A scaled-down workflow for Illumina shotgun sequencing library preparation: lower input and improved performance at small fraction of the cost

Jo Ann Tan[1] and Alexander S. Mikheyev[1*]

Ecology and Evolution Unit, Okinawa Institute of Science and Technology Graduate University, 1919-1 Tancha, Onna-son, Kunigami-gun, Okinawa, Japan 904-0495

[*]corresponding author: mikheyev@homologo.us

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2475v1 | CC BY 4.0 Open Access | rec: 23 Sep 2016, publ: 23 Sep 2016

Abstract

The high cost of library preparation remains a major obstacle to sequencing large numbers of individual genomes. Illumina's proprietary tagmentation technology allows for rapid and easy preparation of sequencing libraries, but remains prohibitively expensive for many users. Here we propose a modified version of the protocol, which uses Illumina reagents at 1/20th the scale. We show that the scaled-down protocol performs comparably to that of the manufacturer on a non-model insect genome. Surprisingly, the scaled-down protocol also produced 14% fewer PCR duplicates that the full-scale protocol. Since PCR duplicates effectively wasted redundant data, our protocol presented here can help save not just library preparation costs, but sequencing costs as well.

Introduction

Genome sequencing costs continue to drop exponentially as a result of new and improved platforms, with the barrier for the $1000 human genome already broken (van Dijk et al., 2014). It is becoming possible to go beyond the single genome, into the genomics of whole populations. Population genomics can be cost-effective, since low-coverage sequencing of genomes can provide accurate inference of ancestry, demographics, selection, and phylogenetic relationships (Fumagalli et al., 2013; Skotte, Korneliussen & Albrechtsen, 2013; Tin, Economo & Mikheyev, 2014; Mikheyev et al., 2015; Vieira et al., 2016). Furthermore, if hundreds or thousands of numbers of genomes are sequenced, even at low depth, they can be used to computationally impute phase, and fill-in sporadically missing data (Browning & Browning, 2009). While sequencing costs make these developments possible, there remains an important obstacle – the cost of preparing sequencing libraries.

By contrast with genome sequencing technology, where platforms proliferate, there are far fewer ways to create genome sequencing libraries. Available from a range of manufacturers, most commercial kits rely on DNA shearing and adaptor ligation, reactions where chemistry is mature and further dramatic reduction in price is unlikely. DNA tagmentation represents a major advance in the preparation of Illumina sequencing libraries, relying on a transposase to simultaneously fragment DNA and insert sequencing adaptors (Adey et al., 2010). Tagmentation is rapid, straightforward to carry out, and requires nanogram-scale DNA input. Unfortunately, this technology is proprietary to Illumina under the "Nextera" brand, and remains prohibitively expensive for large numbers of samples. Though a "do it yourself" protocol has been published (Picelli et al., 2014), it requires substantial protein purification skills, and rigorous quality control on the part of the user to ensure reproducibility.

The most obvious solution to the high cost of Nextera kits is diluting the reaction volume to a fraction of manufacturer's recommendation. This approach has been proposed for sequencing microbial genomes (Lamble et al., 2013; Baym et al., 2015). However, how well dilution performs on eukaryotic genomes, which are orders of magnitude larger, is unknown (Baym et al., 2015). Likewise, its performance compared to full-scale kits is also unknown. Here we propose a further modification of the Nextera XT protocol, which uses 1/20th-scale reactions (Figure 1). We validate this protocol using a non-model eukaryotic genome, and show that it produces comparable, or better performance than the the manufacturer's protocol, at a small fraction of the total price.

Materials and Methods

DNA extraction normalization. DNA was extracted from 45 pharaoh ant thoraxes (*Monomorium pharaonis*) using QIAamp DNA Micro Kit (Qiagen) following the manufacturer's instructions. The genome of this ant is roughly 300 megabases long (Mikheyev & Linksvayer, 2015). The samples contained equal numbers of worker, queen and male castes. The castes vary in body size by a factor of ten, but DNA concentration was normalized to 0.5 ng/μl using PicoGreen® (Invitrogen). Our lab's internal protocol for this assay can be found at (http://ecoevo.unit.oist.jp/site/methods), which closely follows the manufacturer's instructions.

Scaled-down and control reactions. Libraries for 15 randomly selected samples (5 from each caste) were prepared using the full-scale Illumina XT protocol up to the "Size selection" step below. For the rest of the samples, tagmentation and PCR amplification was carried out as follows.

Tagmentation. After thawing reagents on ice, tagmentation was carried out by mixing together 0.5μl TD buffer, 0.25μl each of DNA and ATM buffer in one PCR tube, while ensuring no bubbles form. After a brief centrifugation, the reaction was incubated in a thermocycler at 55°C, followed by a 10°C hold. Finally the reaction was neutralized using 0.25μl of the NT buffer, mixed in by repeated pipetting.

PCR amplification. To the tagmentation reaction (1.25μl) we added 0.25μl each Index 1 and 2 primers, and 0.75μl NPM for a total volume of 2.5μl, again making sure that there are no bubbles, and spinning down the reaction before placing it in the thermocycler. The PCR was carried out with the following conditions: 72°C for 3 minutes, then 95°C for 30 seconds, then 12 cycles with denaturation at 95°C for 10 seconds, annealing at 55°C for 30 seconds, and elongation at 75°C for 10 seconds. The reaction was terminated by a final extension at 72°C for 30 seconds and a 10°C hold.

Size selection. Double solid phase reversible immobilization (dSPRI) allows for reliable size selection of sequencing libraries (Rodrigue et al., 2010). The current protocol closely follows that of Tin et al. (2015), but with lower volumes to accommodate smaller inputs. It requires advance preparation of several stock solutions: (1) two polyethylene glycol (PEG) stocks (13% and 13.5% PEG 6000 dissolved in ultrapure water, 0.9M NaCl and 10mM Tris-HCL, pH 6), (2) 70% aqueous ethanol (EtOH) with 10mM Tris-HCL, pH6 (not more than a week old). The dSPRI procedure was carried out as follows:

1. PCR reaction was adjusted to a total volume of 5μl with water.
2. 10μl of 13% PEG was added and mixed by pipetting.
3. 1μl of Dynabeads® MyOne Carboxylic Acid, CA-Beads (Invitrogen) was added and mixed by vortexing or pipetting.
4. The reaction was incubated at room temperature for 10 minutes and spun down on a benchtop mini-centrifuge for half a second.
5. The reaction was placed on a magnet stand for 5 minutes, after which 15μl of supernatant was transferred to a new tube.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2475v1 | CC BY 4.0 Open Access | rec: 23 Sep 2016, publ: 23 Sep 2016

6. 10µl of 13.5% PEG was added and mixed by pipetting.
7. 1µl of Dynabeads was added and mixed by vortexing or pipetting.
8. The reaction was incubated at room temperature for 5 minutes and spun down.
9. The reaction was placed on a magnet stand for 5 minutes. Steps 10 to 12 were all carried out on the stand.
10. Supernatant was discarded, and 50µl of 70% EtOH was added, followed by a 30 second incubation.
11. Ethanol was discarded and wash step repeated.
12. The beads were air dried for 5-10 minutes. **Important note**: Inexperienced users frequently over-dry the beads, leading to poor sample elution. Beads have reached the correct level of moisture when they change color from dark brown to light brown.
13. The tubes were removed from the magnetic stand, and 5 µl of EB buffer (Qiagen) was added, mixing by pipette. The mixture was incubated for 5 minutes.
14. The mixture was placed back on the magnetic stand for 5 minutes, after which the supernatant was removed.

Library quality control, pooling and sequencing. DNA concentration of each sample was determined using PicoGreen®, and 2 ng of each sample was added to a common pool. We checked size distribution using a Bioanalyzer (Agilent) using a high sensitivity DNA chip, but an equivalent device may be used. Final concentration was checked using ddPCR prior to sequencing at the OIST Sequencing Center. Sequencing was carried out on one lane of an Illumina HiSeq 4000 in PE150 mode.

Bioinformatic and statistical analysis. Reads were mapped to the *M. pharaonis* reference genome (V. 2.0) (Mikheyev & Linksvayer, 2015) using bowtie 2 with the default parameters (Langmead & Salzberg, 2012). Duplicates were detected using Picard tools' MarkDuplicates function. Other statistics, such as average insert size and the percentage of mapping reads, were directly computed from the read alignments. We used Kruskal-Wallis tests to compare small-scale *vs.* full-scale reactions.

Results

There was no difference in the overall number of reads sequenced for the small-scale *vs.* full-scale reactions: 27.6±37.2 million *vs.* 15.5±13.4 million (p = 0.51). There was a slight, though significant difference in the average length of mapped fragments (204±4.0 *vs.* 212±5.9 bp, $p = 2.0\times10^{-5}$), though standard deviations of fragment lengths were not significantly different between the two treatments (p = 0.057). The mapping percentage was negligible but significantly higher in the scaled-down treatment (86.8±0.45% *vs.* 86.4±0.50%, p = 0.029). Although the number of PCR cycles was the same in both treatments, the scaled-down reaction had markedly fewer PCR duplicates (14.3±3.3% *vs.* 28.2±6.6%, $p = 9.4\times10^{-6}$).

Discussion

The simple modification of Illumina's Nextera XT protocol significantly decreases the overall cost, while providing comparable, or even better performance. The overall reagent cost is approximately 400 ¥ (or substantially less – see Table 1), and the protocol takes 4.5 hours to complete. The modified library preparation protocol was not qualitatively different from the full-scale protocol in terms of yield, fragment size distribution or mapping percentage.

The scaled-down protocol is actually superior in several respects to the full-scale protocol. First, the total input DNA amount was only a quarter (0.25 ng) of the original protocol's. Second, there were almost half the number of PCR duplicates, compared to the full-scale protocol. This may be due to the greater efficiency of mixing under low reaction volumes. Interestingly, though only tested on microbes, the scaled-down protocol by Baym et al (2015) also reports low levels of PCR duplicates. Therefore, the scaled-down protocol may be more suited for material where input quantities are limited, since less template is needed, and it can presumably be amplified more cycles, without an excess of PCR duplicates. Importantly, since PCR duplicates represent wasted reads, the scaled-down protocol provides roughly 14% more sequencing data per run, which should result in substantial cost savings for sequencing in addition to savings on library preparation.

For some applications different fragment sizes may be desired. They can be tuned by adjusting the concentration of PEG 6000 solutions used in the size selection section of the protocol, as discussed in detail by Rodrigue *et al.* (2010). Alternatively, other bead-based purification approaches may be substituted, such as AMPure XP (Agencourt) beads used in the Illumina protocol. A notable advantage of the dSPRI method, resulting in a tighter distribution of fragments, compared with Illumina's protocol, which removes only small fragments. Tighter fragment size distributions allow for more accurate quantification of libraries downstream, which is particularly important as Nextera can produce long DNA fragments.

Due to the small volumes involved, automating the tagmentation and PCR amplification steps using conventional liquid handling platforms may be difficult. However, given the apparent improvement of performance at small scales, microfluidic miniaturization of the Nextera protocol may be the way forward for processing very large numbers of samples (Kim et al., 2013).

The Nextera library preparation protocol a wide range of applications beyond genome re-sequencing. For instance, our lab routinely uses them for RNA-seq studies using inexpensive reagents to generate cDNA as input (Aird et al., 2013, 2015). We hope that the straightforward protocol presented here will further democratize the cost of next-generation sequencing, paving the way for even more applications and discoveries.

Acknowledgements. We are grateful to Tim Linksvayer and Michael Warner for providing ants for this study and to the OIST Sequencing Center for sequencing the samples. Funding for this research was provided by the Okinawa Institute of Science and Technology Graduate University Subsidy for Operation and JSPS KAKENHI grants 25221206 and 16H06209 to ASM.

References

Adey A., Morrison HG., Asan., Xun X., Kitzman JO., Turner EH., Stackhouse B., MacKenzie AP., Caruccio NC., Zhang X., Shendure J. 2010. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome biology 11:R119.

Aird SD., Watanabe Y., Villar-Briones A., Roy MC., Terada K., Mikheyev AS. 2013. Quantitative high-throughput profiling of snake venom gland transcriptomes and proteomes (*Ovophis okinavensis* and *Protobothrops flavoviridis*). BMC genomics 14:790.

Aird SD., Aggarwal S., Villar-Briones A., Tin MM-Y., Terada K., Mikheyev AS. 2015. Snake venoms are integrated systems, but abundant venom proteins evolve more rapidly. BMC genomics 16:647.

Baym M., Kryazhimskiy S., Lieberman TD., Chung H., Desai MM., Kishony R. 2015. Inexpensive multiplexed library preparation for megabase-sized genomes. PloS one 10:e0128036.

Browning BL., Browning SR. 2009. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. American journal of human genetics 84:210–223.

van Dijk EL., Auger H., Jaszczyszyn Y., Thermes C. 2014. Ten years of next-generation sequencing technology. Trends in genetics: TIG 30:418–426.

Fumagalli M., Vieira FG., Korneliussen TS., Linderoth T., Huerta-Sánchez E., Albrechtsen A., Nielsen R. 2013. Quantifying population genetic differentiation from next-generation sequencing data. Genetics 195:979–992.

Kim H., Jebrail MJ., Sinha A., Bent ZW., Solberg OD., Williams KP., Langevin SA., Renzi RF., Van De Vreugde JL., Meagher RJ., Schoeniger JS., Lane TW., Branda SS., Bartsch MS., Patel KD. 2013. A microfluidic DNA library preparation platform for next-generation sequencing. PloS one 8:e68988.

Lamble S., Batty E., Attar M., Buck D., Bowden R., Lunter G., Crook D., El-Fahmawi B., Piazza P. 2013. Improved workflows for high throughput library preparation using the transposome-based Nextera system. BMC biotechnology 13:104.

Langmead B., Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nature methods 9:357–359.

Mikheyev AS., Tin MMY., Arora J., Seeley TD. 2015. Museum samples reveal rapid evolution by wild honey bees exposed to a novel parasite. Nature communications 6:7991.

Mikheyev AS., Linksvayer TA. 2015. Genes associated with ant social behavior show distinct transcriptional and evolutionary patterns. eLife 4:e04775.

Picelli S., Björklund AK., Reinius B., Sagasser S., Winberg G., Sandberg R. 2014. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. Genome research 24:2033–2040.

Rodrigue S., Materna AC., Timberlake SC., Blackburn MC., Malmstrom RR., Alm EJ., Chisholm SW. 2010. Unlocking short read sequencing for metagenomics. PloS one 5:e11840.

Skotte L., Korneliussen TS., Albrechtsen A. 2013. Estimating individual admixture proportions from next generation sequencing data. Genetics 195:693–702.

Tin MMY., Rheindt FE., Cros E., Mikheyev AS. 2015. Degenerate adaptor sequences for detecting PCR duplicates in reduced representation sequencing data improve genotype calling accuracy. Molecular ecology resources 15:329–336.

Tin MM-Y., Economo EP., Mikheyev AS. 2014. Sequencing degraded DNA from non-destructively sampled museum specimens for RAD-tagging and low-coverage shotgun phylogenetics. PloS one 9:e96793.

Vieira FG., Lassalle F., Korneliussen TS., Fumagalli M. 2016. Improving the estimation of genetic distances from Next-Generation Sequencing data. Biological journal of the Linnean Society. Linnean Society of London 117:139–149.

Figure 1. Outline of the library preparation procedure. The workflow and this outline is qualitatively similar to that of Baym et al (2015) (see their Fig. 1), except that this protocol uses dSPRI for PCR clean-up and size selection.

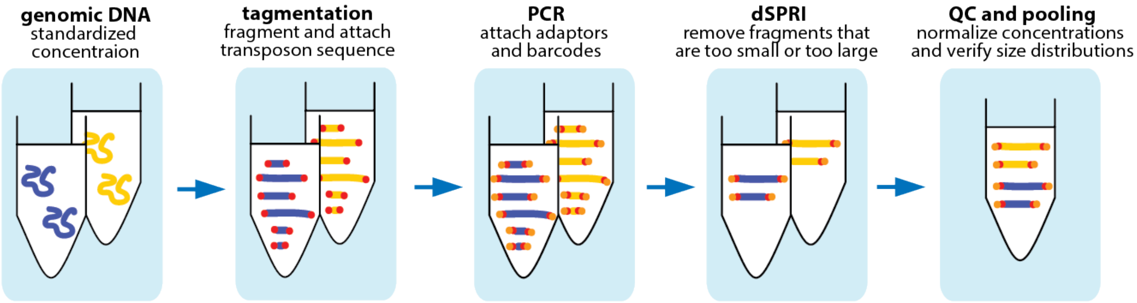
genomic DNA
standardized concentraion
tagmentation
fragment and attach transposon sequence
PCR
attach adaptors and barcodes
dSPRI
remove fragments that are too small or too large
QC and pooling
normalize concentrations and verify size distributions

Table 1. Approximate cost and time requirements for the scaled-down library preparation protocol. These costs do not include the Bioanalyzer run and ddPCR for the final quality control checks, since these costs depend on the number of libraries being run at one time. Sequencing centers typically keep them low by processing multiple libraries at once. The exchange rate fluctuates, but 100 ¥ is approximately 1 US$. Actual prices will vary based on region and institutional agreements with the manufacturers, and may be substantially cheaper, since the reagents we use in Japan are often sold at a substantial markup. For reference, in September 2016 our price for the most expensive reagents, the Illumina XT kit (96 samples, cat # FC-131-1096), was ¥384,400. By contrast, US price on the Illumina web site of US$2,760. This corresponds to a roughly 40% markup, which is typical of US-branded reagents in Japan. Therefore for many users, particularly those based in the US, costs may be considerably less than those listed here.

Protocol step	**Cost (¥)**	**Time (minutes)**
Normalization	50	60
Tagmentation	200	60
PCR amplification and adaptor addition	20	60
Size selection	80	45
Final quantification	40	45
Total	**390**	**270 (4.5 h)**