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

Michael N. Edmonson1,*, Jinghui Zhang2, Chunhua Yan3, Richard P. Finney3, Daoud M. Meerkaman1 and Kenneth H. Buetow1
1Laboratory of Population Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, 2Department of Computational Biology, St. Jude Children’s Research Hospital, 262 Danny Thomas Place, Memphis, TN 38105 and 3Center for Bioinformatics and Information Technology, National Cancer Institute, Rockville, MD 20852, USA

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 alignment 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://cgrweb.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 program (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 programs such as VarScan (Koboldt et al., 2009) which can only work with a single file at a time.

2 VISUALIZATION

Bambino’s assembly viewer displays alignments from one or more BAM data 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.
Variant detection may be configured and run from the command line. TCGA-validated variants found via next-gen sequencing the limiter feature makes it capable of analyzing even very large BAM files containing regions of extremely high-read coverage detail about participating reads and their SAM annotations. For a summary flag indicating whether each variant was observed clone names observed for each variant are provided, as well as somatic. As measures of supporting read diversity, counts of unique reads supporting each variant are provided, broken down by tumor/normal status, allele and strand. This facilitates analysis of tumor/normal pairs, or data from multiple lines 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 sequence reference 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 (≥ 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 mismatch set. This prevents false positive calls based on reads which have an 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


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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Samples</th>
<th>Variants</th>
<th>Detected</th>
<th>Validation rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validation of novel SNPs</td>
<td>3</td>
<td>55</td>
<td>50</td>
<td>90.9</td>
</tr>
<tr>
<td>in liver cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TCGA-validated variants</td>
<td>440</td>
<td>1739</td>
<td>1704</td>
<td>97.9</td>
</tr>
<tr>
<td>found via next-gen sequencing</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TCGA-validated variants</td>
<td>7</td>
<td>1728968</td>
<td>1717830</td>
<td>99.3</td>
</tr>
<tr>
<td>found via SNP6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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. 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 sequence reference 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 (≥ 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 mismatch set. This prevents false positive calls based on reads which have an 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


866