Journal of
Computer Science & Systems Biology

Francesco et al., J Comput Sci Syst Biol 2013, 6:4
DOI: 10.4172/jcsb.1000116

Research Article Open Access

SAM-Profiler: A Graphical Tool for Qualitative Profiling of Next Generation Sequencing Alignment Data

Flora Francesco[1], Pirola Alessandra[2], Spinelli Roberta[2], Redaelli Sara[2], Valletta Simona[2], Gambacorti Passerini Carlo[2] and Piazza Rocco[2]*

[1]Department of Informatics, Systems and Communication, University of Milano-Bicocca, 20900, Monza, Italy
[2]Deptartment of Health Sciences, University of Milano-Bicocca, 20900, Monza, Italy

Abstract

SAM/BAM alignment file formats are extensively used in virtually all the laboratories devoted to high-throughput sequencing. However, limited effort has been yet dedicated to the development of SAM/BAM quality reporting tools. To overcome this problem, we developed SAM-Profiler, a multiplatform tool dedicated to the advanced quality reporting of SAM/BAM files. SAM-Profiler performs qualitative analysis of SAM/BAM alignment data in the context of next-generation sequencing. It is implemented in C# and can be run under Windows, Linux and MacOS operative systems. Two versions are available: fully graphical, event-driven software and a command-line tool.

SAM-Profiler is able to generate an extensive set of qualitative reports on SAM/BAM alignment data, among them: overall, per-base and per-chromosome read quality, mapping quality, duplicate and coverage analyses, bases distribution, perfect, proper and improper mapping, exonic, intronic, intergenic, 5` and 3` UTR coverage, mismatch distribution profile and CG distribution. In presence of paired-end sequencing experiments our tool is able to automatically report the insert size distribution and to analyze the relative pair mapping, reporting absolute and relative distribution of properly, improperly mapped, mapped/unmapped and unmapped pairs. Its modular architecture allows embedding additional analytical monitoring/reporting tools to the already developed list, allowing SAM-Profiler to grow according to the specific requests of the end-users.

Keywords: Next-Generation Sequencing; SAM; BAM; Alignment; Qualitative Analysis; Multiplatform

Abbreviations: NGS: Next-Generation Sequencing; BM: Bone Marrow; PB: Peripheral Blood; gDNA: Genomic DNA; FIFO: First-In First-Out; WES: Whole-Exome Sequencing

Introduction

The development of high-throughput sequencing instruments generated a tremendous amount of data from different sources: from viral and bacterial de novo genomes to human resequencing projects, such as the ambitious 1000 genomes project [1]. This led to a previously unimagined and challenging bioinformatics problem: storing and analyzing huge amounts of sequencing data.

In fact, if just a few years ago the complexity of generating sequencing data was more challenging and time-consuming than analyzing them, the sequencing revolution originated by the development of high-throughput sequencing instruments completely changed this scenario [2]. Actually, with single sequencing instruments capable of a throughput of 500 Gigabases of raw sequences per week, the bottleneck between sequencing and data analysis is clearly identifiable in the latter.

One of the most critical issues in the analysis of sequencing data is the requirement of standardized data-formats. This led to the definition of universally-accepted file formats, such as the Sanger-FastQ [3] (Fasta-with-Quality) and the SAM/BAM [4] formats. Specifically, the SAM (Sequence Alignment/Map format) and the corresponding binary file BAM (Binary Alignment/Map format) formats are actually routinely used in virtually all the sequencing centers devoted to high-throughput sequencing. Their acceptance as the standard alignment file formats led to the development of an extremely large set of bioinformatics tools that are able to process them in order to address many different biological questions. Surprisingly however, limited effort has been yet devoted to the development of software dedicated to the analysis of alignment files in order to generate complete quality reports. Due to the complexity and size of alignment files, the availability of advanced reporting tools is now critical to inspect SAM/BAM files either for routine quality reporting, in order to detect the presence of sequencing problems and library preparation errors in advance, or to perform in-depth post-processing tests, in case of failure of downstream analyses.

To overcome the limited availability of SAM/BAM reporting tools, we developed SAM-Profiler, a bioinformatics tool dedicated to the advanced quality reporting of SAM and BAM files available as fully graphical, event-driven software and as a command-line tool.

Materials and Methods

Patient samples

Bone Marrow (BM) or Peripheral Blood (PB) samples from atypical chronic myeloid leukemia patients were collected at diagnosis, after obtaining written informed consent approved by the local ethics committee [5]. Myeloid lineage cells were collected applying Ficoll Paque Plus gradient (GE Healthcare, UK) to BM samples, or by buffy coat to PB samples. Cytofluorimetric analysis of the cells phenotype, confirmed that the percentage of myeloid cells was greater than 80%. Lymphocytes from the same patients were obtained as described previously [5].

***Corresponding author:** Piazza Rocco, Department of Health Sciences, University of Milano-Bicocca, 20900, Monza, Italy, Tel: 0390264488059; Fax: 0390264488363; E-mail: rocco.piazza@unimib.it

Received July 25, 2013; **Accepted** August 15, 2013; **Published** August 22, 2013

Citation: Francesco F, Alessandra P, Roberta S, Sara R, Simona V, et al. (2013) SAM-Profiler: A Graphical Tool for Qualitative Profiling of Next Generation Sequencing Alignment Data. J Comput Sci Syst Biol 6: 188-193. doi:10.4172/jcsb.1000116

Exon library preparation

Genomic DNA (gDNA) was extracted with the PureLink^{TM} kit (Invitrogen, Life technology, Grand Island, NY, USA) according to manufacturer procedures. 1 µg of gDNA was fragmented to an average size of 500-100 bp and then processed according to the standard protocol for the Illumina TruSeq DNA Sample Preparation kit (FC-121-1001), with selection of fragments of 200–300 bp on 2% agarose gels. Multiplexed genomic libraries were then enriched with the Illumina TruSeq Exome Enrichment kit (FC-121-1008). Libraries were then sequenced on an Illumina Genome Analyzer IIx with 76 bp paired-end reads using the Illumina TruSeq SBS kit v5 (FC-104-5001).

Statistical analysis

All the statistical t-test analyses were run on GraphPad software analysis program (Prism, San Diego, CA).

Algorithm implementation

SAM-Profiler is entirely written in C# and can be run under Windows (using the .NET runtime library), Linux and Apple Mac OS operative systems (using the Mono runtime library: http://www.mono-project.com/Main_Page). It has been developed using streaming techniques and limited memory footprint requirements, so it is typically able to smoothly run on 4 Gigabytes RAM memory desktop or notebook systems. To improve its performance, our tool makes use of parallel programming techniques. Specifically, two different internal pipelines are implemented, depending on whether a SAM or a BAM file is under processing. In presence of SAM files, a parallel, two steps Producer/Consumer blocking First-In First-Out (FIFO) collection is implemented, where the producer, dedicated to the acquisition and preprocessing of each read, feeds the consuming algorithm which qualitatively analyzes the preprocessed data. Although BAM files are preferable over SAM for several reasons, among them the reduced file size and the ability to quickly extract reads within a specified position range, their use is computationally intensive, because of the overhead required by the BGZF/gzip blocks decompression and, in case of paired-read experiments, by the read-matching algorithms. To take into account this problem, in order to process BAM files, SAM-Profiler generates a three-step Producer/Consumer blocking collection, where a first thread processes the BAM/BAI files and extracts individual reads, instantiating dedicated *BamRead* reference objects. The *BamRead* objects are used to internally feed a second thread, which is mainly devoted to the preprocessing and, in case of paired-end reads, to the read-matching algorithms. Finally, a third thread qualitatively analyzes the preprocessed, matched reads. Details about the individual algorithms can be found in the supplementary material.

Results

To demonstrate how SAM-Profiler can be used to generate quality reports (Figure 1) from next generation datasets, we analyzed a set of 13 paired-end Whole-Exome Sequencing (WES) BAM files from our recently published high-throughput sequencing study [5]. On average the size of the BAM files was 8.5 Gigabytes and the number of reads per experiment 141×10^6. The median time required to complete an analysis was two hours on a 4 Gigabytes notebook PC.

Read and mapping quality

Overall the analysis of the 13 WES files revealed a very high read and mapping quality throughout all the 24 chromosomes, with a median read quality of 39.0 and a global median mapping quality of 49.0. The read quality didn't change significantly in the different chromosomes, with a homogeneous quality distribution across all the experiments analyzed. Globally, 37.8% nucleotides scored a read quality (Phred) of 40 and 78.5% a Phred range of 35-40.

In depth analysis of the per-position read quality distribution revealed that, despite the improvements in the sequencing chemistry achieved in the last years, a progressive albeit slow decrease in the overall read quality was still detectable when moving from the first to the last positions of each read (Figure 1a). In our experiments, the median read quality in the first and in the last 10 bases was 39.0 and 36.0, respectively, indicating that the degradation of the read quality, albeit detectable, remained at reasonable levels even in the last positions, at least with 76 bp reads. Comparative analysis of the read quality in the two paired reads indicated that the global read quality of the second read was slightly lower than that of the first one, although the difference was not significant (Mean read quality of first and second read of 35.6 and 34.8, respectively; p=0.20). Indeed, the degradation of the read quality in the last nucleotides was more pronounced in the second than in the first read (Mean read quality of the last base: 31.4 and 29.5, respectively; p=0.011). Taken together these data suggest that monitoring the individual read quality of both reads is important, especially in experiments with long reads, in order to decide whether to perform base trimming or not.

Duplicates analysis

It is known that the amplification steps required to process nucleic acids and generate the sequencing libraries are prone to the development of duplicate artifacts. This typically occurs when limited amounts of nucleic acids are used as a template in PCR reactions. Although this is a well-known phenomenon, biological steps requiring PCR amplification are often necessary because of the limited concentration of the sequencing libraries. Therefore, duplicates identification and removal is now considered a critical step in many high throughput sequencing pipelines. To allow an in-depth analysis of duplicates in alignment data, SAM-Profiler calculates the individual duplicate profiles of all the chromosomes by using two different approaches: the first one is more stringent and requires that two reads have same coordinates and same sequence to label them as a duplicate; the second one only takes into account the position of the mapped reads. Moreover, in presence of paired-end reads SAM-Profiler automatically adapts the duplicate identification strategy by taking into account the combined position of the two paired reads. The analysis of our WES data by using the two approaches led to similar results: in general the duplicate level was acceptable throughout the whole dataset, with a median percentage of duplicate reads of 11.5%. The per-chromosome duplicate distribution was uniform (not shown), indicating that the duplicates are generated homogeneously throughout each WES library.

To further test the duplicate detection algorithms and to assess if our tool was able to detect the presence of abnormal duplicate events in real high-throughput sequencing experiments, we generated new analyses by using alignment data from 4 exomes where a nested PCR was required to generate an adequate amount of sequencing library. As expected, a sharp increase in the relative number of duplicate reads was detected by SAM-Profiler in this subset (Figure 2; p<0.0001), with a mean duplicate percentage of 72.6% when taking into account both sequence and position.

Coverage analysis

Coverage analysis plots the coverage distribution in exonic,

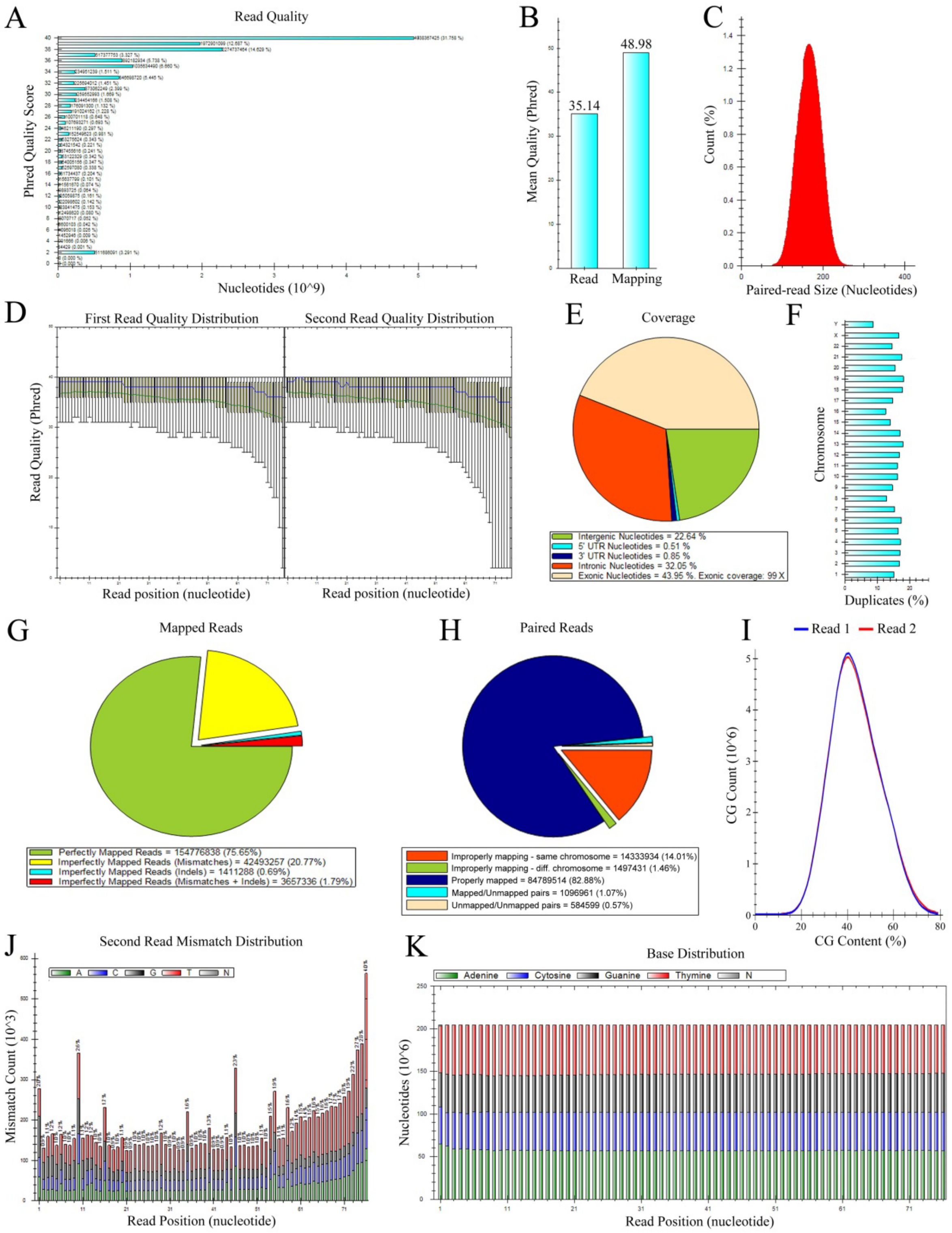

Figure 1: Selection of SAM-Profiler output panels. A) Read quality distribution. B) Overall mapping and read quality. C) Pair-end fragment distribution. D) Read quality distribution across the two reads of a pair-end experiment: mean, median, P10, 25, 75 and 90 are showed for first and second read. E) Coverage distribution across exonic, intronic and intergenic regions and in 3'/5'UTR. Fold-Coverage for exonic regions is also reported. F) Pie chart showing the distribution of perfectly and imperfectly mapped reads. Imperfectly mapped reads are further analyzed in order to report the cause of the suboptimal mapping. G) Pie chart showing the relative mapping of paired reads. Counts and percentage of Properly, Improperly, Mapped/Unmapped and Unmapped/Unmapped pairs are shown. Improperly mapped pairs are further analyzed in order to report improperly mapping occurring in the same or in different chromosomes. H) Bar chart of the per-chromosome duplicates analysis. Here only the per-sequence/per-position duplicate panel is shown. I) Bar graph reporting the per-read and per-base mismatch distribution, with individual reports for each of the four bases plus N. Here only the mismatch distribution for the second read is shown. J) Bar chart showing the overall base distribution across the reads.

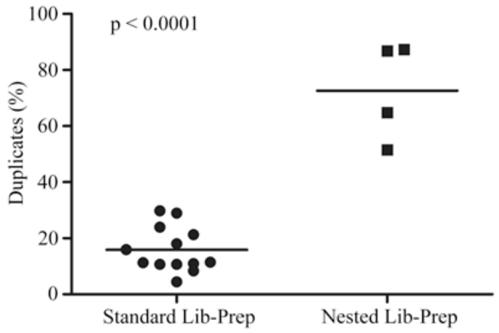

Figure 2: Duplicate analysis performed on Whole-Exome Sequencing samples generated using either the standard library preparation (Standard Lib-Prep) or a nested polymerase chain reaction (Nested Lib-Prep) protocol.

intronic, intergenic, 5' and 3'UTR regions. This analysis, applied to our WES dataset, indicated that the exonic enrichment was effective, although slightly less efficient than expected, with mean and median percentage of nucleotides covering exonic regions of 45.0 and 44.2%, respectively, and with a mean of 32.7 and 21.0% intronic and intergenic sequences. Median exonic coverage throughout the whole dataset was 74x.

Fragment size distribution

By using the mapping coordinates of paired-reads in alignment files, SAM-Profiler calculates the fragment size distribution. To generate this distribution, our tool takes into account the coordinates of all the paired-reads mapping to the same chromosome, either with proper or improper mapping and generates the distribution, capping the maximum distance taken into account at a user defined length (default is 1000 bp).

The analysis of fragment size distribution is particularly useful in order to perform a fine tuning of paired-end experiments. Specifically, it may highlight the presence of too broad distributions, of abnormal peaks with an irregular distribution shape or it may show that on average the fragment size is too short, therefore leading to the generation of overlapping reads. In our tests, the fragment size distribution analysis revealed the presence of the latter problem, with a median fragment size of 159 bp and a 25th percentile of 146 bp in presence of 76 bp paired-end reads, indicating that in a considerable fraction of experiments a limited but consistent overlap between the two reads was present and suggesting that an overall increase in the size of the sheared genomic DNA could be beneficial.

Mapping

The mapping distribution analysis reports the percentage and absolute counts of perfectly mapped reads and of reads with imperfect mapping due to the presence of mismatches, indels or a combination of both. It also reports in-depth analysis of the properly, improperly and Mapped-Unmapped (MU) pairs, together with the percentage and absolute counts of Unmapped-Unmapped (UU) paired reads. Improperly mapping typically occurs when two reads of a mapped pair map too near or too far away from each other when compared to the rest of the aligned reads, when two paired reads map to different chromosomes or when the mapping pattern of the two reads is abnormal, giving rise to a Forward-Forward (FF), Reverse-Reverse (RR) or Reverse-Forward (RF) distribution instead of the correct Forward-Reverse (FR) pattern. To allow an in-depth analysis of improperly mapped reads, SAM-Profiler reports the percentage and absolute counts of improperly mapped reads mapping to the same and to different chromosomes, of FF, RR, RF reads and of too short and too long fragments. For all these data, absolute counts and percentage are reported.

Analysis of the mapping distribution in our WES dataset showed that, in line with what expected, the majority of the paired reads were perfectly mapped (76.3%). The main source of imperfectly mapping was the presence of single nucleotide mismatches (median 19.5%) throughout the whole dataset. Indels and coexisting single nucleotide mismatches/indels contributed only for 1.4 and 1.1% to the imperfectly mapped read counts. According to SAM-Profiler, the median percentage of properly mapped pairs was 82.6%, with a total of 15.6% improperly mapping, 1.2% mapped-unmapped and 0.6% completely unmapped pairs. Of the improperly mapping pairs, the vast majority (14.1%) was caused by pairs mapping to the same chromosome and 1.5% by pairs mapping to different chromosomes.

Mismatch distribution

Although the recent improvements in the sequencing chemistry allowed to significantly increase the length of the short reads that are generated in high throughput experiments, the analysis of the global and per-read mismatch distribution in our dataset revealed at least three phenomena: the first one was an increase in the absolute count of single nucleotide mismatches occurring when moving from the first nucleotides to the end of the sequences (Figure 3a), which suggests that a degradation of the sequencing quality is still detectable over time and

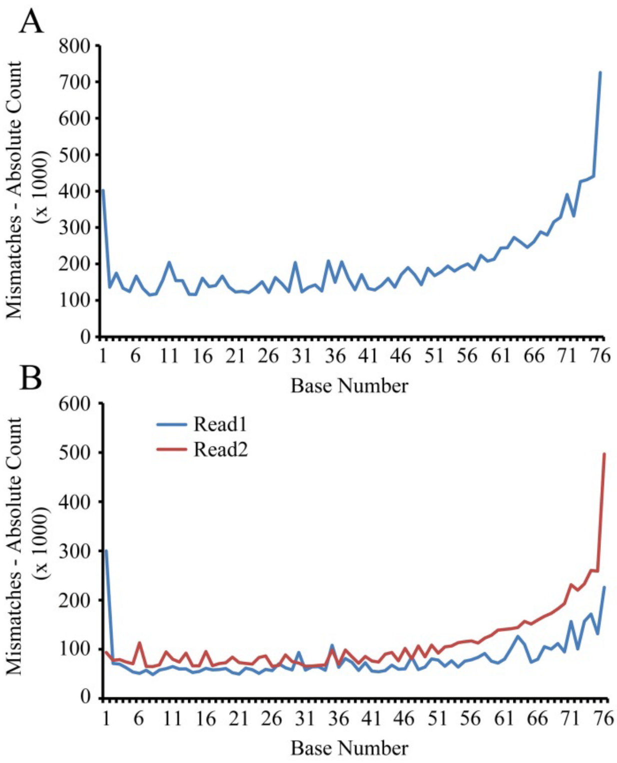

Figure 3: Mismatch analysis. A) Median of the absolute per-base mismatches count occurring in the Whole-Exome Sequencing dataset. B) Median of the absolute per-base mismatches count for the first and second read in the Whole-Exome Sequencing dataset. The blue and red lines indicate the absolute mismatch count of the first and the second read, respectively.

that caution should be taken when deciding to increase the throughput of a sequencing experiment by increasing its read length. These data are in line with the reported degradation of the read quality from the start to the end of each sequence (see Read and Mapping Quality paragraph). The second phenomenon is the global increase in the number of per-base mismatches generated in the second read (Figure 3b; p<0.0001). Similarly to the previous, this phenomenon is in line with the reported overall degradation of the read quality from the first to the second read and suggests that the decrease of the read quality, albeit limited, may be responsible for the functional increase of the error rate. The last phenomenon is a high number of mismatches occurring at the first nucleotide of the first read (Figure 3b). In CAGE experiments involving sequencing of cDNA-based libraries, a high mismatch rate occurring at the first base has been already reported [6] and it is likely due to the cap-dependent deoxycytidyl transferase activity of MMLV reverse transcriptases [7]. In WES, where no reverse transcription takes place, this result was partially unexpected although it has been reported that such event may be due to the higher handling time required when starting the sequencing run [8] (e.g. for focusing and first cycle report). This hypothesis is supported by the lower error rate detected for the first base of the second read (p=0.0079), where the time required to start the sequencing process is shorter than that of the first read. Taken together these data suggest that, in this context, trimming the leading base of the first read may improve the quality of the sequencing data.

Nucleotide and CG distribution

Nucleotide distribution analysis reports the relative distribution of the 4 bases plus N (undetermined nucleotide). As it may be expected, this analysis showed that the distribution of the four bases was very homogeneous and that the number of undetermined nucleotides was very low (<1:500) across the whole dataset. CG distribution indicated a mean CG content of 43 and 42% for read 1 and 2, respectively.

Comparison with existing tools

The availability of bioinformatics tools devoted to an in-depth quality analysis of alignment files, either in SAM or BAM format, is at the moment extremely limited. In the past several tools able to complete individual analytical or, most frequently, filtering tasks, such as TagDust [9], TagCleaner [10] or CANGS [11] have been developed. However, none of them has been built in order to provide complete alignment quality reports. Indeed, only two SAM/BAM reporting tools are actually widely used: SamStat [12] and FastQC [13]. Although both are excellent tools able to generate quality profiles of SAM/BAM files, the information provided by them is limited (Table 1). Specifically, SamStat reports about overall read quality, mapping quality and overrepresented nucleotides but no information is provided about per-base or per-chromosome mapping distribution, paired-reads-based mismatch pattern, exonic, intronic and intergenic coverage, about the presence of duplicates and the in-depth characteristics of improperly mapped reads. Fast QC reports about overall and per-base read quality, nucleotide overrepresentation, nucleotides distribution, per sequence CG distribution and provides partial support for duplicate detection (it analyzes the presence of duplicates for the first 200000 reads) but doesn't provide information about per chromosome read quality, mapping quality, per-base or per-chromosome mapping distribution, mismatch distribution, properly/improperly pairs, fragment size distribution, exonic, intronic and intergenic coverage.

Feature	SAM-Profiler	SamStat	FastQC
Operative System	L/W/M	L	L/W/M
Graphical UI	Yes	No	Yes
Programming Language	C#	C	Java
External Dependencies	.Net/Mono Runtime	No[a]	Java Runtime
Input	SAM/BAM	SAM/BAM	SAM/BAM
Batch Input	Yes	Yes	No
Interactive Graphical Output	Yes	No	No
Parallel programming	Yes	No	Yes
Read Quality	Yes	Yes	Yes
Mapping Quality	Yes	Yes	No
Per-Read Read Quality	Yes	No	No
Per-Base Read Quality	Yes	No	Yes
Per-Chromosome Read Quality	Yes	No	No
Nucleotides Overrepresentation	No	Yes	Yes
Duplicates	Yes	No	Yes[b]
Per-Chromosome Duplicates	Yes	No	No
Exonic/Intronic/Intergenic Coverage	Yes	No	No
Exonic Fold-Coverage	Yes	No	No
Fragment Size Distribution	Yes	No	No
Perfectly/Imperfectly Mapped Reads	Yes	No	No
Properly/Improperly Pairs	Yes	No	No
M/U and U/U Pairs	Yes	No	No
Improperly Mapped Reads Analysis	Yes	No	No
Mismatch Distribution	Yes	Yes	No
Per-Read CG Distribution	Yes	No	Yes
A,C,T,G,N Distribution	Yes	Yes	Yes

L/W/M: Linux/Windows/MacOS.
a: Requires installation.
b: Duplicates calculated only for the first 200000 reads.

Table 1: Summary of the main characteristics of three SAM/BAM reporting tools.

In this study we described SAM-Profiler, a bioinformatics tool dedicated to the qualitative analysis of SAM/BAM alignment data. Our tool is developed by using an entirely graphical, user-friendly interface. It doesn't require programming or scripting knowledge to be run and it is able to report a large set of qualitative data, such as overall, per-base and per-chromosome read quality, mapping quality, duplicate and coverage analyses, bases distribution, perfect, proper and improper mapping, exonic, intronic, intergenic, 5` and 3`UTR coverage, mismatch distribution, CG distribution and fragment size profile in both a textual, tab-spaced format and in a fully graphical, interactive view. To allow a smooth integration with commonly used sequencing pipelines, a command-line version of the software is also provided.

SAM-Profiler is entirely written in C# and can be run under different operative systems, such as Windows, 64 or 32 bit, Linux or MacOS. To test SAM-Profiler in the context of real high-throughput data, we used a set of 13 WES experiments. For all of them a complete quality profile was generated, which allowed us to identify a list of previously undetected problems and potential improvements to our experimental/

analytical pipeline, therefore confirming the importance of routinely using alignment reporting tools such as SAM-Profiler. Although other tools have been developed in the past to perform quality analyses on alignment files, the vast majority of them are dedicated to a single or few specific tasks, such as the identification of duplicates or the analysis of the read and mapping quality values. Unfortunately, running multiple reporting tools over the same alignment is very impractical and time-consuming, given also the huge size of the SAM/BAM files. Moreover, this approach often fails to give a broad, complete and easily accessible view of the global alignment quality. Conversely, SAM-Profiler was built to generate an extensive set of reports within a single analysis. The modular architecture of our software allows us to easily embed other analytical monitoring/reporting tools to the already developed list, allowing SAM-Profiler to grow according to the specific requests of the end-users. Therefore, we propose SAM-Profiler as flexible reporting software in the context of high-throughput sequencing alignment data. SAM-Profiler is freely available for download from: http://www.ngsbicocca.org/html/SAMProfiler.html

Acknowledgements

We gratefully thank Michela Viltadi for technical assistance. This work was supported by Progetti di Ricerca di Interesse Nazionale (PRIN) (20084XBENM_004 to RP), the Italian Association for Cancer Research (AIRC) (IG-10092 to CGP); Regione Lombardia (ID-16871 and ID-14546A to CGP), FSE “Dote ricercatori” (ID 16-AR to SR).

References

1. A Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65.
2. Shendure J, Lieberman Aiden E (2012) The expanding scope of DNA sequencing. Nat Biotechnol 30: 1084-1094.
3. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 38: 1767-1771.
4. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.
5. Piazza R, Valletta S, Winkelmann N, Redaelli S, Spinelli R, et al. (2013) Recurrent SETBP1 mutations in atypical chronic myeloid leukemia. Nat Genet 45: 18-24.
6. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, et al. (2006) Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet 38: 626-635.
7. Schmidt WM, Mueller MW (1999) CapSelect: a highly sensitive method for 5’ CAP-dependent enrichment of full-length cDNA in PCR-mediated analysis of mRNAs. Nucleic Acids Res 27: e31.
8. Kircher MU, Stenzel, Kelso J (2009) Improved base calling for the Illumina Genome Analyzer using machine learning strategies. Genome Biol 10: R83.
9. Lassmann T, Hayashizaki Y, Daub CO (2009) TagDust--a program to eliminate artifacts from next generation sequencing data. Bioinformatics 25: 2839-2840.
10. Schmieder R., Lim YW, Rohwer F, Edwards R (2010) TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets. BMC Bioinformatics 11: 341.
11. Pandey RV, Nolte V, Schlotterer C (2010) CANGS: a user-friendly utility for processing and analyzing 454 GS-FLX data in biodiversity studies. BMC Res Notes 3: 3.
12. Lassmann T, Hayashizaki Y, Daub CO (2011) SAMStat: monitoring biases in next generation sequencing data. Bioinformatics 27: 130-131.
13. Patel RK, Jain M (2012) NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoS One 7: e30619.