 23393031

Agenda

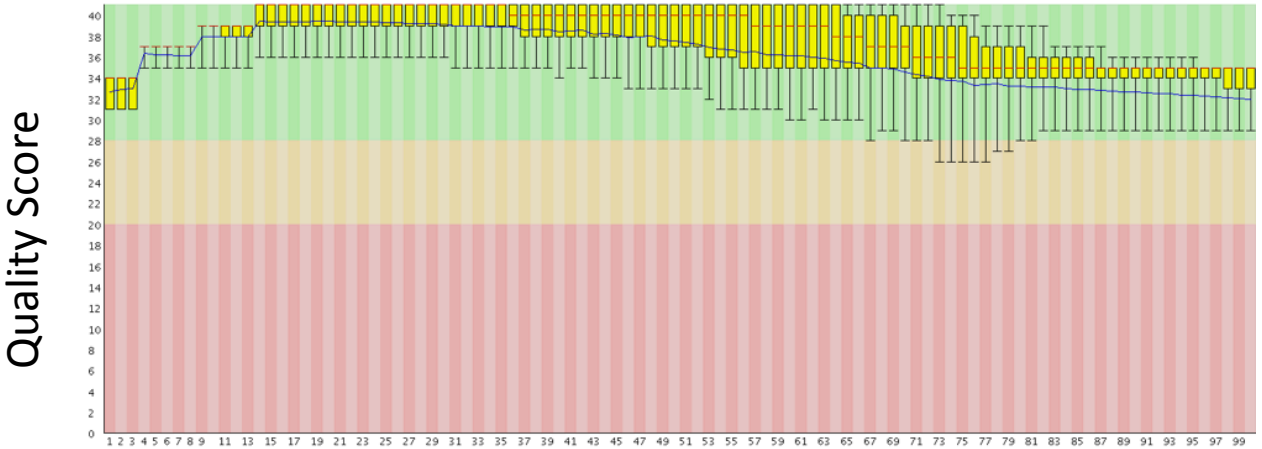
- Genome assembly
  - Traditional (Sanger)
  - Next-Gen Sequencing
  - Hybrid approaches
  - Example: *Arthrobacter* bacteria
- Transcriptome assembly
  - Algorithms and assemblers
  - Example: Zebra finch
- New methods

Agenda

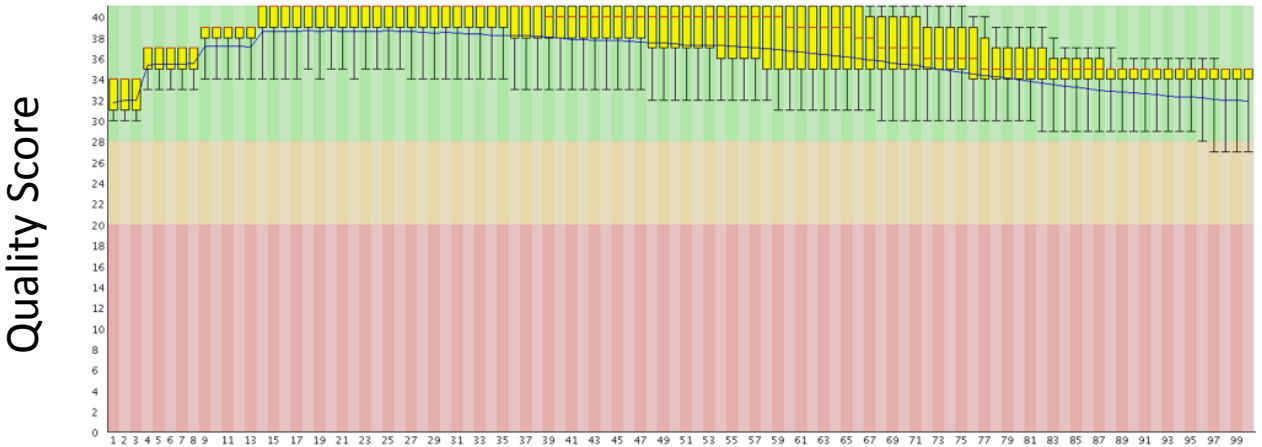
- Genome assembly
  - Traditional (Sanger)
  - Next-Gen Sequencing
  - Hybrid approaches
  - Example: *Arthrobacter* bacteria
- Transcriptome assembly
  - Algorithms and assemblers
  - Example: Zebra finch
- New methods

Example Transcriptome Assembly: Zebra finch

- 20M paired-end, 100nt reads from Illumina
- Preprocess to improve dataset quality



Read 1
(raw)



Read 2
(raw)

Zebra Finch Transcriptome Assembly

Assemble transcripts with Trinity

Script to run Trinity to assemble transcriptome:

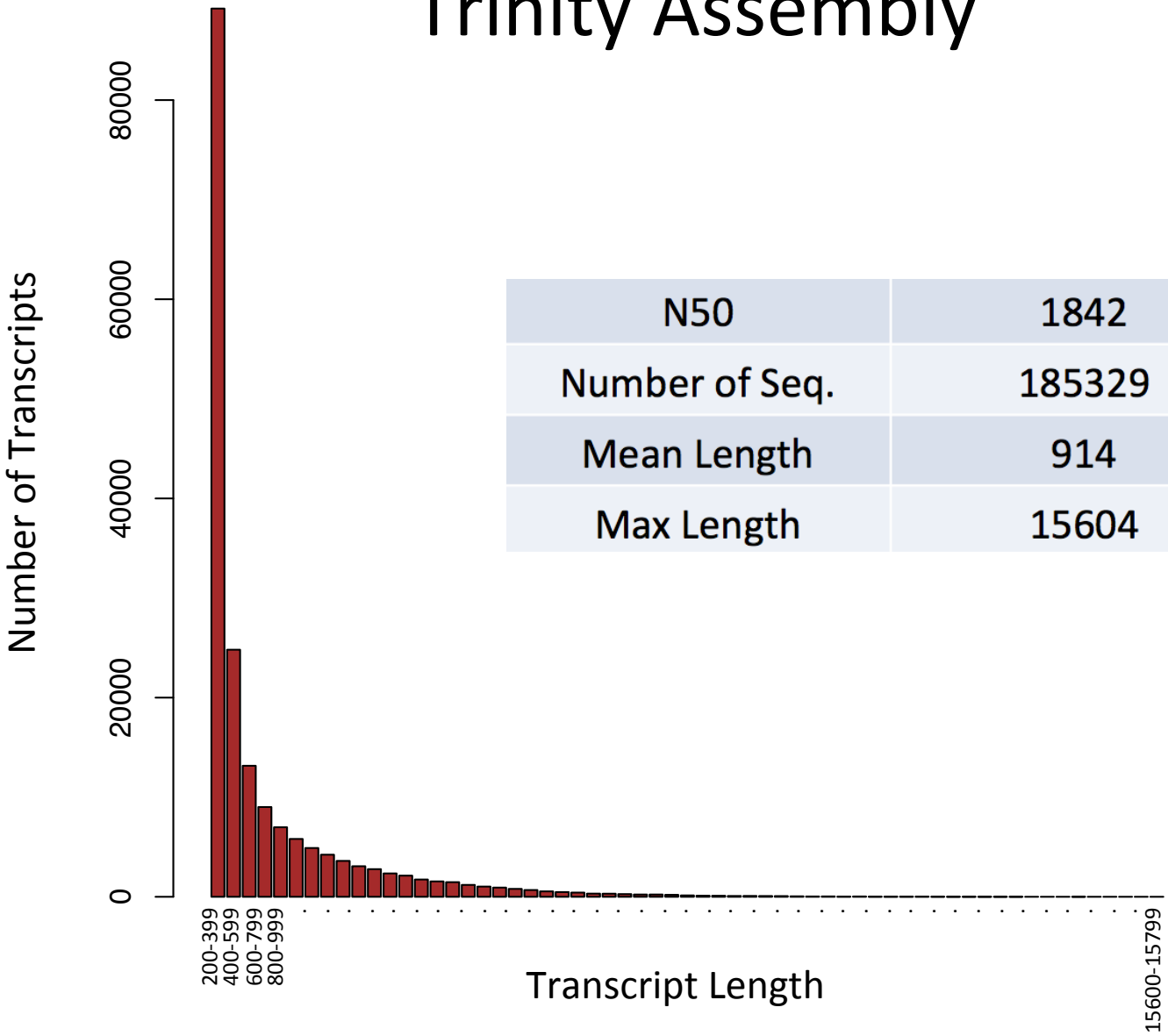
```
#!/bin/bash
#SBATCH --nodes=1                       # Number of nodes
#SBATCH --ntasks=6                      # Number of cores
#SBATCH --time=4-0                      # Runtime in days
#SBATCH --partition=general             # Partition to submit to
#SBATCH --mem=256000                    # Memory in MB

module load centos6/samtools-0.1.19
module load centos6/bowtie2-2.1.0
module load centos6/trinityrnaseq_r20131110
module load centos6/perl-5.8.8

Trinity.pl --seqType fq --JM 20G --CPU 6 \
           --left  R1.paired.and_orphans.fastq \
           --right R2.paired.fastq
```

Zebra Finch Transcriptome Assembly

Trinity Assembly



N50	1842
Number of Seq.	185329
Mean Length	914
Max Length	15604

Zebra Finch Transcriptome Assembly

Compare to reference transcriptome

Script to find overlap with reference genome exons:

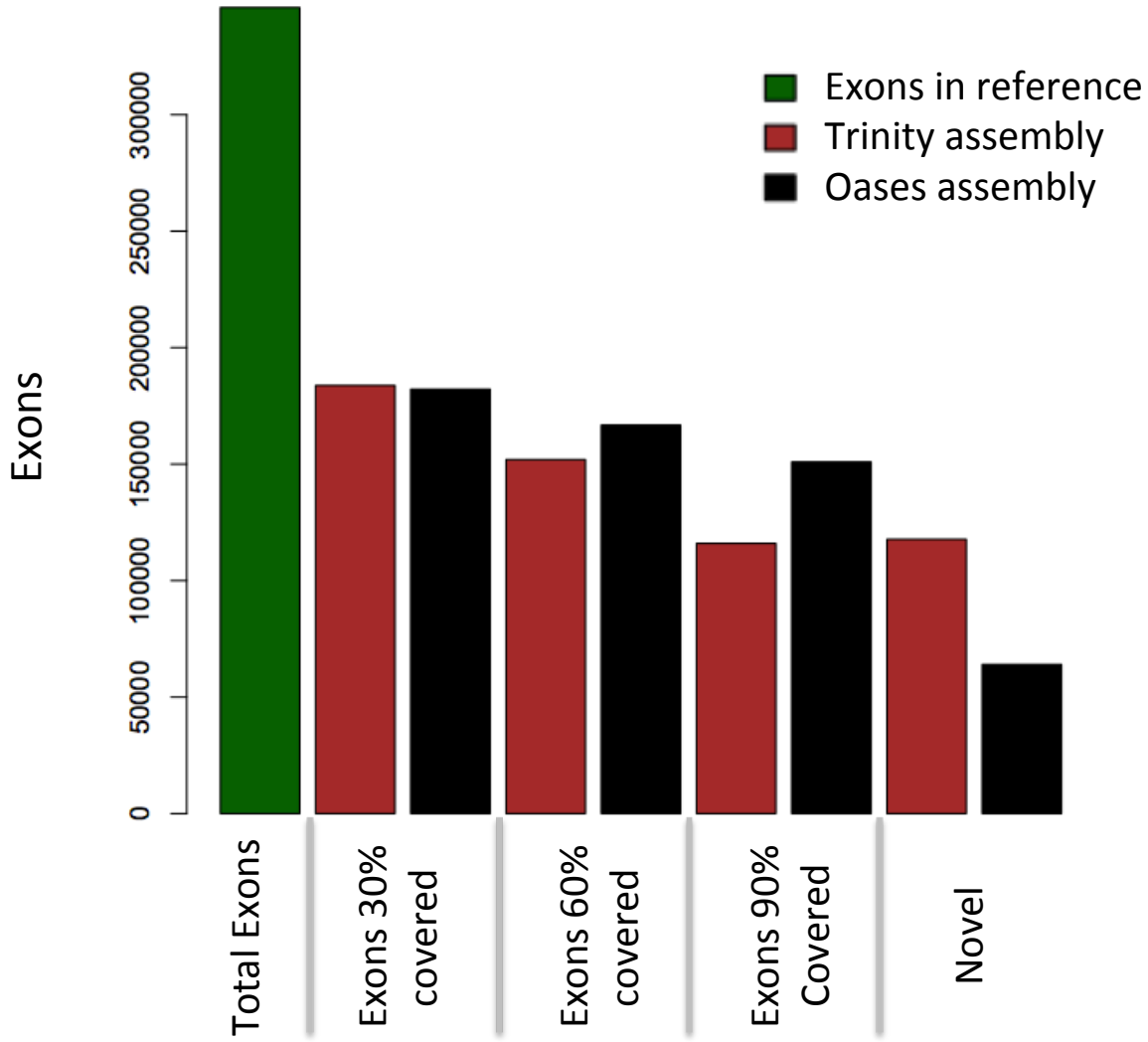
```
#!/bin/bash
#SBATCH --nodes=1                       # Number of nodes
#SBATCH --ntasks=6                      # Number of cores
#SBATCH --time=8:00:00                   # Runtime HH:MM:SS
#SBATCH --partition=serial_requeue      # Partition to submit to
#SBATCH --mem=256000                    # Memory in MB

module load centos6/bedops-2.3.0
module load centos6/bedtools2-2.18.1
module load centos6/samtools-0.1.19
module load centos6/gmap-2013-10-04

gmap -d taeGut3.2.4.73 -f samse Trinity.fasta > aligned.sam
samtools view -b -S $TRINITY_SAM > aligned.bam
bamToBed -i aligned.bam > aligned.bed
intersectBed –a ZF.bed -b aligned.bed -f 0.90 -u > 90perc_overlap.bed
```

Zebra Finch Transcriptome Assembly

Compare to reference transcriptome



Agenda

- Genome assembly
  - Traditional (Sanger)
  - Next-Gen Sequencing
  - Hybrid approaches
  - Example: *Arthrobacter* bacteria
- Transcriptome assembly
  - Algorithms and assemblers
  - Example: Zebra finch
- New methods

Agenda

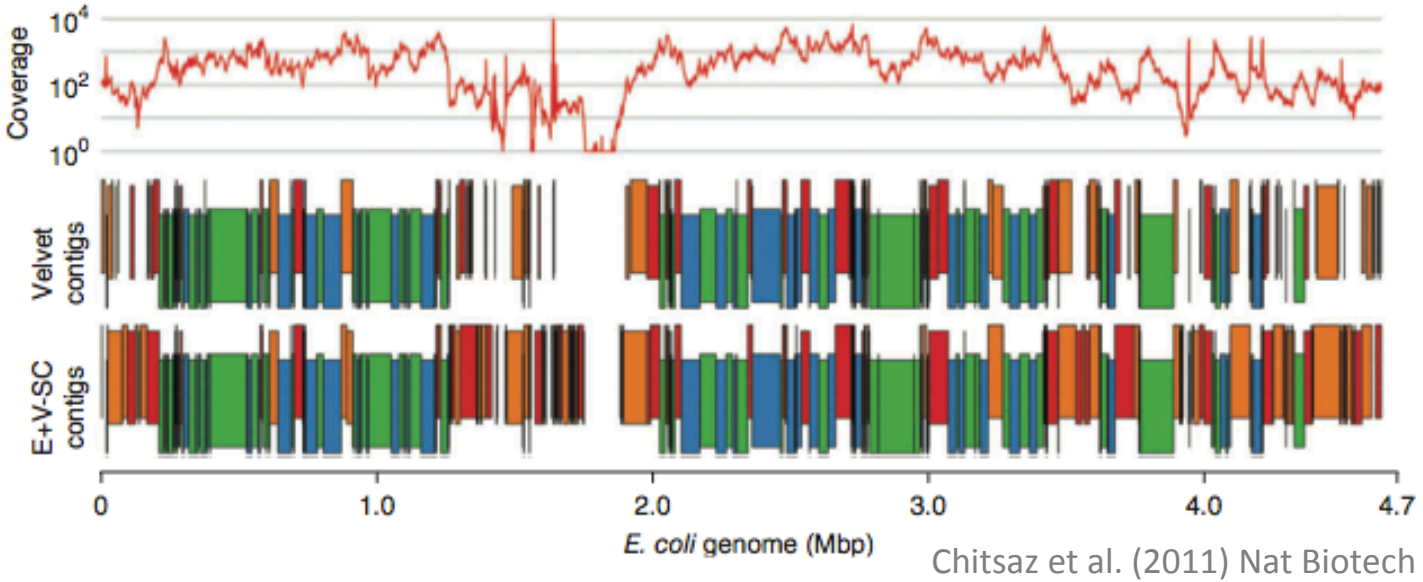
- Genome assembly
  - Traditional (Sanger)
  - Next-Gen Sequencing
  - Hybrid approaches
  - Example: *Arthrobacter* bacteria
- Transcriptome assembly
  - Algorithms and assemblers
  - Example: Zebra finch
- New methods

New Methods

- Single-cell
- Mixed population (metagenomics)
- VLRs: Very Long Reads

New Methods:
Single-cell assembly

- Good for microorganisms that cannot grow in lab
- After extraction, DNA is amplified using multiple displacement amplification (MDA)
- Coverage varies by orders of magnitude
- Traditional assemblers must be altered so they do not filter out low read-count assemblies as spurious



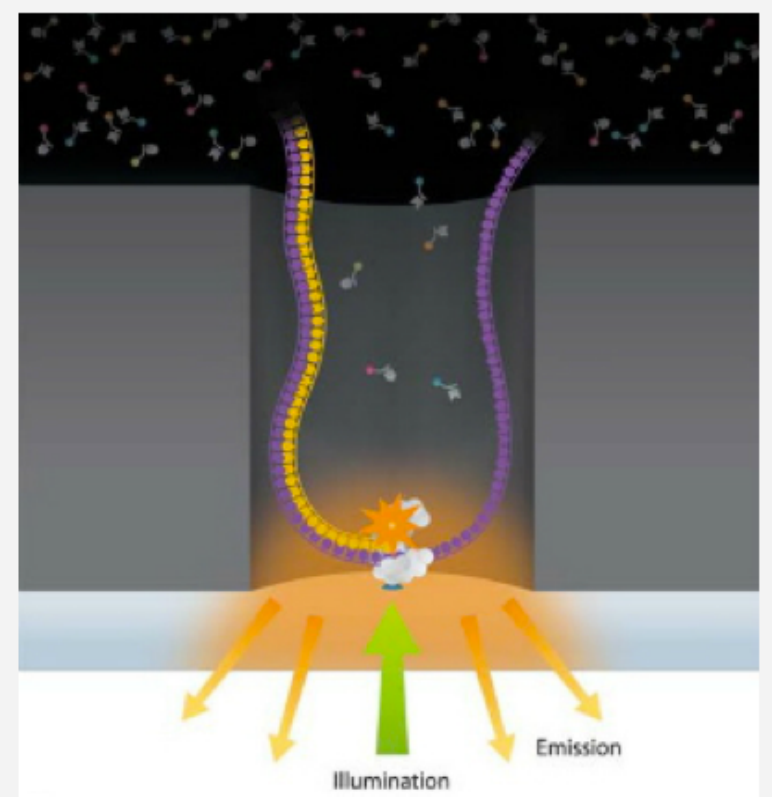
Chitsaz et al. (2011) Nat Biotech PMID: 21926975

New Methods:
Mixed population assembly

- Good for biologically diverse samples taken directly from the environment
- DNA extraction efficiency and kmer frequency is species-dependent
- Validation with essential single-copy genes (e.g., ribosomal genes)
- Validation with cultures mixed *in vitro*

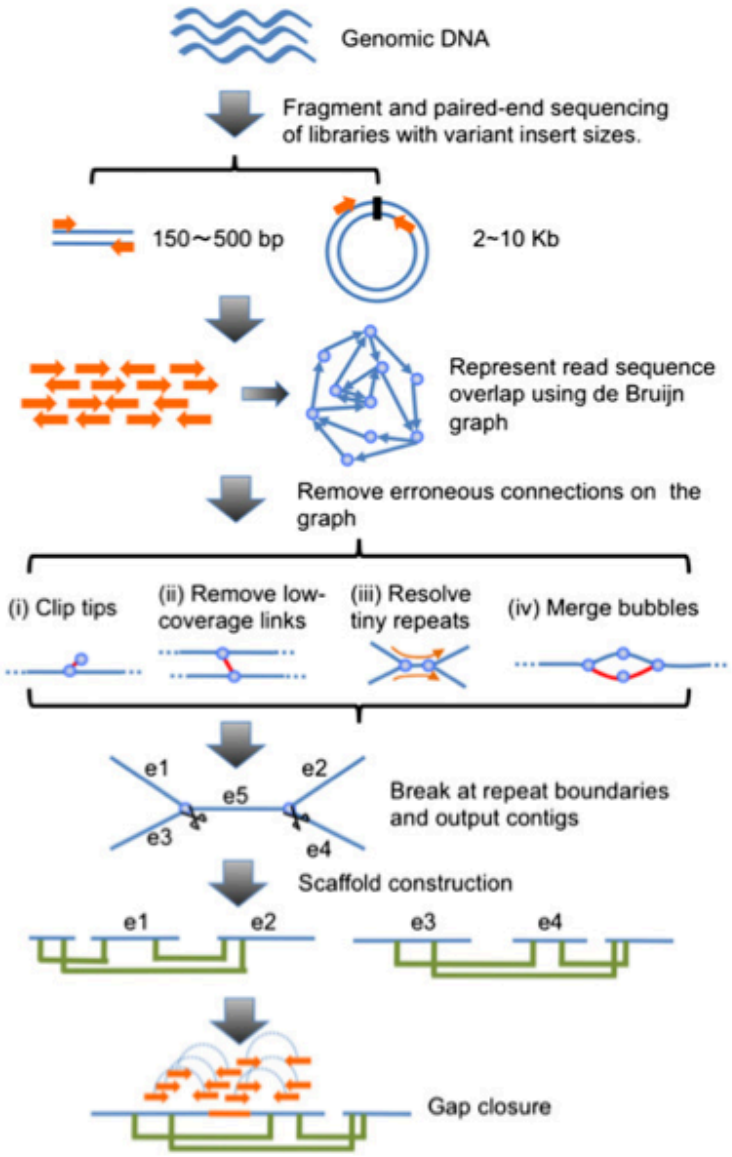
New Methods:
Assembly with Very Long Reads

- PacBio SMRT mean read length:
  - ~2500 bp
- Error rate: ~15%
- Use short NGS reads to "correct" SMRT reads before assembly
- Long reads facilitate contig placement



Hybrid Assembly: Long + Short reads

- Use short reads to build contigs
- Use paired-end reads of long fragments (2-10 kb) to join contigs
- Assemblers:
  - SOAPdenovo
  - AllPaths
  - Velvet



Li et al (2010) Genome Res PMID: 20019144