BIOINFORMATICS APPLICATIONS NOTE

Vol. 27 no. 1 2011, pages 130–131
doi:10.1093/bioinformatics/btq614

Sequence analysis

Advance Access publication November 18, 2010

SAMStat: monitoring biases in next generation sequencing data

Timo Lassmann*, Yoshihide Hayashizaki and Carsten O. Daub

Omics Science Center, Riken Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan

Associate Editor: Martin Bishop

ABSTRACT

Motivation: The sequence alignment/map format (SAM) is a commonly used format to store the alignments between millions of short reads and a reference genome. Often certain positions within the reads are inherently more likely to contain errors due to the protocols used to prepare the samples. Such biases can have adverse effects on both mapping rate and accuracy. To understand the relationship between potential protocol biases and poor mapping we wrote SAMstat, a simple C program plotting nucleotide overrepresentation and other statistics in mapped and unmapped reads in a concise html page. Collecting such statistics also makes it easy to highlight problems in the data processing and enables non-experts to track data quality over time.

Results: We demonstrate that studying sequence features in mapped data can be used to identify biases particular to one sequencing protocol. Once identified, such biases can be considered in the downstream analysis or even be removed by read trimming or filtering techniques.

Availability: SAMStat is open source and freely available as a C program running on all Unix-compatible platforms. The source code is available from http://samstat.sourceforge.net.

Contact: timolassmann@gmail.com

Received on June 20, 2010; revised on October 28, 2010; accepted on October 29, 2010

1 INTRODUCTION

Next generation sequencing is being applied to understand individual variation, the RNA output of a cell and epigenetic regulation. Not surprisingly, the mapping of short reads to the genome has received a lot of attention with over 30 programs published to date [see Trapnell and Salzberg (2009) for a review of the most commonly used approaches]. Nevertheless, commonly a noticeable fraction of reads remains unmatched to the reference genome in each experiment. One possibility is that these reads simply represent the fraction of reads containing more sequencing errors in the form of mismatches, insertions or deletions than the programs can handle. Alternatively, it is conceivable that these reads contain contaminants and therefore do not map to the expected reference sequence. Finally, the unmapped reads may represent novel splice junctions or genomic regions absent from the reference assembly. Understanding the reason behind obtaining unmapped reads is clearly of interest.

Mapping programs like MAQ (Li *et al*., 2008) and BWA (Li and Durbin, 2009) report mapping qualities allowing for further investigation. We wrote SAMstat to contrast properties of unmapped, poorly mapped and accurately mapped reads to understand whether particular properties of the reads influence the mapping accuracy. As the name suggests, our program is designed to work mainly with SAM/BAM files (Li *et al*., 2009) but also not only can be used to visualize nucleotide composition and other basic statistics of fasta and fastq (Cock *et al*., 2009) files.

Table 1. Overview of SAMstat output

Reported statistics
Mapping rate[a]
Read length distribution
Nucleotide composition
Mean base quality at each read position
Overrepresented 10mers
Overrepresented dinucleotides along read
Mismatch, insertion and deletion profile[a]

[a]Only reported for SAM files.

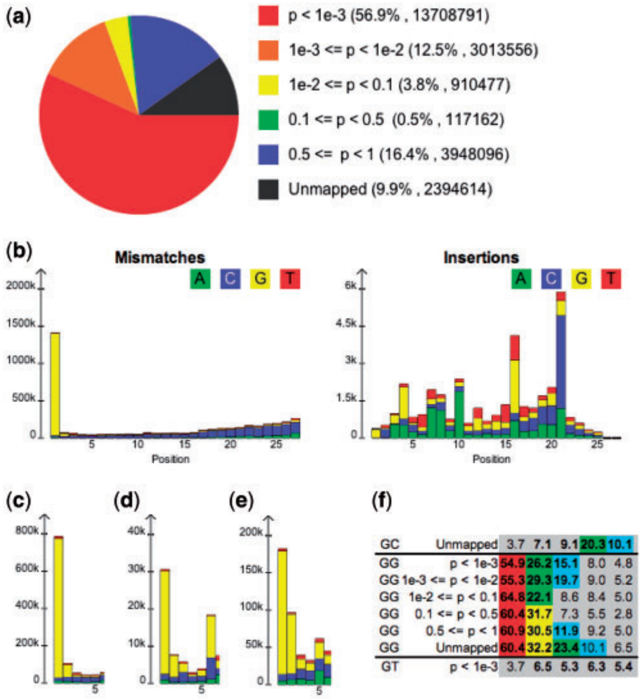
2 METHODS

SAMStat automatically recognizes the input files as either fasta, fastq, SAM or BAM and reports several basic properties of the sequences as listed in Table 1. Multiple input files can be given for batch processing. For each dataset, the output consists of a single html5 page containing several plots allowing non-specialists to visually inspect the results. Naturally, the html5 pages can be viewed both on- and off-line and easily be stored for future reference. All properties are plotted separately for different mapping quality intervals if those are present in the input file. For example, mismatch profiles are given for high-and low-quality alignments allowing users to verify whether poorly mapped reads contain a specific collection of mismatches. The latter may represent untrimmed linkers in a subset of reads. Dinucleotide overrepresentation is calculated as described by Frith *et al*. (2008). Overrepresented 10mers are calculated by comparing the frequency of 10mer within a mapping quality interval compared with the overall frequency of the 10mer.

3 RESULTS AND DISCUSSION

To demonstrate how SAMStat can be used to visualize mapping properties of a next generation datasets, we used data from a recently published transcriptome study (Plessy *et al*., 2010); (DDBJ short read archive: DRA000169). We mapped all 24 million 5′ reads to the human genome (GRCh37/hg19 assembly) using BWA (Li and Durbin, 2009) with default parameters. SAMStat parsed the alignment information in ∼3 min which is comparable to the 2 min it takes to copy the SAM file from one directory to another. The majority of reads can be mapped with very high confidence (Fig. 1a). When inspecting the mismatch error profiles, we noticed that there

*To whom correspondence should be addressed.



GC	Unmapped	3.7	7.1	9.1	20.3	10.1
GG	p < 1e-3	54.9	26.2	15.1	8.0	4.8
GG	1e-3 <= p < 1e-2	55.3	29.3	19.7	9.0	5.2
GG	1e-2 <= p < 0.1	64.8	22.1	8.6	8.4	5.0
GG	0.1 <= p < 0.5	60.4	31.7	7.3	5.5	2.8
GG	0.5 <= p < 1	60.9	30.5	11.9	9.2	5.0
GG	Unmapped	60.4	32.2	23.4	10.1	6.5
GT	p < 1e-3	3.7	6.5	5.3	6.3	5.4

Fig. 1. A selection of SAMStat's html output. (**a**) Mapping statistics. More than half of the reads are mapped with a high mapping accuracy (red) while 9.9% of the reads remain unmapped (black). (**b**) Barcharts showing the distribution of mismatches and insertions along the read for alignments with the highest mapping accuracy [shown in red in (a)]. The colors indicate the mismatched nucleotides found in the read or the nucleotides inserted into the read. (**c**,**d** and **e**) Frequency of mismatches at the start of reads with mapping accuracies $1e^{-3} \leq P < 1e^{-2}$, $1e^{-2} \leq P < 0.5$ and $0.5 \leq P < 1$, respectively (shown in orange, yellow and blue in panel a). The fraction of mismatches involving G's at position 2–5 increases. (**f**) Percentage of 'GG' dinucleotides at positions 1–5 in reads split up by mapping quality intervals. The background color highlights large percentages. The first and last row for nucleotides 'GT' and 'GC' are shown for comparison.

are many mismatches involving a guanine residue at the very start of many reads (yellow bars in Fig. 1b–e). These 5′ added guanine residues are known to originate from the reverse transcriptase step in preparing the cDNAs (Carninci *et al*., 2006). When comparing the mismatch profiles for high (Fig. 1b) to low-quality alignments (Fig. 1e), it is clear that a proportion of reads contain multiple 5′ added G's which in turn pose a problem to the mapping. For example, at the lowest mapping quality (Fig. 1e), there are frequent mismatches involving G's at positions one, two and to a lesser extent until position five while in high-quality alignments the mismatches are confined to the first position of the reads (Fig. 1b).

Alongside the mismatch profiles SAMstat gives a table listing the percentages of each dinucleotide at each position of the reads split up by mapping quality intervals (Fig. 1f). For the present dataset, 60.4% of unmapped reads start with 'GG' and 10.1 percent contain a 'GG' at position 4. Evidently, 5′ G residues are added during library preparation and the start positions of mappings should be adjusted accordingly.

SAMStat is ideally suited to deal with the ever increasing amounts of data from second-and third-generation sequencing projects. Specific applications include the verification and quality control of processing pipelines, the tracking of data quality over time and the visualization of data properties derived from new protocols and approaches which in turn often leads to novel insights.

ACKNOWLEDGEMENT

We thank the reviewers for constructive suggestions.

Funding: Research Grant for the RIKEN Omics Science Center from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government (MEXT) to Y.H.

Conflict of Interest: none declared.

REFERENCES

Carninci,C. *et al*. (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.*, **38**, 626–635.

Cock,P.J. *et al*. (2009) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.*, **38**, 1767–1771.

Frith,M.C. *et al*. (2008) A code for transcription initiation in mammalian genomes. *Genome Res.*, **18**, 1–12.

Li,H. and Durbin,R. (2009) Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics*, **25**, 1754–1760.

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

Li,H. *et al*. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Plessy,C. *et al*. (2010) Linking promoters to functional transcripts in small samples with nanoCAGE and CAGEscan. *Nat. Methods*, **7**, 528–534.

Trapnell,C. and Salzberg,S.L. (2009) How to map billions of short reads onto genomes. *Nat. Biotechnol.*, **27**, 455–457.