Benchmarks

Improving genome assemblies by sequencing PCR products with PacBio

Xiaojing Zhang, Karen W. Davenport, Wei Gu, Hajnalka E. Daligault, A. Christine Munk, Hazuki Tashima, Krista Reitenga, Lance D. Green, and Cliff S. Han

The Genome Science Group, Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM, U.S.A.

BioTechniques 53:61–62 (July 2012) doi 10.2144/0000113891

Keywords: Sequencing; hairpin structure; GC rich; PacBio.

Advances in sequencing technologies have dramatically reduced costs in producing high-quality draft genomes. However, there are still many contigs and possible misassembled regions in those draft genomes. Improving the quality of these genomes requires an efficient and economical means to close gaps and resequence some regions. Sequencing pooled gap region PCR products with Pacific Biosciences (PacBio) provides a significantly less expensive means for this need. We have developed a genome improvement pipeline with this strategy after decreasing a loading bias against larger PCR products in the PacBio process. Compared with Sanger technology, this approach is not only cost-effective but also can close gaps greater than 2.5 kb in a single round of reactions, and sequence through high GC regions as well as difficult secondary structures such as small hairpin loops.

Second-generation sequencing technologies produce more and more draft genomes at an ever faster speed and lower cost. However, finished, high-quality genomes are still preferred by researchers (1). Closing gaps in a draft genome is necessary to improve the quality of the final genome sequence. Picking primers at gap regions for PCR and assembling the resulting PCR sequences into the genome can reduce numbers of both contigs and scaffolds. Since the advancement of less expensive sequencing technologies, Sanger sequencing (2,3) of individual PCR products spanning targeted regions becomes a more expensive method compared with the cost of the draft itself. Pooling dozens of PCR products of various sizes and sequencing them as one library with single molecular sequencing technology from PacBio is a more economical option (4,5). However, there is a loading bias against large DNA fragments in the PacBio sequencing process. The PacBio technique uses single

molecule sequencing done in wells on a chip, which is called a single molecule, real time (SMRT) cell. Smaller PCR products will load into the PacBio wells with a greater efficiency than larger PCR products. When PCR products ranging from 500 bp to 5 kb are pooled and sequenced together using PacBio, the smaller products have a substantially higher coverage than the larger products resulting in poor quality or incomplete sequences for the larger PCR products. To address this problem, we tried adjusting the molar ratio of the PCR products when pooling them together based on the PCR size and concentration. This resulted in a much closer distribution of coverage for the different sizes of PCR products. We used a finished genome to normalize our process, and we chose 18 PCR primer pairs with amplification sizes ranging from 500 bp to 5 kb. Amplifications were performed using two commercial kits: FailSafe PCR System (Epicentre, Madison, WI, USA) for genomes

with mid-range GC content (40%–60%) and GC-Rich PCR System (Roche, Indianapolis, IN, USA) for genomes with GC content higher than 60%. PCR products were cleaned individually (ZR 96 DNA Clean and Concentrator, Zymo Research, Irvine, CA, USA) and pooled PCRs were purified again (Agencourt AMPure XP, Beckman Coulter, Brea, CA, USA). The results were pooled into three groups for three different approaches: group one was our control, where we loaded equal DNA mass for each PCR product; for group two, we pooled each PCR product at the equal molar mass; for group three, we adjusted the molar mass ratio based on the size of the PCR product, increasing the molar mass with size. The results can be seen in Figure 1. The control group resulted in much higher coverage for the smaller PCR products while the longer PCR products were barely covered (shown in the blue bars in Figure 1). The second group had an improvement in coverage for the larger products, but still less than the coverage for the smaller products (shown in the red bars in Figure 1). The third group shows dramatic improvement in the coverage for the larger products (shown in the green bars in Figure 1). For group 3, the formula below was used to make the molar mass adjustment and to calculate the volume needed for each PCR and for robotic pooling:

$$\text{Volume} = [\text{PCR size (kb)}]^2 \times [10\text{ng/PCR concentration (ng/}\mu\text{l)}]$$

We have combined over 200 PCR products into one pool and the process produced good coverage for the products. Since one SMRT cell can produce 0.5 gigabases of data after filtering to remove adapters, the process described in this paper provides us with an efficient method for potentially pooling 500–1000 PCR products into one sequencing library, depending on the sizes of the PCR products. By decreasing the loading bias against larger PCR products in the PacBio technology, we have developed a much more efficient and economical method to close gaps in draft genomes since larger gaps can be closed with the PacBio technology.

We applied this gap closure method to 16 diverse bacterial genome projects in our genome improvement pipeline. We selected primers for 362 regions in these 16 projects and sequenced the resulting products with both Sanger and PacBio technologies. The gap sizes ranged from 500 bp to 5 kb. While the majority of gaps less than 2.5 kb were closed with both Sanger (64%) and PacBio (73%) technologies, none of the gaps larger than 2.5 kb were closed with a

Table 1. PacBio closes larger gaps and hard stops in a single round of PCR.

PCR	# PCR	% closed by Sanger	% closed by PacBio
<2.5 kb	246	64	73
>2.5 kb	113	0	88
hard stop	3	0	100

362 PCR products (each covering a different gap) were sequenced with both Sanger and PacBio technologies. While the majority of gaps less than 2.5 kb were closed with both Sanger (64%) and PacBio (73%) technologies, none of the gaps larger than 2.5 kb were closed with a single round of Sanger technology. Three hard stop gaps were all closed by PacBio.

Figure 1. Changes in coverage of PCR products by PacBio subreads as molar mass of PCR product is changed. Three groups of 18 PCR products with sizes ranging from 500 bp to 5 kb were pooled in three PacBio libraries according to mass or molar mass and sequenced. Group 1 with constant mass for all PCR resulted in much higher coverage for the smaller PCR products while the longer PCR products were barely covered. Group 2 with constant molar mass had an improvement in coverage for the larger products, but still less than the coverage for the smaller products. Group 3 with adjusted molar mass by PCR length shows dramatic improvement in the coverage for the larger products.

single round of Sanger technology (Table 1). PacBio sequencing of the PCR products closed almost 90% of these larger gaps. This method also allows the closure of gaps due to small hairpin structures (typically with higher GC content) called hard stops, since PacBio can successfully sequence through these regions whereas other sequencing technologies usually fail. Because one of our goals is to reduce costs, we pooled more than 200 PCR products in a single PacBio SMRT cell for sequencing. To successfully assemble the PacBio sub-reads (a sub-portion of a read resulted from screening and removing of sequencing adapters that were in the middle of a read) into an accurate consensus for a single PCR product we pulled out sub-reads from the pool of sequenced sub-reads that belonged only to that PCR product. We developed computational scripts to interact with our local database to identify the primer sequences and an additional 150-nucleotide unique sequence next to the primers from the draft assembly to fish

out the sub-reads—using BLAST (6)—that belong to a particular PCR product and, therefore, a particular gap. This is especially necessary for repeat gaps so that if there are slight differences in the repeats, they can be resolved correctly. Of course, since the error rate of PacBio sequencing is typically reported to be about 15%, we needed to increase the accuracy of the consensus we obtained. By choosing 200 sub-reads with the highest sequence matches to the primer-plus-150nt-unique sequences we were able to dramatically improve the quality of the PCR product consensus sequences after assembling the selected sub-reads for an individual PCR product using Allora, the long read assembler for de novo assembly from PacBio. For the smaller gaps where the missing sequences were resolved by both Sanger and PacBio technologies, 91% of the PacBio consensus sequences matched the Sanger sequences with a 98% identity or better. To try to maintain this accuracy rate, we found that for the larger PCR

products we needed to increase the number of sub-reads to 300. We did not see a significant difference in the results based on the GC content of the genomes. For genomes with mid-range GC content (40%–60%), 78% of 51 PCRs closed the gap. For genomes with high GC content (>60%), 86% of 311 PCRs closed gaps. We continue to investigate ways to get the highest accuracy possible from the PacBio PCR consensus sequences for the larger gaps.

Acknowledgments

We want to thank Beverly Parson Quintana and Yuliya A. Kunde for supporting this work in PCR and PacBio sequencing process. The work is funded by DOE Joint Genome Institute through contract W-7405-ENG-36.

Competing interests

The authors have a patent pending for the technology described in this paper.

References

1. Chain, P.S., D.V. Grafham, R.S. Fulton, M.G. Fitzgerald, J. Hostetler, D. Muzny, J. Ali, B. Birren, et al. 2009. Genomics. Genome project standards in a new era of sequencing. *Science* 326:236–237.
2. Sanger, F., G.M. Air, B.G. Barrell, N.L. Brown, A.R. Coulson, C.A. Fiddes, C.A. Hutchison, P.M. Slocumbe, and M. Smith. 1977. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 265:687–695.
3. Sanger, F., J.E. Donelson, A.R. Coulson, H. Kossel, and D. Fischer. 1973. Use of DNA polymerase I primed by a synthetic oligonucleotide to determine a nucleotide sequence in phage fl DNA. *Proc. Natl. Acad. Sci. USA* 70:1209–1213.
4. Eid, J., A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* 323:133–138.
5. Flusberg, B.A., D.R. Webster, J.H. Lee, K.J. Travers, E.C. Olivares, T.A. Clark, J. Korlach, and S.W. Turner. 2010. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* 7:461–465.
6. Altschul, S.F., T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389–3402.

Received 20 April 2012; accepted 11 June 2012.

Address correspondence to either Xiaojing Zhang or Cliff S. Han. The Genome Science Group, Bioscience Division, Mail stop 888, Los Alamos National Laboratory, Los Alamos, NM 87545, USA. E-mail: xlz@lanl.gov (X.Z.) or han_cliff@lanl.gov (CSH).

To purchase reprints of this article, contact: biotechniques@fosterprinting.com