BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 10 2010, pages 1364–1365
doi:10.1093/bioinformatics/btq138

Genome analysis

Advance Access publication April 1, 2010

ACCUSA—accurate SNP calling on draft genomes

Sebastian Fröhler and Christoph Dieterich*

Bioinformatics in Quantitative Biology, The Berlin Institute for Medical Systems Biology at the Max Delbrück Center for Molecular Medicine, Robert-Rössle-Strasse 10, 13125 Berlin-Buch, Germany

Associate Editor: Martin Bishop

ABSTRACT

Summary: Next generation sequencing technologies facilitate genome-wide analysis of several biological processes. We are interested in whole-genome genotyping. To our knowledge, none of the existing single nucleotide polymorphism (SNP) callers consider the quality of the reference genome, which is not necessary for high-quality assemblies of well-studied model organisms. However, most genome projects will remain in draft status with little to no genome assembly improvement due to time and financial constraints. Here, we present a simple yet elegant solution ('ACCUSA') that considers both the read qualities as well as the reference genome's quality using a Bayesian framework. We demonstrate that ACCUSA is as good as the current SNP calling software in detecting true SNPs. More importantly, ACCUSA does not call spurious SNPs, which originate from a poor reference sequence.

Availability: ACCUSA is available free of charge to academic users and may be obtained from ftp://bbc.mdc-berlin.de/software. ACCUSA is programmed in JAVA 6 and runs on any platform with JAVA support.

Contact: christoph.dieterich@mdc-berlin.de

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

Received on December 23, 2009; revised on March 12, 2010; accepted on March 26, 2010

Single nucleotide polymorphisms (SNPs) are single DNA base variations within a population of individuals from the same species. In genetics, SNPs are useful markers for association studies. SNP discovery approaches usually screen a set of candidate regions by analyzing multiple nucleotide alignments. With the advent of next generation sequencing methods, even large genomes can be exhaustively mined for SNPs exploiting the high coverage. Metzker (2010) discusses next generation sequencing platforms and applications in a recent review.

Like traditional sequencing, high-throughput sequencing returns base quality values for each sequenced base. Contrary to traditional Sanger sequencing reads, high-throughput reads are much shorter (e.g. ~50–100 bp for Illumina's Solexa platform). Dohm *et al.* (2008) found that read error rates range from 0.3% at the beginning of reads to 3.8% at the end of reads. However, insertion/deletion errors are negligible in comparison to base substitution errors.

Typically, either a seed-based index or a Burrows–Wheeler transform index is used for read mapping (Metzker, 2010). SNP calling is subsequently performed on the assembled reads.

All these approaches were designed to work on high-quality reference genomes. However, most genome projects will remain in draft status with little to no manual assembly quality checking due to time and financial constraints. A high number of false SNPs may be called due to the presence of low-quality base calls in the draft reference sequence.

SNP discovery on resequencing data critically depends on three aspects: the quality of the read mapper, the quality of the assembler and the quality of the SNP caller. Each of these methods should account for sequencing errors, the SNP caller should also account for genome assembly artefacts.

Here, we present a simple yet elegant solution ('ACCUSA') that considers both the read qualities as well as the reference genome's quality for SNP calling. We use the Bayesian framework of Marth *et al.* (1999) to compute the probability of an SNP (pReseq) in a given alignment column of reads. We have outlined this framework and computational heuristics to improve runtime in the Supplementary Material.

Essentially, we compute pReseq for all aligned short reads at a given genome assembly position (pReseq) and for the complete alignment column including the reference base (pAll; see Supplementary Fig. 1). Both probabilities are subject to filtering steps: one filter imposes a lower bound on the probability of a candidate SNP ($\mathrm{pAll} \geq \mathrm{Thr}_1$) and another filter imposes an upper bound on the homogeneity of the short reads ($\mathrm{pReseq} \leq \mathrm{Thr}_2$). In an ideal haploid/isogenic dataset, all candidate SNPs are pure ($\mathrm{pReseq} \approx 0$). In other words, all short read bases are the same. Evidently, sequencing and read assembly errors generate some 'noise' in candidate SNP positions. Thr_2 removes 'noisy' columns.

Our algorithm reports a list of SNPs, which contains the position of the SNP on the genome assembly as well as additional information (pReseq, bases and quality values). The margin $\mathrm{pAll} - \mathrm{pReseq}$ provides a intuitive ranking of the SNP quality.

A wide range of software can be used to pre-filter, align and assemble resequencing reads. ACCUSA is able to operate on flat file-based assemblies in ACE file format (http://bozeman.mbt.washington.edu/consed/distributions/README.16.0.txt) as well as on stream-based input in the pileup format (http://samtools.sourceforge.net/pileup.shtml).

We will evaluate and compare our method from an experimentalist's point of view. That is why, we call a predicted SNP, which can be verified by independent Sanger sequencing, a 'true positive' (TP) SNP. A predicted SNP, which cannot be verified, is called a 'false positive' (FP) SNP. We would like to emphasize that existing SNP callers cannot distinguish between a 'true positive' SNP as defined by us and a reference sequence error. ACCUSA is unique in the sense of being the first SNP caller, which

*To whom correspondence should be addressed.

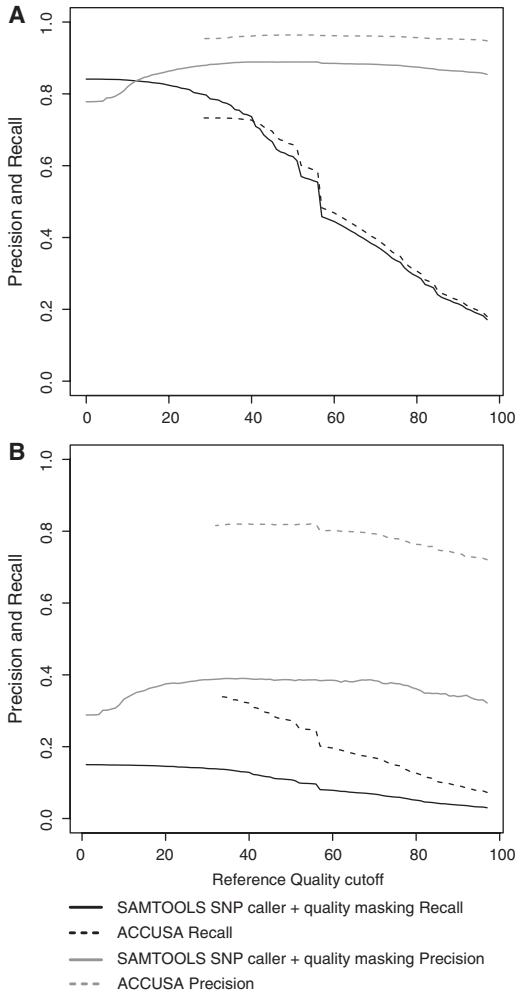

Fig. 1. Reverse cumulative precision and recall curves for ACCUSA and SAMTOOLS SNP caller on all reference genome bases with a quality higher than cutoff (*x*-axis). Two datasets are shown: yeast strain SK1 in (**A**) and yeast strain W303 in (**B**). See text and Supplementary Material.

considers the reference base quality. Any comparison to existing SNP callers will necessarily demonstrate the benefit of including the reference sequence quality in SNP calling on draft genomes. However, we make this comparison a fair competition by including a simple reference sequence quality masking step (see Supplementary Material).

We assessed our software with a large resequencing dataset of *Saccharomyces cerevisiae* strains (Liti *et al.*, 2009). Details of the assessment methods and more results can be found in the Supplementary Material. We selected two Solexa runs for *S.cerevisiae* strains W303 and SK1 and a Sanger-based draft assembly of strain S288c for our evaluation. Liti *et al.* (2009) provide a list of verified SNPs (true positives) for both strains. In total, we could assess 37 247 verified SNPs for SK1 and 6153 verified SNPs for W303. Short reads were mapped using BWA with default parameters. We compared ACCUSA with SNP caller from the SAMTOOLS package (Li *et al.*, 2009). Both programs were run on the BWA output to facilitate a direct comparison. Precision (TP/TP + FP) and recall (TP/TP + FN) values were calculated to compare both methods. In summary, ACCUSA discovered yeast SNPs with 95.41%/81.77% precision (for SK1 and W303, respectively), whereas the SAMTOOLS SNP caller had 77.67%/39.24% precision. The lower precision values for W303 might be explained in part by the lower number of mapped reads for W303 (15.1%) versus SK1 (43.4%). Recall rates were 69.31%/31.85% for ACCUSA on SK1/W303 and 84.11%/21.78% for SAMTOOLS SNP on SK1/W303, respectively. The proportion of true positives in the overlap of both predictors is 97.5% for SK1 and 88.7% for W303. These figures were determined for the whole W303 and SK1 dataset. Recall and precision for different reference base quality thresholds are shown in Figure 1. ACCUSA's precision is constantly higher in this comparison.

We conclude that ACCUSA's recall performance is comparable to the SAMTOOLS SNP calling software. More important, ACCUSA offers a higher precision than SAMTOOLS SNP caller. This is also true if SAMTOOLS SNP caller is combined with a reference quality masking step. Our approach was only tested on Solexa sequencing data, but should be useful in other contexts as well.

ACKNOWLEDGEMENTS

We appreciate valuable comments from three anonymous referees.

Funding: Max Planck Institute for Developmental Biology Tübingen; Max-Delbrück-Center for Molecular Medicine Berlin.

Conflict of Interest: none declared.

REFERENCES

Dohm,J.C. *et al.* (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.*, **36**, e105.

Li,H. *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Liti,G. *et al.* (2009) Population genomics of domestic and wild yeasts. *Nature*, **458**, 337–341.

Marth,G.T. *et al.* (1999) A general approach to single-nucleotide polymorphism discovery. *Nat. Genet.*, **23**, 452–456.

Metzker,M.L. (2010) Sequencing technologies - the next generation. *Nat. Rev. Genet.*, **11**, 31–46.