BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 9 2013, pages 1210–1211
doi:10.1093/bioinformatics/btt118

Sequence analysis

Advance Access publication March 6, 2013

MitoSeek: extracting mitochondria information and performing high-throughput mitochondria sequencing analysis

Yan Guo[1,*], Jiang Li[1], Chung-I Li[1], Yu Shyr[1] and David C. Samuels[2]

[1]Center for Quantitative Sciences, Vanderbilt University, 2220 Pierce Ave, 571 Preston Research Building and [2]Center for Human Genetics Research, Vanderbilt University, 507B Light Hall, Nashville, TN 37232, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Exome capture kits have capture efficiencies that range from 40 to 60%. A significant amount of off-target reads are from the mitochondrial genome. These unintentionally sequenced mitochondrial reads provide unique opportunities to study the mitochondria genome.

Results: MitoSeek is an open-source software tool that can reliably and easily extract mitochondrial genome information from exome and whole genome sequencing data. MitoSeek evaluates mitochondrial genome alignment quality, estimates relative mitochondrial copy numbers and detects heteroplasmy, somatic mutation and structural variants of the mitochondrial genome. MitoSeek can be set up to run in parallel or serial on large exome sequencing datasets.

Availability: https://github.com/riverlee/MitoSeek

Contact: yan.guo@vanderbilt.edu

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

Received on December 18, 2012; revised on February 15, 2013; accepted on March 1, 2013

1 INTRODUCTION

Next-generation sequencing (NGS) has enabled high-throughput production of sequencing data at a low cost. Projects such as The Cancer Genome Atlas (TCGA) and the 1000 Genomes Project have generated huge amounts of sequencing data. NGS data are rich and informative and contain many off-target reads that are often ignored but which may be biologically relevant. Sequencing data outside capture regions can produce reliable variation data (Guo *et al.*, 2012c). Mitochondrial DNA (mtDNA) sequences are recoverable in exome sequencing data (Larman *et al.*, 2012), even when the mtDNA is not included in the target region. Picardi and Pesole (2012) also provided scripts for assembling mitochondrial genome exome sequencing data. Based on those findings, we designed and implemented a tool, MitoSeek, for high-throughput secondary mitochondrial data mining from exome sequencing data or whole genome sequencing data. MitoSeek extracts mitochondrial information from exome sequencing data and performs analyses on four major mitochondrial factors: heteroplasmy, somatic mutations, relative copy number variation and large structural changes.

*To whom correspondence should be addressed.

2 METHODS

2.1 Mitochondrial sequence extraction

Compared with the scripts released by Picardi and Pesole (2012), which extract mtDNA reads and reassemble the mitochondrial genome from exome sequencing data, MitoSeek can extract mitochondrial genome information directly from a BAM file and also perform mitochondrial genome assembly. To deal with mtDNA homologous regions in the nuclear genome, MitoSeek uses a conservative approach, which uses reads unmapped to nuclear genome for the mtDNA assembly. However, choosing to reassemble the mitochondrial genome requires a significantly longer running time.

2.2 Quality control

Before conducting any analysis, MitoSeek will first produce a mitochondrial alignment quality control report, which contains important statistics such as average depth, percent of base pairs covered, base quality distribution, mapping quality distribution and insert size distribution. These quality control parameters serve as important confidence indicators for the downstream mitochondrial analysis. MitoSeek filters reads based on mapping quality score (MQ $\geq$ 20) and base quality score (BQ $\geq$ 20). The threshold of the filters is adjustable by the user.

2.3 Heteroplasmy detection

The most crucial factor for detecting heteroplasmy is depth. The ideal sequencing technique for detecting heteroplasmy in mitochondria is mitochondria-targeted sequencing, which is capable of generating depths of up to 10 000 and detecting heteroplasmy as low as 0.1%. The depth for mtDNA in exome sequencing data is significantly lower, which limits the detectable heteroplasmy to ~1%. Based on the alignment quality control report, MitoSeek will automatically adjust the heteroplasmy detection threshold to the most appropriate level. The heteroplasmy detection threshold is defined on one of two scales: read count or read percentage at a given location. For example, a user can specify the number of raw reads that are required to show support for heteroplasmy, or the percentage of reads that are required to show support for heteroplasmy. The heteroplasmy empirical filters follow Guo *et al.* (2012a).

In addition to the empirical filters, we implemented a statistical framework to assess heteroplasmy. MitoSeek performs a one-tail Fisher's exact test to determine if the rate of heteroplasmy at each site is greater than zero or a user-defined threshold. Phred quality scores of heteroplasmy are also reported by MitoSeek.

2.4 Somatic mutation detection

Current genotype callers such as GATK's Unified Genotyper (McKenna *et al.*, 2010) and glfMultiple are designed for a diploid genome. Using those genotype callers on a haploid genome where only a single allele is

expected will generate inaccurate results. To solve this problem, MitoSeek compares the empirical allele counts between tumor and normal samples directly instead of using a genotype caller. MitoSeek can extract empirical allele counts for every mitochondrial position and then compare the allele counts between tumor and normal to determine somatic mutation status.

2.5 Relative copy number estimation

MtDNA copy number can be obtained through NGS data (Castle *et al.*, 2010). MitoSeek estimates the relative mtDNA copy number through exome and whole genome sequencing data. These are not absolute measurements of mtDNA copy number. The method takes advantage of the proportion of mitochondrial reads captured during exome and whole genome sequencing. The relative mtDNA copy number is computed as $CN = \frac{R_m}{2R_T}$, where R_m is the reads aligned to mitochondria that passed the quality filter and R_T is the total reads that passed the quality filter. Alternatively, relative mtDNA copy number can be computed as $CN = \frac{\overline{D_m}}{2\overline{D_E}}$, where $\overline{D_m}$ is the average depth of mtDNA, and $\overline{D_E}$ is the average depth of the exome. If whole genome data are used, $\overline{D_E}$ will be replaced with $\overline{D_T}$, where $\overline{D_T}$ is the average depth across the whole genome excluding mitochondria.

2.6 Structural mtDNA change

MitoSeek also reports mtDNA structural changes when pair-end sequencing data are used. The structural changes include mitochondria-nuclear genome integration and large deletion in mitochondria. Nuclear genome integration is a known phenomenon, which has been documented by multiple studies (Mourier *et al.*, 2001; Timmis *et al.*, 2004), and it can be detected through discordant read pairs. Large deletion in mtDNA is a well-studied mtDNA dysfunction (Chen *et al.*, 2011), and MitoSeek detects it through identifying abnormal insert sizes.

2.7 Other features

The most common sequencing alignment reference for the human is HG19; however, the most accepted mitochondria reference is the revised Cambridge Reference sequence (rCRS) (Andrews *et al.*, 1999) (GenBank: NC_012920). MitoSeek can interchange genomic positions and reference nucleotides between HG19 and rCRS. Annotation of amino acid changes and identification of known pathogenic mutations are reported by MitoSeek in both HG19 and rCRS coordinates.

3 RESULT

We downloaded exome sequencing data on 10 normal-paired breast cancer tumor samples from TCGA and tested MitoSeek using those data. An example of heteroplasmy identified by MitoSeek can be viewed in Figure 1. An example of a full report from a MitoSeek run can be seen in the Supplementary Material. MitoSeek is written in Perl and Linux Shell Script. The typical run time on exome sequencing data with BAM file size of 11 GB is ~50 min on a 2.4 GHz CPU with 1 GB of memory.

4 DISCUSSION

Owing to the limitation of exome sequencing data, MitoSeek is not capable of calculating absolute copy number of mtDNA, only relative mtDNA copy number. Also, owing to noise in the sequencing data, MitoSeek is more suited to detecting large copy number variation rather than small copy number variation. MitoSeek is designed to work with paired-end sequencing data rather than single-end sequencing data given the large abundance of paired-end sequencing data. MitoSeek is designed with accessibility in mind. It considers many unique parameters that have not been implemented by other genome analysis tools. It is the only sequencing analysis tool that reports allele counts separately by forward and reverse strands, critical information for assessing strand bias (Guo *et al.*, 2012b). MitoSeek creates opportunities for high-throughput mitochondrial sequencing data mining from existing large exome sequencing databases.

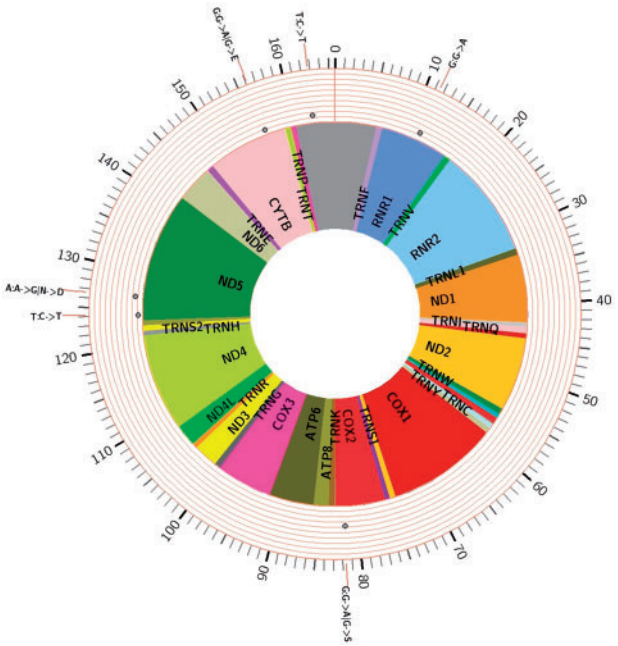

Fig. 1. An example of MitoSeek's somatic mutation graphical output

Conflict of Interest: none declared.

REFERENCE

Andrews,R.M. *et al.* (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat. Genet.*, **23**, 147.

Castle,J.C. *et al.* (2010) DNA copy number, including telomeres and mitochondria, assayed using next-generation sequencing. *BMC Genomics*, **11**, 244.

Chen,T. *et al.* (2011) The generation of mitochondrial DNA large-scale deletions in human cells. *J. Hum. Genet.*, **56**, 689–694.

Guo,Y. *et al.* (2012a) The use of next generation sequencing technology to study the effect of radiation therapy on mitochondrial DNA mutation. *Mutat. Res.*, **744**, 154–160.

Guo,Y. *et al.* (2012b) The effect of strand bias in Illumina short-read sequencing data. *BMC Genomics*, **13**, 666.

Guo,Y. *et al.* (2012c) Exome sequencing generates high quality data in non-target regions. *BMC Genomics*, **13**, 194.

Larman,T.C. *et al.* (2012) Spectrum of somatic mitochondrial mutations in five cancers. *Proc. Natl Acad. Sci. USA*, **109**, 14087–14091.

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

Mourier,T. *et al.* (2001) The human genome project reveals a continuous transfer of large mitochondrial fragments to the nucleus. *Mol. Biol. Evol.*, **18**, 1833–1837.

Picardi,E. and Pesole,G. (2012) Mitochondrial genomes gleaned from human whole-exome sequencing. *Nat. Methods*, **9**, 523–524.

Timmis,J.N. *et al.* (2004) Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat. Rev. Genet.*, **5**, 123–135.