Practical wet-lab advice to generate good quality data

Anoja Perera (agp@stowers.org)

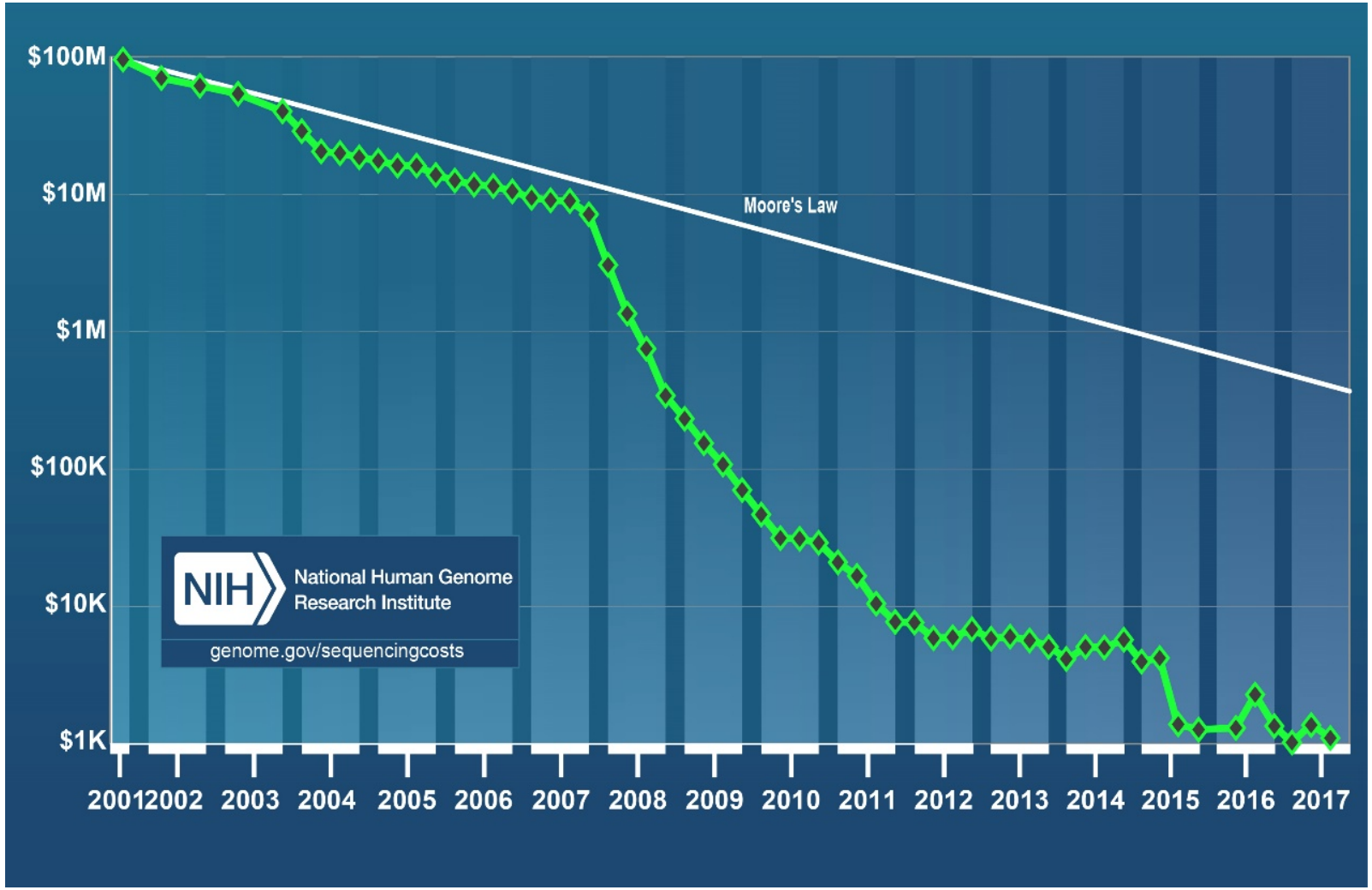
Director, Sequencing and Discovery Genomics

Stowers Institute for Medical Research

Agenda

- Introduction
- Short-Read vs Long-Read Sequencing
- Pacbio Sequencing
- Oxford Nanopore Sequencing
- Illumina Sequencing
- Data Quality Assessment

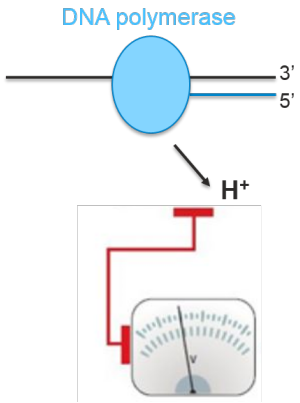
Cost per Human Genome



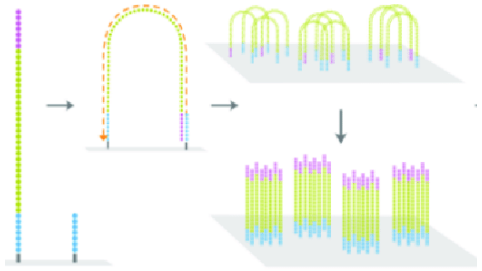
Source: National Human Genome Research Institute

Next Generation Sequencing: Short-Reads vs Long-Reads

Thermo Ion Torrent Proton

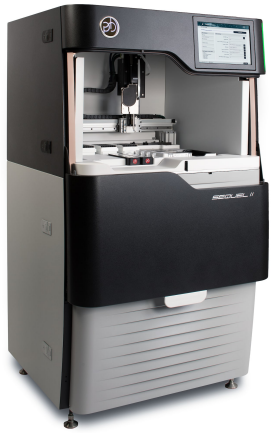


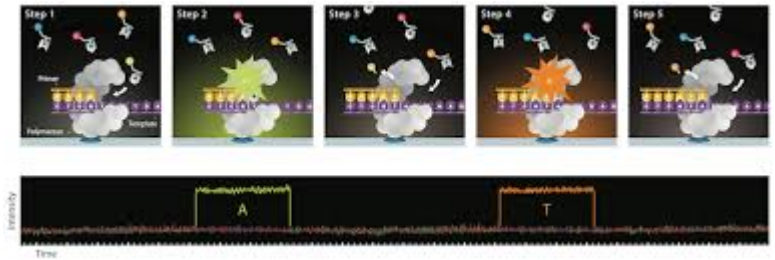
Illumina NovaSeq



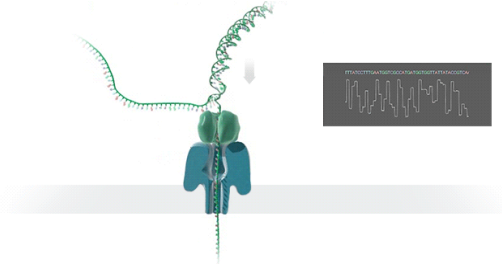
Westbury, Michael. (2018). Unraveling evolution through Next Generation Sequencing.

Pacbio Sequel II





Oxford Nanopore MinION



Pros and Cons of Short-Read data

Pros

- Cost per base is significantly lower than long-reads methods
- Higher accuracy, works well for SNP detection
- Great for counting applications such as Chromatin Immunoprecipitation Sequencing (ChIP-Seq) or gene expression studies

Cons

- PCR bias introduced by clonal amplification
- Difficulty sequencing high GC regions
- Difficulty in resolving substitution, deletions, duplications, haplotypes, palindromic and repetitive regions

Pros and Cons of Long-Read data

Pros

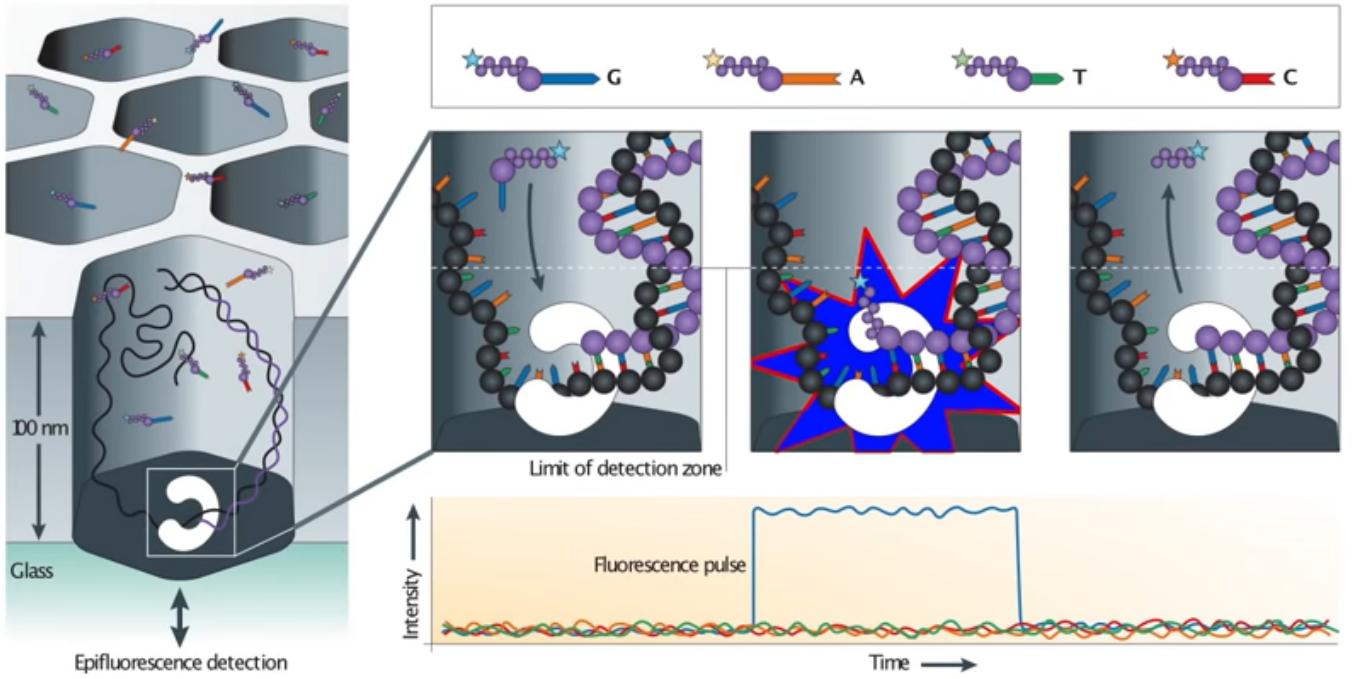
- No amplification bias during sequencing
- Can be used to determine phasing
- Assists in determining haplotypes, structural variations, indels etc.
- Aids in *de novo* assembly by spanning low complexity and repetitive regions
- Can obtain full length transcript information and hence aid isoform discovery
- Determination of base modifications such as methylated bases
- ONT allows sequencing in the field

Cons

- Costly
- ONT has difficulty with homopolymer regions
- Less accurate than short read sequencing

Pacbio Sequencing

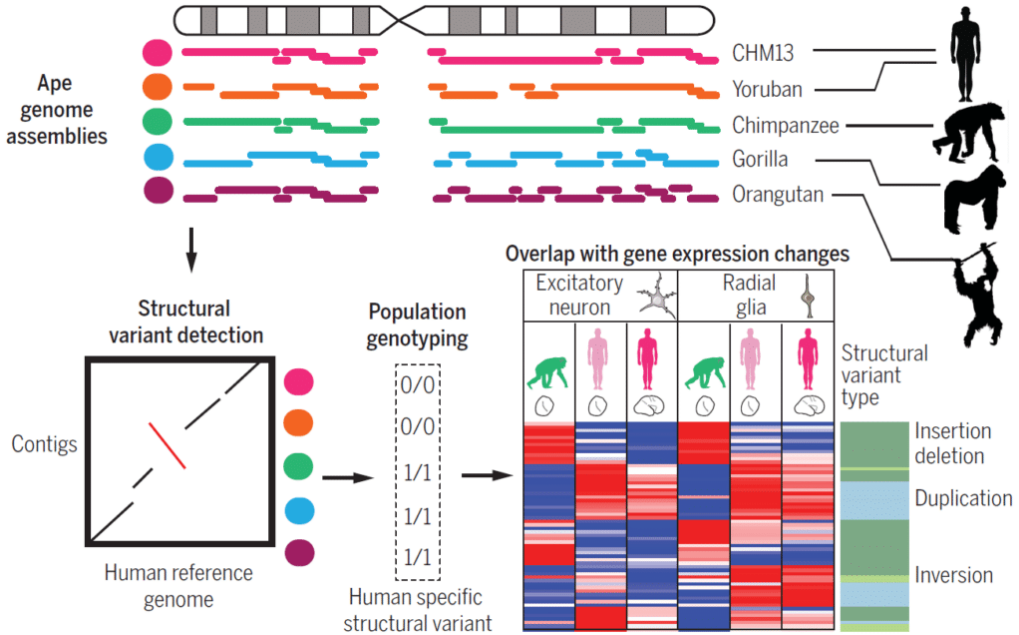
SMRT Cell



Nature Reviews Genetics **11**, 31–46 (2010)

PacBio Sequel II System Specifications and Applications

- Up to 4 million reads
- Run time of up to 30 hrs
- ~$1000/SMRT Cell
- 10-50kb insert sizes
- Applications
  - Whole Genome Sequencing
  - Metagenomics
  - Targeted Sequencing
  - RNA Sequencing
  - Epigenetics

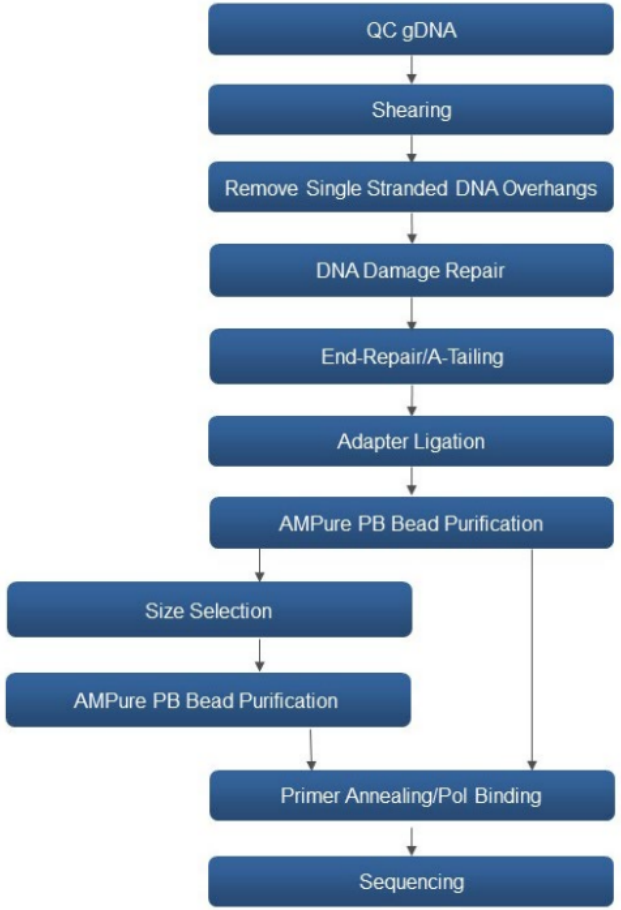


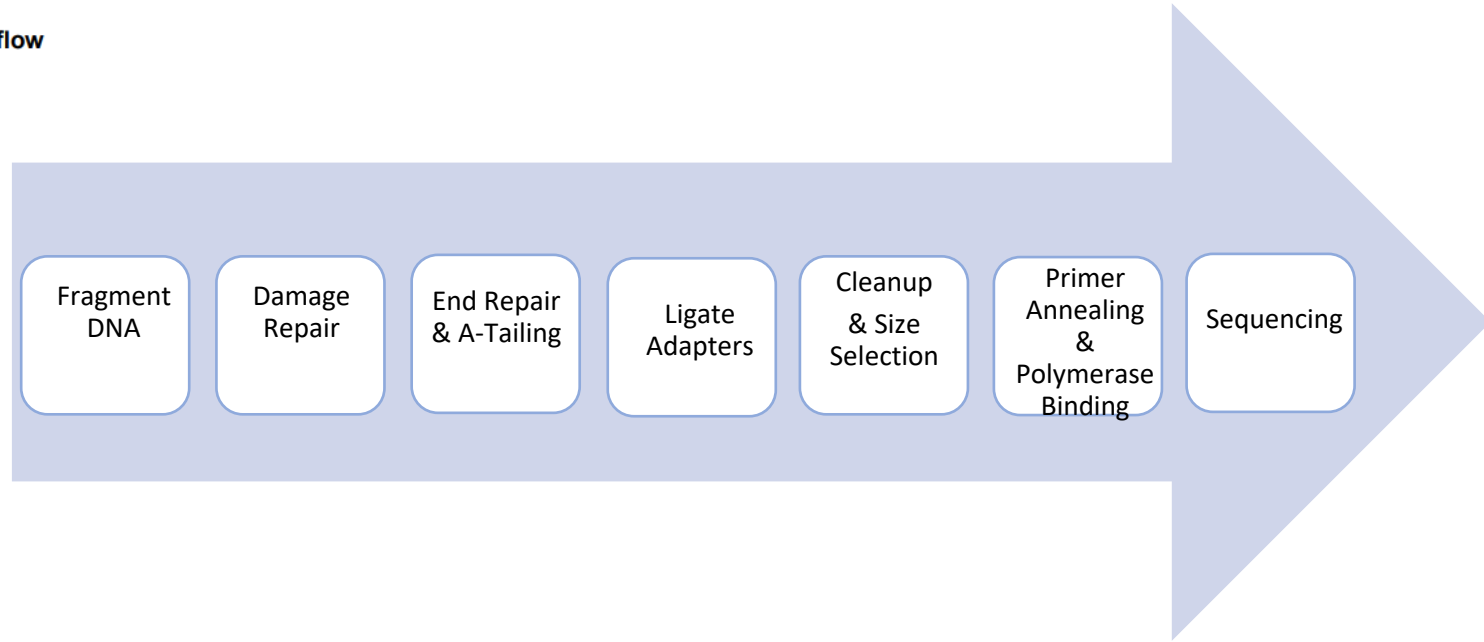
Scientists sequenced chimpanzee, orangutan and human genomes at >65-fold coverage and generated high-quality contiguous genome assemblies that improved gene annotation and understanding of genomic synteny among the species.

Kronenberg, Z. N., et al. (2018). High-resolution comparative analysis of great ape genomes. Science, 360(6393).

Pacbio Library Prep

Overview of the SMRTbell Express 2.0 Large-insert Library Workflow



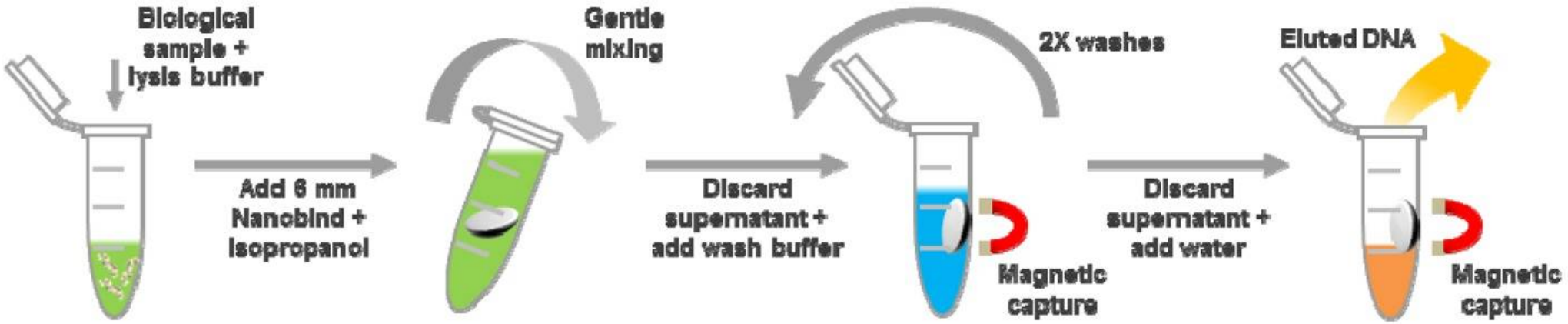


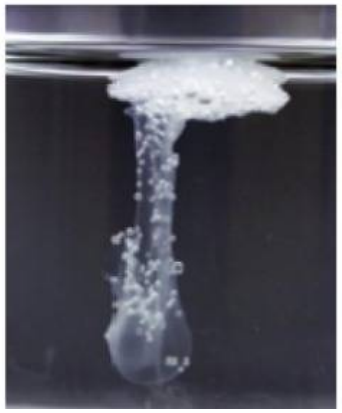
Best practices:

- Start with HMW DNA that has been slightly sheared (multiple approaches available)
- Avoid vortexing, pipette slowly using wide bore tips
- Sage Sciences' Blue Pippin for size selection
- Multiple approaches available for quality and quantity checks

Figure 1: Workflow for Preparing Large-insert Libraries Using the SMRTbell Express Template Preparation Kit 2.0.

Key to long read sequencing: HMW DNA!

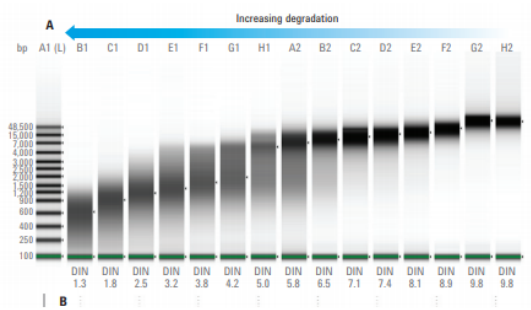




https://www.circulomics.com/

DNA Quality and Quantity Assessment

- Check quality on a pulse field gel, Agilent Tape Station or Advanced Analytical FEMTO



Agilent Tape Station

Advanced Analytical Femto

- Check for impurities via Nanodrop. 260 nm: 280 nm of ~1.8 is accepted as “pure” DNA; a ratio of ~2.0 is accepted as “pure” RNA.

Thermo Nanodrop

- Check for quantity via Invitrogen Qubit Fluorometer

Invitrogen Qubit Fluorometer

Physical / Mechanical Shearing

Method	Time	DNA Input	Fragment Length
Covaris Ultrasonicator	~ 1 min	Up to 5 µg	100 bp – 5 kb
Covaris g-Tube	10 min	~ 10 µg	6 – 20 kb
Needle Shearing	20 min	2 – 10 µg	> 30 kb
Diagenode Megaruptor	10 – 20 min	Up to 8 µg	2 – 75 kb

Ultra Long Read Sequencing

Oxford Nanopore Sequencing (ONT)

Additional Pros:

- No limitations in length of molecules that can be sequenced (>2 Mb)
- Capable of detecting DNA, RNA, protein, and modifications
- Active community improving and developing protocols
- Some platforms (Fongle, MinION) are portable!

Additional Cons:

- Difficulty sequencing homopolymer regions
- Error rate is higher than other methods
- Cannot sequence the same strand multiple times

Oxford Nanopore Sequencing

ONT sequencing platforms

Flongle

MinION Mk1B

GridION X5

PromethION

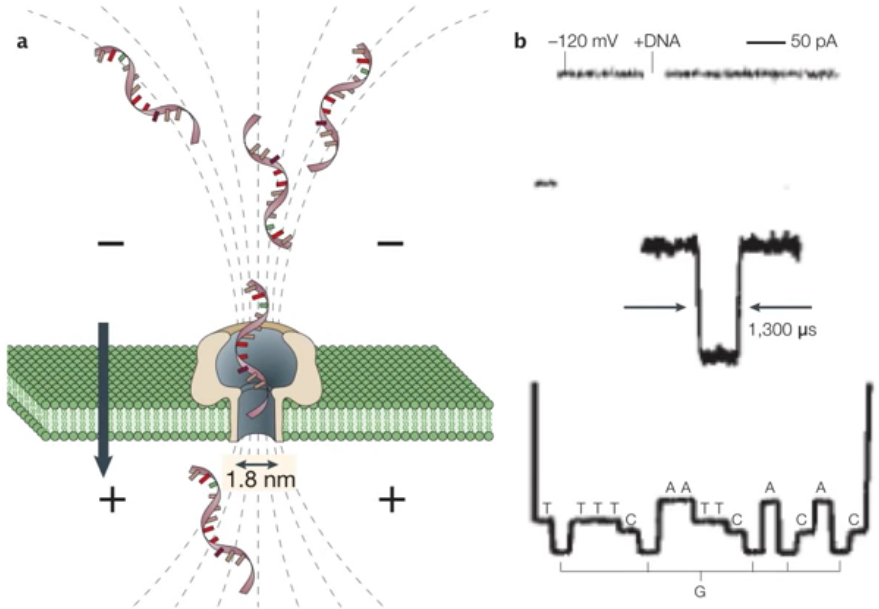
MinION Mk1C

PromethION 24/48

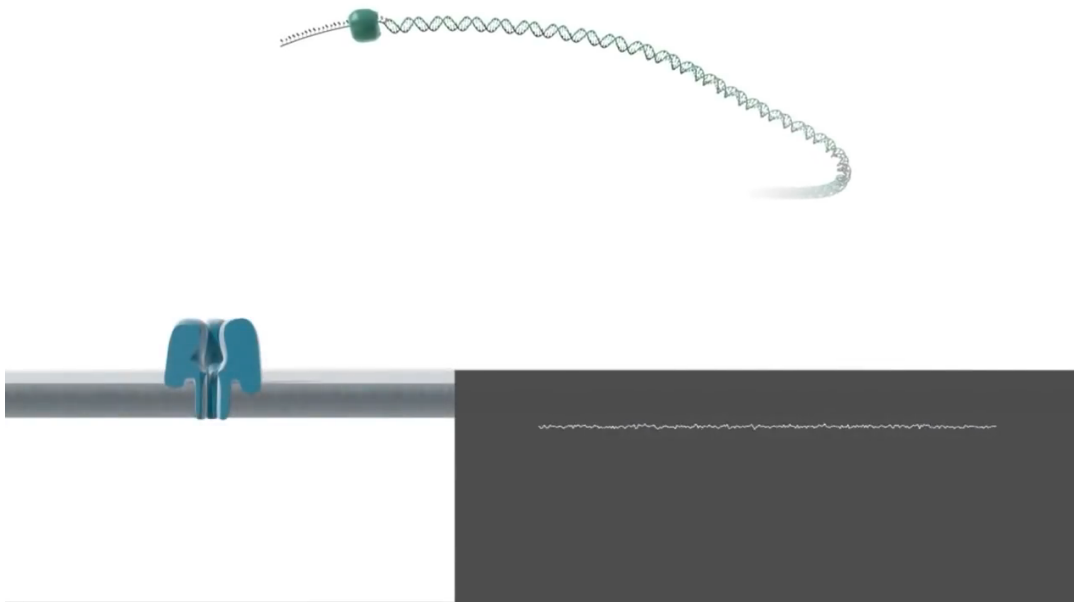
ONT: Platform Comparison

	MinION	GridION	PromethION
Flow cells	1	5	24 (48)
Real-time base calling	Needs MinIT	Yes	Yes
Channels	512	5 x 512	24 (48) x 3,000
Yield per flow cell	15-30 Gb	15-30 Gb	100-180 Gb
Yield per device	15-30 Gb	75-100 Gb	2.4-8.6 Tb

ONT: Sequencing Through Pore

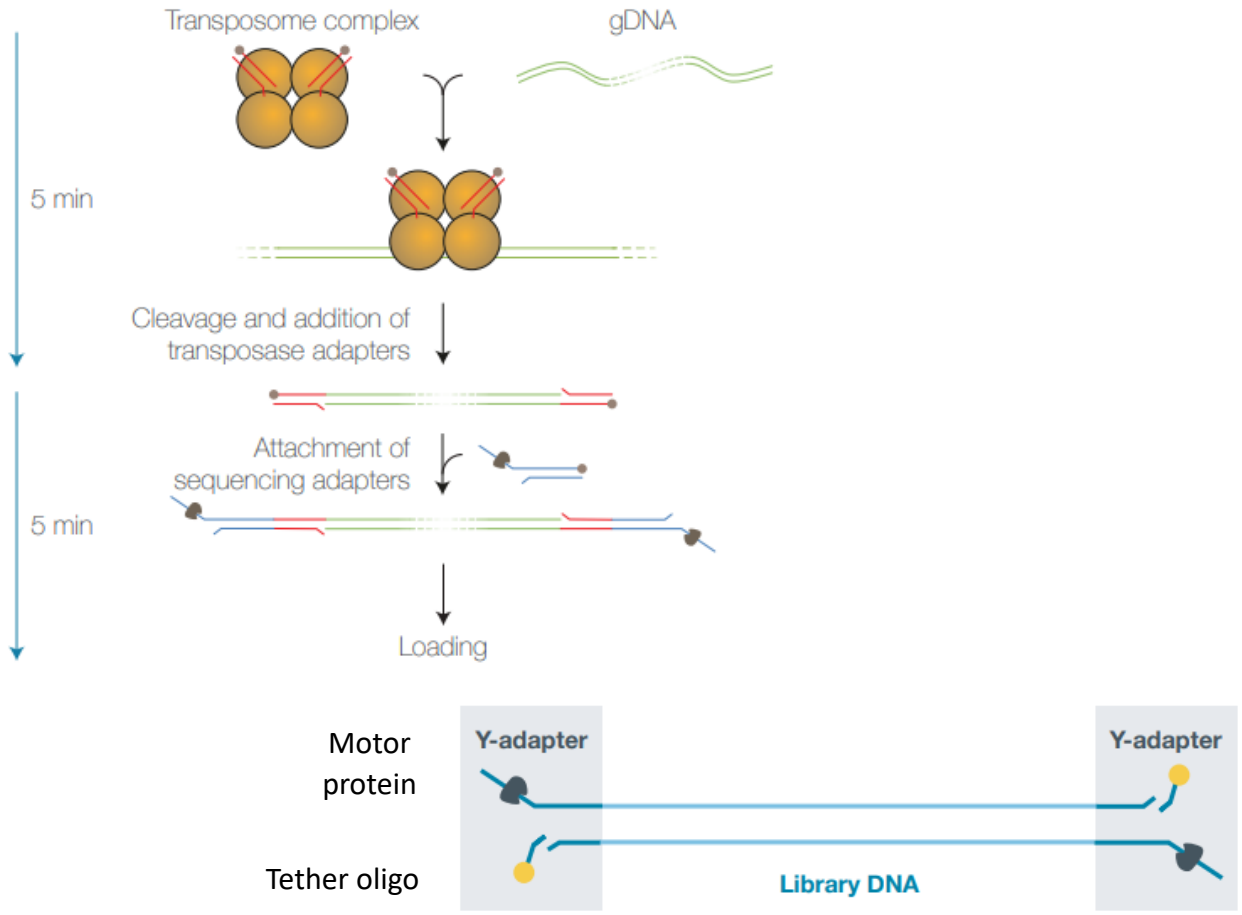


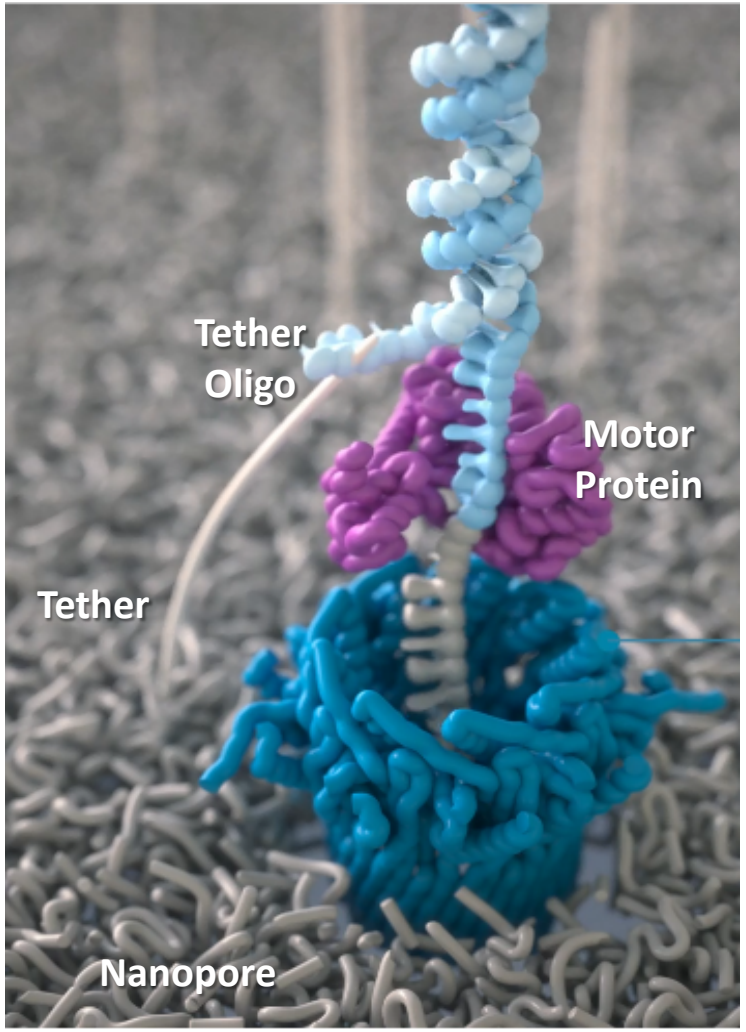
Nature Reviews Drug Discovery **volume 1**, 77-84 (2002)



Video created by Oxford Nanopore Technologies

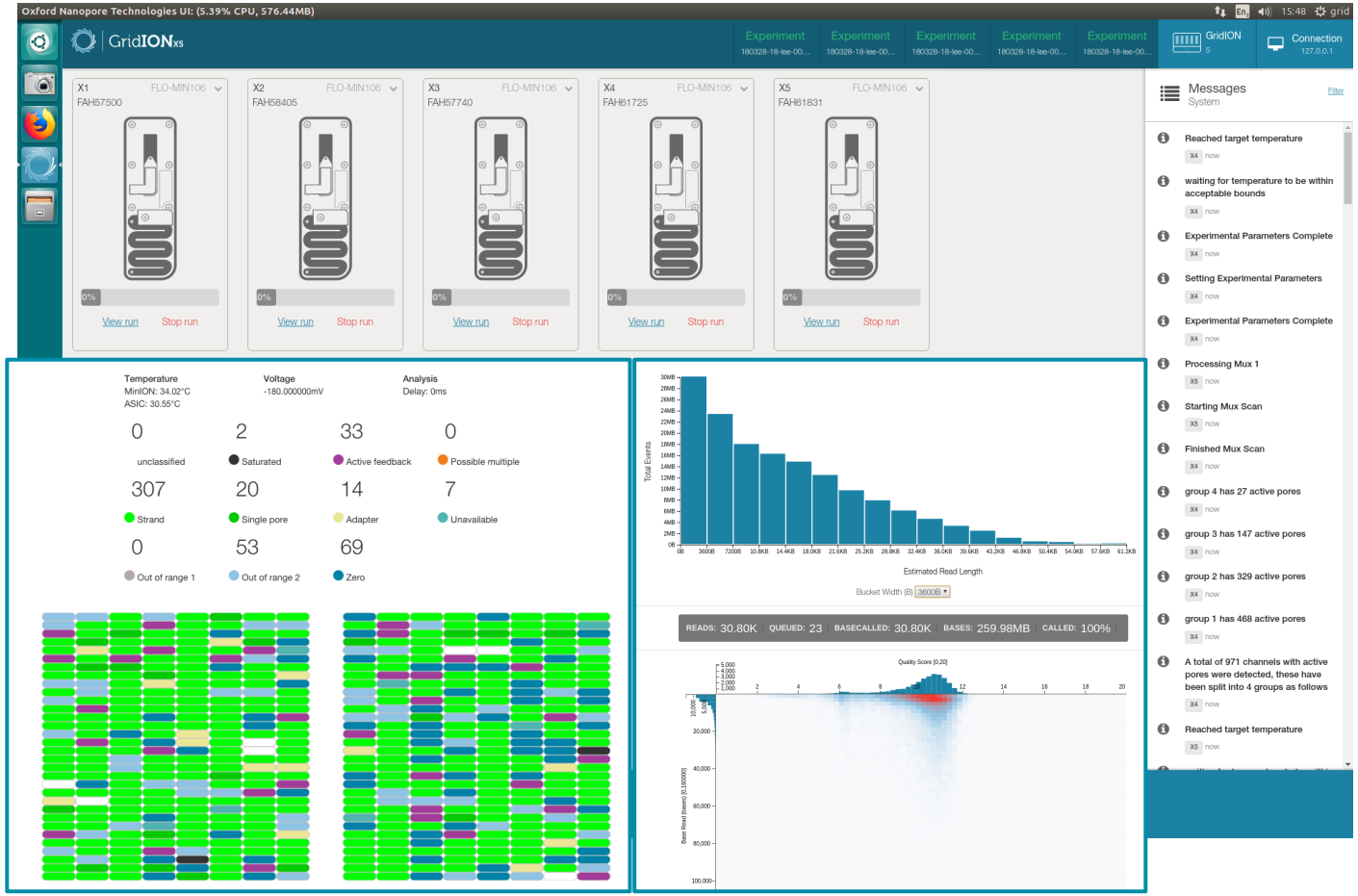
ONT: Library Structure Enables Sequencing





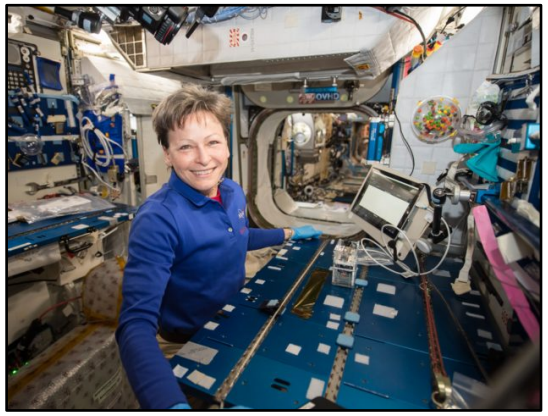
Images created by Oxford Nanopore Technologies
Modified by JAX

ONT: Sequencing



ONT: Sequencing in the Field

Sequencing microbial DNA in Antarctica



NASA astronaut sequencing bacteria in space

Sequencing Ebola virus in Guinea

Sequencing Zika virus in Brazil

LONG-READ SEQUENCING WORKSHOP

Location: **The Jackson Laboratory for Genomic Medicine - JAX Genomic Medicine, Farmington CT**

Long-read sequencing is rapidly becoming the standard in genomics, for assembling genomes, identifying structural variants, sequencing through repetitive regions, and phasing critical variants. Through this 3-day workshop, participants, including graduate students, post-doctoral fellows, and faculty, will learn about the technology and molecular biology driving each sequencing platform, including those from Pacific Biosciences, 10X Genomics, and Oxford Nanopore. Expert users, developers and representatives from long-read sequencing companies will share information about their applications in basic and translational genomic science, and will explore the commonalities and differences in long-read sequencing technologies.

Participants will have the opportunity to tour The Jackson Laboratory's Sequencing Center and network with genomics experts and industry leaders at the Welcome Reception!

Registration is now closed.

Please contact the event organizer for more information

7:00 am Central Daylight Time

APR

23 - 25

2018

Twitter

JUMP TO > ABOUT SCHEDULE SPEAKERS TRAVEL INFORMATION CONTACT SUPPORTERS

Long Read Sequencing Meeting at Jackson Labs

Break

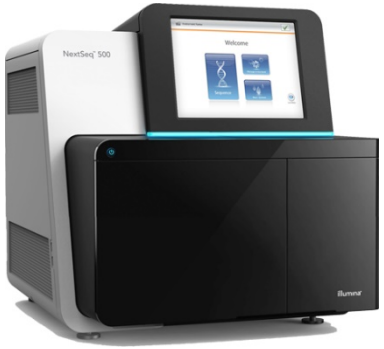
Illumina Sequencing

https://www.youtube.com/watch?v=fCd6B5HRaZ8

Illumina: Short Read Sequencing Platforms

MiSeq
$99K

NextSeq
$275K



HiSeq
$700K

NovaSeq
$985K

Comparison of Illumina Platforms:
https://www.illumina.com/systems/sequencing-platforms.html

Source: Illumina

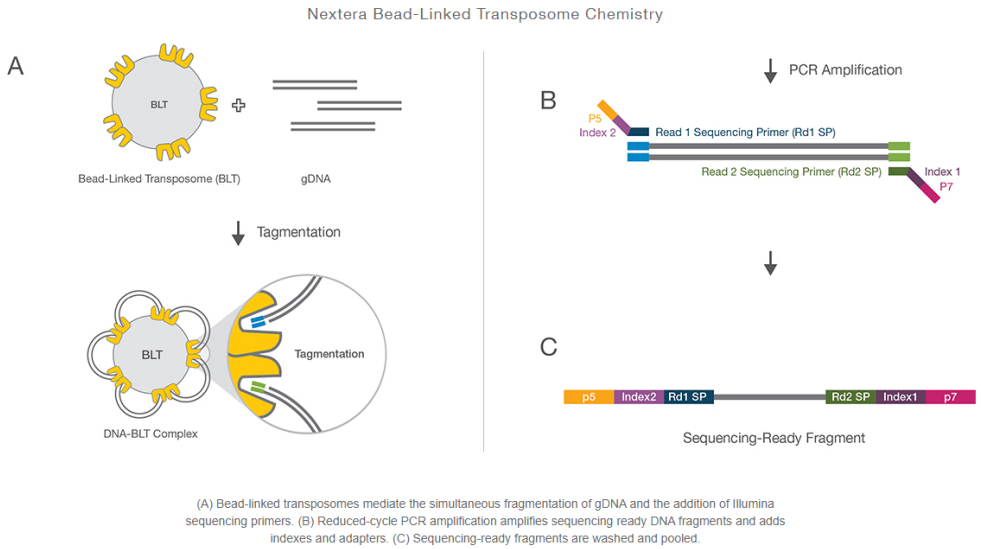
Illumina: Short Read Sequencing Platforms

- MiSeq:
  - Longest reads at 300bp
  - Up to 50 M
  - Good for amplicon sequencing, metagenomics…
  - Handles low diversity samples best!
- NextSeq and HiSeq are widely used and suitable for any short-reads application.
- NextSeq/NovaSeq use 2-channel chemistry and MiSeq/HiSeq use 4-channel
- NovaSeq
  - Generates the most amount of data, up to 20 B.
  - Cost per base is lowest.
  - Can handle most short read applications but best for projects that require lot of reads. For instance, whole genome sequencing or single cell transcriptome projects.

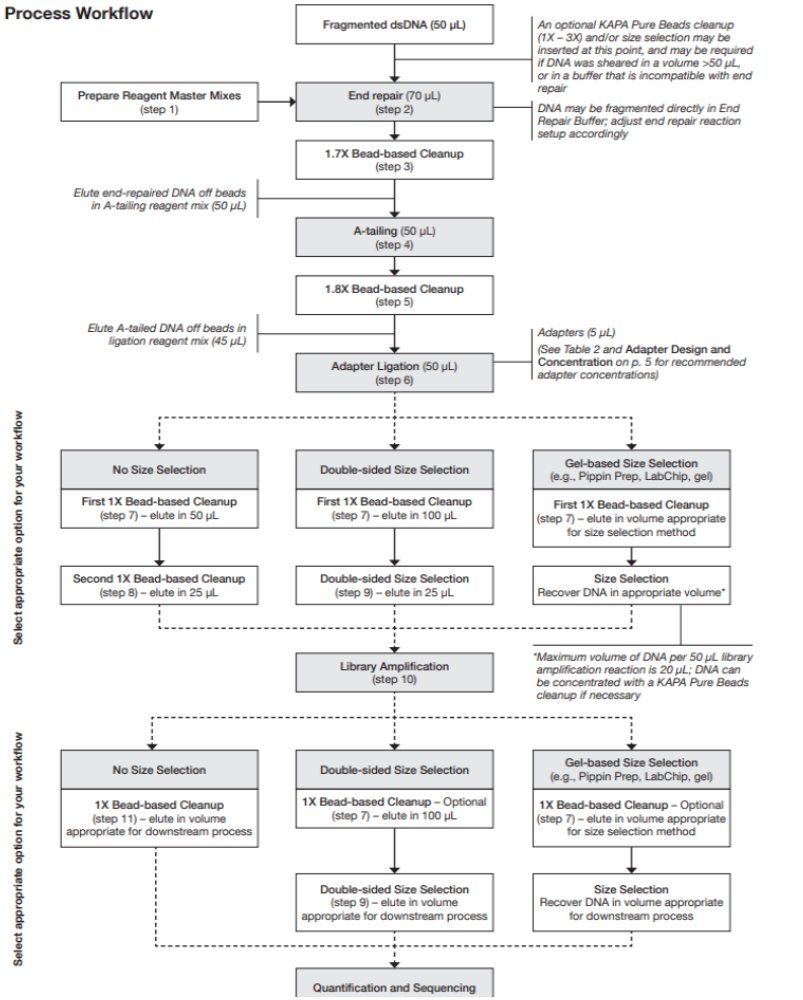
DNA-Seq Library construction

At Stowers, we do two different kinds of DNA-Seq library constructions;

1. Nextera DNA Flex Library Prep Kit - Illumina
2. KAPA HTP Library Preparation Kit Illumina® Platforms – Roche
   - Also used for ChIP-Seq



(A) Bead-linked transposomes mediate the simultaneous fragmentation of gDNA and the addition of Illumina sequencing primers. (B) Reduced-cycle PCR amplification amplifies sequencing ready DNA fragments and adds indexes and adapters. (C) Sequencing-ready fragments are washed and pooled.



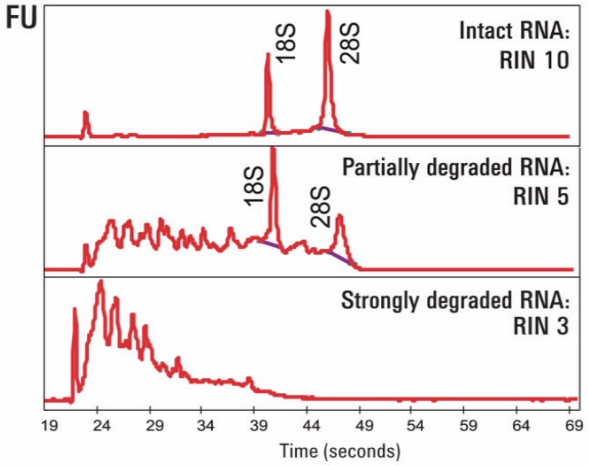
KAPA HTP	Nextera Flex
Can take up to two days	Half a day
Shearing required	No shearing
Size selection may be required	Size selection is not necessary
Customizable protocol (For instance: starting amounts, fragment size selection)	Less flexible protocol
More costly than Nextera Flex	Less costly than KAPA HTP

DNA-Seq Library construction: KAPA HTP vs. Nextera Flex

RNA: Quality is Critical

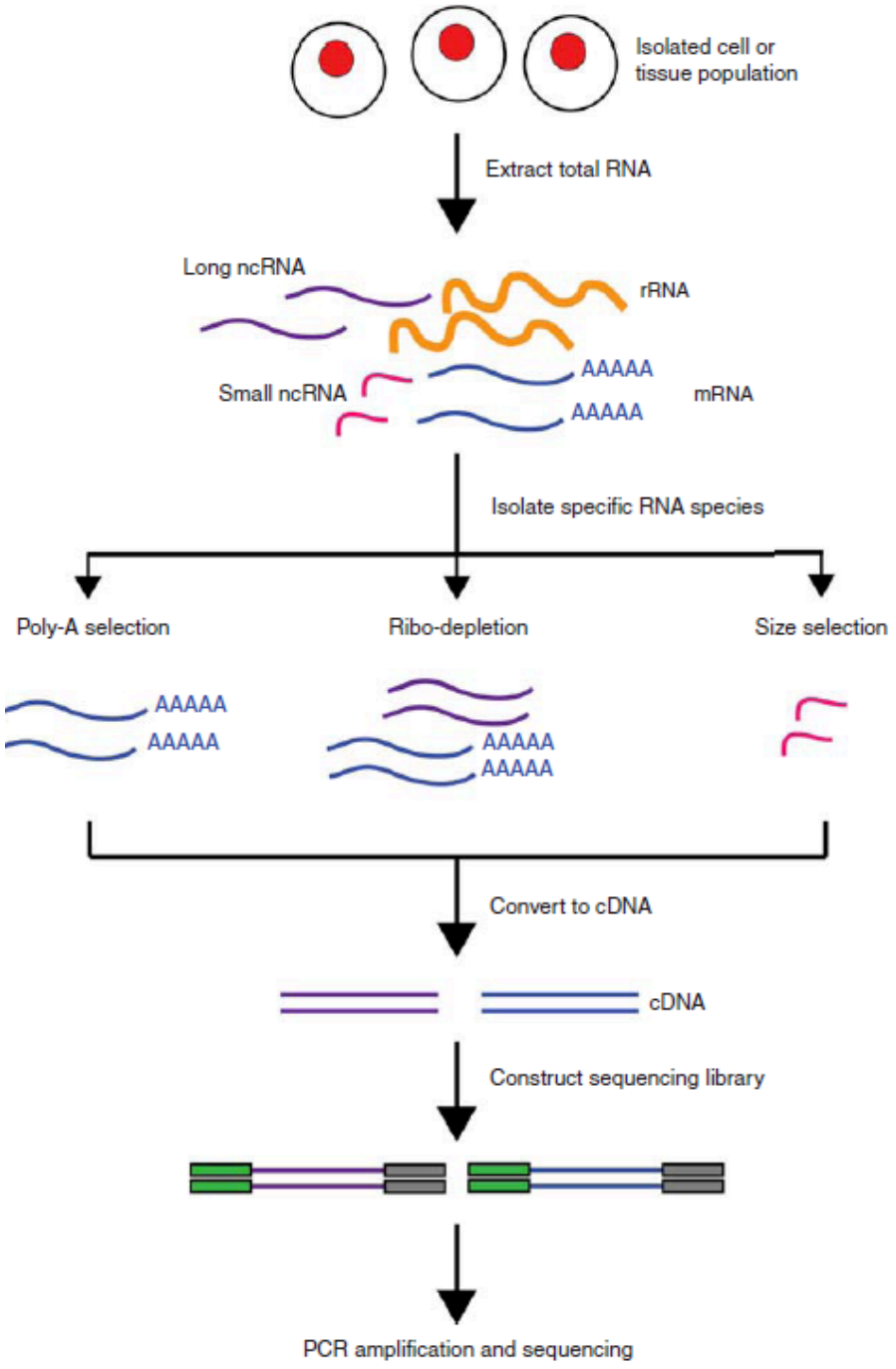
- The Bioanalyzer system provides a RIN (RNA Integrity Number) value for RNA quality ranging from 10 (highly intact RNA) to 1 (completely degraded RNA).
- RNase degradation can be detected by:
  - Decreasing ratio of ribosomal bands
  - Additional peaks below the ribosomal bands
  - Decrease in overall RNA signal
  - Shift towards shorter fragments
- To maintain RNA quality:
  - Avoid Freeze-thaw.
  - Store samples at -80C and transport samples on dry ice.
- Direct-zol from Zymo Research, and many other good kits for RNA extraction

Agilent Bioanalyzer System



RNA Library Construction: Poly A vs. Ribodepletion vs. smRNA-Seq

- Always have replicates! At least three biological replicates.
- If RNA is degraded and there is nothing you can do about it, we recommend doing a ribodepletion protocol with oligos opposed to poly-A selection.
- Some RNA extraction kits remove small RNAs.
- Pick a library construction kit that is suitable for your project.
- Limit technical variables such as kits used, sequencing platform etc.



Takara Clontech SMARTer v4 followed by Illumina Nextera XT for 1 cell to 1000 cells or 10pg to 10ng

If the starting material is between 10 ng to 1ug, NEB Ultra works well

For samples with 100ng and up, most kits including ones from Roche (KAPA), Illumina, NEB will work just fine

RNA Library Construction: The kits Used will Depend on the Starting Material

Library QC

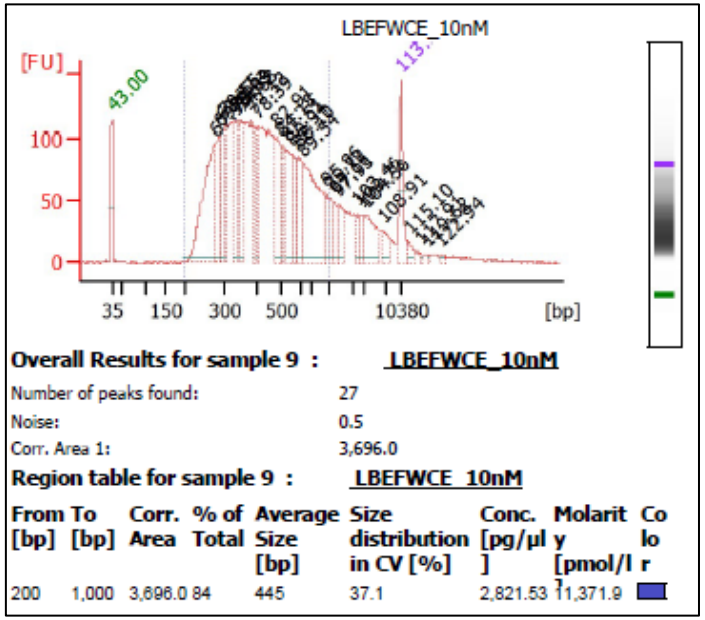
- Quality of libraries is a good indication of how your data will look.
- qPCR will help you quantify fragments that are “sequence-able”
- Quantity assessment via Invitrogen Qubit Fluorometer
- Agilent Bioanalyzer to determine quality; size of fragments, presence of adapter dimer, overamplification etc.

Invitrogen Qubit Fluorometer

Agilent Bioanalyzer System

Good vs. Ugly

A



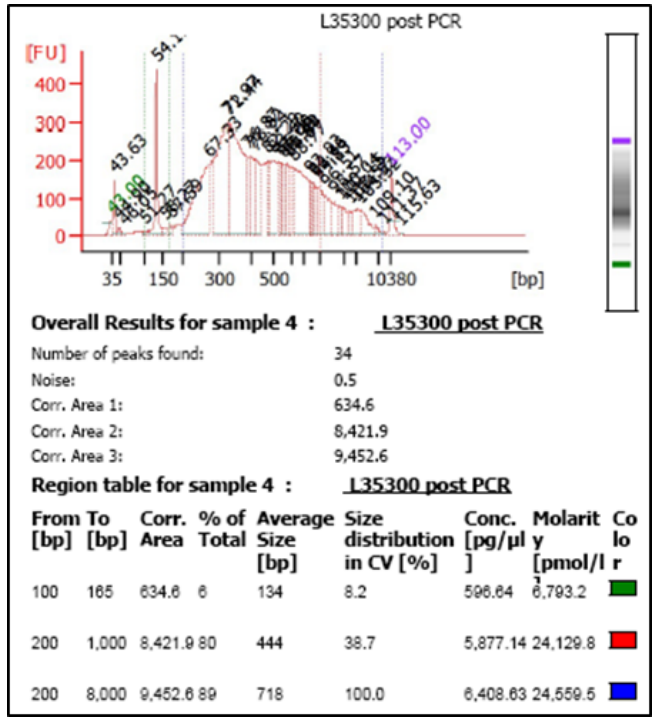
Overall Results for sample 9 : **LBEFWCE_10nM**

Number of peaks found:	27
Noise:	0.5
Corr. Area 1:	3,696.0

Region table for sample 9 : **LBEFWCE_10nM**

From [bp]	To [bp]	Corr. Area	% of Total	Average Size [bp]	Size distribution in CV [%]	Conc. [pg/µl]	Molarity [pmol/l]	Color
200	1,000	3,696.0	84	445	37.1	2,821.53	11,371.9	

B



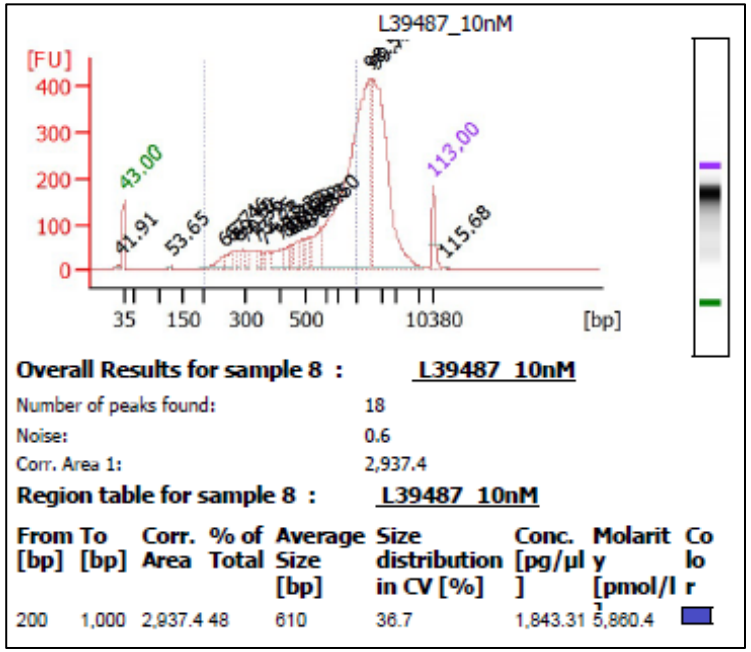
Overall Results for sample 4 : **L35300 post PCR**

Number of peaks found:	34
Noise:	0.5
Corr. Area 1:	634.6
Corr. Area 2:	8,421.9
Corr. Area 3:	9,452.6

Region table for sample 4 : **L35300 post PCR**

From [bp]	To [bp]	Corr. Area	% of Total	Average Size [bp]	Size distribution in CV [%]	Conc. [pg/µl]	Molarity [pmol/l]	Color
100	165	634.6	6	134	8.2	596.64	6,793.2	
200	1,000	8,421.9	80	444	38.7	5,877.14	24,129.8	
200	8,000	9,452.6	89	718	100.0	6,408.63	24,559.5	

C



Overall Results for sample 8 : **L39487_10nM**

Number of peaks found:	18
Noise:	0.6
Corr. Area 1:	2,937.4

Region table for sample 8 : **L39487_10nM**

From [bp]	To [bp]	Corr. Area	% of Total	Average Size [bp]	Size distribution in CV [%]	Conc. [pg/µl]	Molarity [pmol/l]	Color
200	1,000	2,937.4	48	610	36.7	1,843.31	5,860.4	

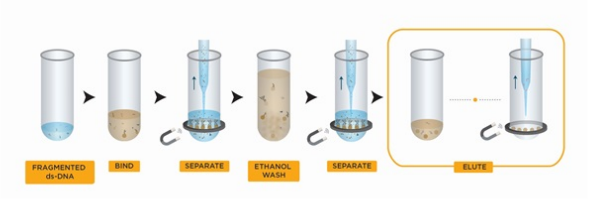
Normal:
Technical
Control

Adapter Dimer

Larger Fragments,
Unsuitable for Illumina
Sequencing

Library Size Selection

Sage Science's Blue Pippin/Pippin Prep

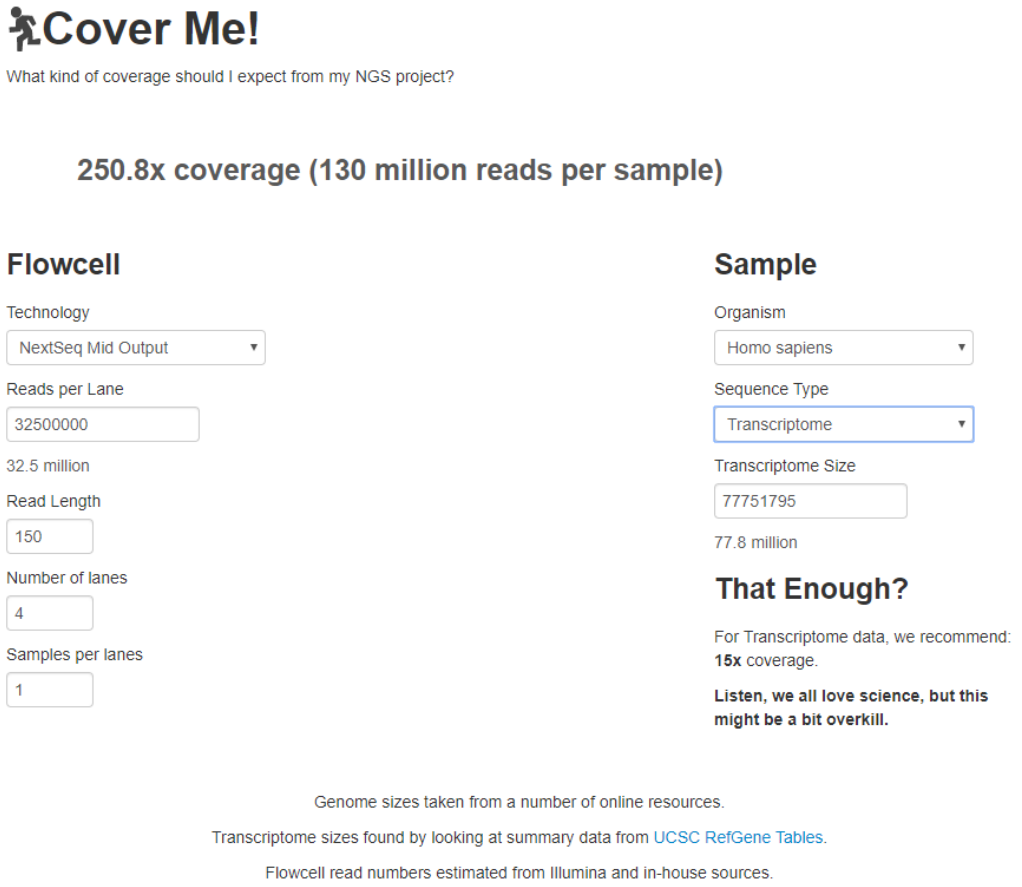


1. Bind DNA to magnetic beads
2. Separate beads from contaminants
3. Wash magnetic beads with 85% ethanol to remove contaminants
4. Elute DNA from magnetic beads
5. Transfer to a storage plate

Beckman Coulter's SPRI Select beads

- Two popular methods of size selection:
  - Sage Sciences' automated instruments
    - Easy to use
    - Costly
    - Specific
  - Bead based size selection
    - Requires optimization
    - Less costly
    - Good for large projects
    - Can be automated

How Much to Sequence?

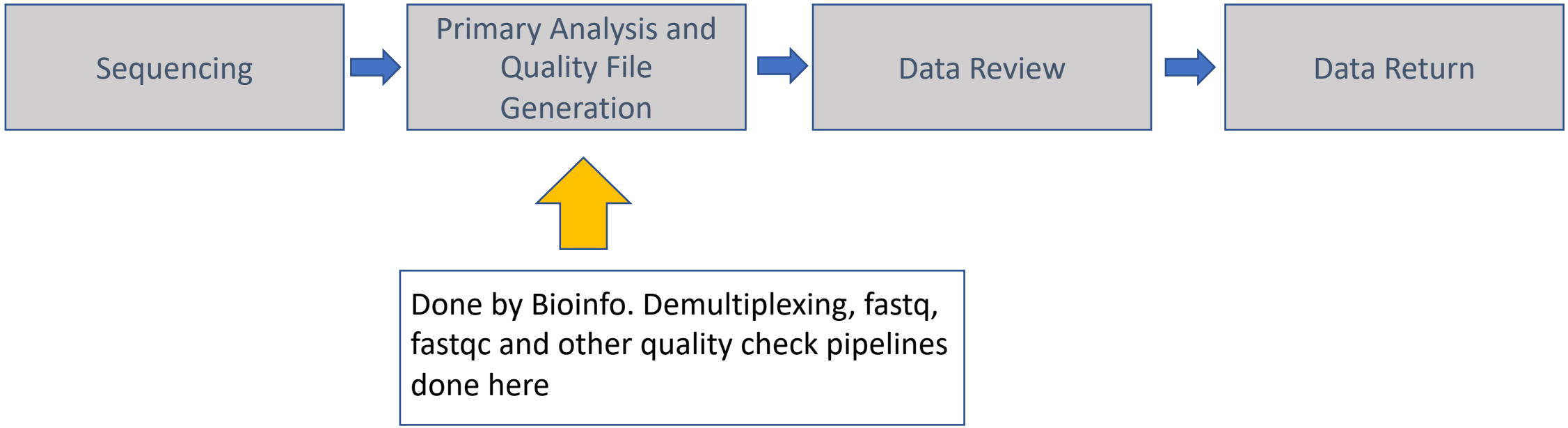


Will depend on:

- Project goal
- Organism
- Machine Used
- Number of libraries

http://metalhelix.github.io/coverme/

NGS Workflow of the Core



Primary Analysis and Quality Checks

- Read counts per index/library
- Index accuracy
- Alignment
- contaminations
- FastQC
  - Quality of reads
  - Contamination
- Other (RNA-specific)

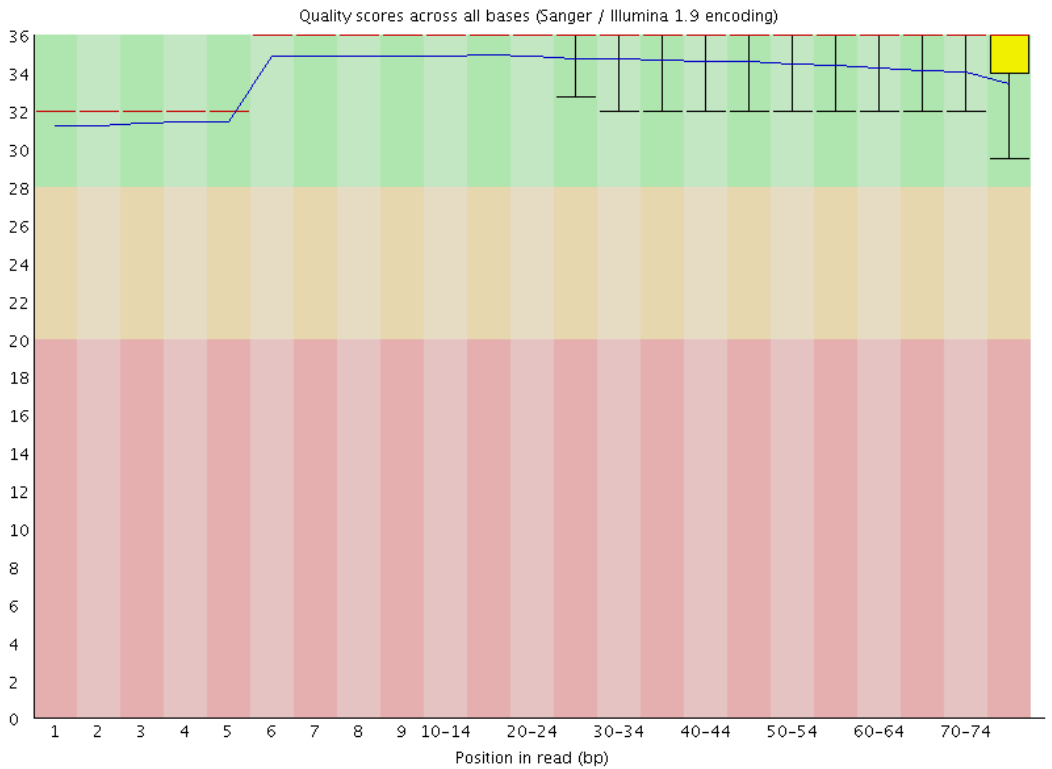
Read Counts per Library

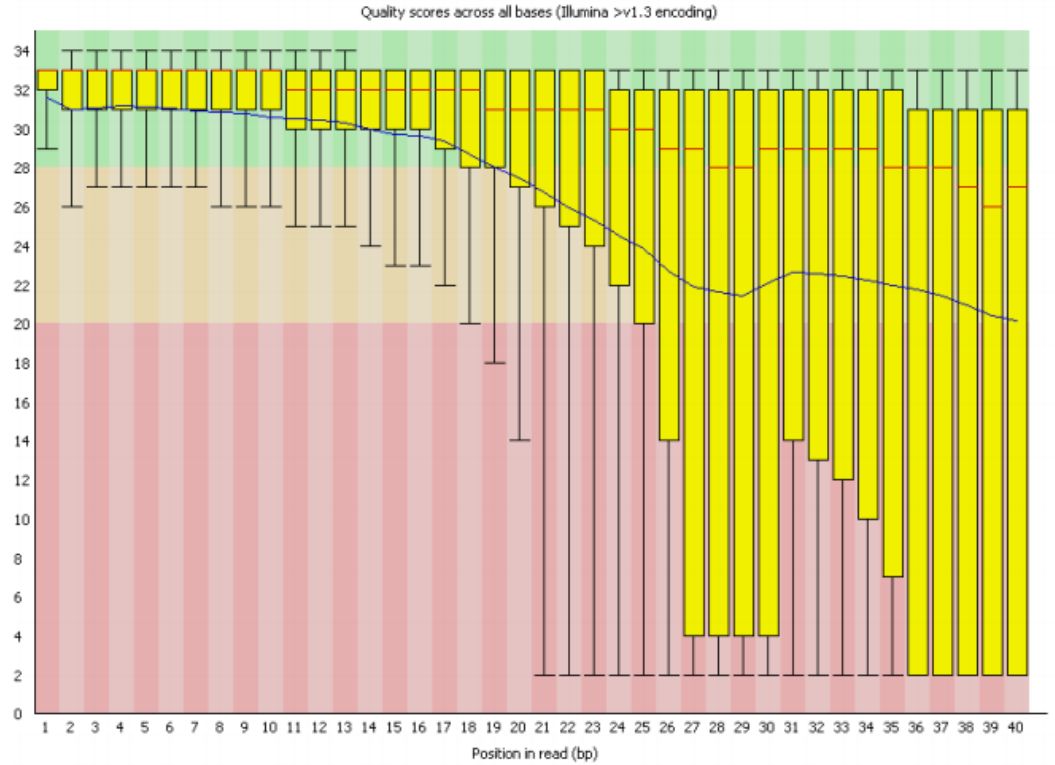
Order Type	Lane	Read No	Sample ID	Sample Name	Cluster Cou...	% Align PF	Index Sequence
poly-A Stranded RNA-Seq	1	1	S39858	V6.5_ES_epiLC_24hr_rep1	20600098	84.75	GATCAG
poly-A Stranded RNA-Seq	1	1	S39857	V6.5_ES_2i_rep2	21669575	85.55	ACAGTG
poly-A Stranded RNA-Seq	1	1	S39861	V6.5_ES_epiLC_48hr_rep2	[illegible]	[illegible]	[illegible]
poly-A Stranded RNA-Seq	1	1	S39856	V6.5_ES_2i_rep1	16414815	85.72	TTAGGC
poly-A Stranded RNA-Seq	1	1	S39860	V6.5_ES_epiLC_48hr_rep1	[illegible]	[illegible]	[illegible]
poly-A Stranded RNA-Seq	1	1	S39859	V6.5_ES_epiLC_24hr_rep2	21083786	84.79	CTTGTA

Alignments per Library

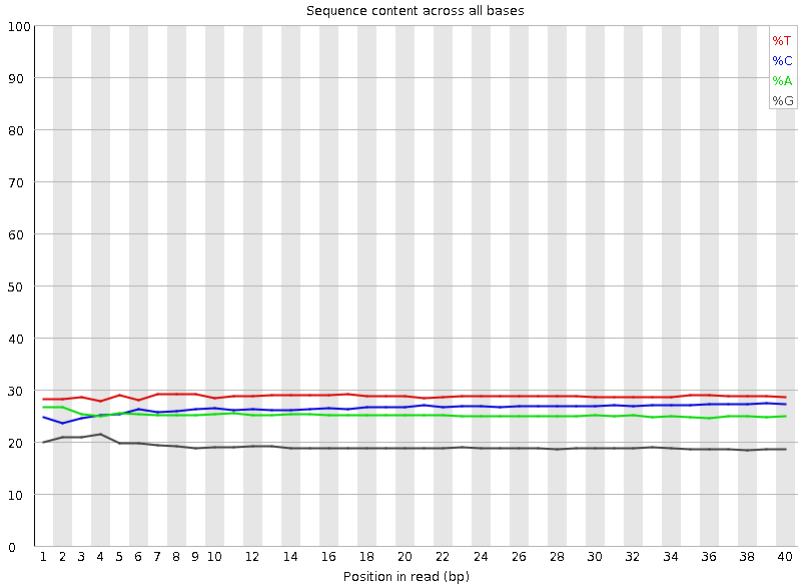
Order Type	Lane	Read No	Sample ID	Sample Name	Cluster Cou...	% Align PF	ndex Sequence
poly-A Stranded RNA-Seq	1	1	S39858	V6.5_ES_epiLC_24hr_rep1	20600098	84.75	GATCAG
poly-A Stranded RNA-Seq	1	1	S39857	V6.5_ES_2i_rep2	21669575	85.55	ACAGTG
poly-A Stranded RNA-Seq	1	1	S39861	V6.5_ES_epiLC_48hr_rep2	22341206	84.82	ATTCCT
poly-A Stranded RNA-Seq	1	1	S39856	V6.5_ES_2i_rep1	16414815	85.72	TTAGGC
poly-A Stranded RNA-Seq	1	1	S39860	V6.5_ES_epiLC_48hr_rep1	21799517	84.71	CGTACG
poly-A Stranded RNA-Seq	1	1	S39859	V6.5_ES_epiLC_24hr_rep2	21083786	84.79	CTTGTA

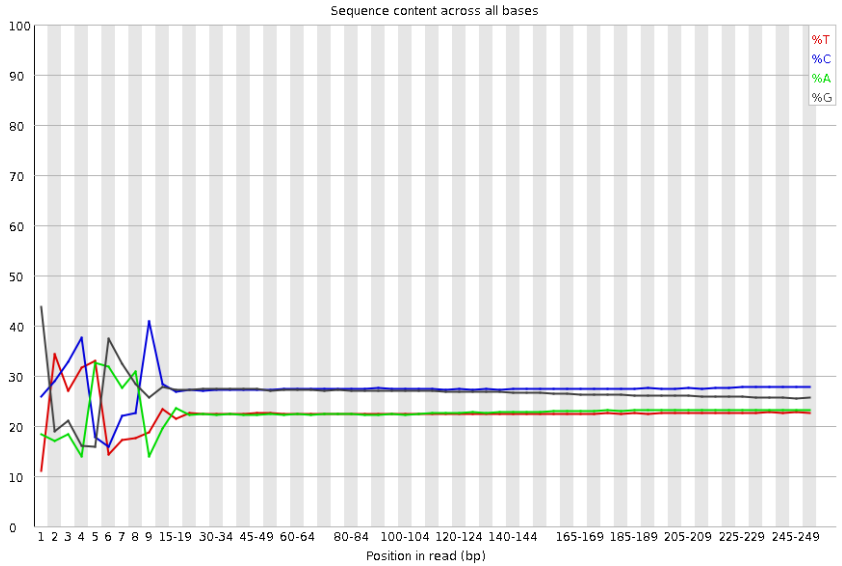
FastQC: Per Base Sequence Quality

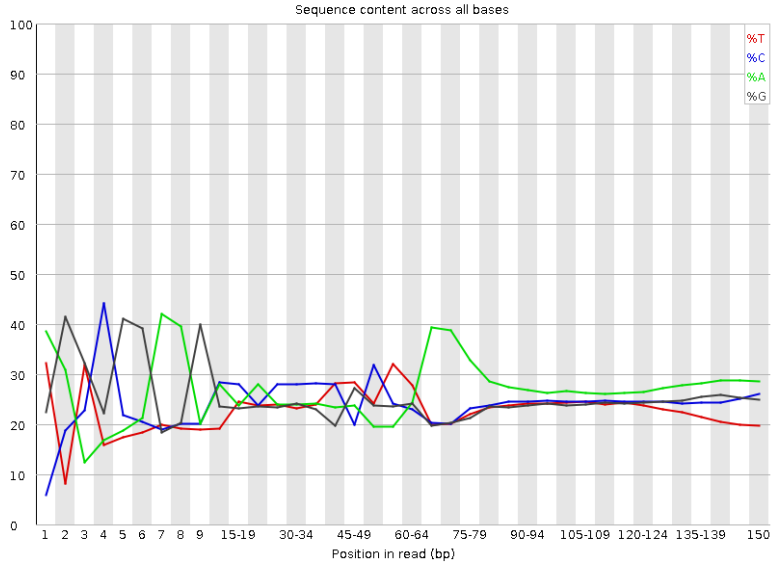




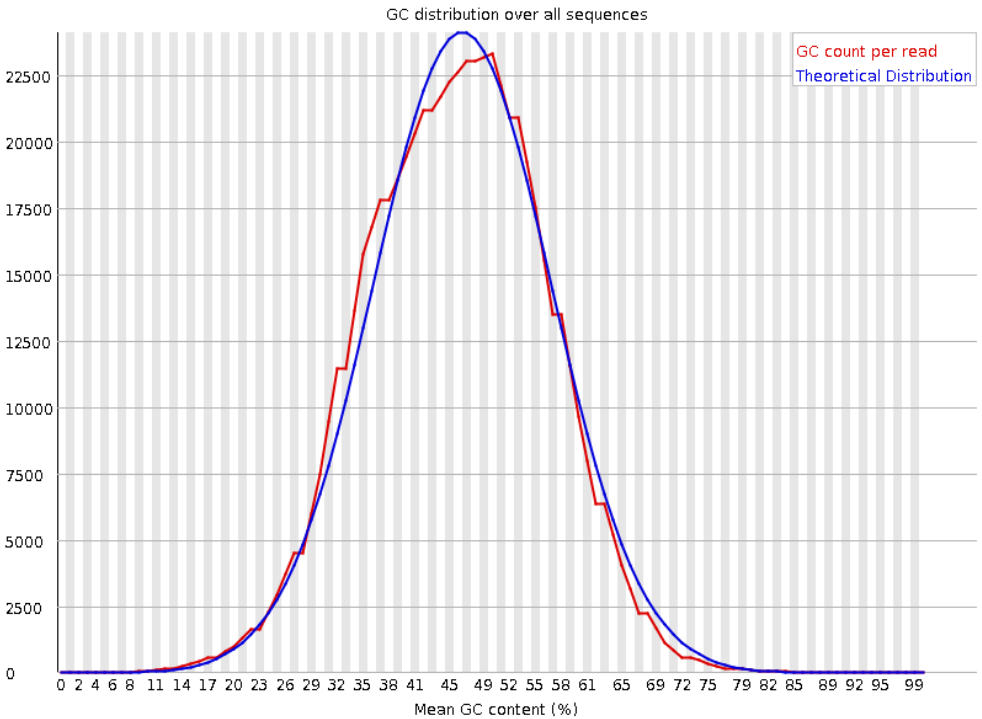
FastQC: Per Base Sequence Content

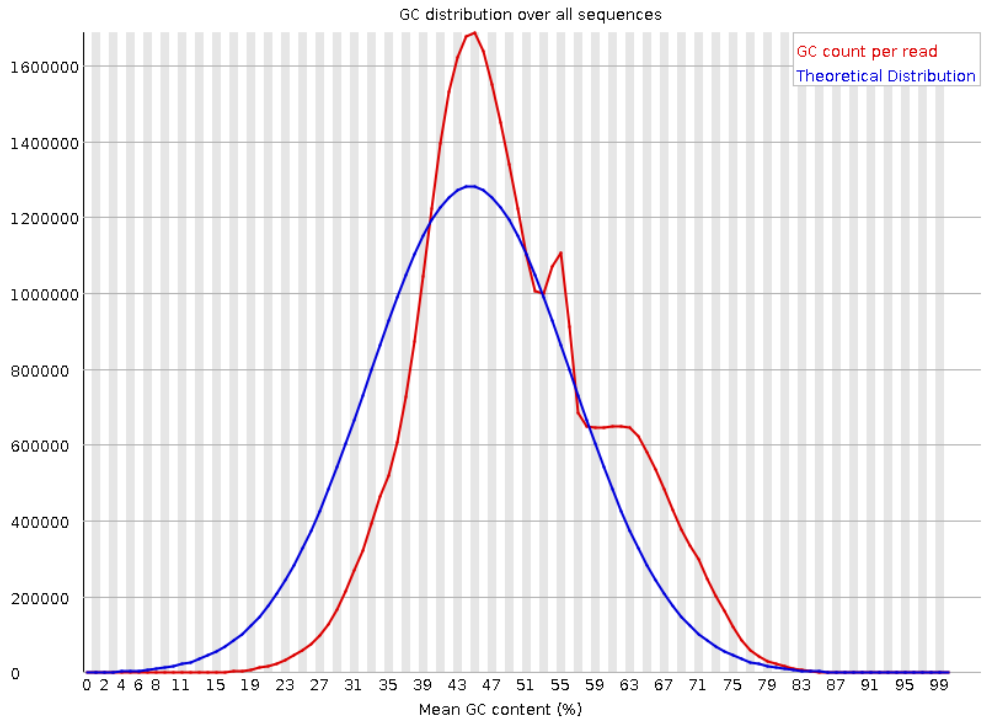




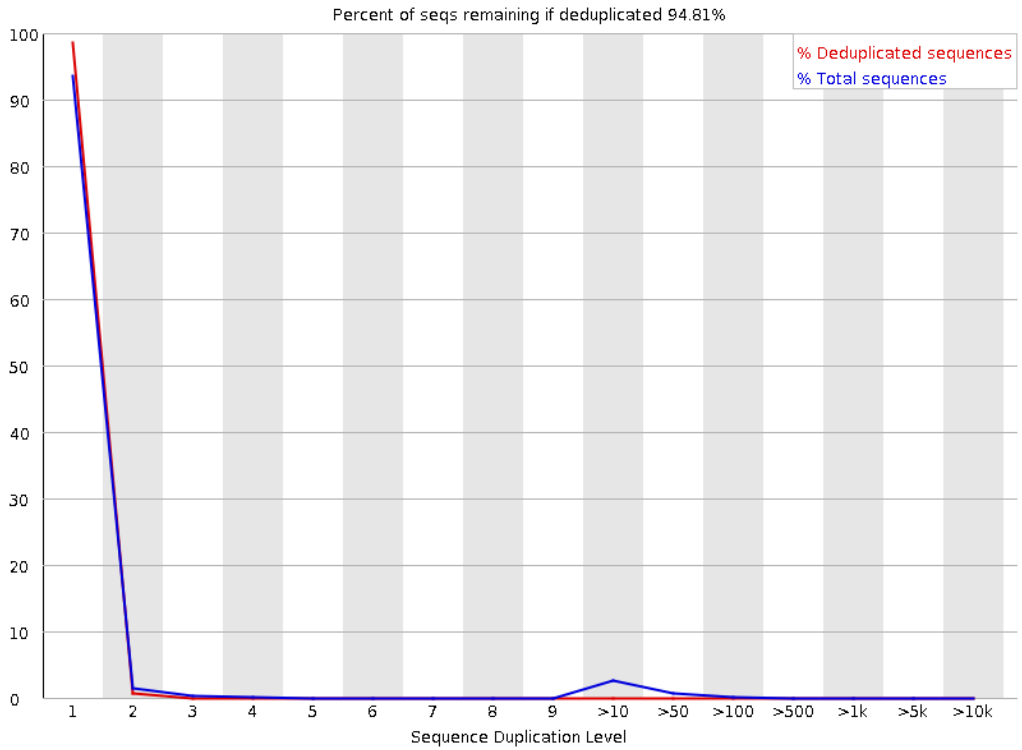


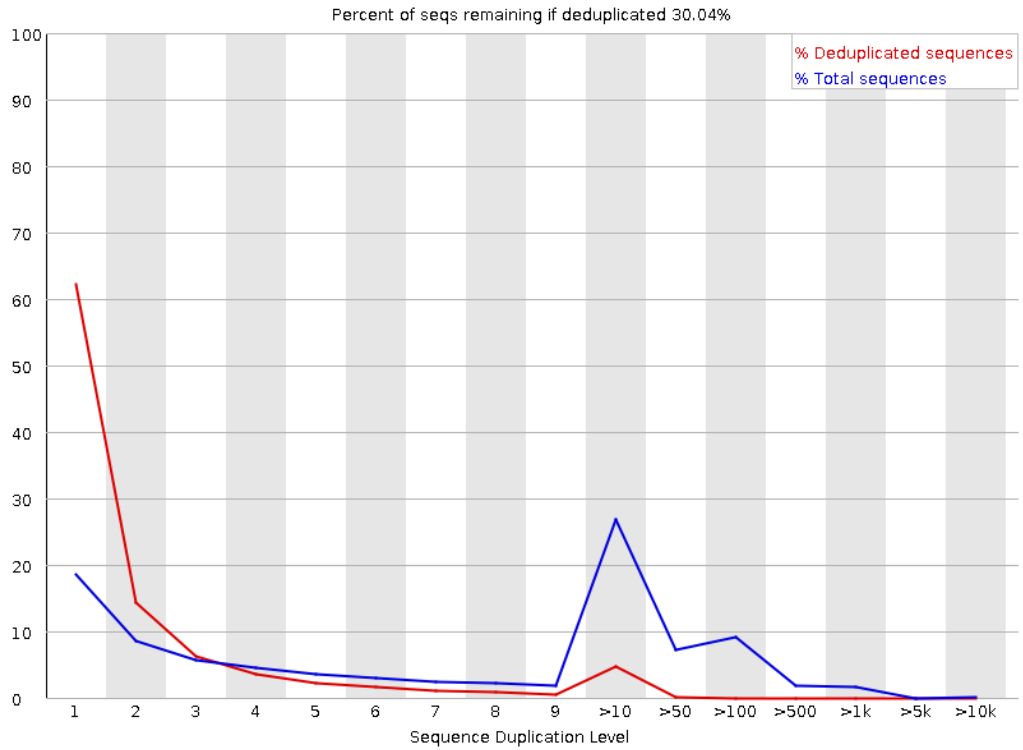
FastQC: Per Sequence GC Content

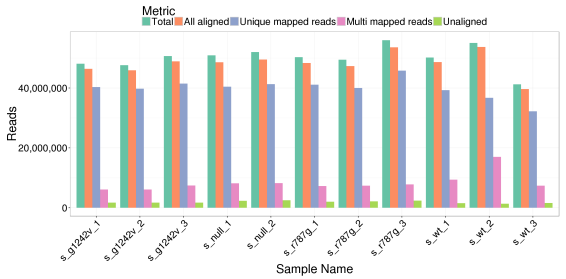




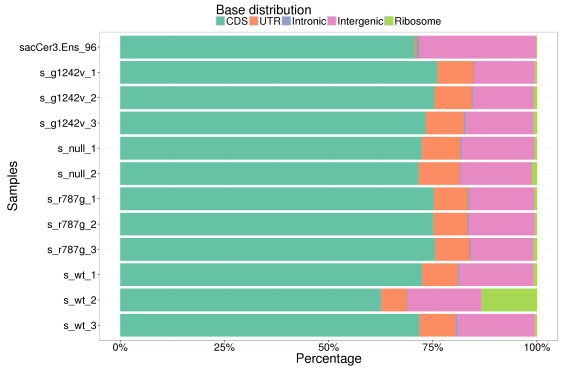
FastQC: Sequence Duplication Levels



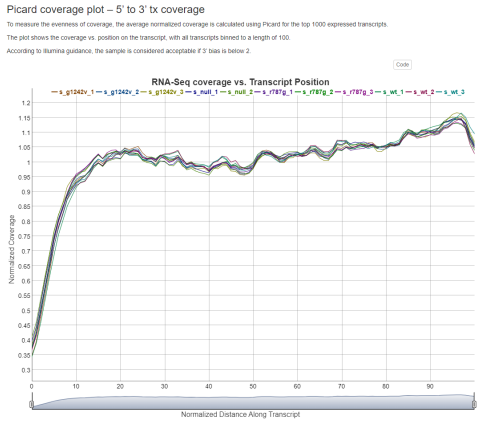




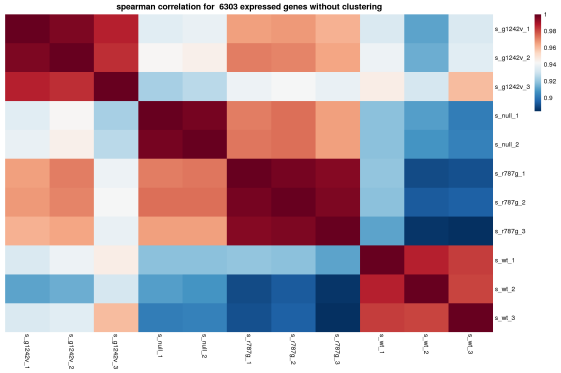
Overall mapping



Regions to which reads map



Evenness of coverage for transcripts



Heat map of sample clustering

Internal Pipeline: Secundo

Few Resources:

- SeqAnswers http://seqanswers.com/
- ABRF http://list.abrf.org/groups/abrf/
- Oxford Nanopore https://nanoporetech.com/community
- FastQC manual https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
- And ton of information on social media; twitter, YouTube