bioRxiv preprint doi: https://doi.org/10.1101/035436; this version posted December 28, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Genomic DNA preparation enabling multiple replicate reads for accurate nanopore sequencing

Dimitra Tsavachidou[1]

[1]Vastogen, Inc., Houston, TX

Correspondence: dimitra@vastogen.com

Web: www.vastogen.com

Running title: Sample prep method for nanopore sequencing

Key words: genome sequencing; nanopore sequencing; serial copies; multiple copies; concatemers; consensus sequence; whole genome sequencing; genomic DNA

Abstract

Sequencing at single-nucleotide resolution using nanopore devices is performed with reported error rates 10.5-20.7% (Ip et al., 2015). Since errors occur randomly during sequencing, repeating the sequencing procedure for the same DNA strands several times can generate sequencing results based on consensus derived from replicate readings, thus reducing overall error rates.

The method presented in this manuscript constructs copies of a nucleic acid molecule that are consecutively connected to the nucleic acid molecule. Such copies are useful because they can be sequenced by a nanopore device, enabling replicate reads, thus improving overall sequencing accuracy.

Serial Copies Method

Long-read sequencing technologies can be benefited significantly by sample preparation methods that enable multiple replicate reads of the same genomic DNA molecule.

For example, the PacBio (Pacific Biosciences) sequencing error rate for a single read is relatively high (around 11%–15%) (Rhoads and Au, 2015). The Circular Consensus Sequencing (CCS) method allows for the repeated sequencing of individual templates (Travers et al., 2010). The errors are distributed randomly in single reads, so that the overall error rate can be reduced by generating CCS reads with sufficient sequencing passes. For example, templates with at least 4 replicate reads (i.e. templates that went through at least 4 CCS passes) have a minimum Phred quality score of 20, templates with 7 replicate reads have a Phred score of at least 40, and those with 9 replicate reads have a Phred score of at least 50 (Larsen et al., 2014).

bioRxiv preprint doi: https://doi.org/10.1101/035436; this version posted December 28, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The MinION nanopore sequencing platform (Oxford Nanopore Technologies) can generate 2D reads (one read for the template and one for its complement), reducing the error rate to approximately 12% (Ip et al., 2015).

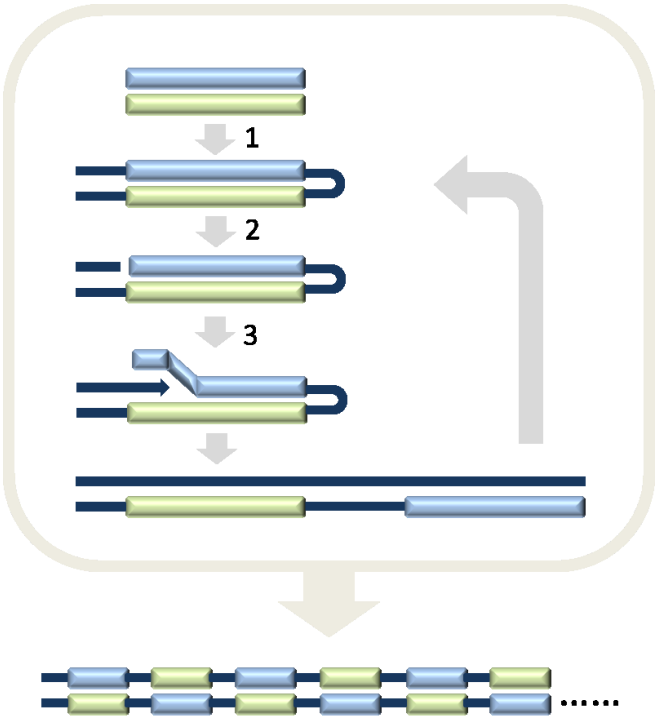

Figure 1: Serial copies method

Clearly, there is a need for further error rate reduction in nanopore sequencing. Methods of sample preparation that generate multiple DNA copies in tandem are desired. One such method (called the “serial copies method”) is shown in Figure 1.

During the 1st step of the method, a blunt-ended doubled-stranded DNA molecule is ligated to two adaptors, one of which is a hairpin. The other adaptor can be free in solution or attached to a solid surface.

During the 2nd step, the DNA molecule is subjected to incubation with nicking endonucleases that recognize a restriction site within the adaptor. The nicking endonucleases may nick within the adaptor or between the 3’ end of the adaptor and the adjacent 5’ end of the DNA molecule.

During the 3rd step, the DNA molecule participates in an extension reaction using strand-displacing polymerases. Owing to the presence of the hairpin, two connected copies of the genomic DNA are generated, one inverted in relation to the other.

The process can be repeated several times by subjecting the product of each extension reaction to a cycle of hairpin adaptor ligation, nicking and polymerization. The total number of copies generated after each cycle is double the number of copies of the previous cycle.

The design of this method allows running all steps concurrently, by including ligases, hairpin adaptors, nicking endonucleases and strand-displacing polymerases in a single reaction.

Importantly, the single reaction is combined with a proprietary procedure running concurrently, that prevents the formation of hairpin-to-hairpin ligation products and removes any such products formed, without using laborious and expensive size selection procedures (e.g.

bioRxiv preprint doi: https://doi.org/10.1101/035436; this version posted December 28, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

AMPure, gel separation). Additional information and data about this approach will be presented in future revisions of this manuscript.

The size of the construct produced by the serial copies method depends on the type of polymerase used in the extension steps. When phi29 is used, the total size can be ~50 kb. The smaller the original genomic fragment to be copied, the more copies can be generated for a fixed total construct size.

Nicking endonuclease sites may be present in the genomic fragments to be copied using the serial copies method. In this event, genomic fragments may be nicked and bias may be introduced, because of underrepresentation of regions at the 5' sides of nicks. Such bias may be prevented by performing two separate reactions for the same genomic sample, using a different type of nicking endonuclease in each reaction. There are other versions of the serial copies method that do not require nicking endonucleases. These versions will be discussed in future revisions of this manuscript.

An advantage of this method is that it produces multiple copies of a genomic fragment in a convenient double-stranded construct that can be subsequently attached to appropriate adaptors for nanopore sequencing. The serial copies method can be more useful than rolling-circle amplification, which produces less convenient single-stranded constructs and requires circularization which is typically an inefficient process. For example, circularization is used in the construction of mate-pair libraries, an expensive process that may require up to 15–20 µg of high molecular weight DNA of which most is lost during the enrichment step of the end-to-end ligated fragments (Knief, 2014). The widely used Nextera mate pair library preparation kit (Illumina) requires a minimum of 1 ug starting material to yield 2-8 kb products, but getting longer fragments (5-10 kb) requires more material (4 ug) and sacrifices library diversity (Illumina, 2014).

References

Illumina, 2014. http://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet_nextera_mate_pair.pdf.

Ip, C.L.C., Loose, M., Tyson, J.R., de Cesare, M., Brown, B.L., Jain, M., Leggett, R.M., Eccles, D.A., Zalunin, V., Urban, J.M., Piazza, P., Bowden, R.J., Paten, B., Mwaigwisya, S., Batty, E.M., Simpson, J.T., Snutch, T.P., Birney, E., Buck, D., Goodwin, S., Jansen, H.J.,

bioRxiv preprint doi: https://doi.org/10.1101/035436; this version posted December 28, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

O'Grady, J., Olsen, H.E., MinION Analysis and Reference Consortium, 2015. MinION Analysis and Reference Consortium: Phase 1 data release and analysis. F1000Research. doi:10.12688/f1000research.7201.1

Knief, C., 2014. Analysis of plant microbe interactions in the era of next generation sequencing technologies. Plant Genet. Genomics 5, 216. doi:10.3389/fpls.2014.00216

Larsen, P.A., Heilman, A.M., Yoder, A.D., 2014. The utility of PacBio circular consensus sequencing for characterizing complex gene families in non-model organisms. BMC Genomics 15, 720. doi:10.1186/1471-2164-15-720

Rhoads, A., Au, K.F., 2015. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics, SI: Metagenomics of Marine Environments 13, 278–289. doi:10.1016/j.gpb.2015.08.002

Travers, K.J., Chin, C.-S., Rank, D.R., Eid, J.S., Turner, S.W., 2010. A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic Acids Res. 38, e159. doi:10.1093/nar/gkq543