Mitochondrion xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Mitochondrion

journal homepage: www.elsevier.com/locate/mito

Review

High-throughput sequencing in mitochondrial DNA research

Fei Ye [a], David C. Samuels [b], Travis Clark [c], Yan Guo [d,*]

[a] Department of Biostatistics, Vanderbilt University, Nashville, TN 37232, USA
[b] Center for Human Genetics, Vanderbilt University, Nashville, TN 37232, USA
[c] Vanderbilt Technology for Advanced Genomics, Vanderbilt University, Nashville, TN 37232, USA
[d] Department of Cancer Biology, Vanderbilt University, Nashville, TN 37232, USA

ARTICLE INFO

Article history:
Received 4 January 2014
Received in revised form 4 April 2014
Accepted 13 May 2014
Available online xxxx

Keywords:
High throughput sequencing
Next generation sequencing
Heteroplasmy
SNP
Mutation
Copy number

ABSTRACT

Next-generation sequencing, also known as high-throughput sequencing, has greatly enhanced researchers' ability to conduct biomedical research on all levels. Mitochondrial research has also benefitted greatly from high-throughput sequencing; sequencing technology now allows for screening of all 16,569 base pairs of the mitochondrial genome simultaneously for SNPs and low level heteroplasmy and, in some cases, the estimation of mitochondrial DNA copy number. It is important to realize the full potential of high-throughput sequencing for the advancement of mitochondrial research. To this end, we review how high-throughput sequencing has impacted mitochondrial research in the categories of SNPs, low level heteroplasmy, copy number, and structural variants. We also discuss the different types of mitochondrial DNA sequencing and their pros and cons. Based on previous studies conducted by various groups, we provide strategies for processing mitochondrial DNA sequencing data, including assembly, variant calling, and quality control.

Contents

1. Introduction

Typically, there are approximately 100 mitochondria in each mammalian cell, and each mitochondrion harbors 2–10 copies of mitochondrial DNA (mtDNA) (Robin and Wong, 1988). Thus, mtDNA mutations are often heteroplasmic, with a mixture of normal and mutant mtDNA copies within a cell (Durbin et al., 2010; Ng et al., 2010). It has been found that heteroplasmies throughout the mitochondrial genome are common in normal individuals and moreover, that the frequency of heteroplasmic variants varies considerably between different tissues in the same individual (He et al., 2010). Mitochondria generate the majority of their cellular energy through oxidative phosphorylation, which produces ATP. Mitochondrial dysfunctions are important causes of many neurological diseases (Fernandez-Vizarra et al., 2007) and drug toxicities (Lemasters et al., 1999; Wallace and Starkov, 2000).

1.1. Older methods to sequence mtDNA

Previously, the two most popular complete mitochondrial genome sequencing methods were direct Sanger sequencing and mitochondrial

* Corresponding author.

http://dx.doi.org/10.1016/j.mito.2014.05.004
1567-7249/

DNA re-sequencing by Affymetrix's MitoChip v.2.0 (referred to as "MitoChip"). The MitoChip is based on microarray technology that contains 25-mer probes complementary to the revised Cambridge Reference Sequence (rCRS) (Andrews et al., 1999). Several methods have been developed to quantify mtDNA heteroplasmy, such as real-time amplification refractory mutation system quantitative PCR (Bai and Wong, 2004), PCR-RFLP analysis (Holt et al., 1990), allele-specific oligonucleotide dot-blot analysis (Liang et al., 1998), and pyrosequencing (White et al., 2005). However, these methods are constrained by the limited number of targets they can scan. The maturity of high-throughput sequencing technology allows us to study the mitochondrial genome, including the level of mtDNA heteroplasmy at all sites in the mtDNA genome, in a reliable and cost-effective manner over large numbers of samples.

2. Direct sequencing of mtDNA

There have been three major sequencing platforms on the market: Illumina's HiSeq platform, Roche's 454 platform, and Applied Biosystems' SOLiD system. Mitochondrial DNA sequencing is possible with all three platforms (Craven et al., 2010; Payne et al., 2013), however the market has clearly been dominated by Illumina's sequencing platform for the last few years with no sign of diminishing. Thus, we focus on Illumina's sequencing technology in this review.

There are two typical ways to obtain information about the mitochondrial genome from high-throughput sequencing technology: direct and indirect. By "direct" we mean methods that sequence mtDNA directly through mtDNA enriched from total cellular DNA. There are several methods to enrich for mtDNA. Prior methods used ultracentrifugation in CsCL density gradients to enrich mtDNA from nuclear DNA but this is a time-consuming and low-throughput procedure. Faster, high-throughput methods for mtDNA enrichment are microarray hybridization and PCR-based enrichment. For example, in the study of mitochondrial disorders by Vasta et al., a custom-designed Agilent microarray was used to capture the entire mitochondrial genome (Vasta et al., 2009). Similarly, in a radiation therapy study by Guo et al., the Affymetrix MitoChip v.2.0 was used to enrich mtDNA, though it was not used for the sequencing. Custom-designed primers can also be used to capture mtDNA (He et al., 2010; Sosa et al., 2012). There is a major drawback for using overlapped primer capturing, however. For example, the MitoChip v.2.0 kit amplifies genomic DNA using PCR with two primer sets, mito3 and mito1-2. The two primer sets generate 7814 bp and 9307 bp long fragments respectively. Since, mtDNA is circular and only 16,569 base pairs long, the two fragments will generate two overlap regions. The sequencing depth of the two overlapped regions is significantly higher than the non-overlapped regions, and the primer sequences need to be trimmed prior to variant calling. Common practice is to discard data obtained from the overlapped regions if overlapped primers are used for enrichment (Guo et al., 2012a). Recently, a new PCR-based method using a single primer pair has been introduced to enrich the entire mitochondrial genome (Cui et al., 2013; Zhang et al., 2012). Using a single pair of primers readily avoids the pitfalls of using two or more sets of primers. Other advantages of this method include more uniform coverage, less interference from nuclear copies of the mitochondrial genome (nuMTs) (Hazkani-Covo et al., 2010; Li et al., 2012), and improved ability to estimate the breakpoints of large deletions. Additionally, several alternative commercial assays are available for mtDNA enrichment. For example, Qiagen SAbiosciences has a highly multiplexed PCR-based capture with an mtDNA GeneRead panel of 199 amplicons less than 300 bp covering 16,146 bases (99.86%) and Integrated DNA Technologies (IDT) offers a solution phase capture of mtDNA with their xGen Lockdown probes.

A recent study comparing DNA isolation kits and mtDNA enrichment with and without PCR found that the Qiagen Miniprep kit had 22% of the reads aligned to mtDNA without a PCR enrichment step and 99% of the reads aligned to the mtDNA with a limited 10-cycle PCR step using a high fidelity enzyme (Quispe-Tintaya et al., 2013). The commercial mtDNA isolation kits from Miltenyi Biotech and BioVision both had ~10% of the reads aligned to mtDNA without PCR. With PCR enrichment, Miltenyi-prepped DNA increased to ~35% aligned to mtDNA, and BioVision increased to only ~15% indicating that the mtDNA isolation kits were inefficient in enrichment of mtDNA directly and the standard Qiagen Miniprep kit isolated a larger fraction of mtDNA, even though it was not optimized for mtDNA enrichment in the extraction procedure.

3. Indirect sequencing of mtDNA

By "indirect sequencing", we mean methods to obtain mitochondrial DNA sequences as byproducts of other types of high-throughput sequencing. Besides performing deep-sequencing specifically targeted at mtDNA, mtDNA sequences can also be extracted from other types of high-throughput sequencing data such as exome and whole genome sequencing data. In exome sequencing data, a significant amount of reads will align to the mitochondrial genome (around 1–5%), even when it is not the intended sequencing target (Samuels et al., 2013). Because the mitochondrial genome is not considered to be part of the exome, it is not included in the set of target DNA for exome sequencing methods in common use today. A recent study has shown that mtDNA content can be extracted from exome sequencing data (Larman et al., 2012) and that the fraction of captured mtDNA sequences is linked to the relative abundance of the corresponding mitochondrial genome in the original total DNA extract (Picardi and Pesole, 2012). The average coverage of the mitochondrial genome from exome sequencing is ~100, easily surpassing the average coverage of even the targeted genomic regions (Picardi and Pesole, 2012). The relatively high coverage is due to the high copy number of mtDNA per cell, on the order of hundreds to several hundred thousand copies per cell, depending on the tissue type (Bogenhagen and Clayton, 1974). The advance of high-throughput sequencing technologies and the typically high coverage of an mtDNA sequence provide a powerful tool for the study of mitochondrial DNA heteroplasmy in unprecedented detail (Durbin et al., 2010; Goto et al., 2011; Guo et al., 2012b; Ng et al., 2010; Tang and Huang, 2010). However, this should be contrasted to techniques that specifically target the mtDNA sequence, which can produce an average depth of tens of thousands of reads across the mitochondrial genome (Ameur et al., 2011; Guo et al., 2012a; He et al., 2010; Tang and Huang, 2010). Researchers have started to infer information about mtDNA mutation from exome sequencing data. The best examples are The Cancer Genome Atlas (TCGA) project, where all mitochondrial DNA somatic mutations were inferred from exome sequencing data. For example, the current somatic mutation results for breast cancer in TCGA (Annon., 2012) contain exome sequencing data from 776 tumors and report 325 mtDNA somatic mutations derived from off-target reads from the exome sequencing data. Furthermore, by assessing mtDNA, exome sequencing mutation data has also been used to diagnose certain mitochondrial disorders (Dinwiddie et al., 2013).

Detecting mtDNA somatic mutation from exome sequencing data might contain false positive results caused by pseudogenes or homologous sequences (nuMTs). Because tumor tissue and the adjacent normal tissue often have different mtDNA content, the false results of heteroplasmic "mutation/variant" calling from the nuMTs could be different between the two samples, and thus at least some of the somatic mutations identified by using the exome data could be false. It is worth pointing out that false positive heteroplasmic variation due to nuMTs might also be improperly confirmed using a different method, unless the confirmation method carefully isolates the mtDNA from the nuclear DNA.

An important complication to consider in aligning DNA reads to the mitochondrial genome is the presence of nuMTs. The nuMTs can cause ambiguity about whether reads map to the nuclear or the mitochondrial genome. Aligning the raw reads against the mitochondrial reference genome directly and then filtering out the non-aligned reads, thus ignoring

Please cite this article as: Ye, F., et al., High-throughput sequencing in mitochondrial DNA research, Mitochondrion (2014), http://dx.doi.org/10.1016/j.mito.2014.05.004