Microbial Multiplexing Workflow on the Sequel System

Introduction

Obtaining microbial genomes with the highest accuracy and contiguity is extremely important when exploring the functional impact of genetic and epigenetic variants on a genome-wide scale. A comprehensive view of the bacterial genome, including genes, regulatory regions, IS elements, phage integration sites, and base modifications is vital to understanding key traits such as antibiotic resistance, virulence, and metabolism. SMRT Sequencing provides complete genomes, often assembled into a single contig.

Our streamlined microbial multiplexing procedure for the Sequel System, from library preparation to genome assembly, can be completed with less than 12 hours bench time (Figure 1). Starting with high-quality genomic DNA (gDNA), samples are sheared to a 10 kb distribution, ligated with barcoded adapters, pooled at equimolar representation, and sequenced. Demultiplexing of samples is automated, allowing for immediate genome assembly on our SMRT Link analysis software solution.

The workflow supports up to 16-plex of *de novo* microbial genomes where the total genome sums up to 30 Mb on each SMRT Cell. As microbial genomes and gDNA samples vary in genetic complexity and quality, respectively, we describe general recommendations and best practices.

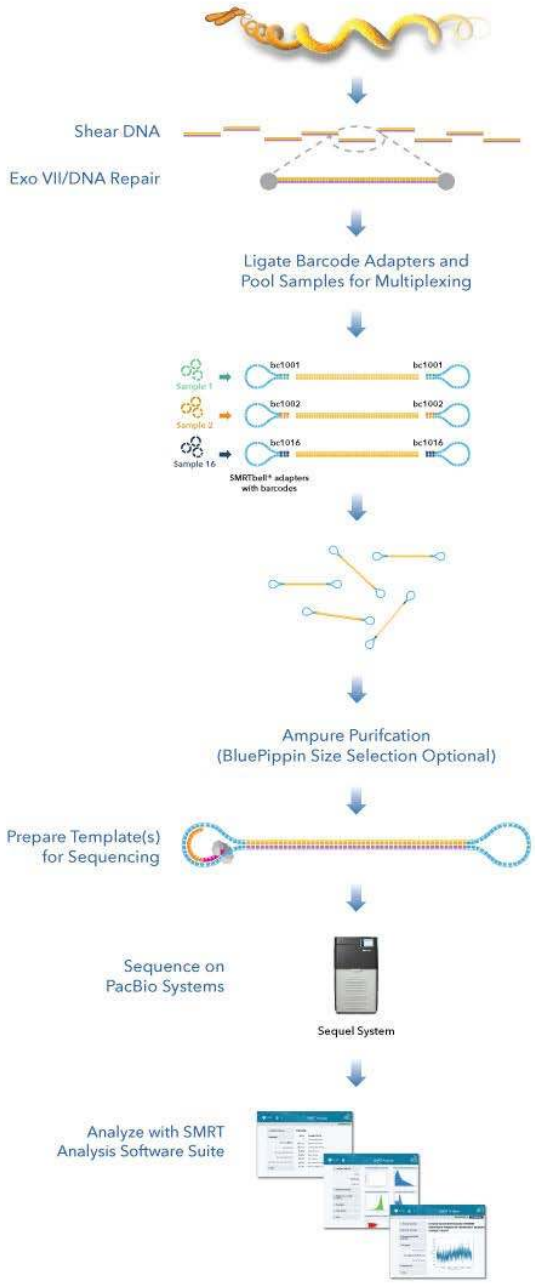


Figure 1 – Overview of multiplexed microbial sequencing workflow on the Sequel System.

Multiplex Samples with Validated Barcoded Adapters and a Streamlined Workflow

Two new barcoded adapter kits, with 8 barcoded SMRTbell adapters in each, are available for multiplexing experiments on the Sequel System. The barcodes used in these kits are specifically validated for the microbial multiplexing application. We recommend using these validated barcodes, particularly when pooled genomes share high homology, to ensure the correct genetic content assignment to the original multiplexed microbe. Other benefits of barcoding the individual microbes include the correct assignment of plasmids to specific strains if captured during sequencing.

Along with the barcoded adapter kits, we've streamlined our sample preparation and analysis workflow to offer:

- Simple, efficient workflow from sample DNA to high-quality genome assemblies
- Highly contiguous assemblies, with main chromosomes captured in 5 contigs or fewer
- High empirical QV scores at 99.999% accuracy to resolve SNPs and structural variants
- Cost savings with a multiplexed workflow involving less than 12 hours hands-on at the bench

Experimental Design: Achieving Closed Microbial Genomes

When working with genomes of varying quality or from unknown sources, we recommend starting with a more conservative experimental design of 30 Mb total microbial genomes, including extrachromosomal genomic sequences. Plasmids may occasionally be sequenced and assembled along with the chromosomal DNA. However, plasmids often need to be isolated separately as these may be excluded depending on the extraction kit used, particularly since supercoiled plasmids are also resistant to shearing.

Genome complexity is another consideration that will impact the ability to achieve closed microbial genomes. A recent survey of microbial genome complexity proposed three classes of genome assembly complexity which we find useful in understanding how to achieve the desired genome contiguity in the most cost-effective way (Koren et al. 2013). We offer the following considerations:

I. **Class I** genomes have few repeats except for the rDNA operon sized 5 kb to 7 kb. These assemble to <5 contigs with multiplexing up to 30 Mb total genomes.
II. **Class II** genomes have many repeats, such as insertion sequence elements, but none greater than 7 kb. These may need higher coverage to close, which can be achieved by lowering per SMRT Cell multiplexing. Size-selection may enrich for longer reads to span repeats.
III. **Class III** genomes contain large, often phage-related repeats >7 kb, including tandem repeats and segmental duplications. These are difficult to close with 10 kb insert libraries. If a closed genome is required, non-multiplexed sequencing with larger insert libraries can be explored. We offer >15 kb and >30 kb insert library options.

While we recommend a starting point of 30 Mb, high quality gDNA and Class I microbes have resulted in closed microbial genomes with a 10-plex experimental design containing a mix of PacBio internal control samples and microbes obtained from Center for Food Safety and Nutrition (CFSAN). Five microbial genomes from different sources were sequenced in duplicate for a total genome size of 42.4 Mb on a single run (Table 1). Our ability to achieve closed genomes with >40 Mb sample pools is highly dependent on obtaining highly intact DNA and adhering to the best practice guidelines described in our library preparation workflow (Figure 2). Key considerations include:

- Starting with predominantly >20 kb gDNA
- Attaining equimolar pooling of microbial samples
- Optimized loading for sequencing yield
- Incorporation of advanced parameters specific for microbial genome assemblies

Barcode ID	Sample ID	Gram Status	Genome Size (bp)
BC1002	*Escherichia coli* [1]	-	4,642,522
BC1015	*Escherichia coli* [1]	-	4,642,522
BC1004	*Bacillus subtilis* [1]	+	4,045,592
BC1016	*Bacillus subtilis* [1]	+	4,045,592
BC1009	*Escherichia coli* [2]	-	4,642,510
BC1018	*Escherichia coli* [2]	-	4,642,510
BC1012	*Shigella sonnei* [3]	-	4,813,454
BC1020	*Shigella sonnei* [3]	-	4,813,454
BC1014	*Listeria monocytogenes* [3]	+	3,032,244
BC1022	*Listeria monocytogenes* [3]	+	3,032,244
		Total Genome Size	**42,352,644**

Table 1 – Experimental design for a 10-plex microbial genome pool with a total of 42.4 Mb genome size. Pooled samples include the following: (1) PacBio control microbial genomes extracted using 'MasterPure Complete DNA and RNA Purification Kit' (Lucigen), and nanobind DNA extraction technology from Circulomics Nanobind CBB Big DNA Kit for (2) Pacbio control and (3) CFSAN microbial genomes. Sequel System Instrument Control Software v5.1 and Sequel Binding and Sequencing Kits 2.1 were used. Movie time was 10 hours.

Microbial Multiplexing Sample Preparation Workflow

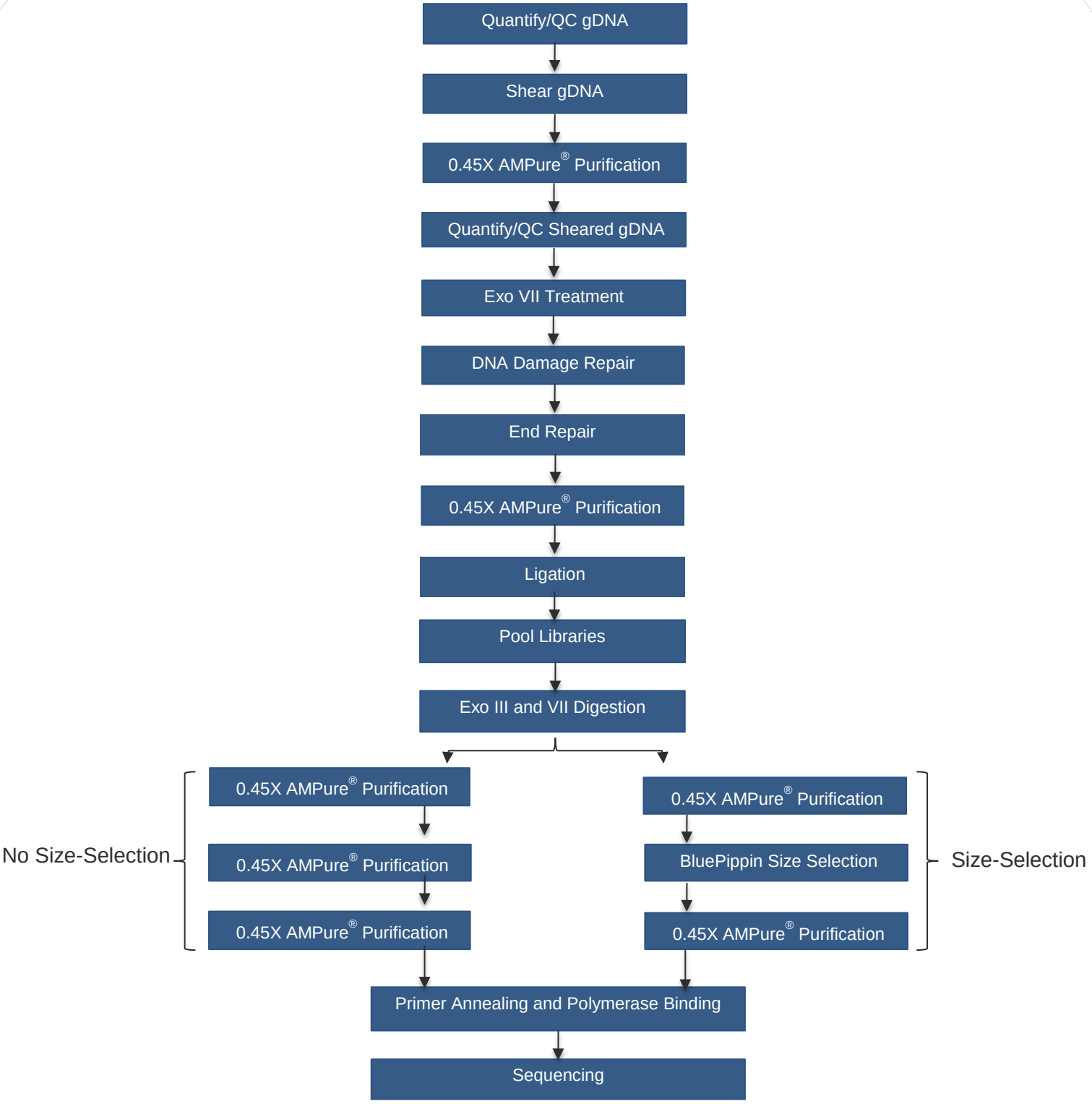


Figure 2 – Microbial Multiplexing Sample Preparation workflow with Barcoded SMRTbell Adapters. Size-selection is optional. Workflow can be completed with <12 hours hands-on bench time.

Sample Preparation: Start with High Quality Microbial DNA

High quality gDNA is critical to achieve the long read lengths needed to span repeats and close microbial genomes. We recommend starting with genomic DNA predominantly >20 kb. For the dataset described in Table 1, we used Lucigen and Circulomics kits to extract gDNA of sufficient quality to effectively produce closed genomes. We characterized the size distribution of the extracted gDNA using either the Bio-Rad CHEF Mapper XA Pulsed Field Electrophoresis or Pippin Pulse Electrophoresis Power Supply from Sage Science. Another alternative is the FEMTO Pulse automated pulsed-field capillary electrophoresis instrument from Advanced Analytical Technologies, Inc. (FEMTO Pulse), which provides accurate sizing of gDNA up to 165 kb in under 1.5 hours. We show here the sizing QC of the samples on the CHEF Mapper system (Figure 3a), showing all DNA migrated predominantly above the 20 kb sizing marker.

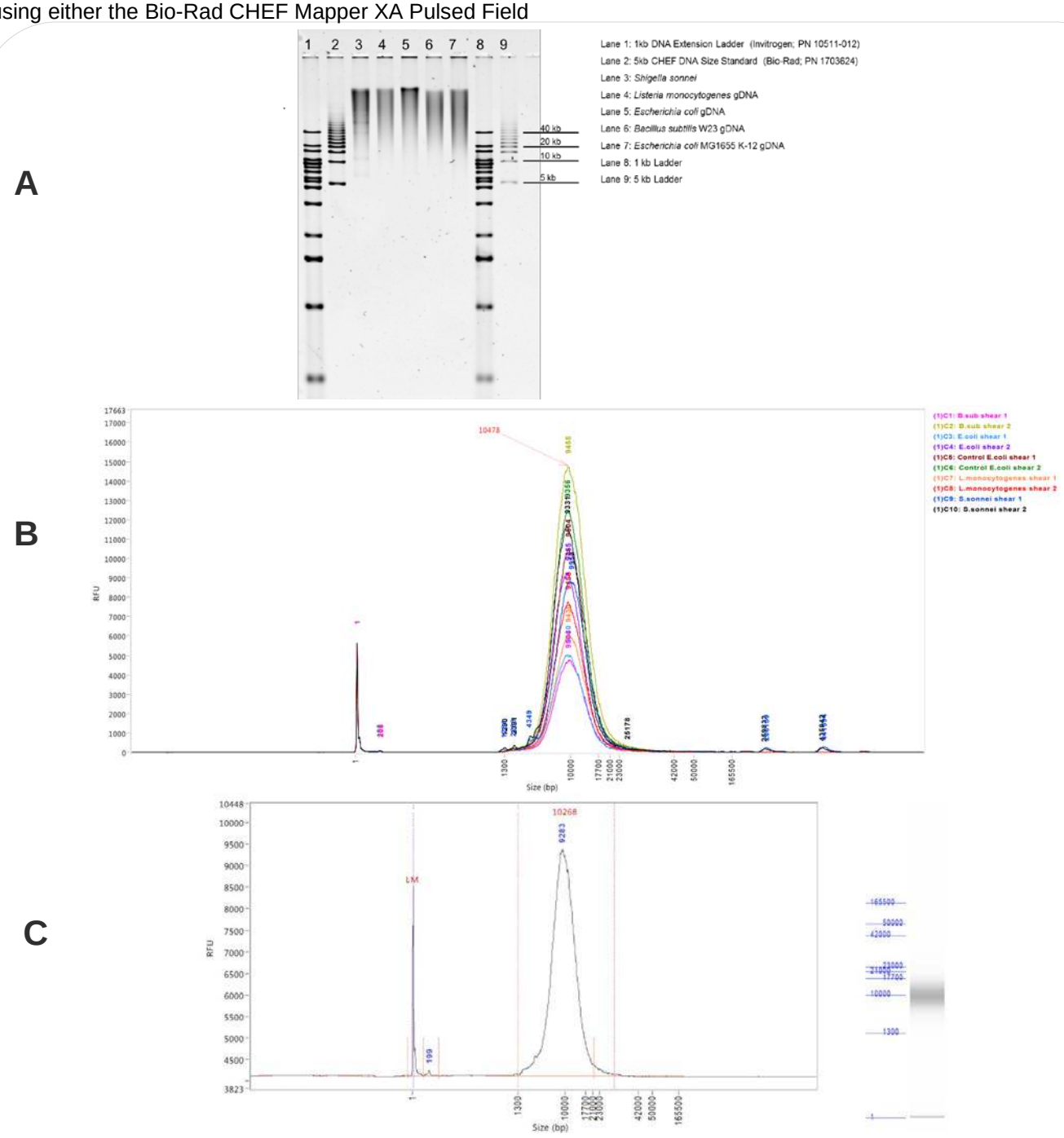


Figure 3 – Genomic DNA and Shear QC. (A) CHEF Mapper gel image showing all gDNA sample quality predominantly >20 kb. (B) FEMTO *Pulse* electropherogram of 10 samples sheared to a target 10 kb distribution, and (C) pooled SMRTbell libraries retained similar size distribution after the library preparation workflow.

Starting with high-quality gDNA predominantly >20 kb helps ensure an even 10 kb average shear when processing samples in high-volumes, as shown in our FEMTO *Pulse* electropherogram in this case study (Figure 3b). Consistent size distribution from shearing gDNA coupled with our Microbial Multiplexing Calculator (Figure 4) helps to ensure equimolar pooling and even sequencing representation across all pooled samples despite different genome sizes, shear sizes, and sample concentrations.

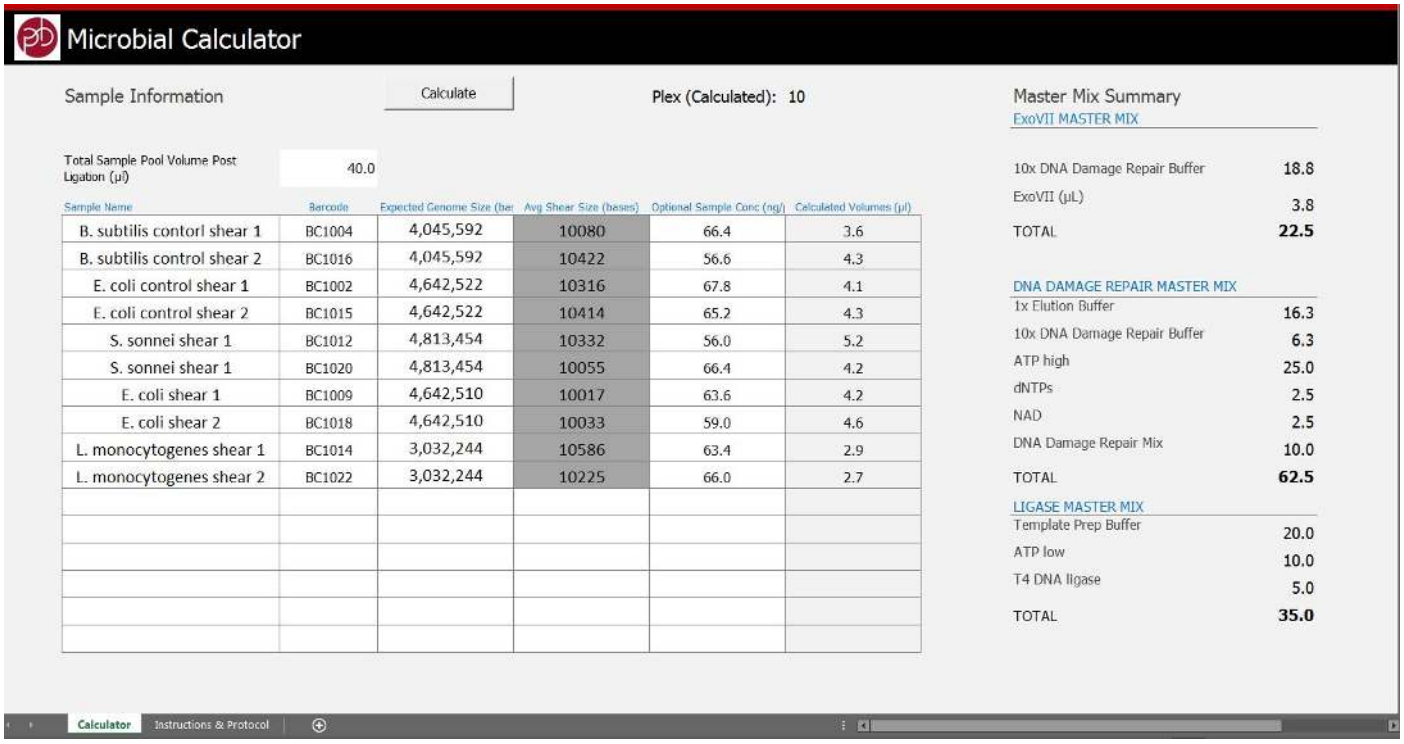
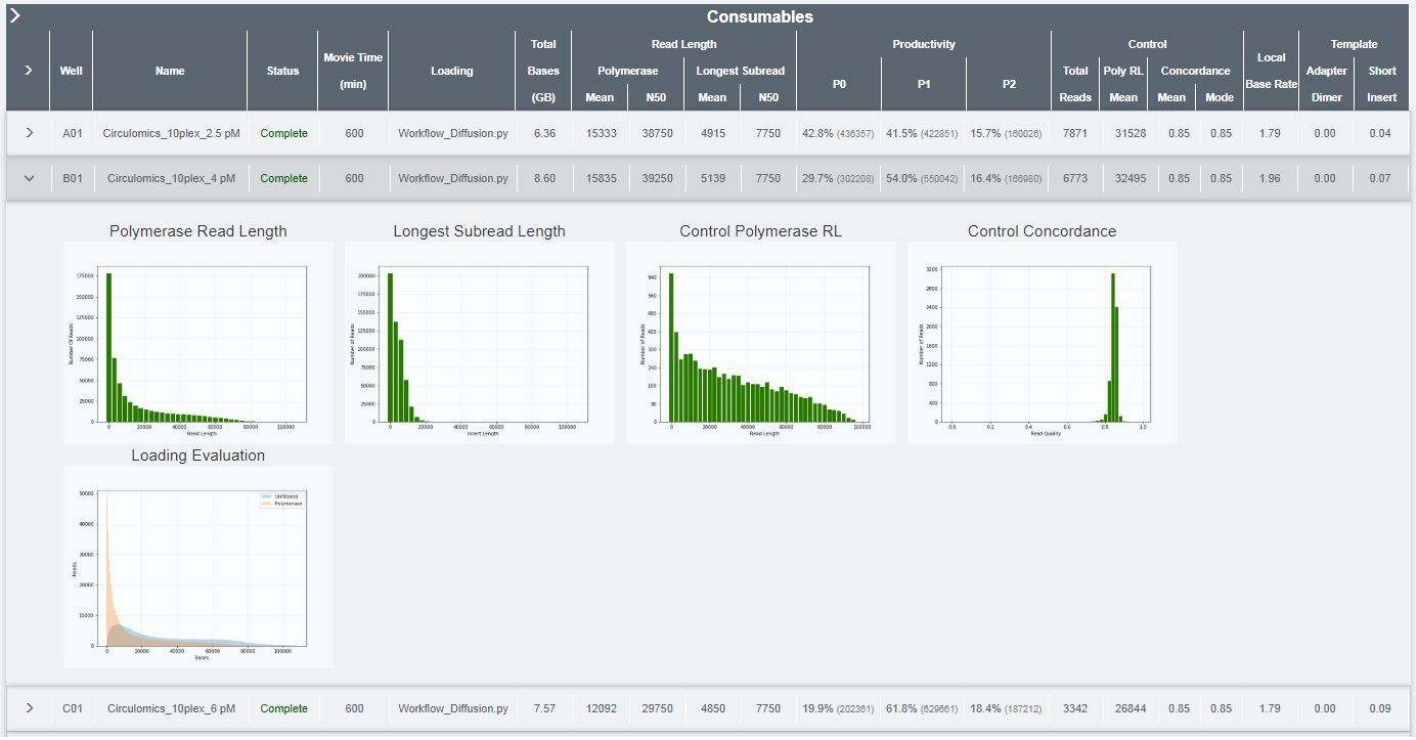


Figure 4 – The Microbial Multiplexing Calculator helps to ensure equimolar pooling of samples with respect to average shear and genome size. Optional concentration input serves to further optimize pooling ratio between samples. Calculator includes 'Master Mix Calculations' for generating SMRTbell library templates.

Size-selection is provided as an option within the protocol and can be useful for microbial genomes containing known repetitive regions spanning >6 kb. Size-selection may confer some benefits to assembly contiguity for more fragmented samples with smaller shear sizes. However, since this is performed after pooling, those samples with a shorter average shear size may be subsequently underrepresented with reduced coverage in the sequencing data. Size-selection will also remove plasmids below the size selection cutoff. We highly recommend starting from high quality gDNA for best results.

SMRT Sequencing: Optimized Loading to Achieve Sufficient Microbial Genome Coverage

For optimal loading, we recommend following the guidance detailed in the Quick Reference Card: Diffusion Loading and Pre-Extension Time Recommendations for the Sequel System. Generally, we highy recommend targeting a productive loading fraction (P1) ranging from 50% – 65%, for a total throughput of approximately 8 Gb per SMRT Cell. Our experiences have shown both under and overloading negatively impacts SMRT Sequencing yield resulting in insufficient genome coverage for high contiguity assemblies (Figure 5). We recommend diffusion loading with a 120 min pre-extension time to ensure highest possible yield with sequencing through at least one barcode for pooled sample assignment. Ten-hour movie collections provide sufficient data collection to achieve the needed coverage across the pooled microbial genomes.



Well	Name	Status	Movie Time (min)	Loading	Total Bases (GB)	Read Length: Polymerase Mean	Read Length: Polymerase N50	Read Length: Longest Subread Mean	Read Length: Longest Subread N50	Productivity: P0	Productivity: P1	Productivity: P2	Control: Total Reads	Control: Poly RL Mean	Control: Concordance Mean	Control: Concordance Mode	Local Base Rate	Template: Adapter Dimer	Template: Short Insert
A01	Circulomics_10plex_2.5 pM	Complete	600	Workflow_Diffusion.py	6.36	15333	38750	4915	7750	42.8% (436357)	41.5% (422851)	15.7% (160026)	7871	31528	0.85	0.85	1.79	0.00	0.04
B01	Circulomics_10plex_4 pM	Complete	600	Workflow_Diffusion.py	8.60	15835	39250	5139	7750	29.7% (302208)	54.0% (550042)	16.4% (166980)	6773	32495	0.85	0.85	1.96	0.00	0.07
C01	Circulomics_10plex_6 pM	Complete	600	Workflow_Diffusion.py	7.57	12092	29750	4850	7750	19.9% (202361)	61.8% (629861)	18.4% (187212)	3342	26844	0.85	0.85	1.79	0.00	0.09

Figure 5 – Run QC from SMRT Link with a range of loading on-plate concentration from 2.5, 4 and 6 pM which yield productive fraction loading of P1 41.5%, 54.0%, 61.8% respectively. Optimal loading at 4 pM shows highest yield at 8.6 Gb and average polymerase read length of 15,835 bp. Underloading may yield insufficient coverage for samples, whereas overloading may result in decrease sequencing quality which will also impact genome assembly results.

Genome Assembly: Automated Demultiplexing and Advanced Parameters Tuned for Multiplexed Microbial Genome Assemblies

Starting with v5.1.0, SMRT Link supports automated demultiplexing of samples upon completion of the sequencing run (Figure 6). After demultiplexing, you can assess variation in pooling by measuring the number of barcoded bases and reads from each barcode. If the pooled library contained genomes of comparable size, a quick assessment could be made by the ratio between the highest and lowest number of barcoded reads. For example, the equimolar pooling variance for this 10-plex pooled library is approximately 2-fold between samples.

Specifically, the observed number of polymerase reads ranged from 15,771 to 35,895, with 28,709 as the average number of reads across the 10 samples (Figure 6). Pools containing genomes of varying sizes should use barcoded bases normalized to expected genome size to more accurately estimate relative coverage for each genome and access equimolar pooling variance.

We are typically able to achieve approximately 2-fold variation in coverage. Larger variations due to inaccurate or indeterminant genome sizing may result in poor assemblies for those genomes with insufficient coverage. Our 30 Mb recommendation for genome pooling is a good starting point to address this variance. With experience, the number of multiplexed genomes can be increased.

Barcode Data

Bio Sample Name	Barcode Index	Barcode Name	Polymerase Reads	Subreads	Bases	Mean Read Length	Longest Subread Length	Mean Barcode Quality	Rank Order (Num. Reads)
E.coli	1--1	bc1002_BAK8A--bc1002_BAK8A	31,888	167,563	827,446,387	39,734	51,908	68.0	5
B.sub	2--2	bc1004_BAK8A--bc1004_BAK8A	27,424	150,411	755,588,853	41,594	48,403	70.0	7
E.coli	4--4	bc1009_BAK8A--bc1009_BAK8A	32,647	170,173	850,150,310	40,021	59,582	70.0	4
Shigella	6--6	bc1012_BAK8A--bc1012_BAK8A	33,895	175,258	892,001,257	40,239	57,109	69.0	3
Listeria	7--7	bc1014_BAK8A--bc1014_BAK8A	20,686	115,263	580,738,539	42,092	42,516	70.0	9
E.coli	8--8	bc1015_BAK8B--bc1015_BAK8B	35,817	187,090	935,531,479	40,079	38,820	66.0	1
B.sub	9--9	bc1016_BAK8B--bc1016_BAK8B	26,892	147,895	732,606,272	41,180	40,018	65.0	8
E.coli	11--11	bc1018_BAK8B--bc1018_BAK8B	33,948	177,558	884,477,818	40,107	36,476	68.0	2
Shigella	13--13	bc1020_BAK8B--bc1020_BAK8B	28,124	144,128	724,830,493	39,455	37,428	70.0	6
Listeria	15--15	bc1022_BAK8B--bc1022_BAK8B	15,771	87,562	439,391,247	41,650	49,411	65.0	10
No Name	None	Not Barcoded	256,090	313,297	862,610,257	14,174	73,676	0.0	NA

Figure 6 – Barcode QC after sample demultiplexing with SMRT Link software. 52% of reads were barcoded with a low equimolar variance (e.g. less than approximately 2-fold) between pooled samples as evidenced by the number of barcoded reads.

After demultiplexing, HGAP4 *de novo* genome assembly runs can be initiated for each sample. We have optimized HGAP4 genome assembly parameters to help ensure a robust experience provided there is a minimum of 30-fold unique read coverage across the microbial genome. HGAP4 automatically filters the sequencing data to remove SMRTbell adapter sequences, recover high-quality genomic content, and access unique read coverage reported as filtered subread coverage. We highly recommend following the analysis guidance optimized for microbial multiplexing experiments. This includes an advanced parameter requiring manual override for microbial genome assemblies as detailed in the Analysis Procedure: Multiplexed Microbial Assembly with SMRT Link v5.1.0 , as well as setting 'Aggressive Mode' to 'ON'.

An important metric to monitor is pre-assembly yield, which is reported in the assembly output and should be >60% (Table 2). Pre-assembly yield serves as a good indicator, as low yield may be an indication of poor SMRTbell library quality, overloading, or both. Poor pre-assembly yield will also likely result in insufficient coverage in some parts of the genome, resulting in reduced assembly contiguity and a higher number of contigs.

The initial HGAP assembly is only the first step in achieving a high-quality assembly. This first pass assembly can be circularized using the Circlator tool. Once circularized, the assembly can be imported into SMRT Link for subsequent polishing with the **Resequencing** analysis job to attain a higher base quality (Table 3).

		# Contigs Per On-Plate Concentration			Pre-Assembly Yield (%) Per On-Plate Concentration		
Barcode ID	Sample ID	2.5 pM	4 pM	6 pM	2.5 pM	4 pM	6 pM
BC1002	*E. coli control 1*	1	1	1	79.3	77	70.59
BC1015	*E. coli control 2*	1	1	1	80.2	77.9	70.89
BC1004	*B. sub control 1*	1	1	1	78	76.4	70.59
BC1016	*B. sub control 2*	2	1	1	77	74.9	69.79
BC1009	*E. coli 1*	1	1	1	79.9	77.4	70.89
BC1018	*E. coli 2*	1	1	1	80.9	77.7	71.39
BC1012	*S. sonnei 1*	1	1	1	78	76.4	71.5
BC1020	*S. sonnei 2*	1	1	3	81	78.1	71.5
BC1014	*L. monocytogenes 1*	Failed Coverage	1	1	Failed Coverage	71.8	66.8
BC1022	*L. monocytogenes 2*	1	1	1	73	73.2	68.7

Table 2 – Summary of HGAP4 Genome assembly results showing impacts of loading and preassembly yield. We recommend a >60% minimum preassembly yield and >30-fold filtered subread coverage per genome for high contiguity assemblies. The on-plate concentration for this run, 2.5, 4, and 6 pM, corresponded to a productive P1 loading of P1 41.5%, 54.0%, 61.8% respectively. A P1 range of 50 – 65% is recommended for optimal loading. Underloading will have negative impacts on the genome assemblies with insufficient coverage. As will overloaded samples which may yield decreased average read lengths and quality.

Barcode ID	Sample ID	GC Content (%)	Genome Size (bp)	Contigs (#)	Concordance w/ Reference (QV)
BC1002	*E. coli control 1*	50.08	4642523	1	55.21
BC1015	*E. coli control 2*	50.08	4642523	1	56.25
BC1004	*B. sub control 1*	43.94	4045593	1	51.16
BC1016	*B. sub control 2*	43.94	4045593	1	51.45
BC1009	*E. coli 1*	50.08	4642523	1	54.63
BC1018	*E. coli 2*	50.08	4642523	1	54.91
BC1012	*S. sonnei 1*	51.03	4813450	1	59.83
BC1020	*S. sonnei 2*	51.03	4813450	1	54.27
BC1014	*L. monocytogenes 1*	37.94	3032269	1	53.06
BC1022	*L. monocytogenes 2*	37.94	3032269	1	49.63

Table 3 – Summary of genome assembly results after circularization and polishing for improved base quality scores. Contig N50 not reported given the single contig assembly, and would therefore equal the assembled genome size. QV scores compared to reference genomes are shown; a QV 50 score indicates 99.999% concordance. *L. monocytogenes* and *S. sonnei* assemblies were compared to CFSAN001178 and CFSAN030807 reference genomes, respectively.

Achieve Cost Efficiency with Optimized Internal Processes for High Quality Genomes

With routine experience, adherence to our suggested best practices, and optimization of internal workflow and processes, cost efficiency can be achieved with higher multiplexed samples. We demonstrate here a 68 Mb pooled library for four microbial genomes (*B. subtilis* W23*, E. coli* MG1655*, R. palustris* CGA009*,* and *S. aureus* USA300_TCH1516) multiplexed in quadruplicate collected with a 20-hour movie run time. The benefits of 20-hour movie data or, in general, longer movie collections compared to our recommended 10-hour movie may improve unique read coverage. Fundamentally, starting with high-quality genomic DNA maximizes your chance of a successful outcome. As shown in Table 3, we have been successful with closing microbial genomes using gDNA extraction kits from Circulomics and Lucigen. All genomes were sequenced with a minimum of 30-fold filtered subread coverage, with the exception of *S. aureus* which fell below the recommended minimum (Figure 7). The majority of samples had good pre-assembly yield, and <5 contig genome assemblies were achieved for all samples excepting one, which was recovered in a size-selected library. Size-selection may potentially improve assemblies, however is not required. In some instances, it may be helpful to further tune HGAP4 assembly parameters to optimize genome assembly, for example by increasing the minimum preassembly coverage. Overall, the workflow demonstrates robustness and reproducibility across the four replicates.

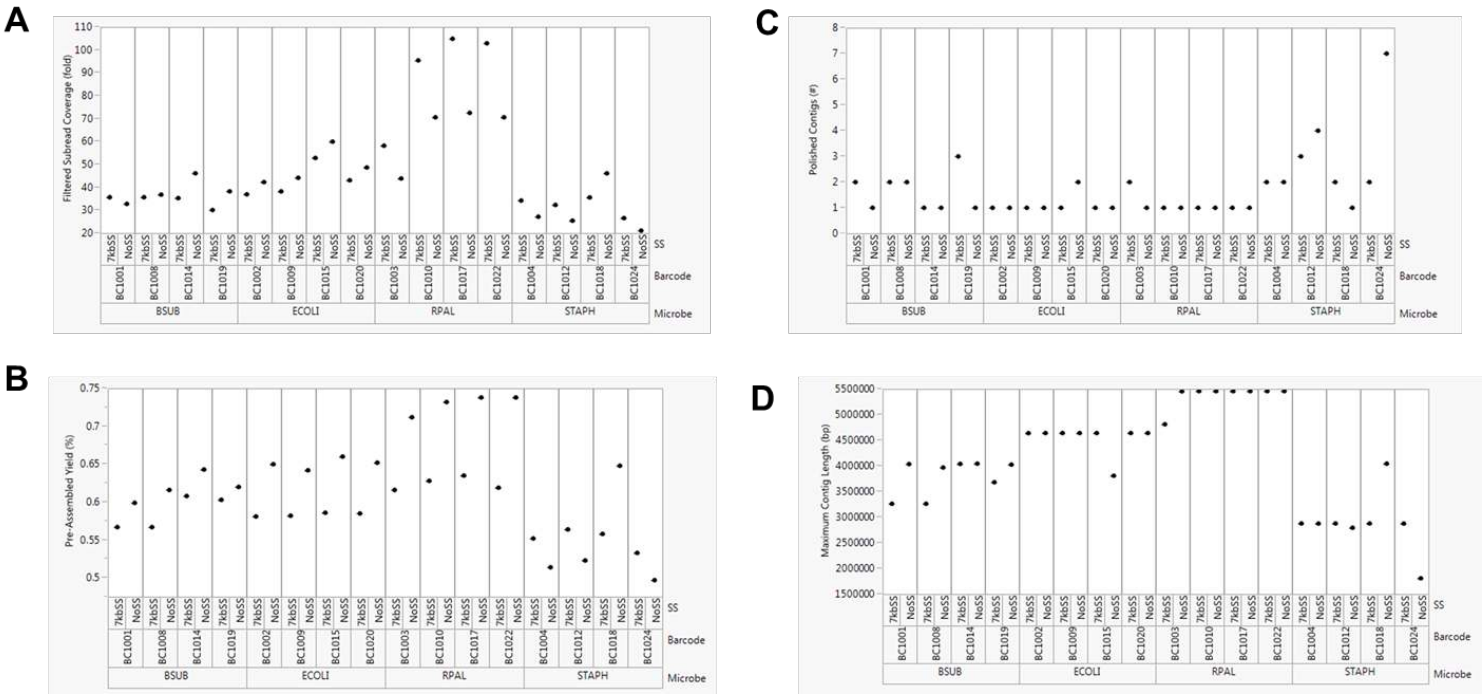


Figure 7 – Genome sequencing and assembly results for a 68 Mb pooled library design and a 20-hour movie data collection. (A) Filtered subread coverage show the importance of equimolar pooling for even genome coverage across multiplexed samples to avoid underrepresentation of samples. (B) Pre-assembly yield recommended at >60% for robust genome assemblies and is dependent on attaining sufficient genome coverage. (C) Majority of genomes were assembled in <5 contigs, and improvements in genome assemblies may be observed from size-selected libraries. (D) Maximum contig length from genome assemblies show high contiguity for main chromosomal genome assemblies.

Conclusions

While PacBio recommends 30 Mb of combined microbial genomes per SMRT Cell as a starting point for successful results with unknown sample gDNA quality and unknown genetic complexity, we have demonstrated that it is possible to start with 40 Mb of genome with high molecular weight DNA. We have even demonstrated as high as >65 Mb with a 16-plex library and 20-hour movie collection. The quality of the DNA defines the robustness of this workflow to consistently deliver genome assemblies with high contiguity, along with achieving sufficient sequencing yield. Moving to higher multiplexing may increase the number of contigs but allow for a significantly reduced cost per microbe. This tradeoff may be very attractive and should be assessed based on purposes of your scientific research.

Resources

- Procedure & Checklist: Preparing Multiplexed Microbial SMRTbell Libraries for the PacBio Sequel System
- Analysis Procedure: Multiplexed Microbial Assembly with SMRT Link v5.1.0
- Microbial Multiplexing Calculator
- Quick Reference Card: Diffusion Loading and Pre-Extension Time Recommendations for the Sequel System
- Circlator Tool for circularizing microbial genomes
- Koren S, et. Al. Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome Biology 2013;14(9):R101.
- Dataset Release: Microbial Multiplexing with Sequencing Chemistry 2.1

PacBio Consumable Part Numbers:

Part Number	Item
101-081-300	PacBio Barcoded Adapter Kit 8A
101-081-400	PacBio Barcoded Adapter Kit 8B
100-259-100	SMRTbell Template Prep Kit 1.0

Ancilliary Part Numbers:

Part Number	Supplier	Item
MC89010	Lucigen	MasterPure Complete DNA and RNA Purification Kit
NB-900-001-01	Circulomics	Circulomics Nanobind CBB Big DNA Kit
170-3670	Bio-Rad	CHEF Mapper XA
PP10200	Sage Sciences	Pippin Pulse Electrophoresis Power Supply
FPv1-CE2	Advanced Analytical Technologies Inc.	FEMTO Pulse Automated Pulsed-Field CE Instrument

PN 101-588-800 Version 01 (May 2018)