Complete genome sequences of pooled genomic DNA from 10 marine bacteria using PacBio long-read sequencing

Weizhi Song[1, 2], Torsten Thomas[2, 3, *] and Richard J. Edwards[1, *]

[1]School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW, Australia

[2]Centre for Marine Bio-Innovation, University of New South Wales, Sydney, Australia

[3]School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, NSW, Australia

***Corresponding author address:**

Torsten Thomas
School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, NSW, Australia.
Tel: +61-293853467
Email: t.thomas@unsw.edu.au

Richard J. Edwards
School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW, Australia.
Tel: +61-293850490
Email: richard.edwards@unsw.edu.au

Abstract

Background: High-quality, completed genomes are of critical importance for understanding of the functions of marine bacteria and their interactions with host organisms. PacBio sequencing technology provides a powerful way to get high-quality completed genomes or closing gaps of current draft genomes.

Findings: Pooled genomic DNA from ten marine bacteria was sequenced with eight SMRT cells on the PacBio sequencing platform. In total, 7.35 Gbp of long reads were generated, which is equivalent to an approximate 168X coverage for the input genomes. Genome assembly showed that eight genomes with average nucleotide identities (ANI) lower than 90.8% can be assembled with high-quality and completion using standard assembly algorithms (e.g. HGAP or Canu). A reference-based read phasing step was developed and incorporated to assemble the complete genomes of the remaining two bacteria that had an ANI > 97% and whose initial assemblies were highly fragmented.

Conclusion: Ten complete high-quality genomes of marine bacteria were generated. The approached and findings made here, including the reference-based read phasing approach for the assembly of highly similar genomes, can be used in the future to design strategies to sequence pooled genomes using long-read sequencing.

Keywords: Marine bacteria; Genome sequencing; PacBio; Long-read sequencing; Reference-based reads phasing; SAMPhaser; Assembly

Introduction

Marine bacteria can play important roles in the development, defense and health of higher host organisms, such as seaweeds or sponges [1, 2]. The availability of high-quality reference genomes of such organims is critical for a better understanding of their functions and interactions with hosts. Short-read sequencing technologies, e.g. Illumina and 454, often fail to generate completed genomes due to sequencing biases, repetitive genomic features or genomic polymorphism [3]. Pacific Biosciences (PacBio) SMRT™ sequencing provides a powerful way to get high-quality complete genome or closing gaps of current draft genomes with its long reads [4]. However, library preparation for individual genomes results in a relatively high cost for PacBio-based microbial sequencing projects. Here, we aimed to overcome this issue by PacBio SMRT™ sequencing of pooled genomic DNA for ten marine bacterial strains with various degrees of genome similarity. We also introduce a reference-based read phasing strategy using SAMPhaser [5] for the assembly of highly similar genomes (ie. average nucleotide identities (ANI) > 97%) from PacBio reads of pooled genomic DNA samples. Using this pooled sequencing, we have produced ten complete high-quality genome assemblies using a single SMRTbell library.

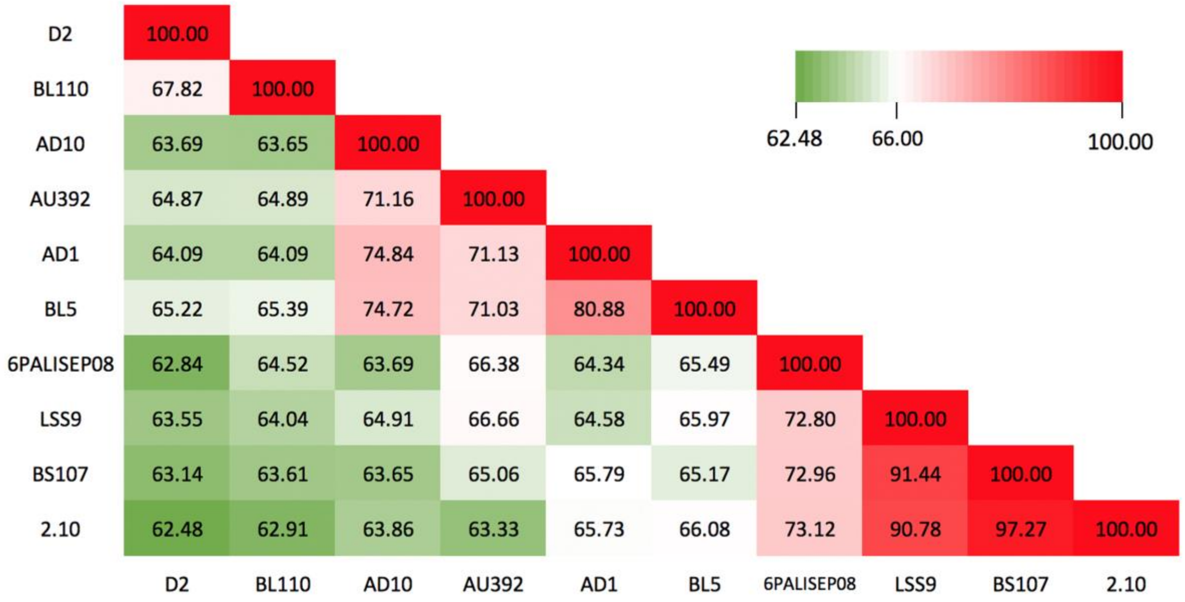
Strain selection

Ten marine bacterial strains, isolated from various marine hosts, were selected for genome sequencing (Table 1): *Aquimarina* sp. AD1, AD10 and BL5 as well as *Alteromonas* sp. BL110, *Phaeobacter* sp. LSS9, and *Ruegeria* sp. AD91A were isolated from the red seaweed *Delisea pulchra* [6, 7]; *Pseudoalteromonas tunicata* D2 was isolated from the surface of the tunicate *Ciona intestinalis* [8]; *Phaeobacter inhibens* 2.10 was isolated from the surface of the green alga *Ulva lactuca* [9]; *Phaeobacter inhibens* BS107 was isolated from the scallop *Pecten maximus* [10]; *Flavobacteriaceae bacterium* AU392 was isolated from the sponge *Tedania anhelens* [11]. All strains had been propagated for multiple rounds in the laboratory and existed as pure cultures. Previous draft genomes (Illumina MiSeq) or completed reference genomes were available for nine of the selected strains (Table 1). As no reference sequences was available for strain *Ruegeria* sp. AD91A, the *Ruegeria* sp. 6PALISEP08 genome (GenBank accession: NZ_LGXZ00000000), which had a 16S rRNA gene similarity of 99.45%, was selected from the National Center for Biotechnology Information (NCBI) RefSeq database and used as reference. Pairwise ANI of the reference genomes were calculated with OrthoANI [12] and ranged from 62.48% to 97.27% (Fig. 1). GC content of the reference genomes ranges from 30.7% to 60.3% (Table 1).

Table 1: Selected strains and the status of their genomes from previous studies

Sequenced strain	Abbrev.	Reference	Contigs	Status	Approximate genome size (Mbp)	GC (%)
Alteromonas sp. BL110	BL110	BL110*	10	Draft	4.2	44.1
Aquimarina sp. AD1	AD1	AD1*	519	Draft	5.1	32.1
Aquimarina sp. AD10	AD10	AD10*	88	Draft	3.2	32.4
Flavobacteriaceae bacterium AU392	AU392	AU392*	12	Draft	6.2	30.7
Aquimarina sp. BL5	BL5	BL5*	353	Draft	5.6	32.9
Pseudoalteromonas tunicata D2	D2	D2*	42	Draft	4.8	39.9
Phaeobacter sp. LSS9	LSS9	LSS9*	50	Draft	3.9	60.3
Phaeobacter inhibens 2.10	2.10	2.10 [10]	4 (3 plasmids)	Completed	4.0	59.8
Phaeobacter inhibens BS107	BS107	BS107 [10]	4 (3 plasmids)	Completed	4.0	59.8
Ruegeria sp. AD91A	AD91A	6PALISEP08	42	Draft	4.3	57.0

*Unpublished.



	D2	BL110	AD10	AU392	AD1	BL5	6PALISEP08	LSS9	BS107	2.10
D2	100.00									
BL110	67.82	100.00								
AD10	63.69	63.65	100.00							
AU392	64.87	64.89	71.16	100.00						
AD1	64.09	64.09	74.84	71.13	100.00					
BL5	65.22	65.39	74.72	71.03	80.88	100.00				
6PALISEP08	62.84	64.52	63.69	66.38	64.34	65.49	100.00			
LSS9	63.55	64.04	64.91	66.66	64.58	65.97	72.80	100.00		
BS107	63.14	63.61	63.65	65.06	65.79	65.17	72.96	91.44	100.00	
2.10	62.48	62.91	63.86	63.33	65.73	66.08	73.12	90.78	97.27	100.00

Figure 1: Pairwise average nucleotide identities (ANI) between reference genomes.

Genomic DNA extraction, library construction and sequencing

Genomic DNA (gDNA) was extracted from pure cultures using the DNeasy Blood & Tissue Kits (QIAGEN, Hilden, Germany). Concentrations were measured by Qubit (Invitrogen, USA) and genomes were mixed together in equal molarity. The mixed gDNA was subjected to 15-50 kb BluePippin size selection (Sage Science, Beverly, MA, USA) and a single library was prepared using the SMRTbell template preparation kit 1.0 (Pacific Biosciences, Menlo Park, CA) according to the manufacturer’s instructions. Recovered fragments were sequenced using the P6C4 sequencing chemistry on the RS II platform (240 min movie time). Sequencing on eight SMRT cells generated 7.35 Gbp raw data (Table 2), which is equivalent to an approximate 168X coverage of input genomes.

Table 2: Statistics of subreads from the eight SMRT cells

SMRT cell	Read count	N50 (bp)	Mean length (bp)	Total length (Gbp)
1	77,208	14,983	10,277	0.74
2	57,267	15,015	10,132	0.54
3	114,378	14,772	10,430	1.11
4	105,227	14,828	10,531	1.03
5	106,107	14,935	10,534	1.04
6	39,942	14,936	10,584	0.39
7	125,170	14,666	10,444	1.22
8	130,439	14,675	10,447	1.27

De novo whole genome assembly

Subreads were assembled into 105 contigs (Table 3) using HGAP v3 [13] (default settings, genome size 50.5 Mbp) and polished with Quiver, as implemented in the SMRT Analysis suite through the SMRT Portal. HGAP contigs were mapped against reference genomes by performing a pairwise BLASTN (BLAST+ 2.6.0) [14], with a local alignment length cut-off of 5,000 bp, an identity cut-off of 80% for genome AD91A and 99% for the other nine genomes. Eight of the genomes produced assemblies in 1-5 pieces, with the two *Phaeobacter inhibens* genomes being highly fragmented (Fig. 2). Further analysis showed that some short contigs with reference assignment (hcq1, hcq9, hcq32 and hcq150) were sequences covering the break-point of their corresponding circular chromosome sequences. These sequences were marked as redundant sequences (Table 4) and excluded from further analysis. 16S rRNA gene sequences were identified from assemblies with Barrnap v0.9 [15] and percentage identities to those in reference genomes was calculated by performing pairwise BLASTN (Fig. 2).

Overlapping end regions of assembled circular chromosomes/plasmids were identified by BLASTN and contig circularisation performed manually by trimming contig ends to the middle of each overlapping region (Table 4). The break-point of contig hcq3 (from genome AD1) was located in a highly repetitive region, which made its circularisation difficult, and so this strain was re-assembled. Subreads were mapped onto the original assemblies with BLASR v3.1.1 [16] and those matching AD1 (hcq3 and hcq32) or failing to map were extracted with an in-house script (get_reads_from_sam.py) [17]. Extracted subreads were reassembled with Canu v1.7 [18], producing a single circular contig with a length of 5,483,011 bp. This contig was then manually circularised (Table 4, AD1_tig1). AD91A (hcq7, hcq43, hcq44) failed to generate a full-length chromosomal contig with overlapping ends. AD91A was therefore also re-assembled with Canu as described above. Two circular contigs (3,685,098 and 766,037 bp, respectively) assigned to reference genome 6PALISEP08 were identified and manually circularised (Table 4, AD9A1_tig1 and AD9A1_tig30). Circularised assemblies were further polished with Quiver and possible indels were corrected with Pilon [19] using their corresponding Illumina reads, if available (Table 5).

Table 3: Summary of HGAP produced assemblies

Type	Measurement
Total length (bp)	46,954,156
Number of Contigs	105
N50 of Contigs (bp)	4,010,148
Average length (bp)	447,182
Shortest contig (bp)	1,216
Longest contig (bp)	6,113,625

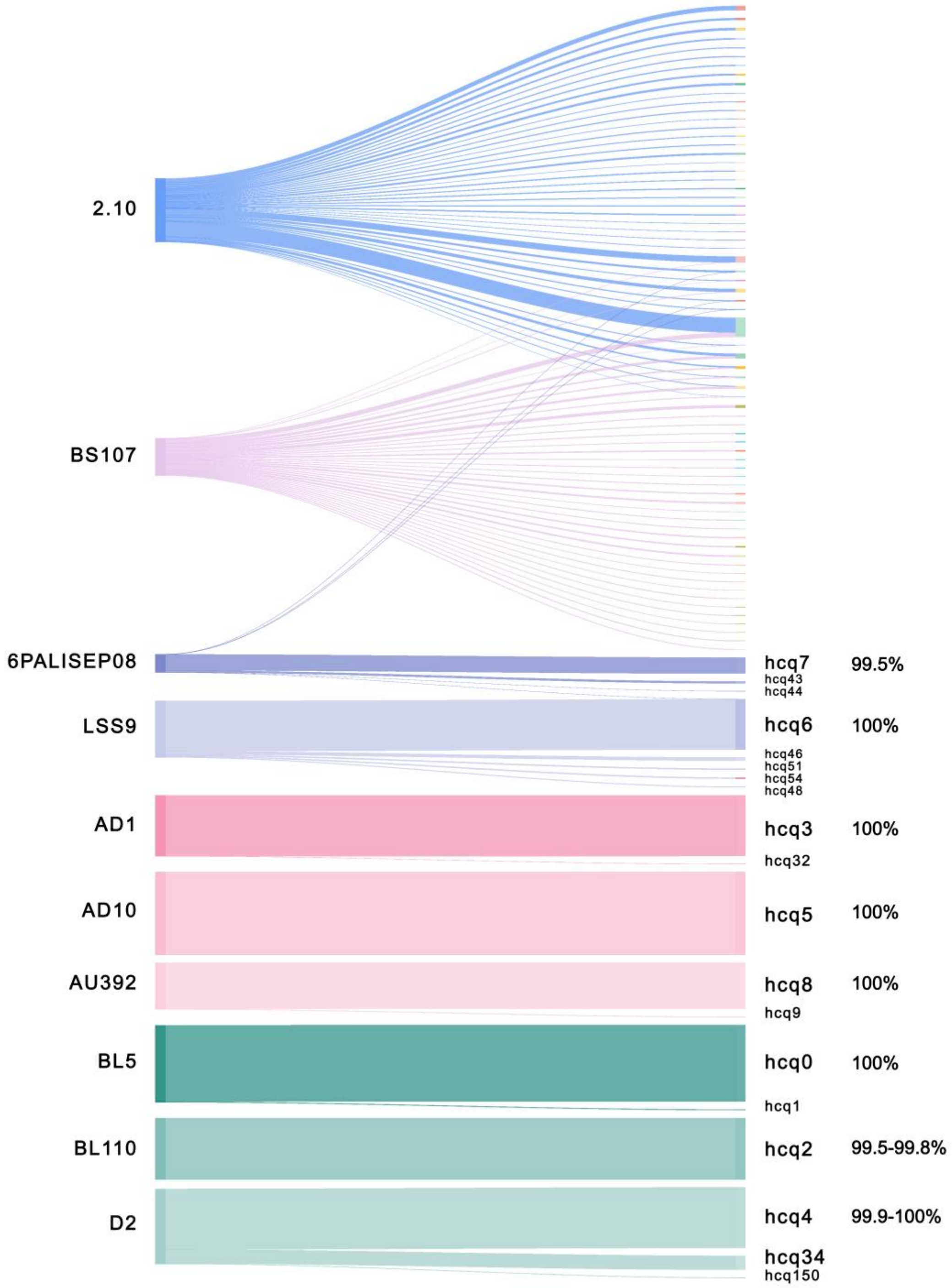


Figure 2: Correlations between the reference sequences (left) and HGAP produced assemblies (right). The alignment length for all blast matches passing the filtering criteria between each pair of reference and contig were summed up. The band width is proportional to the summation of aligned sequences between it connected reference sequences and HGAP produced assembly. The percentage identity between assembly and reference 16S rRNA gene sequences are given on the right.

Table 4: Summary of successful genome assemblies

Genome	Contig ID	Length (bp)	Circularity	Overlapping length (bp)	Overlapping identity (%)	Category
AD1	AD1_tig1[1]	5,461,560	Yes	21,563	99.77	Chromosome
AD10	hcq5	6,097,687	Yes	25,886	99.40	Chromosome
AD91A	AD91A_tig1[1]	3,662,296	Yes	22,889	99.94	Chromosome
	AD91A_tig30[1]	743,267	Yes	22,810	99.75	Plasmid
AU392	hcq8	3,372,961	Yes	20,697	99.73	Chromosome
	hcq9	49078	No	NA	NA	Redundant sequences[2]
BL5	hcq0	5,941,734	Yes	18,918	99.46	Chromosome
	hcq1	56,167	No	NA	NA	Redundant sequences[2]
BL110	hcq2	4,492,109	Yes	31,108	99.79	Chromosome
D2	hcq4	4,870,663	Yes	25,090	99.66	Chromosome 1
	hcq34	1,066,196	Yes	28,119	99.57	Chromosome 2
	hcq150	40,782	No	NA	NA	Redundant sequences[2]
LSS9	hcq6	3,692,517	Yes	24,434	99.73	Chromosome
	hcq46	234,169	Yes	25,974	99.72	Plasmid
	hcq48	93,928	Yes	23,190	99.77	Plasmid
	hcq51	89,716	Yes	21,325	99.71	Plasmid
	hcq54	58,676	Yes	15,980	99.81	Plasmid

[1] Reassembled contigs, not part of original HGAP3 assembly.

[2] Sequences crossing the break-point.

Table 5: The number of Pilon corrected indels

Genome	Corrected indels
AD1	4
AD10	28
AU392	2
AD91A	No NGS reads
BL5	6
BL110	23
D2	No NGS reads
LSS9	No NGS reads
2.10	No NGS reads
BS107	No NGS reads

Reference-guided assembly of two *Phaeobacter inhibens* genomes

Assemblies from two of the three *Phaeobacter inhibens* genomes (2.10 and BS107) were highly fragmented (Fig. 2), presumably due to high sequence similarity between them (Fig. 1 and Fig. 3). We therefore adopted a different, reference-guided assembly strategy for the two *Phaeobacter inhibens* genomes (Fig. 4).

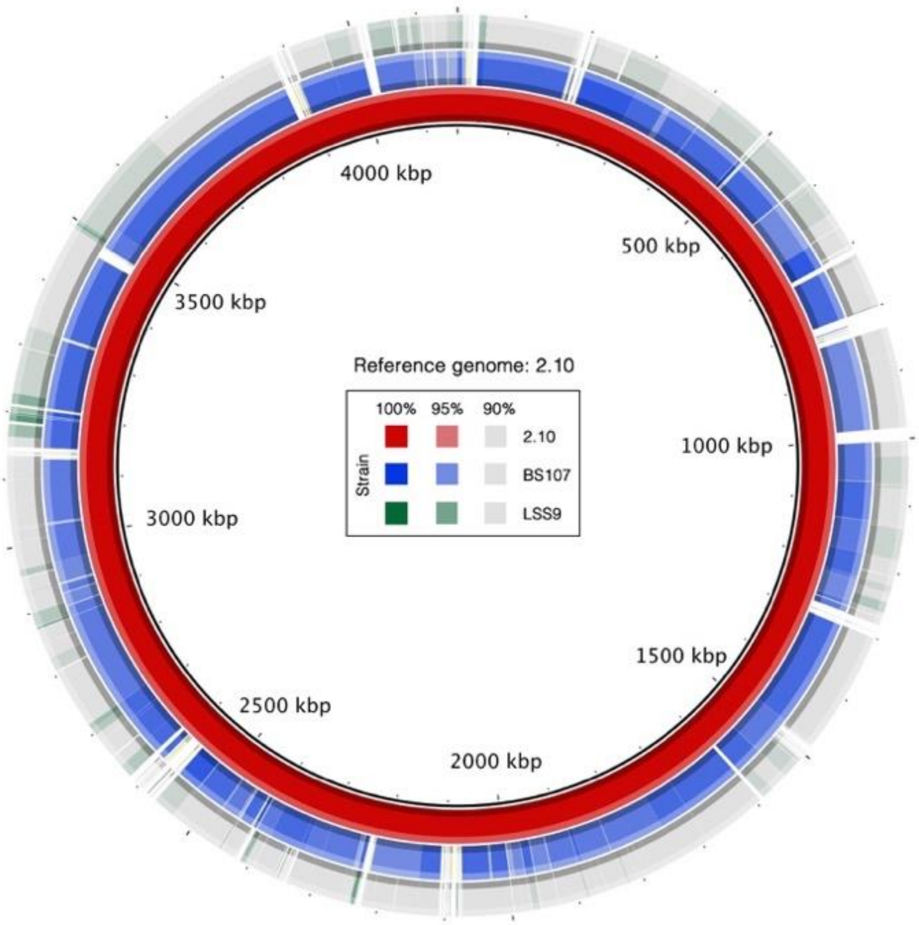


Figure 3: Sequence similarity between the three *Phaeobacter* reference genomes. Sequence similarity was calculated with BLASTN. Plot was generated using BLAST Ring Image Generator (BRIG) [20].

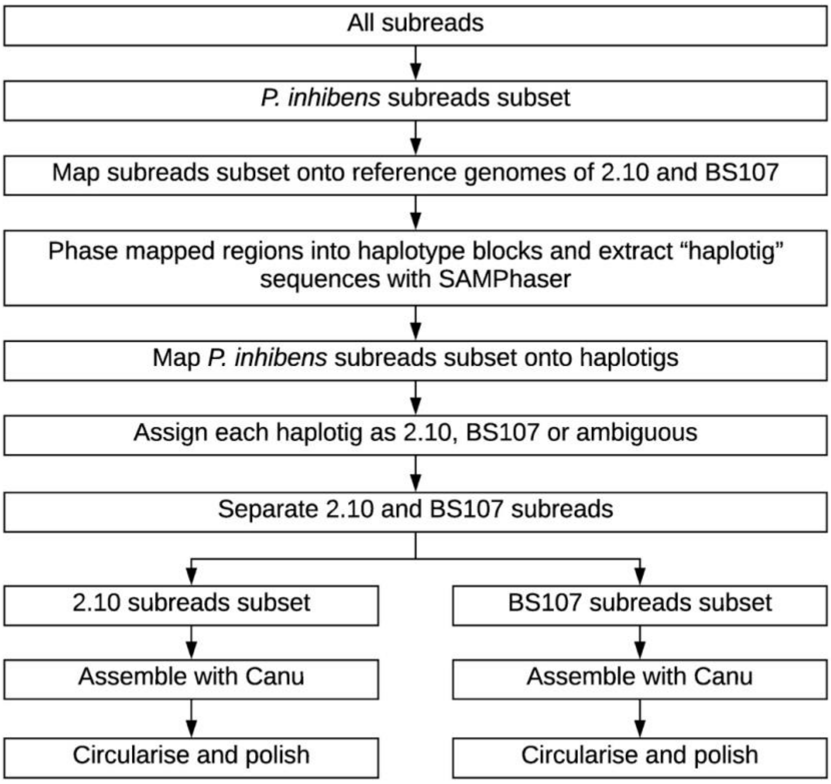

All subreads
P. inhibens subreads subset
Map subreads subset onto reference genomes of 2.10 and BS107
Phase mapped regions into haplotype blocks and extract "haplotig" sequences with SAMPhaser
Map *P. inhibens* subreads subset onto haplotigs
Assign each haplotig as 2.10, BS107 or ambiguous
Separate 2.10 and BS107 subreads
2.10 subreads subset
BS107 subreads subset
Assemble with Canu
Assemble with Canu
Circularise and polish
Circularise and polish

Figure 4: Work schematic of reference-guided genome assembly

In detail, subreads not mapped to the eight completed genome assemblies were extracted as described above as the *P. inhibens* subreads subset (Fig. 4). These subreads were mapped to a combined 2.10 and BS107 reference genomes with BLASR v3.1.1. Because of the high similarity between strains, we were not confident that 2.10 and BS107 subreads would be exclusively mapping to the correct reference. The next step was therefore to phase single nucleotide polymorphisms (SNPs) and extract 2.10 and BS107 haplotype blocks. This was performed using an in-house tool, SAMPhaser v0.5.0 [5].

SAMPhaser first identifies variants from a pileup file, generated from the BLASR BAM output using SAMtools v1.7 [21]. SNPs and indels were called for all positions where the minor allele was supported by at least 10% of the reads, with an absolute minimum of two reads. The subset biallelic SNPs with the minor variant supported by at least five reads at a frequency of at least 25% were used for phasing. Indels, and any SNPs not meeting these criteria, were used for sequence correction, but not phasing. Phasing is performed by iteratively assigning alleles and reads to haplotypes. Initially, each read is given an equal probability of being in haplotype "A" or “B”. The reference allele of the first SNP then defines haplotype A. For each SNP, SAMPhaser iteratively calculates (1) the probability that each allele is in haplotype A given the haplotype A probabilities for reads containing that allele, and then (2) the probability that each read is in haplotype A given the haplotype A probabilities for that read’s alleles at the last ten SNPs. This is performed by modelling a SNP call error rate (set at 5%) and then calculating the relative likelihood of seeing the observed data if a read or allele is really in haplotype A versus haplotype B. This progresses until all SNPs have been processed. If at any point, all reads with processed SNP positions reach their ends before another SNP is reached, a new phasing block is started. Draft phase blocks are then resolved into the final haplotype blocks by assigning reads and SNPs where the probability of assignment of a read to one haplotype exceeds 95%. Ambiguous reads and SNPs are ignored.

The final step is to “unzip” the reference sequence into “haplotigs”. SAMPhaser unzips phase blocks with at least five SNPs. Regions that are not unzipped are output as “collapsed” haplotigs (Fig. 5). First, phased reads are assigned to the appropriate haplotig. Regions of 100+ base pairs without coverage are removed as putative structural variants, and the haplotig split at this point. Haplotigs with an average depth of coverage below 5X are removed. Note that this can result in “orphan” haplotigs, where the minor haplotig did not have sufficient coverage for retention. Haplotigs ending within 10 bp of the end of the reference sequence are extended. Next, collapsed blocks are established by identifying reads that (a) have not been assigned to a haplotype, and (b) are not wholly overlapping a phased block. Finally, unzipped blocks have their sequences corrected. This is performed by starting with the reference sequence and then identifying the dominant haplotype allele (or consensus for collapsed blocks) at all variant positions (not just those used for phasing) providing the variant has at least three reads supporting it. The final haplotig sequence is the original reference sequence with any assigned non-reference alleles substituted in at the appropriate positions. Single base deletions are cut out of the sequence and so it may end up shorter than the original contig. Insertions and longer deletions are not currently handled and are ignored; for this reason, it is important to re-map reads and correct the final haplotig sequences.

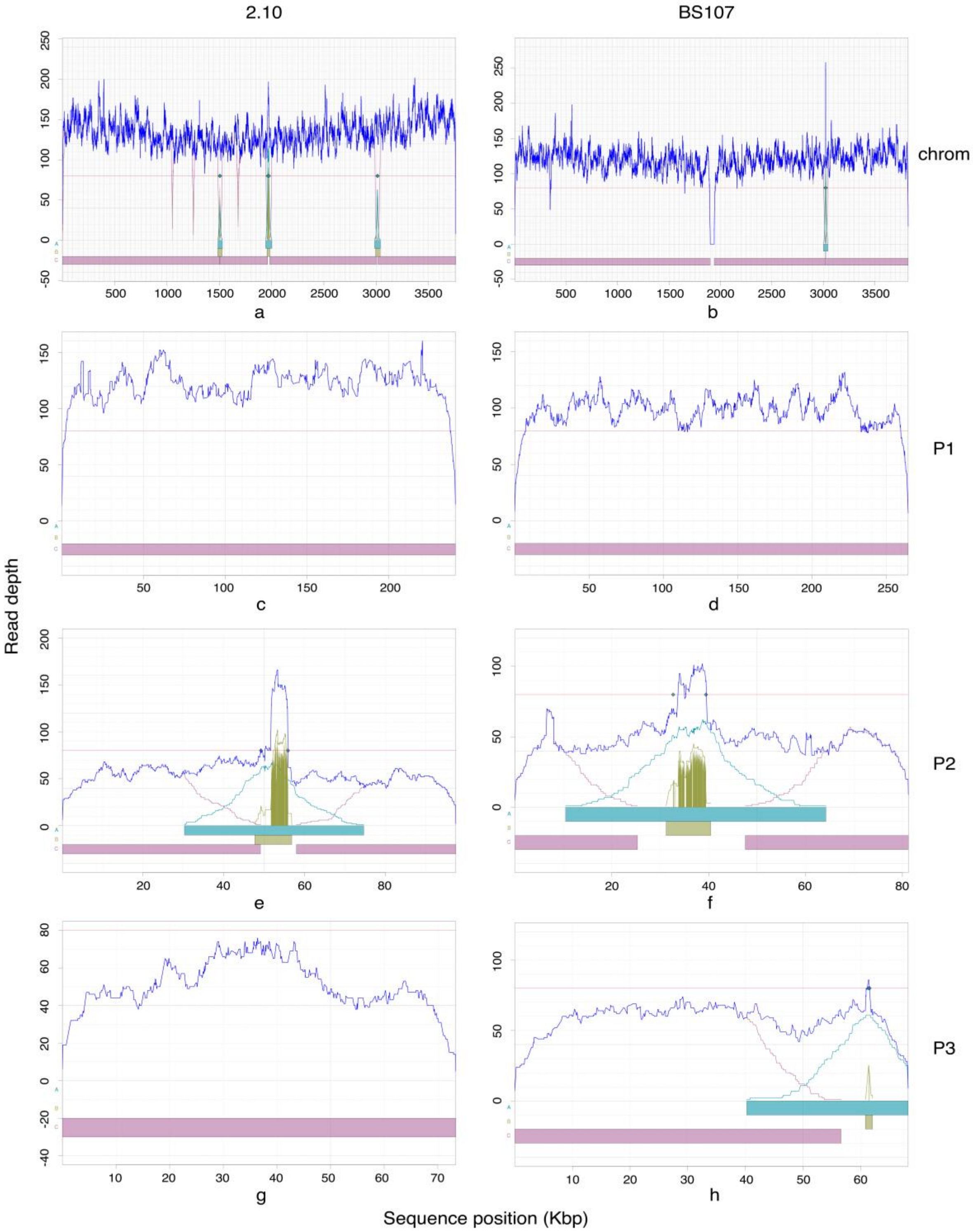


Figure 5: SAMPhaser phasing of combined *P. inhibens* subreads mapped onto combined reference genome of 2.10 and BS107. SAMPhaser phasing plot output of read depth versus chromosome position (Kbp) for **(a)** 2.10 chromosome, **(b)** BS107 chromosome, **(c)** 2.10 plasmid 1, **(d)** BS107 plasmid 1, **(e)** 2.10 plasmid 2, **(f)** BS107 plasmid 2, **(g)** 2.10 plasmid 3, and **(h)** BS107 plasmid 3. Depth traces are for all reads (blue) and reads assigned to phased track A (cyan) or track B (gold). Vertical lines indicate SNP positions, coloured by the track of the minor allele. The extent of phased blocks is shown as coloured bars below the plots, labelled A, B and C, and diamonds on the main depth trace mark the extent of SNPs within these blocks. Phased haplotig blocks themselves extend to the ends of the reads mapping to that haplotig. Track C indicates "collapsed" blocks that lack heterozygosity.

A haplotig "purity" statistic was used to assess the quality of SAMPhaser phased haplotigs. Purity was calculated using an in-house script (get_purity.py) [17] as follows:

1. Simulate short reads from the two reference genomes, with the number of simulated reads being in proportion to the sizes of the reference genomes.
2. Map the simulated reads to haplotigs with BBMAP v35.82 [22]. A read will not be mapped if multiple top-scoring mapping locations were found from the query sequences (specified with "ambiguous=toss").
3. Get purity for each query sequence by calculating the percentage of short reads mapped to it that come from each reference genome. The query sequences will be assigned to a reference genome or rated as "ambiguous" according to pre-defined purity cut-off (e.g. 80%).

The overall purity of all query sequences is calculated by:

$$Overall\ Purity = \frac{\sum_{i \in R} L_i P_i}{\sum_{i \in A} L_i}$$

Where A indicates all query sequences. L_i and P_i indicate the length and purity of query sequence *i*. R indicates the set of query sequences with reference assignments according to the pre-defined cut-off.

The purity of phased haplotigs was assessed by mapping one million 250 bp paired-end reads simulated from the 2.10 and BS107 reference genomes to the haplotigs. Haplotigs with fewer than 100 reads mapped were removed. The overall purity of the remaining SAMPhaser haplotigs was 99.99% (Fig. 6).

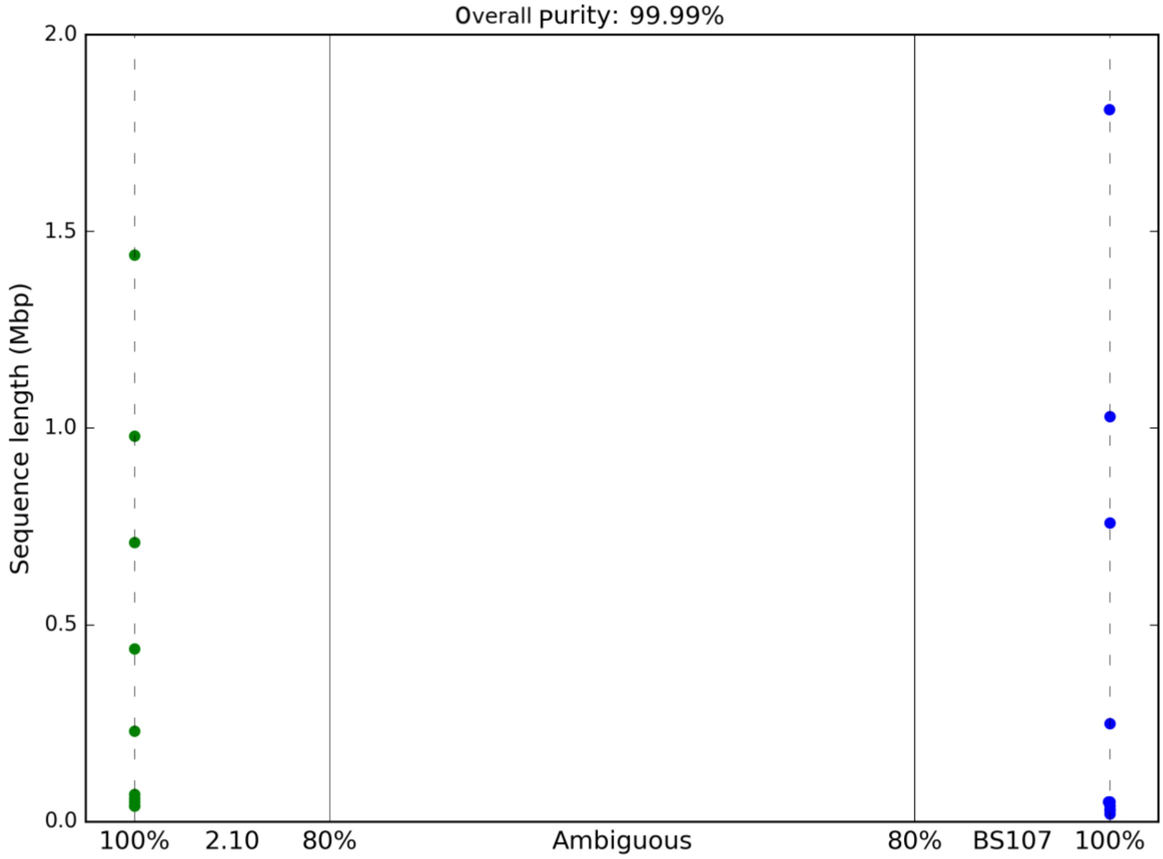


Figure 6: The purity of SAMPhaser produced haplotigs. Haplotigs with the number of mapped reads less than 100 were not shown in the plot.

The *P. inhibens* subreads were then mapped to phased haplotigs, Subreads were extracted from the produced SAM file and exported either to a 2.10 subset or a BS107 subset depending on the assignment of the haplotig they mapped to (Fig. 4 and Table 6). The separated subreads for the two strains were then separately assembled with Canu (Table 7). These assemblies showed excellent contiguity and completeness when compared to the reference genomes. In each case, four contigs had unambiguous assignment to a reference and overlapping ends (Fig. 7). These were circularised and polished using their corresponding subreads with Quiver (Fig. 4). The purity of the polished assemblies was assessed as described above and purities of > 99.98% were obtained (Table 8).

Table 6: Summary of the two subreads subsets

Type	2.10	BS107
Total reads (number)	59,314	54,558
Total reads (Mbp)	541.3	489.7
Coverage	136.4X	122.6X

Table 7: Summary of Canu produced assemblies

Type	2.10	BS107
Total length (bp)	4,244,495	4,283,012
Number of Contigs	6	10
Circularised contigs	4	4
N50 of Contigs (bp)	3,779,804	3,805,069
Average length (bp)	707,416	428,301
Shortest contig (bp)	3,346	1,647
Longest contig (bp)	3,779,804	3,805,069

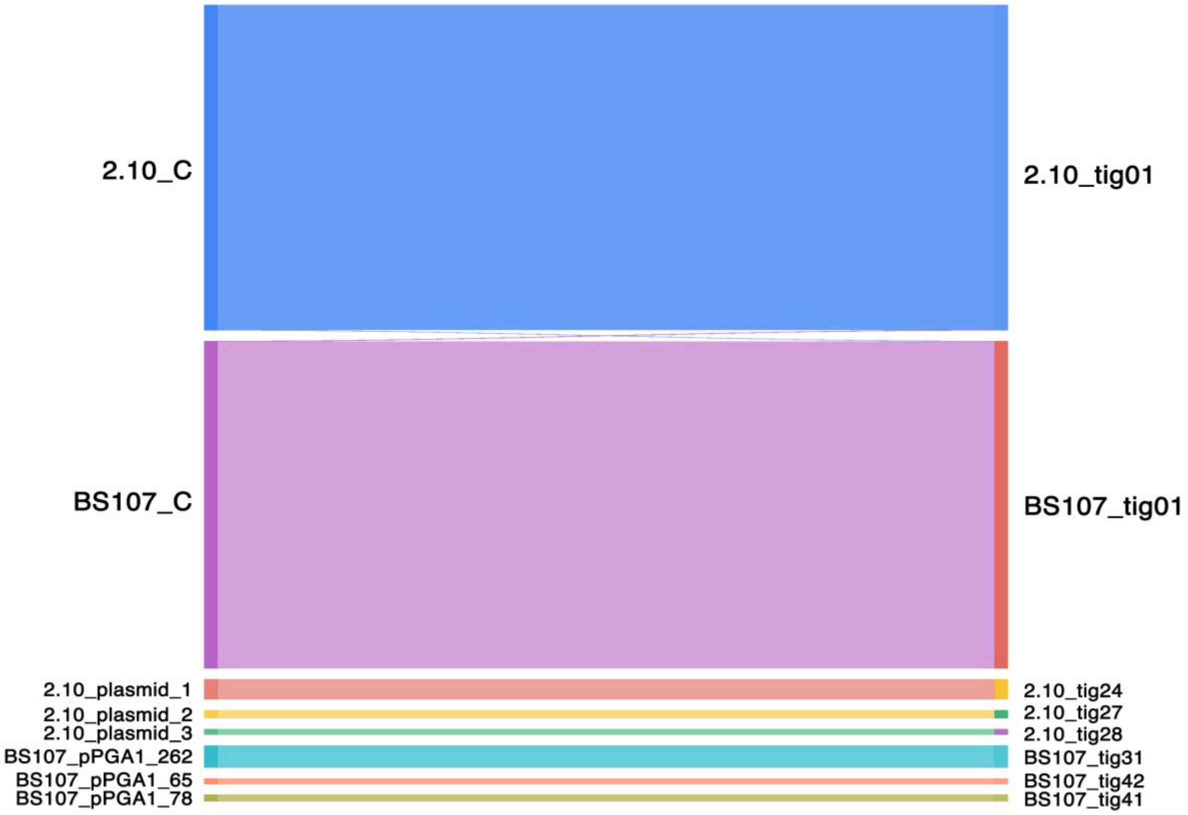


Figure 7: Mapping of combined *P. inhibens* assemblies (right) onto their references (left). BLASTN identity cut-off was set to 99% and alignment length cut-off was set to 5,000 bp to get rid of short fragment matches between the two high similar genomes. The thin cross-linking between 2.10 and BS107 chromosomes refers to a single BLASTN hit with an identity of 99.1% and alignment length of 6,009 bp. Pairwise BLASTN between their reference genomes with same cut-offs identified a single hit with identical identity and alignment length.

Table 8: Purity of the *P. inhibens* final assemblies

Strain	Assembly	Length (Mbp)	Purity (%)	Circularity
2.10	2.10_tig01	3.58	99.99	Yes
	2.10_tig24	0.23	99.98	Yes
	2.10_tig27	0.09	100	Yes
	2.10_tig28	0.07	100	Yes
BS107	BS107_tig01	3.61	99.99	Yes
	BS107_tig31	0.25	100	Yes
	BS107_tig41	0.07	100	Yes
	BS107_tig42	0.06	100	Yes

Genome annotation and gene prediction

Prokka v1.7 [23] was used to annotate the genomes (Table 9). The number of predicted coding sequences (CDS), tRNA and rRNA were given in Table 9. Annotation is available through GigaDB [ref to be added].

Table 9: Annotation summary

Genome	tRNA	rRNA	CDS
2.10	60	12	3,876
AD1	58	9	4,704
AD10	55	9	5,142
AD91A	53	9	4,317
AU392	36	6	3,042
BL5	57	9	5,073
BL110	72	16	3,811
BS107	62	12	3,902
D2	113	31	4,228
LSS9	60	12	3,763

Conclusion

Ten high-quality complete bacterial genomes were assembled from pooled PacBio sequencing. We show that genomes that are sufficiently divergent (i.e. ANI <~ 91%) can be assembled from pooled DNA into high-quality complete genomes using standard assembly algorithms (e.g. HGAP). For highly similar genomes (i.e. ANI > 97%), we found that standard workflow produces highly fragmented assemblies. We present a strategy using references and read phasing to produce final genome products of high quality and purity for these problem genomes. Overall, this information can be used in the future to design strategies to sequence pools of genomes using long-read sequencing.

Availability of supporting data

The raw reads are available at NCBI SRA database (accession: SRP158010). Final assemblies (genome sequences) have been submitted to GenBank (accession: CP031946-CP031967). Assemblies and annotation are available through GigaDB [ref to be added]. SAMPhaser is available as part of SLiMSuite [5]. Other in-house scripts used in this study are available at: https://github.com/songweizhi/metaPacBio.

List of abbreviations

ANI: Average nucleotide identity

BRIG: BLAST Ring Image Generator

CDS: Coding Sequence

NCBI: National Center for Biotechnology Information

PacBio: Pacific Biosciences

rRNA: Ribosomal ribonucleic acid

SNP: Single Nucleotide Polymorphism

tRNA: Transfer ribonucleic acid

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by Australian Research Council. Weizhi Song was funded by the Tuition Fee Scholarship provided by UNSW and the China Scholarship Council (201508200019).

Author’s contributions

TT and RE designed the project. WS extracted genomics DNA. WS and RE performed data analysis. RE designed and implemented the SAMPhaser algorithm. WS designed and implemented the algorithm for purity assessment. WS, RE and TT wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The Authors want to thank Suhelen Egan for providing some of the sequenced bacterial strains, as well as Marwan Majzoub, Qingyu Wang and Ana I. S. Esteves for helping with genomic DNA extraction.

References

1. Marzinelli EM, Campbell AH, Zozaya Valdes E, Vergés A, Nielsen S, Wernberg T, et al. Continental‐scale variation in seaweed host‐associated bacterial communities is a function of host condition, not geography. Environmental microbiology. 2015;17 10:4078-88.
2. Roth‐Schulze AJ, Pintado J, Zozaya‐Valdés E, Cremades J, Ruiz P, Kjelleberg S, et al. Functional biogeography and host specificity of bacterial communities associated with the Marine Green Alga Ulva spp. Molecular ecology. 2018;27 8:1952-65.
3. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PloS one. 2012;7 11:e47768.
4. Rhoads A and Au KF. PacBio sequencing and its applications. Genomics, proteomics & bioinformatics. 2015;13 5:278-89.
5. Edwards RJ. SLiMSuite v1.4.0. 2018; doi:http://doi.org/10.5281/zenodo.1302990.
6. Kumar V, Zozaya‐Valdes E, Kjelleberg S, Thomas T and Egan S. Multiple opportunistic pathogens can cause a bleaching disease in the red seaweed Delisea pulchra. Environmental microbiology. 2016;18 11:3962-75.
7. Fernandes N, Case RJ, Longford SR, Seyedsayamdost MR, Steinberg PD, Kjelleberg S, et al. Genomes and virulence factors of novel bacterial pathogens causing bleaching disease in the marine red alga Delisea pulchra. PloS one. 2011;6 12:e27387.

8. HOLMSTRÖM C, JAMES S, NEILAN BA, WHITE DC and KJELLEBERG S. Pseudoalteromonas tunicata sp. nov., a bacterium that produces antifouling agents. International Journal of Systematic and Evolutionary Microbiology. 1998;48 4:1205-12.

9. Rao D, Webb JS and Kjelleberg S. Competitive interactions in mixed-species biofilms containing the marine bacterium Pseudoalteromonas tunicata. Applied and environmental microbiology. 2005;71 4:1729-36.

10. Thole S, Kalhoefer D, Voget S, Berger M, Engelhardt T, Liesegang H, et al. Phaeobacter gallaeciensis genomes from globally opposite locations reveal high similarity of adaptation to surface life. The ISME journal. 2012;6 12:2229.

11. Esteves AI, Amer N, Nguyen M and Thomas T. Sample processing impacts the viability and cultivability of the sponge microbiome. Frontiers in microbiology. 2016;7:499.

12. Lee I, Kim YO, Park S-C and Chun J. OrthoANI: an improved algorithm and software for calculating average nucleotide identity. International journal of systematic and evolutionary microbiology. 2016;66 2:1100-3.

13. Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nature methods. 2013;10 6:563.

14. Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990;215 3:403-10.

15. Seemann T: Barrnap. https://github.com/tseemann/barrnap.

16. Chaisson MJ and Tesler G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC bioinformatics. 2012;13 1:238.

17. Song W: metaPacBio. https://github.com/songweizhi/metaPacBio (2018).

18. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH and Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome research. 2017;27 5:722-36.

19. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS one. 2014;9 11:e112963.

20. Alikhan N-F, Petty NK, Zakour NLB and Beatson SA. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC genomics. 2011;12 1:1.

21. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25 16:2078-9.

22. Bushnell B. *BBMap: a fast, accurate, splice-aware aligner*. 2014. Ernest Orlando Lawrence Berkeley National Laboratory, Berkeley, CA (US).

23. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30 14:2068-9.