EMBnet.journal
Bioinformatics in Action

Jayasekera et al. (2022) EMBnet.journal 28, e1007
http://dx.doi.org/10.14806/ej.28.0.1007

Research Papers

A Bioinformatics pipeline for variant discovery from Targeted Next Generation Sequencing of the human mitochondrial genome

Lakshika Jayasekera, Kanchana Senanayake✉, Ruwandi Ranasinghe, Kamani Tennekoon

Institute of Biochemistry, Molecular Biology and Biotechnology, University of Colombo, Sri Lanka

Competing interests: LJ none; KS none; RR none; KT none

Abstract

Sequence variants of human mitochondrial DNA (mt DNA) have been implicated in a variety of disorders and conditions. Massive parallel sequencing is becoming increasingly popular due to its efficiency and cost-effectiveness. In relation to acquiring significant sequence information like levels of heteroplasmy in mt DNA, it offers a marked improvement compared to previous methods used. Here we describe a variant calling pipeline for human mitochondrial DNA using Next Generation Sequencing (NGS) data obtained by enriching the sample only for mitochondria prior to sequencing.

Introduction

Human Mitochondrial (mt) genome is a double-stranded and a closed circular DNA molecule of 16,569 base pairs that represent <1% of total cellular DNA with each mitochondrion harboring 2-10 copies of mitochondrial DNA (mtDNA) molecules (Holt *et al.*, 2007; Veltri *et al.*, 1990). It codes for a total of 37 genes, including the 13 involved in electron transport and oxidative phosphorylation, 2 coding for 16rRNA and 12sRNA, and 22 coding for other mitochondrial transfer RNAs (tRNAs) needed for protein translation, thus proving its essential role in cellular function. (Anderson *et al.*, 1981). The presence of dissimilar sequences across different mitochondrial DNA molecules, from a single source, is referred to as heteroplasmy, which could conform to varying degrees among several tissues or different cells of the same tissue (Melton, 2004; Wong *et al.*, 2005). Compared to the nuclear genome, the mt genome has approximately 10 times higher mutation accumulation rate (Elmore, 2007), and it causes maternally inherited mitochondrial dysfunctions in a range of diverse disorders (mtDNA diseases) including diabetes mellitus, hypertension, Alzheimer's disease, heart diseases and cancer (Huang, 2011). There's also increasing evidence suggesting the association of somatic variants of mtDNA to other traits like ageing and cancer (Schon *et al.*, 2012). Thus, the characterisation of mitochondrial genome sequences is necessary for the molecular diagnosis of associated conditions. However, mtDNA analyses methods like PCR-restriction fragment length polymorphism (PCR-RFLP) analyses, Affymetrix's MitoChip, and even the gold standard Sanger sequencing fail to detect heteroplasmy under 10% (Mertens *et al.*, 2019). Further, these methods are hindered by the limited number of targets they can scan in a single run, highlighting the need for an accurate, cost-effective, and more sensitive method to study mtDNA.

Next Generation Sequencing and mtDNA analysis

Massive parallel sequencing has revolutionised the sequencing technology in recent years and proves ideal for small genome sequencing due to its high throughput and low cost (Yao *et al.*, 2019). Level of heteroplasmy detection could be significantly improved through these methods due to resulting high coverage and small size of the human mitochondrial genome. Although the different NGS technologies may use different methods in generating raw data, their final output is nucleotide base calls producing a huge number of 50 – 300 bp short reads (Mardis, 2013), usually combined in a FATSQ file.

Bioinformatics pipelines are an intrinsic aspect of NGS data analysis, to detect genomic alterations derived from these massive amounts of raw sequence data. The computationally intensive and complex nature of NGS data analysis makes many biologists who lack understanding of these computational techniques shy away from personally analysing raw sequence data (Roy

Article history
Received: 12 May 2021
Accepted: 23 September 2021
Published: 30 September 2022

et al., 2018). Several options are available to facilitate the analysis of mtDNA data acquired through NGS. However, their use is constrained by various factors. For example, several available bioinformatics pipeline frameworks like MitoSeek, Mtoolbox, are difficult to install, certain online servers like mit-o-matic, have limited input volumes or generate unreliable results (Weissensteiner *et al.*, 2016) In this paper, we present a bioinformatics pipeline for analyzing NGS data of targeted sequencing of the mitochondrial genome, through a series of command line tools.

Methodology

Quality assessment

When compared with Sanger sequencing, NGS acquire more errors as platforms face a variety of failures in chemistry and instrumentation, resulting in errors such as adaptor contamination, low-quality reads, and base call errors. To assure the conclusions derived through analysis are correct, it is necessary to eliminate these errors as downstream procedures fail to identify them (Pabinger *et al.*, 2014; Cox *et al.*, 2010; Dohm *et al.*, 2008). In this protocol, FastQC (version 0.11.8) (Andrews, 2010), the most preferred tool among the several tools available for checking the quality of raw data, was used for quality assessment. Upon assessment, this tool produces a report of useful information including quality score distribution across bases and across reads. Base quality score is an expression of base calling accuracy (Zhou and Rokas, 2014). A score of >20 is commonly referred to as the threshold for inclusion criteria of sequence reads, and removing sequences lower than 20 is preferable. As a part of the quality control procedure, adaptor sequences and the sequences not meeting the defined standards were removed using the cutadapt (version 1.18) tool (Martin, 2011). Similarly, sequences that were significantly shorter and longer than average were removed as well.

Alignment

Alignment is the process where the previously quality controlled massive amount of short reads, usually around 250 bp in FASTQ format, is mapped to the reference sequence that's in FASTA format, in this scenario, rCRS Human Mitochondrial Genome Reference sequence of Genbank accession No. NC_012920.1 (Andrews *et al.*,1999). It is the paramount step of any variant calling pipeline as even a few inaccurate alignments could produce many false-positive variant calls. For the current pipeline, the BWA-MEM tool of BWA aligner (version 0.7.12, one of the popular aligners that is fast and facilitates indel identification, -r1039) was used (Li, 2013). BWA was also used for the indexing of the reference sequence that was downloaded from Genbank prior to alignment, and the resulting sequence alignment mapping (SAM) format file from the alignment between the reference sequence and trimmed sequences was converted to a binary alignment map (BAM) file using SAMtools (version 1.9) BAM is the default binary format for storing sequence alignment data (Li *et al.*, 2009). Once the aligned BAM file is produced, alignment should be further refined. Accordingly, sorting and indexing of the BAM was also performed using SAMtools (version 1.9), through which BAM data is efficiently arranged and coordinate sorted, so the reads could be retrieved efficiently during further downstream analysis. SAMtools also compresses the aligned BAM file further before being used in deduplication (Li *et al.*, 2009).

Removing/marking duplicates

Duplicate removal is also enabled through SAMtools (version 1.9). This additional refinement process is important to mitigate the effects generated by the over-amplification of certain sequences. Duplicate sequences that are naturally present spanning the interested regions on DNA do not need to be removed. However, optical duplicates that are mistakenly read as separate clusters through signal capture software, when in reality they are generated by the same cluster, and the duplicates caused through PCR should be removed. PCR duplicates occur when the same amplified copies of one original fragment are identified as different fragments and further amplified through high throughput sequencing (Zhou and Rokas, 2014). In the initial steps of the NGS process, mtDNA is fragmented for library preparation and amplified for enrichment. Therefore, the presence of PCR duplicates at some level is usual. But having overly propagated duplicates cause erroneous end results. If a sequence subjected to PCR duplication contains a variant, that variant call would be biased. To further worsen the final output, if an error had occurred during PCR it would be further inflated during high throughput sequencing and cause a false positive. Therefore, compared with other DNA sequencing projects, the effect of duplicates is detrimental for heteroplasmy level detection of mtDNA analysis. Through deduplication algorithms, the groups of duplicate reads are identified and that of the highest sum of base quality scores is marked as a single read (Goto, 2011; Pfeifer, 2017).

Base quality score recalibration

The last stage that produces the final output of variant calling is possible through several different software tools. Just as many tools applied in different stages of the analysis, no single software could perfectly identify all variants in the genome of interest without false positives or negatives, however, according to a survey of variant analysis tools previously performed (Pabinger *et al.*, 2014), Genome Analysis Tool Kit (GATK) by Broad Institute of Harvard and MIT (McKenna *et al.*, 2010; Heldenbrand *et al.*, 2019) offers a satisfactory output for both germline and somatic variant calling. Prior to variant calling, Indel realignment and Base quality recalibration is the recommended practice. In previous versions of GATK, local alignment tools like IndelRealigner were available for Indel realignment.

Jayasekera et al. (2022) EMBnet.journal 28, e1007
http://dx.doi.org/10.14806/ej.28.0.1007

However, in the current version, GATK 4 realignment is no longer recommended (Heldenbrand *et al.*, 2019). As per the GATK best practices pipeline (Van der Auwera *et al.*, 2013), before the input files are processed through GATK tools, they should be preprocessed with utility software. Accordingly, a sequence dictionary for the reference file is created using CreateSequenceDictioanry tool by Picard (version 2.25.1), whereas the read group information of the input BAM file is assigned using the AddOrReplaceReadGroups tool also by Picard (version 2.25.1), a step through which, all the reads of a single file are assigned into one read group. Sorting and indexing of this file, which is necessary for subsequent steps could also be integrated into the same command.

Owing to the systemic technical errors of NGS data processing, base quality score alone may not be a proper indication of true base call errors. To address this issue GATK has introduced base quality score recalibration (BQSR). According to the official website available at GATK[1], BQSR is the machine learning algorithm introduced by GATK that enables improved overall base qualities, that subsequently increases the variant call accuracy. With the latest version of GATK, 4.2.0.0, the process involves two major steps. Initially, a model of covariation along with a recalibration table is produced on the BAM input from the previous step, with the BaseRecalibrator tool, based on an indexed VCF file of known variants and SNP's downloaded from dbSNP for the respective reference sequence and various covariates including read group, machine cycle number, reported quality score, and nucleotide context. Secondly, using the ApplyBQSR tool, a new BAM output is produced with adjusted base quality scores, depending on the built model. Additionally, the effects of recalibration are assessed by building another model for the new BAM output with the BaseRecalibrator tool, and plots are generated with the AnalyzeCovariates tool to compare the effects of the process. It is noteworthy that the AnalyzeCovariates tool requires other tools installed like R libraries, ggplot2, gsalib, and reshape to function.

Variant calling and filtering

Variant calling of Human mitochondrial genome is possible through the mitochondrial mode of GATK (version 4.2.0.0) Mutect2 (Benjamin *et al.*, 2019). Through mitochondrial mode, it allows sensitive calling of short nucleotide variants and indels at high depths with local assembly of haplotypes. The output is a raw highly sensitive VCF call set. Due to various types of errors and biases in the data, it demand the generation of a high-quality set of variants. To identify false positives out of the original VCF files and acquire a balance between sensitivity and specificity, necessary filters should be applied through variant filtering. GATK (version 4.2.0.0) offers several filtering tools based on different strategies used (Pfifer, 2017; Van der Auwera *et al.*, 2013). FilterMutectCalls is the recommended tool for this scenario (Benjamin *et al.*, 2019) through which possible false-positive artifacts will be flagged in the output VCF by its 'failed filter' while the remaining would be marked as 'PASS' (Van der Auwera *et al.*, 2013; Roy *et al.* 2018). With the SelectVariants tool, they can then be excluded from the final VCF output file.

Generating coverage plots and VCF file report

With SAMtools (version 1.9), coverage plots can be created for the final BAM output of GATK (version 4.2.0.0). Additionally, with the DISCVRSeq (version1.21) tool, a VCF file report can be generated using a previously compressed and indexed final VCF file.

Determining heteroplasmy levels, contamination, and haplogroup detection

High throughput sequencing has enabled the detection of heteroplasmy levels that are below 10%, which was not possible with previous sequencing methods (Mertens *et al.*, 2019). mtDNA-Server (Weissensteiner *et al.*, 2016), a free online data analysis server for mtDNA, reliably detects levels of heteroplasmy in NGS data 1% or above. The standalone version of mtDNA-Server, which can be locally installed, Mutserve (version 2.0.0-rc7) was used to detect the level of heteroplasmy in this pipeline and provided a sorted and indexed BAM as input.

Sample cross-contamination during next generation sequencing of mtDNA proves to be challenging at the data analysis phase as they may present themselves as low-level heteroplasmy (Dickins *et al.*, 2014; Li *et al.*, 2010). Haplocheck is a software tool developed to estimate the contamination level of mtDNA samples through mitochondrial phylogeny (Weissensteiner *et al.*,2021). It can be used as a cloud service or be locally installed. Locally installed Haplocheck (version 1.3.2), was used in analyzing NGS data for contamination.

Distinct regions of the mtDNA genome sequence that group together, reflecting phylogenetic origin through different maternal lineages are defined as mitochondrial haplotypes. These haplotypes can be assigned to haplogroups that represent the main branching points of the mitochondrial phylogenetic tree as they show how specific SNPs or variations have been accumulated through a certain matrilineage (Pipek *et al.*, 2019; Samuels *et al.*, 2006). Determining haplogroup in mitochondrial DNA sequence studies is important both to trace these lineages in human population genetics and to identify their various associations with diseases and health conditions. Haplogrep 2 (Weissenteiner *et al.*, 2016) is a popular tool used to classify haplogroups in NGS studies. Local installation of Haplogrep 2 (version 2.1.25) was used to identify haplogroups of preferred VCF input files.

[1]https://gatk.broadinstitute.org/hc/en-us/articles/360035890531-Base-Quality-Score-

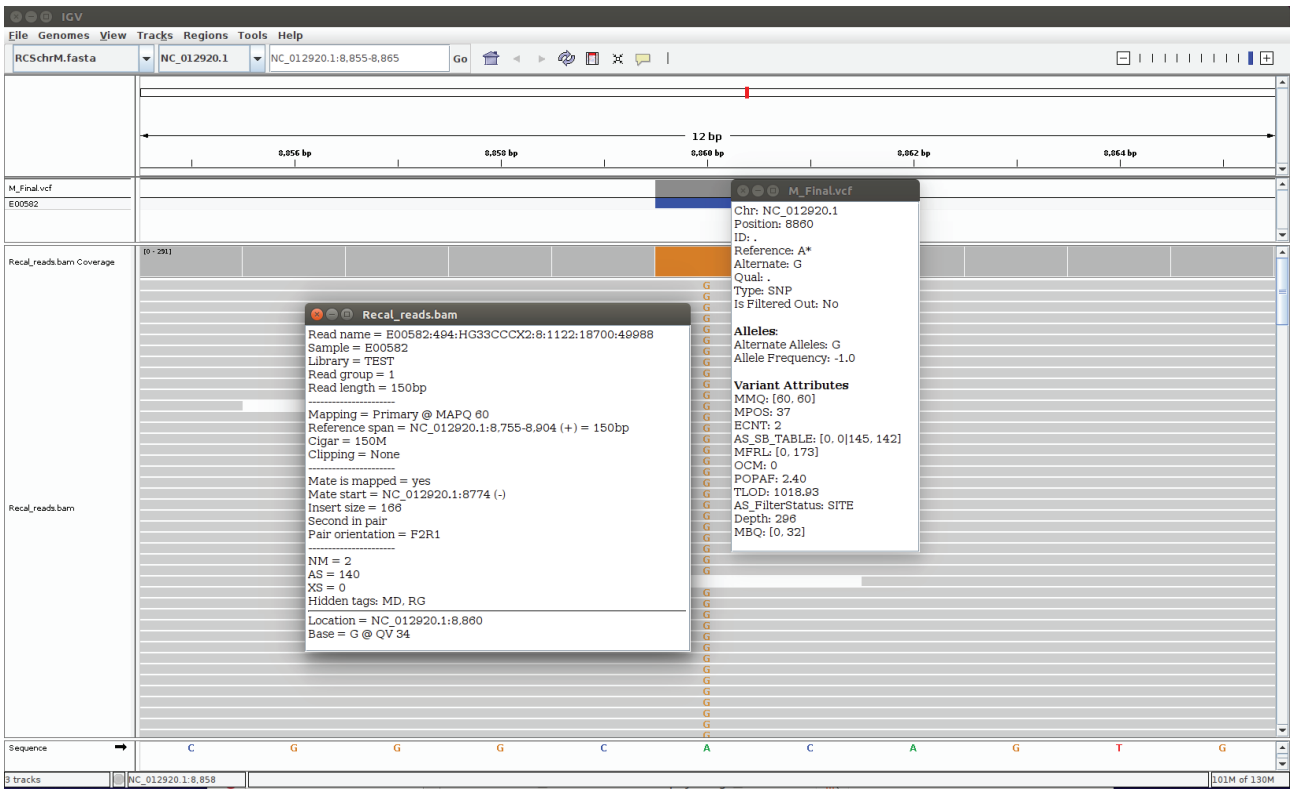

Figure 1. An A>G base substitution at position 8860 of human mitochondrial genome, aligned with rCRS Human Mitochondrial Genome Reference sequence zoomed for 12 base pairs length in IGV 2.9.2.

Results

Variant Annotation and visualisation

Assigning related biological information to identified variants is defined as variant annotation. There are several computational tools for human mitochondrial DNA analysis. Query with these tools provide sequencing data against variant databases and designate a set of associated metadata including the location compared to the reference sequence, respective change in amino acid and cDNA sequence, prediction of functional effects, and their presence in various databases (Roy *et al.*, 2018; Wadapurkar and Vyas, 2018). Variant Effect Predictor (VEP) (McLaren *et al.*, 2016) by ENSEMBLE project is a tool available at http://www.ensembl.org/ that accepts the final VCF file as input and enables the download of the annotated file in several formats. Via integrating SIFT (Kumar *et al.*, 2009) and Polyphen (Adzhubei *et al.*, 2010) it allows prediction of functional effect, as well as the discovery of genomic location, substitution effect of amino acid, and codon change. A txt format output, downloaded from VEP website, following annotation of the final VCF generated with the current pipeline is given in Supplementary file 5[2].

The final step of NGS variant calling pipelines is the visualisation of these data using genome browsers and visualisation tools. This task was performed with Integrative Genomics Viewer (IGV version 2.9.2) (Robinson *et al.*, 2011; Thorvaldsottir *et al.*, 2013) provided by Broad Institute of Harvard and MIT. It is a user-friendly and high performing interactive tool for exploring NGS data. By enabling read alignment examination, this step allowed further confirmation of called variant through visual estimation of it being true or a sequencing artifact. Additionally, through this step, more associated information of variants like mapping quality and variant impact acquired could be viewed individually (Figure 1).

Discussion

Limitations and perspective

Initially, it is important to notice that the sensitivity of heteroplasmy level detection correlates with increasing coverage depths (Holland *et al.*, 2011; Zhang *et al.*, 2012). At the same time, data generated with higher depths of coverage requires significant computer storage and advanced computers for data handling without compromising efficiency. Secondly, a signal from a very low-frequency variant is not discernable from that of a sequencing error. Finding the right balance between precision and sensitivity during data analysis, holds key to identify these variants. As desired, during variant filtering with FilterMutectCalls, the level for -f-score-beta argument could be adjusted between precision and recall, whereas 1 is the default value, 0.5 indicates higher precision over recall, and 2 indicates the higher recall over precision. However, other than bioinformatic analysis,

[2]http://journal.embnet.org/index.php/embnetjournal/article/downloadSuppFile/1007/1007_supp_5

criteria for quality control are also a limiting factor for detecting levels of heteroplasmies. It has been found that optimal conditions for primary PCR amplification and library preparation for high throughput sequencing are critically important as substandard conditions result in inflated variant frequencies (Mertens *et al.*,2019). Evidently, maximum accuracy in variant calling through NGS data demands high proficiency from laboratory practices to computational analysis. Nevertheless, despite the challenges present in the field of mitochondrial high throughput sequencing data processing, with rapidly improving and newly developing analysis tools coupled with other usual benefits offered via all massive parallel sequencing methods, NGS is likely to become predominant over previous sequencing methods in the foreseeable future.

Acknowledgements

This work was supported by National Research Council, Sri Lanka Grant No: 17-020 and constituted part of PhD studies of Jayasekera BMLP.

Key Points

- Targeted Sequencing of Human mitochondrial genome
- Bioinformatics pipeline for variant detection
- Next Generation Sequence data analysis of mtDNA

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A *et al.* (2010) A method and server for predicting damaging missense mutations. Nat Methods **7**(4), 248-249. http://dx.doi.org/10.1038/nmeth0410-248

Anderson S, Bankier AT, Barrell BG, De Bruijin MHL, Coulson AR *et al.* (1981) Sequence and organization of the human mitochondrial genome. Nature **290**, 457-465. http://dx.doi.org/10.1038/ 290457a0

Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM *et al.* (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nature Genetics **23**(2), 147. http://dx.doi.org/10.1038/13779

Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc (accessed 11 April 2021).

Benjamin D, Sato T, Cibulskis K, Getz G, Stewart C *et al.* (2019) Calling Somatic SNVs and Indels with Mutect2. bioRxiv. http://dx.doi.org/10.1101/861054 (accessed 11 April 2021)

Cox MP, Peterson DA, Biggs PJ (2010) SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics **11**, 485. http://dx.doi.org/10.1186/1471-2105-11-485

Dickins B, Rebolledo-Jaramillo B, Su MSW, Paul IM, Blankenberg D *et al.* (2014) Controlling for contamination in re-sequencing studies with a reproducible web-based phylogenetic approach. BioTechniques **56**(3), 134-141. http://dx.doi.org/10.2144/000114146

Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Research **36**(16), e105. http://dx.doi.org/10.1093/nar/gkn425

Dolled-Filhart MP, Lee MJr, Ou-Yang CW, Haraksingh RR, Lin JC (2013) Computational and bioinformatics frameworks for next-generation whole exome and genome sequencing. The Scientific World Journal 2013, 730210. http://dx.doi.org/10.1155/2013/730210

Elmore S (2007) Apoptosis: A review of programmed cell death. Toxicol Pathol **35**(4),495–516. http://dx.doi.org/10.1080/01926230701320337

Goto H, Dickins B, Afgan E, Paul IM, Taylor J *et al.* (2011) Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study. Genome Biology **12**(6), R59. http://dx.doi.org/10.1186/gb-2011-12-6-r59

Heldenbrand JR, Baheti S, Bockol MA, Drucker TM, Hart SN *et al.* (2019) Recommendations for performance optimizations when using GATK3.8 and GATK4. BMC Bioinformatics **20**(1), 557. http://dx.doi.org/10.1186/s12859-019-3169-7

Holland MM, McQuillan MR, O'Hanlon KA. (2011) Second generation sequencing allows for mtDNA mixture deconvolution and high resolution detection of heteroplasmy. Croat Med J **52**(3), 299-313. http://dx.doi.org/10.3325/cmj.2011.52.299

Holt IJ, He J, Mao CC, Boyd-Kirkup JD, Martinsson P *et al.* (2007) Mammalian mitochondrial nucleoids: organizing an independently minded genome. Mitochondrion **7**, 311-321. http://dx.doi.org/10.1016/j.mito.2007.06.004

Huang T (2011) Next generation sequencing to characterize mitochondrial genomic DNA heteroplasmy. Curr Protoc Hum Genet **19**, 19.8. http://dx.doi.org/10.1002/0471142905.hg1908s71

Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc **4**(7), 1073-81. http://dx.doi.org/10.1038/nprot.2009.86

Li M, Schonberg A, Schaefer M, Schroeder R, Nasidze I *et al.* (2010) Detecting heteroplasmy from high-throughput sequencing of complete human mitochondrial DNA genomes. Am J Hum Genet **87**(2), 237-249. http://dx.doi.org/10.1016/j.ajhg.2010.07.014

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J *et al.* (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**(16), 2078–2079. http://dx.doi.org/10.1093/bioinformatics/btp352

Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. https://arxiv.org/abs/1303.3997 (accessed 11 April 2021).

Mardis ER (2013) Next-generation sequencing platforms. Annu Rev Anal Chem (Palo Alto Calif) **6**, 287-303. http://dx.doi.org/10.1146/annurev-anchem-062012-092628

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet journal **17**(1),10–12. http://dx.doi.org/10.14806/ej.17.1.200

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research **20**(9), 297-303. http://dx.doi.org/10.1101/gr.107524.110

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS *et al.* (2016) The Ensembl Variant Effect Predictor. Genome Biol **17**, 122. http://dx.doi.org/10.1186/s13059-016-0974-4

Melton T (2004) Mitochondrial DNA Heteroplasmy. Forensic Sci Rev **16**(1), 1-20.

Mertens J, Zambelli F, Daneels D, Caljon B, Sermon K *et al.* (2019) Detection of Heteroplasmic Variants in the Mitochondrial Genome through Massive Parallel Sequencing. Bio-protocol **9**(13), e3283. http://dx.doi.org/10.21769/BioProtoc.3283

Pabinger S, Dander A, Fischer M, Snajder R, Sperk M *et al.* (2014) A survey of tools for variant analysis of next-generation genome sequencing data. Briefings in Bioinformatics **15**(2), 256-278. http://dx.doi.org/10.1093/bib/bbs086

Pfeifer SP (2017) From next-generation resequencing reads to a high-quality variant data set. Heredity **118**(2),111–124. http://dx.doi.org/10.1038/hdy.2016.102

Pipek OA, Medgyes-Horváth A, Dobos L, Stéger J, Szalai-Gindl J *et al.* (2019) Worldwide human mitochondrial haplogroup distribution from urban sewage. Scientific Reports **9**(1), 11624. http://dx.doi.org/10.1038/s41598-019-48093-5

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES *et al.* (2011) Integrative genomics viewer. Nat Biotechnol **29**(1), 24-26. http://dx.doi.org/10.1038/nbt.1754

Roy S, Coldren C, Karunamurthy A, Kip NS, Klee EW *et al.* (2018) Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists. J Mol Diagn **20**(1), 4-27. http://dx.doi.org/10.1016/j.jmoldx.2017.11.003

Samuels DC, Carothers AD, Horton R, Chinnery PF (2006) The power to detect disease associations with mitochondrial DNA haplogroups. Am J Hum Genet **78**(4), 713-720. http://dx.doi.org/10.1086/ 502682

Schon EA, Dimauro S, Hirano M (2012) Human mitochondrial DNA: roles of inherited and somatic mutations. Nat Rev Genet **13**(12), 878–890. http://dx.doi.org/10.1038/nrg3275

Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform **14**(2), 178-192. http://dx.doi.org/10.1093/bib/bbs017

Van der Auwera G, Carneiro MO, Hartl C, Poplon R, Guillermo Del A *et al.* (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Current Protocols in Bioinformatics **43**(1110), 11.10.1-11.10.33. http://dx.doi.org/10.1002/047 1250953.bi1110s43

Veltri KL, Espiritu M, Singh G (1990) Distinct genomic copy number in mitochondria of different mammalian organs. J Cell Physiol **143**(1), 160-164. http://dx.doi.org/10.1002/jcp.1041430122

Wadapurkar RM and Vyas R (2018) Computational analysis of next generation sequencing data and its applications in clinical oncology. Informatics in Medicine Unlocked **11**,75-82. http://dx.doi.org/10.1016/j.imu.2018. 05.003

Weissensteiner H, Forer L, Fendt L, Kheirkhah A, Salas A *et al.* (2021) Contamination detection in sequencing studies using the mitochondrial phylogeny. Genome Res **31**(2), 309-316. http://dx.doi.org/10.1101/gr.256545.119

Weissensteiner H, Forer L, Fuchsberger C, Schöpf B, Kloss-Brandstätter A *et al.* (2016) mtDNA-Server: next-generation sequencing data analysis of human mitochondrial DNA in the cloud. Nucleic Acids Research **44**(W1), W64–W69. http://dx.doi.org/10.1093/nar/gkw247

Weissensteiner, H, Pacher D, Kloss-Brandstätter A, Forer L, Specht G *et al.* (2016). HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic Acids Research **44**(W1), W58–W63. http://dx.doi.org/10.1093/nar/gkw233

Wong, L-JC and Boles RG (2005) Mitochondrial DNA analysis in clinical laboratory diagnostics. Clin Chim Acta **354**(1-2), 1–20. http://dx.doi.org/10.1016/ j.cccn.2004.11.003

Yao Y, Nishimura M, Murayama K, Kuranobu N, Tojo S *et al.* (2019) A simple method for sequencing the whole human mitochondrial genome directly from samples and its application to genetic testing. Scientific Reports **9**(1), 17411. http://dx.doi.org/10.1038/s41598-019-53449-y

Zhang W, Cui H, Wong LJ (2012) Comprehensive one-step molecular analyses of mitochondrial genome by massively parallel sequencing. Clin Chem **58**(9), 1322-1331. http://dx.doi.org/10.1373/clinchem.2011.181438

Zhou X, Rokas A (2014) Prevention, diagnosis and treatment of high-throughput sequencing data pathologies. Molecular Ecology **23**(7), 1679-1700. http://dx.doi.org/10.1111/mec.1268