Applying Sequel to Genomic Datasets

Alicia Clum[1] (aclum@lbl.gov), Chris Daum[1], Paul Kotturi[2], Kurt LaButti[1], Ronan O'Malley[1], Jasmyn Pangilinan[1], Kristi Spittle[2], Matt Zane[1], Alex Copeland[1]

[1]Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, [2]Pacific Biosciences, 1305 O'Brien Drive, Menlo Park, CA

Abstract

De novo assembly is a large part of JGI's analysis portfolio. Repetitive DNA sequences are abundant in a wide range of organisms we sequence and pose a significant technical challenge for assembly. We are interested in long read technologies capable of spanning genomic repeats to produce better assemblies. We currently have three RS II and two Sequel PacBio machines. RS II machines are primarily used for fungal and microbial genome assembly as well as synthetic biology validation. Between microbes and fungi we produce hundreds of PacBio libraries a year and for throughput reasons the vast majority of these are >10 kb AMPure libraries. Throughput for RS II is about 1 Gb per SMRT Cell. This is ideal for microbial sized genomes but can be costly and labor intensive for larger projects which require multiple cells.

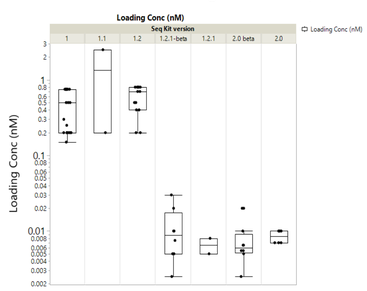
JGI was an early access site for Sequel and began testing with real samples in January 2016. During that time we've had the opportunity to sequence microbes, fungi, metagenomes, and plants. Here we present our experience over the last 18 months using the Sequel platform and provide comparisons with RS II results.

Methods

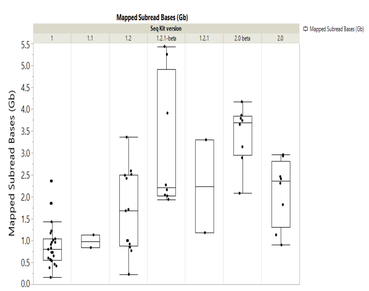
Libraries are a mix of large insert (10-50 kb) with BluePippin size selection and >10 kb with AMPure size selection. Data was generated on Sequel using a variety of chemistries and movie lengths vary from 2 -10 hours, majority of runs are 6 hour movies. RS II data shown is P6/C4 chemistry with 4 hour movies. Sequel chemistry 2.0.0 beta or later is used when compared directly to RS II. Mapping metrics were generated using SMRT Link[1] Resequencing protocol, versions 3.0.2 – 4.0.0.190159. CCS results were generated using SMRT Link 4.0.0.190159 CCS 2 protocol. Fungal assemblies were generated using Falcon 0.7.3[2]. ESTmapper[3] was used to align the transcriptome assembly to the genome assembly to estimate completeness.

Results

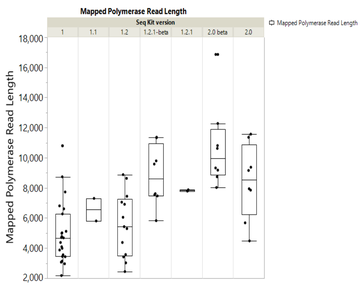
Figure 1: Sequel metrics by chemistry



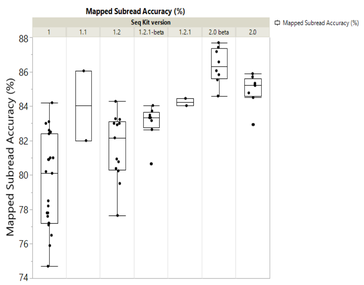
1a. Library loading concentration by Sequel chemistry



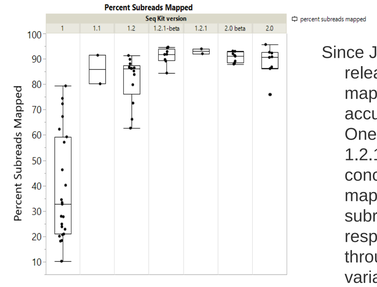
1b. Mapped subread gigabases by Sequel chemistry



1c. Mapped polymerase read length by Sequel chemistry



1d. Mapped subread accuracy by Sequel chemistry



1e. Mapped subread accuracy by Sequel chemistry

Since January 2016 there have been several chemistry releases resulting in an increase in throughput, mapped polymerase read lengths, mapped subread accuracy, and amount of mapped subread bases. One large improvement was the release of chemistry 1.2.1 which dramatically reduced the required loading concentrations. The variability has been reduced for mapped subread accuracy and percentage of subreads mapping, stabilizing at 85% and 90% respectively. We continue to struggle with run throughput. With the current chemistry it is quite variable and well short of the expected 6-7 Gb per SMRT Cell.

Results cont'd

Table 1: Interaction between loading concentration, throughput, and read length

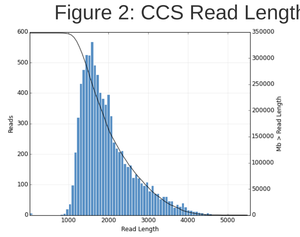
Sample	Loading Conc (nM)	P1 Productivity (%)	Mapped Reads	Mapped Polymerase Read Length	Mapped Subread Bases (Gb)	Mapped Subread Length	Mapped Subread Accuracy (%)	Percent Subreads Mapped
E. Coli K12-MG1655	0.005	18	178,335	11,331	2.02	9,530	82.64	92.90
E. Coli K12-MG1655	0.01	35	344,748	11,364	3.91	9,540	83.72	94.38
E. Coli K12-MG1655	0.02	56	537,263	9,795	5.25	8,297	83.16	91.83
E. Coli K12-MG1655	0.03	59	567,887	9,584	5.43	8,029	83.33	91.72

The table above demonstrates how increasing the loading concentration can generate more total gigabases but overloading can negatively impact read lengths and percent of mapped subreads. Optimal P1 for Sequel is ~40%.

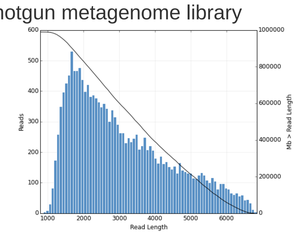
Table 2: Sequel vs. RS II for a shotgun metagenome library

Platform	Number of Polymerase Reads (mapped)	Polymerase Read Length Mean (mapped)	Polymerase Read N50 (mapped)	Number of Subreads (mapped)	Number of Subread Bases (mapped)	Subread Length Mean (mapped)	Mean Concordance (mapped)
RS II	72,091	5,418	11,290	184,987	355,962,125	1,924	85.40%
Sequel	159,063	7,787	14,857	353,062	1,179,642,453	3,341	84.45%
Sequel	437,578	7,875	14,010	895,907	3,299,941,373	3,683	84.03%

Figure 2: CCS Read Length for a shotgun metagenome library



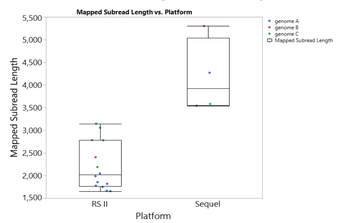
2a. RS II CCS read lengths



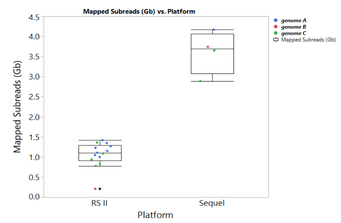
2b. Sequel CCS read lengths

We sequenced an existing low input 10 kb PacBio library for which we had RS II P6/C4 data on a synthetic community containing 304 organisms. The results show increased throughput, 1-3 Gb, and increased read lengths with similar mapped concordance (table 2). CCS read lengths are approximately 3 kb vs 2 kb on the Sequel, likely containing at least 1 full length CDS per consensus read (fig. 2).

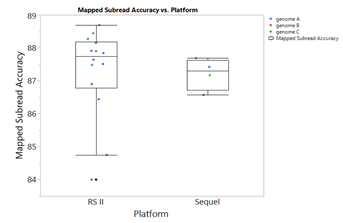
Figure 3: Sequel vs. RS II read metrics for isolates



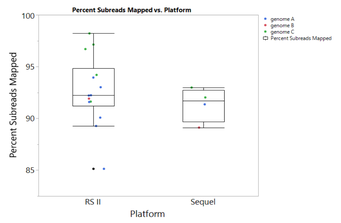
3a. Mapped subread lengths by platform



3b. Mapped subread gigabases by platform



3c. Mapped subread accuracy by platform

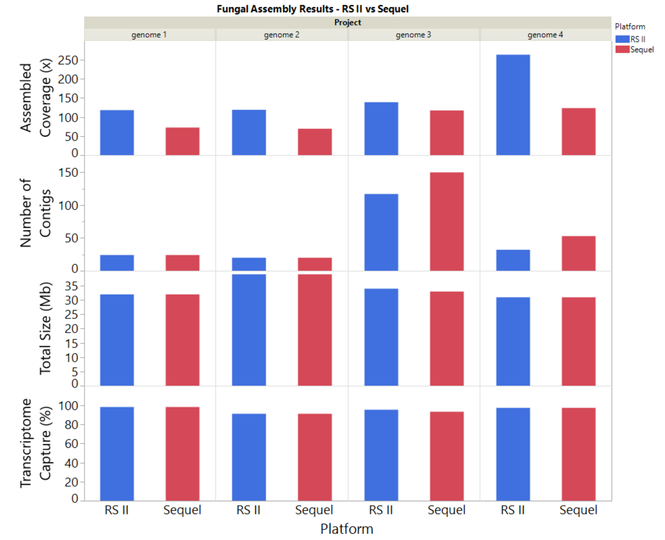


3d. Percent of subreads mapped by platform

The figure above compare >10 kb AMPure size selected libraries sequenced on both RS II and Sequel. One unexpected benefit of the Sequel platform has been reduced loading bias, mapped subread lengths for these libraries doubled. Sequel also produces a larger total number of mapped subread bases resulting in higher project throughput and reduced labor per project costs. Sequel mapped subread accuracy and percent of subreads mapping are slightly lower than RS II but acceptable for our analysis applications.

Results cont'd

Figure 3: Platform impact on fungal assembly



Due to the reduction in loading bias in the Sequel platform it is possible to produce comparable assembly results with less total sequencing coverage. The four projects above are 31-40 MB in size and were assembled with just a single Sequel SMRT Cell compared to 5-8 RS II SMRT Cells. Transcriptome capture as a measure of completeness is comparable.

Conclusions

- JGI implemented Sequel in production in April 2017 after more than a year of validation
- We are planning to migrate large projects including fungi, metagenomes, and plants to this platform
- Increased throughput and read lengths on the Sequel compared to RS II will produce comparable or better assemblies with fewer resources
- Average size for fungi is 40 Mb, ideal for a 1 SMRT Cell approach given current Sequel throughput
- The increased throughput makes small plant genomes more trackable
- Metagenome shotgun results look promising despite issues obtaining high molecular weight DNA
- We expect significant labor savings per project from Sequel
- For significant reagent savings throughput needs to be higher and more stable as Sequel SMRT Cells are more costly than RS II SMRT Cells

Next steps

- Improving gigabase throughput and reducing variability per SMRT Cell
- testing additional metagenome samples
- optimizing loading concentrations across a broad range of samples
- Testing barcoding to allow for pooling of multiple samples per SMRT Cell to keep up with projected throughput roadmap

References

[1]PACBIO®. http://www.pacb.com/products-and-services/analytical-software/smrt-analysis/. Accessed 5 May 2017

[2]Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, Cramer GR, Delledonne M, Luo C, Ecker JR, Cantu D, Rank DR, Schatz MC (2016) Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods 13 (12):1050-1054. doi:10.1038/nmeth.4035

[3]Xue Wu, Woei-Jyh (Adam) Lee, Chau-Wen Tseng, (2005) ESTmapper: Efficiently Aligning DNA Sequences to Genomes. IPDPS 7(8) 196a. doi: 10.1109/IPDPS.2005.204

Acknowledgments

We'd like to acknowledge the JGI sequencing and library preparation teams and the PacBio customer support team.

The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported under Contract No. DE-AC02-05CH11231.