BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 21 2014, pages 3115–3117
doi:10.1093/bioinformatics/btu483

Genome analysis

Advance Access publication July 14, 2014

MToolBox: a highly automated pipeline for heteroplasmy annotation and prioritization analysis of human mitochondrial variants in high-throughput sequencing

Claudia Calabrese[1,†], Domenico Simone[2,†], Maria Angela Diroma[3], Mariangela Santorsola[4], Cristiano Guttà[3], Giuseppe Gasparre[1], Ernesto Picardi[3,5,6], Graziano Pesole[3,5,7] and Marcella Attimonelli[3,*]

[1]Department of Medical and Surgical Sciences, University of Bologna, 40138 Bologna, [2]Department of Biosciences, University of Milan, 20133 Milan, [3]Department of Biosciences, Biotechnologies and Biopharmaceutics, University of Bari, 70126 Bari, [4]Department of Sciences and Technologies, University of Sannio, 82100 Benevento, [5]Institute of Biomembranes and Bioenergetics, National Research Council, 70126 Bari, [6]National Institute of Biostructures and Biosystems, 00136 Rome and [7]Center of Excellence in Genomics for Biomedicine and Agri-food (CEGBA), University of Bari, 70126 Bari, Italy

Associate Editor: Inanc Birol

ABSTRACT

Motivation: The increasing availability of mitochondria-targeted and off-target sequencing data in whole-exome and whole-genome sequencing studies (WXS and WGS) has risen the demand of effective pipelines to accurately measure heteroplasmy and to easily recognize the most functionally important mitochondrial variants among a huge number of candidates. To this purpose, we developed MToolBox, a highly automated pipeline to reconstruct and analyze human mitochondrial DNA from high-throughput sequencing data.

Results: MToolBox implements an effective computational strategy for mitochondrial genomes assembling and haplogroup assignment also including a prioritization analysis of detected variants. MToolBox provides a Variant Call Format file featuring, for the first time, allele-specific heteroplasmy and annotation files with prioritized variants. MToolBox was tested on simulated samples and applied on 1000 Genomes WXS datasets.

Availability and implementation: MToolBox package is available at https://sourceforge.net/projects/mtoolbox/.

Contact: marcella.attimonelli@uniba.it

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

Received on March 19, 2014; revised on July 8, 2014; accepted on July 9, 2014

1 INTRODUCTION

Emerging discoveries in human mitochondrial genetics, driven by the advent of next-generation sequencing, have revealed that individuals exhibit a complex mixture of mitochondrial genotypes (He *et al.*, 2010) and carry low-level heteroplasmic variants (Payne *et al.*, 2013). On the other hand, the deeper the sequencing coverage, the higher the number of mitochondrial DNA (mtDNA) variants and the variety of heteroplasmic ranges found *per* individual (Diroma *et al.*, 2014; He *et al.*, 2010; Payne *et al.*, 2013). In this frame, the deep sequencing of mtDNA raises the demand of effective pipelines to accurately measure heteroplasmy and to easily recognize the most functionally important variants among a huge number of candidates. To this purpose, we developed MToolBox, a highly automated bioinformatics pipeline to reconstruct and analyze human mtDNA from high-throughput sequencing (HTS) data. The MToolBox workflow includes a computational strategy to assemble mitochondrial genomes from whole-exome sequencing (WXS) and/or whole-genome sequencing (WGS) data (Picardi and Pesole, 2012), which was further updated to detect insertions and deletions (ins/dels) and to assess the heteroplasmic fraction (HF) of each variant allele with the related confidence interval (CI), reported as sample-specific meta-information in an enhanced version of the Variant Call Format (VCF) file (version 4.0). The MToolBox pipeline analyzes the reconstructed genomes for haplogroup assignment (Rubino *et al.*, 2012) and variant prioritization.

2 METHODS

2.1 Mitochondrial reads extraction, genome reconstruction and VCF file generation

The MToolBox pipeline integrates in a unique automatic workflow a computational strategy for mtDNA data extraction from WXS and WGS data (Picardi and Pesole, 2012), where new important features have been added. MToolBox can accept as input raw data or prealigned reads (Fig. 1a). In both cases, reads are mapped/remapped by the *mapExome.py* script (Fig. 1b) at user's choice either onto the Reconstructed Sapiens Reference Sequence (RSRS; Behar *et al.*, 2012) or the revised Cambridge Reference Sequence (rCRS; Andrews *et al.*, 1999). Subsequently, reads mapped on mtDNA are realigned onto the nuclear genome (GRCh37/hg19), to discard Nuclear mitochondrial Sequences (NumtS; Simone *et al.*, 2011; Fig. 1c) and amplification artifacts. The resulting Sequence Alignment/Map (SAM) file (Fig. 1d) can be optionally processed for ins/dels realignment around a set of

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

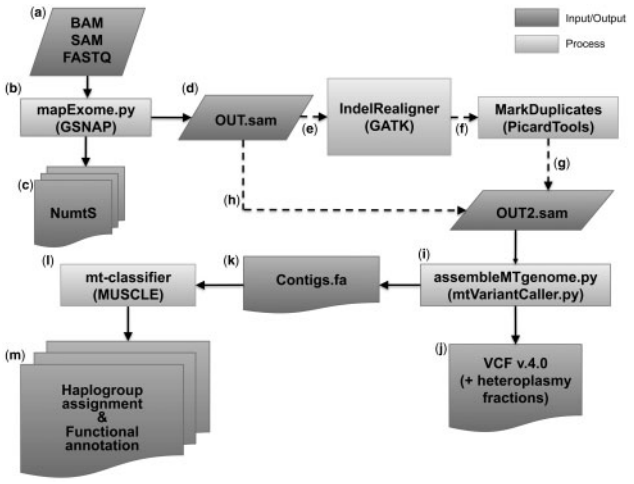

Fig. 1. The main steps of the MToolBox workflow: (a–d) read mapping and NumtS filtering; (e–h) post-mapping processing; (i–m) genome assembly, haplogroup prediction and variant annotation. In brackets, programs or modules particularly important for the associated process. Solid connectors indicate mandatory pipeline steps; dashed connectors (**e–g**) indicate that the corresponding post-mapping steps can be optional, otherwise the OUT2.sam file directly undergoes the assembly process (**h**). Please refer to Supplementary Information for a detailed description of MToolBox workflow steps

known ins/dels, annotated in HmtDB (Rubino *et al*., 2012) and MITOMAP (Ruiz-Pesini *et al*., 2007), and for putative PCR duplicates removal (Fig. 1e–h and Supplementary Information). This step generates a dataset of highly reliable mitochondrial aligned reads, which is used to reconstruct a complete mitochondrial genome by the *assembleMTgenome.py* script (Fig. 1i), now integrating the *mtVariantCaller.py* module for nucleotide mismatches and ins/dels detection. All the genomic variants are filtered based on the quality scores and read depth, and annotated in a VCF file (v.4.0), with the corresponding HF and CI values (Fig. 1j and Supplementary Information).

2.2 Haplogroup prediction and prioritization analysis of mitochondrial variants

MToolBox provides an output file with reconstructed contig sequence(s) (*Contigs.fa*) (Fig. 1k and Supplementary Information). Each set of contigs is subjected to haplogroup prediction, relying on the RSRS-based Phylotree resource (van Oven and Kayser, 2009), by *mt-classifier* (Fig. 1l), an updated version of the *fragment-classify* tool (Rubino *et al*., 2012), which now includes a module to perform functional annotation and prioritization of mitochondrial variants (Fig. 1m and Supplementary Information). This latter analysis is carried out by aligning each sample-specific reconstructed contig against the related macro-haplogroup-specific consensus sequence (Supplementary Information) to recognize, *via* a prioritization process, private variants, deserving further clinical investigation. The prioritization takes into account also the pathogenicity of each mutated allele, determined with different algorithms, and the nucleotide variability of each variant site; amino acid variability is also considered if the variant site is codogenic (Supplementary Information). For each mutated allele, additional annotations are also reported, i.e. annotation from HmtDB and MITOMAP resources and their occurrence among 1000 Genomes Project samples (Supplementary Information). Variants of assembled genomes are also reported with respect to rCRS (Supplementary Information), to ensure a full compatibility of the resulting annotation with the current clinical literature (Bandelt *et al*., 2014).

3 RESULTS

The MToolBox performance in heteroplasmy detection was tested on four artificial heteroplasmic samples, whose sequencing was simulated at different mean depth (Supplementary Information). MToolBox showed high specificity and sensitivity in detecting all the artificial heteroplasmy tested, with an average coverage depth equal or above 1000×. MToolBox was extensively applied on WXS data from 1000 Genomes (Genomes Project *et al*., 2012 and Supplementary Information), to obtain a VCF file of mtDNA variants from 2419 individuals (available at https://sourceforge.net/projects/mtoolbox/files/1000Genomes_data/). Reliability of reconstructed mitochondrial genomes was confirmed by their haplogroup predictions, the majority of which coherent with the ancestry of the related individual (Supplementary Information). The accuracy in heteroplasmy detection and quantification was confirmed by the results from four mother–child pairs that showed the expected pattern of mtDNA inheritance (Supplementary Information).

4 DISCUSSION

A highly automated pipeline for mtDNA analysis from HTS data is not available to date. To fill this gap, we developed MToolBox, an effective workflow with customizable parameters and able to analyze multiple samples in a single run. MToolBox is the only tool that generates as output a VCF file, the standard format for large-scale genotyping information, suitably customized for mitochondrial data, by including the heteroplasmy fraction and its related CI. In fact, also the MitoSeek tool (Guo *et al*., 2013) performs mitochondrial HTS data analyses, including somatic and structural variant recognition. Additionally, MToolBox provides the user with essential analyses of reconstructed mitochondrial genomes, i.e. haplogroup assignment and variant prioritization, exploiting a broad collection of annotation resources. Thus, MToolBox may provide a valuable support for the recognition of candidate mitochondrial mutations in clinical studies.

Funding: This work was supported by Progetto Strategico 'Invecchiamento' e 'Medicina Personalizzata' (CNR, Italy) and the PRIN2009 fund assigned to M.A. The computational work has been executed on the IT resources made available by the ReCaS project (PONa3_00052).

Conflicts of interest: none declared.

REFERENCES

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

Bandelt,H.J. *et al*. (2014) The case of the continuing use of the revised Cambridge Reference Sequence (rCRS) and the standardization of notation in human mitochondrial DNA studies. *J. Hum. Genet*., **59**, 66–77.

Behar,D.M. *et al*. (2012) A "Copernican" reassessment of the human mitochondrial DNA tree from its root. *Am. J. Hum. Genet*., **90**, 675–684.

Diroma,M.A. *et al.* (2014) Extraction and annotation of human mitochondrial genomes from 1000 Genomes Whole Exome Sequencing data. *BMC Genomics.*, **15** (**Suppl. 3**), S2.

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

Guo,Y. *et al.* (2013) MitoSeek: extracting mitochondria information and performing high-throughput mitochondria sequencing analysis. *Bioinformatics*, **29**, 1210–1211.

He,Y. *et al.* (2010) Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. *Nature*, **464**, 610–614.

Payne,B.A. *et al.* (2013) Universal heteroplasmy of human mitochondrial DNA. *Hum. Mol. Genet.*, **22**, 384–390.

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

Rubino,F. *et al.* (2012) HmtDB, a genomic resource for mitochondrion-based human variability studies. *Nucleic Acids Res.*, **40**, D1150–D1159.

Ruiz-Pesini,E. *et al.* (2007) An enhanced MITOMAP with a global mtDNA mutational phylogeny. *Nucleic Acids Res.*, **35**, D823–D828.

Simone,D. *et al.* (2011) The reference human nuclear mitochondrial sequences compilation validated and implemented on the UCSC genome browser. *BMC Genomics*, **12**, 517.

van Oven,M. and Kayser,M. (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.*, **30**, E386–E394.