Vol. 27 no. 6 2011, pages 865–866
doi:10.1093/bioinformatics/btr032

Advance Access publication January 28, 2011

Bambino: a variant detector and alignment viewer for next-generation sequencing data in the SAM/BAM format

Michael N. Edmonson[1,*], Jinghui Zhang[2], Chunhua Yan[3], Richard P. Finney[3],
Daoud M. Meerzaman[1] and Kenneth H. Buetow[1]

[1]Laboratory of Population Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, [2]Department of Computational Biology, St. Jude Children's Research Hospital, 262 Danny Thomas Place, Memphis, TN 38105 and [3]Center for Bioinformatics and Information Technology, National Cancer Institute, Rockville, MD 20852, USA

Associate Editor: Joaquin Dopazo

ABSTRACT

Summary: Bambino is a variant detector and graphical alignment viewer for next-generation sequencing data in the SAM/BAM format, which is capable of pooling data from multiple source files. The variant detector takes advantage of SAM-specific annotations, and produces detailed output suitable for genotyping and identification of somatic mutations. The assembly viewer can display reads in the context of either a user-provided or automatically generated reference sequence, retrieve genome annotation features from a UCSC genome annotation database, display histograms of non-reference allele frequencies, and predict protein-coding changes caused by SNPs.

Availability: Bambino is written in platform-independent Java and available from https://cgwb.nci.nih.gov/goldenPath/bamview/documentation/index.html, along with documentation and example data. Bambino may be launched online via Java Web Start or downloaded and run locally.

Contact: edmonson@nih.gov.

Received on October 22, 2010; revised on December 29, 2010; accepted on January 16, 2011

1 INTRODUCTION

The SAM format (Li *et al.*, 2009a) along with BAM, its compressed equivalent, is an emerging standard for efficient storage and retrieval of second-generation sequencing data and their associated mapping results. Reference sequence alignment programs such as mapping and assembly with qualities (MAQ) and burrows-wheeler alignment tool (BWA) (Li *et al.*, 2009b) support SAM either directly or via conversion with samtools. Our program, 'Bambino', can detect SNPs, insertions and deletions directly from BAM files, leveraging a wide range of SAM sequence annotations during the process. This is an advantage over tools such as VarScan (Koboldt *et al.*, 2009) which are driven from intermediate files in the sparser 'pileup' format.

Because SAM provides an ideal interchange format, it is desirable to have visualization and analysis tools that not only support SAM, but also can work with collections of files. Bambino's variant detector and assembly viewer are capable of pooling and analyzing data from multiple BAM files simultaneously. Dynamic pooling of BAM data facilitates integrated analysis of paired tumor/normal samples as well as data from multiple runs or platforms, and obviates the need for expensive construction of combined files for these analyses. This is an advantage over programs such as MagicViewer (Hou *et al.*, 2010) which can only work with a single file at a time.

2 VISUALIZATION

Bambino's assembly viewer displays alignments from one or more BAM files against a reference sequence either loaded from a file (FASTA, UCSC .2bit and .nib formats are supported), or generated from the underlying reads. Nucleotides are displayed color-shaded based on quality values, in a style similar to the consed program (Gordon *et al.*, 1998). Display of various SAM alignment tags is supported, including soft and hard clipping and spliced alignments. Padding characters are added to the reference sequence and alignments as necessary to provide complete visualization of insertions and short tandem repeats. The viewer generates a summary histogram of non-reference allele frequencies for both tumor and normal samples, providing a quick impression of whether potential SNP sites are homozygous or heterozygous, germline or somatic. Another panel displays a bird's eye view of the wider region, showing normalized depth of coverage and exon positions. The viewer also displays dbSNP entries and NCBI RefSeq protein translations retrieved from a configurable MySQL UCSC genome annotation database (Rhead *et al.*, 2009) and can predict whether a given variant alters protein coding.

3 VARIANT DETECTION

Bambino includes a variant detector, which can identify single nucleotide variants (SNVs), insertions and deletions directly from one or more BAM files. SAM-specific features include the ability to specify a minimum read mapping quality, mate-pair read consistency checks and SAM tag filtering. The latter feature allows the user to leverage even custom SAM tags: for example, if the BAM data were generated using BWA, filters using the X0 or XT tags could ensure variant calling was performed exclusively with uniquely mapped reads. Minimum read quality, depth of coverage and allele frequency are also configurable. Each variant is assigned a Bayesian quality score (Buetow *et al.*, 1999) based on the conversion of associated SAM reads' phred-scaled (Ewing *et al.*, 1998) quality

[*]To whom correspondence should be addressed.

Published by Oxford University Press on behalf of the US Government 2011.

Table 1. Summary of results: validation of novel variants, and detection of variants confirmed by other groups

Dataset	Samples	Variants	Detected	Validation rate (%)
Validation of novel SNPs in liver cancer	3	55	50	90.9
TCGA-validated variants found via next-gen sequencing	440	1739	1704	97.9
TCGA-validated variants found via SNP6	7	1 728 968	1 717 830	99.3

scores into probability-of-error values. This score is most helpful for evaluating calls in low-coverage regions. A variety of low-level options and settings are available for configuration by the user, increasing transparency and making it easier to adapt the detector to different use cases, for example, the analysis of assemblies of long Sanger-based reads aligned with BWA's 'bwasw' command. Variant detection may be configured and run from the command line or interactively from within the assembly viewer. Various techniques are used to avoid false positive variation calls, several of which focus on ambiguously mapped or mismapped reads. A given read may be rejected altogether for variant calling if it contains more than a maximum number of mismatches to the reference sequence for one of two sequence quality thresholds (the default settings permit a maximum of three high-quality mismatches and six low-quality mismatches). Mismatches of extremely low quality ($q \leq 2$) may be optionally ignored to accommodate Illumina's reserved usage of these values. A mismapped read filter tracks high-quality mismatches in these rejected reads, disqualifying candidate variants elsewhere if their alleles appear too frequently in the mismatch set. This prevents false positive calls based on reads which even partially overlap problematic regions. Another filter discounts reference mismatches near read termini occurring within regions deleted from other reads, considering them possible indel alignment errors. Read mate-pair consistency checks are also performed, excluding overlapping reads from variant calling if their base calls disagree.

The variant detector can pool data from multiple BAM files, facilitating analysis of tumor/normal pairs, or data from multiple runs or platforms. Each file may be annotated as tumor or normal. Detailed counts of reads supporting each variant are provided, broken down by tumor/normal status, allele and strand. This adds an additional level of granularity beyond that provided by pileup-format files, and can be used to determine whether detected variants are homozygous or heterozygous, germline or somatic. As measures of supporting read diversity, counts of unique clone names observed for each variant are provided, as well as a summary flag indicating whether each variant was observed bidirectionally. An optional read-level report provides extended detail about participating reads and their SAM annotations. For BAM files containing regions of extremely high-read coverage (e.g. liver albumin), an optional limiter may be employed to restrict memory usage during processing. The variant detector uses a streaming model to manage memory usage, which along with the limiter feature makes it capable of analyzing even very large whole-genome datasets. Memory consumption is dependent on the limiter settings used; the program generally runs well using 1 or 2 GB RAM.

4 RESULTS

To assess the accuracy of Bambino's variant detector, we sequenced a set of 55 SNPs called from three liver cancer samples (Meerzaman *et al*., 2011, in submission), selecting candidates with a coverage depth of at least 10 reads and a predicted minor allele frequency of ≥30%. Of these SNPs, 50 (90.9%) were validated. For the subset of SNPs having a coverage depth of 20 reads or more, 30 (96.7%) of 31 were validated.

We also compared Bambino's predicted calls with two collections of validated variants from The Cancer Genome Atlas (TCGA) project, available from its data access portal (http://tcga-data.nci.nih.gov/tcga/). In a set of 1739 somatic variants identified in next-generation sequencing data of 440 ovarian cancer samples, Bambino was able to detect 1704 (97.9%) of the same sites. Of the 35 variants which were not found, 20 showed no supporting reads in the associated BAM files, and 10 showed only a single supporting read, often of low quality. Because BAM files distributed by the TCGA project are subject to replacement by updated versions, it is possible we did not have the same versions used by the sequencing centers for their calls. If we eliminate these sites from consideration, our detection rate approaches 100%. Additionally, we obtained a set of seven normal samples from the TCGA ovarian cancer project which have both whole-genome next-generation sequencing and Affymetrix SNP6 array data. Bambino identified >99% of heterozygous SNPs called from Affymetrix SNP6 array data in the equivalent whole-genome next-generation sequencing data. It is important to note that these TCGA dataset comparisons checked only for false negatives: because we did not perform any validation experiments in these data, it is difficult to estimate a false positive call rate. Validation rates generally will be heavily influenced by the settings used for variant detection and any subsequent filtering of putative sites.

Funding: This work was supported by the National Institutes of Health (ZIH BC 010470).

Conflict of Interest: none declared.

REFERENCES

Buetow,K.H. *et al.* (1999) Reliable identification of large numbers of candidate SNPs from public EST data. *Nat. Genet*., **21**, 323–325.

Ewing,B. *et al.* (1998) Base-calling of automated sequencer traces using phred. *Genome Res*., **8**, 175–185.

Gordon,D. *et al.* (1998) Consed: a graphical tool for sequence finishing. *Genome Res*., **8**, 195–202.

Hou,H. *et al.* (2010) MagicViewer: integrated solution for next-generation sequencing data visualization and genetic variation detection and annotation. *Nucleic Acids Res*., **38** (Suppl. 2), W732–W736.

Koboldt,D.C. *et al.* (2009) VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, **25**, 2283–2285.

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

Li,H. *et al.* (2009b) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Rhead,B. *et al.* (2009) The UCSC Genome Browser database: update 2010. *Nucleic Acids Res*., **38** (Suppl. 1), D613–D619.