A personal and population-based Egyptian genome reference

Inken Wohlers, Axel Künstner, Matthias Munz, Michael Olbrich, Anke Fähnrich, Caixia Ma, Misa Hirose, Shaaban El-Mosallamy, Mohamed Salama, Hauke Busch & Saleh Ibrahim

Abstract

The human genome is composed of 23 chromosomal DNA sequences of bases `A`, `C`, `G` and `T` -- the blueprint to implement the molecular functions that are at the basis of every individual's life. Deciphering the first human genome was a consortium effort that took more than a decade and cost about 3 billion dollars. With latest technological advances, determining an individual's entire personal genome at manageable cost and effort comes into reach. Although the benefit of all-encompassing genetic information that entire genomes provide is widely noted, so far only a small number of *de novo* assembled human genomes have been reported. Even less have been characterized and complemented with respect to population-specific variation. Here we combine long- and short-read whole genome next-generation sequencing data together with the recent assembly approaches for the first *de novo* assembly of the genome of an Egyptian individual, which we merged with Egyptian variant data into a population reference genome. The resulting genome assembly demonstrates overall well-balanced quality metrics and comes along with high quality variant phasing into maternal and paternal haplotypes. Further, we assayed population-specific variations genome-wide within a representative cohort of more than 100 Egyptian individuals. By annotation of these genetic data and integration with public databases we showcase genetic variants that alter protein sequence and that are linked to allelic gene expression. This is one of a handful of studies that comprehensively describe a population reference genome based on a high-quality personal genome and which highlights population-specific variants of interest. It is a proof-of-concept to be considered by the many national genome initiatives underway. And, more importantly, we anticipate that the Egyptian reference genome will be a valuable resource for precision medicine initiatives targeting the Egyptian population and beyond.

All summary data of the Egyptian genome reference is available at www.egyptian-genome.org. The Egyptian genome reference will be publicly available upon journal publication.

Main

In the last years, several high-quality *de novo* human genome assemblies (1–3) and, more recently, pan-genomes (4) extended human sequence information and improved the *de facto* reference genome. At present, many national genome initiatives are established which aim to genetically characterize human populations (5).

Population-specific genetic variation as part of an individual's personal genetic variation is indispensable for precision medicine (PM). Currently, genomics-based PM compares the patients' genetic make-up to a reference genome, a genome model inferred from people of mostly European descent, to detect risk mutations that are related to disease. However, genetic and epidemiologic studies have long recognized the importance of ancestral origin in conferring risk genes for disease. Risk alleles and structural variants (6) can be missing from the reference genome or can have different population frequencies such that alternative pathways become disease related in patients of different ancestral origin, which motivates to establish national genome projects. At present, there are several population-based sequencing efforts that aim at

mapping out specific variants in the 100,000 genomes projects in Asia (7) or England (8). Further, large-scale sequencing efforts currently explore population, society and history-specific genomic variations in Northern and Central Europe (9,10), North America, Asia (1) and recently the first sub-Saharan Africans (4). However, it is still expensive to obtain all-embracing genetic information such as high-quality *de novo* assemblies for many individuals. Currently a subset of population variation is readily assessable, e.g. single-nucleotide polymorphisms (SNPs) on genotyping arrays, variation in exonic regions by use of exome sequencing (11,12) or variation detectable by short-read sequencing (10,13–17).

In this study we have generated a phased *de novo* assembly of an Egyptian individual and used it as a basis to identify single-nucleotide variants (SNVs) and structural variants (SVs) from an additional 109 Egyptian individuals obtained from short-read sequencing. Those were integrated to generate a consensus reference Egyptian genome. We anticipate that an Egyptian population reference genome will strengthen precision medicine efforts that may eventually benefit nearly 100 million Egyptians. Likewise, our genome will be of universal value for research purposes, since it contains both European and African variant features, and could thus be used to investigate the validity of genetic disease risk transfer across populations. As most genetic association studies are performed in Europeans (18), an Egyptian genome will be well suited to identify (i) genetic loci with shared or with distinct disease susceptibility across populations (ii) haplotypes that influence gene expression and (iii) variants that are likely protein-damaging and putatively related to disease.

Our Egyptian genome is based on a high-quality human *de novo* assembly for one Egyptian individual (see workflow in Suppl. Figure 1). This assembly was generated from PacBio, 10x Genomics and Illumina paired-end sequencing data at overall 270x genome coverage (Suppl. Table 1). For this personal genome, we constructed two draft assemblies, one based on long-read assembly by an established assembler, `FALCON` (19), and another one based on the assembly by a novel assembler, `WTDBG2` (20), that has a much lower runtime at comparable accuracy (cf. Suppl. Fig. 1). Both assemblies were polished using short-reads and various polishing tools. For the `FALCON`-based assembly, scaffolding was performed, whereas we found that the `WTDBG2`-based assembly was of comparable accuracy without scaffolding (cf. dotplots in Suppl. Figs. 2-3). We compared our two draft assemblies to the publicly available assemblies of a Korean (1) and a Yoruba individual (GeneBank assembly accession GCA_001524155.4, unpublished) with respect to various quality control (QC) measures using `QUAST-LG` (21) (Table 1). The `WTDBG2`-based assembly was selected as base, because it performs comparable or better concerning various QC measures (Suppl. Table 2).

Where larger gaps outside centromere regions occurred, we complemented this assembly with sequence from the `FALCON`-based assembly (Suppl. Table 3) to obtain a final Egyptian meta assembly, denoted as EGYPT (for overall assembly strategy, see Suppl. Figure 1). The comparative assembly statistics are summarized in Table 1. Suppl. Figure 2 compares the assemblies NA-values and Suppl. Figures 3-7 show dot plots of alignment with reference GRCh38. We performed repeat annotation and repeat masking for all assemblies (Suppl. Table 4).

The meta assembly was complemented with high-quality phasing information (Suppl. Table 5). Variants and small insertions and deletions (indels) called using short-read sequencing data were phased using high-converge linked-read sequencing data. This resulted in 98.99% of variants being phased. Further, nearly all (99.41%) of genes with length less than 100kb and more than one heterozygous SNP were phased into a single phase-block.

Based on the personal Egyptian genome, we constructed an Egyptian population genome by considering genome-wide SNV allele frequencies in 109 additional Egyptians (Suppl. Table 6). This enabled the characterization of the major allele (i.e. the allele with highest allele frequency) in the given Egyptian cohort. For this, we called variants using short-read data of 12 Egyptians sequenced at high coverage and 97 Egyptians sequenced at low coverage. Although sequence coverage affects variant-based statistics (Suppl. Fig. 8), due to combined genotyping most variants could also be called reliably in low coverage samples (Suppl. Fig. 9). Altogether, we called a total of 19,758,992 SNVs and small indels (Suppl. Fig. 10) in all 110 Egyptian individuals (Table 2). The number of called variants per individual varied between 2,901,883 to 3,934,367 and was correlated with sequencing depth (see Suppl. Figs. 8-9). This relation was particularly pronounced for low coverage samples. The majority of variants was intergenic (53.5%) or intronic (37.2%) (Suppl. Fig. 11). Only about 0.7% of variants were located within coding exons, of which 54.4% were non-synonymous and thus have an impact on protein structure (Suppl. Fig. 12).

Using short-read sequencing data of 110 Egyptians, we called 121,141 structural variants, which were mostly deletions (Suppl. Fig. 13), but also inversions, duplications, insertions and translocations of various orders of magnitudes (Table 2, Suppl. Fig. 14). Similar to SNVs, also SV calls vary between individuals (Suppl. Fig. 15) and are slightly affected by coverage (Suppl. Fig. 16). After merging overlapping SV calls we obtained on average 2,773 SVs per Egyptian individual (Suppl. Table 7, Suppl. Figs. 17-19).

To characterize the Egyptian population with respect to European and African populations which have been genotyped within the 1000 Genomes Project (22) (Suppl. Table 8), we used SNVs and short indels for a genotype-based principal component analysis. According to this analysis, Egyptians are a genetically homogenous population compared to other populations, sharing genetic variants with both Europeans and Sub-Saharan Africans (see Fig. 1 and Suppl. Figs. 20-32). So far, there are no North-African populations with high-quality genome-wide genotype data available, and from the European and Sub-Saharan African populations reported by the 1000 Genomes Project, Egyptians are closest to the European Tuscany population (see Fig. 1 and Suppl. Figs. 20-32), which has been previously proposed through the genetic studies of ancient Egyptian mummies (23).

The mixed European and African ancestry of Egyptians is further supported by mitochondrial haplogroup assessment from literature (17) and our own analyses. We found that Egyptians have haplogroups most frequent in Europeans (e.g. H,V,T,J etc.; more than 60%), but many also had African (e.g. L with 24.8%) or Asian/East Asian haplogroups (e.g. M with 6.7%), indicating that the Egyptian genome contains genetic variations from various major human population (Suppl. Fig. 33).

In total we identified 2,270,642 common Egyptian SNVs (MAF > 5%) of which 26,564 are population-specific, i.e., they are rare (MAF < 1%) to non-existent in all other continental populations according to the 1000 Genomes data (Table 2). This is comparable to population-specific variant numbers reported previously for 1000 Genomes populations (24). Additionally, we found 4,807 African, 2 Ad Mixed American, 11 East Asian, 3 European and 77 South Asian SNVs that are population-specific in the Egyptian cohort and the respective continental population (Figure 1). These numbers clearly indicate an insufficient coverage of the genetic heterogeneity of the world's population for precision medicine and thus the need for local reference genomes.

To detect a putative genetic predisposition of Egyptian population-specific SNPs towards molecular pathways, phenotypes or disease, we selected all genes having a Combined Annotation-Dependent Depletion (CADD) phred score > 15 (25). This resulted in 361 associated genes out of which we discarded 159 non-protein coding or anti-sense genes. The resulting 202 genes were uploaded to Enrichr, a gene list enrichment tool incorporating 153 gene set and pathway databases (26). Among the most enriched pathways we found 4 out of 23 body fat percentage related genes from the GWAS catalogue 2019 (adj. p-value =0.038; Genes CRTC1; IGF2BP1; WDR41; SULT1A2) as well as Glycolysis in humans from the 2016 Panther database (adj.p-val = 0.017, 3 out of 17 Genes: TPI1;BPGM;GAPDH), which was confirmed by the HumanCyc 2016 database. There, we found the terms glycolysis, gluconeogeneis und superpathway of conversion of glucose to acetyl CoA (pathway IDs PWY-6313, PWY66-400, PWY66-407) significant (adj. p-value=0.013; Genes TPI1; SUCLG2; BPGM; GAPDH). Lastly, there are 7 out of 103 genes frequently mutated which are related to obesity according to the DISEASES resource (27) (adj. p-value=0.019; Genes: PKHD1; ANKDD1B; SV2C; NRXN3; CDH12; ZNF248; SLC30A10). These results might hint at population-specific metabolism regulation that is linked to body weight.

Variants that are not protein-coding may have a regulatory effect on gene and eventually protein expression. Using blood expression data obtained from RNA sequencing for the assembly individual in conjunction with the phased variant data, we identified genes whose expression differs between maternal and paternal haplotype (see Suppl. Fig. 34 for the analysis overview and Suppl. Figs. 35-36 for results). We report 1,180 such genes (see Suppl. Table 9).

Through our analysis it will be possible to perform integrated genome and transcriptome comparisons for Egyptian individuals based on our reference genome, which might shed light on personal as well population-wide common genetic variants. Figure 2 depicts an example for such an integrated analysis. Here we use the DNA repair associated gene BRCA2, which is linked to breast and other cancer types, if mutated. The figure depicts the sample coverage based on different PacBio, 10x Genomics and Illumina whole genome short-read sequencing for a personal genome together with previously identified risk loci and common Egyptian SNPs. The bottom compares the identified SNVs and Indels from the Korean and Yoruba reference genome with our *de novo* EGYPT assembly. Visual inspection already yields significantly different variants. Furthermore, note the three significant GWAS SNPs between position 32,390 and 32,400kb. These examples support the need for whole genome sequencing analysis to shed light on both mutations and structural variations on the personal and population-based genome level.

In conclusion, we have constructed the first Egyptian reference genome, which is a hitherto unprecedented substantial step towards compiling a comprehensive, genome-wide knowledge base of personal and population-specific genetic variation. The wealth of information it provides can be immediately utilized to evaluate, on a genome-wide scale, whether a genetic region of interest is affected by personal or population-specific variation. A comprehensive annotation of these variations indicates their impact on molecular phenotypes such as RNA abundance or protein structure and therefore their potential relevance in disease and will pave the way towards a better understanding of the genomic landscape of the Egyptian population for precision medicine.

Methods

Sample acquisition

Samples were acquired from 10 Egyptian individuals. For nine individuals, high coverage Illumina short-read data was generated. For the assembly individual, high coverage short-read data was generated as well as high-coverage PacBio data and 10x data. Further, we used public Illumina short-read data from 100 Egyptian individuals from Pagani et al (17). See Supplementary Tables 1 and 6 for an overview of the individuals and the corresponding raw and result data generated in this study.

PacBio data generation

For Pacbio library preparation, the SMRTbell DNA libraries were constructed following the manufacturer's instructions (Pacific Bioscience, www.pacb.com). The SMRTbell DNA libraries were sequenced on the PacBio Sequel and generated 298.2GB of data.
Sequencing data from five PacBio libraries was generated at overall 99x genome coverage.

Illumina short-read data generation

For 350bp library construction, the genomic DNA was sheared, and fragments with sizes around 350bp were purified from agarose gels. The fragments were ligated to adaptors and PCR amplified. The generated libraries were then sequenced on the Illumina HiSeq X Ten using PE150 and generated 312.8GB of data.
For the assembly individual, sequencing data from five libraries was generated at 90x genome coverage. For nine additional individuals, one library each was generated amounting to overall 305x coverage of sequencing data. For the 100 individuals of Pagani et al (17), three were sequenced at high coverage (30x) and 97 at low coverage (8x). Average coverage over SNV positions for all 110 samples is provided in Supplementary Table 6.

RNA sequencing data generation

For RNA sequencing, ribosomal RNA was removed from total RNA, and double-stranded cDNA were synthesized, and then adaptors were ligated. The second strand of cDNA was then degraded to generate a directional library. The generated libraries with insert size of 250-300 bp were selected and amplified, and then sequenced on the Illumina HiSeq using PE150. Overall, 64,875,631 150-bp paired-end sequencing reads were generated.

10x sequencing data generation

For 10x genomic sequencing, the Chromium Controller was used for DNA indexing and barcoding according to the manufacturer's instructions (10x Genomics, www.10xgenomics.com). The generated fragments were sheared, and then adaptors were ligated. The generated libraries were sequenced on the Illumina HiSeq X Ten using PE150 and generated 272.7 GB of data.
Sequencing data from four 10x libraries was generated at 80x genome coverage.

Construction of draft *de novo* assemblies and meta assembly

We used `WTDBG2` (20) 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 data (cf. Suppl. Methods: *`WTDBG2`-based assembly*).
An alternative assembly was generated by using `FALCON`, `QUIVER`, `SSPACE-LONGREAD (28)`, `PBJELLY` (29), `FRAGSCAFF` (30) and `PILON` (31) (cf. Suppl. Methods: *`FALCON`-based assembly*).
Proceeding from the `WTDBG2`-based assembly, we constructed a meta assembly. Regions larger than 800kb that were not covered by this base assembly and were not located within centromere regions were extracted from the alternative `FALCON`-based assembly (Suppl. Table 3). See

253 Supplementary Figure 1 for an overview of our assembly strategy including meta-assembly
254 construction (cf. Suppl. Methods: *Meta assembly construction*).
255 Assembly quality and characteristics were assessed with `QUAST-LG` (cf. Suppl. Methods:
256 *Assembly comparison and QC*). The extraction of coordinates for meta-assembly construction
257 was performed using `QUAST-LG` output.
258
259 **Repeatmasking**
260 Repeatmasking was performed by using `REPEATMASKER` (32) with RepBase version 3.0
261 (Repeatmasker Edition 20181026) and Dfam_consensus (http://www.dfam-consensus.org) (cf.
262 Suppl. Methods: *Repeat annotation*).
263
264 **Phasing**
265 Phasing was performed for the assembly individual's SNVs and short indels obtained from
266 combined genotyping with the other Egyptian individuals, i.e. based on short-read data. These
267 variants were phased using 10x data and the 10x Genomics `LONGRANGER WGS` pipeline with
268 four 10x libraries provided for one combined phasing. See Supplementary Methods *Variant*
269 *Phasing* for details.
270
271 **SNVs and small indels**
272 Calling of SNVs and small indels was performed with `GATK` 3.8 (33) using the parameters of
273 the best practice workflow. Reads in each read group were trimmed using `Trimmomatic` (34)
274 and mapped against reference genome hg38 using `BWA`, subsequently. Then the alignments for
275 all read groups were merged sample-wise and marked for duplicates. After the base
276 recalibration, we run the variant calling using `HaplotypeCaller` to obtain GVCF files.
277 These files were then combined into batches and inputted into GenotypeGVCFs to perform
278 joint genotyping. Lastly, the variants in the outputted VCF file were recalibrated and only
279 considered only those variants that were flagged as "PASS" were kept for further analyses. We
280 used `FastQC` (35), `Picard Tools` (36) and `verifyBamId` (37) for QC (cf. Suppl.
281 Methods: *Small variant QC*).
282
283 **Variant annotation**
284 Variant annotation was performed using `ANNOVAR` (38) and `VEP` (39) (cf. Suppl. Methods:
285 *Small variant annotation*)
286
287 **Structural variants**
288 Structural variants were called using `DELLY2` (40) with default parameters as described on the
289 `DELLY2` website for germline SV calling (https://github.com/dellytools/delly) (cf. Suppl.
290 Methods: *Structural variant QC*). Overlapping SV calls in the same individual were collapsed
291 by the use of custom scripts. See Supplementary Methods *Collapsing structural variants* for
292 details.
293
294 **Genotype principal components**
295 1000 Genomes phase 3 variant data was obtained for all European and African individuals and
296 merged with the Egyptian variant data. Variants were excluded if their minor allele frequency
297 was less than 5% in 1000 Genomes individuals, they violate Hardy-Weinberg-Equilibrium, are
298 multi-allelic or within regions of high LD and/or known inversions. LD pruning was performed
299 and remaining SNPs passed on to the `SMARTPCA` program (41) of the `EIGENSOFT` package
300 for PC computation. See Supplementary Methods *Genotype principal components* for details.
301
302 **Mitochondrial haplogroups**

Haplogroup assignment was performed for 227 individuals using `HAPLOGREP` 2 (42). Further, mitochondrial haplogroups have been obtained from Pagani et al. (17) for 100 individuals. See Suppl. Methods *Mitochondrial haplogroups* for details.

Population-specific variants

SNVs that are common in the 110 Egyptians and otherwise rare in the 1000 Genomes populations were considered Egyptian-specific. We considered a variant common if it has a minor allele frequency of at least 5% and as rare if it has a minor allele frequency of less than 1%.

Haplotypic expression analysis

RNA-Seq reads were mapped and quantified using `STAR` (Version 2.6.1.c) (43). Haplotypic expression analysis was performed by using `PHASER` and `PHASER GENE AE` (version 1.1.1) (44) with Ensembl version 95 annotation on the 10x-phased haplotypes using default parameters. See Supplementary Methods *Allelic expression* for details.

Integrative genomics view

We implemented a workflow to extract all Egyptian genome reference data for view in the Integrative Genomics Viewer (IGV) (45). This includes all sequencing data mapped to GRCh38 (cf. Suppl. Methods *Sequencing read mapping to GRCh38*) as well as all assembly differences (cf. Suppl. Methods *Alignment to GRCh38* and *Assembly-based variant identification*) and all Egyptian variant data. See Suppl. Methods *Gene-centric integrative data views* for details.

Ethics statement

The study was approved by the Mansoura Faculty of Medicine Institutional Review Board (MFM-IRB) Approval Number RP/15.06.62. All subjects gave written informed consent in accordance with the Declaration of Helsinki.

References

1. Seo J-S, Rhie A, Kim J, Lee S, Sohn M-H, Kim C-U, et al. De novo assembly and phasing of a Korean human genome. Nature. 2016 Oct 13;538(7624):243–7.

2. Cho YS, Kim H, Kim H-M, Jho S, Jun J, Lee YJ, et al. An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes. Nat Commun. 2016 24;7:13637.

3. Shi L, Guo Y, Dong C, Huddleston J, Yang H, Han X, et al. Long-read sequencing and *de novo* assembly of a Chinese genome. Nature Communications. 2016 Jun 30;7:12065.

4. Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, Rafaels N, et al. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nature Genetics. 2019 Jan;51(1):30.

5. Stark Z, Dolman L, Manolio TA, Ozenberger B, Hill SL, Caulfied MJ, et al. Integrating Genomics into Healthcare: A Global Responsibility. Am J Hum Genet. 2019 Jan 3;104(1):13–20.

6. Levy-Sakin M, Pastor S, Mostovoy Y, Li L, Leung AKY, McCaffrey J, et al. Genome maps across 26 human populations reveal population-specific patterns of structural variation. Nature Communications. 2019 Mar 4;10(1):1025.

7. GenomeAsia 100k [Internet]. GenomeAsia 100k. [cited 2019 Mar 4]. Available from: http://www.genomeasia100k.com/

8. Turnbull C, Scott RH, Thomas E, Jones L, Murugaesu N, Pretty FB, et al. The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. BMJ. 2018 Apr 24;361:k1687.

9. Schneider VA, Lindsay TG, Howe K, Bouk N, Chen H-C, Kitts PA, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. bioRxiv. 2016 Aug 30;072116.

10. Maretty L, Jensen JM, Petersen B, Sibbesen JA, Liu S, Villesen P, et al. Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. Nature. 2017 03;548(7665):87–91.

11. Henn BM, Botigué LR, Gravel S, Wang W, Brisbin A, Byrnes JK, et al. Genomic Ancestry of North Africans Supports Back-to-Africa Migrations. PLOS Genetics. 2012 Jan 12;8(1):e1002397.

12. Scott EM, Halees A, Itan Y, Spencer EG, He Y, Azab MA, et al. Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. Nature Genetics. 2016 Sep;48(9):1071.

13. Bai H, Guo X, Narisu N, Lan T, Wu Q, Xing Y, et al. Whole-genome sequencing of 175 Mongolians uncovers population-specific genetic architecture and gene flow throughout North and East Asia. Nat Genet. 2018 Nov 5;

14. Chiang CWK, Mangul S, Robles C, Sankararaman S. A Comprehensive Map of Genetic Variation in the World's Largest Ethnic Group-Han Chinese. Mol Biol Evol. 2018 Nov 1;35(11):2736–50.

15. ElHefnawi M, Jeon S, Bhak Y, ElFiky A, Horaiz A, Jun J, et al. Whole genome sequencing and bioinformatics analysis of two Egyptian genomes. Gene. 2018 Aug 20;668:129–34.

16. Choudhury A, Ramsay M, Hazelhurst S, Aron S, Bardien S, Botha G, et al. Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. Nat Commun. 2017 Dec 12;8(1):2062.

17. Pagani L, Schiffels S, Gurdasani D, Danecek P, Scally A, Chen Y, et al. Tracing the Route of Modern Humans out of Africa by Using 225 Human Genome Sequences from Ethiopians and Egyptians. Am J Hum Genet. 2015 Jun 4;96(6):986–91.

18. Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic Studies. Cell. 2019 Mar 21;177(1):26–31.

19. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased diploid genome assembly with single-molecule real-time sequencing. Nature Methods. 2016 Dec;13(12):1050–4.

20. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. bioRxiv. 2019 Jan 26;530972.

388 21. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome
389 assembly evaluation with QUAST-LG. Bioinformatics. 2018 01;34(13):i142–50.

390 22. The 1000 Genomes Project Consortium. A global reference for human genetic variation.
391 Nature. 2015 Oct;526(7571):68–74.

392 23. Schuenemann VJ, Peltzer A, Welte B, Pelt WP van, Molak M, Wang C-C, et al. Ancient
393 Egyptian mummy genomes suggest an increase of Sub-Saharan African ancestry in post-
394 Roman periods. Nature Communications. 2017 May 30;8:ncomms15694.

395 24. Choudhury A, Hazelhurst S, Meintjes A, Achinike-Oduaran O, Aron S, Gamieldien J, et
396 al. Population-specific common SNPs reflect demographic histories and highlight
397 regions of genomic plasticity with functional relevance. BMC Genomics [Internet]. 2014
398 Jun 6 [cited 2019 Jun 20];15(1). Available from:
399 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4092225/

400 25. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the
401 deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019 Jan
402 8;47(D1):D886–94.

403 26. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al.
404 Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic
405 Acids Res. 2016 08;44(W1):W90-97.

406 27. Pletscher-Frankild S, Pallejà A, Tsafou K, Binder JX, Jensen LJ. DISEASES: Text
407 mining and data integration of disease–gene associations. Methods. 2015 Mar 1;74:83–
408 9.

409 28. Boetzer M, Pirovano W. SSPACE-LongRead: scaffolding bacterial draft genomes using
410 long read sequence information. BMC Bioinformatics [Internet]. 2014 Dec [cited 2019
411 Jun 24];15(1). Available from:
412 https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-15-211

413 29. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, et al. Mind the Gap: Upgrading
414 Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. Liu Z,
415 editor. PLoS ONE. 2012 Nov 21;7(11):e47768.

416 30. Adey A, Kitzman JO, Burton JN, Daza R, Kumar A, Christiansen L, et al. In vitro, long-
417 range sequence information for de novo genome assembly via transposase contiguity.
418 Genome Research. 2014 Dec;24(12):2041–9.

419 31. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an
420 integrated tool for comprehensive microbial variant detection and genome assembly
421 improvement. PLoS ONE. 2014;9(11):e112963.

422 32. SMIT AFA. Repeat-Masker Open-3.0. http://www.repeatmasker.org [Internet]. 2004
423 [cited 2019 Jun 21]; Available from: https://ci.nii.ac.jp/naid/10029514778/

424 33. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The
425 Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation
426 DNA sequencing data. Genome Research. 2010 Sep 1;20(9):1297–303.

34. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114–20.

35. Andrews S. FASTQC - A quality control tool for high throughput sequence data. [Internet]. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

36. Picard Toolkit [Internet]. Available from: http://broadinstitute.github.io/picard/

37. Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis GR, et al. Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data. The American Journal of Human Genetics. 2012 Nov;91(5):839–48.

38. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010 Sep 1;38(16):e164–e164.

39. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biology. 2016 Jun 6;17(1):122.

40. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012 Sep 15;28(18):i333–9.

41. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006 Dec;2(12):e190.

42. Kloss-Brandstätter A, Pacher D, Schönherr S, Weissensteiner H, Binna R, Specht G, et al. HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. Hum Mutat. 2011 Jan;32(1):25–32.

43. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2012 Oct 25;bts635.

44. Castel SE, Mohammadi P, Chung WK, Shen Y, Lappalainen T. Rare variant phasing and haplotypic expression from RNA sequencing with phASER. Nat Commun. 2016 08;7:12817.

45. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011 Jan;29(1):24–6.

46. Michailidou K, Lindström S, Dennis J, Beesley J, Hui S, Kar S, et al. Association analysis identifies 65 new breast cancer risk loci. Nature. 2017 Nov;551(7678):92–4.

Supplementary Information

Acknowledgements

We acknowledge support on coordination of the project and assembly work through Ms. Lu Wang from the Novogene (UK) Company Limited.

Author information

Medical Systems Biology Division, Lübeck Institute of Experimental Dermatology and Institute for Cardiogenetics, University of Lübeck, Lübeck, Germany
Inken Wohlers, Axel Künstner, Matthias Munz, Michael Olbrich, Anke Fähnrich & Hauke Busch

Novogene (UK) Company Limited, Babraham Research Campus, Cambridge, United Kingdom
Caixa Ma

Medical Experimental Research Institute, Mansoura University and the American University in Cairo, Egypt
Mohamed Salama & Shaaban El-Mosallamy

Genetics Division, Lübeck Institute of Experimental Dermatology, University of Lübeck, Lübeck, Germany
Misa Hirose and Saleh Ibrahim

Contributions

H.B, S.I., M.S. conceived the study. I.W, A.K, M.M., H.B. and S.I. designed the study. I.W., A.K., M.M., M.O and A.F. performed data analysis. C.M. constructed the `FALCON`-based assembly. M.S. and S.E-M. compiled the Egyptian cohort and provided samples. I.W., H.B. and S.I. wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

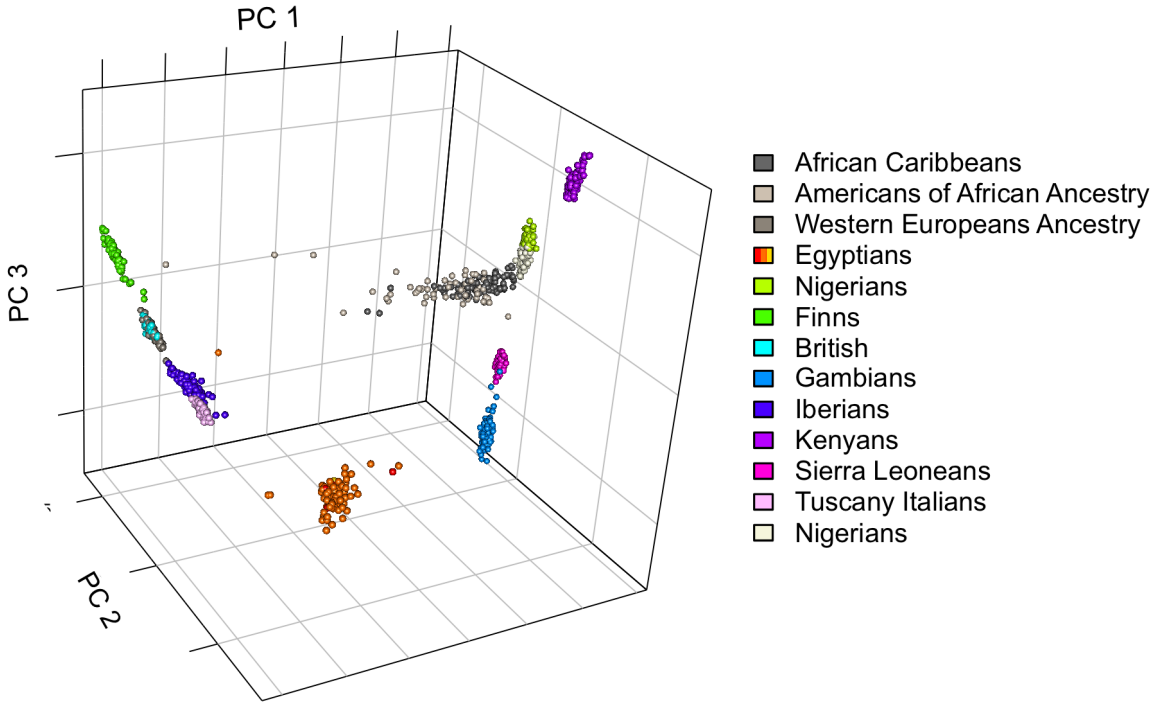
The authors declare no competing interests

Corresponding authors

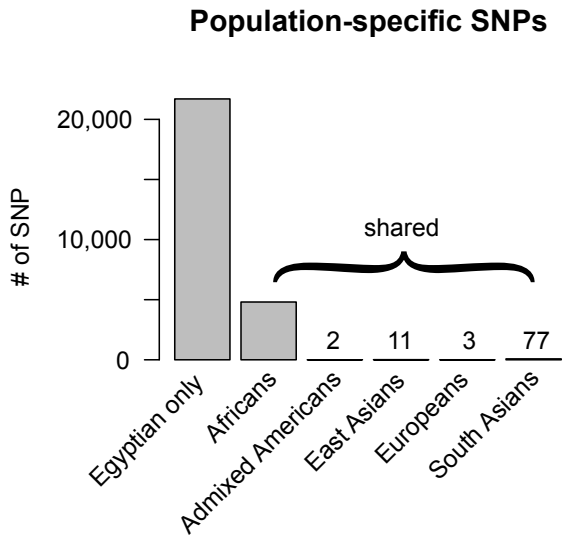
Correspondence to Hauke Busch or Saleh Ibrahim

495 **Figures**

496



497



498

499 *Figure 1: Top: PCA plot of different populations from the 1000 Genomes Project and 110 Egyptian genomes from Pagani et*
500 *al. as well as from our own study. Bottom: Egyptian population-specific SNPs and SNPs that are common in Egyptians and*
501 *specific to a single continental population.*

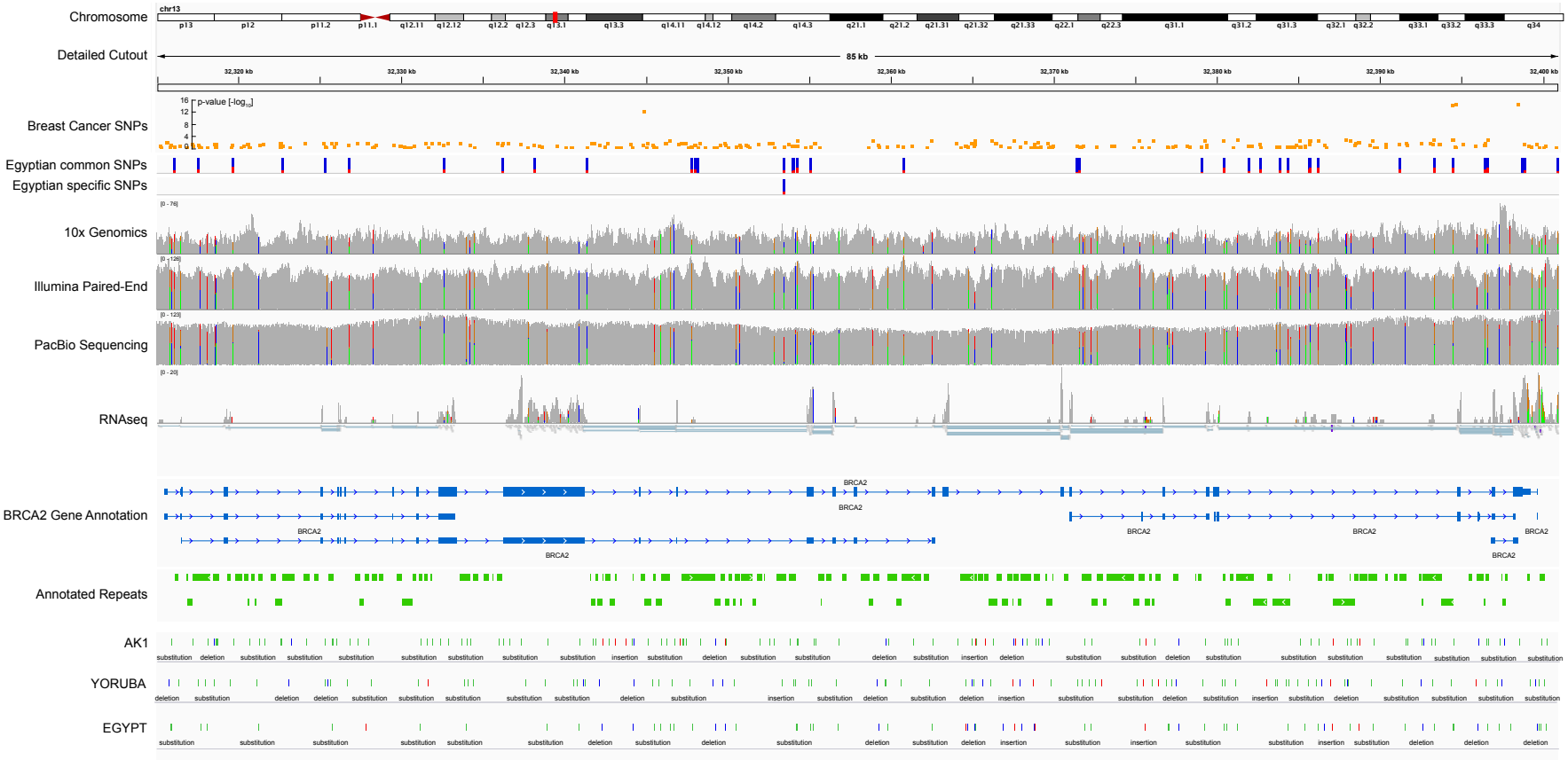


Figure 2: Integrative view of all data utilized and generated within the Egyptian genome project for the gene BRCA2, which is associated with breast cancer. The rows denote from top to bottom: Genome location on chromosome 13 of the magnified region for BRCA2 (first and second row), GWAS data for breast cancer risk loci (46), variants that are common in the cohort of 110 Egyptians, variants that are Egyptian population-specific, Coverage of DNA section based on 10x Genomics, Illumina paired-end and PacBio sequencing data, coverage and reads of RNA sequencing data, BRCA2 gene annotation from Ensembl GTF file, Repeats annotated by REPEATMASKER, SNVs and Indels identified by comparison of assemblies AK1, YOURUBA and EGYPT to GRCh38. The colors denote base substitutions (green), deletions (blue) and insertions (red)

510 **Tables**

511
512 *Table 1: Default assembly quality measures according to Quast-LG. The extended Quast-LG report is provided in*
513 *Supplementary Table 2. Yoruba is a chromosome-level assembly.*

Genome statistics	**EGYPT**	**EGYPT_wtdbg2**	**EGYPT_falcon**	**AK1**	**YORUBA**
Genome fraction (%)	94.174	92.247	95.924	95.177	**95.391**
Duplication ratio	1.01	**0.999**	1.018	1.023	1.088
# genomic features	20,908 (3,226 part)	20,613 (3,229 part)	**21,176 (1578 part)**	21,047 (1,396 part)	21,077 (1,721 part)
Largest alignment	**75,492,126**	**75,492,126**	56,458,009	58,219,133	65,512,502
Total aligned length	2,800,100,449	2,713,712,375	**2,865,356,241**	2,829,006,639	2,832,740,986
NGA50	**11,187,777**	**11,187,777**	8,226,500	13,028,687	19,529,238
LGA50	71	71	95	66	**43**
Misassemblies					
# misassemblies	**1,276**	**1,276**	3,499	1,952	1,756
Misassembled contigs length	**2,137,050,584**	2,137,050,584	2,851,404,290	2,657,569,650	3,053,643,982
Mismatches					
# mismatches per 100 kbp	139	138.72	143.64	**126.92**	141.56
# indels per 100 kbp	32.09	**31.74**	40.06	32.77	46.95
# N's per 100 kbp	**0**	**0**	209.01	1285.7	7180.2
Statistics without reference					
# contigs	3,235	3,106	**1,615**	2,832	1,647
Largest contig	88,566,048	88,566,048	84,324,762	113,921,103	**248,986,603**
Total length	2,836,714,529	2,750,324,638	2,916,268,178	2,904,207,228	**3,088,335,497**
Total length (>= 1000 bp)	2,837,367,164	2,750,799,236	2,916,433,762	2,904,207,228	**3,088,485,407**
Total length (>= 10000 bp)	2,828,723,737	2,742,501,225	2,914,302,309	2,904,207,228	**3,086,359,078**
Total length (>= 50000 bp)	2,803,817,652	2,718,165,929	2,895,137,452	2,855,011,855	**3,059,626,724**
K-mer-based statistics					
K-mer-based compl. (%)	86.01	85.15	**87.75**	87.68	85.82
# k-mer-based misjoins	1,654	1,649	1,786	**1,345**	1,453

514

515

Table 2: Numbers of short and structural variants identified within the cohort of 110 Egyptian individuals. The Percentages refer to the respective higher-order variant category. The number of multi-allelic variants included in these numbers is given.

	Number	Percent	Multi-allelic
Small variants	19,758,992		672,781
→ Indels	2,858,821	14.47	
→ SNVs	16,900,171	85.53	
→ Common SNVs	2,270,642	13.44	1,506
→ Population-specific SNVs	26,564	1.17	37
Structural variant calls	121,141		
→ Deletions	95,889	79.15	
→ Inversions	11,477	9.47	
→ Duplications	10,092	8.33	
→ Translocations	3,275	2.70	
→ Insertions	408	0.34	