Introduction to Genome/Transcriptome Assembly, using Sequencing Technologies

Noushin Ghaffari, PhD
Bioinformatics Scientist, Genomics and Bioinformatics, Texas A&M AgriLife Research
Research Scientist, Texas A&M High Performance Research Computing

DIVISION OF RESEARCH
TEXAS A&M UNIVERSITY

Outline

- Background
  - Sequencing
- Application of Sequencing in Research
  - Transcriptome assembly
  - Evaluations
  - Hands-on experiments

Primary NGS Applications

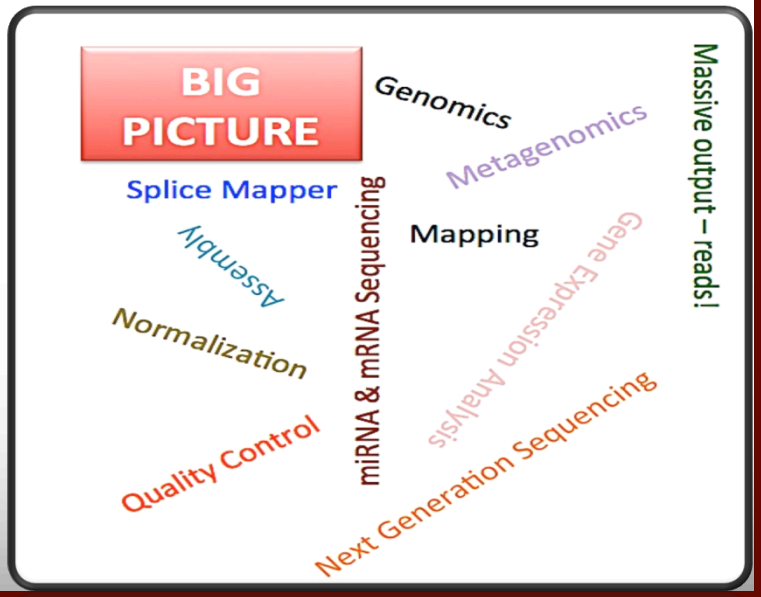
Today
1. Alignment
2. **Assembly** (no reference, with a reference)
   - **Genome**
   - **Transcriptome**

Last week → 3. RNA-Seq

Next Week → 4. Metagenomics

5. ChIP-Seq

Next Week → 6. RADSeq



Why sequencing?

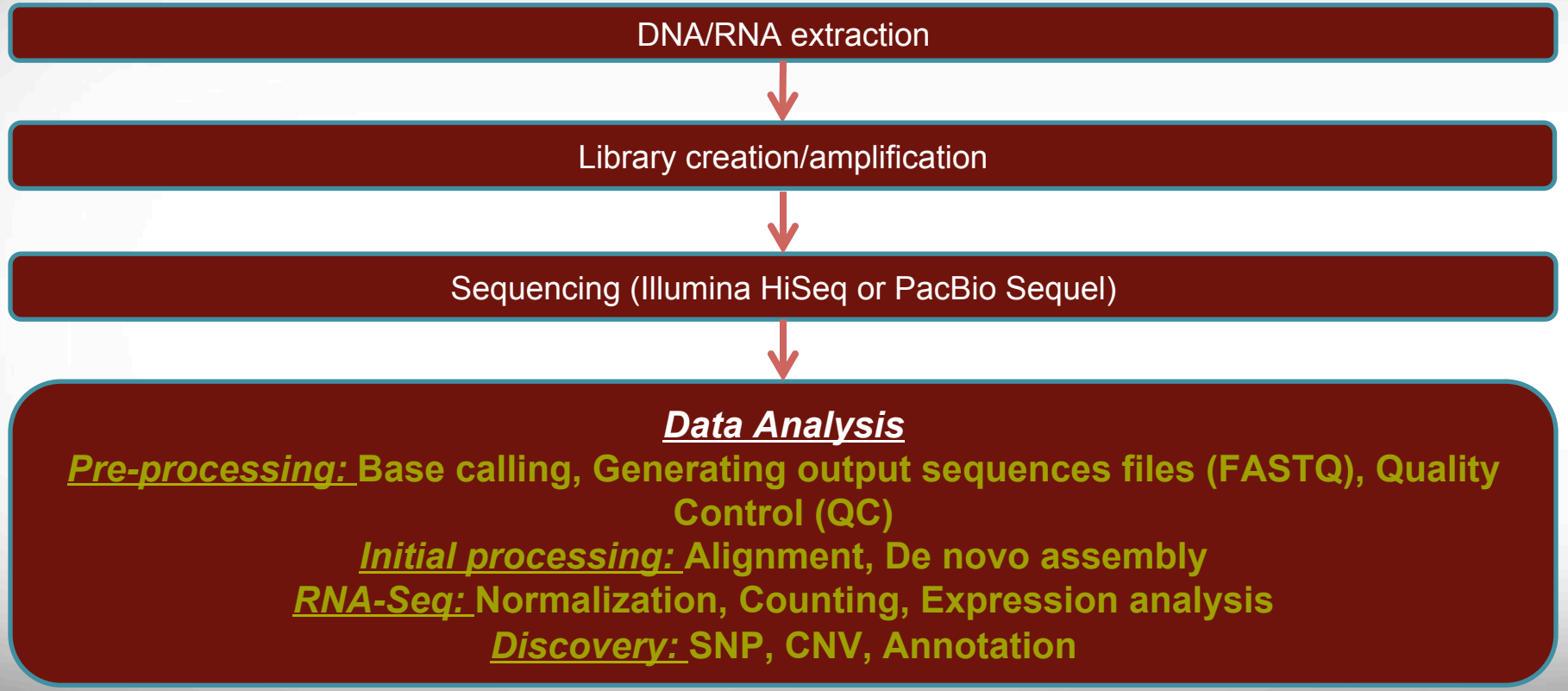
Determining the sequence of nucleotides within a DNA (or RNA) fragment

- Ultimately completing the genome of non-model organisms, e.g. *Pacific whiteleg shrimp*
- Human genome project, $3.8 Billion, 13 years to complete (1990-2003), 8-9x, coverage, 27 GBases

How?

Using sequencing methods, such as Sanger sequencing, next generation sequencing and single-molecule techniques

NGS Sequencing Workflow



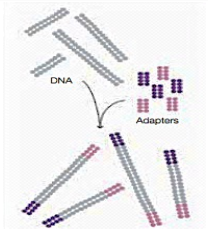
SHORT READS

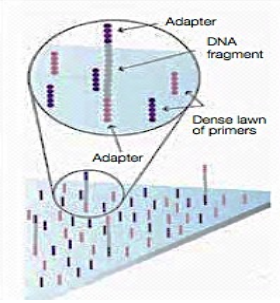
- Illumina

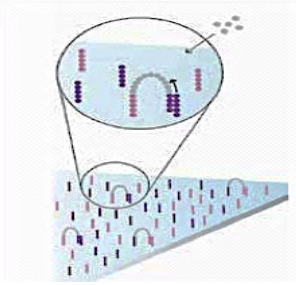
Illumina next-generation sequencing

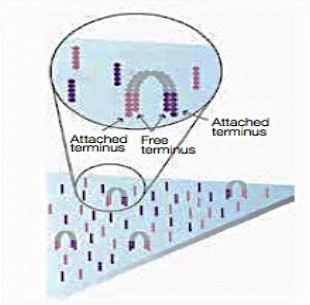
Sequencing by Synthesis (SBS) Technology

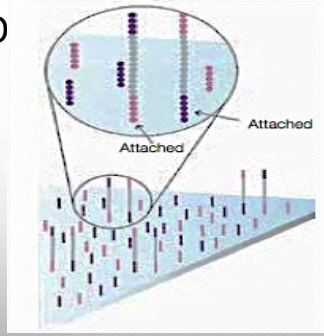
- Randomly shearing DNA
- Attaching DNA fragments to the flowcell surface
- Cluster generation, “Bridge Amplification”
- Adding four labelled *reversible terminators*, primers, and D polymerase
- Determining the attached nucleotide, based on the emitted fluorescence

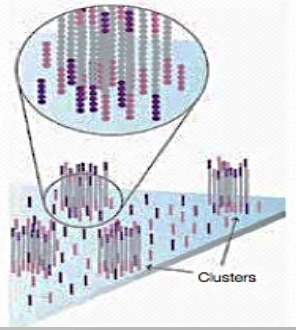


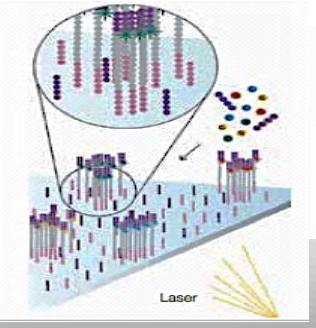


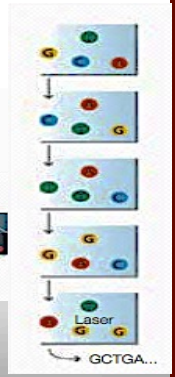












Sequence and Quality Scores

Quality scores measure the probability that a base is called incorrectly.

Quality Score

- Illumina Quality Score
  - Phred-like algorithm: similar to scoring for Sanger sequencing
  - Quality score of a given base, Q, is defined as:
  - e: estimated probability of the base call being wrong

$$Q = -10\log_{10}(e)$$

Quality Score	Probability of Incorrect Base Call	Inferred Base Call Accuracy
10 (Q10)	1 in 10	90%
20 (Q20)	1 in 100	99%
30 (Q30)	1 in 1000	99.90%

FASTQ Format

- Illumina 1.8+, Phred+33, raw reads typically (0, 41)

Read 1

@HWI-EASXXX:96:96:1:1:7939:13150 *1*:N:0:

TTCTCCCCCCTTCTCCGTTTCATTCCACCCGCCCTATTCCTTCGCCTCCTCTTCCTTG

+

BEHBHGDA(DA>CCAEAHHHHGGHGHADCF@CDCE@EGGGDHH?HG@GGDGFGGGGE=

@HWI-EASXXX:96:96:1:1:14632:1706 *1*:N:0:

CACGAGAACGAGAAGAAGAAATGGGGAGGAGTCACAGAGAGAGAGAGGGGAAGGGGGGAGGGAGAGGATGGAGGAGAAGGG

+

HHHHHFGD(GCGECGGHHHBDGEGGGGGG>HFHDHBG2D8C>C)C-@D?;A>ECECAA0A=;+B0A?+;AD<@DB>5=A@@

Read 2

@HWI-EASXXX:96:96:1:1:7939:13150 *2*:N:0:

CAAGGAAGAGGAGGCGAAGGAATAGGGCGGGTGGAATGAAACGGAGAAGAGGGGAGAA

+

4111166664@@@@@@@@@@@@@8@@@;@44284477778+4666575228884444@

@HWI-EASXXX:96:96:1:1:14632:1706 *2*:N:0:

ACCTTCTCCTCCATCCTCTCCCTCCCCCCTTCCCCTCTCTCTCTCTGTGACTCCTCCCCATTTCTTCTTCTCTTTCTCGTG

+

-555598888@C@@C@@@@@@@@@C444444@@@@@:40::6465689998@:@@@::4447677544:::@@;@@######

Choosing Illumina Sequencer!

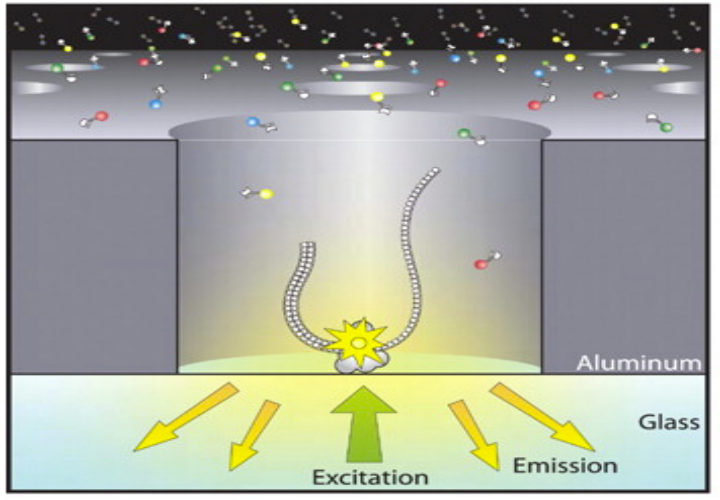
http://core-genomics.blogspot.com/2016/01/meet-newest-members-of-family-miniseq.html

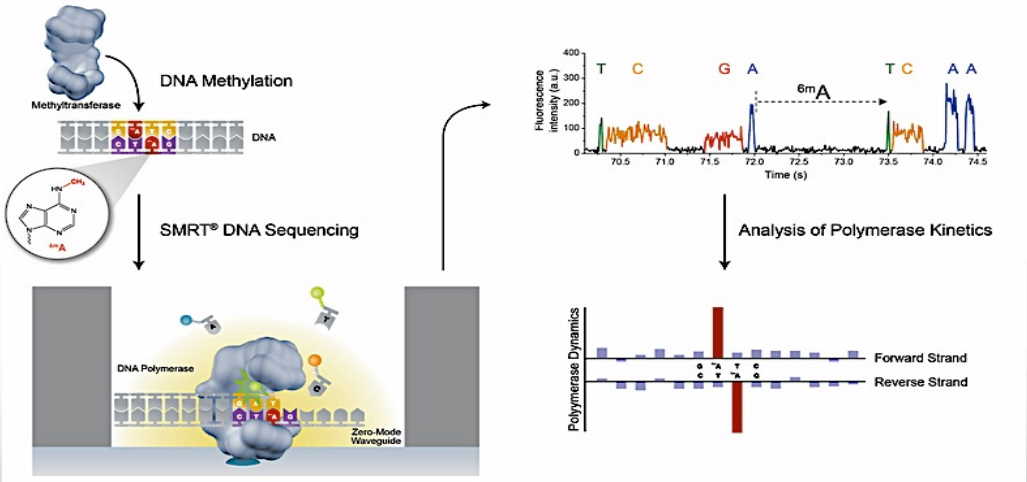
LONG READS

- Pacific Biosciences (PacBio)
- Oxford Nanopore Technologies - MinION

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

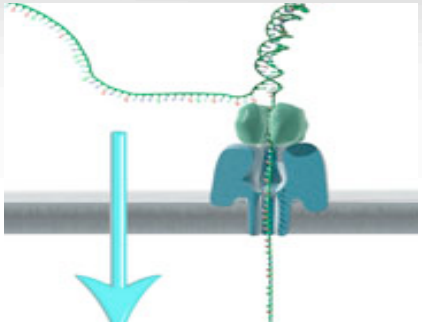
PacBio Sequel

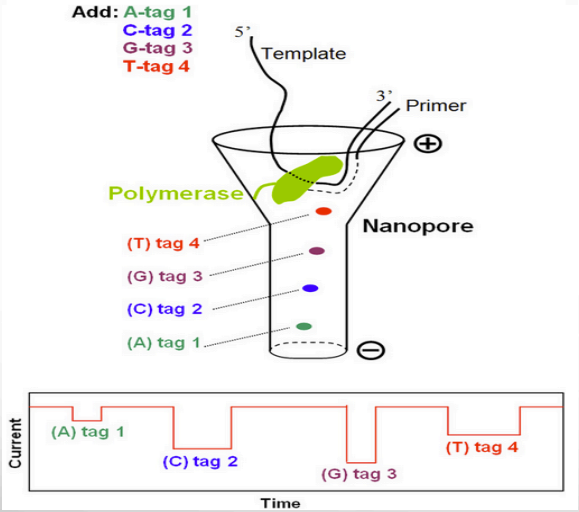
~10 GB per SMRT Cell
1M ZMW/SMRT Cell
Up 16 SMRT/week
10 hour run time/SMRT
Avg. read 10-15kb

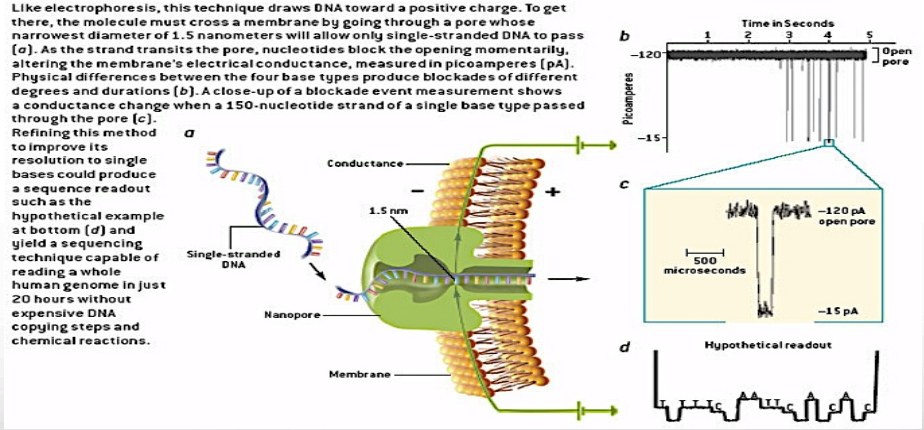
~10x jump over RSII

Long reads – Oxford NanoPore

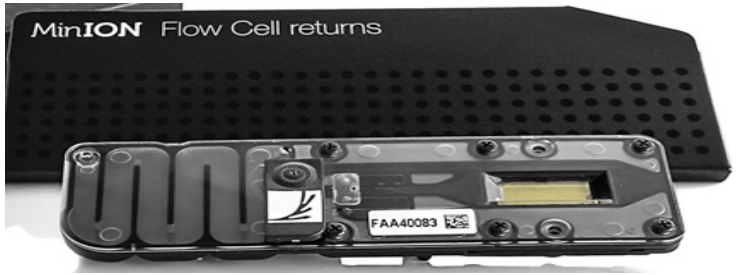
- Oxford Nanopore Technologies
- Nanopore: a small hole (nanometer)
  - used to identify DNA sequence, passing through nanopore
- Single DNA molecule is sequenced







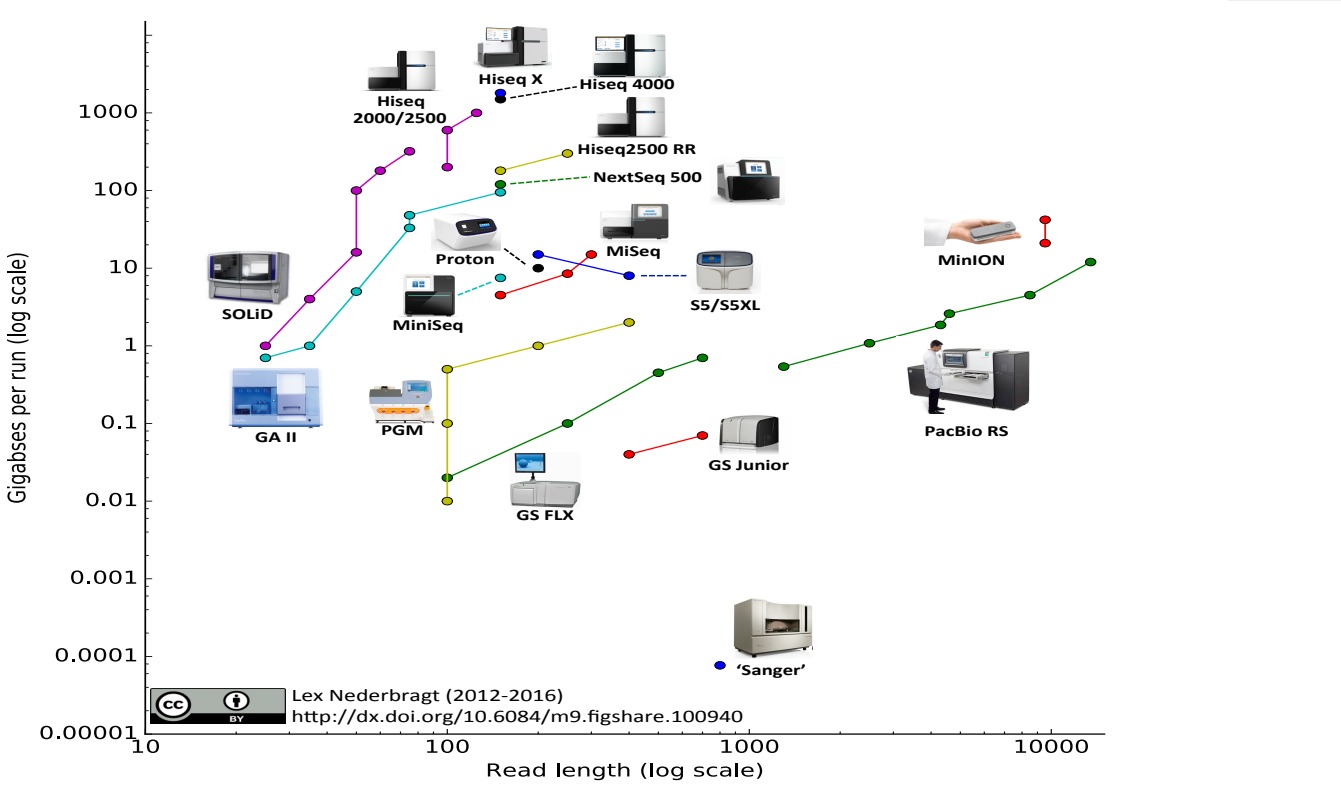
http://www.nature.com/scientificamerican/journal/v294/n1/full/scientificamerican0106-46.html
http://www.kurzweilai.net/single-molecule-electronic-dna-sequencing#!prettyPhoto



TAMU holds patent
Dr. Higgin Bailey

TEXAS A&M AGRILIFE RESEARCH

NGS Read Specifications



Lex Nederbragt blog: https://flxlexblog.wordpress.com/2016/07/08/developments-in-high-throughput-sequencing-july-2016-edition/

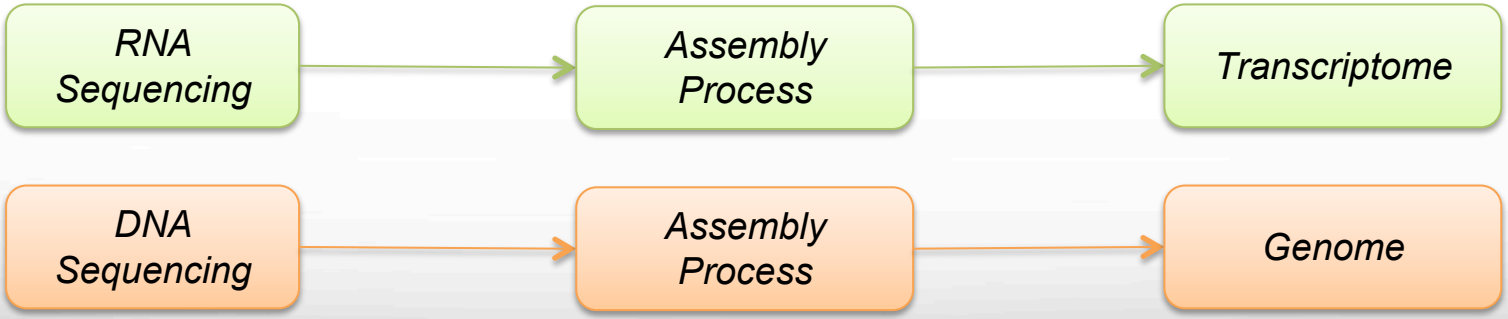
Comparing Sequencing Technologies

Platform	Read length	Error rates	Technology	Portable?
Illumina	< 400 bp	Low	Sequencing by synthesis	No
PacBio	~ 10-15 Kb	High	SMRT – ZMW	No
Oxford Nanopore Technologies	~ 5-8 Kb	High	Nanopore protein – strand sequencing	Yes

Why assembly?

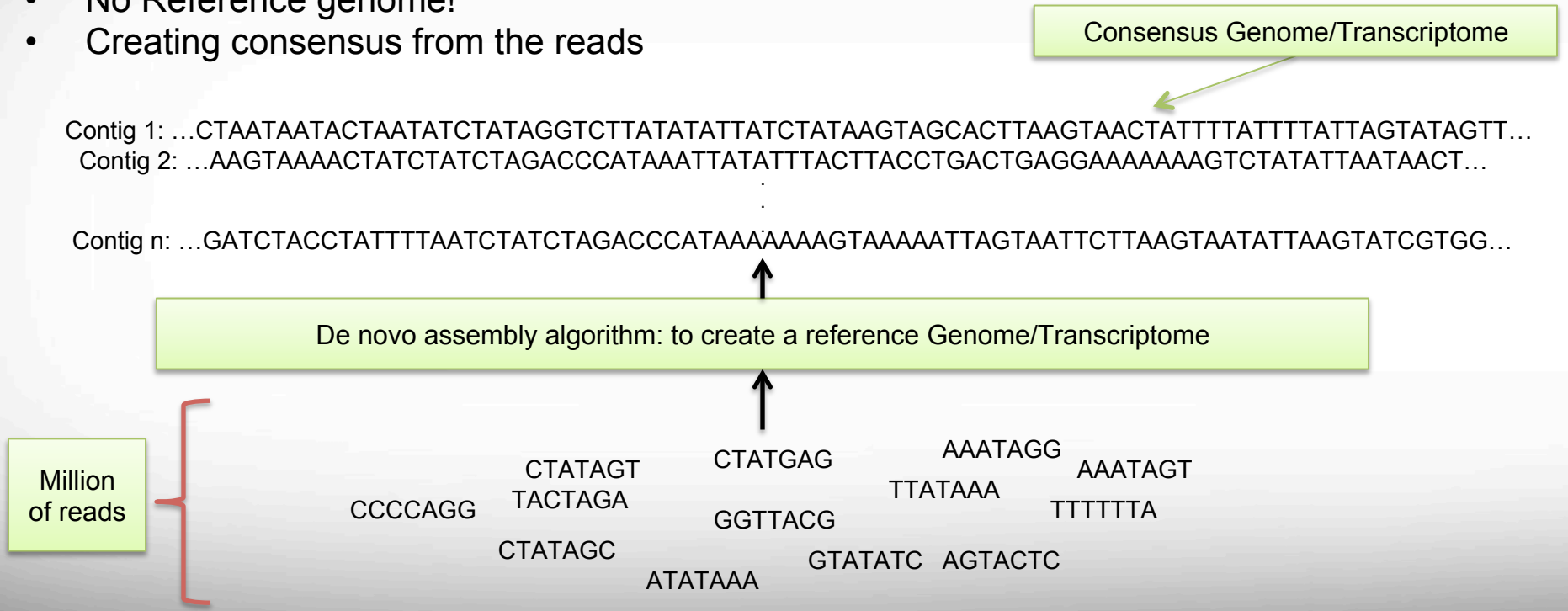
Generating the consensus of transcriptome or genome of non-model species

Reconstructing the genome and transcriptome of non-model species are essential steps in expanding our understanding of the organism and developing therapeutic approaches to fight disease



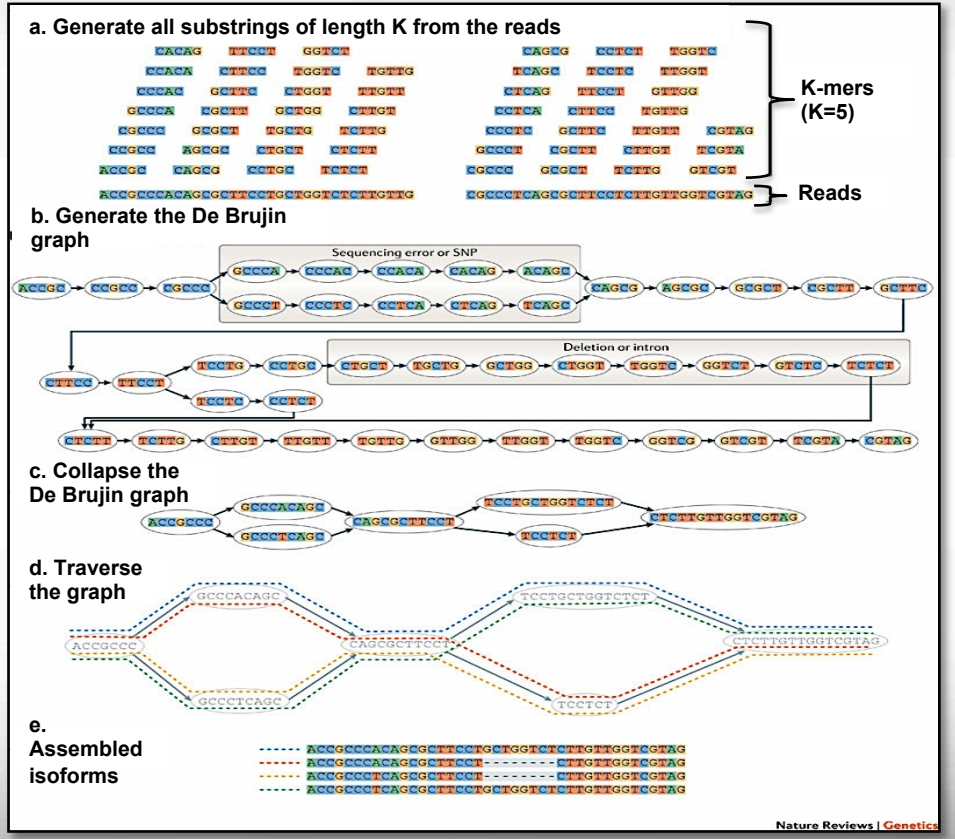
De novo Assembly

- Pool of reads
- No Reference genome!
- Creating consensus from the reads



De novo Assembly - 2

- Connection reads by finding common sections of kmers
  - Kmers are made from reads!
- Resolving conflicts
- Complicated process!
- Highly computational resource demanding!



Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. Genomics. 2010 Dec;95(6):315-27. doi: 10.1016/j.ygeno.2010.03.001. Epub 2010 Mar 17. PMID: 20211242; PMCID: PMC2874646.

Compeau PE, Pevzner PA, Tesler G. How to apply de Bruijn graphs to genome assembly. Nat Biotechnol. 2011 Nov 15;29(11):987-91. doi: 10.1038/nbt.2023. PMID: 22068540; PMCID: PMC5531759.

Nagarajan N, Pop M. Sequence assembly demystified. Nat Rev Genet. 2013 Mar;14(3):157-67. doi: 10.1038/nrg3367. Epub 2013 Jan 22. PMID: 23358380.

Martin JA and Wang Z. Next-generation transcriptome assembly. Nature Reviews Genetics 12, 671-682, 2011

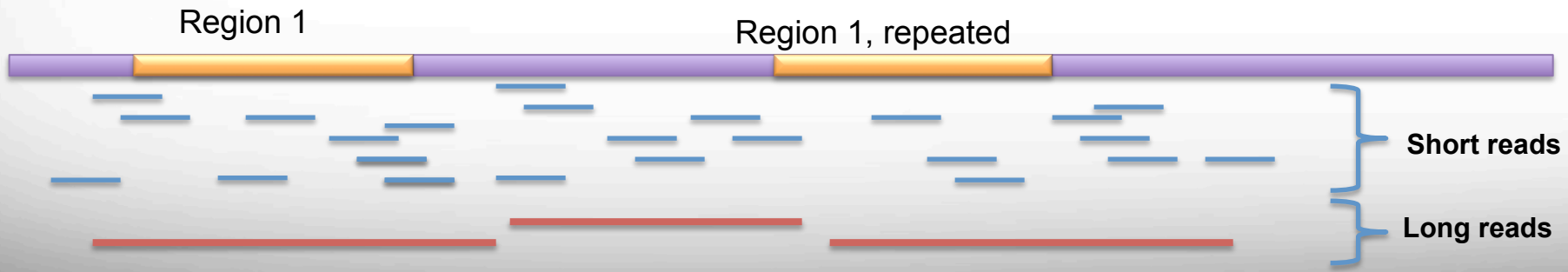
De novo Assembly - 3

Reference Genome Generation

- Goal: generating the reference genome for a new species, using the genomic DNA data, generated by NGS/Single Molecule platforms
- Main tool: *de novo* assembly algorithms
- Output: annotated reference genome

De novo Assembly - 4

Sanger	Next Generation Sequencing
Low coverage depth	High coverage depth
High cost for large genomes	Relatively low cost, even for large genomes
Slow	Fast
Handles repeats well	Need long reads for repeated regions (e.g. PacBio, Illumina Mate-Pair)



Genome Assembly Tools:

- ALLPATHS
- ALLPATHS-LG (Special recipe: fragments + jumping libraries)
- DISCOVAR *de novo*
- ABySS
- EULER-SR
- SOAPDenovo
- VCAKE
- Velvet
- MaSuRCA
- **Canu (will cover in practical portion today)**
- CLC Bio Genomics Workbench

Transcriptome Assembly Tools:

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

High Quality Assembly

- Hybrid Approach
- High Coverage
- Merging
  - Metassembler

Practical Portion

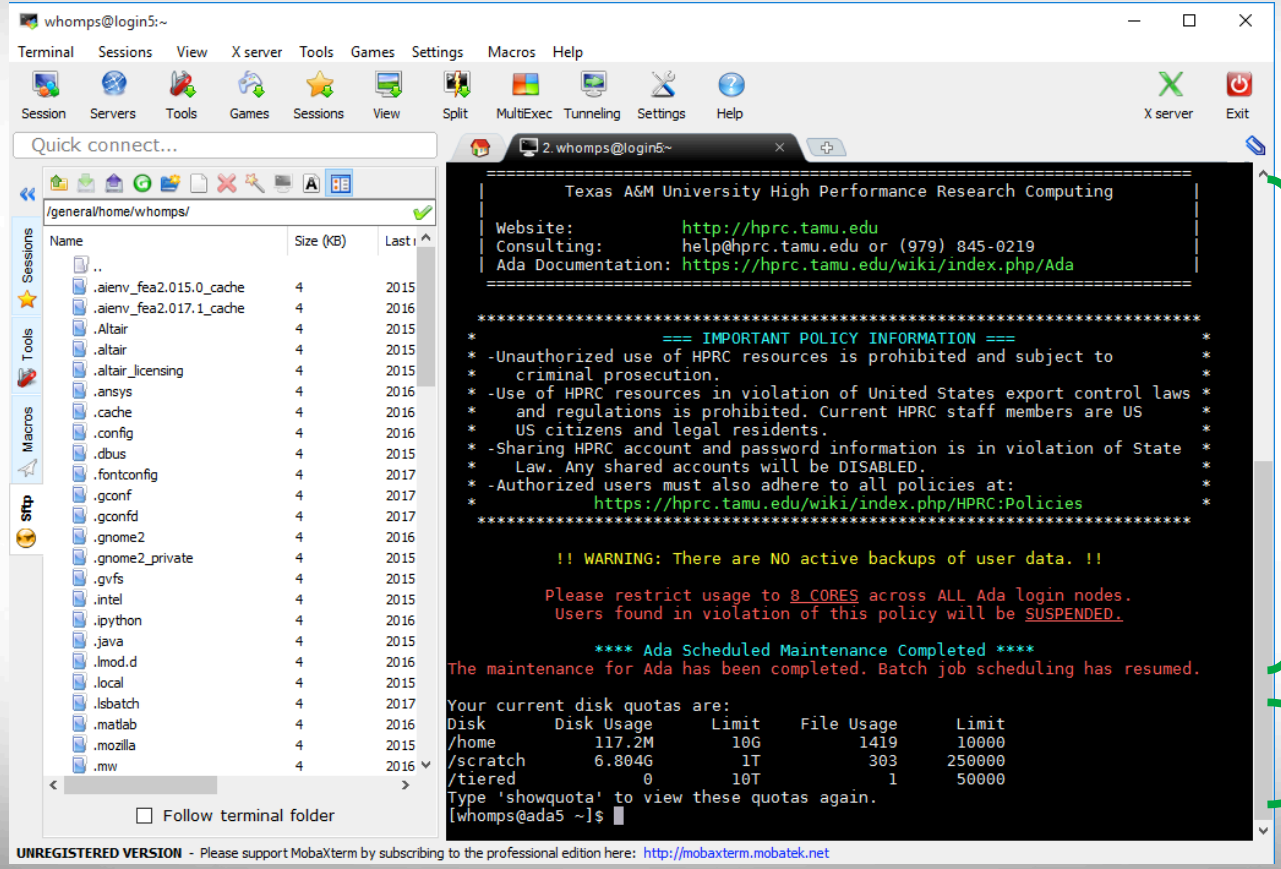
Logging in to the system

- SSH (secure shell)
  - The only program allowed for remote access; encrypted communication; freely available for Linux/Unix and Mac OS X hosts;
- For Microsoft Windows PCs, use *MobaXterm*
    - **https://hprc.tamu.edu/wiki/HPRC:MobaXterm**
      - You are able to view images and use GUI applications with MobaXterm
  - or *Putty*
    - **https://hprc.tamu.edu/wiki/HPRC:Access#Using_PuTTY**
      - You can not view images or use GUI applications with PuTTY

Your Login Password

- Both state of Texas law and TAMU regulations prohibit the sharing and/or illegal use of computer passwords and accounts
- Don’t write down passwords
- Don’t choose easy to guess/crack passwords
- Change passwords frequently

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:
```

Contact the HPRC Helpdesk

Website: hprc.tamu.edu

Email:

help@hprc.tamu.edu

Telephone: (979) 845-0219

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

Any question?

nghaffari@tamu.edu