Introduction to NGS genome/transcriptome assembly on HPRC

Shichen Wang, PhD
Bioinformatics Scientist
Genomics and Bioinformatics, AgriLife research
High Performance Research Computing

DIVISION OF RESEARCH
TEXAS A&M UNIVERSITY

Hprc.tamu.edu

TEXAS A&M HIGH PERFORMANCE RESEARCH COMPUTING

Home | User Services | Resources | Research | Policies | Events | About

Quick Links

- New User Information
- Accounts
  - Apply for Accounts
  - Manage Accounts
- User Consulting
- Training
- Documentation
- Software
- FAQ

User Guides

- Ada
- Terra
- Curie
- Portal
- Galaxy

Cluster Status

Effect of methylation on local mechanics and hydration structure of DNA by *Xiaojing Teng and Wonmuk Hwang*, Department of Biomedical Engineering, Texas A&M.

News

Events

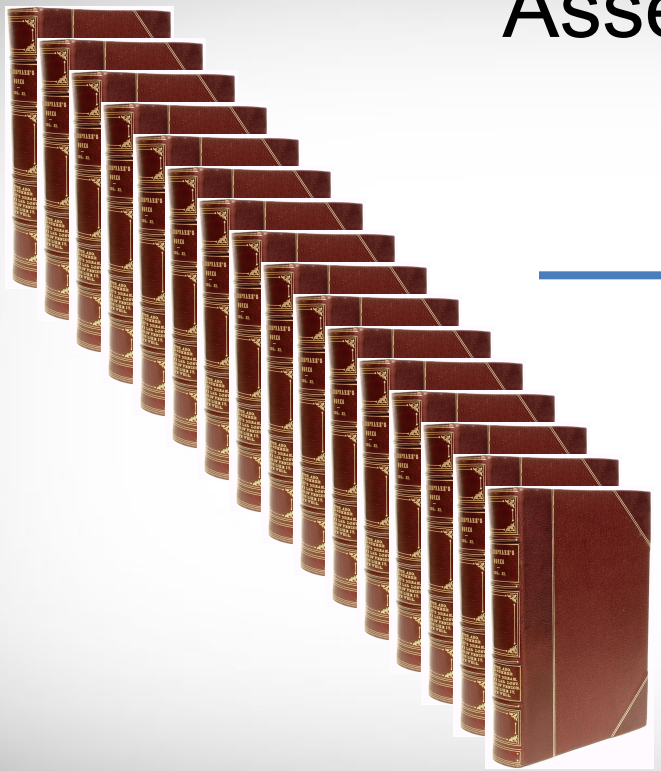
Genome Assembly

What is a genome assembly?

The **genome assembly** is simply the **genome sequence** produced after chromosomes have been fragmented, those fragments have been sequenced, and the resulting sequences have been put back together.

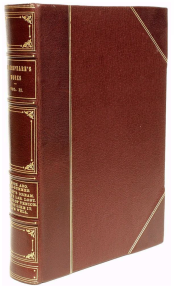
Source: ensembl.org

Assembly is difficult



Shredded

Assembler



The Human Genome Project was reported to have cost **$3 billion**, from 1990-2003.

Assembly is more feasible now than before

With the advances in sequencing technologies (NGS, long read sequencing), genome assembly has become much more feasible, and affordable, to assemble and annotate the genomic sequence of most organisms, including large eukaryote genomes;

Assembly and annotation of small genomes e.g., bacterias and fungi, can often be performed with fairly small resources and a limited time commitment, but eukaryotic genome projects often take months or even years to finish, especially when no reference genomes can be used for these tasks.

The properties of the genome you study

1. Genome size
To assemble a genome, a certain amount of sequences (also called reads) is needed. For example, for Illumina sequencing (see Illumina Genome Assembly below), a number of >60x sequence depth is often mentioned.

2. Repeats
Amount and distribution of repeats in a genome hugely influences the genome assembly results, simply because reads from these different repeats are very similar, and the assembly tools cannot distinguish between them. This can lead to mis-assemblies.

To resolve the assembly of repeats, reads need to be long enough to also include the unique sequences flanking the repeats.

The properties of the genome you study

3. Heterozygosity
Highly heterozygous genomes can lead to more fragmented assemblies, or create doubt about the homology of the contigs. It is recommended to sequence inbred individuals, if possible

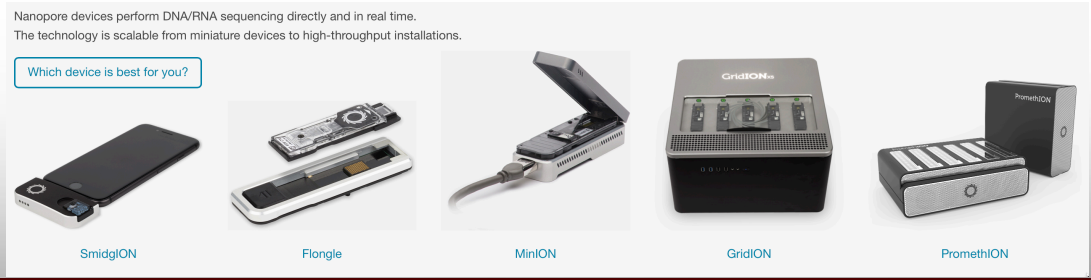
4. Ploidy level
Diploid tissues, which will be the case for most animals and plants, is fine and usually manageable, while tetraploidy and above has the potential to greatly increase the number of present alleles, which likely will result in a more fragmented assembly (see heterozygosity above). Diploid-aware assemblers using long reads can help, but keep in mind that correct assembly of diploid genomes might require higher coverage.

5. GC-content
Extremely low or extremely high GC-content in a genomic region is known to cause a problem for Illumina sequencing, resulting in low or no coverage in those regions. This can be compensated by an increased coverage, or the use of a sequencing technology that does not exhibit that bias (i.e., PacBio or Nanopore).

Short reads VS long reads

Short read sequencing platform
- Illumina NovaSeq
- High throughput, high accuracy

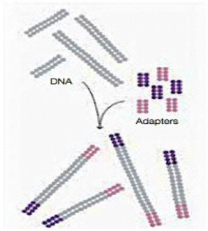
Long read sequencing platform
- Pacbio Sequel II; Oxford nanopore
- Generate long reads (>30Kb), relatively low accuracy but could achieve high accuracy with consensus building or error correction

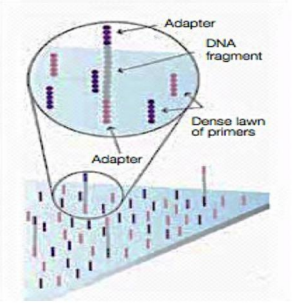


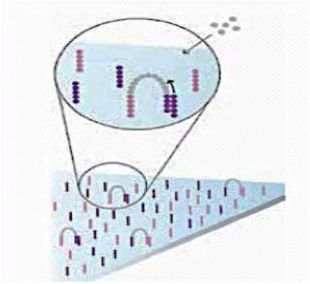
Illumina next-generation sequencing

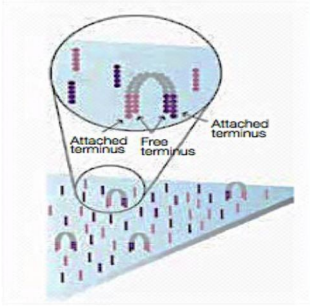
Sequencing by Synthesis (SBS) Technology

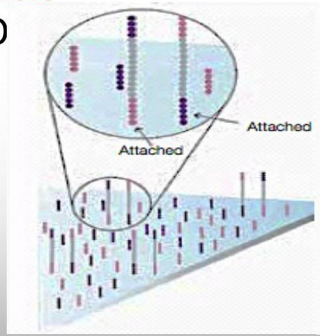
- Randomly shearing DNA
- Attaching DNA fragments to the flowcell surface
- Cluster generation, “Bridge Amplification”
- Adding four labelled *reversible terminators*, primers, and D polymerase
- Determining the attached nucleotide, based on the emitted fluorescence

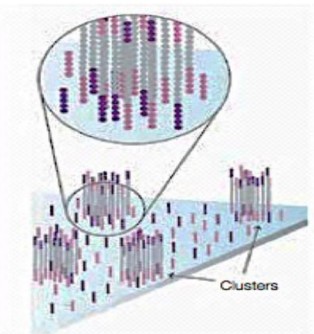


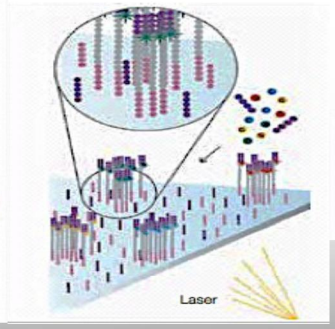


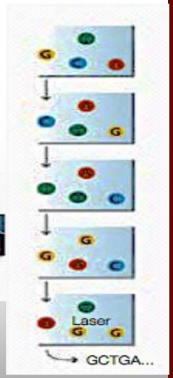






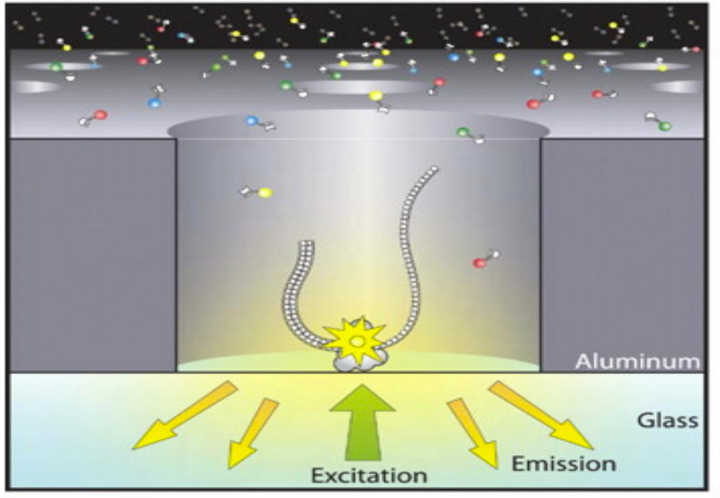


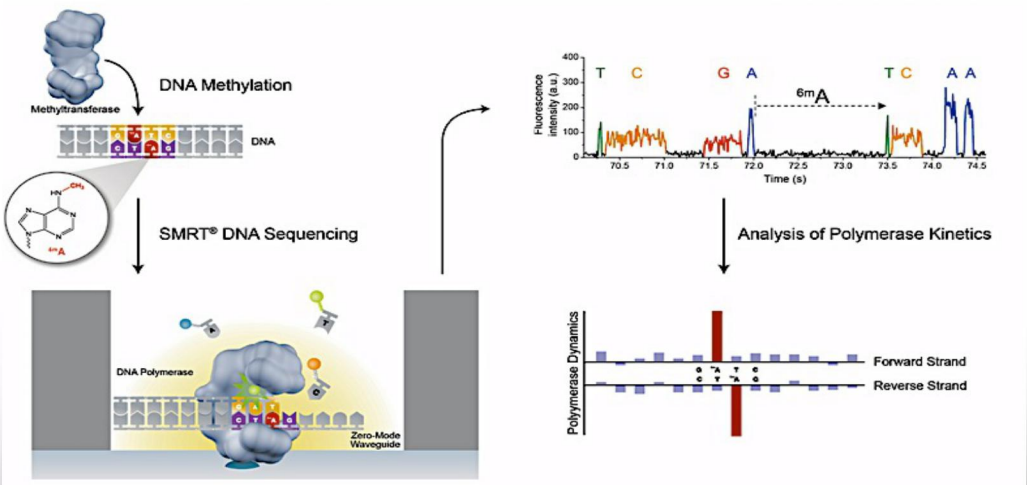




Long reads - PacBio

- Single Molecule Real Time Sequencing (SMRT) Methodology
- Fluorescent dyes
- Zero Mode Waveguide (ZMW)
- DNA polymerase is immobilized at the bottom of a ZMW

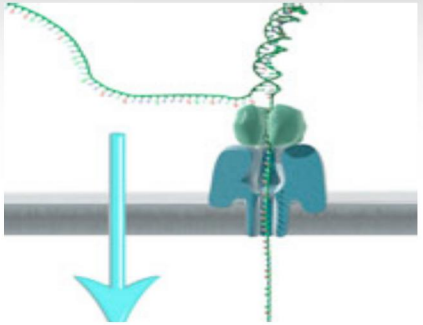


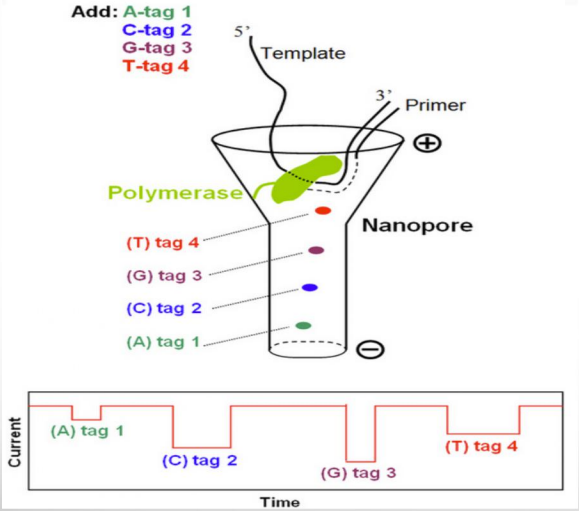


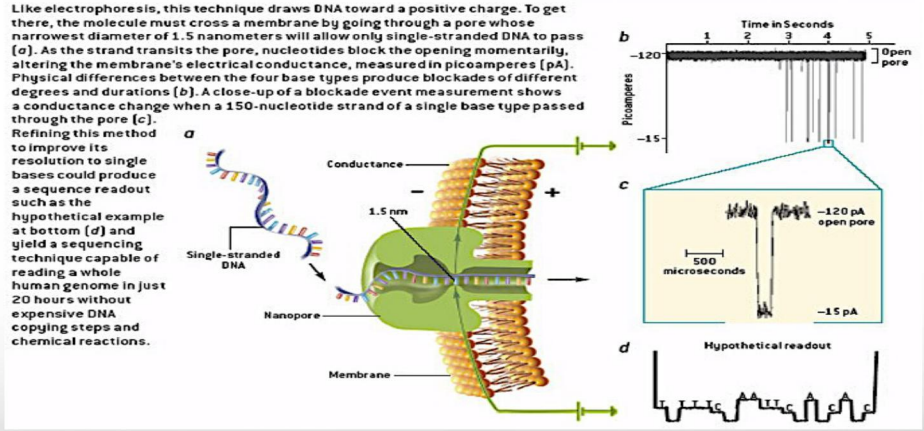
http://www.nature.com/scientificamerican/journal/v294/n1/full/scientificamerican0106-46.html
http://science.sciencemag.org/content/323/5910/133.full

Long reads – Oxford NanoPore

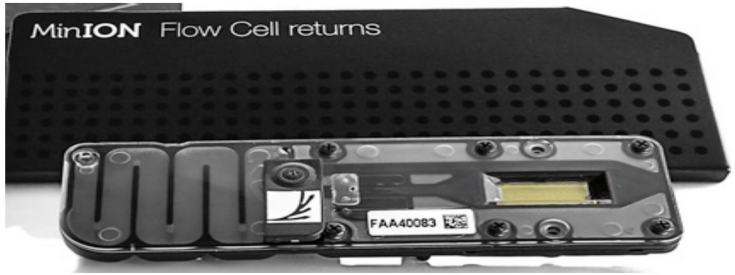
- Oxford Nanopore Technologies
- Nanopore: a small hole (nanometer)
  - used to identify DNA sequence, passing through nanopore
- Single DNA molecule is sequenced







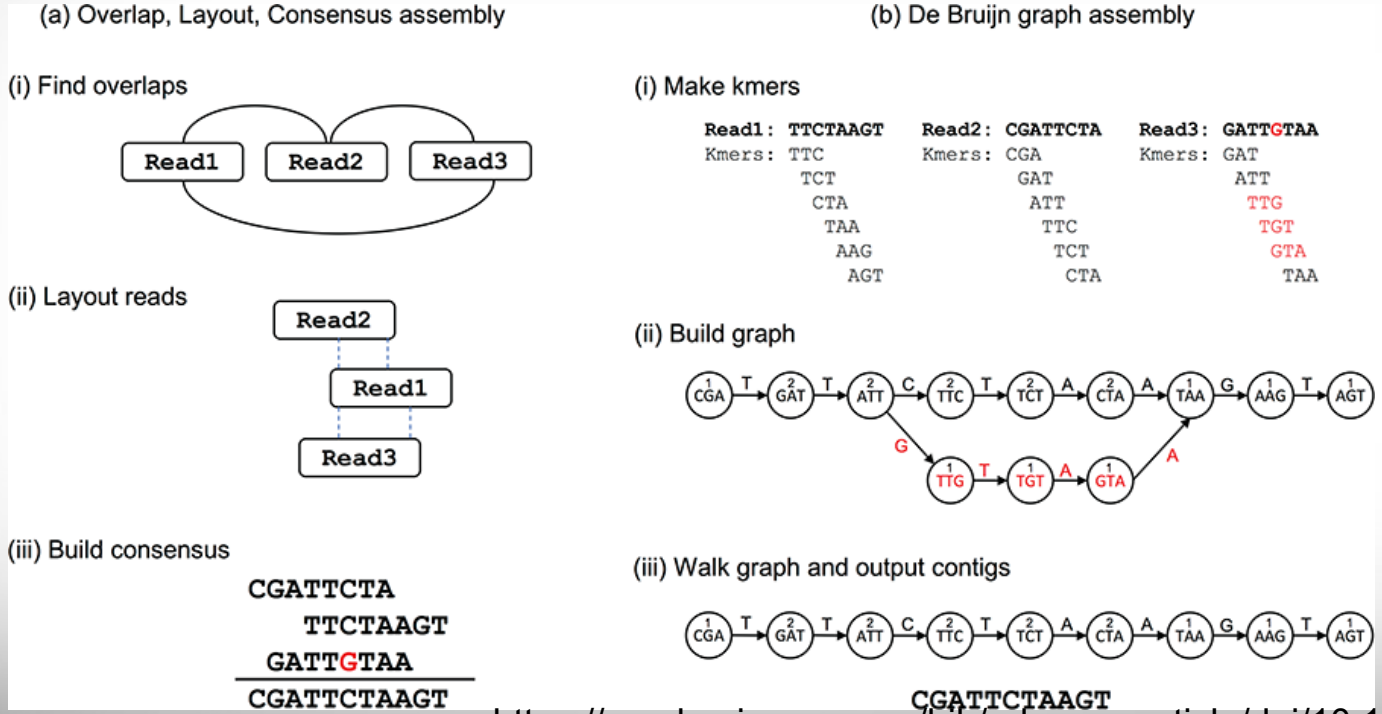
http://www.nature.com/scientificamerican/journal/v294/n1/full/scientificamerican0106-46.html
http://www.kurzweilai.net/single-molecule-electronic-dna-sequencing#!prettyPhoto



TAMU holds patent
Dr. Higgin Bailey

TEXAS A&M AGRILIFE RESEARCH

Genome assembly algorithms



https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbz020/5363831

Genome assembly algorithms for **short** reads

1. Overlap/Layout/Consussus (OLC)

Celera assembler, CAP and Arachne et al.

2. The de Brujin Graph (DBG) methods

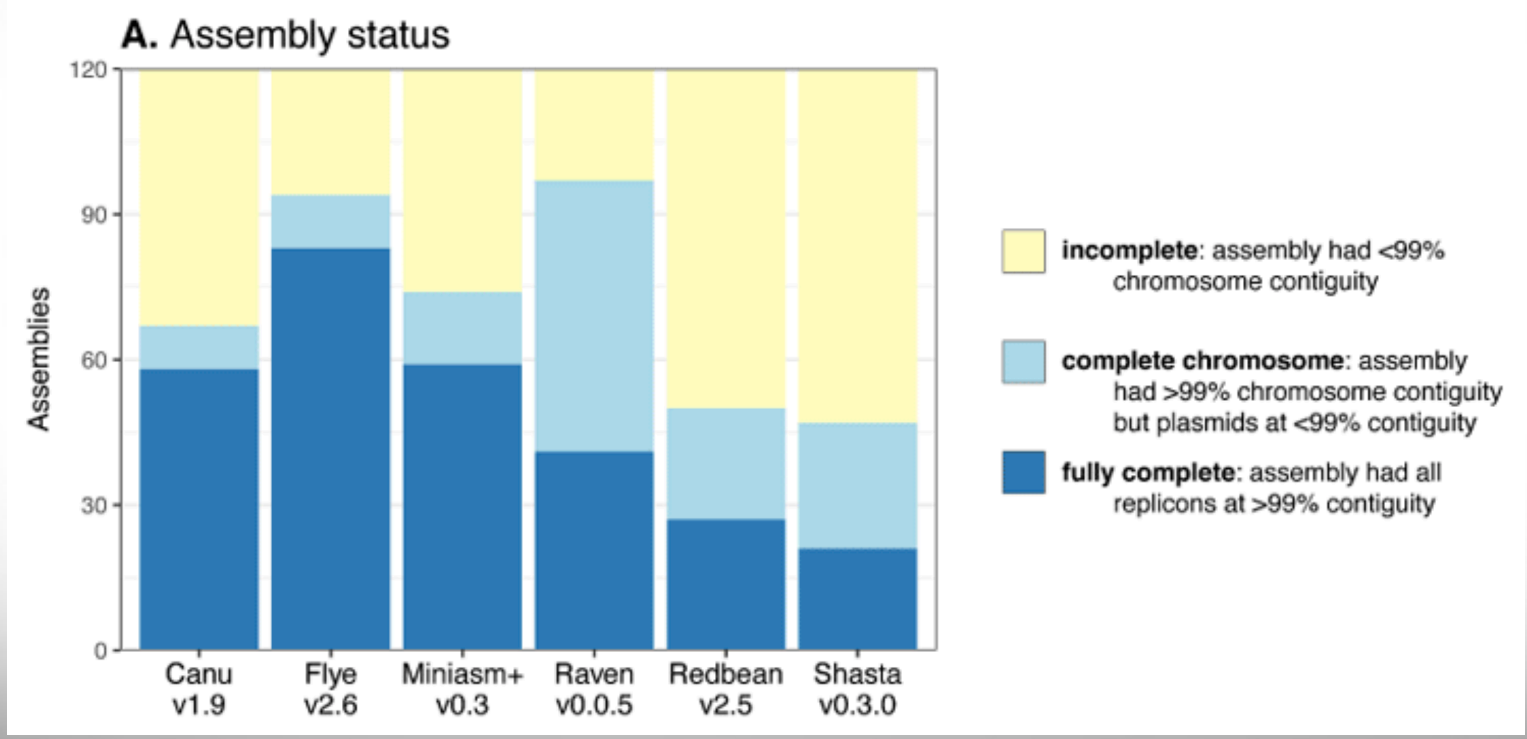
Velvet, ABySS, AllPATHS, SOAPdenovo, DISCOVAR

Genome assembly algorithms for **Long** reads

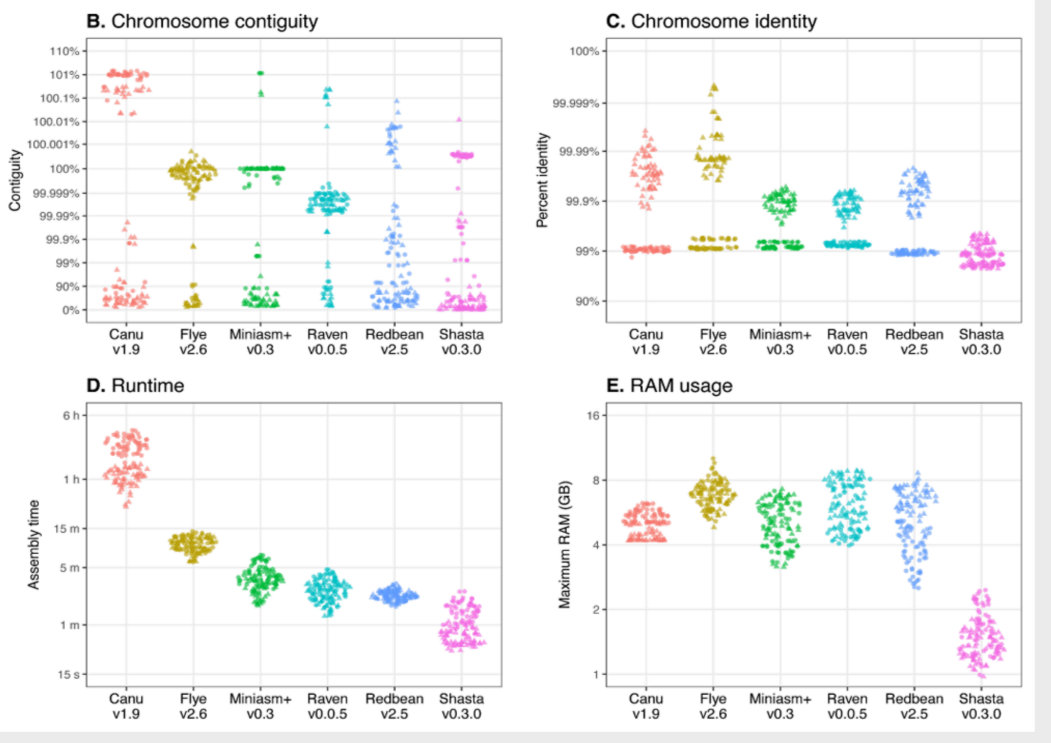
Graph algorithms with attentions on error correction

- General long read assembler: Canu, Flye, Miniasm/Minipolish, Raven, Redbean and Shasta
- Only works on PacBio: HGAP, FALCON
- Hybrid assembler: MaSuRCA, Unicycler

Benchmarking of long-read assemblers for prokaryote whole genome sequencing, https://f1000research.com/articles/8-2138



Benchmarking of long-read assemblers for prokaryote whole genome sequencing, https://f1000research.com/articles/8-2138



Transcriptome Assembly Tools:

- SOAPdenovo-Trans
- Trans-ABySS
- Velvet + Oases
- **Trinity (will cover in practical portion today)**
- Rnnotator
- CLC Bio Genomics Workbench

Assembly quality assessment

Genome assembly

1. Contig size: #contigs, Largest contig, total length, N50
2. Misassemblies and structural variations: # misassemblies, # misassbmled contigs, Length of misassembled contigs, # un-alinged contigs
3. Genome representation: Genome fraction, duplicatoin ratio, GC%, # variations Per 100Kb, # of genes covered

Tools: QUAST, CAGE

Assembly quality assessment

Transcriptome assembly

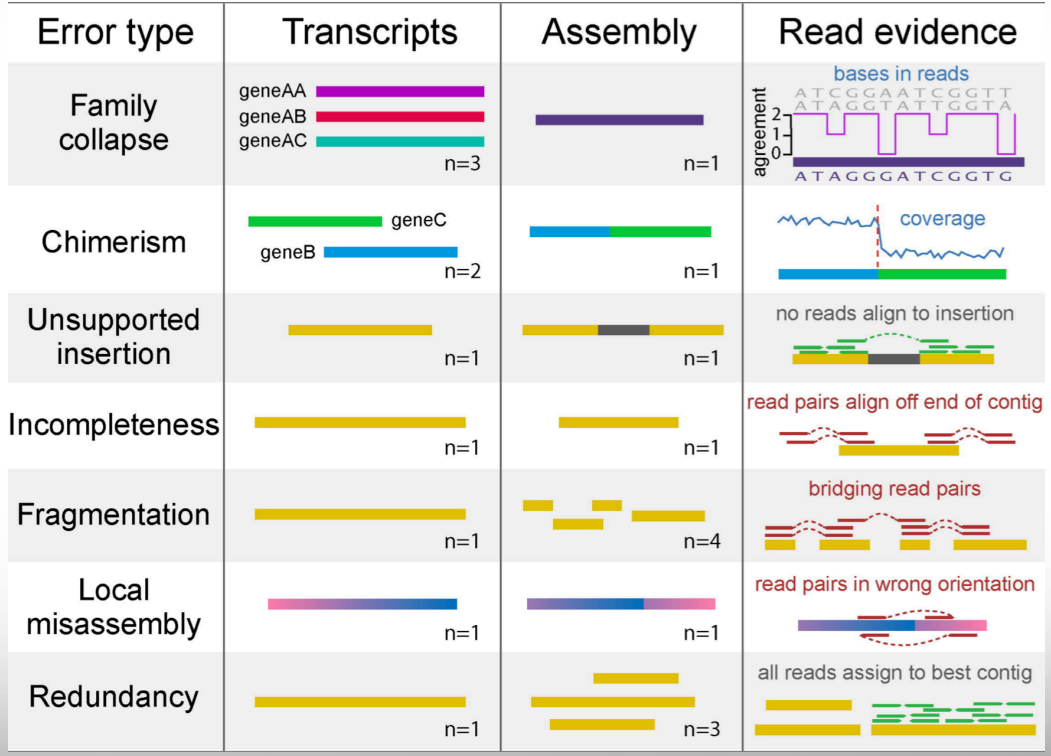
1. RNA-Seq read representation of the assembly, ~80%
2. Representation of full-length genes, by searching know protein sequences
3. Calculate E90N50, or the DETONATE scores
4. Recovery rate of the conserved genes

Tools: RnaQUAST, BUSCO, DETONATE, Transrate,

Assembly quality assessment

Transcriptome assembly

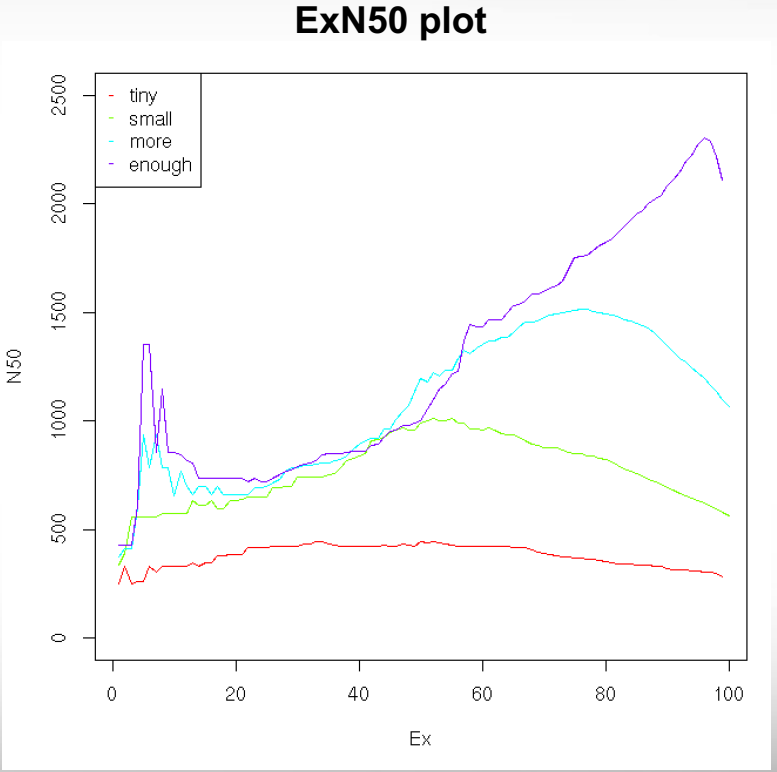
TransRate: reference free quality assessment of *de-novo* transcriptome assemblies



Assembly quality assessment

Transcriptome assembly

As the read depth is increased, the ExN50 peak begins to shift towards ~90%. In addition to exploring saturation of full-length reconstructed transcripts as a function of read depth, the ExN50 profiles can provide a useful guide towards understanding whether deeper sequencing might be expected to provide for a higher quality assembly.



https://github.com/trinityrnaseq/trinityrnaseq/wiki/Transcriptome-Contig-Nx-and-ExN50-stats

Metagenomics genome assembly

Assumptions made by the single genome assembly algorithms do not apply when assembling multiple genomes.

1. Unknown abundance and diversity
2. Related species

Assemble tools:
metaVelvet, metaSpades, MEGAHIT, et al

MetaQUAST for assessing the quality of the assembly

https://github.com/trinityrnaseq/trinityrnaseq/wiki/Transcriptome-Contig-Nx-and-ExN50-stats

Practice on Ada

- E. Coli genome assembly with PacBio data

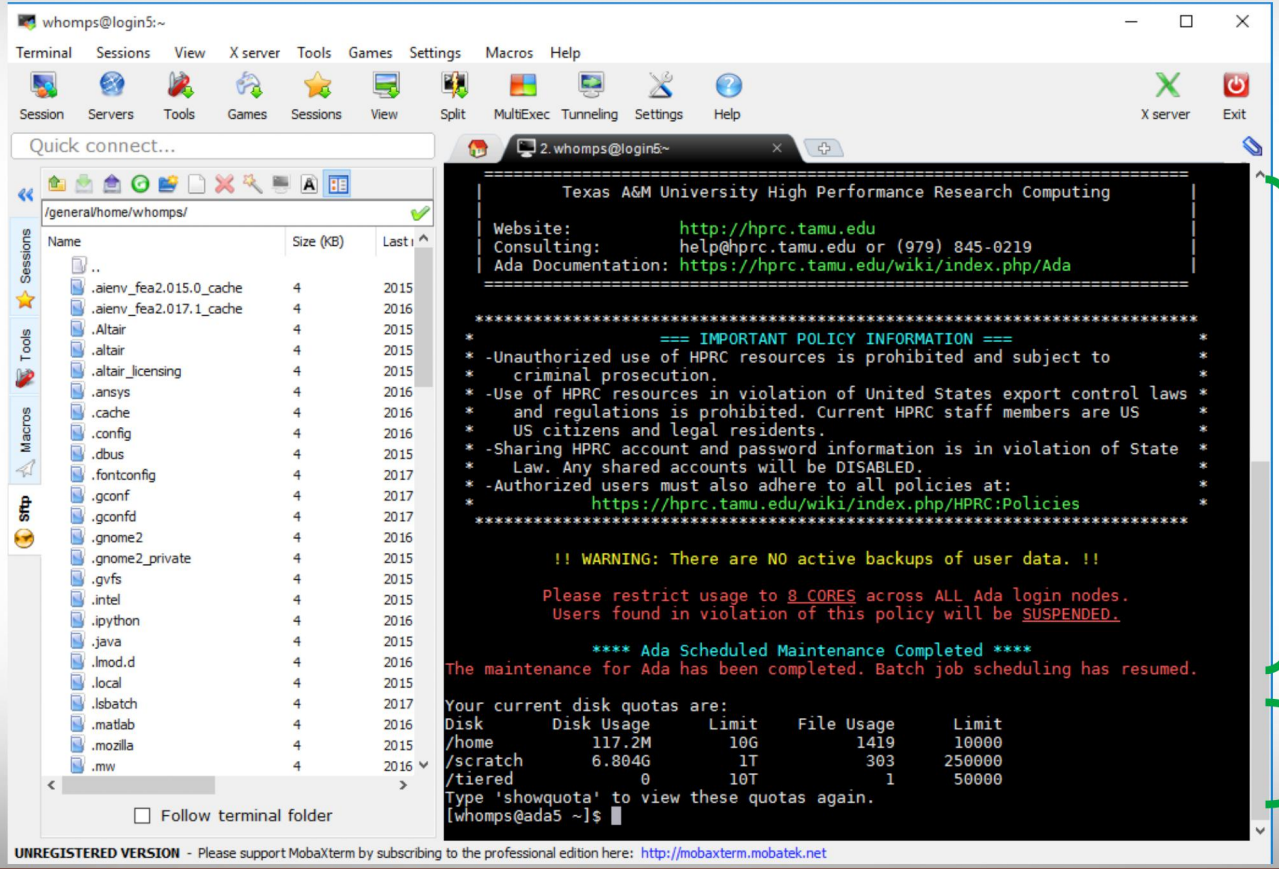
Command line, running Canu

- Small transcriptome assembly

https://galaxy-terra.hprc.tamu.edu/bdf/

https://github.com/trinityrnaseq/trinityrnaseq/wiki/Transcriptome-Contig-Nx-and-ExN50-stats

Using SSH - MobaXterm (on Windows)



message of the day

your quotas

Using SSH to Access Ada

```
ssh user_NetID@ada.tamu.edu
```

https://hprc.tamu.edu/wiki/Ada:Access

You may see something like the following the first time you connect to the remote machine from your local machine:

```
Host key not found from the list of known hosts.
Are you sure you want to continue connecting (yes/no)?
```

Type yes, hit enter and you will then see the following:

```
Host 'ada.tamu.edu' added to the list of known hosts.
user_NetID@ada.tamu.edu's password:
```

Any questions?

For More Help…

Website:	hprc.tamu.edu
Email:	help@hprc.tamu.edu
Telephone:	(979) 845-0219
Visit us in person:	Henderson Hall, Room 114A

Help us, help you -- we need more info

- Which Cluster
- UserID/NetID
- Job id(s) if any
- Location of your jobfile, input/output files
- Application used if any
- Module(s) loaded if any
- Error messages
- Steps you have taken, so we can reproduce the problem

Pacbio data assembly

https://github.com/swang8/assembly

Transcriptome assembly

https://galaxy-terra.hprc.tamu.edu/bdf/