BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 22 2014, pages 3264–3265
doi:10.1093/bioinformatics/btu513

Genome analysis

Advance Access publication July 29, 2014

Trowel: a fast and accurate error correction module for Illumina sequencing reads

Eun-Cheon Lim*, Jonas Müller, Jörg Hagmann, Stefan R. Henz, Sang-Tae Kim and Detlef Weigel*

Department of Molecular Biology, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany

Associate Editor: Michael Brudno

ABSTRACT

Motivation: The ability to accurately read the order of nucleotides in DNA and RNA is fundamental for modern biology. Errors in next-generation sequencing can lead to many artifacts, from erroneous genome assemblies to mistaken inferences about RNA editing. Uneven coverage in datasets also contributes to false corrections.

Result: We introduce Trowel, a massively parallelized and highly efficient error correction module for Illumina read data. Trowel both corrects erroneous base calls and boosts base qualities based on the *k*-mer spectrum. With high-quality *k*-mers and relevant base information, Trowel achieves high accuracy for different short read sequencing applications.The latency in the data path has been significantly reduced because of efficient data access and data structures. In performance evaluations, Trowel was highly competitive with other tools regardless of coverage, genome size read length and fragment size.

Availability and implementation: Trowel is written in C++ and is provided under the General Public License v3.0 (GPLv3). It is available at http://trowel-ec.sourceforge.net.

Contact: euncheon.lim@tue.mpg.de or weigel@tue.mpg.de

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

Received on October 24, 2013; revised on July 8, 2014; accepted on July 23, 2014

1 INTRODUCTION

Reads produced by current next-generation sequencing technologies typically suffer from relatively high base error rates. To improve downstream analyses, it is desirable to correct sequencing errors directly after base calling. The most widely applied error correction methods rely on *k*-mer spectrum-based algorithms, following the spectral alignment (SA) approach (Pevzner *et al.*, 2001). A *k*-mer occurring more often than a given threshold is called *solid* (or 'trusted') and a less frequent *k*-mer is named *weak*. The goal of the SA approach is to maximize the number of *solid* calls. Quake (Kelley *et al.*, 2010), a widely used error correction module, applies a mixed model of the distributions of *solid* and *weak* calls incorporating quality values. It identifies the set of corrections that maximizes the number of *k*-mers using a maximum likelihood approach. Another *k*-mer–based method, Musket (Liu *et al.*, 2013), uses two-stage corrections with similarity to Trowel's two methods. Musket corrects bases depending on frequencies of *k*-mers and does not use base qualities. Coral (Salmela and Schröder, 2011) collects similar reads into groups and performs multiple alignments on them. It can correct indels by using the Needleman–Wunsch algorithm. Owing to the alignment complexity, this method is not favorable for large datasets. Yet, another approach is to build a suffix array or trie. For example, Hybrid SHREC (Salmela, 2010) can correct indels by detecting and replacing low-weight nodes. The weight of a node indicates the number of cohort edges in a suffix trie. Here, we introduce a new *k*-mer–based error correction module, Trowel, suitable for Illumina datasets. The key difference to other tools is that instead of relying on the uniformity of sequencing coverage, which fluctuates stochastically or inherently, Trowel trusts in sequences with continuously high-quality values. We demonstrate the accuracy and efficiency of Trowel with several datasets and compare it with other available read correction tools.

2 METHOD

Unlike other methods, Trowel solely relies on quality values to identify *solids*. Trowel selects a quality threshold, $q^{\wedge}$, such that all *k*-mer bases with a quality of $q^{\wedge}$ or higher represent at least 8% (empirical) of the entire dataset. These *solids* are called *bricks*, i.e. consecutive stretches of high-quality bases ($\geq q^{\wedge}$) and are stored in the *brick* index as keys. We sequentially make use of two *brick* indices with different *k*-mer compositions. To each key, the first index associates high-quality bases enclosed by two *bricks*. The *Double Bricks & Gap* (*DBG*) algorithm exploits an asymmetric k_1-gap-k_2 structure, where gap is a single base, $k = k_1 + k_2$. This *k*-mer structure has advantages over symmetrical or single *k*-mer patterns at repeat element boundaries (Supplementary Fig. S7). The quality of a gap is boosted to the maximum quality value when the index relevant to gap-enclosing *bricks* contains the gap with high quality. The gap is corrected to another base when it is uniquely associated with a different high-quality base (Supplementary Fig. S6). Initially, the high-quality regions in the raw reads are usually fragmented (Supplementary Fig. S4), and hence, only a limited number of trusted *k*-mers can be used. Owing to the quality boosting (Supplementary Fig. S5), the *brick* index is iteratively expanded, leading to better sensitivity. Finally, because bases at read ends cannot be accessed by the described *brick* index, the second algorithm, *Single Brick & Edges* (*SBE*), uses a new edge-*k*-edge index to correct edges, where an edge is a single base, or increase their quality values as in the *DBG* algorithm. Full details are explained in Supplementary Data.

*To whom correspondence should be addressed.

Table 1. Sum-of-rank of performance metrics

Tool	Read accuracy	Base accuracy	Transcriptome data accuracy	Genome assembly (overall)	Assembly (high coverage)	Assembly (low coverage)	Runtime	Memory usage
Uncorrected	–	–	–	6	6	4	–	–
Trowel	**1**	**1**	2	**1**	**1**	2	**1**	2
Coral	2	4	**1**	2	4	**1**	3	6
Musket	3	5	3	4	4	2	2	**1**
SOAPec	4	3	4	4	3	4	4	4
Quake	5	2	5	2	2	4	5	3
Assembler	–	–	–	–	–	–	–	5

Note. The best in each column is highlighted in bold (see Supplementary Table S1. for datasets information).

Table 2. Runtime of error correction modules (min)

Data	Quake	Coral	Musket	SOAPec	Trowel
D1	261.1	43.9	4.1	130.5	**2.9**
D2	377.9	228.9	18.0	194.1	**9.9**
D3	119.8	36.6	8.8	91.8	**3.7**
D4	696.3	255.2	21.9	335.6	**14.1**
D5	262.2	156.0	26.0	186.4	**6.4**
D6_1	490.6	256.1	33.0	275.7	**18.4**
D6_2	243.3	105.5	15.7	137.5	**8.4**
D6_3	460.2	72.8	10.6	141.0	**4.2**
D6_4	359.9	50.0	8.0	96.9	**3.1**
D7	1088.2	886.7	145.4	595.6	**40.0**

Note. The fastest in each row is highlighted in bold (see Supplementary Table S1. for datasets information).

3 RESULTS

We performed our evaluations on paired-end *Arabidopsis thaliana* reads (86×) generated in-house and on Illumina datasets referenced in (Yang *et al*., 2012) for *Escherichia coli* (163–618×), *Staphylococcus aureus* (691×), *Saccharomyces cerevisiae* (319×) and *Drosophila melanogaster* (6–26×) (Supplementary Table S1). We assessed runtime and accuracy in several applications, and summarized all local ranks of each metric in a sum-of-rank table (Table 1). Trowel was scored consistently as one of the top two tools and was among the two most accurate tools for high-coverage datasets (Supplementary Section S4.1). Trowel has a better performance than the other tools with genome assemblies on high-coverage datasets, while on low-coverage datasets, the alignment-based tool Coral outperformed all other *k*-mer–based methods including Trowel. When the coverage is highly variable, i.e. for transcriptomes, the dataset is particularly hard to correct because low-coverage sequences cannot be equated with higher likelihood of errors. In transcriptome mapping evaluation for a human dataset (Yang *et al*., 2011), SEECER (Le *at al*., 2013) achieved the best performance (Supplementary Section S2.3). Aside from this specialized tool, Trowel obtained the best read accuracy, but was slightly worse in base accuracy than Coral. Only Trowel improved transcriptome quantification; for the other tools including SEECER, the performance was worse than the uncorrected cases. For SNP calling evaluations, Trowel showed the highest concordance with high-coverage datasets in three of four cases (Supplementary Section S2.4). Finally, Trowel was the only tool that consistently generated expected results with simulated datasets for an erroneous-base-next-to-repeats problem (Supplementary Section S2.5). Concerning runtime, Trowel outperformed all other tools for all datasets, being up to 100 times faster (Table 2). Detailed results are provided in Supplementary Data.

4 CONCLUSION AND DISCUSSION

We have developed a new error correction module for Illumina short reads, Trowel, which draws its power to correct bases solely from high-base qualities rather than coverage estimates. High-base qualities are a requirement for datasets in practice, whereas equally distributed coverage cannot always be expected depending on sequencing performance or quantitative datasets, e.g. transcriptome studies. We assessed different error correction tools on numerous commonly used applications for read mapping, resequencing, genome assembly and gene expression analyses. Trowel consistently obtained best or second-best accuracy on all applications, while achieving on average 25 times faster runtimes than the other tools. Only on genome assemblies from low-coverage datasets (≤10×), an alignment-based method was superior. In practice, the low-coverage problem is inevitable, and further improvements for *k*-mer–based methods are possible.

ACKNOWLEDGEMENTS

The authors thank Christa Lanz for sequencing and Xi Wang for making the *A. thaliana* dataset available.

Funding: This work was supported by the Max Planck Society.

Conflict of interest: none declared.

REFERENCES

Kelley,D.R. *et al.* (2010) Quake: quality-aware detection and correction of sequencing errors. *Genome Biol*., **11**, R116.

Le,H.S. *et al.* (2013) Probabilistic error correction for RNA sequencing. *Nucleic Acids Res*., **41**, e109.

Liu,Y. *et al.* (2013) Musket: a multistage k-mer spectrum based error corrector for Illumina sequence data. *Bioinformatics*, **29**, 308–315.

Pevzner,P.A. *et al.* (2001) An Eulerian path approach to DNA fragment assembly. *Proc. Natl Acad. Sci. USA*, **98**, 9748–9753.

Salmela,L. (2010) Correction of sequencing errors in a mixed set of reads. *Bioinformatics*, **26**, 1284–1290.

Salmela,L. and Schröder,J. (2011) Correcting errors in short reads by multiple alignments. *Bioinformatics*, **27**, 1455–1461.

Yang,L. *et al.* (2011) Genomewide characterization of non-polyadenylated RNAs. *Genome Biol*., **12**, R16.

Yang,X. *et al.* (2012) A survey of error-correction methods for next-generation sequencing. *Brief. Bioinform*, **14**, 56–66.