Computational Biology and Chemistry 65 (2016) 178–184

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research article

SnpFilt: A pipeline for reference-free assembly-based identification of SNPs in bacterial genomes

Carmen H.S. Chan[a], Sophie Octavia[a], Vitali Sintchenko[b,c], Ruiting Lan[a,*]

[a] School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, New South Wales, 2052, Australia
[b] Centre for Infectious Diseases and Microbiology–Public Health, Institute of Clinical Pathology and Medical Research, Westmead Hospital, New South Wales, Australia
[c] Marie Bashir Institute for Infectious Diseases and Biosecurity, The University of Sydney, New South Wales, Australia

ARTICLE INFO

Article history:
Received 31 August 2016
Accepted 7 September 2016
Available online 9 September 2016

Keywords:
Next generation sequencing
Genome assembly
Single nucleotide polymorphisms
Reference free SNP discovery

ABSTRACT

De novo assembly of bacterial genomes from next-generation sequencing (NGS) data allows a reference-free discovery of single nucleotide polymorphisms (SNP). However, substantial rates of errors in genomes assembled by this approach remain a major barrier for the reference-free analysis of genome variations in medically important bacteria. The aim of this report was to improve the quality of SNP identification in bacterial genomes without closely related references. We developed a bioinformatics pipeline (SnpFilt) that constructs an assembly using SPAdes and then removes unreliable regions based on the quality and coverage of re-aligned reads at neighbouring regions. The performance of the pipeline was compared against reference-based SNP calling for Illumina HiSeq, MiSeq and NextSeq reads from a range of bacterial pathogens including *Salmonella*, which is one of the most common causes of food-borne disease. The SnpFilt pipeline removed all false SNP in all test NGS datasets consisting of paired-end Illumina reads. We also showed that for reliable and complete SNP calls, at least 40-fold coverage is required. Analysis of bacterial isolates associated with epidemiologically confirmed outbreaks using the SnpFilt pipeline produced results consistent with previously published findings. The SnpFilt pipeline improves the quality of de-novo assembly and precision of SNP calling in bacterial genomes by removal of regions of the assembly that may potentially contain assembly errors. SnpFilt is available from https://github.com/LanLab/SnpFilt.

1. Introduction

The development of next-generation sequencing (NGS) technologies has made rapid and high-throughput whole-genome sequencing of microbial genomes broadly accessible to clinical microbiology laboratories and has transformed the public health microbiology for epidemiological typing and outbreak investigation (Dallman et al., 2015; den Bakker et al., 2014; Kohl et al., 2014). NGS has been employed for the detection and control of epidemiological outbreaks in clinically actionable time (Tang and Gardy, 2014). It provides a universal solution for high-resolution typing of pathogens and sequence data are amenable to shared analysis between laboratories (Mardis, 2008).

However, the analysis of NGS genome data is often hampered by the error-prone nature of these datasets (Chain et al., 2009; Tang and Gardy, 2014). The standard approach for identifying SNPs is to map short reads to a closely related reference assembly (Nielsen et al., 2011). The accuracy of this method depends strongly on the availability of the reference; there is a bias towards calling the reference base, particularly in regions where SNPs are densely located (Pightling et al., 2014). A commonly used reference-free approach, based on the analysis of the structure of de-Brujin graphs without explicitly constructing an assembly (Gardner and Hall, 2013; Leggett and MacLean, 2014; Uricaru et al., 2015), has similar limitations and is known to have problems when multiple, closely located SNPs occur. An alternative approach for reference-free identification of SNPs is to use de-novo assemblies to identify SNPs, but the assembly of contigs using short reads is a complex process that can introduce additional errors (Alkan et al., 2011; Kelley and Salzberg, 2010; Phillippy et al., 2008; Ricker et al., 2012).

* Corresponding author.
E-mail address: r.lan@unsw.edu.au (R. Lan).

http://dx.doi.org/10.1016/j.compbiolchem.2016.09.004
1476-9271/

Previous studies (Li, 2014; Liu et al., 2012; O'Rawe et al., 2013; Reumers et al., 2012) have demonstrated the necessity of applying filters on SNP discoveries by standard variant calling programs using quality and coverage data but their applicability for SNP-calling using de-novo assemblies is unclear. Alternatively, existing tools that were developed for evaluating assembly and contig quality can be employed. One of the earliest tools developed, such as amosvalidate (Phillippy et al., 2008; Schatz et al., 2013) allows the user to visually examine regions of the contig, but is not easily automated and is incompatible with the output format of the more recent assemblers. These tools also generally do not evaluate quality per site for distinguishing true SNPs from assembly errors. More recent software such as REAPR (Hunt et al., 2013) and ALE (Clark et al., 2013) output a score describing the quality at each contig site but the suitability of applying these tools for SNP-calls remains untested.

Here, in this study, we present a pipeline for identifying SNPs in microbial genomes using de-novo assembly, which we named SnpFilt. We show that the choice of assembly method can heavily influence accuracy of the SNPs called. For any choice of assembly method, assembly errors can introduce a large number of false SNPs. Consequently, in our pipeline, we introduce a reference-independent filtering step to remove ambiguous regions of the contigs before SNPs are called. That is, our pipeline can be applied to any read dataset to obtain a set of high-confidence contigs; in our analysis we re-align these contigs to known reference genomes, but this is only for the purpose of obtaining for positional information and to verify their accuracy.

To demonstrate the applicability of SnpFilt for public health laboratory surveillance and outbreak detection, we applied the pipeline to isolates of *Salmonella enterica* serovar Typhimurium, one of the most common and widely distributed food-borne bacterial pathogens (Flint et al., 2005). We analysed isolates from strain SARA2 a derivative of lab strain LT2, which we sequenced using two platforms, as well as a set of more distantly related *S.* Typhimurium isolates that has previously been epidemiologically characterised (Octavia et al., 2015). We also applied SnpFilt on three sets of paired-end Illumina reads with fully sequenced genomes and GC content ranging from 40 to 65%, to evaluate its performance on a broader range of bacterial genomes.

2. Methods

2.1. SnpFilt pipeline

SnpFilt is a SNP detecting pipeline that employs a de-novo assembly program to assemble the contigs and SNPs are detected by re-mapping the reads to the assemblies and applying a set of filters to identify high quality SNPs. Alignment to a reference genome is used to obtain the base positions relative to the reference but SNP detection is independent of the reference.

Contigs were constructed using SPAdes v3.1.0 (Bankevich et al., 2012) with default parameters. Sequence variations relative to a reference genome were then determined by aligning the contigs to the reference genome using LASTZ (Harris, 2007). For each contig, we used only the highest scoring set of non-overlapping (< 100 bp) alignments. We re-mapped the reads to the contigs using BWA-mem v0.7.112 (Li, 2013) with default parameters and obtained coverage and quality data for each site using mpileup in samtools v1.1 (Li, 2011). Variant sites were reported if they were associated with mutations or indels in the LASTZ alignment and were not removed due to any of the filters (F1-F6) listed below.

F1) The running mean of the read coverage, or the running mean of any site within a neighbourhood of 600 bases (i.e. 300 bases on either side), is greater than the median + 2 median absolute deviation (MAD) of read depth across the whole assembly. We used the running mean to remove noise due to isolated sites with higher coverage. It was computed by taking the mean read depth across a window of 100 bases on both sides of the focal site.

F2) Mapping quality < 58, for any site within a neighbourhood of 400 bases.

F3) The consensus base is supported by < 20 reads, or there is no supporting read in either the forward or reverse direction, for any site within a neighbourhood of 400 bases.

F4) The consensus base is supported by <10 reads in the forward direction, for any site within a neighbourhood of 20 bases.

F5) The number of reads supporting a non-consensus base >0.3 times the number of aligned reads, for any site within a neighbourhood of 20 bases.

F6) At least 50 bases within a window of 2000 bases have base quality < *q.thres,* where *q.thres* is the mean-3 standard deviations of quality scores across the whole assembly.

Filters F1–F6 have several novel aspects. Firstly, the filters remove sites based not only on the coverage and quality of the focal site, but also the local context of the surrounding bases to account for dependencies introduced during the assembly process. Secondly, cut-offs for mapping quality (F2) and base quality (F6) differ from standard values (Li, 2014; Magi et al., 2012; Octavia et al., 2015). These reflect the higher scores expected for reads mapped to their own assembly, and cut-off values were empirically chosen. Further explanation regarding the filters and cut-off values is given in the Additional file 1

2.2. Evaluation of pipeline

To evaluate the performance of the pipeline, we applied filters F1-F6 to a set of assembled contigs (Dataset 1, see description below), and we applied the full SnpFilt pipeline to several sets of paired-end Illumina reads. We included two sets of reads from an *S.* Typhimurium isolate with a closely related reference strain (Dataset 2) and three isolates, one each, from 3 different species with fully sequenced genomes (Dataset 3) and an outbreak dataset (Dataset 4).

For Datasets 1 and 2, the accuracy of the assemblies was evaluated based on their consistency with SNPs reported by mapping the reads to a closely related reference using BWA-mem and samtools. We also imposed the standard requirements that SNPs have base quality > = 20, at least 20 reads covering the SNP site and that > = 70% of the reads support the SNP (Octavia et al., 2015). We considered only SNPs in this comparison as indel-calling is known to be unreliable (O'Rawe et al., 2013). SNPs called by both reference-based mapping and SnpFilt are considered to be true positives (TP), while SNPs called by SnpFilt but not mapping are false positives (FP). SNPs called by mapping but omitted by SnpFilt are false negatives (FN), and sites identified as non-variant by both methods are considered to be true negatives (TN).

We measured the performance of the pipeline in terms of *sensitivity* (the probability that a true SNP is called; TP/(TP + FN)) and *precision* (the probability that a called SNP is true; TP/(TP + FP)) (Olson et al., 2015). Note that we use precision, which conditions on the variant call, instead of the more common metric, *specificity* (TN/(TN + FP)), which conditions on whether or not the site is truly a SNP. Because the majority of sites in the genome will be non-variants, the number of TNs is always very large, but for SNP-calling, a small number of FPs may have a disproportionately large effect on downstream analysis, particularly if the isolate is closely related to the reference. Consequently, we employ precision, which accounts better for the small number of SNPs relative to the total size of the genome (Davis and Goadrich, 2006).

For Dataset 3, isolates match the reference genome exactly so all reported SNPs are known to be false. This allows us to evaluate