for standard library prep. This is not the main advantage, though it does now offer a route from sample to sequence within an average working day. This protocol may be of benefit to the direct sequencing of plasmids, single-standed or double-stranded viruses, mitochondrial DNA, and microbial pathogens in a clinical setting.

Materials and methods

M13mp18 viral DNA (both single-stranded and double-stranded; catalog no. N4040S and N4018S, respectively) and M13 forward (5′-GTTTTCCCAGTCACGAC-3′) and reverse sequencing primers (5′-AACAGCTATGACCATG-3′) were from New England Biolabs (Hitchin, UK). Methicillin-resistant *Staphylococcus aureus* (MRSA) plasmids were purified from a solution prep of *S. aureus* TW20 using a Qiagen (Crawley, UK) Plasmid Midi Kit with Qiagen Genomic-tip 100/G following the manufacturer's "very low-copy plasmid/cosmid purification protocol" from a 500 ml culture. Plasmid Safe DNase (Epicentre Biotechnologies, Madison, WI, USA) was used to reduce the amount of linear single- and double-stranded molecules from the TW20 plasmid prep. Random hexamer primers from Roche (Welwyn Garden City, UK) were used, as provided in the Transcriptor First Strand cDNA Synthesis Kit. pET28a plasmid vectors encoding EcoDamI methyltransferase (Dam constructs) expressed in dam-/dcm *Escherichia coli* cells, were prepared in-house. Components from the DNA/Polymerase Binding Kit 2.0 from Pacific Biosciences were used during the annealing and binding reactions. The Annealing and Binding Calculator (version 1.3.1) provided by Pacific Biosciences was used to calculate the concentration of bound complex to be loaded onto the sample plate for the instrument. An MJ PTC-225 thermocycler from MJ Research (Watertown, MA, USA) was used for the annealing and binding reactions. The PacBio DNA Sequencing Kit 2.0 (8Rxn) and SMRT Cell 8Pac v2 (8 Cells) were used for sequencing. Sequence analysis was performed with SMRT portal, SMRT pipe, and SMRT View, version 1.3.1, and Motif Finder, version 0807, all from Pacific Biosciences.

Annealing reaction

Standard library preparation was omitted; the DNA templates were used directly in the annealing reaction. For each experiment, a quantity of DNA between 1 ng and 100 ng was annealed with suitable primers. With ssDNA, the annealing reaction used the standard PacBio protocol; i.e., 2 min at 80°C followed by cooling at 0.1°C/s to 25°C. With dsDNA, a different annealing protocol was used; the reaction was heated to 95°C for 5 min, then immediately snap-cooled on wet ice. As an example, when using ds M13mp18 DNA, 2.2 μL of DNA at 46 ng/μl (~100 ng), 0.9 μL PacBio Primer Buffer (10×), and both 0.9 μL forward primer (10 μM) and 0.9 μL reverse primer (10 μM) were mixed in a final annealing reaction volume of 9 μL. The final concentration of DNA template was therefore ~2.5 nM, with 1000 nM primer (~400×). In order to use the PacBio Annealing and Binding calculator, we assumed that denatured M13mp18 DNA is comparable to a SMRT bell, with half the original double-stranded M13mp18 molecule's nominal length; i.e., one double-stranded 7.2-kb molecule, when denatured, becomes two 3.6 kb SMRT bells. A 2-fold dilution series of DNA was used to create additional annealing reactions in the range of 0.8-100 ng of DNA. There was a massive excess of forward and reverse primers at the lower concentrations of DNA in these reactions.

Binding reaction, loading, and sequencing

In the binding reaction, the ratio of polymerase to template DNA used was 3:1. First, 1.5 μL of polymerase (1600 nM) was combined with 25 μL of binding buffer giving a 90 nM polymerase solution. Four μl of a 1:1:1 DTT:dNTP:binding buffer mix (each from the PacBio Binding Kit) was added to the annealed template DNA and 1.5 μL of 90 nM

polymerase was added. This was mixed gently by pipetting and then incubated at 30°C for 4 h.

The bound complex was loaded at 1 nM onto the instrument. Typically this is achieved by diluting the bound complexes with a mixture of 1:10 DTT:Complex Dilution Buffer. In this experiment, however, it was only possible to achieve a 1 nM loading concentration for the samples containing 100 ng and 50 ng input DNA. For the other samples in the 2-fold dilution series, the calculated concentration was <1 nM before dilution. The total volume of 14.6 μL of binding reaction was therefore loaded directly into the sample plate wells for each of these dilute samples.

Two × 45 min sequencing movies were acquired for each sample in this study. Mapping, de novo assembly, and modification analysis, were carried out with PacBio's SMRT Analysis pipeline run via the SMRT Portal interface. PacBio's Motif Finder was used in the final step of analysis for the pet28a plasmid vector to characterize the sequence specific motif at which base modifications were observed.

Results and discussion

At the outset of this study, an experiment using single-stranded M13mp18 viral DNA and the M13 forward sequencing primer (5′-GTTTTCCCAGTCACGAC-3′) showed that it was possible to generate sequence data directly from circular DNA molecules without library preparation; i.e., fragmentation, end repair, and adapter ligation. From 25 ng of ssDNA and 100-fold molar excess of primer, it was possible to map the data generated against the 7.2 kb

the polymerization speed and longevity does account for some of the observed distribution. Additionally, the SMRT pipe software might have difficulty in mapping some reads, especially those that extend beyond the end of the linear FASTA reference. The DNA molecules sequenced were circular, but the reference used is a single linear sequence. Therefore, a number of the reads generated in these runs will, in fact, extend beyond the artificially imposed boundaries of the reference file. Some of the longest reads will also span the entire circular genome, further complicating the automatic analysis. The SMRT analysis software is not designed to deal with reads of this nature; although the initial filtering of the data are unaffected, as it's based on read quality metrics only. None of the reads have PacBio adapters that signal the end of a DNA template fragment so the standard re-sequencing (mapping) protocol in SMRT portal possibly contributes to the uneven coverage profiles generated (mapping thresholds were a maximum of one hit per read, 30% maximum divergence, and minimum anchor size of 12). Some reads were longer than the entire genome as evidenced by the maximum read lengths in the SMRT Portal raw read-length histogram (i.e., any reads >7.2 kb). These very long reads could be observed using PacBio's SMRT view software by concatenating two M13mp18 references in tandem into a single FASTA file (Figure 2).

The sequencing metrics for a 2-fold dilution series of M13mp18 input genomic DNA

provided by PacBio, for mining polymerization kinetics for motifs associated with base modifications. In this vector, 50 instances of GATC methylated at the A position were identified; there are 25 GATC sites in the sequence and wild type EcoDam was expected to methylate each one of them.

Direct sequencing was then tested using a DNA extract of *Staphylococcus aureus* TW20, a MRSA strain and well-known nosocomial infection (13). The plasmids of this bacterial sample were of interest as an example of the application of the PacBio RS to infectious disease identification through sequencing. Antibiotic resistance genes are often carried on plasmids (14,15) and can spread very quickly in heterogeneous bacterial communities (16–19). DNA was extracted from a solution culture of TW20 and digested with Plasmid Safe DNase to reduce the amount of linear fragments and effectively increase the concentration of plasmids in the sample. An electropherogram of the final sample showed that the DNA preparation also contained a smear of linear double stranded fragments ranging from 100 bp to >25 kbp, with a peak at approximately 20 kb (Supplementary Figure 1). The two plasmids in TW20 are double-stranded and circular, with lengths of 3 kb and 30 kb. We generated sequence data using random hexamer primers in the annealing reaction. Four reactions containing 50 ng of the *S. aureus* DNA preparation with various amounts of hexamer primers, from 10-fold to 600-fold, i.e., 500 ng, 1 μg, 10 μg, and 30 μg per annealing reaction, were performed in 9 μL reaction volumes. A single SMRT cell was sequenced for each reaction and the trend observed across these four reactions showed fewer mapped reads as the amount of random hexamer primers increased. This is perhaps because of the proximity of annealed primers on the DNA strand at higher concentrations, leading to polymerases colliding with one another, or simply the reduction of signal to noise as two fluorescent signals could be observed concurrently. The annealing reaction with 10-fold primers generated 3240 mapped reads, 20-fold generated 3085 mapped reads, 200-fold generated 2911 mapped reads, and 600-fold generated 2011 mapped reads, all with a mean mapped read length of approximately 500 bp. There was also a difference in coverage depth between the two plasmids; the mean coverage for the 3-kb plasmid was 35×, but only 5× coverage was obtained for the 30-kb plasmid, which is due mostly to the difference in plasmid length. There is a loading inefficiency of larger molecules because of their lower diffusion coefficient, as well as the disparity between the molecule's hydrodynamic radius and the very small zero-mode waveguide (ZMW). Future upgrades to the loading mechanism on the PacBio instrument (MagBead loading) which should eliminate this problem are very close to release. The combined sequence data from these four SMRT cells produced 13,724 reads; 479 reads mapped to the plasmids and 11,247 to the genome (5.3 Mb mapped providing a mean 1.6× coverage), an overall mapping rate of 85% which is not dissimilar to standard mapping rates of SMRT bell libraries we have made (from a recent single SMRT cell of *S. aureus* TW20 1 kb SMRT bell library 39,478 reads were mapped from 47,465 filtered reads, a mapping rate of 83%).

Finally, the technique was used to sequence linear molecules of *Candidatus* Phytoplasma mali, a plant-pathogenic mycoplasma with a small genome of ~600 kb that is 21.4% GC and characterized by large terminal inverted repeats and covalently closed hairpin ends (20). The DNA was sheared to approximately 3-kb fragments and a 25-ng aliquot was sequenced using random hexamers in a similar manner to that described previously. From a single SMRT cell with 2 × 45 min movies, only 870 post-filter reads were generated of which 63 reads mapped, with a mean consensus accuracy of 84.4%. The mean mapped read-length was 817 bp and the coverage only 0.08%. The poor mapping rate is most likely due to a greater percentage of low-quality reads from this particular sample. Although the yield is poor, direct sequencing of these linear DNA molecules shows some promise too. A blastn (21) search using the NCBI server against the refseq_genomic database called out *Candidatus* Phytoplasma mali as the most likely taxonomic hit (Supplementary Table 1).

This suggests it is possible to obtain enough information from very few mapped reads to begin to identify the genomes present in a sample. However, comparing the difference in data yield between the *S. aureus* and *Ca.* Phytoplasma mali, it is clear that further optimization of the method is required to improve the number of reads that can be mapped when sequencing linear molecules from a variety of genomic samples.

The method described here utilizes the PacBio RS platform for direct sequencing, enabling the generation of sequence data from small single- and double-stranded DNA genomes. Potentially this technique also could be applied to circularized molecules, e.g., amplicons or sheared fragments that have been circularized. However, the additional circularization step and clean up would mean relatively minor time and DNA savings compared with current PacBio protocols. The direct sequencing technique could allow the identification of plasmids present in a bacterial sample in an extremely straightforward and fast manner. Although there is an indication that different genomes may be more or less accessible with this method, we have demonstrated its application to sequencing ssDNA and dsDNA viruses, plasmid vector models for methylation studies, antibiotic resistance gene-carrying plasmids, and the entire genome of a clinically relevant microbial pathogen. All of these were performed without the need for library preparation, and it is possible to generate sequence data within 8 h from <1 ng of DNA without a PCR amplification step. The fact that our method can be performed without a priori knowledge of any sequence and with no organism-specific reagents, coupled with its simplicity and speed, makes it particularly well suited for use in acute disease and infectious outbreak scenarios.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank NEB for providing the M13 primers which were not currently available for purchase. Sascha Sauer of the Max-Planck-Institute for Molecular Genetics for isolating and providing the Candidatus Phytoplasma mali sample. Theresa Feltwell of The Wellcome Trust Sanger Institute for culturing S. aureus TW20 and performing the plasmid prep. Albert Jeltsch and Tomasz Jurkowski for providing the original Dam constructs. This work was supported by the Wellcome Trust grant 098051 (PC, MQ, HS) and The Cambridge Cancer Center (TC, WR). ESGI – The research leading to these results has received funding from the Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 262055.

References

1. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, et al. Real-Time DNA Sequencing from Single Polymerase Molecules. Science. 2009; 323:133–138. [PubMed: 19023044]
2. Korlach J, Bibillo A, Wegener J, Peluso P, Pham TT, Park I, Clark S, Otto GA, Turner SW. Long, processive enzymatic DNA synthesis using 100% dye-labeled terminal phosphate-linked nucleotides. Nucleosides Nucleotides Nucleic Acids. 2008; 27:1072–1083. [PubMed: 18711669]
3. Korlach J, Bjornson KP, Chaudhuri BP, Cicero RL, Flusberg BA, Gray JJ, Holden D, Saxena R, et al. Real-time DNA sequencing from single polymerase molecules. Methods Enzymol. 2010; 472:431–455. [PubMed: 20580975]
4. Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW. Zero-mode waveguides for single-molecule analysis at high concentrations. Science. 2003; 299:682–686. [PubMed: 12560545]
5. McCarthy A. Third generation DNA sequencing: pacific biosciences' single molecule real time technology. Chem. Biol. 2010; 17:675–676. [PubMed: 20659677]
6. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. Hum. Mol. Genet. 2010; 19:R227–R240. [PubMed: 20858600]

7. Carneiro MO, Russ C, Ross MG, Gabriel SB, Nusbaum C, Depristo MA. Pacific biosciences sequencing technology for genotyping and variation discovery in human data. BMC Genomics. 2012; 13:375. [PubMed: 22863213]
8. Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N, Scheutz F, Paxinos EE, Sebra R, et al. Origins of the E. coli strain causing an outbreak of hemolytic-uremic syndrome in Germany. N. Engl. J. Med. 2011; 365:709–717. [PubMed: 21793740]
9. Travers KJ, Chin CS, Rank DR, Eid JS, Turner SW. A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic Acids Res. 2010; 38:e159. [PubMed: 20571086]
10. Clark TA, Murray IA, Morgan RD, Kislyuk AO, Spittle KE, Boitano M, Fomenkov A, Roberts RJ, Korlach J. Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. Nucleic Acids Res. 2012; 40:e29. [PubMed: 22156058]
11. Song CX, Clark TA, Lu XY, Kislyuk A, Dai Q, Turner SW, He C, Korlach J. Sensitive and specific single-molecule sequencing of 5-hydroxymethylcytosine. Nat. Methods. 2012; 9:75–77. [PubMed: 22101853]
12. Murray IA, Clark TA, Morgan RD, Boitano M, Anton BP, Luong K, Fomenkov A, Turner SW, et al. The methylomes of six bacteria. Nucleic Acids Res. 2012
13. Holden MT, Lindsay JA, Corton C, Quail MA, Cockfield JD, Pathak S, Batra R, Parkhill J, et al. Genome sequence of a recently emerged, highly transmissible, multi-antibiotic- and antiseptic-resistant variant of methicillin-resistant Staphylococcus aureus, sequence type 239 (TW). J. Bacteriol. 2010; 192:888–892. [PubMed: 19948800]
14. Robicsek A, Jacoby GA, Hooper DC. The worldwide emergence of plasmid-mediated quinolone resistance. Lancet Infect. Dis. 2006; 6:629–640. [PubMed: 17008172]
15. Svara F, Rankin DJ. The evolution of plasmid-carried antibiotic resistance. BMC Evol. Biol. 2011; 11:130. [PubMed: 21595903]
16. Haenni M, Saras E, Metayer V, Doublet B, Cloeckaert A, Madec JY. Spread of the blaTEM-52 gene is mainly ensured by IncI1/ST36 plasmids in Escherichia coli isolated from cattle in France. J Antimicrob Chemother. 2012
17. Miró E, Segura C, Navarro F, Sorli L, Coll P, Horcajada JP, Alvarez-Lerma F, Salvadó M. Spread of plasmids containing the bla(VIM-1) and bla(CTX-M) genes and the qnr determinant in Enterobacter cloacae, Klebsiella pneumoniae and Klebsiella oxytoca isolates. J. Antimicrob. Chemother. 2010; 65:661–665. [PubMed: 20089541]
18. Valverde A, Canton R, Garcillan-Barcia MP, Novais A, Galan JC, Alvarado A, de la Cruz F, Baquero F, Coque TM. Spread of bla(CTX-M-14) is driven mainly by IncK plasmids disseminated among Escherichia coli phylogroups A, B1, and D in Spain. Antimicrob. Agents Chemother. 2009; 53:5204–5212. [PubMed: 19786598]
19. Dionisio F, Matic I, Radman M, Rodrigues OR, Taddei F. Plasmids spread very fast in heterogeneous bacterial communities. Genetics. 2002; 162:1525–1532. [PubMed: 12524329]
20. Kube M, Schneider B, Kuhl H, Dandekar T, Heitmann K, Migdoll AM, Reinhardt R, Seemuller E. The linear chromosome of the plant-pathogenic mycoplasma ‘Candidatus Phytoplasma mali’. BMC Genomics. 2008; 9:306. [PubMed: 18582369]
21. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25:3389–3402. [PubMed: 9254694]

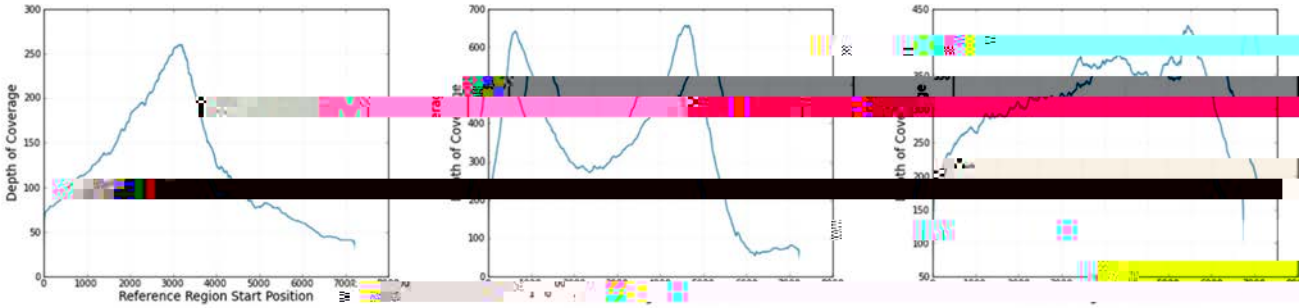

Figure 1. The coverage plot of M13mp18 viral genome after sequencing wih a single SMRT cell
(Left) M13mp18 ssDNA sequenced using the M13 forward sequencing primer. (Middle) M13mp18 dsDNA sequenced using the M13 forward and reverse sequencing primers. (Right) Coverage of M13mp18 dsDNA sequenced using random hexamer primers.

Europe PMC Funders Author Manuscripts

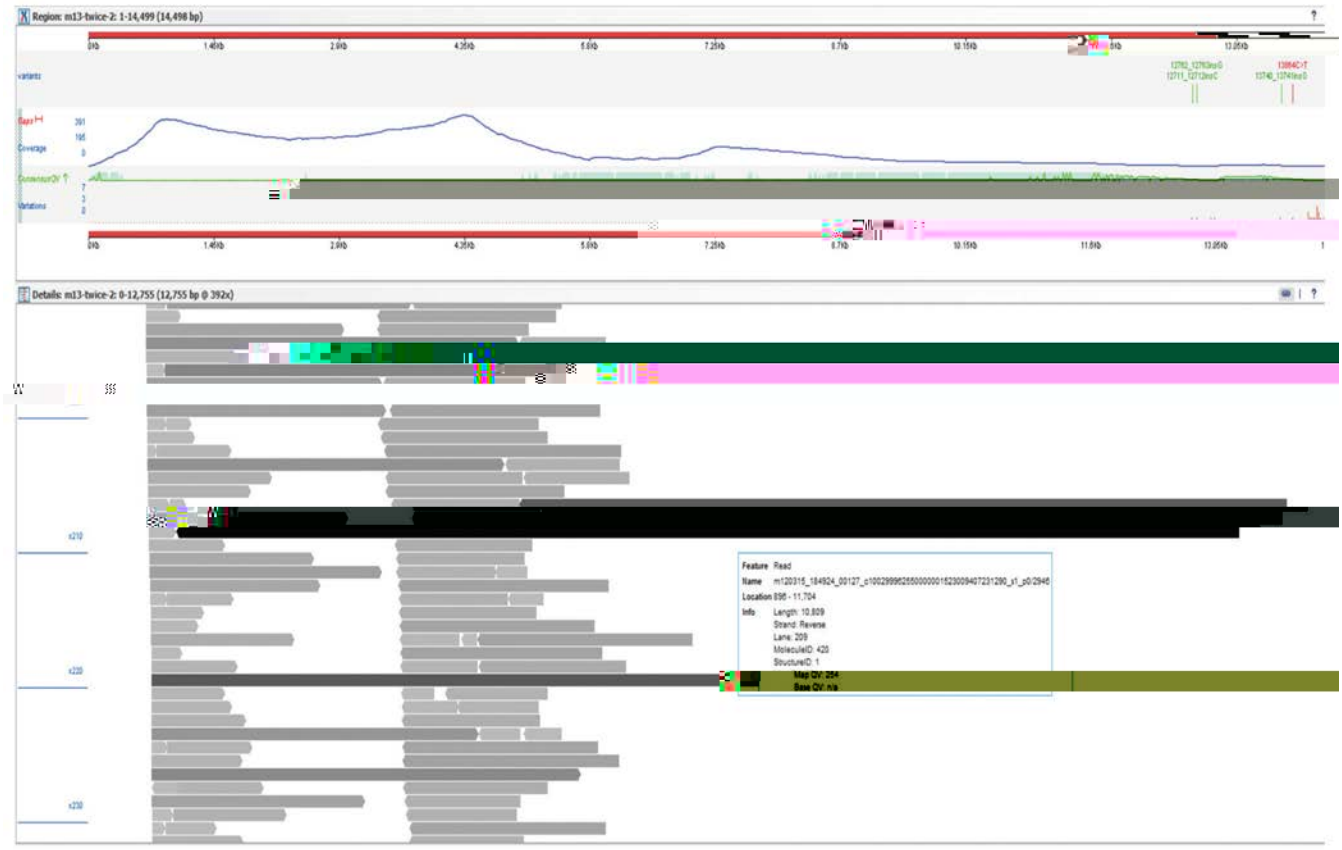

Figure 2. SMRT View genome browser, showing sequence data mapped against M13mp18
A single >10 kb read, longer than the entire M13mp18 circular genome, is highlighted.

Europe PMC Funders Author Manuscripts

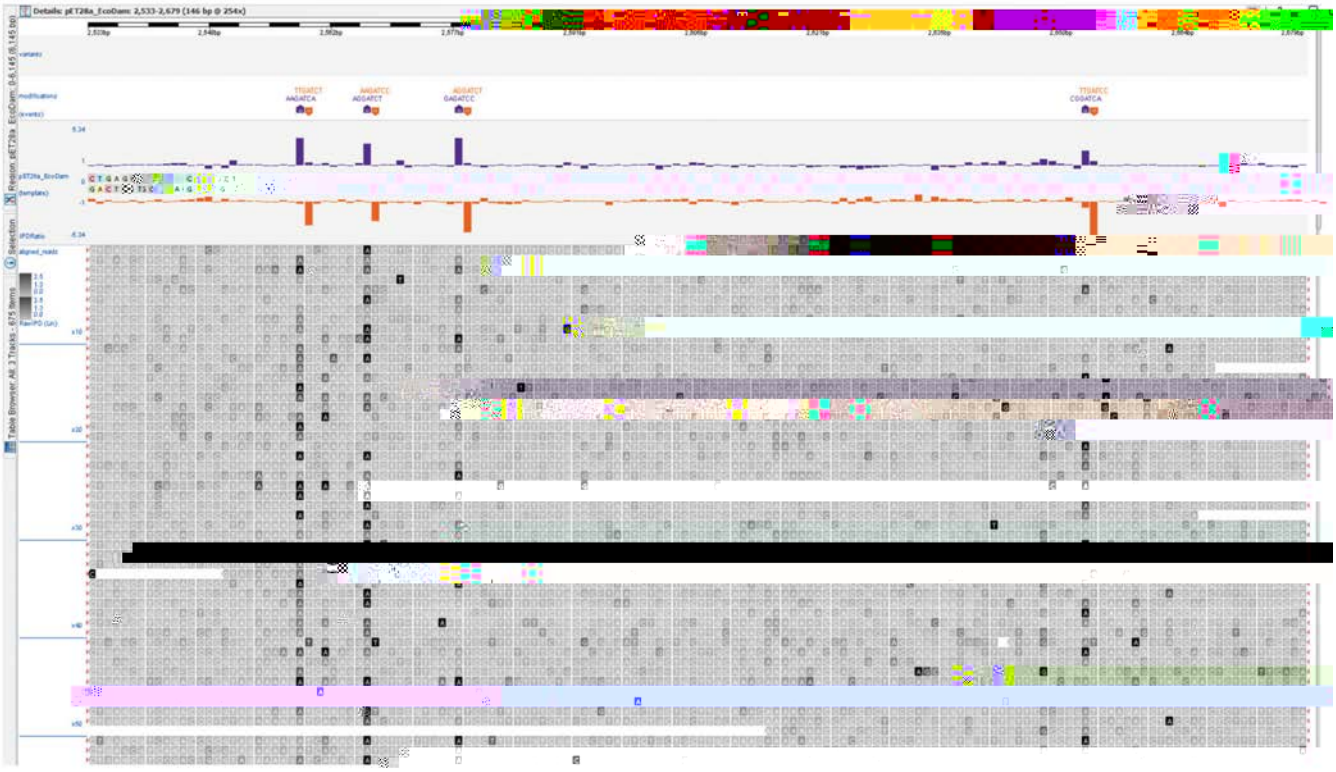

Figure 3. SMRT View genome browser, showing kinetic information elucidating modified bases

Four instances of m6A methylation in GATC motifs as detected with the PacBio base-modification analysis software (Motif Finder).

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Table 1

Sequencing metrics for a dilution series of genomic input **DNA**.

Input DNA (ng)	# of Filtered Reads	Mean Mapped Read Length	% Bases Called	Coverage Depth (mean)	Consensus Accuracy (%)
100	1917	1559	100	329	100
50	2514	1484	100	412	100
25	2716	1378	100	384	100
12.5	2395	1294	100	297	100
6.3	1168	1058	100	103	100
3.1	613	1119	100	56	100
1.6	224	1046	99.9	18	99.8
0.8	74	674	91.4	3	93.4