BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 14 2013, pages 1809–1810
doi:10.1093/bioinformatics/btt268

Sequence analysis

Advance Access publication May 16, 2013

ACCUSA2: multi-purpose SNV calling enhanced by probabilistic integration of quality scores

Michael Piechotta 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: Alfonso Valencia

ABSTRACT

Summary: Direct comparisons of assembled short-read stacks are one way to identify single-nucleotide variants. Single-nucleotide variant detection is especially challenging across samples with different read depths (e.g. RNA-Seq) and high-background levels (e.g. selection experiments). We present ACCUSA2 to identify variant positions where nucleotide frequency spectra differ between two samples. To this end, ACCUSA2 integrates quality scores for base calling and read mapping into a common framework. Our benchmarks demonstrate that ACCUSA2 is superior to a state-of-the-art SNV caller in situations of diverging read depths and reliably detects subtle differences among sample nucleotide frequency spectra. Additionally, we show that ACCUSA2 is fast and robust against base quality score deviations.

Availability: ACCUSA2 is available free of charge to academic users and may be obtained from https://bbc.mdc-berlin.de/software.

Contact: christoph.dieterich@mdc-berlin.de

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on September 5, 2012; revised on March 20, 2013; accepted on May 6, 2013

The introduction of next-generation sequencing platforms has paved the way to a new understanding of biological systems. Currently, available sequencing set-ups deliver billions of short reads in one run. This unprecedented sequencing depth is ideally suited for variant detection, e.g. SNP detection in DNA or RNA samples. For example, the variant caller in the samtools package (Li, 2011) infers genotypes by comparing a sequencing sample to a reference sequence. Our solution is more general, as it implements head-to-head cross-sample comparisons with the aim of detecting single-nucleotide variant (SNV) positions where nucleotide frequency spectra differ considerably. In our approach, short reads are first aligned to a reference sequence, and subsequently, read stacks from two samples are directly compared with one another.

The process of SNV detection typically uses sequencing base calls (BCs) and associated quality scores (QSs) from short reads ($\geq$50 nt). This information is anchored on a reference genome by a short-read mapper [e.g. MAQ (Li *et al.*, 2008)]. The mapping step yields quality scores for each short read. A quality score expresses the uncertainty that the given mapping is true (see Li *et al.*, 2008 for details). The consideration of mapping qualities in an SNV calling pipeline will improve SNV calling precision (Nielsen *et al.*, 2011) as false calls, which are induced by incorrect mappings, are filtered out.

Early SNV callers solely used BCs to identify variant sites and are likely to produce many false calls. An improvement to this situation is attained by including QSs for BCs and short-read mappings into elaborate frameworks. These QSs estimate the uncertainty from various error sources, such as wrong base calling, false read mappings or poor reference sequence quality.

Lately, probabilistic SNP callers that use QSs (see review in Nielsen *et al.*, 2011) to distinguish true variant positions from false SNPs have been implemented. First approaches focused on BC quality filtering to retain only high-quality BCs. In the course of these developments, it has been observed that raw BC QSs reported by sequencing platforms tend to deviate from the true BC error rates. Nowadays, raw BC quality scores are recalibrated with the help of known polymorphic sites to mitigate this effect [e.g. GATK (DePristo *et al.*, 2011)].

Our proposed method ACCUSA2 performs cross-sample comparisons to identify variant sites based on as few previous assumptions as possible and tests whether the underlying hypothesis of equal nucleotide frequencies in the two compared samples can be rejected.

In brief, we model the observed nucleotide frequencies Φ^i of each sample $i \in \{A, B\}$ by a Dirichlet distribution $\mathrm{Dir}(\alpha^i) : \alpha = (\alpha_A, \alpha_C, \alpha_G, \alpha_T)$ where the BC QSs and mapping QSs are integrated as priors. ACCUSA2 performs a head-to-head comparison of two BAM files. To this end, pairwise read stacks are built for every reference position that passes a user-defined minimal read coverage and quality threshold. First, we transform and combine the base quality Q^{BC}_{Phred} and mapping quality Q^{MAP}_{Phred} for each BC in the current pileup to a probability vector $\mathbf{p} = (p(A), p(C), p(G), p(T))$ by transforming the Phred quality scores to error probabilities (see Supplementary Material for details).

$$Q^{BC}_{Phred} = -10 \log_{10} P(\text{wrong BC}) \quad (1)$$

$$Q^{MAP}_{Phred} = -10 \log_{10} P(\text{wrong mapping}) \quad (2)$$

In a second step, we compile a $n \times 4$ probability matrix M^i over individual base probability vectors $\mathbf{p^i}$ from each read in the pileup (n equals read coverage). Matrix column averages are used to parameterize $\mathrm{Dir}(\alpha^i)$. We use a log-odds (LOD) score Z that is

*To whom correspondence should be addressed.

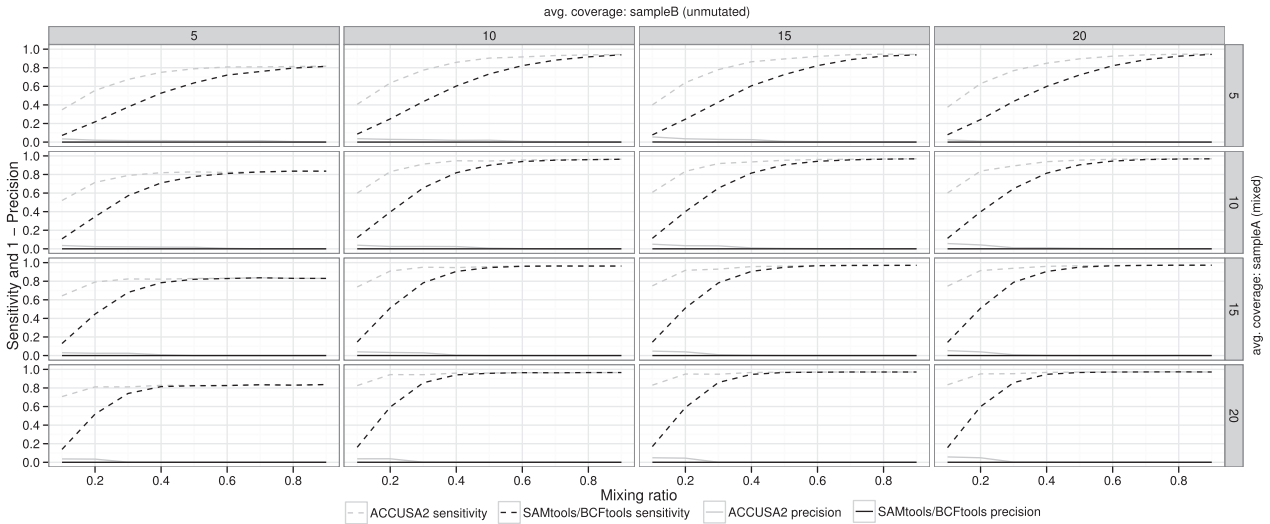

Fig. 1. Benchmark (**a**) results for 'DNA mixed ratio' set-up without base quality recalibration showing the sensitivity and (1: precision) of ACCUSA2 in comparison with SAMtools/BCFtools for the respective target coverage and target mixing ratio combinations

defined as the ratio of likelihoods of $\mathrm{Dir}(\Phi^i;\alpha^j)$ to rank our SNV calls. This score becomes zero if $\alpha^A = \alpha^B$:

$$Z = \log_{10}\frac{\mathrm{Dir}(\Phi^B;\alpha^A)\cdot\mathrm{Dir}(\Phi^A;\alpha^B)}{\mathrm{Dir}(\Phi^A;\alpha^A)\cdot\mathrm{Dir}(\Phi^B;\alpha^B)} \quad (3)$$

We implemented our approach in a JAVA program called ACCUSA2 and contrasted it with SAMtools/BCFtools (Li, 2011) on simulated datasets DNA-seq and RNA-seq. Briefly, we used the MAQ simulation tools and the Flux simulator (Griebel *et al.*, 2012) to built our benchmark read datasets from the reference sequence of *Caenorhabditis elegans*. Our benchmark is designed to contrast two samples of either DNA (a) or RNA short reads (b). We account for the different read properties by using MAQ for benchmark (a) and the Flux simulator for benchmark (b). Our benchmark implements a read mixing scenario (i.e. one sample contains only reads that originate from a reference sequence and another sample contains reads from a diverged sequence and mutation-less reads) in combination with a scan over different input read coverages. In both benchmarks, each of the two input samples was sequenced to a different read depth (see Supplementary Material for details). We explore different read mixing ratios (amount of mutated versus unmutated reads) for the second sample (10–90%).

We compared our method with the SAMtools/BCFtools pipeline. Figure 1 shows the performance of both methods in benchmark (a) without quality score recalibration. ACCUSA2 performs better in situations where a combination of different sample read coverages and a low-mixing ratio is encountered. We observed a superior sensitivity (average sensitivity of 86.71% achieved compared with 72.30% of SAMtools/BCFtools) for ACCUSA2 while being comparable in SNP calling precision (on average 98.76% compared with 99.99%). A detailed assessment of both benchmarks is given in the Supplementary Materials. Generally, ACCUSA2 shows higher sensitivity and a greater robustness against QS deviations than the SAMtools/BCFtools pipeline.

In summary, ACCUSA2 is a multi-purpose tool for SNV discovery and will perform best in situations where read coverages differ between samples, and SNVs of interest might only be lowly represented. Higher robustness of ACCUSSA2 against QS deviations enables applications where variant positions are *a priori* unknown, and recalibration can not be easily performed. Future versions of ACCUSA2 will target specialized applications such as RNA editing event analysis, selection experiments, bulk segregant analysis, and novel assays for mapping of RNA–protein interactions (e.g. PAR-CLIP).

ACKNOWLEDGEMENT

Both authors appreciate numerous discussions with Andreas Ipsen. They thank Sebastian Fröhler for his support in transitioning from ACCUSA to ACCUSA2.

Funding: As part of the Berlin Institute for Medical Systems Biology at the MDC, the research group of C.D. is funded by the Federal Ministry for Education and Research (BMBF) and the Senate of Berlin, Berlin, Germany (0315362A).

Conflict of Interest: none declared.

REFERENCES

DePristo,M.A. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–498.

Griebel,T. *et al.* (2012) Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic acids research*, **40**, 10073–10083.

Li,H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–2993.

Li,H. *et al.* (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18**, 1851–1858.

Nielsen,R. *et al.* (2011) Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.*, **12**, 443–451.