Benchmarks

Optimization and cost-saving in tagmentation-based mate-pair library preparation and sequencing

Kaori Tatsumi, Osamu Nishimura, Kazu Itomi, Chiharu Tanegashima, and Shigehiro Kuraku

Phyloinformatics Unit, RIKEN Center for Life Science Technologies, Kobe, Japan

BioTechniques 58:253-257 (May 2015) doi 10.2144/000114288

Keywords: mate pair library; genome assembly; genome scaffolding; CEGMA; tagmentation

Supplementary material for this article is available at www.BioTechniques.com/article/114288.

In de novo genome sequencing, mate-pair reads are crucial for scaffolding assembled contigs. However, preparation of mate-pair libraries is not a trivial task, even when using one of the latest approaches, the Nextera Mate Pair Sample Prep Kit from Illumina. To reduce cost and enhance library yield and fidelity when using this kit, we have modified the manufacturer's protocol based on (i) variable tagmentation conditions, (ii) intensive DNA shearing to decrease library insert length, and (iii) sequencing on an Illumina HiSeq with >150 cycles. Finally, we provide additional suggestions for further improvement in the application of this kit.

While the genome sequences of many traditional model species have become available, de novo sequencing of unexplored species remains challenging (reviewed in Reference 1). Particularly crucial is the use of mate-pair reads for joining contigs, a process known as genome scaffolding. Traditional approaches for obtaining mate-pair reads often require large amounts of genomic DNA, especially for libraries with long mate distances, as well as lengthy laboratory procedures (2,3). The release of the Nextera Mate Pair Sample Prep Kit (Illumina, San Diego, CA), based on tagmentation technology, has significantly reduced the difficulty, preparation time

and cost of preparing mate-pair libraries (4). However, there remain opportunities to improve the efficiency and reduce costs for this preparation technique. Here we present several key modifications of the manufacturer's protocol that address these issues.

The first key step in our optimization is a modification of the tagmentation conditions. The standard protocol provides only one condition with a fixed volume of tagment enzyme (12 μ l), but we performed the reaction with reduced volumes of the enzyme (4 and 8 μ l). We also tested self-made tagment buffer prepared according to another published protocol (5). For this test case, we used 4

μ g of genomic DNA from the Madagascar ground gecko *Paroedura picta*, quantified with Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific, Waltham, MA), in each reaction. The tagmented DNA was analyzed with pulse-field electrophoresis (Figure 1). The self-made buffer exhibited comparable tagmentation efficiency to the buffer supplied in the kit, while reactions with reduced volumes of enzyme yielded tagmented DNA with altered length distributions (Figure 1). The use of a reduced volume of enzyme and self-made buffer allows for a higher number of tagmentation reactions to be attempted using variable conditions. This strategy enables the identification of ideal reaction conditions that maximize the yield of DNA in the intended size range for any given genomic DNA sample. In particular, this modification, which yields larger DNA molecules, is advantageous in preparing mate-pair libraries spanning large distances (>10 kb), which tends to be hindered by low yields.

In addition to the tagmentation step, the strand displacement step employs some reagents whose amounts limit the number of preparations. To address this limitation, the strand displacement reaction was performed after size selection on a BluePippin (Sage Science, Beverly, MA) (see the Figure 2 legend)—not before size selection as in the standard protocol—which yielded a sufficient quantity of mate-pair libraries. Reversing the order of strand displacement and size selection enables significantly smaller volumes of strand displacement polymerase and buffer for reduced amounts of size-selected DNA, which allows for a larger number of preparations (up to 3-fold) than the standard protocol. This approach provides substantial cost savings, as other reagents for downstream steps are bundled for as many as 48 reactions, but the Nextera Mate Pair Sample Prep Kit is only intended for 12 reactions based on the “gel-plus” protocol.

When preparing mate-pair libraries, it has traditionally been a major concern that a high proportion of library inserts

METHOD SUMMARY

Here we describe several key modifications to the Nextera Mate Pair Sample Prep Kit, including optimizing DNA tagmentation conditions and improved coordination between insert size and sequence read length, that result in higher mate-pair library yield and fidelity while reducing costs.

can be derived from genomic fragments that are not true mate pairs. Thus, for scaffolding, we adhere to a conservative policy, using only read pairs in which a junction adapter, the landmark of proper mate-pair structure, is successfully identified. To increase the proportion of read pairs harboring the junction adapter, we optimized library insert length and read length. First, to adjust insert length, we performed shearing of circularized DNA with a Covaris S220 sonicator (Covaris, Woburn, MA). We used the same setting as instructed in the standard protocol, except that shearing was repeated from three to seven times, depending on the species, instead of once. The prepared libraries had distinct length distributions, with a considerably higher proportion of DNA molecules shorter than 600 bp, compared with those prepared following the standard protocol (Figure 2). Second, regarding read length, we sequenced the prepared libraries using the Illumina HiSeq rapid run mode with paired-end (PE) 171 cycles with single indexing ($2 \times 171 + 7 = 349$ cycles). This was achieved by using 4 of the TruSeq Rapid SBS v1 kits for 50 cycles, each of which allows no less than 87.5 cycles, including extra cycles for dual indexing, according to the standard protocol in the manual for the SBS kit ($4 \times 87.5 = 350$ cycles). This strategy, employing multiple 50-cycle kits, provides the lowest cost per cycle

Figure 1. Size distributions of tagged genomic DNA. Genomic DNA was tagged with the buffer supplied in the kit and with self-made buffer, as well as with variable amounts of enzyme. Electrophoresis was performed using a gel prepared with SeaKem GOLD agarose (LONZA, Basel, Switzerland) in 0.5× TBE on a PippinPulse (Sage Science, Beverly, MA) using its “0.5–15 kb” preset program.

and will also be economical when used with the latest Rapid SBS v2 kit.

Overall read properties were analyzed by FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc/), whose recent version (v0.11 or later) reports the proportions of detected Nextera and TruSeq adapters along their base positions in the reads (Supplementary Figure S1). In terms of these measures, an ideal library would exhibit a high proportion of reads in which

the Nextera junction adapter is detected, in addition to a low proportion with the TruSeq adapter. Removal of junction adapters was done using the NextClip (version 1.3) (6) program with default parameters. We compared reads from Library 1, which was prepared according to the standard protocol targeting the mate distance range of 1–6 kb, to those of Library 2, the library prepared according to our modified protocol targeting the same mate distance range (Table 1).

Figure 2. Size distribution of prepared libraries. The red curve shows the length distribution of the library prepared with intensive shearing with Covaris (Library 2), while the blue curve shows the library prepared according to the standard protocol (Library 1). Both libraries were prepared using DNA fragments between 1 and 6 kb, size-selected on a BluePippin (Sage Science, Beverly, MA) using the 0.75% gel cassette for 1–10 kb (Catalog # BLF7503) with the built-in program “0.75%DF2–6kb” and measured with a High Sensitivity DNA Kit on Bioanalyzer (Agilent Technologies, Santa Clara, CA).

BLOW UP THE OLD WAY OF TESTING gDNA

Whole-genome QC
now has one workflow.

FULLY AUTOMATED FRAGMENT ANALYZER™ DOES IT ALL.

- Assesses gDNA concentration up to 40,000 bp
- Generates a user-defined Genomic Quality Number
- Identifies RNA contamination in gDNA
- Works with samples as small as 0.1 µL

No chips. No tapes. No compromises.

More at AATI-US.COM

Table 1. Properties of the prepared mate-pair libraries.

Library	Targeted mate distance in size selection (kb)	Peak in library size distribution* (bp)	# of PCR cycles	Read length (nt)**	Duplicate read pairs (%)	Proportion of usable read pairs*** (%)
Library 1	1–6	869	10	101	0.48	33.6
				171	0.38	59.9
Library 2	1–6	545	8	171	0.60	65.4
Library 3	11–18	501	10	171	3.08	68.1

Note that for each library 15 million read pairs were randomly sampled with the program seqtk (<https://github.com/lh3/seqtk>) and used in the measures of duplicate read pair % and the proportions of usable reads by the program NextClip.

*This column shows the values reported by the Bioanalyzer (see Supplementary Figure S1).

**Reads with a length of 101 nucleotides (nt) were obtained by trimming the end of 171 nt reads using the program fastx_trimmer (http://hannonlab.cshl.edu/fastx_toolkit/).

***This column shows the proportion of read pairs that have a junction adaptor and are longer than 25 nt after the removal of the adaptor.

The following statistics are based on random sampling of an adjusted number of sequenced reads among the different libraries (15 million read pairs). In comparison to Library 1, sequencing Library 2 with PE 171 cycles resulted in the largest proportion of read pairs (65.4%) that are regarded as usable in genome scaffolding (those in Categories A, B and C as defined by NextClip) (Table 1). This comparison shows that longer reads largely contribute to the increased proportion of usable reads, an effect that is enhanced with intensive shearing. Moreover, we sequenced a library prepared according to our modified protocol targeting the mate distance range of 11–18 kb (Library 3), and this library yielded a comparable proportion of usable reads with only 10 PCR cycles (Table 1).

Using the mate-pair reads, we computed tentative genome scaffolding using the Platanus (version 1.2.1) program (7). Scaffolding with mate-pair reads extensively increased the overall sequence continuity, as shown by the increased N50 values. Moreover, these

libraries provided improved coverage of the genome (Table 2) as indicated by the proportion of core eukaryotic genes (CEG) detected with the program CEGMA version 2.4 (8). Scaffolding using reads from Library 2 sequenced with 171 cycles had the best score (Table 2), albeit evaluation of genome assemblies requires multifaceted assessment (9). The power of our modified protocol was most evident when multiple libraries with varied mate distances were employed in scaffolding (Table 2).

In our trials using *P. picta* genomic DNA, we targeted relatively broad size ranges of tagged DNA fragments (1–6 kb and 11–18 kb), but for a different species, we succeeded in preparing mate-pair libraries with narrower mate distance ranges (e.g., 12–16 kb). It is of note that our protocol enabled library preparation with no more than 10 PCR cycles even for >10 kb mate distance to yield 20 µl of library solution of up to 7 nM (0.14 pmol), which kept the PCR duplicate rates down to less than 4% among 15 million read pairs (Table 1).

Table 2. Contribution of mate-pair reads to genome scaffolding.

Reads used*	Scaffold N50 (kb)	Max scaffold length (kb)	Gene coverage (%)**
PE only	13.3	382.9	58.9
PE + MP (Library 1, 101 nt)	25.9	1420.0	71.3
PE + MP (Library 1, 171 nt)	41.0	1790.6	75.4
PE + MP (Library 2, 171 nt)	51.8	1792.6	77.4
PE + MP (Libraries 2 and 3, both 171 nt)	1819.1	2425.0	82.7

*This column indicates read types [paired-end (PE) or mate-pair (MP)] and choice of MP libraries used in the scaffolding step after assembling only paired-end (PE) reads.

**This column shows the numbers of core eukaryotic genes (CEGs) detected as “complete” by CEGMA runs with the “-vrt” option.

Overall, good control of DNA shearing allows flexible coordination between library insert length and read length to maximize the efficiency of mate-pair sequencing. This coordination has a large impact on the proportion of reads recognized as true mates and their contribution to scaffolding that depends on read lengths, in addition to the cost of sequencing reagents. Most importantly, we propose a new cost-saving scheme for library preparation (allowing for approximately 1/3 the current cost) that does not require the purchase of additional commercial products or laborious preparation of self-made items (as described in Reference 10), except for the tagment buffer.

The mate-pair reads from this study have been deposited in the NCBI Sequence Read Archive (SRA) under BioProject ID PRJDB3317. The genome sequences of the Madagascar ground gecko after scaffolding will be released in a designated database, after sequencing more reads and rebuilding scaffolds, which will be announced on our laboratory web site (www.clst.riken.jp/phylo/imate.html).

Author contributions

K.T., C.T., K.I., O.N., and S.K. conceived the original idea of the study. K.T. carried out the experiments, and S.K., O.N., and K.T. analyzed the data. K.T., C.T., K.I., O.N., and S.K. edited the manuscript. S.K. supervised the project.

Acknowledgments

We thank all other members of GRAS, the sequencing core of RIKEN Kobe Institute in the Biosystem Dynamics Group, RIKEN CLST, for insightful discussion as well as Eriko Kajikawa, Shinichi Aizawa, Yasuhide Furuta, and Hiroshi Kiyonari in the Biosystem Dynamics Group, RIKEN CLST, for the supply of animal materials. Our gratitude extends to Helen Gunter, Yuichiro Hara, Sean Keeley, and Mitsutaka Kadota for critical reading of the manuscript.

Competing interests

The authors declare no competing interests.

References

1. Baker, M. 2012. *De novo* genome assembly: what every biologist should know. *Nat. Methods* 9:333-337.
2. Van Nieuwerburgh, F., R.C. Thompson, J. Ledesma, D. Deforce, T. Gaasterland, P. Ordoukhanian, and S.R. Head. 2012. Illumina mate-paired DNA sequencing-library preparation using Cre-Lox recombination. *Nucleic Acids Res.* 40:e24.
3. Peng, Z., Z. Zhao, N. Nath, J.L. Froula, A. Clum, T. Zhang, J. Cheng, A.C. Copeland, et al. 2012. Generation of long insert pairs using a Cre-loxP inverse PCR approach. *PLoS ONE* 7:e29437.
4. Park, N., L. Shirley, Y. Gu, T.M. Keane, H. Swerdlow, and M.A. Quail. 2013. An improved approach to mate-paired library preparation for Illumina sequencing. *Methods in Next Generation Sequencing* 1:10-20.
5. Wang, Q., L. Gu, A. Adey, B. Radlwimmer, W. Wang, V. Hovestadt, M. Bahr, S. Wolf, et al. 2013. Tagmentation-based whole-genome bisulfite sequencing. *Nat. Protoc.* 8:2022-2032.
6. Leggett, R.M., B.J. Clavijo, L. Clissold, M.D. Clark, and M. Caccamo. 2014. NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries. *Bioinformatics* 30:566-568.
7. Kajitani, R., K. Toshimoto, H. Noguchi, A. Toyoda, Y. Ogura, M. Okuno, M. Yabana, M. Harada, et al. 2014. Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* 24:1384-1395.
8. Parra, G., K. Bradnam, and I. Korf. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23:1061-1067.
9. Bradnam, K.R., J.N. Fass, A. Alexandrov, P. Baranay, M. Bechner, I. Birol, S. Boisvert, J.A. Chapman, et al. 2013. Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. *Gigascience*. 2:10.
10. Picelli, S., Å.K. Björklund, B. Reinius, S. Sagasser, G. Winberg, and R. Sandberg. 2014. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* 24:2033-2040.

Received 28 October 2014; accepted 24 February 2015.

Address correspondence to Shigehiro Kuraku, Phyloinformatics Unit, RIKEN Center for Life Science Technologies, Kobe, Japan. E-mail: shigehiro.kuraku@riken.jp

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

NEW VERSION
SEQUENCHER® 5.3
 with Sequencher® Connections

POWER WITH SIMPLICITY

- New RNA-Seq analysis using Cufflinks Suite
- Perform multiple analyses on DNA sequences
- New visualizations

Gene Codes Corporation
 T C A G A G T G E N E C O D E S

734.769.7249
 GeneCodes.com