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

Weizhi Song\textsuperscript{a,b}, Torsten Thomas\textsuperscript{b,c,∗∗}, Richard J. Edwards\textsuperscript{a,*}

\textsuperscript{a} School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW, Australia
\textsuperscript{b} Centre for Marine Bio-Innovation, University of New South Wales, Sydney, Australia
\textsuperscript{c} School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, NSW, Australia

\textbf{A B S T R A C T}

**Background:** High-quality, completed genomes are important to understand the functions of marine bacteria. PacBio sequencing technology provides a powerful way to obtain high-quality completed genomes. However, individual library production is currently still costly, limiting the utility of the PacBio system for high-throughput genomics. Here we investigate how to generate high-quality genomes from pooled marine bacterial genomes.

**Results:** Pooled genomic DNA from 10 marine bacteria were subjected to a single library production and sequenced with eight SMRT cells on the PacBio RS II sequencing platform. In total, 7.35 Gbp of long-read data was generated, which is equivalent to an approximate 168× average coverage for the input genomes. Genome assembly showed that eight genomes with average nucleotide identities (ANI) lower than 91.4% can be assembled with high-quality and completion using standard assembly algorithms (e.g. HGAP or Canu). A reference-based reads phasing step was developed and incorporated to assemble the complete genomes of the remaining two marine bacteria that had an ANI > 97% and whose initial assemblies were highly fragmented.

**Conclusions:** Ten complete high-quality genomes of marine bacteria were generated. The findings and developments 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.

**1. Introduction**

Marine bacteria can play important roles in biogeochemical cycles of the ocean (Azam and Malfatti, 2007) and the development, defense and health of host organisms, such as seaweeds or sponges (Egan et al., 2013; Webster and Thomas, 2016). The availability of high-quality reference genomes of bacteria is critical for a better understanding of their functions and interactions with hosts. Short-read sequencing technologies, e.g. Illumina, often fail to generate completed genomes due to sequencing biases, repetitive genomic features or genomic polymorphism (English et al., 2012). Pacific Biosciences (PacBio) SMRT™ sequencing provides a powerful way to get high-quality complete genome or closing gaps of current draft genomes with long reads (Rihoods and Au, 2015). However, library preparation for individual genomes results in a relatively high cost for PacBio-based microbial sequencing projects. For example, current pricing for a Sequel 10-20 kb library preparation is ~AU$1800 and this library could be sequenced on a Sequel v3 SMRT cell (10 h movie) for ~AU$2600 (Ramaciotti Centre for Genomics, UNSW). While PacBio offers barcoding of individual samples, it would still result in a cost for a size-selected library production and sequencing of ~AU$1300 per genome, and if 10 genomes are multiplexed results in a total cost ~AU$13,000. This is substantial more than an approach where 10 bacterial genomes would be pooled and then sequenced as a single library (i.e. ~AU$4400, see numbers above). Furthermore, multiplexed PacBio sequencing involves a smaller library insert size, which can present challenges for the assembly of “Class II” and “Class III” genomes with high repeat content (Koren et al., 2013).

Here, we investigated this strategy for the PacBio system by sequencing pooled genomic DNA for ten marine bacterial strains with

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

* Corresponding author.
** Correspondence to: T. Thomas, School of Biological, Earth and Environmental Science, University of New South Wales, Sydney, Australia.

E-mail addresses: songwz03@gmail.com (W. Song), t.thomas@unsw.edu.au (T. Thomas), richard.edwards@unsw.edu.au (R.J. Edwards).

https://doi.org/10.1016/j.margen.2019.05.002
Received 28 March 2019; Received in revised form 15 May 2019; Accepted 15 May 2019
1874-7787/ © 2019 Published by Elsevier B.V.

Please cite this article as: Weizhi Song, Torsten Thomas and Richard J. Edwards, Marine Genomics, https://doi.org/10.1016/j.margen.2019.05.002
various degrees of genome similarity. We also introduce a reference-based read phasing strategy using SAMPhaser (Edwards, 2018) for the assembly of highly similar genomes (i.e. average nucleotide identities (ANI) > 97%). Using this sequencing strategy, we have produced ten complete high-quality genome assemblies using a single SMRTbell library.

2. Materials and method

2.1. Strain selection

Ten marine bacterial strains, isolated from various marine hosts, were selected for genome sequencing: *Aquimarina* sp. AD1, AD10 and BL5 as well as *Alteromonas* sp. BL110, *Phaeobacter* sp. LS9, and *Ruegeria* sp. AD91A were isolated from the red seaweed *Delisea pulchra* (Kumar et al., 2016; Fernandez et al., 2011); *Pseudoalteromonas tunica* D2 was isolated from the surface of the tunicate *Ciona intestinalis* (Holmstrom et al., 1998); *Phaeobacter inhibens* 2.10 was isolated from the surface of the green alga *Ulva australis* (Rao et al., 2005); *Phaeobacter inhibens* BS107 was isolated from the scallop *Pecten maximus* (Thole et al., 2012); *Flavobacteriaceae bacterium* AU392 was isolated from the sponge *Tedania anhelans* (Esteves et al., 2016). All strains had been propagated for multiple rounds of subculturing in the laboratory and were pure cultures. Previous draft or completed genomes produced with Illumina MiSeq, Roche FLX or Sanger sequencing technologies were available for nine of the selected strains (Table 1). As no reference genomes were available for nine of the selected strains (Table 1).

Table 1

<table>
<thead>
<tr>
<th>Sequenced strain</th>
<th>Abbrev.</th>
<th>Reference</th>
<th>Reference contig number</th>
<th>Status</th>
<th>Approximate genome size (Mbp)</th>
<th>GC (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Alteromonas</em> sp. BL110</td>
<td>BL110</td>
<td>BL110</td>
<td>10</td>
<td>Draft</td>
<td>4.2</td>
<td>44.1</td>
</tr>
<tr>
<td><em>Aquimarina</em> sp. AD1</td>
<td>AD1</td>
<td>AD1</td>
<td>519</td>
<td>Draft</td>
<td>5.1</td>
<td>32.1</td>
</tr>
<tr>
<td><em>Aquimarina</em> sp. AD10</td>
<td>AD10</td>
<td>AD10</td>
<td>88</td>
<td>Draft</td>
<td>3.2</td>
<td>32.4</td>
</tr>
<tr>
<td><em>Flavobacteriaceae bacterium</em> AU392</td>
<td>AU392</td>
<td>AU392</td>
<td>12</td>
<td>Draft</td>
<td>6.2</td>
<td>30.7</td>
</tr>
<tr>
<td><em>Aquimarina</em> sp. BL5</td>
<td>BL5</td>
<td>BL5</td>
<td>353</td>
<td>Draft</td>
<td>5.6</td>
<td>32.9</td>
</tr>
<tr>
<td><em>Pseudoalteromonas Tunicata</em> D2</td>
<td>D2</td>
<td>D2 (Thomas et al., 2008)</td>
<td>42</td>
<td>Draft</td>
<td>4.8</td>
<td>39.9</td>
</tr>
<tr>
<td><em>Phaeobacter</em> sp. LS9</td>
<td>LS9</td>
<td>LS9</td>
<td>50</td>
<td>Draft</td>
<td>3.9</td>
<td>60.3</td>
</tr>
<tr>
<td><em>Phaeobacter inhibens</em> 2.10</td>
<td>2.10</td>
<td>2.10 (Thole et al., 2012)</td>
<td>4 (3 plasmids)</td>
<td>Complete</td>
<td>4.0</td>
<td>59.8</td>
</tr>
<tr>
<td><em>Phaeobacter inhibens</em> BS107</td>
<td>BS107</td>
<td>BS107 (Thole et al., 2012)</td>
<td>4 (3 plasmids)</td>
<td>Complete</td>
<td>4.0</td>
<td>59.8</td>
</tr>
<tr>
<td><em>Ruegeria</em> sp. AD91A</td>
<td>AD91A</td>
<td>6PALISEP08</td>
<td>42</td>
<td>Draft</td>
<td>4.3</td>
<td>57.0</td>
</tr>
</tbody>
</table>

* Unpublished, see “Availability of supporting data” for database accession numbers.

As assemblies from the two *Phaeobacter inhibens* genomes (2.10 and BS107) were highly fragmented (see below), a reference-guided partitioning and re-assembly strategy was adopted. In detail, subreads that were not mapped to the eight completed genome assemblies were extracted as the *P. inhibens* subreads subset (Fig. 1). These subreads were then 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 mapped 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 (Edwards, 2018). SAMPhaser first identifies variants from a pileup file generated from the BLASR BAM output using SAMTools v1.7 (Li et al., 2009). 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 removed as putative structural variants, and the haplotig split at this point. Haplotigs with an average depth of coverage below 5× 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.

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) (Song, n.d.) as follows:

- Simulate short reads from the two reference genomes, with the number of simulated reads being in proportion to the sizes of the reference genomes.
- Map the simulated reads to haplotigs with BBMAP v35.82 (Bushnell, 2014). A read will not be mapped if multiple top-scoring mapping locations were found from the query sequences (specified with “ambiguous = toss”).
- 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:

\[
\text{Overall Purity} = \sum_{i \in A} \frac{L_i P_i}{\sum_{i \in R} L_i}
\]

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 \(P.\) inhibens subreads were then mapped to the 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. The separated subreads for the two strains were then separately de novo assembled with Canu. Canu assemblies from the two genomes were then circularised and polished using their corresponding subreads with Quiver and the purity of the polished assemblies was assessed as described above.
3. Results and discussions

3.1. De novo assembly of pooled genomes

A summary of the reference genome of the 10 selected strains from previous studies is given in Table 1. Their GC content and pairwise ANI were ranged from 30.7% to 60.3% (Table 1) and 62.48% to 97.27% (Fig. 2), respectively.

Sequencing of the mixed gDNA from the 10 bacterial strains on eight SMRT cells generated 7.35 Gbp raw data (Table S1), which is equivalent to an approximate 168× average coverage of the input genomes.

De novo assembly of generated subreads using HGAP v3 produced 105 contigs (Table S2). Mapping of the HGAP contigs to the 10 reference genomes revealed that eight of the genomes produced assemblies in 1–5 pieces, with two Phaeobacter inhibens genomes being highly fragmented (Fig. 3). Further analysis showed that some short contigs with reference assignment (hcq1, hcq9, hcq32 and hcq150) were...
actually sequences covering the break-point of their corresponding circular chromosome sequences. These sequences were marked as redundant sequences (Table 2) and excluded from further analysis.

The break-point of contig hcq3 (from the genome AD1) was located in a highly repetitive region, which made its circularisation difficult, and so this strain was re-assembled as described above. Re-assembly of the AD1 subreads produced a single circular contig with a length of 5,483,011 bp. This contig was then manually circularised (Table 2, AD1_tig1). AD91A (hcq7, hcq43, hcq44) failed to generate a full-length chromosomal contig with overlapping ends. AD91A was therefore also re-assembled, and two circular contigs (3,685,098 and 766,037 bp, respectively) assigned to the reference genome 6PALISEP08 were identified and manually circularised (Table 2, AD91A_tig1 and AD91A_tig30). A summary of the number of Pilon corrected indels for...
the circularised assemblies were given in Table S3.

3.1.1. Reference-guided data partitioning and de novo assembly of two highly similar Phaeobacter inhibens genomes

Assemblies from two of the three Phaeobacter inhibens genomes (2.10 and BS107) were highly fragmented (Fig. 3), presumably due to high sequence similarity between them (Figs. 2 and 4). A reference-guided re-assembly strategy was therefore adopted as described in Fig. 1.

Phasing results of combined P. inhibens subreads mapped onto combined reference genome of 2.10 and BS107 are given in Fig. 5. Purity assessment of the remaining SAMPhaser haplotigs as described in the Methods revealed that their overall purity was 99.99% (Fig. 6). We anticipate that alternative phasing tools, such as WhatsHap (Martin et al., 2016) could be used with similar success, but it is beyond the scope of this study to perform such a comparison. The separated subreads for the two strains were then separately assembled with Canu (Table 3). These assemblies showed excellent contiguity and completeness when compared to the reference genomes. In each case, four contigs (covering the chromosome and three plasmids) had unambiguous assignment to a reference (Fig. 7). The purity of the polished assemblies was assessed as described above and estimated to be > 99.98% (Table 4).

3.2. Genome annotation

Genomes were submitted to Genbank and summary of genome annotation (e.g. coding sequences, tRNA and rRNA) of the 10 marine bacterial genomes from GenBank is given in Table 5.

4. Conclusion

Ten high-quality complete bacterial genomes were assembled from pooled PacBio sequencing. We show that genomes that are sufficiently divergent (i.e. ANI < ~ 92%) 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 a standard workflow produces highly fragmented assemblies. We present a strategy using references and read phasing to separate subreads for assembly without the need for barcoding, and produce final genome products of high quality and purity for such genomes. Overall, this information can be used in the future to design strategies to sequence pools of genomes using long-read sequencing. The information gained here might however require further refinement as read length and error rates are improving (e.g. with the newer versions of the PacBio Sequel platform). And finally, our sequencing of pooled
genomes is analogous to metagenome sequencing of a mixture of bac-
teria. Our results indicate that genomes with ANI < 92% should readily
assemble from metagenomic shotgun sequencing data generated with
the PacBio platform, if suffi-
tient coverage is obtained. For highly si-
milar genomes there will be likely limited success in assembly unless
suitable reference genomes are available.

Availability of supporting data

The raw reads are available at NCBI SRA database (accession: SRP158010). The final complete genomes have been submitted to GenBank (accession: CP031946-CP031967, see Table 5). The unpublished draft genomes and their corresponding Illumina short-reads are available at GenBank (accession: RAZZ00000000, RAZY00000000, RAX00000000, RAZV00000000 and RAZW00000000) and SRA
database (accession: PRJNA493226). SAMPhaser is freely available under a GNU GPL v3 license as part of SLiMSuite (Edwards, 2018) (https://github.com/slimsuite/SLiMSuite) and as a reduced set of code and documentation at https://github.com/slimsuite/SAMPhaser. Other in-house scripts and commands used in this study are available at: https://github.com/songweizhi/metaPacBio.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Declarations of interest

None.

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.
Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.margen.2019.05.002.

References

Webster, N.S., Thomas, T., 2016. The sponge hologenome. MBio. 7 (2), e00135–e00216.