Sequence analysis

Detecting SNPs and estimating allele frequencies in clonal bacterial populations by sequencing pooled DNA

Kathryn E. Holt1*, Yik Y. Teo2, Heng Li1, Satheesh Nair3, Gordon Dougan1, John Wain3 and Julian Parkhill1

1Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA,
2Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN and
3Laboratory of Gastrointestinal Pathogens, Centre for Infections, Health Protection Agency, 61 Colindale Avenue, London NW9 5HT, UK

Received on March 19, 2009; revised on May 25, 2009; accepted on May 29, 2009
Advance Access publication June 3, 2009
Associate Editor: Joaquin Dopazo

ABSTRACT

Summary: Here, we present a method for estimating the frequencies of SNP alleles present within pooled samples of DNA using high-throughput short-read sequencing. The method was tested on real data from six strains of the highly monomorphic pathogen Salmonella Paratyphi A, sequenced individually and in a pool. A variety of read mapping and quality-weighting procedures were tested to determine the optimal parameters, which afforded ≥80% sensitivity of SNP detection and strong correlation with true SNP frequency at poolwide read depth of 40x, declining only slightly at read depths 20–40x.

Availability: The method was implemented in Perl and relies on the opensource software Maq for read mapping and SNP calling. The Perl script is freely available from ftp://ftp.sanger.ac.uk/pub/pathogens/pools/.

Contact: kh2@sanger.ac.uk

Supplementary information: Supplementary data are available at Bioinformatics online.

The discovery of assayable genetic variation is essential in order to study the population structure of bacteria, which is crucial for addressing important research questions including evolution, transmission and epidemiology of bacterial pathogens and associated disease (Keim et al., 2004; Nübel et al., 2008). Multilocus sequence typing (MLST) has been widely adopted for the study of bacterial population structure (Maiden, 2006), however this technique is not sensitive enough to detect variation among highly monomorphic bacterial species such as Salmonella Typhi (Kidgell et al., 2004). Furthermore, while many bacterial pathogens display a much more diverse population structure, disease outbreaks are often associated with the spread of a single clone associated with a single sequence type—e.g. methicillin resistant Staphylococcus aureus MRSA-15 and MRSA-16 (Enright et al., 2000). Here, we propose the high-throughput sequencing of pools containing equal amounts of genomic DNA extracted from multiple bacterial isolates. This approach facilitates genome-wide SNP detection among closely related isolates and is cheaper than sequencing isolates individually (even using multiplex libraries), allowing extensive sampling of a population at low cost. This is crucial for unbiased detection of genetic variation within a population, which in turn is required to give an unbiased picture of the underlying population structure (Pearson et al., 2004). Two likely applications of SNP detection in pooled DNA samples from bacteria are: (i) identifying SNPs for typing in a much larger population (Pearson et al., 2004), and (ii) identification of SNPs associated with a particular phenotype, e.g. increased virulence (Falush and Bowden, 2006). In either application it will be important to estimate the allele frequencies of each SNP within each pool.

We used the Illumina GAI to sequence pools of DNA, and Maq (http://maq.sourceforge.net/) to map short reads to a reference sequence and make initial SNP calls. Maq is ideally suited to this task as it employs a user-specified number of haplotypes (in this case the number of strains in the pool) to detect the presence of SNPs (Li et al., 2008). The frequency of each SNP k in pool p containing Sp strains was estimated using information (read from Maq’s pileup output) on each read i of N reads mapped to the SNP locus, including the phred-like base quality wk,i, which was used to calculate weights wk,p,i (see below). Frequencies were calculated according to the following formulae, implemented in a Perl script which calls Maq to do the initial read mapping and SNP calling (here xk,p,i = 1 if SNP allele, 0 otherwise):

\[
propk,p = \frac{\sum_{i=1}^{N} wk,i \times xk,i}{\sum_{i=1}^{N} wk,i} (1) \quad freqk,p = propk,p \times Sp (2)
\]

To optimize and validate the method for allele frequency estimation, we chose six strains of the monomorphic bacterial pathogen Salmonella Paratyphi A to be sequenced individually and in pools. The strains include ATCC9150 and five novel isolates. Illumina sequencing was used to generate 35 bp reads at 20–40x depth (European Read Archive: ERA000083). Maq was used to map reads to the reference genome AKU_12601 (EMBL:FM200053) and call SNPs in each isolate: SNPs between the finished sequence ATCC9150 and the reference were also included [identifed using MUMmer (Kurtz et al., 2004)]. SNPs lying in repetitive or phage sequences (5% of genomic sequence) were excluded from the analysis, as were SNPs with low quality (quality score <20 or read depth <10). For each SNP locus identified in any strain (550 loci) alleles were checked in all six strains, resulting in a set of 403 SNPs with a reliable frequency estimate among the six strains.

To whom correspondence should be addressed.

© 2009 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

[14:20 14/7/2009 Bioinformatics-btp344.tex]
experimentally observed frequencies were strongly correlated with
was quite sensitive (78–84% of expected SNPs detected) and the
(1–7 per read) and mapping qualities (10–50), SNP detection
Fig. 1). Analysis of variance tables for each measure are given in
the rate of incorrect frequency estimates. Five additional weighting
R
frequency at
SNP detection, i.e. proportion of the SNPs detected with estimated
was reduced by using any of the weighting methods (4–6) and was
also dependent on mapping quality. The lowest rate of incorrect
mismatches (Supplementary Fig. 3). However, setting the number
of mismatches ≤
mapping (maq assemble -m option) was varied from 1 to 7 bases
(i.e. up to 20% mismatches per 35 bp read); the minimum mapping
quality for reads to be included in frequency estimation (maq pileup -q) was varied from 10 to 50. In order to avoid interpreting base
calling errors as SNPs, the contribution of each base to the estimate
of prop,y,p (Equation (1)) was weighted according to its phred-like
quality score qk,p,i. Since quality scores are calibrated for each
sequencing run, the simple and squared ratios of qk,p,i to the maximum possible quality score maxQ were considered, resulting
in four alternative weighting schemes:

\[
w_{2,1,p,i} = 1 \quad (3) \\
w_{2,1,p,i} = \frac{q_{k,p,i}}{q_{k,p,\text{max}}} \quad (5) \\
w_{2,2,p,i} = q_{k,p,i} \quad (4) \\
w_{2,2,p,i} = \frac{q_{k,p,i}}{Q_{i}} \quad (6)
\]

All combinations of parameters and weighting measures were
tested, and the following measures calculated (after removing SNP
calls in repetitive or phage sequences): (i) sensitivity of SNP
detection, i.e. proportion of the 550 known SNPs that were detected
with an estimated frequency of ≥1 strain, (ii) false positive rate of
SNP detection, i.e. proportion of the SNPs detected with estimated
frequency at ≥1 strain that were not expected to be present in
the pool, (iii) correlation (Pearson R²) between the expected and
estimated allele frequencies and (iv) the proportion (among the
403 SNPs with reliable frequency estimates) of loci for which
estimated and expected allele frequencies differed by ≥1 strain, i.e.
the rate of incorrect frequency estimates. Five additional weighting
measures were trialed, but were excluded from further analysis as
they gave highly insensitive or inaccurate results (see Supplementary
Fig. 1). Analysis of variance tables for each measure are given in
Supplementary Tables 1–4.

Using any combination of weights [Equations (3–6)], mismatches
(1–7 per read) and mapping qualities (10–50), SNP detection
was quite sensitive (78–84% of expected SNPs detected) and the
experimentally observed frequencies were strongly correlated with
the expected frequencies (Pearson R² = 0.92–0.95) (Supplementary
Fig. 2). Specifying the number of haplotypes in the pool (maq assemble -N parameter) was crucial to maintain sensitivity as
without this, sensitivity dropped to 70%. Detection sensitivity was
highly dependent on SNP frequency, with 37% detection for SNPs
present in just one strain, compared with 95% and 100% detection,
respectively, for SNPs present in 2 or ≥3 strains. The false positive
(f.p.) rate varied between 5% and 18% using different methods and
was closely correlated with number of mismatches allowed during
mapping (Supplementary Fig. 3). However, setting the number
of mismatches ≤1 reduced sensitivity too low (78%), thus the
optimal setting was ≤2 mismatches per read (mean f.p. 8.8%, mean
sensitivity 82.7%). The proportion of incorrect frequency estimates
was reduced by using any of the weighting methods (4–6) and was
also dependent on mapping quality. The lowest rate of incorrect
estimates (19%) was seen with a minimum mapping quality 40;