

Overview of Assembly Techniques for Next Generation Sequencing Data

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 Next Generation Sequencing in Research
 - Transcriptome assembly
 - Evaluations
 - Hands-on experiments



Primary NGS Applications

1. Alignment

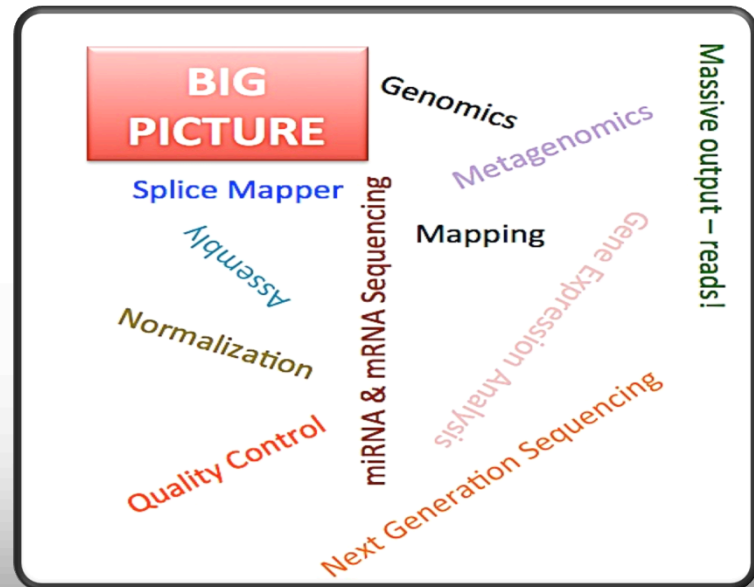
Today { **2. Assembly** (no reference, with a reference)
• **Genome**
• **Transcriptome**

This Morning → 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

Sanger



Classic Sequencing

Third Generation Sequencing Platforms

PacBio



MinION



Next Generation Sequencing Platforms

Illumina



© 2014 Illumina, Inc. All rights reserved.

<http://nextgenseek.com/2014/01/illumina-announces-new-sequencers-hiseqx-nextseq-500-at-jpm-2014/>



Choosing Illumina Sequencer!

MiniSeq



MAX OUTPUT

8 Gb

MAX READ NUMBER

25 million

MAX READ LENGTH

2x150 bp

MiSeq



MAX OUTPUT

15 Gb

MAX READ NUMBER

25 million

MAX READ LENGTH

2x300 bp

NextSeq



MAX OUTPUT

120 Gb

MAX READ NUMBER

400 million

MAX READ LENGTH

2x150 bp

HiSeq 4000



MAX OUTPUT

1500 Gb

MAX READ NUMBER

5 billion

MAX READ LENGTH

2x150 bp

HiSeq X Ten



MAX OUTPUT

1800 Gb

MAX READ NUMBER

6 billion

MAX READ LENGTH

2x150 bp

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

NGS Sequencing Workflow

DNA/RNA extraction



Library creation/amplification



Sequencing (Illumina HiSeq or Roche 454)



Data Analysis

Pre-processing: Base calling, Generating output sequences files (FASTQ), Quality Control (QC)

Initial processing: Alignment, De novo assembly

RNA-Seq: Normalization, Counting, Expression analysis

Discovery: SNP, CNV, Annotation

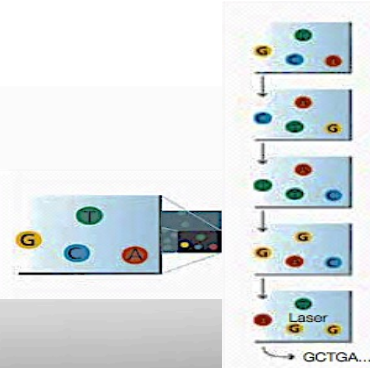
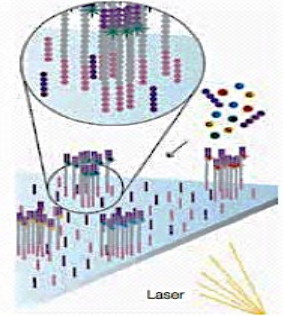
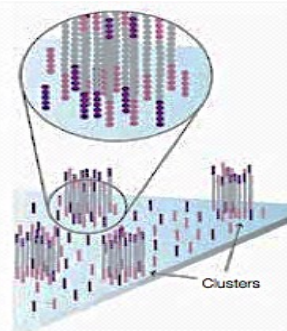
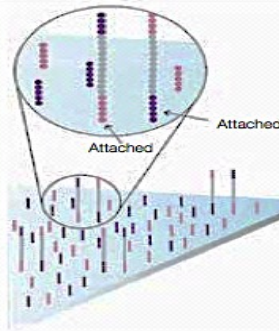
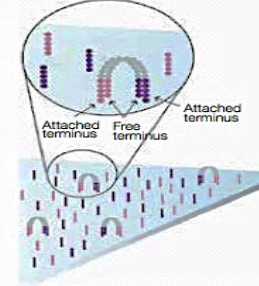
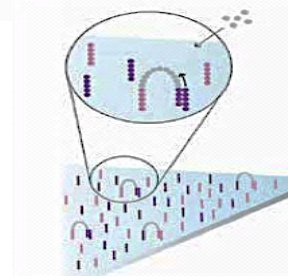
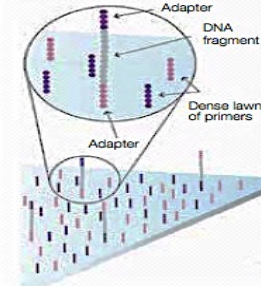
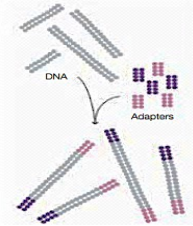
SHORT READS

- Illumina

Illumina next-generation sequencing

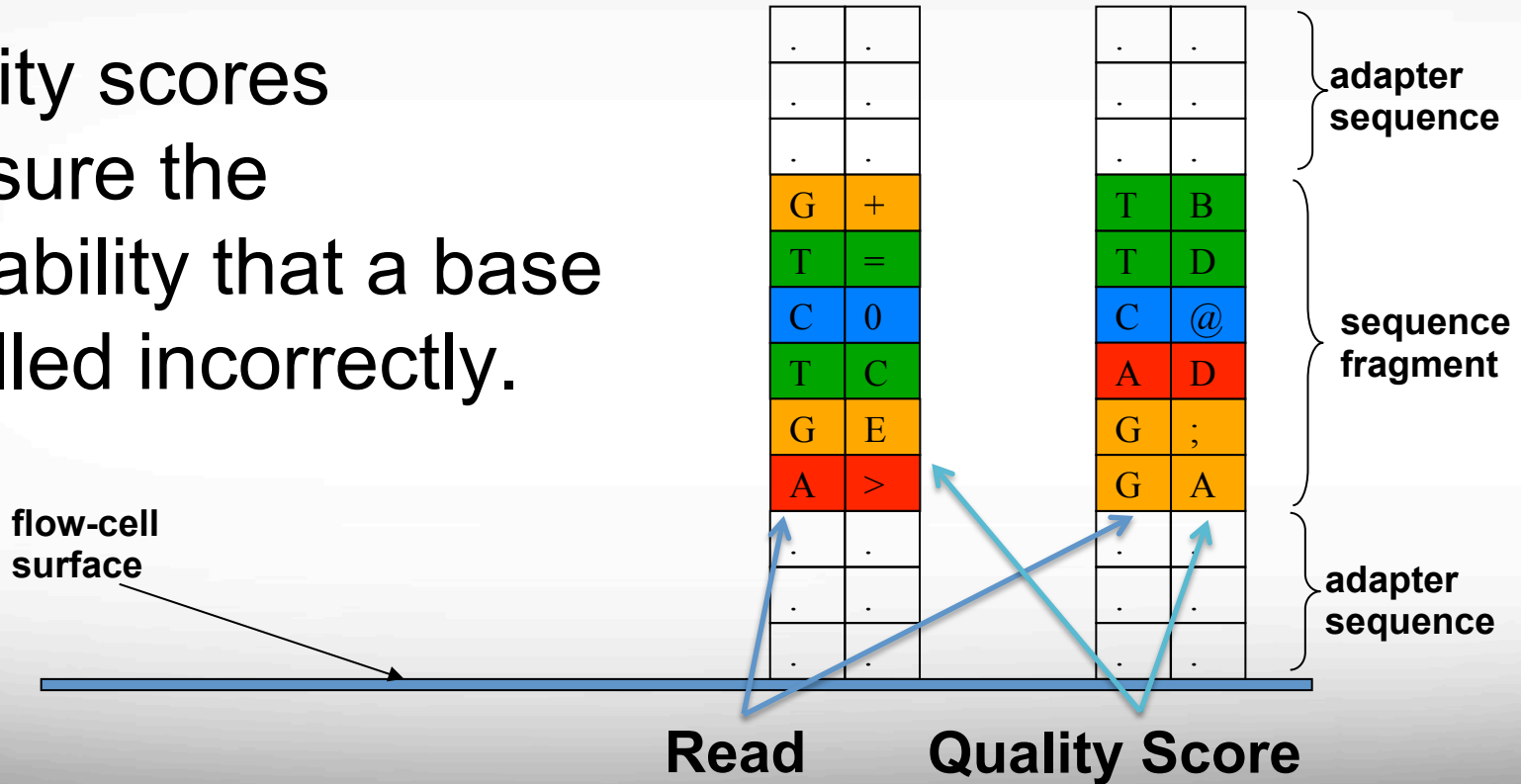
Sequencing by Synthesis (SBS) Technology

- 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:

TTCTCCCCCTTCTCCGTTTCATTCCACCCGCCCTATTCCTTCGCCTCCTCTTCCTTG

+

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:

CAAGGAAGAGGAGGCGAAGGAATAGGGCGGGTGAATGAAACGGAGAAGAGGGGAGAA

+

4111166664@@@@@@@@@@@@@8@@@@:44284477778+4666575228884444@

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

ACCTTCTCCTCCATCCTCTCCCTCCCCCTCCCCTCTCTCTCTGTGACTCCTCCCCATTTCTTCTTTCTTTCTCGTG

+

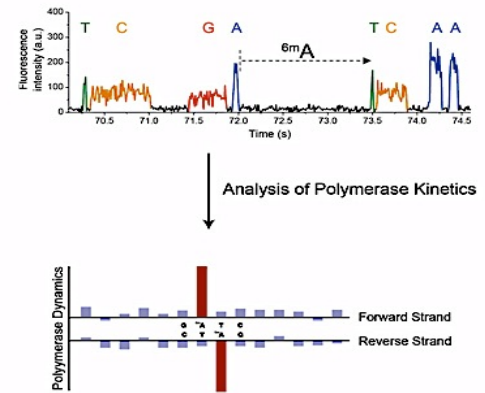
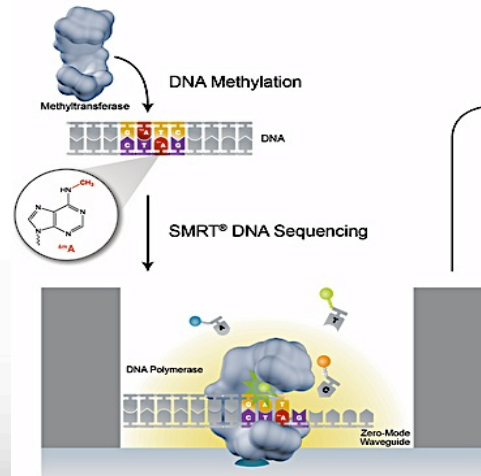
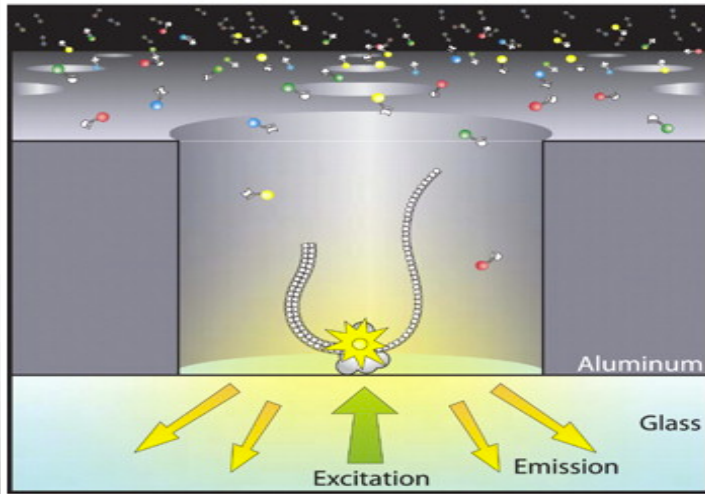
-555598888@C@@C@@@@@@@@@C444444@@@@@:40::6465689998@:@@@:4447677544::@@:@@#####

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

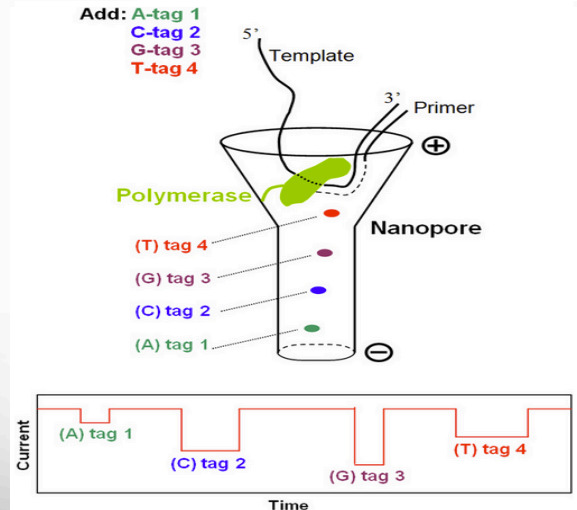
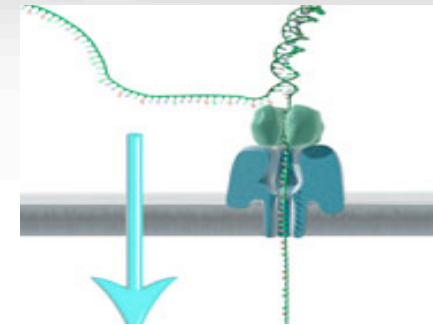


~10 GB per SMRT Cell
1M ZMW/SMRT Cell
Up 16 SMRT/week
10 hour run time/SMRT
Avg. read 10-15kb

~10x jump over RSII

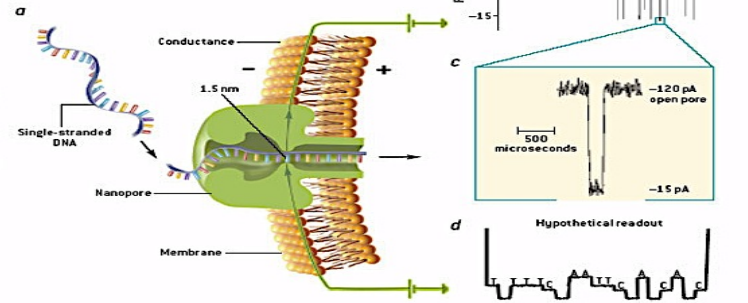
Long reads – Oxford NanoPore

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



Like electrophoresis, this technique draws DNA toward a positive charge. To get there, the molecule must cross a membrane by going through a pore whose narrowest diameter of 1.5 nanometers will allow only single-stranded DNA to pass [a]. As the strand transits the pore, nucleotides block the opening momentarily, altering the membrane's electrical conductance, measured in picoamperes [pA]. Physical differences between the four base types produce blockades of different degrees and durations [b]. A close-up of a blockade event measurement shows a conductance change when a 150-nucleotide strand of a single base type passed through the pore [c].

Refining this method to improve its resolution to single bases could produce a sequence readout such as the hypothetical example at bottom [d] and yield a sequencing technique capable of reading a whole human genome in just 20 hours without expensive DNA copying steps and chemical reactions.



<http://www.nature.com/scientificamerican/journal/v294/n1/full/scientificamerican0106-46.html>

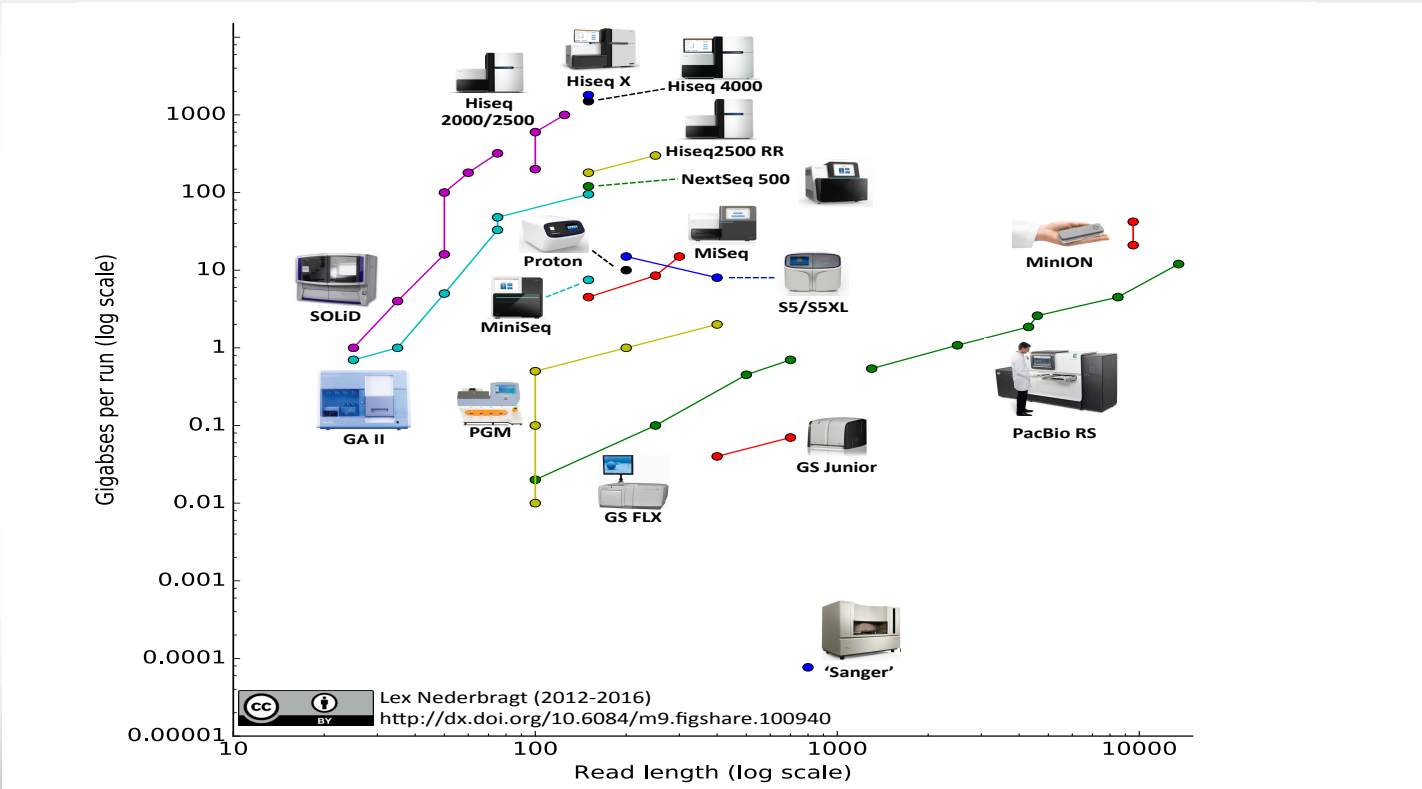
<http://www.kurzweilai.net/single-molecule-electronic-dna-sequencing#!prettyPhoto>



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/>

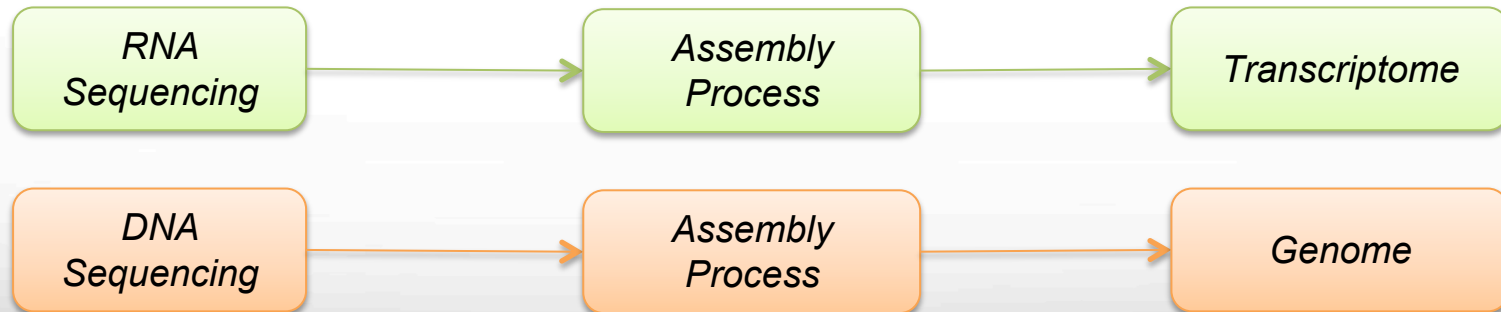
Comparing Sequencing Technologies

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

Consensus Genome/Transcriptome

Contig 1: ...CTAATAACTAATATCTATAGGTCTTATATATTATCTATAAGTAGCACTTAAGTAACTATTTTATTTTATTAGTATAGTT...

Contig 2: ...AAGTAAACTATCTATCTAGACCCATAAATTATATTTACTTACCTGACTGAGGAAAAAAGTCTATATTAATAACT...

⋮

Contig n: ...GATCTACCTATTTTAATCTATCTAGACCCATAAAAAAAGTAAAAATTAGTAATTCTTAAGTAATATTAAGTATCGTGG...

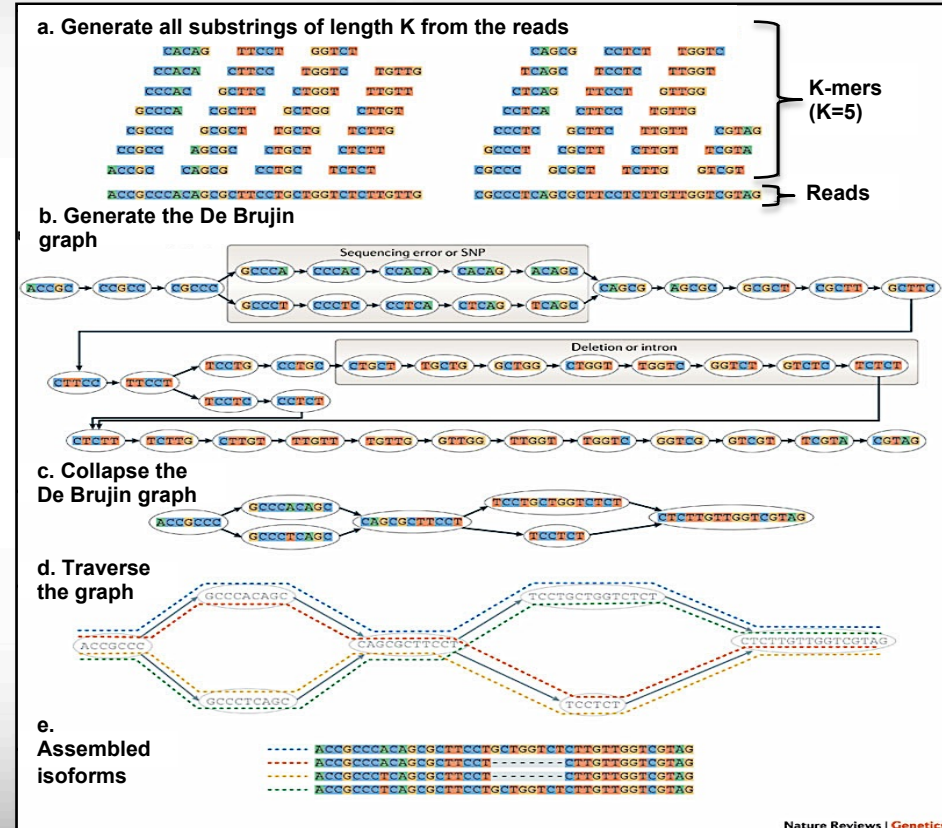
De novo assembly algorithm: to create a reference Genome/Transcriptome

Million of reads

CCCCAGG CTATAGT CTATGAG AAATAGG AAATAGT
 TACTAGA
 GGTTACG TTATAAA TTTTTTA
 CTATAGC ATATAAA GTATATC AGTACTC

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!



De novo Assembly - 3

Reference Genome Generation

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

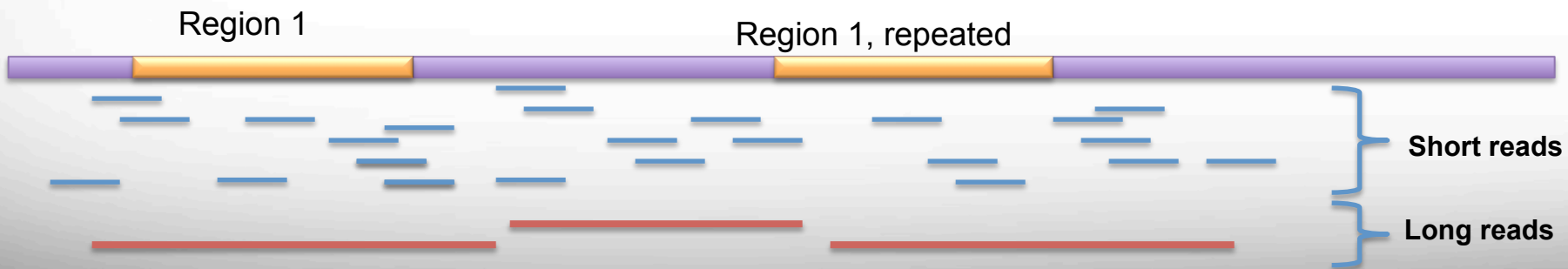
Major steps

- Step 1: Assembly
 - ALLPATHS-LG (large genomes, recently DISCOVAR *de novo*): Broad institute
- Step 2: Annotation
 - PASA: Broad institute



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
- **Canu**
- CLC Bio Genomics Workbench

Transcriptome Assembly Tools:

- SOAPdenovo-Trans
- Trans-ABYSS
- Velvet + Oases
- **Trinity**
- 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

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

Using SSH - MobaXterm (on Windows)

The screenshot shows the MobaXterm interface. On the left is a file explorer showing the local file system. The main terminal window displays the following output:

```
whomps@login5:~
Terminal Sessions View X server Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MultiExec Tunneling Settings Help X server Exit

Quick connect...
/general/home/whomps/

Name      Size (KB)  Last
..         4          2015
.aienv_fea2.015.0_cache 4          2015
.aienv_fea2.017.1_cache 4          2015
..altair   4          2015
..altair   4          2015
..altair_licensing 4          2015
..ansys    4          2016
..cache    4          2016
..config   4          2016
..dbus     4          2015
..fontconfig 4          2017
..gconf    4          2017
..gconfd   4          2017
..gnome2   4          2016
..gnome2_private 4          2015
..gvfs     4          2015
..intel    4          2015
..ipython  4          2016
..java     4          2015
..lmod.d   4          2016
..local    4          2015
..lsbatch  4          2017
..matlab   4          2016
..mozilla  4          2015
..mw       4          2016

[ ] Follow terminal folder

UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: http://mobaxterm.mobatek.net

=====
Texas A&M University High Performance Research Computing
Website:      http://hprc.tamu.edu
Consulting:   help@hprc.tamu.edu or (979) 845-0219
Ada Documentation: https://hprc.tamu.edu/wiki/index.php/Ada
=====

*****
                == IMPORTANT POLICY INFORMATION ==
*****
* -Unauthorized use of HPRC resources is prohibited and subject to
*   criminal prosecution.
* -Use of HPRC resources in violation of United States export control laws
*   and regulations is prohibited. Current HPRC staff members are US
*   US citizens and legal residents.
* -Sharing HPRC account and password information is in violation of State
*   Law. Any shared accounts will be DISABLED.
* -Authorized users must also adhere to all policies at:
*   https://hprc.tamu.edu/wiki/index.php/HPRC:Policies
*****

!! WARNING: There are NO active backups of user data. !!

Please restrict usage to 8 CORES across ALL Ada login nodes.
Users found in violation of this policy will be SUSPENDED.

**** Ada Scheduled Maintenance Completed ****
The maintenance for Ada has been completed. Batch job scheduling has resumed.

Your current disk quotas are:
Disk      Disk Usage  Limit  File Usage  Limit
/home     117.2M      10G    1419         10000
/scratch  6.8046G    1T     303         250000
/tiered   0           10T    1           50000
Type 'showquota' to view these quotas again.
[whomps@ada5 ~]$
```

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 question?
nghaffari@tamu.edu