BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 6 2010, pages 849–850
doi:10.1093/bioinformatics/btq045

Sequence analysis

Filtering error from SOLiD Output

Ariella Sasson[1,2] and Todd P. Michael[1,2,3,*]

[1]Waksman Institute of Microbiology, [2]BioMaPS Institute for Quantitative Biology and [3]Department of Plant Biology and Pathology, Rutgers, The State University of New Jersey, Piscataway, NJ 08554, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: Here, we report the development of a filtering framework designed for efficient identification of both polyclonal and independent errors within SOLiD sequence data. The filtering utilizes the quality values reported by SOLiD's primary analysis for the identification of the two different types of errors. The filtering framework facilitates the passage of high-quality data into a variety of functional genomics applications, including *de novo* assemblers and sequence matching programs for SNP calling, improving the output quality and reducing resources necessary for analysis.

Availability: This error analysis framework is written in Perl and runs on Mac OS and Linux/Unix systems. The filter, documentation and sample Excel files for quality analysis are available at http://hts.rutgers.edu/filter and are distributed as Open Source software under the GPLv3.0.

Contact: tmichael@waksman.rutgers.edu

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

Received on November 12, 2009; revised on January 31, 2010; accepted on February 1, 2010

High-throughput sequencing (HTS) technologies are revolutionizing biological research by their ability to generate millions of short read sequences in a relatively short time frame. Not only do these platforms expand our knowledge by their ability to sequence genomes, they also expand studies in the related fields of transcriptome and proteome research (Ondov and Varadarajan *et al.*, 2008; Pettersson and Lundeberg *et al.*, 2009). One of the most recent HTS platforms comes from Applied Biosystems (ABI, now Life Technologies, Foster City, CA, USA), the Sequence by Oligonucleotide Ligation and Detection (SOLiD) HTS platform (http://solid.appliedbiosystems.com). This system uses a ligation-mediated sequencing strategy that is less prone to some of the problems associated with the polymerase-based sequencing platforms. Unlike other platforms, SOLiD data are collected using colorspace, a representation of two-base encoding, thus requiring a conversion step to attain DNA sequence (Supplement of McKernan and Peckham *et al.*, 2009). This two base encoding requires dual interrogation of each base during SOLiD's sequencing process aiding in distinguishing sequencing errors from true polymorphisms (McKernan and Peckham *et al.*, 2009; Smith and Quinlan *et al.*, 2008; Valouev and Ichikawa *et al.*, 2008). The SOLiD platform outputs two types of files after primary analysis: a sequence file in colorspace and a quality file containing the corresponding quality values (QVs). The QVs are calculated by training the sequencing process parameters (image intensity, a noise to signal value and angle) against several annotated datasets. The QVs exhibit a linear relationship between observed and predicted phred-scale quality scores (Hyland and Wessel *et al.*, 2009). In essence, the QVs represent the probability of the color call being inaccurate, i.e. the higher the QV the higher the confidence in the color call's accuracy.

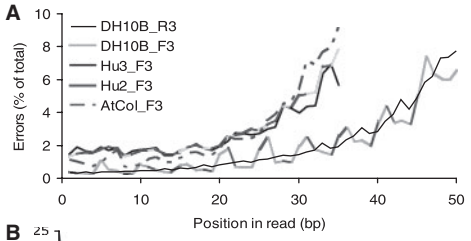
In contrast to other HTS platforms, the SOLiD platform does not prefilter low-quality reads. Therefore, all reads where the two-base transition can be identified, even if with poor quality, are reported after primary analysis. This becomes a crucial problem when dealing with *de novo* assembly of short read sequences since their assembly is highly sensitive to sequencing errors. Therefore, it becomes critical to mitigate these errors prior to assembly. There are two types of sequencing errors commonly observed, polyclonal/correlated errors and independent, erroneous color calls (Valouev and Ichikawa *et al.*, 2008). Polyclonal and correlated errors occur when the entire read is of poor quality or missequenced due to a bead level problem such as in a polyclonal bead or poor resolution of a particular bead. A polyclonal bead occurs when two different templates are amplified on a single bead and then sequenced, resulting in a hybrid sequence that has no match in the true genome. While the original goal was to identify polyclonal beads, there is no guarantee that all the reads identified by this part of the filter are due to polyclonality. A more robust filtering system using the information gathered during the sequencing run, image intensity, noise to signal and angle, could be designed to distinguish between polyclonal beads versus other types of correlated errors. Single color call errors are independent and can occur multiple times in the sequence also leading to an inaccurate sequence.

Here, we present a filtering framework that attempts to optimize the preprocessing step by identification and removal of error prone reads for the SOLiD platform using the QVs provided from the SOLiD's primary analysis. This algorithm flexibly targets the two different types of errors that can occur during SOLiD sequencing.

The goal of the preprocessor is to eliminate the low-quality reads and pass only the high-quality data into downstream applications, saving both resources and improving final output quality.

To identify sequencing reads with either polyclonal calls or miscalls, we utilized several resequencing datasets where few mismatches were expected between the reference sequence and the sequencing reads. Using these datasets, profiles for both polyclonal/correlated errors and erroneous color calls were determined through various QV analyses. Using the SOLiD mapping pipeline, the reads were matched to the reference sequence, and similar to other HTS platforms, the number of errors increased

*To whom correspondence should be addressed.



C

C. elegans resequencing

mm	mapped	All reads	%	mapped	Passed filter	%	mapped	Falied filter	%
0	13,466,335	40,038,192	33.63	4,635,350	5,375,338	86.23	3,709,972	24,445,204	15.18
1	4,972,849	40,038,192	12.42	484,983	5,375,338	9.02	2,817,988	24,445,204	11.53
2	4,094,724	40,038,192	10.23	85,517	5,375,338	1.59	2,950,904	24,445,204	12.07

Fig. 1. The relationship between error and location in SOLiD HTS reads. (**A**) Location of errors in the SOLiD reads. The errors increase on the 3′-end of the read, while the 5′-end of the read remains relatively error free. (**B**) The QVs of the identified errors from the SOLiD matching pipeline. QVs lower than 10 overwhelmingly correspond to detected errors based on their identification by the matching pipeline. (**C**) Results for *Caenorhabditis elegans* matching before and after filtering (filter settings p = 3, p_QV = 22, e = 3, e_QV = 10). This table shows the results of matching divided by the number of mismatches for the F3 read of the mate pair. DH10B_R3 (*Escherichia coli*, reverse mate), DH10B_F3 (forward read), Hu3 (Human 3), Hu2 and AtCol_F3 (*Arabidopsis thaliana*, Columbia—color version, Supplementary Fig. 2).

toward the 3′-end (Chaisson and Brinza *et al.*, 2009). In spite of the genome being sequenced (*Escherichia coli, DH10B*; Arabidopsis, AtCol; or human, Hu3), fewer errors at the 5′-end of the read were identified (Fig. 1A). The lower error rate between DH10B/AtCol and Hu3 is due to a difference in sequencing chemistry (SOLiD v2 vs. v3). From the graph plotting error as a function of QV, it was very clear that color calls with QVs of 10 or less had a higher probability of being erroneous (Fig. 1B). Single color call errors occur randomly throughout the sequence; polyclonal beads seem to reflect subpar color calls all throughout the read, i.e. lower than expected QV values at the 5′-end of the sequenced read. Analysis of the QVs show that early color calls can be highly predictive for the remainder of the read (Supplementary Fig. 1). Therefore, polyclonal analysis focuses on the quality of the first 10 color calls, requiring that some portion of them be of high quality (QV $\geq$ 25).

While filter defaults exist (min polyclonal counts: p = 1, polyclonal min QV: p_QV = 25, max errors identified: e = 3, max QV to identify independent errors: e_QV = 10), the settings of the parameters should depend upon the error tolerance of the downstream applications, such as mapping, *de novo* assembly or transcriptome analysis. The user has the ability to define both the counts and the QV which determine the removal of a read from the dataset. The more conservative the parameters, the smaller the resulting dataset which successfully passes the filtering criteria (Supplementary Tables 2–4).

When the filter is applied to a *Caenorhabditis elegans* resequencing dataset using the stringent settings (p = 3, p_QV = 22, e = 3, e_QV = 10), the raw reads were reduced from 40 M reads to 5 M. Mapping of these reads increased from 56% to 96%. Of the reads that failed the filter, 38% still mapped with 0–2 mm. However, for both the unfiltered and the failed reads, many reads matched with 1 or 2 mm, while the filtered reads had the highest percentage of reads mapping with 0 mm. These results demonstrate that the filter can effectively identify perfect reads, which would be necessary for applications like *de novo* sequence assembly. While reducing the errors within the dataset is highly critical for a quality assembly, most assemblers contain error identification protocols and will attempt removal even if absent. In addition, extreme reduction of coverage could potentially be more harmful than the presence of few errors. In a practical *de novo* assembly project, we found that the settings should be much more relaxed (A.Dayarian *et al.* (2010) SOPRA: an algorithm for high quality de novo assembly of paired reads via statistical optimization. Available at www.physics.rutgers.edu/ anirvans/SOPRA/).

This error analysis framework was designed to overcome memory constraints imposed by uploading both the read and the quality files into memory. The program sacrifices runtime to have a minimal memory footprint. Additional script details, including user-defined input table, figures, tables, functionality and SOLiD data run details are described in the Supplementary Material section.

ACKNOWLEDGEMENTS

We would like to thank Anirvan Sengupta, Randall Kerstetter, Adel Dayarian and the bioinformaticians at Applied Biosystems for discussions around this work.

Funding: Department of Energy's Computational Science Graduate Fellowship (DE-FG02-97ER25308 to A.S.); Charles and Johanna Busch Memorial Fund at Rutgers, The State University of New Jersey to T.P.M.

Conflict of Interest: none declared.

REFERENCES

Chaisson,M.J.P. *et al.* (2009) De novo fragment assembly with short mate-paired reads: Does read length matter? *Genome Res.*, **19**, 336–346.

Hyland,F.C.L. *et al.* (2009) Dibase sequencing allows accurate SNP detection at moderate and low coverage with diBayes algorithm. Available at http://www3.appliedbiosystems.com/cms/groups/mcb_marketing/documents/generaldocuments/cms_065554.pdf.

McKernan,K.J. *et al.* (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.*, **19**, 1527–1541.

Ondov,B.D. *et al.* (2008) Efficient mapping of Applied Biosystems SOLiD sequence data to a reference genome for functional genomic applications. *Bioinformatics*, **24**, 2776–2777.

Pettersson,E. *et al.* (2009) Generations of sequencing technologies. *Genomics*, **93**, 105–111.

Smith,D.R. *et al.* (2008) Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Res.*, **18**, 1638–1642.

Valouev,A. *et al.* (2008) A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Res.*, **18**, 1051–1063.