Bioinformatics, 33(9), 2017, 1399–1401
doi: 10.1093/bioinformatics/btw835
Advance Access Publication Date: 29 January 2017
Applications Note

OXFORD

Sequence analysis

fastMitoCalc: an ultra-fast program to estimate mitochondrial DNA copy number from whole-genome sequences

Yong Qian[1], Thomas J. Butler[1], Krista Opsahl-Ong[1], Nicholas S. Giroux[1], Carlo Sidore[2], Ramaiah Nagaraja[1], Francesco Cucca[2], Luigi Ferrucci[3], Gonçalo R. Abecasis[4], David Schlessinger[1] and Jun Ding[1,*]

[1]Laboratory of Genetics and Genomics, National Institute on Aging, NIH, Baltimore, MD, USA, [2]Istituto di Ricerca Genetica e Biomedica, Consiglio Nazionale delle Ricerche, Monserrato, Cagliari, Italy, [3]Translational Gerontology Branch, National Institute on Aging, NIH, Baltimore, MD, USA and [4]Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Received on July 21, 2016; revised on December 11, 2016; editorial decision on December 27, 2016; accepted on December 31, 2016

Abstract

Mitochondrial DNA (mtDNA) copy number is tightly regulated in tissues, and is both a critical determinant of mitochondrial function and a potential biomarker for disease. We and other groups have shown that the mtDNA copy number per cell can be directly estimated from whole-genome sequencing. The computation is based on the rationale that sequencing coverage should be proportional to the underlying DNA copy number for autosomal and mitochondrial DNA, and most computing time is spent calculating the average autosomal DNA coverage across ~3 billion bases. That makes analyzing tens of thousands of available samples very slow. Here we present *fastMitoCalc*, which takes advantage of the indexing of sequencing alignment files and uses a randomly selected small subset (0.1%) of the nuclear genome to estimate autosomal DNA coverage accurately. It is more than 100 times faster than current programs. *fastMitoCalc* also provides an option to estimate copy number using a single autosomal chromosome, which could also achieve high accuracy but is slower. Using *fastMitoCalc*, it becomes much more feasible now to conduct analyses on large-scale consortium data to test for association of mtDNA copy number with quantitative traits or nuclear variants.

Availability and Implementation: *fastMitoCalc* is available at https://lgsun.irp.nia.nih.gov/hsgu/software/mitoAnalyzer/index.html

Contact: jun.ding@nih.gov

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

1 Introduction

The mitochondrial cellular power plant is partially encoded in its own constituent circular genome of ~16kb—mitochondrial DNA (mtDNA). In humans, every cell has 100–10 000 copies of mtDNA. The copy number is tightly regulated in various tissues, and is both a critical determinant of the level of mitochondrial function and a potential biomarker for various diseases. For example, elevated mtDNA copy number has been reported to be associated with cancer risk (Lan *et al.*, 2008; Thyagarajan *et al.*, 2013) and major depression (Cai *et al.*, 2015b).

With the collection of whole-genome sequencing data in large-scale studies, we and others (Cai *et al.*, 2015a, b; Ding *et al.*, 2015;

Table 1. Estimating mtDNA copy number using randomly selected autosomal DNA fragments, individual chromosomes and whole genome

# of fragments	Fragment length (bp)	Total length	Time	R^2	R^2 Range in 20 simulations
1	100	0.1K	4.6 s	0.154	[0.095, 0.243]
100	1	0.1K	6.6 s	0.932	[0.900, 0.945]
1	1000	1K	4.6 s	0.311	[0.183, 0.480]
1000	1	1K	22 s	0.989	[0.975, 0.993]
300	1000	300K	12 s	0.991	[0.984, 0.996]
10 000	100	1M	96 s	0.999	[0.998, 0.999]
3000	**1000**	**3M**	**59 s**	**0.998**	**[0.997, 0.999]**
30 000	1000	30M	386 s	0.999	[0.999, 0.999]
Chromosome 20		63M	6 min	0.981	N/A
Chromosome 1		249M	23 min	0.999	N/A
Whole Genome		~3B	3 h	1.0	N/A

The different options are ordered by the total length of autosomal DNA being considered in the analysis from the smallest to the largest. The option in bold is selected as the default option in *fastMitoCalc*.

Chu *et al.*, 2012; Wachsmuth *et al.*, 2016) have shown that the mtDNA copy number per cell can be directly estimated from whole-genome sequencing. The computation is based on the rationale that average sequencing coverage should be proportional to underlying DNA copy number for autosomal and mitochondrial DNA. Current programs (e.g. our program *mitoCalc*) infer the average autosomal coverage across the entire genome, so that about 3 h are required to calculate the average for one sample with low average coverage at 4X using one AMD Opteron 2 GHz CPU on a Linux workstation with 32GB memory. That makes it awkward to investigate the tens of thousands of samples already available in the public domain and the further large-scale population data coming online.

Here we present an improved version of our program, '*fastMitoCalc*', that is more than two orders of magnitude faster. It takes advantage of the indexing of sequencing alignment files, focusing on a small subset of the nuclear genome to estimate autosomal DNA coverage accurately (correlation > 0.999 with the full genome estimate). Consequently, a computer cluster with 500 CPUs can now finish analyzing 50 000 deeply-sequenced samples [the current sample size for NHLBI Trans-Omics for Precision Medicine (TOPMed) program] in less than a day rather than the months originally required. As a result, analyses can be conducted more efficiently to test for association of mtDNA copy number with quantitative traits or to look for variants that regulate mtDNA copy number.

2 Implementation

Current programs, including *mitoCalc*, use the following formula to infer the mtDNA copy number (see Ding *et al.*, 2015 for details):

$$mtDNA \text{ copy number} = \frac{mtDNA \text{ average coverage}}{\text{autosomal } DNA \text{ average coverage}} \times 2$$

This holds if, as observed, autosomal and mtDNA are sequenced at comparable intrinsic efficiencies. Current programs obtain the coverage estimate at each base in the genome from the aligned bam files (Li *et al.*, 2009) and then calculate the average coverages for autosomal DNA and mtDNA accordingly. Because the mitochondrial genome is comparatively small, most computing time is spent calculating the average autosomal DNA coverage. Here, we propose two strategies using a small proportion of autosomal DNA to get a comparably accurate estimate: (i) using a single individual chromosome and (ii) using randomly selected autosomal DNA fragments.

2.1 Using individual chromosomes to estimate copy number

One straightforward way to improve speed is to use one chromosome to represent the nuclear genome. We estimated copy number using each individual chromosome from 1 to 22 (X and Y sex chromosomes were excluded to avoid different treatment for males and females), and compared it to the whole-genome estimate regarded as the standard. The estimates were calculated for 400 SardiNIA project participants with low-pass sequencing data (Ding *et al.*, 2015). We used the square of the correlation (R^2) between the two estimates as the measure of concordance. Table 1 lists the concordance and computing time for two representative chromosomes (1 and 20), while Supplementary Table S1 provides information for all chromosomes. Most chromosomes (16 of 22) provide accurate estimates (with concordance $R^2 > 0.99$), but several provide low-concordance estimates (e.g. chromosomes 19 and 22 have estimates with $R^2 < 0.90$; a likely explanation is in Supplementary Information). Using any chromosome to estimate copy number significantly increases computing speed (Supplementary Table S1), but, as we show below, we can achieve the same estimation accuracy while accelerating computing by using randomly selected fragments.

2.2 Using randomly selected autosomal DNA fragments to estimate copy number

In the second strategy, we randomly pick a certain number of autosomal DNA fragments (each with a certain length) to represent the autosomal DNA genome. We have assessed options ranging from very aggressive [e.g. the use of one randomly selected 100-base fragment (considering only 0.0000033% of the genome)] to the less aggressive use of 30 000 fragments, each with 1000 bases (considering 1% of the genome). Table 1 lists the computing time for each option, with the average R^2 and its range for 20 simulations of randomly selecting DNA fragments.

When considering one 100-base fragment, the analysis of a bam file (4X average coverage) takes on average only 4.6 s, but the average R^2 is also very poor, at 0.15. By contrast, considering 100 1-base fragments, the analysis takes 6.6 s and the average R^2 improves to 0.93. Similarly, we see a significant increase in accuracy from one 1000-base fragment to 1000 1-base fragments. These results show that randomly selecting multiple short fragments across the genome is much better than selecting one long fragment at one place. The concordance further improves with an increasing percentage of genome considered, so that when considering 3000 fragments, each with 1000 bases (0.10% of the genome), the analysis

finishes in 59 s with very high accuracy ($R^2 > 0.998$). We choose this as the default option for *fastMitoCalc*, striking a balance between accuracy and speed. It is worth noting that it takes almost same time (4.6s on average) to analyze one 100-base fragment or one 1000-base fragment. This is because the aligned bam files are commonly sorted and indexed, so that once the start position is identified in a bam file, it takes almost no time to scan the consecutive region for coverage information. The default option runs very fast for the same reason. The analyses here are based on low-pass (4X average coverage) sequencing data; but as we show in Supplementary Information, the impact of sequencing coverage on copy number estimates should be minimal.

2.3 Other options for *fastMitoCalc*

Besides the default option, users can choose alternatives for randomly selected fragments (not limited to the options listed in Table 1). Per requests from users, we also provide in *fastMitoCalc* the option of using individual chromosomes to estimate copy number, which will be useful when researchers have sequence data for only one specific chromosome available in a bam file. Users should then avoid chromosomes with high GC content (see Supplementary Information). We also add an option that allows users to specify a list of nuclear DNA and/or mitochondrial DNA regions to be used to estimate copy number, giving further flexibility. This last option can be applied to whole-exome sequencing data to estimate mtDNA copy number based on off-target reads. More detailed discussion and preliminary results are in Supplementary Information.

3 Conclusions

fastMitoCalc can estimate mtDNA copy number highly accurately using 0.1% of the genome, and hence speed up the estimation ~180 fold compared to current programs. The program can thus easily analyze hundreds of thousands of genomes currently being sequenced by large research consortia, thereby facilitating association studies of mtDNA copy number with quantitative trait values or nuclear variants.

Funding

This research was supported by the Intramural Research Program of the National Institute on Aging, NIH with grants Z01-AG000675 (to D.S.) and Z01-AG000693 (to J.D.).

Conflict of Interest: none declared.

References

Cai,N. *et al.* (2015a) Genetic control over mtDNA and its relationship to major depressive disorder. *Curr. Biol.*, **25**, 3170–3177.

Cai,N. *et al.* (2015b) Molecular signatures of major depression. *Curr. Biol.*, **25**, 1146–1156.

Chu,H.T. *et al.* (2012) Quantitative assessment of mitochondrial DNA copies from whole genome sequencing. *BMC Genomics*, **13**, S5.

Ding,J. *et al.* (2015) Assessing mitochondrial DNA variation and copy number in lymphocytes of ~2,000 Sardinians using tailored sequencing analysis tools. *PLoS Genet.*, **11**, e1005306.

Lan,Q. *et al.* (2008) A prospective study of mitochondrial DNA copy number and risk of non-Hodgkin lymphoma. *Blood*, **112**, 4247–4249.

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

Thyagarajan,B. *et al.* (2013) Mitochondrial DNA copy number is associated with breast cancer risk. *PLoS ONE*, **8**, e65968.

Wachsmuth,M. *et al.* (2016) Age-related and heteroplasmy-related variation in human mtDNA copy number. *PLOS Genet.*, **12**, e1005939.