Bioinformatics, 31(12), 2015, 2035–2037
doi: 10.1093/bioinformatics/btv057
Advance Access Publication Date: 5 February 2015
Applications Note

OXFORD

Sequence analysis

NxTrim: optimized trimming of Illumina mate pair reads

Jared O'Connell[1], Ole Schulz-Trieglaff[1], Emma Carlson[2], Matthew M. Hims[2], Niall A. Gormley[2] and Anthony J. Cox[1,*]

[1]Computational Biology Group and [2]Technology Development, Illumina Cambridge Ltd., Chesterford Research Park, Little Chesterford, Essex CB10 1XL, UK

*To whom correspondence should be addressed.
Associate Editor: Alfonso Valencia

Received on August 6, 2014; revised on December 3, 2014; accepted on January 26, 2015

Abstract

Motivation: Mate pair protocols add to the utility of paired-end sequencing by boosting the genomic distance spanned by each pair of reads, potentially allowing larger repeats to be bridged and resolved. The Illumina Nextera Mate Pair (NMP) protocol uses a circularization-based strategy that leaves behind 38-bp adapter sequences, which must be computationally removed from the data. While 'adapter trimming' is a well-studied area of bioinformatics, existing tools do not fully exploit the particular properties of NMP data and discard more data than is necessary.

Results: We present NxTrim, a tool that strives to discard as little sequence as possible from NMP reads. NxTrim makes full use of the sequence on both sides of the adapter site to build 'virtual libraries' of mate pairs, paired-end reads and single-ended reads. For bacterial data, we show that aggregating these datasets allows a single NMP library to yield an assembly whose quality compares favourably to that obtained from regular paired-end reads.

Availability and implementation: The source code is available at https://github.com/sequencing/NxTrim

Contact: acox@illumina.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

A common design for DNA sequencing experiments is to sequence from both the 5′ and 3′ ends of the templates in a library to obtain *paired-end reads* for which the genomic distance between the two halves of each pair is approximately known, information which is useful for *de novo* assembly, alignment and variant calling.

Mate pair libraries add further value by increasing the effective genomic distance (EGD) between the two reads. The Nextera Mate Pair (NMP) protocol is typical: a library of longer DNA molecules is circularized and then fragmented to a size suitable for sequencing, the ends of each circle being joined by an *adapter* sequence tag that is biotinylated to allow enrichment for only those templates that span the join. Read pairs from these templates have an EGD that is determined by the length of the molecule that was circularised, thus yielding longer-range scaffolding information than can be deduced from a standard paired-end read library.

Before further processing, the known adapter sequence must be removed *in silico* to leave only genomic sequence behind. *Adapter trimming* is a generic task in bioinformatics for which a plethora of tools exist (comprehensively surveyed by Jiang *et al.*, 2014), including some, which are specialized to the particular needs of mate pairs (Leggett *et al.*, 2013). However, to our knowledge, all of them work by trimming the adapter and everything to the right of it (i.e. the 3′ side), retaining only the portion of the read that lies to the left of the adapter (its 5′ side).

Here we present a tool NxTrim which demonstrates that the 3′-ward portion of the read constitutes valuable 'real estate' that can be retained to improve coverage and *de novo* assembly quality.

More specifically, the sequence to the 3′ side of the adapter, together with the other half of the read pair, can be reinterpreted as a standard paired-end read. Depending on where the adapter lies in the read, we reinterpret the whole read pair as a single read plus either a mate pair or a paired-end read, choosing between the latter two options so as to maximize the number of bases that are paired. NxTrim converts raw NMP reads into four 'virtual libraries:'

- MP: a set of *known mate pairs* having an outward-facing relative orientation and an EGD whose distribution mirrors the size distribution of the circularised DNA.
- UNKNOWN: a set of read pairs for which the adapter could not be found within either read. Most likely the adapter will lie in the unsequenced portion of the template, although we note (Supplementary Table S5) some contamination with paired-end reads.
- PE: a set of *paired-end reads*, having an inward-facing relative orientation and an EGD whose distribution mirrors the size distribution of the sequenced templates.
- SE: a set of *single reads*.

Trimming tools following a '5′-only' strategy would produce output similar to the MP and UNKNOWN libraries combined. However, the versatile Velvet *de novo* assembler (Zerbino and Birney, 2008) can accept all four of these libraries as input to a single assembly and is able to treat the MP and UNKNOWN libraries differently in anticipation of a proportion of non-mate paired reads in the latter.

2 Methods

NxTrim's logic is described in the Supplementary Materials. Briefly, if the adapter is not found, we place the pair in the UNKNOWN library. If the adapter is detected at the end of one (or both) of the reads, the adapter is removed and the pair is placed in the MP library. If the adapter is at the beginning of a read, the adapter is removed and the pair is placed in the PE library. An adapter in the middle of the read gives rise to a split read. The longest of the split segments is paired with the other read, the pair being added to either the MP or PE library according to which of the 5′-ward or 3′-ward segments is longest. The remaining segment goes into the SE library if its length exceeds a configurable threshold that defaults to 21 bp.

We analyzed two replicates of each of nine common bacterial samples, all prepared according to the NMP protocol then sequenced as paired 150 bp reads during a single run of a MiSeq instrument (Supplementary Table S1). NxTrim's output was compared with that produced by the MiSeq instrument's on-board trimming routine, this being exemplary of the '5′-only' approach to adapter trimming used by all other tools we are aware of.

For each trimmer/sample combination, the reads were assembled by Velvet (version 1.2.10) for all odd *k*-mer sizes between 21 and 119, from which we chose the assembly that maximized contig N50. Contig N50 was strongly correlated with the number of genes detected (Supplementary Figure S2), so this appears to be a reasonable way of selecting an optimal *k*-mer size. We found scaffold N50 to be less correlated with the number of genes found, since it is possible to combine a large number of very small contigs into a long scaffold that has numerous gaps. Assemblies were evaluated using QUAST (Gurevich *et al.*, 2013).

Table 1. Effect of trimming strategies on coverage and assembly statistics for bacterial samples

Sample	'5′-only' adapter trimming			NxTrim		
	Cov	NG50	NGA50	Cov	NG50	NGA50
Bcer1	19.97	914	334	**23.48**	**1707**	**547**
Bcer2	22.69	1404	1264	**26.64**	**1698**	**1658**
EcDH1	40.58	726	361	**47.23**	**4182**	**730**
EcDH2	29.47	1447	308	**34.44**	**4200**	**548**
EcMG1	28.63	3924	401	**34.72**	**4602**	**666**
EcMG2	30.23	4120	332	**34.72**	**4589**	**499**
list1	57.88	2923	1500	**66.65**	**2928**	**2195**
list2	45.24	1494	1494	**52.31**	**2929**	**2424**
meio1	45.79	2539	630	**53.86**	**3002**	**1717**
meio2	40.81	2998	**1710**	**46.90**	**3000**	844
ped1	30.34	5147	**1658**	**35.22**	**5155**	1271
ped2	22.89	**4927**	886	**26.72**	4125	**2822**
pneu1	27.66	3934	539	**31.67**	**5289**	**698**
pneu2	24.97	**3709**	**694**	**28.77**	2934	642
rhod1	32.52	**4127**	2516	**38.26**	3191	**2737**
rhod2	37.87	**3196**	2934	**43.52**	3194	**3179**
TB1	39.14	2551	155	**45.99**	**4365**	**213**
TB2	32.57	**4368**	155	**38.01**	2534	**157**
Average	33.85	3025	993	**39.39**	**3535**	**1308**

Cov is the coverage of trimmed reads having length at least 35 bp. NG50 is the scaffold size such that atleast half of the reference genome is covered by scaffolds that size or larger. The definition of NGA50 is as for NG50, except scaffolds are broken at any mis-assemblies greater than 1 kbp in length. Scaffold sizes are given in kbp. The '5′-only' figures are computed from the reads output by the MiSeq instrument's on-board trimming routine. The trimmer with best metric for each sample is highlighted in bold.

3 Results

Supplementary Figure S3 shows the insert size distributions of the three virtual libraries, as estimated from alignments to the relevant reference genomes with BWA-MEM (Li, 2013). The MP (centre) and UNKNOWN (right) libraries display pleasingly similar distributions, having median insert sizes of 3.85 and 3.69 kbp, respectively, with a trace of small-insert contamination visible in the latter.

Assembly comparisons are summarized in Table 1 (with more detail in Supplementary Tables S2 and S3): on average, NxTrim achieves 39.39× coverage, an 16.4% improvement on the 33.85× obtained by the standard trimming routine. Reads from the standard trimming routine assemble to an average NG50 and NGA50 of 3.025 and 0.993 Mbp, respectively, while NxTrim improves these metrics to 3.535 and 1.308 Mbp. In many cases, the NxTrim assembly has scaffolded nearly the entire bacterial genome. While the lower NGA50 values suggest a number of mis-assemblies, most of these are due to mis-estimated gap sizes rather than more serious inversions or translocations (as illustrated by Supplementary Figure S4).

4 Discussion

NxTrim's ability to extract both long and short insert read pairs from mate-pair data helps enable a high-quality bacterial assembly to be obtained from a single NMP library: despite shorter reads and lower coverage, our *B. cereus* and *R. sphaeroides* assemblies compare favourably with those reported by the GAGE-B study (Magoc *et al.*, 2013, see Supplementary Table S6). The additional coverage retrieved by NxTrim becomes proportionately larger for longer NMP reads and, while we have used Velvet assemblies to demonstrate its value, should prove equally helpful to other assemblers such as SPAdes (Bankevich *et al.*, 2012).

Funding

All authors are employees of Illumina Inc., a public company that develops and markets systems for genetic analysis, and receive shares as part of their compensation.

Conflict of Interest: none declared.

References

Bankevich,A. *et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.,* **19**, 455–477.

Gurevich,A. *et al.* (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics,* **29**, 1072–1075.

Jiang,H. *et al.* (2014). Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics,* **15**, 182.

Leggett,R.M. *et al.* (2013) NextClip: an analysis and read preparation tool for nextera long mate pair libraries. *Bioinformatics*, **30**, 566–568.

Li,H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

Magoc,T. *et al.* (2013) GAGE-B: an evaluation of genome assemblers for bacterial organisms. *Bioinformatics*, **29**, 1718–1725.

Zerbino,D.R. and Birney,E. (2008) Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome Res.,* **18**, 821–829.