2015

HAsh-MaP-ERadicator: Filtering Non-Target Sequences from Next Generation Sequencing Reads

Jonathon Brenner
Loyola University Chicago, jbrenner@luc.edu

Catherine Putonti
Loyola University Chicago, cputonti@luc.edu

Recommended Citation

Brenner, Jonathon and Putonti, Catherine. HAsh-MaP-ERadicator: Filtering Non-Target Sequences from Next Generation Sequencing Reads. Proceedings of the 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), , : 1100-1101, 2015. Retrieved from Loyola eCommons, Bioinformatics Faculty Publications, http://dx.doi.org/10.1109/BIBM.2015.7359835

HAsh-MaP-ERadicator:

Filtering Non-Target Sequences from Next Generation Sequencing Reads

Jonathon Brenner
Department of Computer Science
Bioinformatics Program
Loyola University Chicago
Chicago, IL 60660
Email: jbrenner@luc.edu

Catherine Putonti
Departments of Biology and Computer Science
Bioinformatics Program
Loyola University Chicago
Chicago, IL 60660
Email: cputonti@luc.edu

Abstract*—Contemporary DNA sequencing technologies are continuously increasing throughput at ever decreasing costs. Moreover, due to recent advances in sequencing technology new platforms are emerging. As such computational challenges persist. The average read length possible has taken a giant leap forward with the PacBio and Nanopore solutions. Regardless of the platform used, impurities within the DNA preparation of the sample – be it from unintentional contaminants or pervasive symbiots – remains an issue. We have developed a new tool, HAsh-MaP-ERadicator (HAMPER), for the detection and removal of non-target, contaminating DNA sequences. Integrating hash-based and mapping-based strategies, HAMPER is both memory and time efficient while maintaining a high level of sensitivity. Moreover, HAMPER was designed for flexibility: reads of any size can be efficiently examined and the user can set parameters specific for the analysis of reads produced by a particular sequencer. To evaluate our method, mock sequencing runs were generated including various contaminating species and with variable rates of mutation revealing a high level of sensitivity and specificity. Reads that are not of interest can quickly be removed using HAMPER thus improving downstream analyses.

Keywords*—*sequence contamination, sequence homology, hash-mapping

I. Introduction

The past decade's explosion of genomic data continues to increase as advances in sequencing technology lead to improved speed and throughput at lower cost [1-2]. Emerging technologies promise to push beyond current limitations of read length, albeit at a cost; longer-read technologies presently have been found to have higher error rates [3]. Regardless of the technology used, contamination remains an issue. While foreign DNAs can be introduced during sample preparation, the simple fact of the matter is that many species (ourselves included) have numerous symbionts [4-7]. Contaminants can have confounding affects leading to erroneous conclusions [4].

Three computational strategies have been explored for identifying and removing foreign non-target DNA sequences from raw datasets. The first approach utilizes a hash-table based approach in which the sequencing reads are broken up into words and scanned against a pre-hashed library of the intended target's sequence; this is often executed using NCBI's BLAST [8]. A second relies on a suffix/prefix trie representation of the target sequence for which all reads are compared. The third strategy is founded on short-read mapping approaches, e.g. the Burrows-Wheeler Transformation followed by Smith-Waterman alignment (BWA-SW) [9]. All three techniques were evaluated by Schmieder *et al.* [7], finding the BWA-SW (employed by the authors in their tool DeconSeq) to be the fastest.

Herein we present a blended-approach, integrating the expediency of hash tables with the sensitivity of short-read mapping approaches, called HAsh-MaP-ERadicator or HAMPER for short. The approach is computationally lightweight, yet robust, and the hash-based utility is capable of evaluating large datasets for contamination rapidly. This tool is freely available at: http://www.putonti-lab.com/software.html.

II. Implementation

At the heart of HAMPER is a hash table, *S*. The target sequence(s) are parsed for a given word size *k*. The hash function uniquely maps *k* to a single index within *S*; thus the size of *S* is 4^k. Each element in *S* consists of a Boolean value (indicating presence/absence within the target sequence(s)) and a vector of pointers to an array *U* which contains the words of length *k* as they appear within the target sequence(s). Selection of word size *k* is adjustable by the user with the most optimal value dependent on size of the target genome. Thus, a balance between sensitivity and density of the data structures (memory usage) can be achieved.

Comparison of each read against the target sequence(s) is then facilitated utilizing a hash-mapped approach. Each substring *w* of length *k* from the read sequence is transformed by the same hash function. By checking for the occurrence in the target genome of each unique *k*-mer in *S*, a seeded match of any given *k*-mer is able at this step to be determined in constant time complexity. Following this seed via *S*'s link(s) to *U* allows for an extension of this match in linear time by mere comparison of the next indexed *k*-mer in each genome. Upon reaching a user defined tolerance threshold of sequence divergence, the matching hit is then reported. All overlapping *w* in the read sequence are considered for both strands of the target sequence(s). In the event that the target sequence(s) does not contain *w* such that $S[f(w)]$=NULL, *w* one, two, etc.

J.B. was funded by a Loyola University Chicago Carbon Research Fellowship. C.P. is funded by the NSF (1149387).

mismatches away can be considered by chaining *S*. Greater sensitivity in the search can be achieved by either the chained hash table approach or a smaller value *k*.

III. Proof-of-Concept

To test the efficiency of HAMPER, several simulated high throughput sequencing samples were created. Five genomes were used to generate these samples: the human genome hs_alt_CHM1_1.1, *Escherichia coli* KO11 (GenBank: NC_016902), *Mycobacterium tuberculosis* CTRI-2 (GenBank: NC_017524), *Streptococcus suis* T15 (GenBank: NC_022665), and *Lactococcus lactis* subsp. lactis IO-1 (GenBank: NC_020450). These four bacterial genomes were selected given their variation in nucleotide composition (*E. coli* GC=50.8%, *M. tuberculosis* GC=66.6%, *S. suis* GC=41.0%, *L. lactis* GC=35.1%). All genomes were retrieved from NCBI via the FTP site: ftp://ftp.ncbi.nlm.nih.gov/genomes/.

To parallel high throughput sequencing technologies, a dataset of 250Mbp containing reads from all five samples and of varying lengths (150, 400, 500, 1000, and 2500) was created. HAMPER had 0% false positives (i.e. none of the reads from the microbial genomes were identified as human) and 100% true positives (i.e. all of the reads from the human genome were identified as human). This analysis was carried out in less than 5 minutes on an Intel® Core^{TM} i7 2.20GHz processor with 8GB of available RAM. This first dataset exemplifies the specificity possible given high sequence homology to the reference sequence. To ascertain the sensitivity of HAMPER, six additional datasets were generated, each containing only sequences from the human genome (again of varying lengths). Each dataset was mutated using Geneious (Biomatters Ltd., Auckland, New Zealand) introducing: 1, 2, or 5 point mutations or indels per read. Furthermore, the effects of two thresholds used in guiding the search were explored: query coverage and sequence identity. The results shown in Fig. 1 indicate that HAMPER is significantly better equipped to detect contamination when point mutations rather than indels are present. Nevertheless, fine-tuning the parameters increases the tools sensitivity with no effect on run-time (< 20sec. per Mbp).

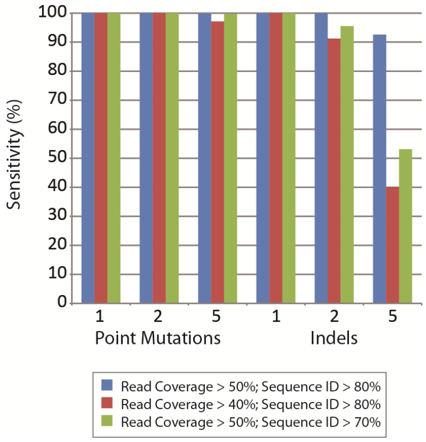

Fig. 1. Sensitivity in detection of reads from contaminating DNAs given mutational variation between read sequence and reference genome. Different thresholds for read coverage and sequence ID were explored.

Memory usage is a function of the size of the target sequence(s) *M* and the *k* selected; the former dictates the size of the array *U* while the latter determines the size of the hash table *S*. Thus the memory usage required is $M \times 4^k$. Run-time is dependent upon the number of reads *n* and the length of the reads *l*. In the worst case scenario in which all reads map to the target sequence(s), the run-time will be $O(n \times l)$. However, as exemplified by the analysis of the mock communities presented here, this is negligible. Expansion of *S* to include chains one-, two-, etc. mismatches away will increase the number of comparisons performed.

IV. Conclusions and Future Directions

The approach presented here provides an efficient and reliable means for identifying reads from contaminating DNAs. This tool was specifically designed with long reads in mind, in particular those possible from the minION in which SNPs far outnumber indels [3] and has previously been shown to produce unidentifiable reads of unknown origin [10]. Future directions include modifications to the underlying data structures to more readily handle indels. Furthermore, HAMPER can be used both pre- and post-assembly to assist in resolving contigs.

Acknowledgment

The authors would like Dr. George Thiruvathukal for his assistance in early data structure and algorithm design.

References

[1] P. Flicek, and E. Birney, “Sense from sequence reads: methods for alignment and assembly,” Nature Methods, vol. 6, pp. S6-S12, 2009.

[2] M.L. Metzker, “Sequencing technologies – the next generation,” Nat. Rev. Genet., vol. 11, pp. 31-46, 2010.

[3] T. Laver, J. Harrison, P.A. O’Neill, K. Moore, A. Farbos, K. Paszkiewicz, and D.J. Studholme, “Assessing the performance of the Oxford nanopore Technologies MinION,” Biomolecular Detection and Quantification, vol. 3, pp. 1-8, 2015.

[4] R.W. Lusk, “Diverse and widespread contamination evident in the unmapped depths of high throughput sequencing data,” PLoS One, vol. 9, e110808, 2014.

[5] K. Gruber, “Here, there, and everywhere: From PCRs to next-generation sequencing technologies and sequence databases, DNA contaminants creep in from the most unlikely places,” EMBO Rep., vol. 16, pp. 898-901, 2015.

[6] S. Mukherjee, M. Huntemann, N. Ivanova, N.C. Kyrpides, and A. Pati, “Large-scale contamination of microbial isolate genomes by Illumina PhiX control,” Stand. Genomic Sci., vol. 10, 18, 2015.

[7] R. Schmieder, and R. Edwards, “Fast identification and removal of sequence contamination from genomic and metagenomic datasets,” PLoS One, vol. 6, e17288, 2011.

[8] S.F. Altschul, W. Gish, E.W. Myers, and D.J. Lipman, “Basic local alignment search tool,” J. Mol. Biol., vol. 215, pp. 403-410, 1990.

[9] H. Li, and R. Durbin, “Fast and accurate long-read alignment with Burrows-Wheeler transformation,” Bioinformatics, vol. 26, pp. 589-595, 2010.

[10] A. Kilianski, J.L. Haas, E.J. Corriveau, A.T. Liem, K.L. Willis, D.R. Kadavy, C.N. Rosenzweig, and S.S. Minot, “Bacterial and viral identification and differentiation by amplification sequencing on the MinION nanopore sequencer, Gigascience, vol. 4, 12, 2015.