BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 21 2009, pages 2839–2840
doi:10.1093/bioinformatics/btp527

Genome analysis

TagDust—a program to eliminate artifacts from next generation sequencing data

Timo Lassmann*, Yoshihide Hayashizaki and Carsten O. Daub*

Omics Science Center, Riken Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan

Received on February 23, 2009; revised on August 25, 2009; accepted on September 1, 2009

Advance Access publication September 7, 2009

Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: Next-generation parallel sequencing technologies produce large quantities of short sequence reads. Due to experimental procedures various types of artifacts are commonly sequenced alongside the targeted RNA or DNA sequences. Identification of such artifacts is important during the development of novel sequencing assays and for the downstream analysis of the sequenced libraries.

Results: Here we present TagDust, a program identifying artifactual sequences in large sequencing runs. Given a user-defined cutoff for the false discovery rate, TagDust identifies all reads explainable by combinations and partial matches to known sequences used during library preparation. We demonstrate the quality of our method on sequencing runs performed on Illumina's Genome Analyzer platform.

Availability: Executables and documentation are available from http://genome.gsc.riken.jp/osc/english/software/.

Contact: timolassmann@gmail.com

1 INTRODUCTION

Next-generation sequencing is applied to address a whole range of biological questions (Mardis, 2008; von Bubnoff, 2008). A widely recognizable challenge lies in the computational treatment of the huge volumes of data being generated. An initial step is to verify whether a sequencing run was successful. A low mapping rate to a reference genome is commonly a good indicator of the run quality, however, it fails to explain the source of the unmapped sequences. From experience we know that large fractions of the unmapped sequences often correspond to artifacts arising from linker and adaptor sequences used in the library construction. Such artifacts are comparable with vector sequences found in traditional Sanger sequencing (White *et al.*, 2008).

The identification of these artifacts is important during the development of novel sequencing assays. More importantly, a fraction of artifacts commonly maps to reference genomes and can thus influence the biological interpretation of the libraries. The situation is particularly problematic when comparing two RNA samples sequenced at different biological states. If the total number of sequences from states A and B is the same but the fraction of artifacts is increased in state B, it may appear that non-artifactual sequences are downregulated compared with state A.

*To whom correspondence should be addressed.

Identification of known library sequences in sequenced reads should be trivial. However, sequencing errors, PCR errors, short read lengths, combinations of several fragmented sequences and their reverse complements complicate this task dramatically. To resolve this basic issue we developed TagDust, a program employing a fast, fuzzy string-matching algorithm to identify partial matches to library sequences in the reads. A read is annotated as an artifact if a large fraction of its residues can be explained by matches to library sequences.

2 METHODS

We previously employed the Muth–Manber algorithm (Muth and Manber, 1996) in the context of multiple alignments to quickly assess sequence similarity (Lassmann *et al.*, 2008). It allows for multiple string matching with up to one error (mismatch, insertion or deletion). The latter is achieved by creating libraries of k-mers from both query and target strings. The library is then extended to patterns of length $k-1$ by deleting each character in all the original k-mers in turn. For example, the 4mer *ACGT* will be converted into *CGT*, *AGT*, *ACT* and *CGT*. We will refer to these extended patterns as lk-mers. A comparison of these libraries via fast exact string matching reveals all matches with up to one error. For example, a mismatch in the original sequences is detected with an exact match of the lk-mers lacking the mismatched residues. As a default, TagDust used a k-mer length of 12.

For detecting artifacts, we are not really interested in the individual matches to a read but instead whether a large proportion of a read can be labeled as matching library sequences. Hence, we altered the default Muth–Manber algorithm to return the percentage of nucleotides involved in matches to library sequences and to run efficiently on very large datasets. Briefly, we record all lk-mers derived from the library sequences in a bitfield, scan all reads and identify matches with quick bit-lookups.

Since the sequenced reads are currently short, between 30–50 nt in length, spurious hits often occur. Discarding reads based on these matches is obviously undesirable. Therefore, it is crucial to select a suitable cutoff on the percentage of residues covered by library sequences. We approach this problem in a manner analogous to recent work by Zhang *et al.* (2008) relating to the interpretation of ChIP-sequencing data. Initially, we simulate a sequencing dataset with the same length distribution and nucleotide composition as the input dataset. Secondly, we apply the modified Muth–Manber algorithm to the simulated reads to derive a distribution of the number of reads labeled as 5%, 10%, …,100% library sequences. The distribution reflects how often we expect reads to be labeled as $X\%$ library sequence by chance. Finally, we obtain P-values from this null distribution and adjust them using the Benjamini–Hochberg method to reflect the controlled false discovery rate (FDR; Benjamini and Hochberg, 1995). The lowest sequence coverage that gives the requested FDR is then used as the cutoff value.

Table 1. Percentages of reads identified as artifacts in five sequencing runs at varying FDR thresholds

Description	Accession	Sequences	FDR 0.05 (%)	FDR 0.01 (%)	FDR 0.001 (%)	CPU sec.
Genomic PE (18 nt)	ERR000017	6 381 596	1.4 (98.79)	0.4 (98.91)	0.1 (98.61)	28
Genomic PE (36 nt)	ERR000130	10 209 914	3.2 (84.05)	0.8 (52.72)	0.4 (11.44)	84
Genomic (25 nt)	SRR000723	7 230 975	1.7 (57.64)	0.5 (54.26)	0.1 (36.44)	45
Chip-Seq (25 nt)	SRR000731	6 011 079	3.7 (29.15)	2.5 (12.81)	2.0 (1.73)	37
RNA-Seq (33 nt)	SRR002052	12 099 833	1.8 (23.32)	0.6 (22.30)	0.1 (20.38)	103

The mapping rates of the artifactual sequences to the human genome are indicated in brackets. The last column lists the runtime of TagDust in CPU seconds for the 0.05 FDR cutoff.

For efficiency, TagDust is implemented in the C programming language. TagDust uses <5 MB of memory since only single reads are read into memory at a time for processing. Hence, it is applicable to current datasets and the large volume of data expected with future next-generation sequencing instruments. A computational bottleneck is the calculation of the adjusted *P*-values since this step, in principle, requires sorting of millions of *P*-values. However, since sequence lengths are natural numbers, only a selection of coverage cutoffs and associated *P*-values is possible. For example, a 20-nt sequence can be 95% or 100% labeled as library sequences but not by 97%. We take advantage of this and use a bit-sort-like algorithm to perform this step in linear memory and time. TagDust is freely available from the OMICS software repository or by request from the author.

3 RESULTS AND DISCUSSION

Obtaining suitable datasets for benchmarking our method is not trivial since partially failed sequencing runs are commonly not deposited in public databases. Nevertheless, we obtained five datasets sequenced by the Illumina Genome Analyzer from the NCBI short read archive. We used the standard Illumina adaptors and primers used in the different sequencing assays as target sequences to be filtered out from the reads. As expected, only a relatively small percentage of the deposited reads can be explained by library sequences (Table 1). To determine whether the same sequences could be filtered out by simply mapping to the reference genome, we mapped all artifactual sequences with up to two mismatches to the human genome (hg18 assembly) using nexalign (T.Lassmann, manuscript in preparation). Evidently, a varying percentage of the artifactual sequences map to the genome. In the absence of replicates it is difficult to determine whether such tags are actual artifacts and hence we recommend users to merely flag such reads and their mapping positions.

TagDust processes even the largest dataset here in <2 min on a standard desktop PC while using <5 MB of memory. Conceivably, the time it takes to map libraries can be reduced by using TagDust to filter out artifacts before the mapping.

The two main applications for TagDust are to troubleshoot failed large-scaled sequencing runs and to filter out artifactual sequences from successful ones. The latter may affect the biological interpretation of the produced data since some artifactual sequences map to the respective reference genomes.

Funding: Research Grant for the RIKEN Omics Science Center from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government (MEXT to Y.H.); a grant of the Genome Network Project from the Ministry of Education, Culture, Sports, Science and Technology, Japan; the Strategic Programs for R&D of RIKEN Grant for the RIKEN Frontier Research System, Functional RNA research program.

Conflict of Interest: none declared.

REFERENCES

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.

Lassmann,T. *et al*. (2008) Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. *Nucleic Acids Res.*, **37**, 858–865.

Mardis,E. (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet.*, **24**, 133–141.

Muth,R. and Manber,U. (1996) Approximate multiple string search. In Hirschberg,D.S. and Myers,E.W. (eds) *Proceedings of the 7th Annual Symposium on Combinatorial Pattern Matching*, Number 1075. Springer, Berlin, pp. 75–86.

von Bubnoff,A. (2008) Next-generation sequencing: the race is on. *Cell*, **132**, 721–723.

White,J. *et al*. (2008) Figaro: a novel statistical method for vector sequence removal. *Bioinformatics*, **24**, 462–467.

Zhang,Z. *et al*. (2008) Modeling ChIP sequencing in silico with applications. *PLoS Comput. Biol.*, **4**, e1000158.