Bioinformatics Advance Access published June 22, 2016

Bioinformatics, 2016, 1–3
doi: 10.1093/bioinformatics/btw239
Advance Access Publication Date: 3 May 2016
Applications Note

Sequence analysis

BAM-matcher: a tool for rapid NGS sample matching

Paul P.S. Wang[1,2,*], Wendy T. Parker[1,2,3], Susan Branford[1,3,4,5] and Andreas W. Schreiber[2,4]

[1]Department of Genetics and Molecular Pathology, and [2]ACRF Cancer Genomics Facility, Centre for Cancer Biology, SA Pathology, Adelaide, Australia, [3]School of Pharmacy and Medical Science, University of South Australia, Adelaide, Australia, [4]School of Biological Sciences and [5]School of Medicine, University of Adelaide, Adelaide, Australia

*To whom correspondence should be addressed.

Received on January 11, 2016; revised on March 16, 2016; accepted on April 25, 2016

Abstract

The standard method used by high-throughput genome sequencing facilities for detecting mislabelled samples is to use independently generated high-density SNP data to determine sample identity. However, as it has now become commonplace to have multiple samples sequenced from the same source, such as for analysis of somatic variants using matched tumour and normal samples, we can directly use the genotype information inherent in the sequence data to match samples and thus bypass the need for additional laboratory testing. Here we present BAM-matcher, a tool that can rapidly determine whether two BAM files represent samples from the same biological source by comparing their genotypes. BAM-matcher is designed to be simple to use, provides easily interpretable results, and is suitable for deployment at early stages of data processing pipelines.

Availability and implementation: BAM-matcher is licensed under the Creative Commons by Attribution license, and is available from: https://bitbucket.org/sacgf/bam-matcher.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Contact: paul.wang@sa.gov.au

1 Introduction

Sample mislabelling or mix-up is a common problem, particularly in projects and facilities that have high sample throughput (Koboldt *et al.*, 2010), with several recent studies placing the error rate at around 1% (Grimm *et al.*, 2010). When working with next generation sequencing (NGS) data, a mislabelled sample can lead to incorrect data processing and analysis, resulting in conflicting results or false conclusions. Even if the error is identified, valuable data resources have typically been wasted, delaying analysis or diagnosis.

A common practice in major genome centres is to use custom high density SNP panels to generate a SNP profile for each sample (Koboldt *et al.*, 2010), and SNP panels specifically designed for sample identification have recently become available commercially. The principle behind these panels is to provide an independently generated SNP profile of each sample that can be used for sample matching.

However, as the cost of NGS technology has steadily decreased, it has become affordable and commonplace for research laboratories to sequence multiple samples from the same individual such as for analysis of somatic variants using tumour and normal samples, monitoring disease progression using longitudinal samples, or discovery of pathogenic variants in rare diseases using samples from multiple individuals in the same family. As these NGS data already contain significant amount of genotype information, we can directly use these data for sample matching without expending additional resources or sample material to perform additional profiling.

To facilitate sample matching using existing NGS data, we have developed BAM-matcher, a tool that provides rapid pair-wise comparison of binary Sequence Alignment/Map (BAM) files (http://samtools.github.io/hts-specs) by comparing the sample genotypes at pre-determined genomic locations. BAM-matcher has the following features: first, it is easy to use and can be deployed at early stages of processing pipelines. Second, BAM-matcher is very fast: by limiting genotype-calling to predetermined positions, a comparison between two samples can be made in ~2 min using the provided default set of variants (Intel Xeon E5, 3.6 GHz). As BAM-matcher caches sample genotype data, subsequent comparisons involving previously calculated samples can be much faster (~1 s), thus significantly reducing overall processing time for large cohorts. Third, BAM-matcher is flexible; it can compare different types of NGS data, including whole-genome sequencing (WGS), whole-exome sequencing (WES) and RNA-sequencing (RNA-seq) data. If an appropriate genome reference and suitable list of SNPs are provided, BAM-matcher can also be used for non-human genomes.

2 Implementation and features

BAM-matcher is a Python command line tool (Python v2.7) for Linux operating systems. It relies on external third party tools for genotype calling, and currently supports the Genome Analysis Toolkit (McKenna *et al.*, 2010) and Freebayes (Garrison et al., 2012).

The only sample-specific input data required by BAM-matcher are mapped read data, thus it can be used at early stages of processing pipelines to facilitate early detection of sample mislabelling. By default, BAM-matcher compares sample genotypes at 1500 exonic SNP sites extracted from the 1000 Genomes database (The 1000 Genomes Project Consortium, 2012) with global minor allele frequencies between 0.45 and 0.55. If required, users can also substitute a customised list of loci.

BAM-matcher can write the output report (Supplementary Figure S1) to a file (TXT or HTML) or to the command line. A short-form tab-separated output is also available, useful for batch processing. Although 1500 SNPs were used for comparison, genotype comparison at any site is only carried out if the coverage is above a depth threshold (default 15) in both samples, thus the reported number of sites compared is typically fewer than 1500. The sub-classification of genotype discordance is particularly useful when comparing WGS or WES data against RNA-seq data because the latter can involve allele-specific expression (Supplementary Text S1).

3 Results

To demonstrate the utility of BAM-matcher, we examined three sets of NGS data: (i) WES data of 120 samples from 39 patients with at least two samples from each individual, (ii) RNA-seq data of 63 samples from the same group of patients as (i), (iii) WES data of 53 samples belonging to 48 individuals from eight families. NimbleGen SeqCap EZ Exome kit was used for exome enrichment in the WES samples, and Illumina Ribo-Zero rRNA removal kit was used for rRNA depletion in the RNA-seq samples. All samples were sequenced on Illumina HiSeq platforms. Research involving human tissue was conducted with institutional ethics review board approval and in conformance with the Declaration of Helsinki.

Results from WES-versus-WES comparison (Fig. 1A) show a clear demarcation between pairs of samples from the same individual (matching) and from different individuals (non-matching). Matching samples shared >95% identical genotype calls, which reduced to ~80–95% if one of the samples involved had low coverage depth (marked 'LD', with average coverage of ~20× compared to typical ~70× coverage). Non-matching samples only had about ~30–50% common genotype calls. These results are consistent with our choice of global allele frequencies of loci to compare (Supplementary Text S2), and similar to the performance reported for high-density SNP array (Koboldt *et al.*, 2010).

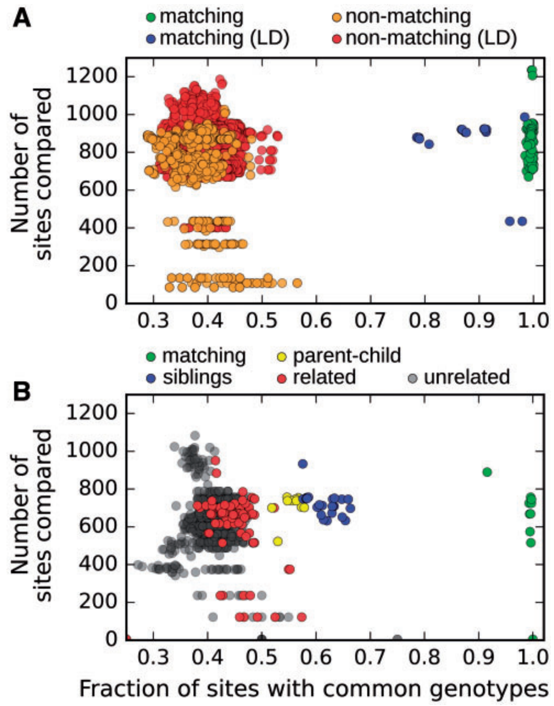

Fig. 1. Scatter plots showing results of **(A)** 7140 unique pair-wise comparisons between 120 BAM files representing samples from 39 individuals ('LD', low depth, denotes comparisons involving samples with significantly lower coverage than expected), **(B)** 1378 pair-wise comparisons between 53 BAM files representing samples belonging to 48 individuals from 8 families. The familial relationships between samples are classified into 'siblings', 'parent–child', 'related' and 'unrelated'. 'Related' includes all familial relationships other than 'siblings' and 'parent–child'

BAM-matcher is also able to match WES samples with RNA-seq samples (Supplementary Figure S2). However, the levels of genotype identity are generally lower between matching WES-RNA-seq samples than matching WES samples due to allele-specific expression in RNA-seq data, and more careful examination of the results may be necessary for some samples (Supplementary Text S1).

BAM-matcher can also identify familial relationships between samples, as shown by the results for familial samples (Fig. 1B). For this type of application, better separation between different groups can be achieved if more sites are used (Supplementary Figure S3).

Acknowledgements

We thank David Yeung, Doris Stangl, Professor Hamish Scott and the patients involved for providing the sequencing data used in this study.

Funding

This work was funded by the National Health and Medical Research Council of Australia (APP 1027531).

Conflict of Interest: none declared.

References

Garrison,E. and Marth,G. (2012) Haplotype-based variant detection from short-read sequencing. *arXiv*, **1207.3907 [q-bio.GN]**.

Grimm,E. *et al*. (2010) Blood bank safety practices: mislabeled samples and wrong blood in tube—a Q-probes analysis of 122 clinical laboratories. *Arch. Pathol. Lab. Med*., **134**, 1108–1115.

Koboldt,D.C. *et al*. (2010) Challenges of sequencing human genomes. *Brief Bioinform*., **11**, 484–498.

McKenna,A. *et al*. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*., **20**, 1297–1303.

The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.