Abstract and Introduction:

- Genome refers to the complete set of genetic information in an organism, Human genome contains approximately 3 billion base pairs, which reside in the 23 chromosomes.

- Alleles occur with different frequencies within different human populations. The more distant groups are geographically and linearly, the greater the differences between them. The differences between human groups represent a small percentage of the general human genetic variation. The degree of difference varies between members of a single population as well

  • Previously, It is estimated that more than 80% of the samples in the (GWAS) studies came from people of European ancestry. The Egyptian Genome Repertoire will support efforts to cover the North African population, who were among the small remaining gaps in the current Human Genome Project (GRCh38).

  •We implemented a workflow to extract all Egyptref data for view in the IGV. This includes all sequencing data mapped to GRCh38 GRCh38) as well as all assembly differences.

- A small number of de novo assembled human genomes have been reported to date, and few have been complemented with population-based genetic variation, which is particularly important for North Africa, a region underrepresented in current genome-wide references. Here, we combine long- and short-read whole-genome sequencing data with recent assembly approaches into a de novo assembly of an Egyptian genome. The assembly demonstrates well-balanced quality metrics and is complemented with variant phasing via linked reads into haploblocks, which we associate with gene expression changes in blood. To construct an Egyptian genome reference, we identify genome-wide genetic variation within a cohort of 109 Egyptian individuals. We show that differences in allele frequencies and linkage disequilibrium between Egyptians and Europeans may compromise the transferability of European ancestry-based genetic disease risk and polygenic scores, substantiating the need for multi-ethnic genome references. Thus, the Egyptian genome reference will be a valuable resource for precision medicine.

- The Human Genome Project contributed to the production of the first complete genome sequence of a human in 2003. Since then, thousands of human genomes have been deciphered as a result of the startling development of Next Generation Sequencing technology that has made this process much cheaper and faster. The resulting data is used worldwide in biomedical sciences , anthropology , forensic medicine and other branches of science. There are widespread expectations that genetic studies will lead to advances in disease diagnosis and treatment, and new insights into many areas of biology , including human evolution .

- We integerated the the sequences of an additional 109 individuals to generate an Egyption Reference ,It will help us to know what are the diseases prevalent among Egyptians, find a treatment for these diseases, the effects of the environment on people, and learn about genetic changes between them and other peoples.

- With Egyptref, it will be possible to perform comprehensive integrated genome and

transcriptome comparisons for Egyptian individuals in the future. This will shed light on personal as well as population-wide common genetic variants.

Related Work:

- ***Whole Genome Sequencing and Bioinformatics analysis of two Egyptian Genomes.*** 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.

- ***Complete Genome Sequence and Bioinformatics analysis of nine Egyptian Females with clinical information from different geographic regions in Egypt.*** Here we report the analysis of whole genomes of nine Egyptian females from different regions using Illumina short-read sequencers. At 30x sequencing coverage, we identified 12 SNPs that were shared in most of the subjects associated with obesity which are concordant with their clinical diagnosis. Haplogroup and Admixture analyses revealed that most Egyptian samples are close to the other north Mediterranean, Middle Eastern, and European, respectively, possibly reflecting the into-Africa influx of human migration. In conclusion, we present whole-genome sequences of nine Egyptian females with personal clinical information that cover the diverse regions of Egypt. Although limited in sample size, the whole genomes data provides possible geno-phenotype candidate markers that are relevant to the region's diseases.

- ***Gail Model Utilization in predicting Breast Cancer Risk in Egyptian women: a cross-sectional study.*** Herein, our purpose was to calculate the 5-year and lifetime risk of breast cancer and to assess new breast cancer potential contributors among Egyptian women utilizing the modified Gail model, while presenting a global comparison of risk assessment. This study included 7009 women from both urban and rural areas scattered across 40% of the Egyptian provinces. We revealed that modified Gail model had a well-fitting and discrimination accuracy in Egyptian women when compared with other countries. In Egypt, breast cancer comprises the most common of all cancer types in females; with 28,000 confirmed cases each year as reported by the National Cancer Institute (NCI), Egypt . While the incidence in Egypt seems to be slightly lower than the corresponding rates in the USA and other Western societies, Egyptian breast cancer patients are

characterized by higher mortality rate (20.1 per 100,000) compared to the USA (14.7 per 100,000) . The current recommendations enacted by the WHO are to commence screening to women at early ages in the hope of an early detection to reduce such mortality rates and minimize the burden of breast cancer .

- **There are many projects in all countries of the world that are interested in building a reference genome that depends on explaining the similarities and differences between the people of the same genome and explaining the difference between them and the rest of the peoples to find out the genetic changes that occur over days and the changes that occur due to their being affected by the weather or some medications.**

- **We genetically characterized the Egyptian population with respect to 143 other populations of the world using variant data of 5429 individuals in total. For this, we combined five different data sets:**

(1) a recently published whole-genome sequencing (WGS)-based variant data set from 929 individuals of the Human Genome Diversity Project (HGDP), covering 51 populations.

(2) 2504 individuals from 26 populations of the 1000 Genomes project for which phase 3 genotypes are available.

(3) WGS-based variant data from 108 Qatari individuals.

(4) SNP array-based variant data of 478 individuals from five countries of the Arabian Peninsula.

(5) 1305 individuals from 68 African, European, Western and Southern Asian populations that were compiled from eight different publications into a recent SNP array-based variant data set.

Methodology:

-Sample acquisition

Samples were acquired from 10 adult Egyptian individuals. These were recruited from healthy relatives escorting patients admitted to Mansoura University hospital. Medical history was taken to ascertain no history of chronic diseases, followed by a full clinical examination by a medical doctor alongside routine laboratory investigations (Liver and kidney function tests and complete blood count (CBC)). For nine individuals, high-coverage illumina short-read data were generated. For the assembly individual, high-coverage short-read data were generated as well as high-coverage PacBio data and 10x Genomics data. Furthermore, we used public Illumina short-read data from 100 Egyptian individuals .

-PacBio Data Generation & illumina short read data generation& RNA sequencing data generation& 10x Genomics sequencing data generation.

-Construction of draft de novo data assemblies and meta-assembly.

We used WTDBG227 for human de novo assembly followed by its accompanying polishing tool WTPOA-CNS with PacBio reads and in a subsequent polishing run with Illumina short reads. This assembly was further polished using **PILON** with short-read

- **we constructed two draft assemblies**

- **FALCON** are *de novo* genome assemblers for PacBio long reads, also known as single-molecule real-time (SMRT) sequences. **FALCON** is a diploid-aware assembler which follows the hierarchical genome assembly process (HGAP) and is optimized for large genome assembly (e.g. non-microbial). **FALCON** produces a set of primary contigs (a-contigs), which represent divergent allelic variants. Each a-contig is associated with a homologous genomic region on an p-contig.

- **WTDBG2** is a de novo sequence assembler for long noisy reads produced by PacBio or Oxford Nanopore Technologies (ONT). It assembles raw reads without error correction and then builds the consensus from intermediate assembly output. **Wtdbg2** is able to assemble the human and even the 32Gb Axolotl genome at a speed tens of times faster than CANU and FALCON while producing contigs of comparable base accuracy.

-Repeatmasking

Repeatmasking was performed by using REPEATMASKER61 with RepBase version 3.0

- **RepeatMasker is a program that screens DNA sequences for interspersed repeats and low complexity DNA sequences. ... Currently over 56% of human genomic sequence is identified and masked by the program.**

-Unique inserted sequences

We trimmed Illumina short sequencing reads of 110 Egyptian individuals using FASTP 0.20.0 with default parameters, mapped the output reads to GRCh38 and GATK bundle sequences using BWA 0.7.15-r1140 and sorted by chromosomal position using SAMTOOLS 1.3.1. Subsequently, we extracted reads that did not map to GRCh38 using SAMTOOLS with parameter F13 (i.e., read paired, read unmapped, mate unmapped) and repeated the mapping and sorting using the Egyptian de novo assembly. We merged the read-group specific BAM files for each sample and calculated the per base read depth using SAMTOOLS. Afterwards, we aggregated the results via custom scripts and extracted uniquely inserted sequences from the Egyptian de novo assembly. Insertions were defined as contiguous regions of at least 500 bp having a coverage of more than 5 reads per base in 10 or more samples. Lastly, we BLASTed the obtained sequences against the standard databases (option nt) for

highly similar sequences (option megablast) using a custom script. For the uniquely inserted sequences that we identified, we created a pileup over all BAM files containing the reads that did not map to GRCh38 using SAMTOOLS. Based on these pileups, we then called the variants using BCFTOOLS. Variants with quality of more than 10 were kept.

-Phasing

Phasing was performed for the assembly individual's SNVs and short indels obtained from combined genotyping with the other Egyptian individuals, i.e., based on short-read data. These variants were phased using 10x Genomics data and the 10x Genomics LONGRANGER WGS pipeline with four 10x libraries provided for one combined phasing. See Supplementary Methods: Variant phasing for details.

-SNVs and small indels

Calling of SNVs and small indels was performed with GATK 3.832 using the parameters of the best practice workflow. Reads in each read group were trimmed using Trimmomatic62 and subsequently mapped against reference genome hg38 Then, the alignments for all read groups were merged sample-wise and marked for duplicates. After the base recalibration, we performed variant calling using HaplotypeCaller to obtain GVCF files. These files were input into GenotypeGVCFs to perform joint genotyping. Finally, the variants in the outputted VCF file were recalibrated, and only those variants that were flagged as PASS were kept for further analyses.

-Variant annotation & Structural variants

Variant annotation was performed using ANNOVAR67 and VEP47, SVs were called using DELLY2 with default parameters according to instructions on the DELLY2 website for germline SV calling , Overlapping SV calls in the same individual were collapsed by the use of custom scripts.

- Mitochondrial haplogroups

Haplogroup assignment was performed for 227 individuals using HAPLOGREP2.

Haplogroups are used to represent the major branch points on the mitochondrial phylogenetic tree. Understanding the evolutionary path of the female lineage has helped population geneticists trace the matrilineal inheritance of modern humans back to human origins in Africa and the subsequent spread around the globe.

Haplotypic expression analysis

RNA sequencing reads were mapped and quantified using STAR. Haplotypic expression analysis was performed by using PHASER and PHASER GENE AE with Ensembl annotation on the 10x-phased haplotypes using default parameters.

GWAS catalog data integration

GWAS catalog associations for GWAS of European ancestry were split into trait-specific data sets using experimental factor ontology terms. For every trait, a locus was defined as an associated variant +/− 1 Mb, and only loci that were replicated were retained. For proxy computation, we used our

Egyptian cohort ($n = 110$) and the European individuals of 1000 Genomes ($n = 503$). For details, see Supplementary Methods: Data integration with the GWAS catalog.

in conclusion, we constructed the first Egyptian—and North African—genome reference, which is an essential step towards a comprehensive, genome-wide knowledge base of the world's genetic variations. The wealth of information it provides can be immediately utilized to study in-depth personal genomics and common Egyptian genetics and its impact on molecular phenotypes and disease. This reference will pave the way towards a better understanding of the Egyptian, African and global genomic landscape for precision medicine.