An integrated 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*

* These authors contributed equally to this work

Abstract

The human genome is composed of chromosomal DNA sequences consisting of bases A, C, G and T – the blueprint to implement the molecular functions that are the basis of every individual's life. Deciphering the first human genome was a consortium effort that took more than a decade and considerable cost. With the latest technological advances, determining an individual's entire personal genome with manageable cost and effort has come within reach. Although the benefits of the all-encompassing genetic information that entire genomes provide are manifold, only a small number of *de novo* assembled human genomes have been reported to date [1–3], and few have been complemented with population-based genetic variation [4], which is particularly important for North Africans who are not represented in current genome-wide data sets [5–7]. Here, we combine long- and short-read whole-genome next-generation sequencing data with recent assembly approaches into the first *de novo* assembly of the genome of an Egyptian individual. The resulting genome assembly demonstrates well-balanced quality metrics and comes with high-quality variant phasing into maternal and paternal haplotypes, which are linked to various gene expression changes in blood. To construct an Egyptian genome reference, we further assayed genome-wide genetic variation

occurring in the Egyptian population within a representative cohort of more than 100 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 genetic studies and corresponding genome references. The Egyptian genome reference represents a comprehensive population data set based on a high-quality personal genome. It is a proof of concept to be considered by the many national and international genome initiatives underway. More importantly, we anticipate that the Egyptian genome reference will be a valuable resource for precision medicine targeting the Egyptian population and beyond.

Main

With the advent of personal genomics, population-based genetics as part of an individual's genome is indispensable for precision medicine. Currently, genomics-based precision medicine compares the patients' genetic make-up to a reference genome [9], a genome model inferred from individuals 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 genetic risk for disease. Risk alleles and structural variants (SVs) [10] 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 the establishment of national or international multi-ethnic genome projects [6,7,11]. At present, there are several population-based sequencing efforts that aim to map specific variants in the 100,000 genome projects in Asia [12] or England [13]. Furthermore, large-scale sequencing efforts currently explore population, society and history-specific genomic variations in individuals in Northern and Central Europe [14,15], North

America [7], Asia [16,17] and, recently, the first sub-Saharan Africans [18,19]. Nonetheless, there is still little genetic data available for many regions of the world. In particular, North African individuals are not adequately represented in current genetic data sets, such as the 1000 Genomes [5], TOPMED [7] or gnomAD [6] databases. Consequently, imminent health disparities between different world populations have been noted repeatedly for a decade. [20–23]

In recent years, several high-quality *de novo* human genome assemblies [1–4] and, more recently, pan-genomes [8] have extended human sequence information and improved the *de facto* reference genome GRCh38 [9]. Nonetheless, it is still prohibitively expensive to obtain all-embracing genetic information, such as high-quality *de novo* assembled personal genomes for many individuals. Indeed, previous genetic studies assess only a subset of variants occurring in the Egyptian population, e.g., single nucleotide polymorphisms (SNPs) on genotyping arrays [24,25], variants in exonic regions via exome sequencing [26] or variants detectable by short-read sequencing [27,28].

In this study, we generated a phased *de novo* assembly of an Egyptian individual and identified single nucleotide variants (SNVs) and SVs from an additional 109 Egyptian individuals obtained from short-read sequencing. Those were integrated to generate an Egyptian genome reference. We anticipate that an Egyptian population genome reference will strengthen precision medicine efforts that eventually benefit nearly 100 million Egyptians, e.g., by providing allele frequencies (AFs) and linkage disequilibrium (LD) between variants, information that is necessary for both rare and common disease studies. Likewise, our genome will be of universal value for research purposes, since it contains both European and African variant features. Most genome-wide association studies (GWAS) are performed in Europeans [29], but genetic disease risk may differ, especially for individuals of African ancestry [30]. Consequently, an Egyptian genome reference will be well suited to support recent

efforts to include Africans in such genetic studies, for example, by serving as a benchmark data set for SNP array construction and variant imputation or for fine-mapping of disease loci.

Our Egyptian genome is based on a high-quality human *de novo* assembly for one Egyptian individual (see workflow in Suppl. Fig. 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` [31], and another based on the assembly by a novel assembler, `WTDBG2` [32], which has a much shorter run time with 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. dot plots in Suppl. Figs. 3-4). The `WTDBG2`-based assembly was selected as the base because it performs comparable or better according to various quality control (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. Fig. 1). We found our assembly to be comparable to the publicly available assemblies of a Korean [33] and a Yoruba individual (GenBank assembly accession GCA_001524155.4, unpublished) with respect to various QC measures [34] (Table 1). Suppl. Fig. 2 compares the assemblies' NA-values, and Suppl. Figs. 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). EGYPT SNVs and small insertions and deletions (indels) called using short-read sequencing data were phased using high-coverage 10x linked-read sequencing data. This

104 resulted in 98.99% of variants being phased. Furthermore, nearly all (99.41%) of the genes
105 with lengths less than 100 kb and more than one heterozygous SNP were phased into a single
106 phase block.
107
108 Based on the personal Egyptian genome, we constructed an Egyptian population genome by
109 considering genome-wide SNV AFs in 109 additional Egyptians (Suppl. Table 6). This
110 approach enabled the characterization of the major allele (i.e., the allele with highest AF) in
111 the given Egyptian cohort. To accomplish this, we called variants using short-read data of 13
112 Egyptians sequenced at high coverage and 97 Egyptians sequenced at low coverage. Although
113 sequence coverage affects variant-based statistics (Suppl. Fig. 8), due to combined
114 genotyping, most variants could also be called reliably in low coverage samples (Suppl. Fig.
115 9). Altogether, we called a total of 19,758,992 SNVs and small indels (Suppl. Fig. 10) in all
116 110 Egyptian individuals (Fig. 1). The number of called variants per individual varied
117 between 2,901,883 to 3,934,367 and was correlated with sequencing depth (see Suppl. Figs.
118 8-9). This relationship was particularly pronounced for low coverage samples. The majority
119 of variants were intergenic (53.5%) or intronic (37.2%) (Suppl. Fig. 11). Only approximately
120 0.7% of the variants were located within coding exons, of which 54.4% were non-
121 synonymous and thus cause a change in protein sequence and possibly structure (Suppl. Fig.
122 12).
123
124 Using short-read sequencing data of 110 Egyptians, we called 121,141 SVs, most of which
125 were deletions but also included inversions, duplications, insertions and translocations of
126 various orders of magnitude (Fig. 1, Suppl. Fig. 13-14). Similar to SNVs, the number of SV
127 calls also varied between individuals (Suppl. Fig. 15) and is slightly affected by coverage
128 (Suppl. Fig. 16). After merging overlapping SV calls, we obtained an average of 2,773 SVs
129 per Egyptian individual (Suppl. Table 7, Suppl. Figs. 17-19).

To characterize the Egyptian population with respect to European and African populations that have been genotyped within the 1000 Genomes Project [5] (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 components with both Europeans and sub-Saharan Africans (see Fig. 2 and Suppl. Figs. 20-32). Thus far, there are no North African populations with high-quality whole-genome sequencing-based genotype data available, and in the European and sub-Saharan African populations reported by the 1000 Genomes Project, Egyptians are closest to the European Tuscany population (see Fig. 2 and Suppl. Figs. 20-32), which has been previously proposed through genetic studies of ancient Egyptian mummies [35].

The mixed European and African ancestry of Egyptians is further supported by mitochondrial haplogroup assessment from the literature [27] and our own mtDNA sequencing data (overall n=327). We found that Egyptians have haplogroups most frequently found in Europeans (e.g., H, V, T, J, etc.; more than 60%), and many Egyptians also have African (e.g., L with 24.8%) or Asian/East Asian haplogroups (e.g., M with 6.7%). This indicates that Egyptian genomes contain genetic components from various major human populations (Suppl. Fig. 33), as has been shown recently for a few Egyptian individuals in a study that performed genome-wide admixture analysis of populations from the Arabian Peninsula using SNP arrays [25].

In total, we identified 6,599,037 common Egyptian SNVs (minor allele frequency (MAF) > 5%, genotypes in a minimum of 100 individuals), of which 1,198 are population-specific; i.e., they are either rare (MAF < 1%) or not detected in any other population in the 1000 Genomes [5] and gnomAD databases [6] as well as TOPMed [7] (Suppl. Table 9). These numbers are comparable to population-specific variant numbers reported previously for 1000 Genomes

156 populations [36]. Four SNVs likely have a molecular impact (Suppl. Table 10), indicated by a
157 CADD [37] deleteriousness score greater than 20. SNP rs143563851 (CADD 24.2) has recently
158 been identified in 1% of individuals of a cohort of 211 Palestinians in a study that performed
159 targeted sequencing of blood group antigen synthase GBGT1 [38]. SNP rs143614333 (missense
160 variant in gene CR2, CADD 23.6) is in ClinVar [39], with three submitters reporting that the
161 variant is of uncertain clinical significance. Additionally, we obtained 49 variants with no
162 dbSNP [40] rsID (Suppl. Table 11). These numbers of population-specific SNPs, of which some
163 likely have an immediate impact on clinical characteristics and diagnostics, indicate
164 insufficient coverage of the genetic diversity of the world's population for precision medicine
165 and thus the need for local genome references. To detect a putative genetic contribution of
166 Egyptian population-specific SNPs towards molecular pathways, phenotypes or disease, we
167 performed gene set enrichment analysis for all 461 protein-coding genes that were annotated
168 to population-specific SNPs by Ensembl VEP [41]. Enrichr, a gene list enrichment tool
169 incorporating 153 gene sets and pathway databases [42], reports that genes from obesity-related
170 traits of the GWAS catalog 2019 collection are over-represented (adj. p-value: 1.02E-6; 49 of
171 804 genes), which might hint at population-specific metabolism regulation that is linked to
172 body weight.
173
174 Variants that are not protein coding may have a regulatory effect that affects gene and
175 eventually protein expression. Using blood expression data obtained from RNA sequencing
176 for the EGYPT assembly individual in conjunction with 10x sequencing-based phased variant
177 data, we identified genes whose expression differs between maternal and paternal haplotypes
178 (see Suppl. Fig. 34 for the analysis overview and Suppl. Figs. 35-36 for the results). We
179 report 1,180 such genes (Suppl. Table 12). Of these, variants contained in haplotypes of 683
180 genes (58%) have previously reported expression quantitative trait loci (eQTLs) in blood
181 according to Qtlizer [43], for 380 genes supported by multiple studies. For 370 genes (31%), the

182 strongest associated blood eQTL SNV is haplotypically expressed, and for 131 genes, the best
183 eQTL has been previously reported by multiple studies. Concordance of haplotypic
184 expression with eQTLs indicates that a common variant may affect gene expression;
185 discordance hints towards a rare variant.
186
187 We investigated the impact of Egyptian ancestry on disease risk by integrating Egyptian
188 variant data with the GWAS catalog [44], a curated database of GWAS. According to the
189 GWAS catalog, most published GWAS are performed in Europeans [29], and only a single
190 study has been performed in Egyptians [45] (by one of our groups). Furthermore, only 2% of
191 individuals included in GWAS are of African ancestry [29]. AFs, LD and genetic architecture
192 can differ between populations, such that results from European GWAS cannot necessarily be
193 transferred [30]. This lack of transferability also compromises the prediction of an individual's
194 traits and disease risk using polygenic scores: such scores are estimated to be approximately
195 one-third as informative in African individuals compared to Europeans [46]. From the GWAS
196 catalog, we constructed a set of 4,008 different, replicated, high-quality tag SNPs (i.e., one
197 strongest associated SNP per locus) from European ancestry GWAS for 584 traits and
198 diseases. We compared the tag SNPs' AFs and proxy SNPs in the Egyptian cohort (n=110)
199 and Europeans from 1000 Genomes (n=503) (Suppl. Table 13). Egyptian AFs of tag SNPs are
200 comparable to European AFs, with a tendency to be lower (Fig. 3a). There are variants
201 common in Europeans (AF>5%) but rare in Egyptians (AF<5%) (Suppl. Fig. 37). A total of
202 261 tag SNPs are not present in the Egyptian cohort (~7%), clearly indicating a need to
203 perform GWAS in non-European populations to further elucidate disease risk conferred by
204 these loci. We investigated differences in LD structure using an approach that is used for fine-
205 mapping of GWAS data, which identifies proxy variants (illustrated in Fig. 3c). Proxy
206 variants are variants correlated with the tag GWAS SNP, i.e., in high LD (here, R^2>0.8). The
207 post-GWAS challenge is the identification of a causal variant from a set of variants in LD (tag

208 SNP and proxy variants). We found that the number of proxy variants was much lower in the
209 Egyptian cohort (Fig. 3b), likely due to shorter haplotype blocks known from African
210 populations. This indicates that LD differences between Egyptians and Europeans may
211 compromise GWAS transferability and European ancestry-based polygenic scores. However,
212 Egyptian proxy variants are usually included in the larger set of European proxy variants (Fig.
213 3d). An example is variant rs2075650 (a locus sometimes attributed to gene TOMM40),
214 which has been linked to Alzheimer's disease in seven GWASs (cf. Suppl. Fig. 38). This tag
215 SNP has seven proxy variants in Europeans but only two proxy variants in Egyptians. One
216 European proxy, rs72352238, has also been reported as a GWAS tag SNP, but it is not a
217 proxy of rs2075650 in Egyptians and may thus fail replication and transfer of GWAS results
218 from the European to the Egyptian population.
219
220 With our Egyptian genome reference, it will be possible to perform comprehensive integrated
221 genome and transcriptome comparisons for Egyptian individuals and shed light on personal as
222 well as population-wide common genetic variants. Fig. 4 visualizes the various types of data
223 of this resource in the integrative genomics viewer [47] (IGV). Here, we selected the DNA
224 repair-associated gene BRCA2, which, if mutated, is linked to breast cancer and other cancer
225 types [48]. IGV depicts the sample coverage based on sequencing data from PacBio, 10x
226 Genomics and Illumina (whole genome as well as RNA) for the personal EGYPT genome
227 together with common Egyptian SNPs. Variants previously assessed in a breast cancer
228 GWAS [48] are displayed as Manhattan plot; note the three significant GWAS SNPs between
229 positions 32,390 and 32,400 kb. The bottom compares the identified SNVs and indels from
230 the Korean and Yoruba *de novo* assembly with our *de novo* EGYPT assembly. Visual
231 inspection already yields significantly different variants. This integrative view sheds light on
232 both small and structural variations at the personal and population-based genome levels.
233

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.

Methods

Sample acquisition

Samples were acquired from 10 Egyptian individuals. 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 data. Furthermore, we used public Illumina short-read data from 100 Egyptian individuals from Pagani *et al.* [27]. 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 298.2 GB of data.

Sequencing data from five PacBio libraries were generated at overall 99x genome coverage.

Illumina short-read data generation

For 350 bp library construction, the genomic DNA was sheared, and fragments with sizes of approximately 350 bp were purified from agarose gels. The fragments were ligated to

adaptors and amplified using PCR. The generated libraries were then sequenced on the Illumina HiSeq X Ten using PE150 and generated 312.8 GB of data.

For the assembly individual, sequencing data from five libraries was generated at overall 90x genome coverage. For nine additional individuals, one library each was generated, amounting to an overall 305x coverage of sequencing data. For the 100 individuals of Pagani *et al.* [27], three were sequenced at high coverage (30x) and 97 at low coverage (8x). The 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, double-stranded cDNA was synthesized, and then adaptors were ligated. The second strand of cDNA was then degraded to generate a directional library. The generated libraries with insert sizes 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 overall 80x genome coverage.

Construction of draft *de novo* assemblies and meta-assembly

We used `WTDBG2` [32] 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 [49] with short-read data (cf. Suppl. Methods: *WTDBG2-based assembly*).

An alternative assembly was generated by using FALCON [50], QUIVER [51], SSPACE-LONGREAD [52], PBJELLY [53], FRAGSCAFF [54] and PILON [49] (cf. Suppl. Methods: *FALCON-based assembly*).

Proceeding from the WTDBG2-based assembly, we constructed a meta-assembly. Regions larger than 800 kb 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 Suppl. Fig. 1 for an overview of our assembly strategy, including meta-assembly construction (cf. Suppl. Methods: *Meta-assembly construction*).

Assembly quality and characteristics were assessed with QUAST-LG [55] (cf. Suppl. Methods: *Assembly comparison and QC*). The extraction of coordinates for meta-assembly construction was performed using QUAST-LG output.

Repeatmasking

Repeatmasking was performed by using REPEATMASKER [56] with RepBase version 3.0 (Repeatmasker Edition 20181026) and Dfam_consensus (http://www.dfam-consensus.org) (cf. Suppl. Methods: *Repeat annotation*).

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 data and the 10x Genomics LONGRANGER WGS pipeline with four 10x libraries provided for one combined phasing. See Supplementary Methods *Variant phasing* for details.

312 **SNVs and small indels**

313 Calling of SNVs and small indels was performed with `GATK` 3.8 [57] using the parameters of
314 the best practice workflow. Reads in each read group were trimmed using `Trimmomatic` [58]
315 and subsequently mapped against reference genome hg38 using `BWA-MEM` [59] version 0.7.17.
316 Then, the alignments for all read groups were merged sample-wise and marked for duplicates.
317 After the base recalibration, we performed variant calling using `HaplotypeCaller` to
318 obtain GVCF files. These files were input into GenotypeGVCFs to perform joint genotyping.
319 Finally, the variants in the outputted VCF file were recalibrated, and only those variants that
320 were flagged as "PASS" were kept for further analyses. We used `FastQC` [60], `Picard`
321 `Tools` [61] and `verifyBamId` [62] for QC (cf. Suppl. Methods: *Small variant QC*).
322
323 **Variant annotation**
324 Variant annotation was performed using `ANNOVAR` [63] and `VEP` [41] (cf. Suppl. Methods: *Small*
325 *variant annotation*)
326
327 **Structural variants**
328 SVs were called using `DELLY2` [64] with default parameters as described on the `DELLY2`
329 website for germline SV calling (https://github.com/dellytools/delly) (cf. Suppl. Methods:
330 *Structural variant QC*). Overlapping SV calls in the same individual were collapsed by the
331 use of custom scripts. See Supplementary Methods *Collapsing structural variants* for details.
332
333 **Genotype-based principal components**
334 1000 Genomes phase 3 variant data were obtained for all European and African individuals
335 and merged with the Egyptian variant data. Variants were excluded if their MAF was less
336 than 5% among individuals in the 1000 Genomes database, they violated Hardy-Weinberg
337 equilibrium, or they were multi-allelic or within regions of high LD and/or of known

338 inversions. LD pruning was performed, and the remaining SNPs passed on to the `SMARTPCA`
339 program [65] of the `EIGENSOFT` package for PC computation. See Supplementary Methods
340 *Genotype-based principal components* for details.
341
342 **Mitochondrial haplogroups**
343 Haplogroup assignment was performed for 227 individuals using `HAPLOGREP 2` [66].
344 Furthermore, mitochondrial haplogroups were obtained from Pagani *et al.* [27] for 100
345 individuals. See Supplementary Methods *Mitochondrial haplogroups* for details.
346
347 **Population-specific variants**
348 Our set of common Egyptian SNVs comprises variants with genotypes in a minimum of 100
349 individuals whose alternative allele has a frequency of more than 5%. Those common
350 Egyptian SNVs that are otherwise rare, i.e., have an AF of less than 1% in the 1000 Genomes,
351 and gnomAD populations as well as in TOPMed were considered Egyptian-specific. AFs
352 were annotated using the Ensembl API. Furthermore, a list of Egyptian common variants
353 without dbSNP rsID was compiled, see Supplementary Methods *Small variant annotation* for
354 details.
355
356 **Haplotypic expression analysis**
357 RNA sequencing reads were mapped and quantified using `STAR` (Version 2.6.1.c) [67].
358 Haplotypic expression analysis was performed by using `PHASER` and `PHASER GENE AE`
359 (version 1.1.1) [68] with Ensembl version 95 annotation on the 10x-phased haplotypes using
360 default parameters. See Supplementary Methods *Haplotypic expression* for details.
361
362
363

GWAS catalog data integration

GWAS catalog associations for GWAS of European ancestry were split into trait-specific data sets using Experimental Factor Ontology (EFO) 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.*

Integrative genomics view

We implemented a workflow to extract all Egyptian genome reference data for view in the IGV [47]. 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 Supplementary Methods *Gene-centric integrative data views* for details.

Ethics statement

This 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. This study and its results are in accordance with the Jena Declaration (https://www.uni-jena.de/unijenamedia/Universit%C3%A4t/Abteilung+Hochschulkommunikation/Presse/Jenaer+Erkl%C3%A4rung/Jenaer_Erklaerung_EN.pdf).

References

1. Cao, H. *et al.* De novo assembly of a haplotype-resolved human genome. *Nat. Biotechnol.* **33**, 617–622 (2015).

2. Seo, J.-S. *et al. De novo* assembly and phasing of a Korean human genome. *Nature* **538**, 243–247 (2016).
3. Shi, L. *et al.* Long-read sequencing and de novo assembly of a Chinese genome. *Nat Commun* **7**, 12065 (2016).
4. Cho, Y. S. *et al.* An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes. *Nat Commun* **7**, 13637 (2016).
5. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
6. Karczewski, K. J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv* 531210 (2019) doi:10.1101/531210.
7. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *bioRxiv* 563866 (2019) doi:10.1101/563866.
8. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).
9. Levy-Sakin, M. *et al.* Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nature Communications* **10**, 1025 (2019).
10. Stark, Z. *et al.* Integrating Genomics into Healthcare: A Global Responsibility. *Am. J. Hum. Genet.* **104**, 13–20 (2019).
11. GenomeAsia 100k. *GenomeAsia 100k* http://www.genomeasia100k.com/.
12. Turnbull, C. *et al.* The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. *BMJ* **361**, k1687 (2018).
13. Maretty, L. *et al.* Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. *Nature* **548**, 87–91 (2017).

14. Leslie, S. *et al.* The fine-scale genetic structure of the British population. *Nature* **519**, 309–314 (2015).

15. Chiang, C. W. K., Mangul, S., Robles, C. & Sankararaman, S. A Comprehensive Map of Genetic Variation in the World's Largest Ethnic Group-Han Chinese. *Mol. Biol. Evol.* **35**, 2736–2750 (2018).

16. Bai, H. *et al.* Whole-genome sequencing of 175 Mongolians uncovers population-specific genetic architecture and gene flow throughout North and East Asia. *Nat. Genet.* (2018) doi:10.1038/s41588-018-0250-5.

17. Sherman, R. M. *et al.* Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat. Genet.* (2018) doi:10.1038/s41588-018-0273-y.

18. Choudhury, A. *et al.* Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. *Nat Commun* **8**, 2062 (2017).

19. Bustamante, C. D., Burchard, E. G. & De la Vega, F. M. Genomics for the world. *Nature* **475**, 163–165 (2011).

20. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–164 (2016).

21. Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).

22. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).

23. Sherman, R. M. *et al.* Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nature Genetics* **51**, 30 (2019).

24. Henn, B. M. *et al.* Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genet.* **8**, e1002397 (2012).

25. Fernandes, V. *et al.* Genome-Wide Characterization of Arabian Peninsula Populations: Shedding Light on the History of a Fundamental Bridge between Continents. *Mol. Biol. Evol.* **36**, 575–586 (2019).

26. Scott, E. M. *et al.* Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. *Nat. Genet.* **48**, 1071–1076 (2016).

27. Pagani, L. *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* **96**, 986–991 (2015).

28. ElHefnawi, M. *et al.* Whole genome sequencing and bioinformatics analysis of two Egyptian genomes. *Gene* **668**, 129–134 (2018).

29. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 1080 (2019).

30. Kim, M. S., Patel, K. P., Teng, A. K., Berens, A. J. & Lachance, J. Genetic disease risks can be misestimated across global populations. *Genome Biol.* **19**, 179 (2018).

31. Chin, C.-S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods* **13**, 1050–1054 (2016).

32. Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *bioRxiv* 530972 (2019) doi:10.1101/530972.

33. Seo, J.-S. *et al.* De novo assembly and phasing of a Korean human genome. *Nature* **538**, 243–247 (2016).

34. Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D. & Gurevich, A. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* **34**, i142–i150 (2018).

35. Schuenemann, V. J. *et al.* Ancient Egyptian mummy genomes suggest an increase of Sub-Saharan African ancestry in post-Roman periods. *Nature Communications* **8**, ncomms15694 (2017).

464 36. Choudhury, A. *et al.* Population-specific common SNPs reflect demographic histories and
465 highlight regions of genomic plasticity with functional relevance. *BMC Genomics* **15**, 437
466 (2014).
467 37. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting
468 the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**,
469 D886–D894 (2019).
470 38. Abusibaa, W. A. *et al.* Expression of the GBGT1 Gene and the Forssman Antigen in Red
471 Blood Cells in a Palestinian Population. *Transfusion Medicine and Hemotherapy* (2019)
472 doi:10.1159/000497288.
473 39. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting
474 evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
475 40. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**,
476 308–311 (2001).
477 41. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biology* **17**, 122
478 (2016).
479 42. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server
480 2016 update. *Nucleic Acids Res.* **44**, W90-97 (2016).
481 43. Munz, M. *et al.* Qtlizer: comprehensive QTL annotation of GWAS results. *bioRxiv*
482 495903 (2019) doi:10.1101/495903.
483 44. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide
484 association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**,
485 D1005–D1012 (2019).
486 45. Bejaoui, Y. *et al.* Genome-wide association study of psoriasis in an Egyptian population.
487 *Exp. Dermatol.* **28**, 623–627 (2019).
488 46. Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse
489 human populations. *Nat Commun* **10**, 3328 (2019).

47. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).

48. Michailidou, K. *et al.* Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92–94 (2017).

49. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).

50. Chin, C.-S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).

51. Chin, C.-S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).

52. Boetzer, M. & Pirovano, W. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics* **15**, (2014).

53. English, A. C. *et al.* Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. *PLoS ONE* **7**, e47768 (2012).

54. Adey, A. *et al.* In vitro, long-range sequence information for de novo genome assembly via transposase contiguity. *Genome Research* **24**, 2041–2049 (2014).

55. Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D. & Gurevich, A. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* **34**, i142–i150 (2018).

56. SMIT, A. F. A. Repeat-Masker Open-3.0. *http://www.repeatmasker.org* (2004).

57. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297–1303 (2010).

58. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

59. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).

60. Andrews, S. FASTQC - A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

61. Picard Toolkit. http://broadinstitute.github.io/picard/.

62. Jun, G. *et al.* Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data. *The American Journal of Human Genetics* **91**, 839–848 (2012).

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

64. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).

65. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).

66. Kloss-Brandstätter, A. *et al.* HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Hum. Mutat.* **32**, 25–32 (2011).

67. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* bts635 (2012) doi:10.1093/bioinformatics/bts635.

68. Castel, S. E., Mohammadi, P., Chung, W. K., Shen, Y. & Lappalainen, T. Rare variant phasing and haplotypic expression from RNA sequencing with phASER. *Nat Commun* **7**, 12817 (2016).

Supplementary information

Supplementary Tables 1-13: An_Egyptian_genome_reference_supplementary_tables.xlsx

Supplementary Methods and Supplementary Figures 1-38:

An_Egyptian_genome_reference_supplement.pdf

Acknowledgements

We acknowledge support on coordination of the project and assembly work w.r.t the `FALCON`-based assembly through Ms. Lu Wang from the Novogene (UK) Company Limited. HB and IW acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany`s Excellence Strategy – EXC 22167-390884018. All authors acknowledge computational support from the OMICS compute cluster at the University of Lübeck.

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

Caixia Ma

Medical Experimental Research Center (MERC), Mansoura University, Mansoura, Egypt

Mohamed Salama & Shaaban El-Mosallamy

Institute of Global Health and Human Ecology, The American University in Cairo, Cairo, Egypt

Mohamed Salama

Genetics Division, Lübeck Institute of Experimental Dermatology, University of Lübeck, Lübeck, Germany

Misa Hirose & Saleh Ibrahim

Contributions

H.B, S.I. and 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. M.H. performed mtDNA library preparation and sequencing. 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.

Data availability

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

Code availability

Computational tools and parameters used are specified in the Supplementary Methods. Workflows have been implemented to permit reproducible data analyses by using Snakemake as workflow management system, Git for version control of workflow code and Conda (especially Bioconda) for managing software environments.

593 **Corresponding authors**

594 Correspondence to Hauke Busch or Saleh Ibrahim.

595

596 *Table 1: Default assembly quality measures according to* `QUAST-LG`*. The extended* `QUAST-LG` *report is*
597 *provided in Suppl. Table 2. Yoruba is a chromosome-level assembly. Best quality for every measure is denoted in*
598 *bold.*

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
	20,908	20,613	**21,176**	21,047	21,077
# genomic features	(3,226 part)	(3,229 part)	**(1578 part)**	(1,396 part)	(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

599

600

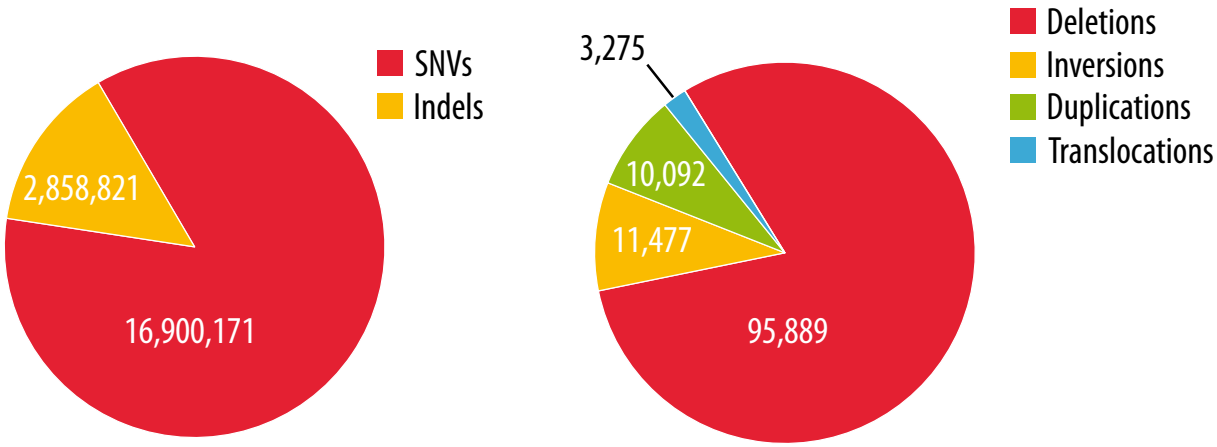
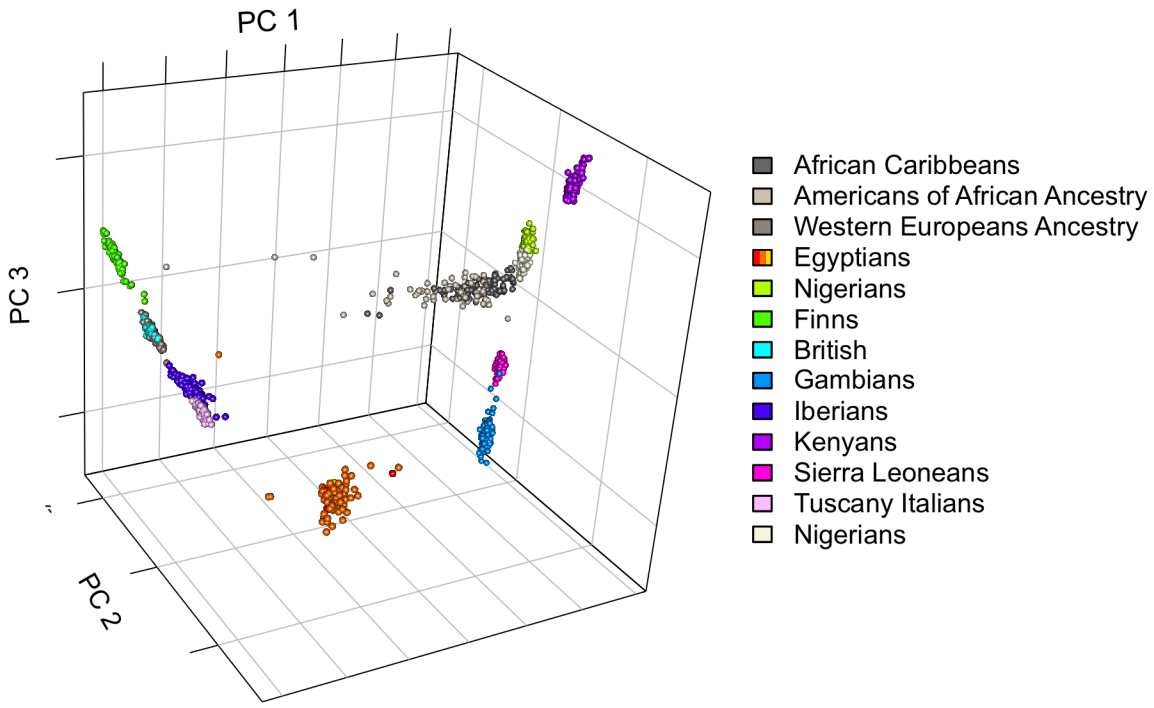


Figure 1: Number of various genetic variant types identified in the Egyptian cohort. Left: The number of SNVs and indels. Right: The number of SV calls: deletions, inversions, duplications and translocations. Additionally, 408 insertions have been called.



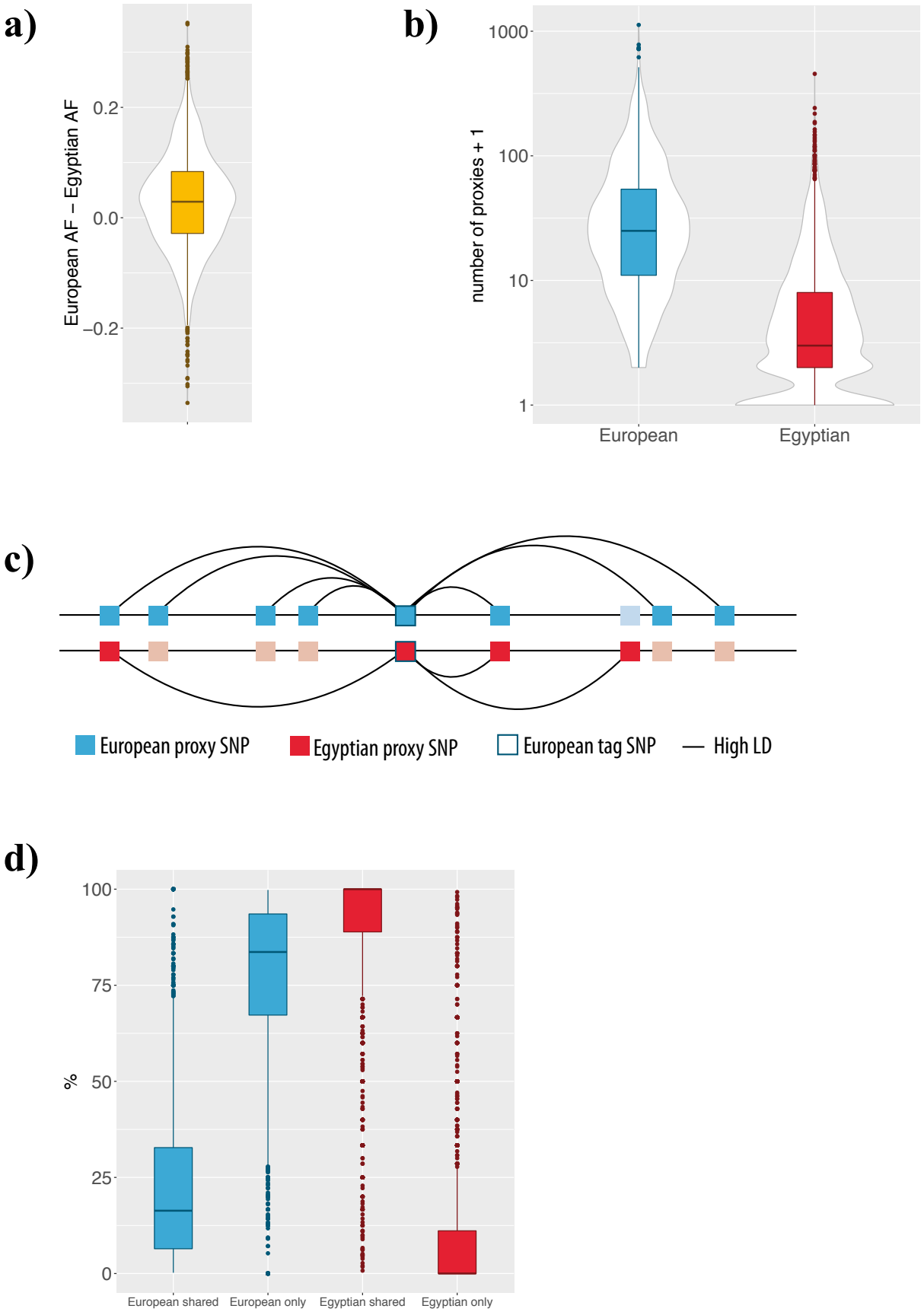
606

607 *Figure 2: PCA plot of different populations from the 1000 Genomes Project and 110 Egyptian genomes from*

608 *Pagani et al. as well as from our own study.*

609

610



611

612 *Figure 3: AF and proxy SNP comparisons for 3,698 GWAS tag SNPs called in a minimum of 100 Egyptians. a)*
613 *AF differences. b) Number of proxies. c) Illustration of the proxy SNP comparison. A European GWAS tag SNP*
614 *(center) and variants in Europeans (top) and Egyptians (bottom). Lines denote variants in high LD. The tag SNP*
615 *has 7 proxy variants in Europeans and 3 in Egyptians. Light blue/red variants are no proxy variants in*
616 *Europeans/Egyptians. Two proxy variants are shared. Thus 2 of 7 European (~29%) and 2 of 3 Egyptian (~67%)*
617 *variants are shared. Further 5 of 7 European proxies are European-only (~71%) and 1/3 Egyptian proxies are*
618 *Egyptian-only (~33%). d) European shared: Percentage of European proxy SNPs shared with Egyptian proxy*
619 *SNPs. European only: Percentage of European proxy SNPs not shared with Egyptian proxies. Egyptian shared /*
620 *Egyptian only respectively.*

621

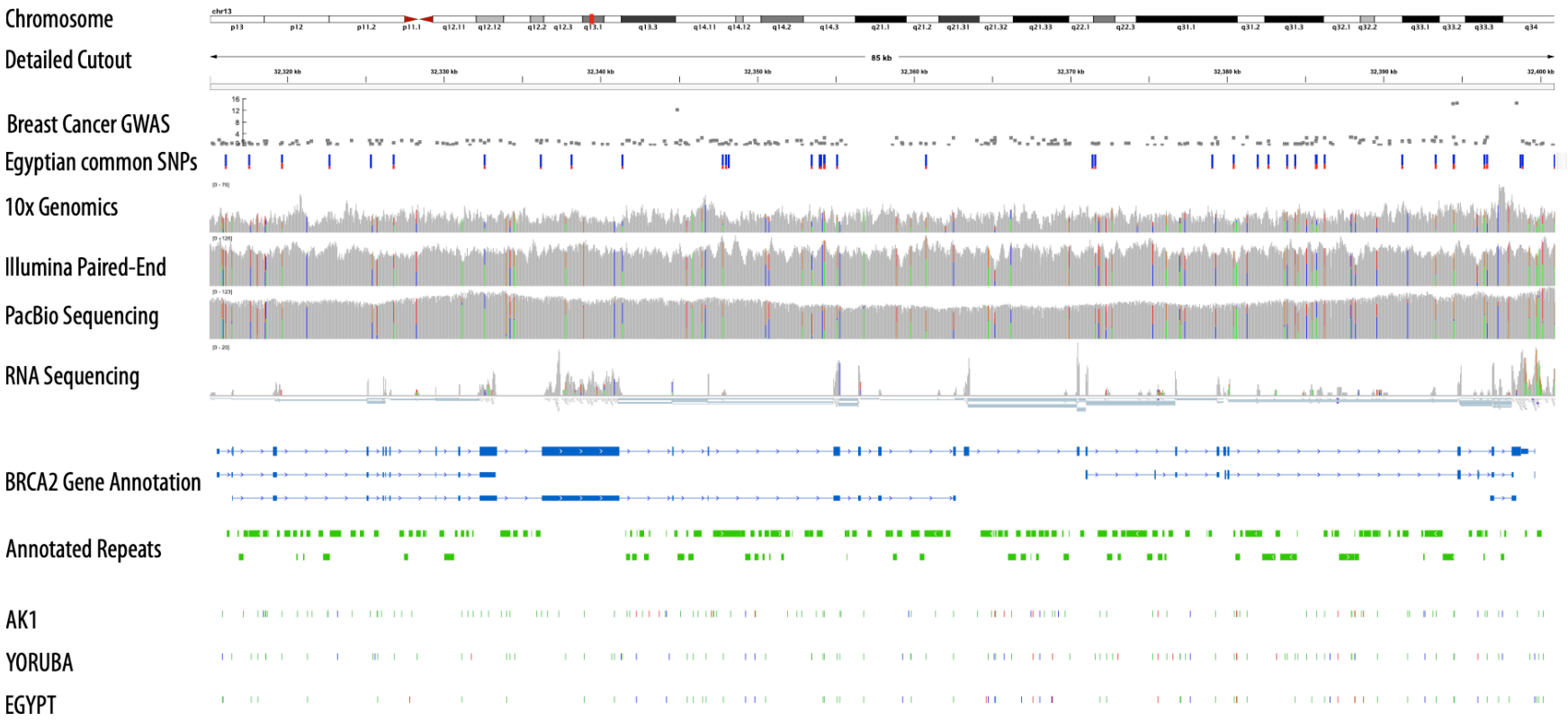


Figure 4: Integrative view of Egyptian genome reference data for the gene BRCA2, which is associated with breast cancer. The rows denote from top to bottom: Genome location on chromosome 13 of the magnified region for BRCA2 (first and second row); GWAS data for breast cancer risk [48]; Variants that are common in the cohort of 110 Egyptians; Read coverage of genetic region based on 10x Genomics, Illumina paired-end and PacBio sequencing data; Coverage and reads of RNA sequencing data; BRCA2 gene annotation from Ensembl; Repeats annotated by `REPEATMASKER`*; SNVs and indels identified by comparison of assemblies AK1, YOURUBA and EGYPT with GRCh38. The colors denote base substitutions (green), deletions (blue) and insertions (red).*