Best Practices for Large Insert Libraries and Sequel System Sequencing

PacBio UGM
June 27, 2017

Melissa Laird Smith, Ph.D.
Assistant Professor,
Assistant Director of Technology Development
Icahn School of Medicine at Mount Sinai
Dept. of Genetics & Genomics Sciences
Icahn Institute for Genomics & Multiscale Biology

Our Technology Development team & capabilities

Mission: To enable investigators to perform translational and interdisciplinary biomedical research using cutting edge genomics tools

Multi-Platform DNA Sequencing:

- 3 PacBio RS IIs
- 2 PacBio Sequels
- 8 Illumina HiSeq 2500/4000
- 2 MiSeq
- 8 Ion Protons
- 7 Ion S5XL
- 2 Ion PGM
- 11 Ion Chefs
- 1 10X Genomics Chromium
- 1 BioNano Genomics Irys
- 1 Applied Biosystems 3730xl
- ONT MinION

Single Cell Technology:

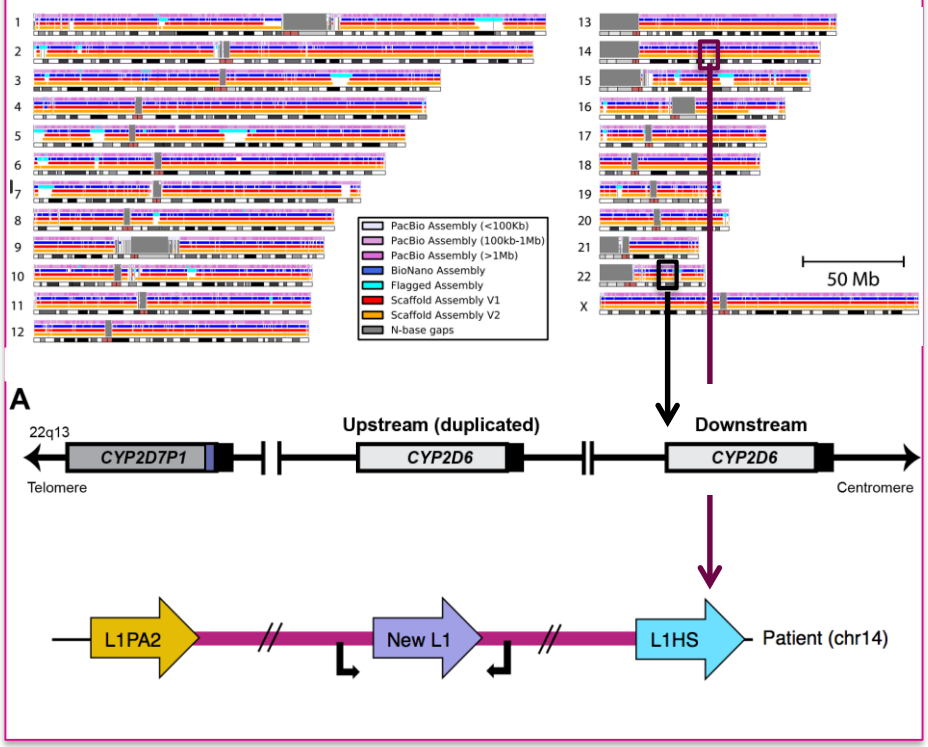
- 1 10X Genomics Chromium
- 1 BLI Beacon
- 1 BioRad/Illumina ddSeq

Using long read sequencing as a translational research tool

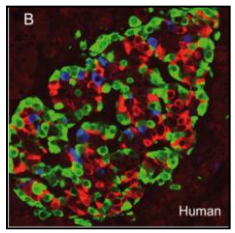
Technology Toolbox and Expertise:

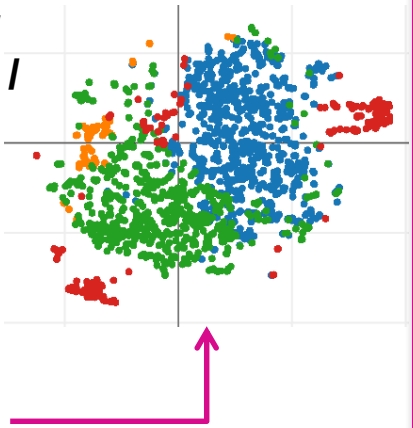
- 2 Sequel & 3 RSII sequencing systems with >3000 RSII & >250 Sequel SMRTcells in 2016
- 10X Chromium and BLI Beacon for unique single cell genomics
- BioNano Genomics optical mapping and scaffold generation expertise
- Novel data integration, spanning all technologies to inform R&D and novel diagnostics

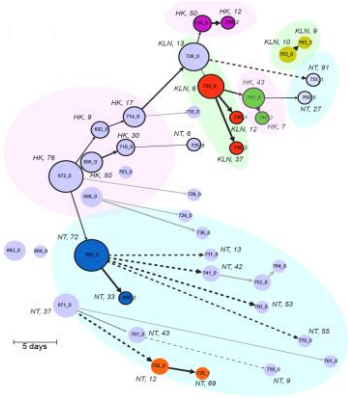
Addressing complex structure & targeting comprehensive diagnostics



Rapid validation of low freq variants in somatic / heterogeneous tissue







Infectious Disease & Immune Repertoire Profiling and Surveillance

Our Goal: To Address Structural Complexity by WGS with Comprehensive Long Read Sequencing Methods

Progress in Human SMRT-WGS Across 7 Genomes

(1) **NA12878 (P5-C3)**
~800 kb de novo PB-FLCN
>1000 SMRTcells
~ 1 year
~ 5kb subRL N50

(2-4) **AJ Trio – GIAB (P6-C4)**
~5 Mb de novo polished PB-FLCN
~200 SMRTcells for child
~ 4 months
~ 11kb subRL N50

(5) Afr-haplotype (P6-C4)
~4 Mb de novo PB-FLCN
~200 SMRTcells
~ 4 months
~ 13.5kb subRL N50

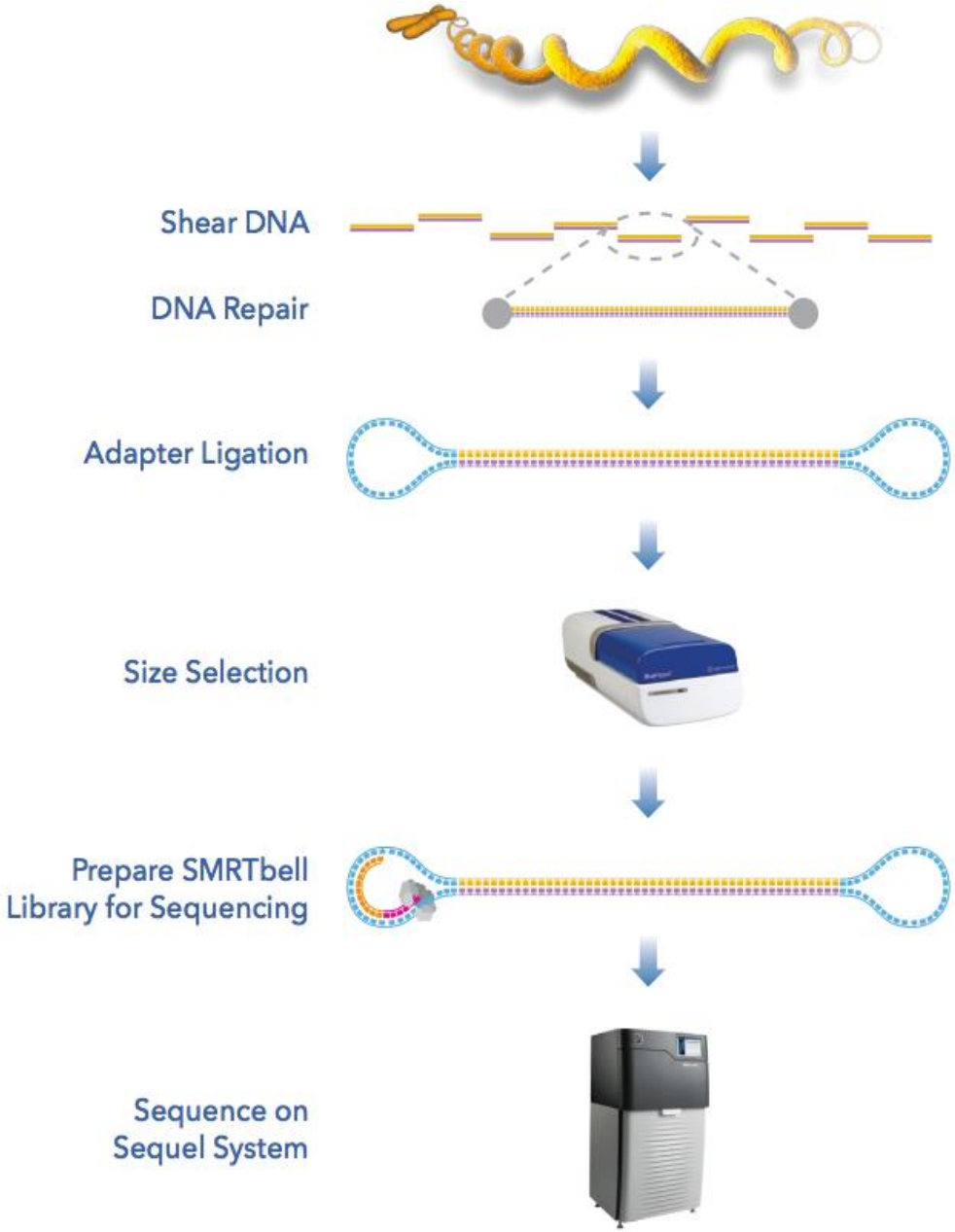
(6) Cau-3 haplotype (SQL-6hr)
~1 Mb* de novo PB-FLCN
~50 SMRTcells
~ weeks
~ 13 kb subRL N50

(7) Cau-haplotype (SQL-10 hr)
~2 Mb* de novo PB-FLCN
~50 SMRTcells
~ weeks
~ 17.5kb subRL N50

* Unpolished / 1st pass in FALCON v0.3
* 4.5 day job on cluster using ~178,000 compute hours (#7)

- From previous experience with BNG, can lift scaffold N50 to 22-32 Mb
- Current max contig sizes, even on pass #1 FALCON 0.3 with Sequel are ~10 Mb
- Polished max contigs up to ~20 Mb on RSII completed assemblies
- Speed and cost on Sequel greatly reduced for generating more haplotypes faster
- Additional large genomes prepped for Sequel – mammalian, plants & insects

Overview of Large Insert Library Prep Process (w/Notes)



Upfront QC: request at least 20 μg; clean (1X AMPure), BA, Qubit

Shear: Covaris g-tubes (20 kb shear)

Pre-library repair: DR, ER + exoVII always!

Library Prep: standard protocols

Size Selection: Based on quality & quantity of library, select 10-50 kb or 20-50 kb (BluePippin and/or PippinHT)

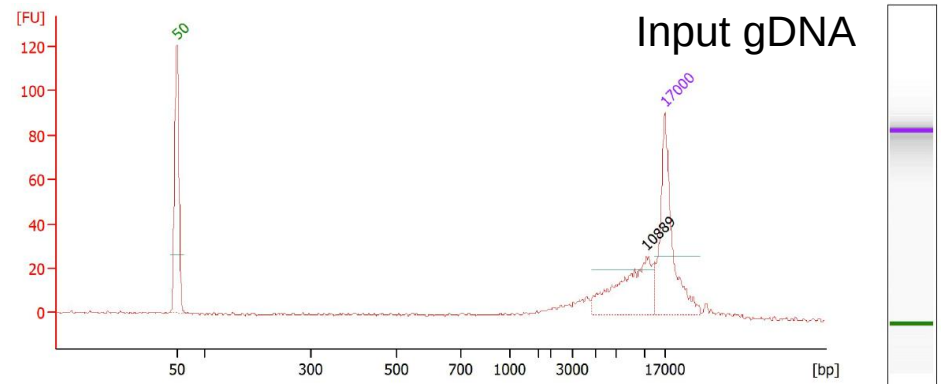
Annealing/Binding: Primer annealing time (30min); Pol binding time (4h)

Sequel: *Titrate* for large projects; 10h movies; 2.0 chemistry

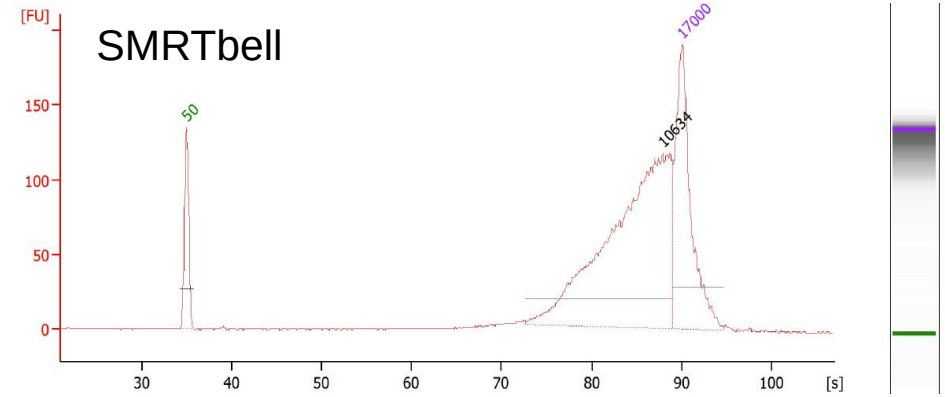
Sample Data Sets: Input, Library Prep and Size Selection

“Just Fine” sample

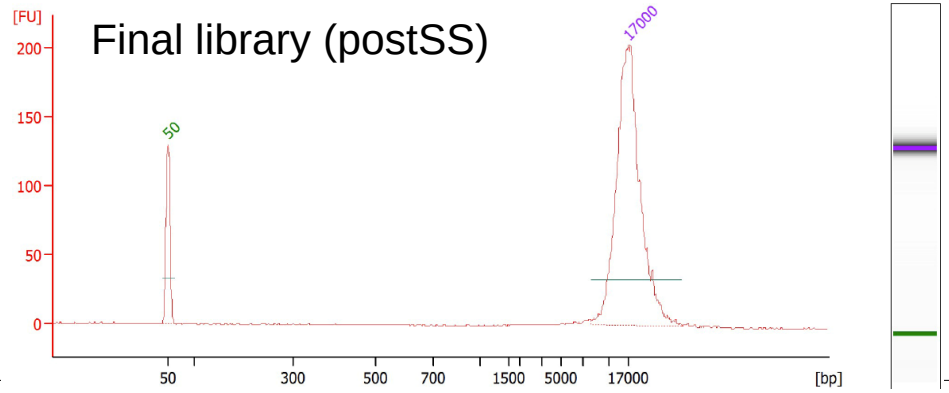
Input gDNA



SMRTbell

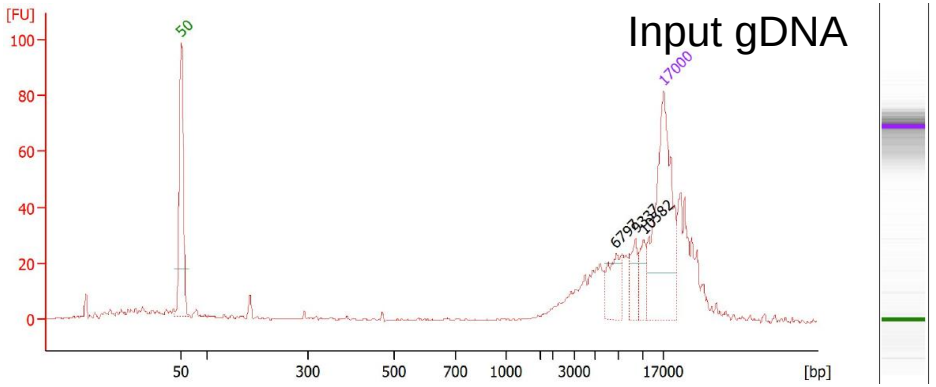


Final library (postSS)

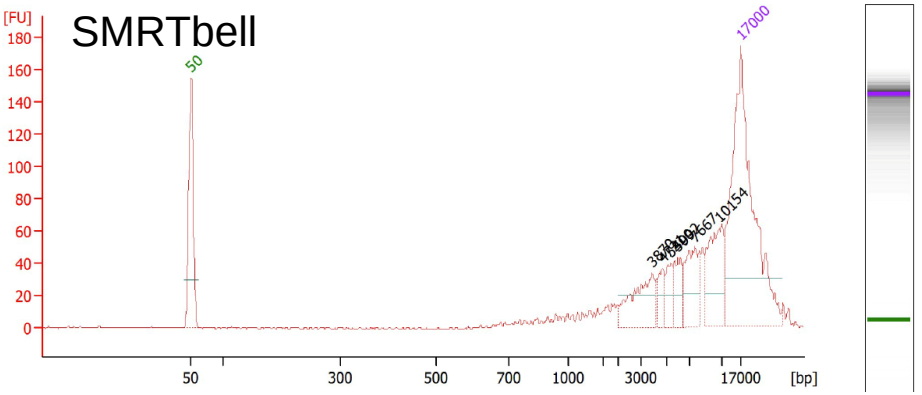


“Great” sample

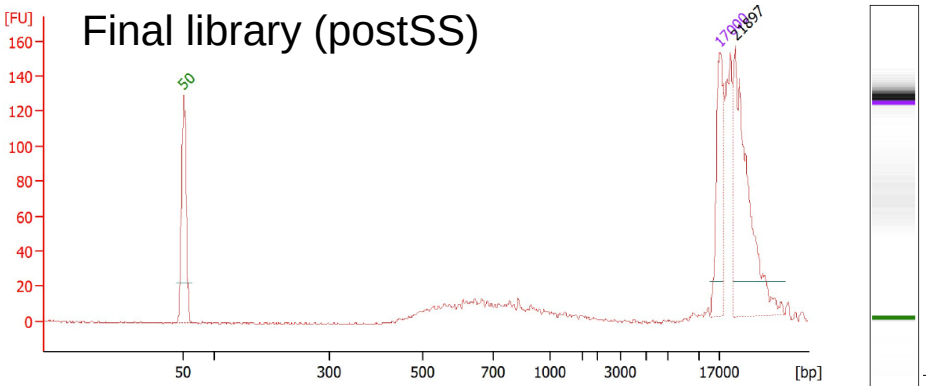
Input gDNA



SMRTbell



Final library (postSS)

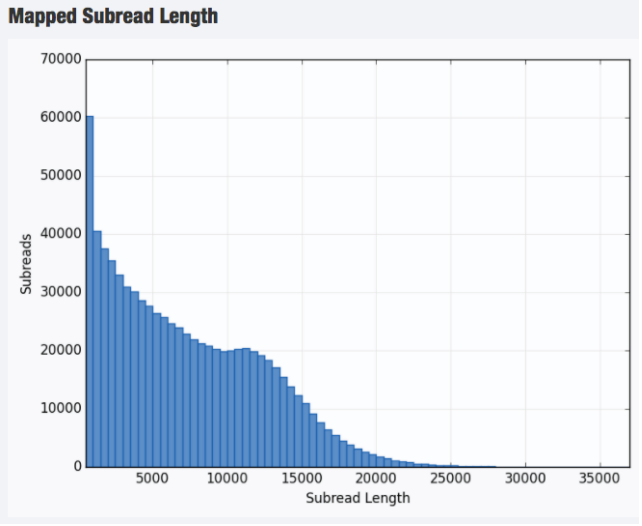


Sample Data Sets : Sequel Stats

"Just Fine (2 cell pilot)"

Mapping Report

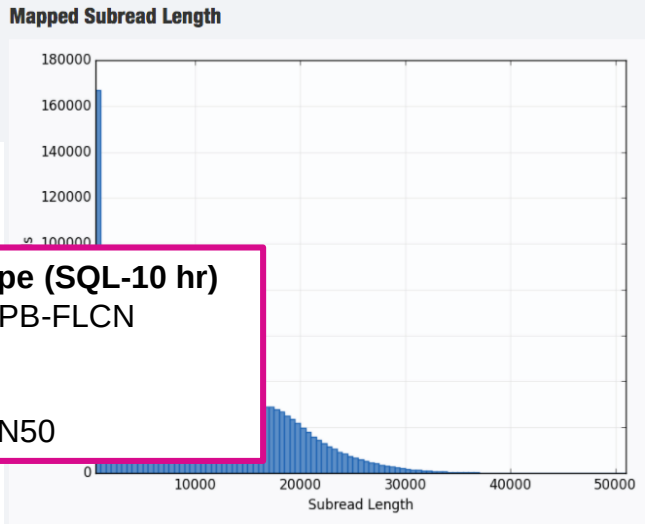
Value	Analysis Metric
83.52%	Mean Concordance (mapped)
794,400	Number of Subreads (mapped)
5,394,918,161	Number of Subread Bases (mapped)
6,791	Subread Length Mean (mapped)
10,718	Subread Length N50 (mapped)
15,840	Subread Length 95% (mapped)
36,431	Subread Length Max (mapped)
642,617	Number of Polymerase Reads (mapped)
8,555	Polymerase Read Length Mean (mapped)
13,008	Polymerase Read N50 (mapped)
24,010	F
77,546	F



"Great (4 cell pilot)"

Mapping Report

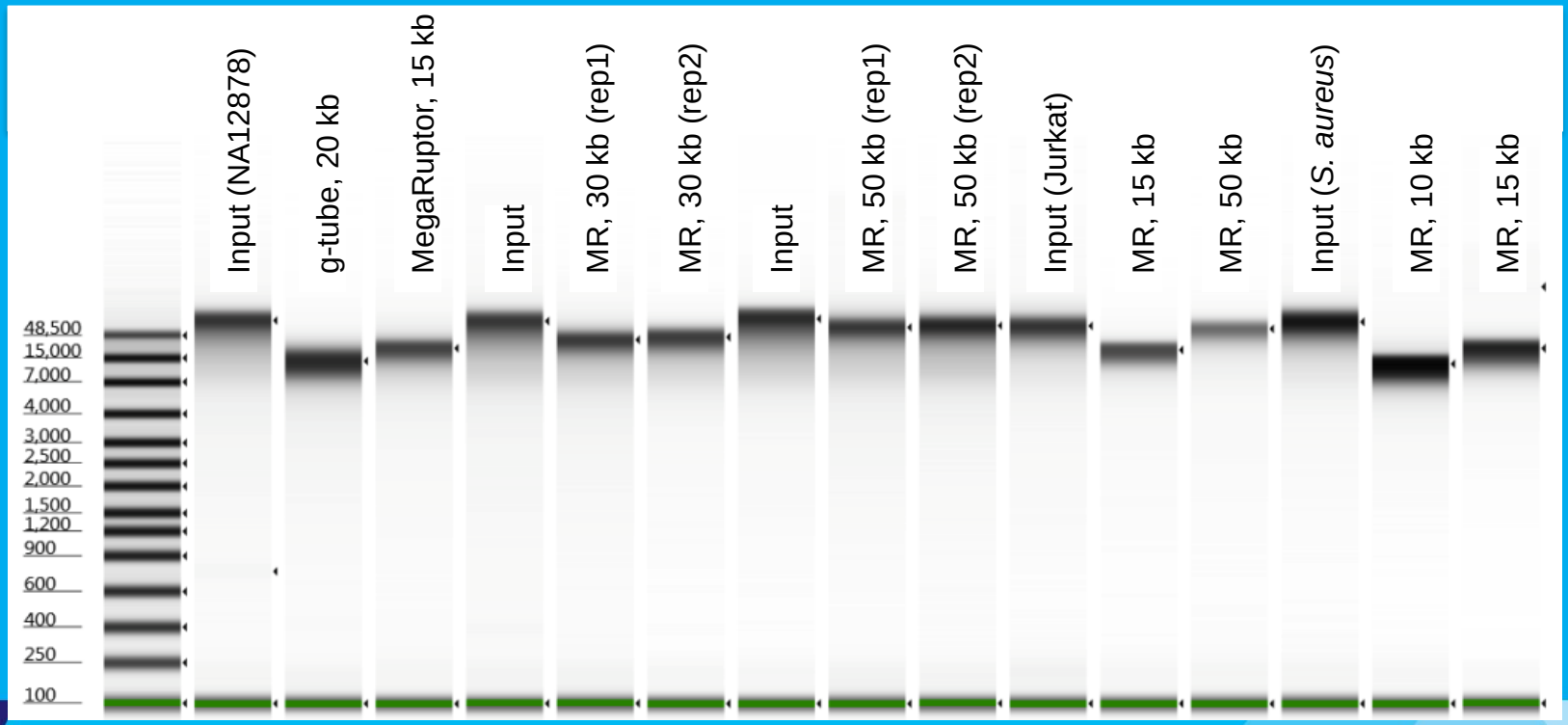
Value	Analysis Metric
84.04%	Mean Concordance (mapped)
2,172,791	Number of Subreads (mapped)
18,590,421,740	Number of Subread Bases (mapped)
8,556	Subread Length Mean (mapped)
13,904	Subread Length N50 (mapped)
21,620	Subread Length 95% (mapped)
50,116	Subread Length Max (mapped)
1,758,655	Number of Polymerase Reads (mapped)
10,732	Polymerase Read Length Mean (mapped)
17,498	Polymerase Read N50 (mapped)
32,120	
75,133	



(7) Cau-haplotype (SQL-10 hr)
~2 Mb* de novo PB-FLCN
~50 SMRTcells
~ weeks
~ 17.5kb subRL N50

Looking Forward

- Awaiting SMRTLink 5.0 and IPS upgrades soon
- Currently:
  - 5 whole human genomes in progress/in queue
  - Multiple additional large genomes, including mammalian, insect and plants
- Expanding to MegaRuptor (≥ 30 kb)-derived libraries



Many Thanks!

Tech Dev Team

Bobby Sebra

Gintaras Deikus

Irina Oussenko

Hardik Shah

Mondale Wang

Kimaada Allette
Maya Strahl
Ethan Ellis
Matthew Emery
Diane Castillo
Leah Newman
James Powell
Alesia Antoine
Wissam Hamou
Hanane Arib
David Sachs

Open to all the Q&A!

melissa.smith@mssm.edu