Gene 668 (2018) 129–134

Contents lists available at ScienceDirect

Gene

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

Research paper

Whole genome sequencing and bioinformatics analysis of two Egyptian genomes

Mahmoud ElHefnawi[a,*,1], Sungwon Jeon[b,c,1], Youngjune Bhak[b,c], Asmaa ElFiky[a,d], Ahmed Horaiz[a], JeHoon Jun[e,f], Hyunho Kim[f], Jong Bhak[b,c,e,f,**]

[a] Biomedical Informatics and Chemo-Informatics Group, Centre of Excellence for Advanced Sciences (CEAS), and Informatics and Systems Department, National Research Centre, Cairo 12622, Egypt
[b] Korean Genomics Industrialization and Commercialization Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea
[c] Department of Biomedical Engineering, School of Life Sciences, Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea.
[d] Environmental and Occupational Medicine Department, National Research Centre, Cairo 12622, Egypt
[e] Personal Genomics Institute, Genome Research Foundation, Cheongju 28160, Republic of Korea
[f] Geromics, Ulsan 44919, Republic of Korea.

ARTICLE INFO

Keywords:
Whole-genome sequencing
Egyptian
Variants
Human migration
Bioinformatics

ABSTRACT

We report two Egyptian male genomes (EGP1 and EGP2) sequenced at ~ 30× sequencing depths. EGP1 had 4.7 million variants, where 198,877 were novel variants while EGP2 had 209,109 novel variants out of 4.8 million variants. The mitochondrial haplogroup of the two individuals were identified to be H7b1 and L2a1c, respectively. We also identified the Y haplogroup of EGP1 (R1b) and EGP2 (J1a2a1a2 > P58 > FGC11). EGP1 had a mutation in the NADH gene of the mitochondrial genome ND4 (m.11778 G > A) that causes Leber's hereditary optic neuropathy. Some SNPs shared by the two genomes were associated with an increased level of cholesterol and triglycerides, probably related with Egyptians obesity. Comparison of these genomes with African and Western-Asian genomes can provide insights on Egyptian ancestry and genetic history. This resource can be used to further understand genomic diversity and functional classification of variants as well as human migration and evolution across Africa and Western-Asia.

1. Introduction

A human genome holds an extensive amount of data on human evolution, diversity, health, physiology, and medicine (Lander et al., 2001). Whole genome sequencing (WGS) data can be used for the deepest possible genetic analyses for various purposes such as common and rare disorder association studies. Genomes and their diverse variation information can also be used effectively for estimating risk factors of common diseases (Bick & Dimmock, 2011; Thompson et al., 2012). Currently, massively-parallel next-generation sequencing (NGS) methods are the most widely used method for analyzing the whole human genomes. Programs to map short reads of a genome and to call the subsequent variations are being rapidly improved and upgraded (Lupski et al., 2010). In addition, the cost of analyzing a genome has become very low and WGS is becoming more common in detecting uncommon, disease-causing variants by scrutinizing affected people's genomes (Lupski et al., 2010; Sobreira et al., 2010; Roach et al., 2010). For example, it can be useful for screening women who have BRCA1 and BRCA2 genes mutations to assess the risk of breast and ovarian cancers (Campeau et al., 2008).

The Egyptian population is diverse due to its position between Africa and Asia. It has two long banks along the Nile River, which is the longest African River, and has hosted various populations throughout history. Ancient Egyptian traditions, such as mummification, play an important role in preserving genomes and subsequent analysis of DNA variants (Paabo, 1985). Egyptian DNA have been studied for a long time

Abbreviations: EGP, Egyptian person; HQ, High quality; LD, Linkage disequilibrium; LHON, Leber's hereditary optic neuropathy; mtDNA, Mitochondrial DNA; NGS, Next generation sequencing; rCRS, revised Cambridge reference sequence; SNP, Single nucleotide polymorphism; WGS, Whole genome sequencing; Y-STR, short tandem repeat (STR) on the Y-chromosome

* Corresponding author.
** Correspondence to: J. Bhak, Korean Genomics Industrialization and Commercialization Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea.
E-mail addresses: mahef@aucegypt.edu (M. ElHefnawi), jongbhak@genomics.org (J. Bhak).
[1] Equal contributors.

https://doi.org/10.1016/j.gene.2018.05.048
Received 28 March 2018; Accepted 13 May 2018
Available online 25 May 2018

(Hawass et al., 2010). Research on the DNA of modern North Africans revealed that their gene frequencies are central to those of southern Europe, the Near East, and Sub Saharan Africa (Cavalli-Sforza et al., 1994). However, the frequency distribution of the non-recombining portion of the Y chromosome of modern Egyptian population is extremely similar to those of the middle northern African population, implying a more extensive portion of Eurasian genetic components (Cavalli-Sforza et al., 1994; Bosch et al., 1997; Manni et al., 2002; Arredi et al., 2004; Luis et al., 2004).

Pagani et al. showed, after analyzing 225 Egyptians and Ethiopians, that the correct root out of Africa is the northern one through Sinai to Eurasia (Pagani et al., 2015). Khairat, Ball et al., found that one of the mummified Egyptian people pertained to the L2 mtDNA haplogroup (Li & Durbin, 2009), a maternal clade that is accepted to have origins from Western Asia.

Here, we report the analyses of two male Egyptian whole genomes at high sequencing depth. A systematic genomic analysis, including the analysis of functional and deleterious mutations, was conducted and we provided the genomic structure and phylogenetic tree in comparison with populations in Africa and the Middle-East.

2. Material and method

2.1. DNA extraction and ethical approval

Blood samples were extracted from two healthy people whose parents originate from the Delta (North of Egypt) and Saied (South of Egypt), respectively. This study was approved by Institutional Review Board at Genome Research Foundation with IRB-REC-2011-10-003 and the written consent was signed by the participants.

2.2. Sample preparation and whole genome sequencing

Genomic DNA was extracted from blood with the Gene JETBlood genomic DNA purification Kit (Thermo Scientific, USA), according to manufacturer's protocol. A library of 400–500 bp insert size was created. The genomic DNA was sheared utilizing Covaris S series (Covaris, MS, USA). The sheared DNA was end-repaired A-tailed, and ligated to paired-end adapters, according to the manufacturer's protocol (Truseq DNA Sample Prep Kit v2, Illumina, San Diego, CA, USA). Adapter-ligated fragments were then size selected on a 2% Agarose gel, with the 520–620 bp band being removed. Gel extraction and column purification was executed by applying the Minelute Gel Extraction Kit (Qiagen), following the manufacturer's protocol. The ligated DNA parts which contained adapter sequences were enhanced via PCR using adapter specific primers. Library quality and concentration were resolved using theAgilent 2100 BioAnalyzer. The libraries were evaluated utilizing a KAPA library quantification kit (KapaBiosystems, MA, USA), as indicated by Illumina's library quantification protocol. According to the qPCR quantification, the libraries were standardized to 2 nM and after that denatured using 0.1 N NaOH. Cluster amplification of denatured templates was completed in flow cells, according to the manufacturer's protocol (Illumina). Flow cells were paired-end sequenced (2 × 100 bp) on an Illumina HiSeq2000 machine. In order to process the raw fluorescent images and the called sequences, the base-calling pipeline (Sequencing Control Software (SCS), Illumina) was applied. The rest of our analysis was initiated from the FASTQ files maintained by Illumina's downstream analysis CASAVA software suite. The raw data can be accessed at NCBI SRA, with accession number **SRR5738871** and **SRR5738872**.

2.3. Alignment of reads to reference and variants detection

The NGSQC toolkit v 2.3.3 (Patel & Jain, 2012) was applied to filter low quality reads with an '-l 70 – s 20' options (cutoff read length for HQ = 70%, cutoff quality score = 20). Subsequently, we aligned the filtered reads onto the hg19 human reference using BWA-MEM 0.7.8 (Li & Durbin, 2009) with the default option. SAM files were then restored to the BAM file using Samtools 0.1.19 (Li et al., 2009a). To remove PCR duplicated reads, MarkDuplicate subroutine in Picard v1.9.2 (http://broadinstitute.github.io/picard/) was used. We also conducted IndelRealigner and BaseRecalibration using GATK v2.3.9 (McKenna et al., 2010a) in order to increase the accuracy of variants calling. The variants were called by GATK UnifiedGenotyper with '–heterozygosity 0.0010 -dcov 200 -stand_call_conf 30.0 -stand_emit_conf 30.0' options.

2.4. Annotation and functional analysis of variants

We annotated the type and genomic regions of variants using snpEff v4.3i (Cingolani et al., 2012). To predict mutations which possibly make function altering amino acid changes, PROVEAN was used (Choi & Chan, 2015). These protein-damaging mutations were further annotated with OMIM (McKusick, 2007) and ClinVar databases (Landrum et al., 2016).

2.5. Construction of phylogenetic tree and ADMIXTURE analysis

We first merged Affymetrix human origin single nucleotide polymorphism (SNP) panel (HOSP) data (Lazaridis et al., 2014) with the two Egyptian genomes using PLINK 1.90 (Purcell et al., 2007), generating 591,356 autosomal SNPs. We pruned the panel with linkage disequilibrium (LD). The final dataset included 289,287 SNPs. We ran *ADMIXTURE* 1.3.0 (Alexander et al., 2009) with default cross-validation and used the number of ancestral population K value from 2 to 10. To construct a phylogenetic tree across populations from Africa and the Middle-east, we calculated a pairwise nucleotide distance with the same SNP panel for *ADMIXTURE*. We then constructed a neighbor-joining tree with pairwise pi distance (Nei & Li, 1979).

2.6. Identification of mitochondrial DNA and Y chromosome haplotype

Before identifying the mitochondrial haplogroup, we extracted the short sequencing reads which were mapped to chrM of hg19 reference using in-house scripts. We then mapped the reads to rCRS mtDNA reference using BWA-MEM (Li & Durbin, 2009) and have generated consensus mtDNA sequences for the samples using samtools (Li et al., 2009a). The mitochondrial haplogroup of each sample was identified by MitoTool (Fan & Yao, 2013). The haplogroup of Y chromosome was identified by using Nevgen predictor.

3. Results and discussion

3.1. The donors

EGP1's clinical history shows bilateral visual loss, worse on the left consistent with Leber's hereditary optic neuropathy (LHON), a mitochondrial disorder which disturbs the optic nerves specifically. A visual field examination (Table 1), MRI scan, and genetic testing all confirmed the diagnosis of LHON. Additionally, there was a family history of cardiovascular and pulmonary diseases in EGP2.

Table 1
Clinical data of ocular examination of EGP1.

	Right eye	Left eye
Vision (cc)	20/400	2/300
Color	RE 45/12	LE 2.5/12
Refraction	−4.5 sphere	−4.5 sphere
Pupils	Sluggishly reactive, There is probably a left relative afferent pupillary defect	
Ocular motility	Ductions and Versions: Full There is no internuclearophthalmologia	