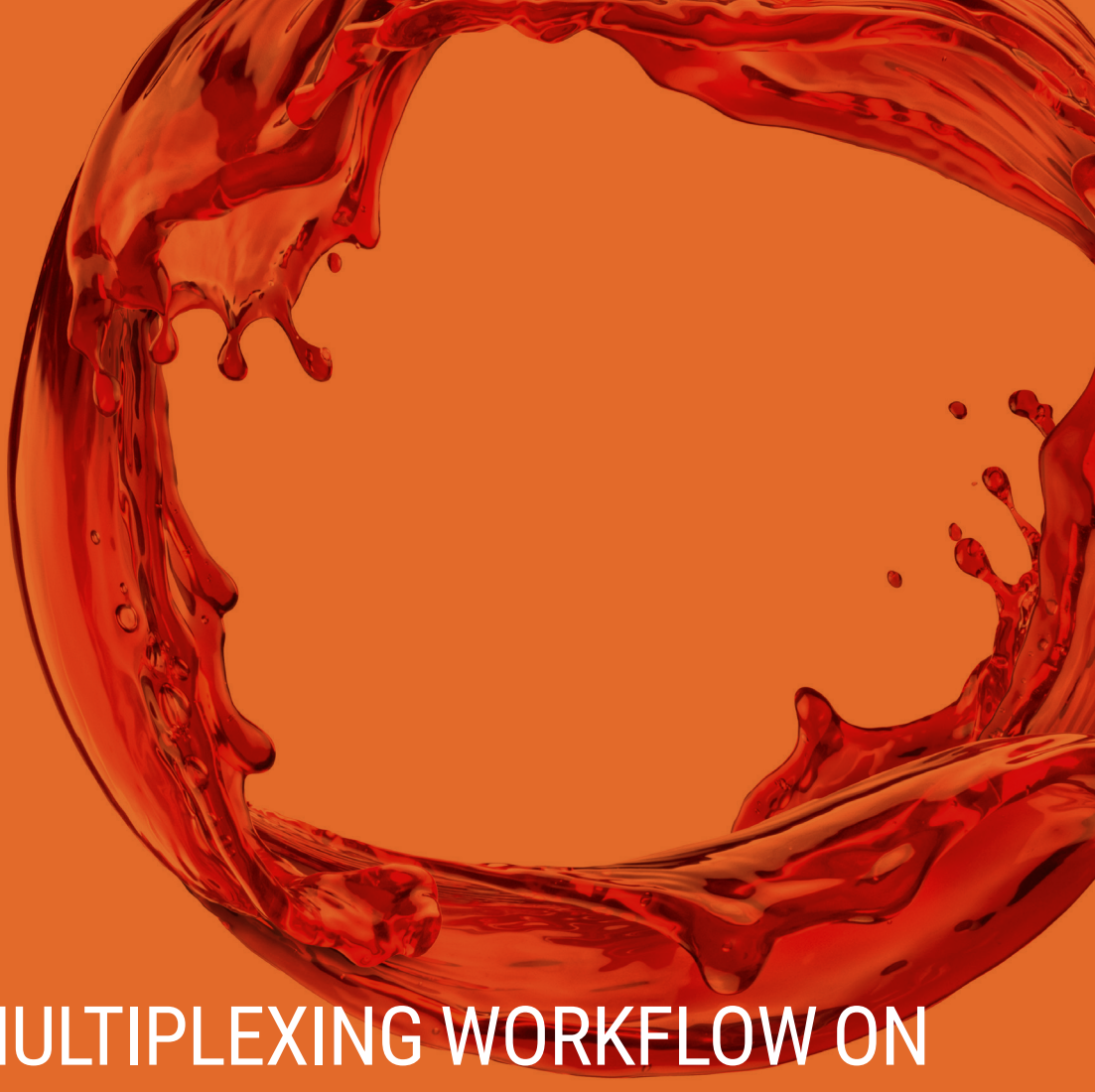


Application note

MICROBIAL MULTIPLEXING WORKFLOW ON THE SEQUEL® SYSTEM USING THE SMRTBELL® EXPRESS TEMPLATE PREP KIT 2.0

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.

PacBio

Our streamlined **microbial multiplexing procedure** for the Sequel system, from library preparation to genome assembly, can be completed with less than 8 hours bench time (Figure 1). Starting with high-quality genomic DNA (gDNA), samples are sheared to approximately 12 kb distribution, ligated with barcoded overhang 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.

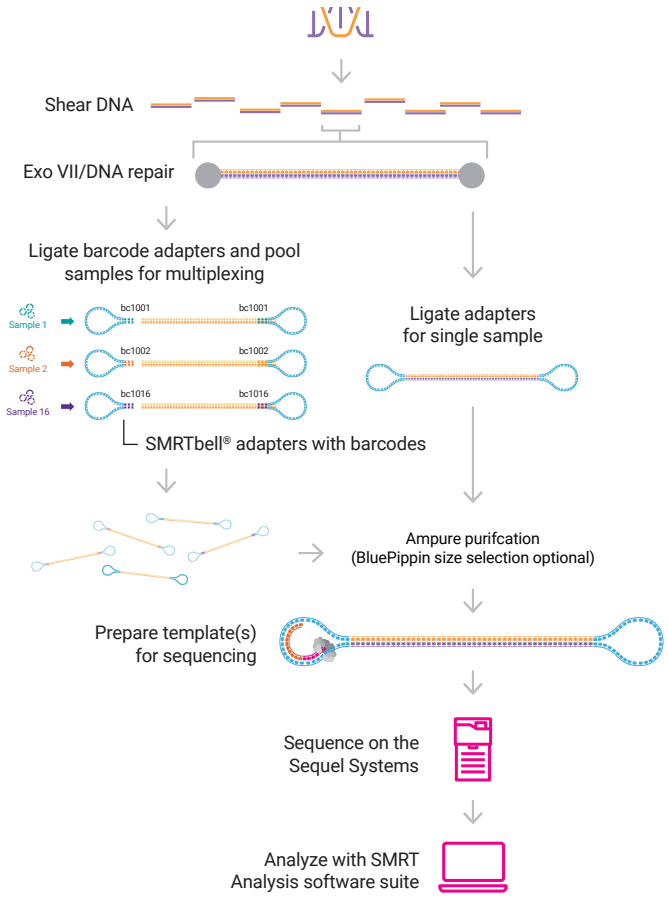


Figure 1. Overview of multiplexed microbial sequencing workflow on the Sequel systems.

Multiplex samples with validated barcoded adapters and a streamlined workflow

Two new barcoded SMRTbell adapter kits, with 8 barcoded overhang 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 overhang 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 >99.99% accuracy to resolve SNPs and structural variants
- Cost savings with a multiplexed workflow that takes advantage of SMRTbell library generation in ~4 hours

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 may be difficult to close with 12 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 for microbes with a range of genome complexity have resulted in closed microbial genomes with a 10-plex experimental design containing a mix of PacBio® internal control sample and microbes obtained from other sources such as Center for Food Safety and Nutrition (CFSAN). Ten microbial genomes from different sources representing a range of genome complexity and size, as well as gram status, were pooled to a total genome size of 42.3 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)
BC1001	*Escherichia coli K12 MG1655*[1]	-	4,653,240
BC1002	*Escherichia coli K12*[2]	-	4,653,240
BC1009	*Klebsiella pneumoniae ATCC BAA-2146*[3]	-	5,781,501
BC1010	*Bacillus cereus strain 971*[3]	+	5,427,083
BC1012	*Listeria monocytogenes CFSAN008100*[4]	+	3,032,269
BC1015	*Shigella sonnei CFSAN030807*[4]	-	5,062,953
BC1016	*Neisseria meningitidis FAM18*[3]	-	2,194,961
BC1018	*Staphylococcus aureus subsp. aureus ATCC 25923*[3]	+	2,806,345
BC1019	*E. coli strain W*	-	5,005,347
BC1022	*Staphylococcus aureus HPV107*[3]	+	2,901,406
		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 (lowercase) for (2) Pacbio control, (3) microbes derived from ATCC, (4) and CFSAN microbial genomes. Sequel system Instrument control software v6.0 and sequel binding and sequencing kits 3.0 were used. Movie time was set to 10 hours.

Microbial multiplexing sample preparation workflow

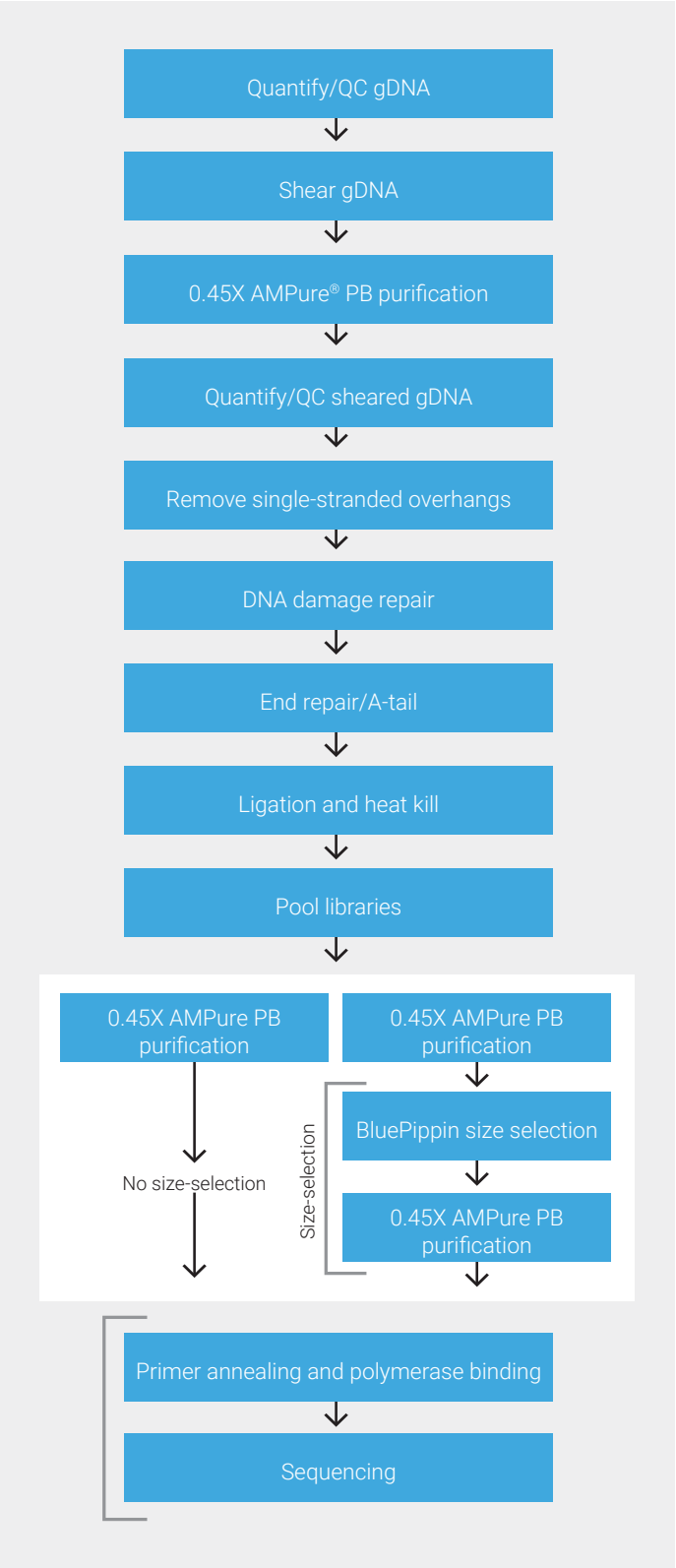
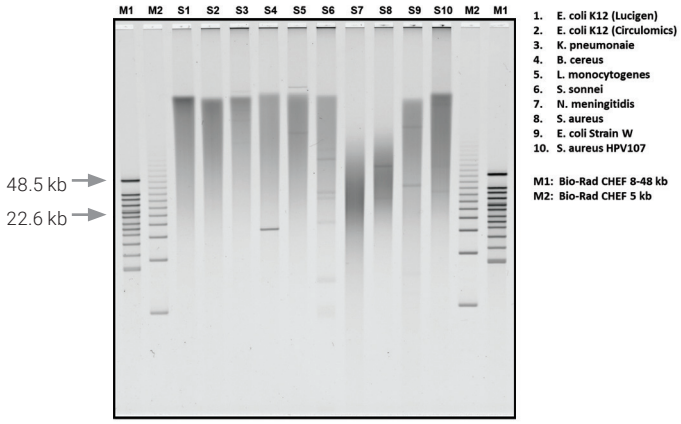


Figure 2. Microbial multiplexing sample preparation workflow with barcoded overhang adapters. Size-selection is optional. Overall workflow can be completed in approximately 8.5 hours (lowercase), but is dependent on multiplexing design for pooled libraries, QC method used, and whether size-selection is employed. Library construction of SMRTbell libraries is completed in 4 hours.

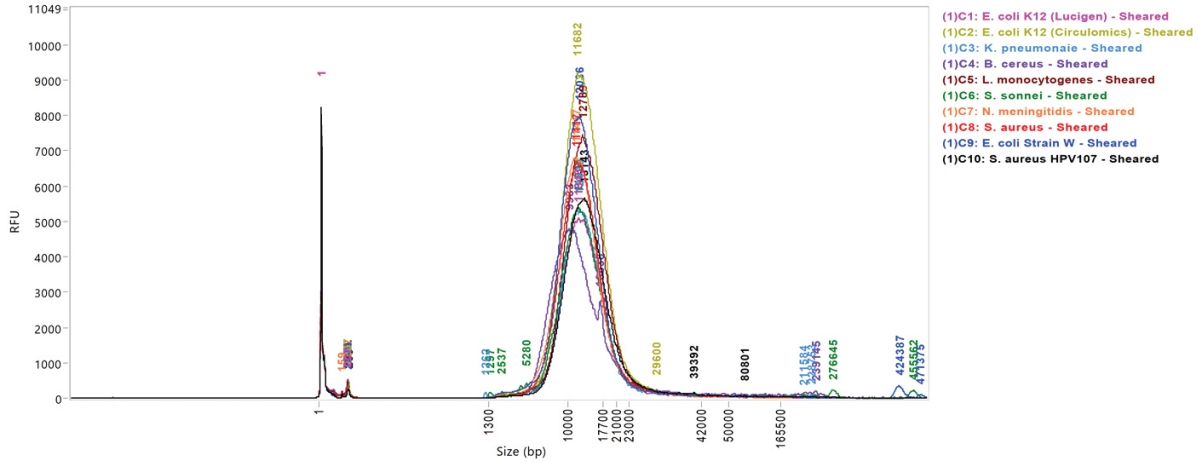
Sample preparation: Start with high-quality microbial DNA

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.

(a)



(b)



(c)

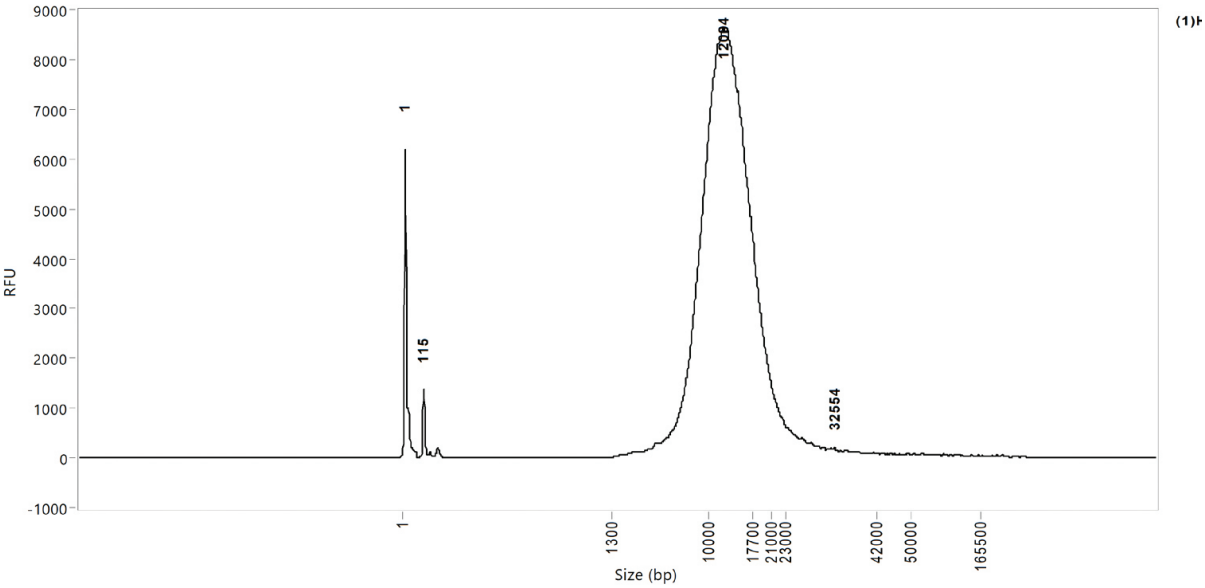


Figure 3. Genomic DNA and shear QC. **(a)** CHEF Mapper gel image showing all gDNA sample quality predominantly >20 kb. **(b)** FEMTO *Pulse* electropherogram of 10 samples sheared to a target 12 kb distribution, and **(c)** pooled SMRTbell libraries retained similar size distribution after the library preparation workflow.

Starting with high-quality gDNA predominantly >20 kb helps ensure an even 12 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 **Express 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.

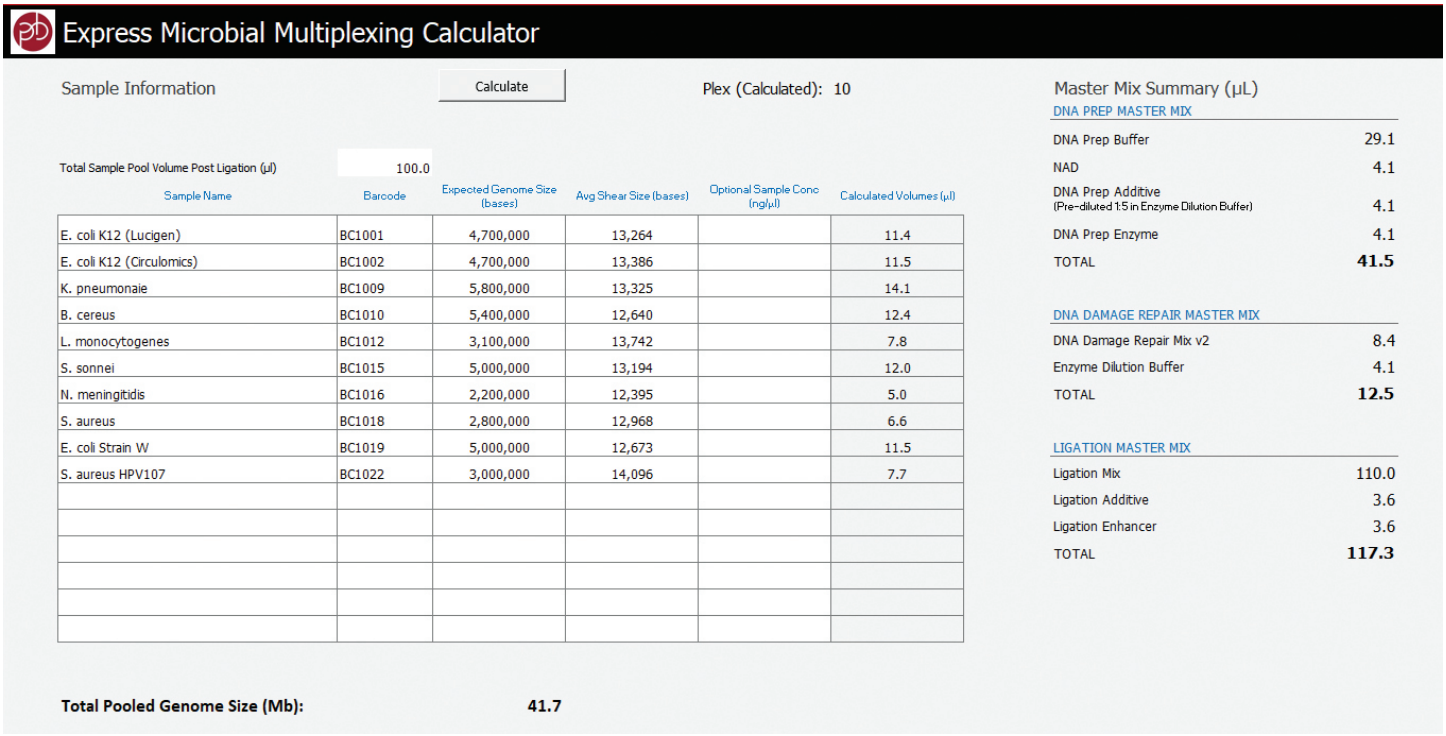


Figure 4. The Express microbial multiplexing calculator helps to ensure equimolar pooling of samples with respect to average shear and genome size. Optional concentration input serves to further optimize pooling ratio between samples. Calculator includes master mix calculations for generating SMRTbell 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 **Loading and pre-extension time recommendations for the Sequel system — Quick reference card.** Generally, we highly recommend targeting a productive loading fraction (P1) ranging from 50–65% for a total throughput of approximately 18 Gb per SMRT Cell (Figure 5). Our experiences have shown both underloading and overloading negatively impact SMRT sequencing yield, resulting in insufficient genome coverage for high-contiguity assemblies. 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.

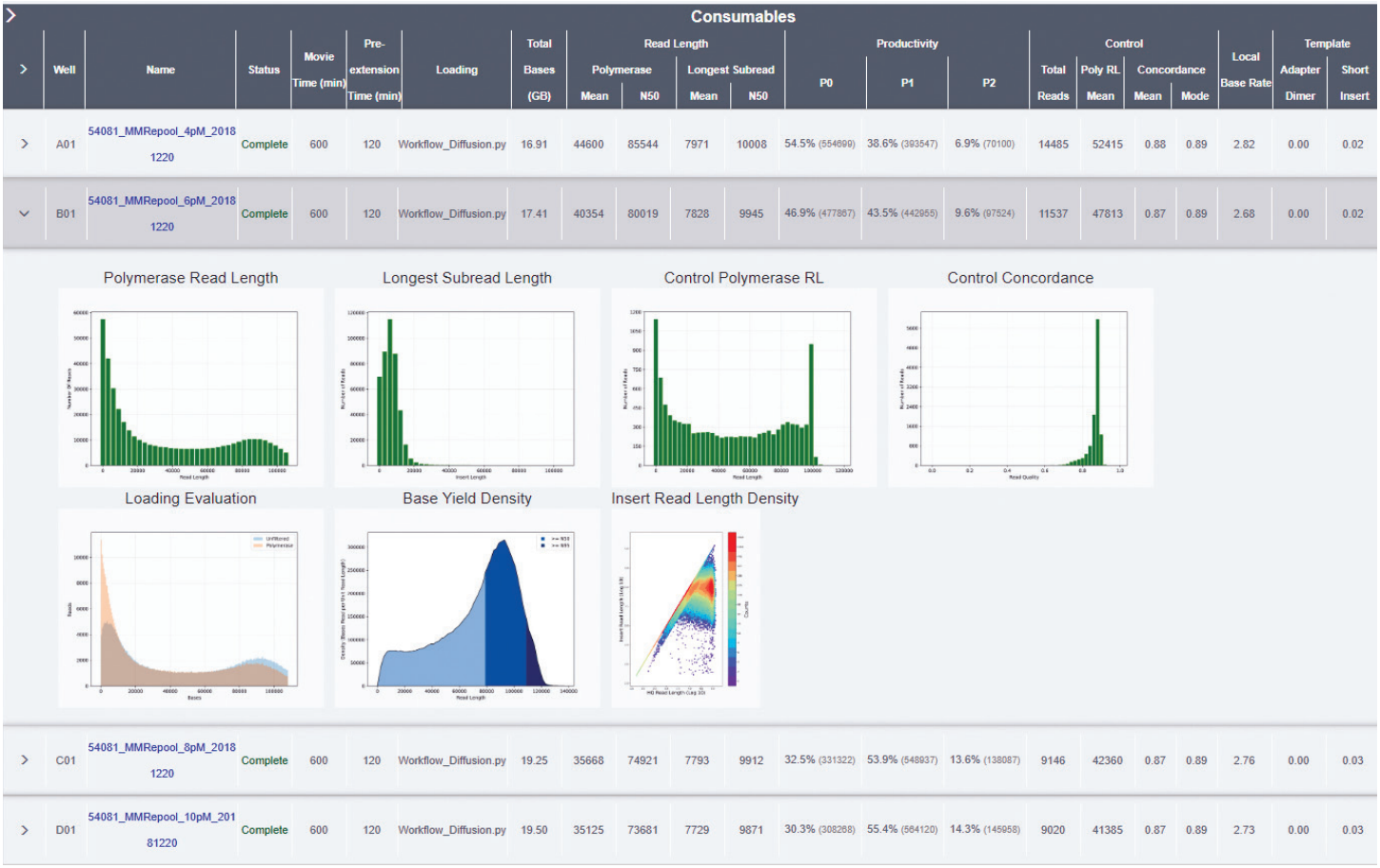


Figure 5. Run QC from SMRT Link with a range of loading on-plate concentration from 4 to 10 pM which yield productive fraction loading of P1 from 38.6 to 55.4% respectively. Optimal loading at 6 to 8 pM shows good yield at approximately 18 Gb and average polymerase read length of 40 kb. Underloading may yield insufficient coverage for samples, whereas overloading may result in a 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 21,520 to 43,465, with 31,769 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 indeterminate 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)
bc1001_BAK8A_OA--bc1001_BAK8A_OA	0--0	bc1001_BAK8A_OA--bc1001_BAK8A_OA	34,390	249,854	1,703,923,994	50,045	49,180	81.0	3
bc1002_BAK8A_OA--bc1002_BAK8A_OA	1--1	bc1002_BAK8A_OA--bc1002_BAK8A_OA	30,038	221,011	1,444,257,809	48,595	87,490	81.0	7
bc1009_BAK8A_OA--bc1009_BAK8A_OA	4--4	bc1009_BAK8A_OA--bc1009_BAK8A_OA	33,122	222,004	1,522,010,412	46,409	62,931	82.0	4
bc1010_BAK8A_OA--bc1010_BAK8A_OA	5--5	bc1010_BAK8A_OA--bc1010_BAK8A_OA	43,465	403,591	2,429,149,703	56,545	69,536	82.0	1
bc1012_BAK8A_OA--bc1012_BAK8A_OA	7--7	bc1012_BAK8A_OA--bc1012_BAK8A_OA	22,261	180,436	1,250,808,368	56,760	93,324	82.0	9
bc1015_BAK8B_OA--bc1015_BAK8B_OA	8--8	bc1015_BAK8B_OA--bc1015_BAK8B_OA	31,232	237,573	1,604,537,231	51,906	94,725	80.0	6
bc1016_BAK8B_OA--bc1016_BAK8B_OA	9--9	bc1016_BAK8B_OA--bc1016_BAK8B_OA	21,520	188,806	1,041,570,446	49,036	69,828	81.0	10
bc1018_BAK8B_OA--bc1018_BAK8B_OA	11--11	bc1018_BAK8B_OA--bc1018_BAK8B_OA	32,777	302,187	1,906,515,431	58,822	84,677	81.0	5
bc1019_BAK8B_OA--bc1019_BAK8B_OA	12--12	bc1019_BAK8B_OA--bc1019_BAK8B_OA	42,243	360,017	2,131,087,130	51,044	78,554	81.0	2
bc1022_BAK8B_OA--bc1022_BAK8B_OA	15--15	bc1022_BAK8B_OA--bc1022_BAK8B_OA	26,650	221,194	1,563,927,397	59,280	58,001	80.0	8
No Name	None	Not Barcoded	113,707	130,021	624,961,364	11,505	119,323	0.0	NA

Figure 6. Barcode QC after sample demultiplexing with SMRT Link software. In this example, 74% of polymerase 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. More relevantly, we observed 95% barcoded subreads and 96% barcoded bases which will be used for the microbial genome assembly.

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 using the SMRTbell express template prep kit 2.0 as detailed in the **Analysis procedure – Multiplexed microbial assembly with SMRT Link v6.0.0 and express template prep kit 2.0.**

An important metric to monitor is pre-assembly yield, which is reported in the assembly output and should be >60% and >30-fold filtered subread coverage per genome for high-contiguity assemblies (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).

Barcode ID	Sample ID	Expected genome size (bp)	Polished contigs (#)	Max contig length (bp)	N50 contig length (bp)	Sum of contig lengths (bp)	Pre-assembled yield (%)	Filtered subread coverage (fold)	Assembly notes
BC1001	*E. coli K12 (Lucigen)*	4,653,240	1	4,642,499	4,642,499	4,642,499	93.1%	64	Complete chromosomal assembly
BC1002	*E. coli K12 (Circulomics)*	4,653,240	1	4,642,500	4,642,500	4,642,500	92.9%	55	Complete chromosomal assembly
BC1009	*K. pneumoniae*	5,781,501	5	5,435,746	5,435,746	5,746,850	92.5%	50	Complete chromosomal assembly, and 140 kb and 85 kb plasmid assemblies. 118 kb plasmid captured in 2 contigs and missing 2 kb plasmid
BC1010	*B. cereus*	5,427,083	2	5,408,315	5,408,315	5,423,588	92.7%	59	Complete chromosomal and 16 kb plasmid assembly
BC1012	*L. monocytogenes*	3,032,269	2	3,043,149	3,043,149	3,137,529	93.7%	66	Complete chromosomal assembly in a single contig
BC1015	*S. sonnei*	5,062,953	1	4,813,454	4,813,454	4,813,454	93.1%	53	Complete chromosomal assembly. Missing eight expected plasmids
BC1016	*N. meningitidis*	2,194,961	1	2,213,947	2,213,947	2,213,947	92.2%	74	Complete chromosomal assembly
BC1018	*S. aureus*	2,806,345	2	2,778,860	2,778,860	2,806,350	92.6%	92	Complete chromosomal and 27 kb plasmid assembly
BC1019	*E. coli strain W*	5,005,347	2	4,898,327	4,898,327	5,004,399	93.0%	68	Complete chromosomal and 103 kb plasmid assembly. Missing 5 kb plasmid
BC1022	*S. aureus HPV107*	2,901,406	2	2,962,786	2,962,786	2,994,972	93.4%	82	Complete chromosomal and 24 kb plasmid assembly

Table 2. Summary of HGAP4 genome assembly results for a 10-plex microbial pool run. We recommend a >60% minimum pre-assembly yield and >30-fold filtered subread coverage per genome for high-contiguity assemblies. Complete microbial chromosomal assemblies were captured for all microbes in the pooled run, along with some plasmids.

Barcode ID	Sample ID	Genome complexity class	GC content (%)	Max contig length (bp)	# of contigs (main chromosomal assembly)	Concordance w/ NCBI reference (QV)
BC1001	*E. coli K12 (Lucigen)*	I	50.8	4,642,499	1	56
BC1002	*E. coli K12 (Circulomics)*	I	50.8	4,642,500	1	57
BC1009	*K. pneumoniae*	II	57.0%	5,435,746	1	48
BC1010	*B. cereus*	I	35.3%	5,408,315	1	35
BC1012	*L. monocytogenes*	I	37.9%	3,043,149	1	46
BC1015	*S. sonnei*	II	51.0%	4,813,454	1	62
BC1016	*N. meningitidis*	I	51.6%	2,213,947	1	50
BC1018	*S. aureus*	I	32.9%	2,778,860	1	56
BC1019	*E. coli strain W*	III	50.8%	4,898,327	1	44
BC1022	*S. aureus HPV107*	III	32.9%	2,962,786	1	33

Table 3. Summary of genome assembly results after circularization and polishing for improved base quality scores. Maximum contig length reported as assumption for genome size for main chromosomal genomes captured in a single contig assembly. QV scores compared to available NCBI reference genomes are shown; a QV 50 score indicates 99.999% concordance. Calculation of QV scores will be depending on the quality of reference genome available for comparison, and discrepancies will need to be validated by a secondary technology such as PCR based approaches. Genome complexity class defined by survey conducted by **Koren et al. 2013.**

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 to achieve 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 except one, which was recovered in a size-selected library. Size-selection may potentially improve assemblies, but 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 pre-assembly coverage. Overall, the workflow demonstrates robustness and reproducibility across the four replicates.

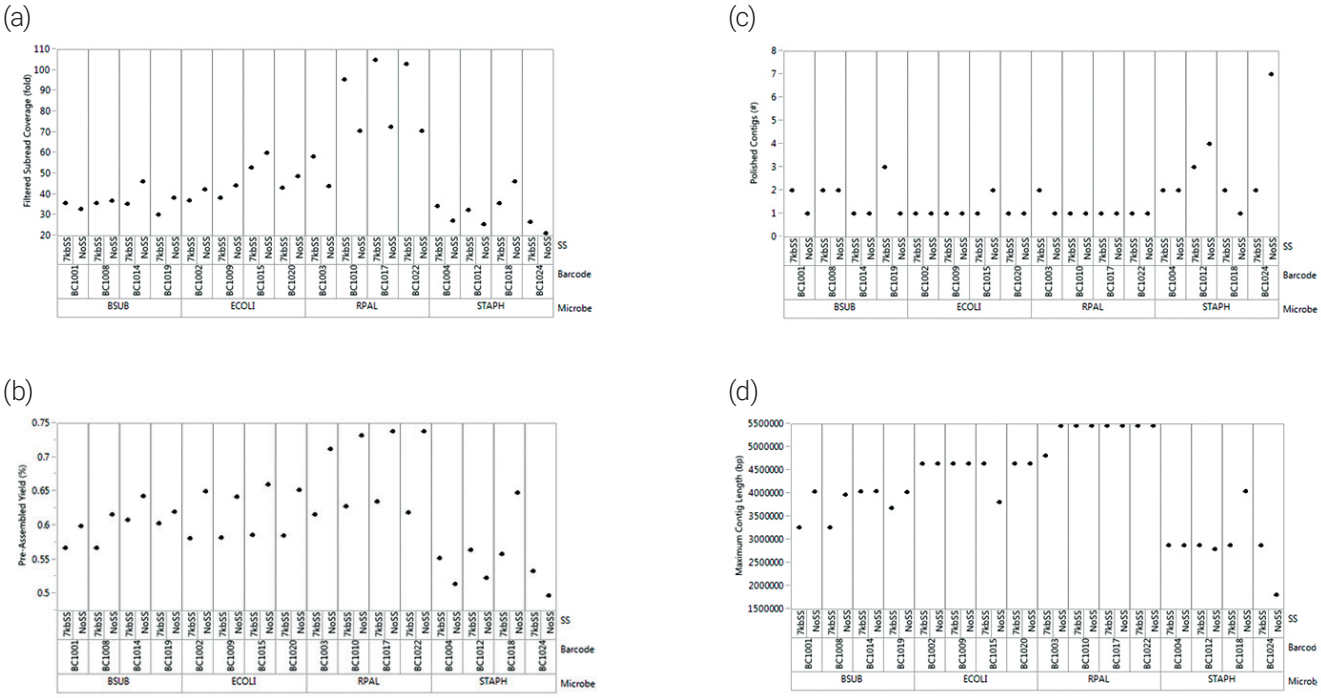


Figure 7. Genome sequencing and assembly results for a 68 Mb pooled library design and a 20-hour movie data collection. **(a)** Filtered subread coverage shows the importance of equimolar pooling for even genome coverage across multiplexed samples to avoid underrepresentation of samples. **(b)** Pre-assembly 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 allows for a significantly reduced cost per microbe. This tradeoff may be very attractive and should be assessed based on purposes of your scientific research.

Pacbio consumable part numbers

Part number	Item
101-628-400	PacBio barcoded overhang adapter kit 8a
101-628-500	PacBio barcoded overhang adapter kit 8b
100-938-900	SMRTbell express template prep kit 2.0
101-633-500	Elution buffer (50 mL)

Ancillary 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

Resources

- **Procedure + checklist — Preparing multiplexed microbial libraries using SMRTbell express template prep kit 2.0**
- **Analysis procedure — Multiplexed microbial assembly with SMRT Link v6.0.0 and express template prep kit 2.0**
- **Express microbial multiplexing calculator**
- **Quick reference card — 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 data set**

Headquarters
1305 O'Brien Drive
Menlo Park, CA 94025
United States
Phone: 1.650.521.8000

Customer service
Phone: 1.877.920.PACB (7222);
option 1
Fax: 1.650.618.2699
orders@pacb.com

Technical support
Phone: 1.877.920.PACB (7222);
option 2
support@pacb.com

As a global company, we have offices in countries around the world. Visit **pacb.com/contact-us** for contact info.

READY TO GET STARTED WITH HIFI SEQUENCING?

Explore the applications:
pacb.com/applications

Connect with PacBio for more info:
North America: **nasales@pacb.com**
South America: **sasales@pacb.com**
EMEA: **emea@pacb.com**
Asia Pacific: **apsales@pacb.com**

Contact a certified service provider
pacb.com/CSP

101-588-800 V3 17DEC2021