An integrated personal and population-based Egyptian genome reference

Inken Wohlers, Axel Künstner, Matthias Munz, Michael Olbrich, Anke Fähnrich, Verónica Calonga-Solís, 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 assembly demonstrates well-balanced quality metrics and is complemented with high-quality variant phasing via linked reads into haploblocks, which we can associate with gene expression changes in blood. To construct an Egyptian genome

26 reference, we further assayed genome-wide genetic variation occurring in the Egyptian
27 population within a representative cohort of 110 Egyptian individuals. We show that differences
28 in allele frequencies and linkage disequilibrium between Egyptians and Europeans may
29 compromise the transferability of European ancestry-based genetic disease risk and polygenic
30 scores, substantiating the need for multi-ethnic genetic studies and corresponding genome
31 references. The Egyptian genome reference represents a comprehensive population data set
32 based on a high-quality personal genome. It is a proof of concept to be considered by the many
33 national and international genome initiatives underway. More importantly, we anticipate that
34 the Egyptian genome reference will be a valuable resource for precision medicine targeting the
35 Egyptian population and beyond.
36
37
38 **Main**
39 With the advent of personal genomics, population-based genetics as part of an individual's
40 genome is indispensable for precision medicine. Currently, genomics-based precision medicine
41 compares the patients' genetic make-up to a reference genome [8], a genome model inferred from
42 individuals of mostly European descent, to detect risk mutations that are related to disease.
43 However, genetic and epidemiologic studies have long recognized the importance of ancestral
44 origin in conferring genetic risk for disease. Risk alleles and structural variants (SVs) [9] can be
45 missing from the reference genome or can have different population frequencies, such that
46 alternative pathways become disease related in patients of different ancestral origin, which
47 motivates the establishment of national or international multi-ethnic genome projects [6,7,10]. At
48 present, there are several population-based sequencing efforts that aim to map specific variants
49 in the 100,000 genome projects in Asia [11] or England [12]. Furthermore, large-scale sequencing
50 efforts currently explore population, society and history-specific genomic variations in
51 individuals in Northern and Central Europe [13,14], North America [7], Asia [15,16] and, recently, the

52 first sub-Saharan Africans [17,18]. Nonetheless, there is still little genetic data available for many
53 regions of the world. In particular, North African individuals are not adequately represented in
54 current genetic data sets, such as the 1000 Genomes [5], TOPMED [7] or gnomAD [6] databases.
55 Consequently, imminent health disparities between different world populations have been
56 noted repeatedly for a decade. [19–22]
57
58 In recent years, several high-quality *de novo* human genome assemblies [1–4] and, more recently,
59 pan-genomes [23] have extended human sequence information and improved the *de facto*
60 reference genome GRCh38 [8]. Nonetheless, it is still prohibitively expensive to obtain all-
61 embracing genetic information, such as high-quality *de novo* assembled personal genomes for
62 many individuals. Indeed, previous genetic studies assess only a subset of variants occurring in
63 the Egyptian population, e.g., single nucleotide polymorphisms (SNPs) on genotyping arrays
64 [24,25], variants in exonic regions via exome sequencing [26] or variants detectable by short-read
65 sequencing [27,28].
66
67 In this study, we generated a *de novo* assembly of an Egyptian individual and identified single
68 nucleotide variants (SNVs) and SVs from an additional 109 Egyptian individuals obtained from
69 short-read sequencing. Those were integrated to generate an Egyptian genome reference. We
70 anticipate that an Egyptian population genome reference will strengthen precision medicine
71 efforts that eventually benefit nearly 100 million Egyptians, e.g., by providing allele
72 frequencies (AFs) and linkage disequilibrium (LD) between variants, information that is
73 necessary for both rare and common disease studies. Likewise, our genome will be of universal
74 value for research purposes, since it contains both European and African variant features. Most
75 genome-wide association studies (GWAS) are performed in Europeans [29], but genetic disease
76 risk may differ, especially for individuals of African ancestry [30]. Consequently, an Egyptian
77 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 male 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 further polishing tools. For the `FALCON`-based assembly, scaffolding was performed, whereas we found that the `WTDBG2`-based assembly was of comparable accuracy without scaffolding (Table 1). Sex chromosomal sequences have not been manually curated. The `WTDBG2`-based assembly was selected as the meta assembly basis, because it performs similarly or better than the `FALCON`-based assembly, according to various quality control (QC) measures. The former did not require scaffolding, and thus there are no N bases or scaffolding errors. Overall, it has about 50% fewer misassemblies. This QC measure holds true even when ignoring misassemblies in centromeres and in segmental duplications and after correction for structural variants (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). Both the base assemblies and the final meta assembly are of high quality and complementary and they are comparable to the publicly available assemblies of a Korean [2] and a Yoruba (GenBank assembly accession GCA_001524155.4) individual in terms of genome length and various quality control (QC) measures, (Table 1, extended version in Suppl. Table 2). Assembly quality is confirmed by quality control (QC) measures assessed by `QUAST-LG`

104 [33] (Suppl. Table 2), NA-values (Suppl. Fig. 2), k-mer multiplicity with `KAT` [34] (Suppl. Fig. 3,
105 4 and 5), QV values of more than 40 and by dot plots of alignment with reference GRCh38
106 (Suppl. Figs. 6-10).
107 We performed repeat annotation and repeat masking for all assemblies (Suppl. Table 4).
108
109 The meta-assembly was complemented with high-quality phasing information (Suppl. Table
110 5). EGYPT SNVs and small insertions and deletions (indels) called using short-read sequencing
111 data were phased using high-coverage 10x linked-read sequencing data. This resulted in
112 3,834,900 of 4,008,080 autosomal variants being phased (95.7%). Furthermore, nearly all
113 (99.41%) of the genes with lengths less than 100 kb and more than one heterozygous SNP were
114 phased into a single phase block. We identified 22 runs of homozygosity (ROH) (Suppl. Table
115 6), out of which 16 are larger than 5 Mb and sum up to overall 192 MB, which indicates
116 consanguinity at the level of parental third-degree relationship [35].
117 Based on the personal Egyptian genome, we constructed an Egyptian population genome by
118 considering genome-wide SNV AFs in 109 additional Egyptians (Suppl. Table 7). This
119 approach enabled the characterization of the major allele (i.e., the allele with highest AF) in the
120 given Egyptian cohort. To accomplish this, we called variants using short-read data of 13
121 Egyptians sequenced at high coverage and 97 Egyptians sequenced at low coverage. Although
122 sequence coverage affects variant-based statistics (Suppl. Fig. 11), due to combined
123 genotyping, most variants could also be called reliably in low coverage samples (Suppl. Fig.
124 12). We called a total of 19,758,992 SNVs and small indels (Suppl. Fig. 13) in all 110 Egyptian
125 individuals (Fig. 1). The number of called variants per individual varied between 2,901,883 to
126 3,934,367 and was correlated with sequencing depth (see Suppl. Figs. 11-12). This relationship
127 was particularly pronounced for low coverage samples. The majority of variants were intergenic
128 (53.5%) or intronic (37.2%) (Suppl. Fig. 14). Only approximately 0.7% of the variants were

129 located within coding exons, of which 54.4% were non-synonymous and thus cause a change
130 in protein sequence and, possibly, structure (Suppl. Fig. 15).
131
132 Using short-read sequencing data of 110 Egyptians, we called 121,141 SVs, most of which
133 were deletions, but also included inversions, duplications, insertions and translocations of
134 various orders of magnitude (Fig. 1, Suppl. Fig. 16-17). Similar to SNVs, the number of SV
135 calls also varied between individuals (Suppl. Fig. 18) and is slightly affected by coverage
136 (Suppl. Fig. 19). After merging overlapping SV calls, we obtained an average of 2,773 SVs per
137 Egyptian individual (Suppl. Table 8, Suppl. Figs. 20-22).
138
139 Using the EGYPT *de novo* assembly, we searched for unique insertions that are common in
140 Egyptians. Towards this, we first mapped all short-read data against the GRCh38 reference
141 genome and to other decoy or alternative haplotype sequences from the GATK bundle. All
142 reads that could not be mapped were subsequently mapped against the EGYPT *de novo*
143 assembly. A similar approach was recently applied to identify novel, unique insertions in *de*
144 *novo* assemblies of 17 individuals from 5 populations using 10x genomics sequencing [36].
145 Altogether we identified 40 unique insertions longer than 500 bp with a total length of 40kb,
146 for which we required for every base in the identified region to have a minimal coverage of 5
147 reads in at least 10 Egyptian individuals (Suppl. Table 9). Of these sequences, 28 have been
148 mentioned before by Wong *et al.* [36], and 10 more in different studies within the last 15 years [37]
149 [38] [39] [40]. Two out of the 40 insertions are most likely novel. In addition, one region contains three
150 unique insertions, of which two contain additional, novel sequences longer than 500 bases.
151 Closer inspection reveals that these sequences are located within a region of two 50 kb gaps
152 (i.e. N sequences) in the GRCh38 reference genome at chromosome loci chr13:111,703,856-
153 111,753,855 and chr13:111,793,442-111,843,441 with about 40 kb of reference sequence
154 between the gaps. The EGYPT, AK1 and YORUBA assembly sequences that cover this 140

155 kb reference sequence from chr13:111,703,856 to 111,843,441 are very similar (Suppl. Figs.
156 23, 24 and 25). They all align about 4 kb from the 40 kb reference sequence between the gaps,
157 only, but at the very beginning of the respective assembly sequence (Suppl. Figs. 26, 27 and
158 28). Performing a BLAST search of the 140 kb EGYPT assembly sequence reveals an overall
159 44 kb alignment in five, mainly consecutive, large alignment blocks to "Homo sapiens
160 chromosome 13 clone WI2-2182D8" (AC188786.1) from position 1 to 44,382, see Suppl. Fig.
161 29. This large reference genome region that contains the largest gap covering sequence reported
162 for AK1 [2] is not resolved yet.
163 Overall, we identified 330 single nucleotide variants and indels in 36 of 40 non-reference
164 sequences (Suppl. Table 10). The percentage of reads that could not be mapped to GRCh38 or
165 GATK bundle sequences, but which were mappable against the *de novo* assembly is on average
166 8.6%, but for some individuals up to 34.2% (cv. Suppl. Fig. 30). Previously unmapped short
167 reads of 110 Egyptians covered positions for more than 19 Mb of the Egyptian *de novo*
168 assembly. Unique sequences that are commonly shared among Egyptians illustrate that
169 additional reference genomes are needed to capture the genetic diversity that are neither
170 assessable by short read sequencing nor with the current human reference genome.
171 In addition, the large number of assembly positions to which such short reads map which could
172 not be mapped to the reference genome GRCh38 (including widely used supplementary
173 sequences included in the GATK bundle), indicate a need for further assembly-based reference
174 data and for new approaches to better capture genetic diversity.
175
176 Genotype principal component analysis of the Egyptian cohort shows a homogeneous group
177 for which the assembly individual is representative (Suppl. Figs. 31-37).
178 We genetically characterized the Egyptian population with respect to 143 other populations of
179 the world using variant data of 5,429 individuals in total. For this, we combined five different
180 data sets: (1) a recently published whole genome sequencing (WGS)-based variant data set from

181 929 individuals of the Human Genome Diversity Project (HGDP), covering 51 populations [41];
182 (2) 2,504 individuals from 26 populations of the 1000 Genomes project for which phase 3
183 genotypes are available [5]; (3) WGS-based variant data from 108 Qatari individuals [42]; (4) SNP
184 array-based variant data of 478 individuals from five countries of the Arabian Peninsula [25]; (5)
185 1,305 individuals from 68 African, European, Western and Southern Asian populations that
186 were compiled from 8 different publications into a recent SNP array-based variant data set [43].
187 All individuals and their annotations are provided in Suppl. Table 11, data sources are described
188 in Suppl. Table 12. A principal component analysis of the data shows a genetic continuum
189 between Europeans, Africans, East Asians and Americans along the first three principal
190 components, see Suppl. interactive HTML-based Fig. PCA_interactive.html. Egyptians are
191 located on the European-African axis and close to Europeans. Their genetic variance spreads
192 to a small degree in the direction of the Asian axis, akin to further individuals from the Middle
193 East (see Fig. 2c). To preclude a technical bias when intersecting WGS with SNP array data,
194 we compared the analysis results when using whole genome data, only, or when intersecting
195 WGS data with SNP arrays and found comparable results in both cases (Suppl. Fig. 38). The
196 Egyptian PCA location is further supported by an admixture analysis. Our analysis specifies
197 k=24 as the optimal number of genetic components for the entire data set, i.e. having the
198 smallest cross validation error (see Suppl. Fig. 39 for results for k=10 to k=25). Accordingly,
199 the genetics of Egyptian individuals comprises four distinct population components that sum
200 up to 75% on average. Egyptians have a Middle Eastern, a European / Eurasian, a North African
201 and an East African component with 27%, 24%, 15% and 9% relative influence, respectively
202 (see Fig. 2a). According to our cohort, Egyptians show genetically little heterogeneity, with
203 little variance in the proportion of individual components between the individuals (Suppl. Figs
204 40 and 41). With a focus on populations from the Horn of Africa, the four components we
205 identified have been described before by Hodgeson *et al.* [44] in a cohort of 2,194 individuals
206 from 81 populations (mainly 1000 Genomes and HGDP) and substantially fewer variants

207 (n=16,766), but including also 31 Egyptians. They and others hypothesize that most non-
208 African ancestry, i.e. the Eurasian / European and Middle Eastern components in the
209 populations from North Africa and the Horn of Africa is resulting from prehistoric back-to-
210 Africa migration [44] [24]. Recently, Serra-Vidal *et al*. describe North Africa as a "melting pot of
211 genetic components", attributing most genetic variation in the region also to prehistoric times
212 [45]. Here, we confirm previously identified genetic components, yet using 2.5 times as many
213 individuals, and using WGS data for the majority of them. This is thus the hitherto most
214 comprehensive data set on genetic diversity world-wide and in this region.
215 The European, African and Asian ancestry components of Egyptians are further supported by
216 mitochondrial haplogroup assessment from mtDNA sequencing of 227 individuals in additiona
217 to 100 available from the literature [27]. mtDNA sequencing revealed that Egyptians have
218 haplogroups most frequently found in Europeans (e.g., H, V, T, J, etc.; >60%), African (e.g.,
219 L with 24.8%) or Asian/East Asian haplogroups (e.g., M with 6.7%). Overall, this supports the
220 admixture and PCA analysis and the notion that Egypt's transcontinental geographical location
221 shaped Egyptian genetics (Suppl. Fig. 42).
222 Lastly, we characterized the Egyptian population with respect to runs of homozygosity. The
223 distribution of overall length of ROHs larger than 5 Mb is comparable for the Egyptian
224 population and Middle Eastern populations and, to lesser extent, also for other North African
225 and Western Asian populations. In comparison, Europeans and Sub-Saharan Africans have
226 usually shorter ROHs, see Fig. 2b. Abundance of long ROHs is typical for the Greater Middle
227 East [26] and reflects the common practice of consanguineous marriages in this region.
228
229 In total, we identified 6,599,037 common Egyptian SNVs (minor allele frequency (MAF) >
230 5%, genotypes in a minimum of 100 individuals), of which 1,198 are population-specific; i.e.,
231 they are either rare (MAF < 1%) or not detected in any other population in the 1000 Genomes
232 [5], gnomAD database [6] or TOPMed [7] (Suppl. Table 13). These numbers are comparable to

233 population-specific variant numbers reported previously for 1000 Genomes populations [46].
234 Four SNVs likely have a molecular impact (Suppl. Table 14), indicated by a CADD [47]
235 deleteriousness score greater than 20. SNP rs143563851 (CADD 24.2) has recently been
236 identified in 1% of individuals of a cohort of 211 Palestinians in a study that performed targeted
237 sequencing of blood group antigen synthase GBGT1 [48]. SNP rs143614333 (missense variant in
238 gene CR2, CADD 23.6) is in ClinVar [49], with three submitters reporting that the variant is of
239 uncertain clinical significance. Additionally, we obtained 49 variants with no dbSNP [50] rsID
240 (Suppl. Table 15). These numbers of population-specific SNPs, of which some are likely to
241 have an immediate impact on clinical characteristics and diagnostics, indicate insufficient
242 coverage of the genetic diversity of the world's population for precision medicine and thus the
243 need for local genome references. To detect a putative genetic contribution of Egyptian
244 population-specific SNPs towards molecular pathways, phenotypes or disease, we performed
245 gene set enrichment analysis for all 461 protein-coding genes that were annotated to population-
246 specific SNPs by Ensembl VEP [51]. Enrichr, a gene list enrichment tool incorporating 153 gene
247 sets and pathway databases [52], reports that genes from obesity-related traits of the GWAS
248 catalog 2019 collection are over-represented (adj. p-value: 1.02E-6; 49 of 804 genes), which
249 might hint at population-specific metabolism regulation that is linked to body weight.
250
251 Variants that are not protein coding may have a regulatory effect that affects gene and
252 eventually protein expression. Using blood expression data obtained from RNA sequencing for
253 the EGYPT assembly individual in conjunction with 10x sequencing-based phased variant data,
254 we identified genes with haplotype-dependent expression patterns (see Suppl. Fig. 43 for the
255 analysis overview and Suppl. Figs. 44-45 for the results). We report 1,180 such genes (Suppl.
256 Table 16). Of these, variants contained in haplotypes of 683 genes (58%) have previously
257 reported expression quantitative trait loci (eQTLs) in blood according to Qtlizer [53], for 380
258 genes supported by multiple studies. For 370 genes (31%), the strongest associated blood eQTL

259 SNV is haplotypically expressed, and for 131 genes, the best eQTL has been previously
260 reported by multiple studies. Concordance of haplotypic expression with eQTLs indicates that
261 a common variant may affect gene expression; discordance hints towards a rare variant.
262
263 We investigated the impact of Egyptian ancestry on disease risk by integrating Egyptian variant
264 data with the GWAS catalog [54], a curated database of GWAS. According to the GWAS catalog,
265 most published GWAS are performed on Europeans [29], and only a single study has been
266 performed on Egyptians [55] (by one of the co-authors). Furthermore, only 2% of individuals
267 included in GWAS are of African ancestry [29]. AFs, LD and genetic architecture can differ
268 between populations, such that results from European GWAS cannot necessarily be transferred
269 [30]. This lack of transferability also compromises the prediction of an individual's traits and
270 disease risk using polygenic scores: such scores are estimated to be approximately one-third as
271 informative in African individuals compared to Europeans [56]. From the GWAS catalog, we
272 constructed a set of 4,008 different, replicated, high-quality tag SNPs (i.e., one strongest
273 associated SNP per locus) from European ancestry GWAS for 584 traits and diseases. We
274 compared the tag SNPs' AFs and proxy SNPs in the Egyptian cohort (n=110) and Europeans
275 from 1000 Genomes (n=503) (Suppl. Table 17). Egyptian AFs of tag SNPs are comparable to
276 European AFs, with a tendency to be lower (Fig. 3a). There are variants common in Europeans
277 (AF>5%) but rare in Egyptians (AF<5%) (Suppl. Fig. 46). A total of 261 tag SNPs are not
278 present in the Egyptian cohort (~7%), clearly indicating a need to perform GWAS in non-
279 European populations to further elucidate disease risk conferred by these loci. We investigated
280 differences in LD structure using an approach that is used for fine-mapping of GWAS data,
281 which identifies proxy variants (illustrated in Fig. 3c). Proxy variants are variants correlated
282 with the tag GWAS SNP, i.e., in high LD (here, R^2>0.8). The post-GWAS challenge is the
283 identification of a causal variant from a set of variants in LD (tag SNP and proxy variants). We
284 found that the number of proxy variants was much lower in the Egyptian cohort (Fig. 3b), likely

285 due to shorter haplotype blocks known from African populations. This indicates that LD
286 differences between Egyptians and Europeans may compromise GWAS transferability and
287 European ancestry-based polygenic scores. However, Egyptian proxy variants are usually
288 included in the larger set of European proxy variants (Fig. 3d). An example is variant rs2075650
289 (a locus sometimes attributed to gene TOMM40), which has been linked to Alzheimer's disease
290 in seven GWASs (cf. Suppl. Fig. 47). This tag SNP has seven proxy variants in Europeans but
291 only two proxy variants in Egyptians. One European proxy, rs72352238, has also been reported
292 as a GWAS tag SNP, but it is not a proxy of rs2075650 in Egyptians and may thus fail
293 replication and transfer of GWAS results from the European to the Egyptian population.
294
295 With our Egyptian genome reference, it will be possible to perform comprehensive integrated
296 genome and transcriptome comparisons for Egyptian individuals in the future. This will shed
297 light on personal as well as population-wide common genetic variants. As an example for
298 personalized medicine for Egyptian specific genetics we visualize the complete genetic
299 information of the DNA repair-associated gene BRCA2 from our study in the integrative
300 genomics viewer [57] (IGV) and the variant phasing information within the 10x Genomics
301 browser `LOUPE` in Fig. 4 and Suppl. Fig 48, respectively. BRCA2 is linked to the progression
302 and treatment of breast cancer and other cancer types [58], if mutated. The IGV depicts the sample
303 coverage based on sequencing data from PacBio, 10x Genomics and Illumina (whole genome
304 as well as RNA) for the personal EGYPT genome together with common Egyptian SNPs.
305 Variants previously assessed in a breast cancer GWAS [58] are displayed as Manhattan plot; note
306 the three significant GWAS SNPs between positions 32,390 and 32,400 kb. The bottom
307 compares the identified SNVs and indels from the Korean and Yoruba *de novo* assembly with
308 our *de novo* EGYPT assembly. Visual inspection of both small and structural variations at the
309 personal and population-based genome levels already yields significantly different variants,
310 which might be important for genetic counselling and detection of inherited risks for cancer.

311

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

318

319 **Methods**

320

321 **Sample acquisition**

322 Samples were acquired from 10 Egyptian individuals. For nine individuals, high-coverage
323 Illumina short-read data were generated. For the assembly individual, high-coverage short-read
324 data were generated as well as high-coverage PacBio data and 10x data. Furthermore, we used
325 public Illumina short-read data from 100 Egyptian individuals from Pagani *et al.* [27]. See
326 Supplementary Tables 1 and 7 for an overview of the individuals and the corresponding raw
327 and result data generated in this study.

328

329 **PacBio data generation**

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

334

335 **Illumina short-read data generation**

336 For 350 bp library construction, the genomic DNA was sheared, and fragments with sizes of
337 approximately 350 bp were purified from agarose gels. The fragments were ligated to adaptors
338 and amplified using PCR. The generated libraries were then sequenced on the Illumina HiSeq
339 X Ten using PE150 and generated 312.8 GB of data.
340 For the assembly individual, sequencing data from five libraries was generated at overall 90x
341 genome coverage. For nine additional individuals, one library each was generated, amounting
342 to an overall 305x coverage of sequencing data. For the 100 individuals of Pagani *et al.* [27], three
343 were sequenced at high coverage (30x) and 97 at low coverage (8x). The average coverage over
344 SNV positions for all 110 samples is provided in Supplementary Table 7.
345
346 **RNA sequencing data generation**
347 For RNA sequencing, ribosomal RNA was removed from total RNA, double-stranded cDNA
348 was synthesized, and then adaptors were ligated. The second strand of cDNA was then
349 degraded to generate a directional library. The generated libraries with insert sizes of 250-300
350 bp were selected and amplified and then sequenced on the Illumina HiSeq using PE150.
351 Overall, 64,875,631 150 bp paired-end sequencing reads were generated.
352
353 **10x sequencing data generation**
354 For 10x genomic sequencing, the Chromium Controller was used for DNA indexing and
355 barcoding according to the manufacturer's instructions (10x Genomics,
356 www.10xgenomics.com). The generated fragments were sheared, and then adaptors were
357 ligated. The generated libraries were sequenced on the Illumina HiSeq X Ten using PE150 and
358 generated 272.7 GB of data. Sequencing data from four 10x libraries was generated at overall
359 80x genome coverage.
360
361 **Construction of draft *de novo* assemblies and meta-assembly**

362 We used WTDBG2 [32] for human *de novo* assembly followed by its accompanying polishing tool
363 WTPOA-CNS with PacBio reads and in a subsequent polishing run with Illumina short reads.
364 This assembly was further polished using PILON [59] with short-read data (cf. Suppl. Methods:
365 *WTDBG2-based assembly*).
366 An alternative assembly was generated by using FALCON [60], QUIVER [61], SSPACE-
367 LONGREAD [62], PBJELLY [63], FRAGSCAFF [64] and PILON [59] (cf. Suppl. Methods: *FALCON-*
368 *based assembly*).
369 Proceeding from the WTDBG2-based assembly, we constructed a meta-assembly. Regions larger
370 than 800 kb that were not covered by this base assembly and were not located within centromere
371 regions were extracted from the alternative FALCON-based assembly (Suppl. Table 3). See
372 Suppl. Fig. 1 for an overview of our assembly strategy, including meta-assembly construction
373 (cf. Suppl. Methods: *Meta-assembly construction*).
374 Assembly quality and characteristics were assessed with QUAST-LG [33]. Additionally, we
375 removed misassemblies in centromeres or in segmental duplication regions from the QUAST-
376 LG report and furthermore removed structural variants from misassemblies (cf. Suppl. Methods:
377 *Assembly comparison and QC*). The extraction of coordinates for meta-assembly construction
378 was performed using QUAST-LG output. K-mer multiplicity was assessed with KAT [34].
379 Following Porubsky *et al*. [65], we computed QV as the number of homozygous variants divided
380 by the effective genome size. Towards this, we mapped all short reads to the assembly using
381 BWA MEM and perform variant calling using FREEBAYES with default parameters. We kept
382 only homozygous variants with a minimum quality of 10 using VCFTOOLS. Single-nucleotide
383 differences were counted as difference of 1 bp, indel differences as the length differences
384 between reference and alternative allele. Based on SAMTOOLS command "stats", we computed
385 the sum of bases with short read coverage as effective genome size.
386
387 **Repeatmasking**

388 Repeatmasking was performed by using REPEATMASKER [66] with RepBase version 3.0
389 (Repeatmasker Edition 20181026) and Dfam_consensus (http://www.dfam-consensus.org) (cf.
390 Suppl. Methods: *Repeat annotation*).
391
392 **Unique inserted sequences**
393 We trimmed Illumina short sequencing reads of 110 Egyptian individuals using FASTP 0.20.0
394 with default parameters, mapped the output reads to GRCh38 and GATK bundle sequences
395 using BWA 0.7.15-r1140 and sorted by chromosomal position using SAMTOOLS 1.3.1.
396 Subsequently, we extracted reads that did not map to GRCh38 using SAMTOOLS with
397 parameter F13 (i.e. read paired, read unmapped, mate unmapped) and repeated the mapping
398 and sorting using the Egyptian *de novo* assembly. We merged the read-group specific BAM
399 files for each sample and calculated the per base read depth using SAMTOOLS. Afterwards, we
400 aggregated the results via custom scripts and extracted uniquely inserted sequences from the
401 Egyptian *de novo* assembly. Insertions were defined as contiguous regions of at least 500 bp
402 having a coverage of more than 5 reads per base in 10 or more samples. Lastly, we BLASTed
403 the obtained sequences against the standard databases (option nt) for highly similar sequences
404 (option megablast) using a custom script. For the uniquely inserted sequences that we identified,
405 we created a pileup over all BAM files containing the reads that did not map to GRCh38 using
406 SAMTOOLS. Based on these pileups, we then called the variants using BCFTOOLS. Variants
407 with quality of more than 10 were kept.
408
409 **Phasing**
410 Phasing was performed for the assembly individual's SNVs and short indels obtained from
411 combined genotyping with the other Egyptian individuals, i.e., based on short-read data. These
412 variants were phased using 10x data and the 10x Genomics LONGRANGER WGS pipeline with

413 four 10x libraries provided for one combined phasing. See Supplementary Methods *Variant*
414 *phasing* for details.
415
416 **SNVs and small indels**
417 Calling of SNVs and small indels was performed with `GATK` 3.8 [67] using the parameters of the
418 best practice workflow. Reads in each read group were trimmed using `Trimmomatic` [68] and
419 subsequently mapped against reference genome hg38 using `BWA-MEM` [69] version 0.7.17. Then,
420 the alignments for all read groups were merged sample-wise and marked for duplicates. After
421 the base recalibration, we performed variant calling using `HaplotypeCaller` to obtain
422 GVCF files. These files were input into GenotypeGVCFs to perform joint genotyping. Finally,
423 the variants in the outputted VCF file were recalibrated, and only those variants that were
424 flagged as "PASS" were kept for further analyses. We used `FastQC` [70], `Picard Tools` [71]
425 and `verifyBamId` [72] for QC (cf. Suppl. Methods: *Small variant QC*).
426
427 **Variant annotation**
428 Variant annotation was performed using `ANNOVAR` [73] and `VEP` [51] (cf. Suppl. Methods: *Small*
429 *variant annotation*)
430
431 **Structural variants**
432 SVs were called using `DELLY2` [74] with default parameters as described on the `DELLY2` website
433 for germline SV calling (https://github.com/dellytools/delly) (cf. Suppl. Methods: *Structural*
434 *variant QC*). Overlapping SV calls in the same individual were collapsed by the use of custom
435 scripts. See Supplementary Methods *Collapsing structural variants* for details.
436
437 **Population genetics**

438 For population genetic analyses, we compared the Egyptian variant data with variant data from
439 five additional sources specified in Suppl. Table 12. Individuals together with their annotations
440 are listed in Suppl. Table 11. Variant data was merged to contain only variants present in all
441 data sets and subsequently filtered and LD pruned. Genotype principal component analysis was
442 computed using `SMARTPCA` [75] from the `EIGENSOFT` package. Admixture was computed with
443 `ADMIXTURE` [76] (cf. Suppl. Methods: *Population genetics* and *SNP array-based Egyptian*
444 *variant data*). Runs of homozygosity were computed on the same files that were used for PC
445 computation and admixture using `PLINK –homozyg`. ROHs with size larger than 5 Mb were
446 summed to obtain overall length of ROHs per individual.
447
448 **Mitochondrial haplogroups**
449 Haplogroup assignment was performed for 227 individuals using `HAPLOGREP 2` [77].
450 Furthermore, mitochondrial haplogroups were obtained from Pagani *et al.* [27] for 100
451 individuals. See Supplementary Methods *Mitochondrial haplogroups* for details.
452
453 **Population-specific variants**
454 Our set of common Egyptian SNVs comprises variants with genotypes in a minimum of 100
455 individuals whose alternative allele has a frequency of more than 5%. Those common Egyptian
456 SNVs that are otherwise rare, i.e., have an AF of less than 1% in the 1000 Genomes, and
457 gnomAD populations as well as in TOPMed were considered Egyptian-specific. AFs were
458 annotated using the Ensembl API. Furthermore, a list of Egyptian common variants without
459 dbSNP rsID was compiled, see Supplementary Methods *Small variant annotation* for details.
460
461 **Haplotypic expression analysis**
462 RNA sequencing reads were mapped and quantified using `STAR` (Version 2.6.1.c) [78].
463 Haplotypic expression analysis was performed by using `PHASER` and `PHASER GENE AE`

464 (version 1.1.1) [79] with Ensembl version 95 annotation on the 10x-phased haplotypes using
465 default parameters. See Supplementary Methods *Haplotypic expression* for details.
466
467 **GWAS catalog data integration**
468 GWAS catalog associations for GWAS of European ancestry were split into trait-specific data
469 sets using Experimental Factor Ontology (EFO) terms. For every trait, a locus was defined as
470 an associated variant +/- 1 MB, and only loci that were replicated were retained. For proxy
471 computation, we used our Egyptian cohort (n=110) and the European individuals of 1000
472 Genomes (n=503). For details, see Supplementary Methods *Data integration with the GWAS*
473 *catalog.*
474
475 **Integrative genomics view**
476 We implemented a workflow to extract all Egyptian genome reference data for view in the IGV
477 [57]. This includes all sequencing data mapped to GRCh38 (cf. Suppl. Methods *Sequencing read*
478 *mapping to GRCh38*) as well as all assembly differences (cf. Suppl. Methods *Alignment to*
479 *GRCh38* and *Assembly-based variant identification*) and all Egyptian variant data. See
480 Supplementary Methods *Gene-centric integrative data views* for details.
481
482 **Ethics statement**
483 This study was approved by the Mansoura Faculty of Medicine Institutional Review Board
484 (MFM-IRB) Approval Number RP/15.06.62. All subjects gave written informed consent in
485 accordance with the Declaration of Helsinki. This study and its results are in accordance with
486 the Jena Declaration (https://www.uni-
487 jena.de/unijenamedia/Universit%C3%A4t/Abteilung+Hochschulkommunikation/Presse/Jenae
488 r+Erkl%C3%A4rung/Jenaer_Erklaerung_EN.pdf).
489

490 **References**

491 1. Cao, H. *et al.* De novo assembly of a haplotype-resolved human genome. *Nat. Biotechnol.*
492 **33**, 617–622 (2015).
493 2. Seo, J.-S. *et al. De novo* assembly and phasing of a Korean human genome. *Nature* **538**,
494 243–247 (2016).
495 3. Shi, L. *et al.* Long-read sequencing and de novo assembly of a Chinese genome. *Nat*
496 *Commun* **7**, 12065 (2016).
497 4. Cho, Y. S. *et al.* An ethnically relevant consensus Korean reference genome is a step
498 towards personal reference genomes. *Nat Commun* **7**, 13637 (2016).
499 5. The 1000 Genomes Project Consortium. A global reference for human genetic variation.
500 *Nature* **526**, 68–74 (2015).
501 6. Karczewski, K. J. *et al.* Variation across 141,456 human exomes and genomes reveals the
502 spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv*
503 531210 (2019) doi:10.1101/531210.
504 7. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed
505 Program. *bioRxiv* 563866 (2019) doi:10.1101/563866.
506 8. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies
507 demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864
508 (2017).
509 9. Levy-Sakin, M. *et al.* Genome maps across 26 human populations reveal population-
510 specific patterns of structural variation. *Nature Communications* **10**, 1025 (2019).
511 10. Stark, Z. *et al.* Integrating Genomics into Healthcare: A Global Responsibility. *Am. J.*
512 *Hum. Genet.* **104**, 13–20 (2019).
513 11. GenomeAsia 100k. *GenomeAsia 100k* http://www.genomeasia100k.com/.
514 12. Turnbull, C. *et al.* The 100 000 Genomes Project: bringing whole genome sequencing to
515 the NHS. *BMJ* **361**, k1687 (2018).

516 13. Maretty, L. *et al.* Sequencing and de novo assembly of 150 genomes from Denmark as a
517 population reference. *Nature* **548**, 87–91 (2017).
518 14. Leslie, S. *et al.* The fine-scale genetic structure of the British population. *Nature* **519**,
519 309–314 (2015).
520 15. Chiang, C. W. K., Mangul, S., Robles, C. & Sankararaman, S. A Comprehensive Map of
521 Genetic Variation in the World's Largest Ethnic Group-Han Chinese. *Mol. Biol. Evol.* **35**,
522 2736–2750 (2018).
523 16. Bai, H. *et al.* Whole-genome sequencing of 175 Mongolians uncovers population-specific
524 genetic architecture and gene flow throughout North and East Asia. *Nat. Genet.* (2018)
525 doi:10.1038/s41588-018-0250-5.
526 17. Sherman, R. M. *et al.* Assembly of a pan-genome from deep sequencing of 910 humans of
527 African descent. *Nat. Genet.* (2018) doi:10.1038/s41588-018-0273-y.
528 18. Choudhury, A. *et al.* Whole-genome sequencing for an enhanced understanding of genetic
529 variation among South Africans. *Nat Commun* **8**, 2062 (2017).
530 19. Bustamante, C. D., Burchard, E. G. & De la Vega, F. M. Genomics for the world. *Nature*
531 **475**, 163–165 (2011).
532 20. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–164
533 (2016).
534 21. Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves discovery for
535 complex traits. *Nature* **570**, 514–518 (2019).
536 22. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health
537 disparities. *Nat. Genet.* **51**, 584–591 (2019).
538 23. Levy-Sakin, M. *et al.* Genome maps across 26 human populations reveal population-
539 specific patterns of structural variation. *Nat Commun* **10**, 1025 (2019).
540 24. Henn, B. M. *et al.* Genomic ancestry of North Africans supports back-to-Africa
541 migrations. *PLoS Genet.* **8**, e1002397 (2012).

542 25. Fernandes, V. *et al.* Genome-Wide Characterization of Arabian Peninsula Populations:
543 Shedding Light on the History of a Fundamental Bridge between Continents. *Mol. Biol.*
544 *Evol.* **36**, 575–586 (2019).
545 26. Scott, E. M. *et al.* Characterization of Greater Middle Eastern genetic variation for
546 enhanced disease gene discovery. *Nat. Genet.* **48**, 1071–1076 (2016).
547 27. Pagani, L. *et al.* Tracing the Route of Modern Humans out of Africa by Using 225 Human
548 Genome Sequences from Ethiopians and Egyptians. *Am J Hum Genet* **96**, 986–991
549 (2015).
550 28. ElHefnawi, M. *et al.* Whole genome sequencing and bioinformatics analysis of two
551 Egyptian genomes. *Gene* **668**, 129–134 (2018).
552 29. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic
553 Studies. *Cell* **177**, 1080 (2019).
554 30. Kim, M. S., Patel, K. P., Teng, A. K., Berens, A. J. & Lachance, J. Genetic disease risks
555 can be misestimated across global populations. *Genome Biol.* **19**, 179 (2018).
556 31. Chin, C.-S. *et al.* Phased diploid genome assembly with single-molecule real-time
557 sequencing. *Nature Methods* **13**, 1050–1054 (2016).
558 32. Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *bioRxiv* 530972
559 (2019) doi:10.1101/530972.
560 33. Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D. & Gurevich, A. Versatile
561 genome assembly evaluation with QUAST-LG. *Bioinformatics* **34**, i142–i150 (2018).
562 34. Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J. & Clavijo, B. J. KAT: a
563 K-mer analysis toolkit to quality control NGS datasets and genome assemblies.
564 *Bioinformatics* **33**, 574–576 (2017).
565 35. Sund, K. L. & Rehder, C. W. Detection and reporting of homozygosity associated with
566 consanguinity in the clinical laboratory. *Hum. Hered.* **77**, 217–224 (2014).

567 36. Wong, K. H. Y., Levy-Sakin, M. & Kwok, P.-Y. De novo human genome assemblies
568 reveal spectrum of alternative haplotypes in diverse populations. *Nat Commun* **9**, 3040
569 (2018).
570 37. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat. Genet.* **37**, 727–
571 732 (2005).
572 38. Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human
573 genomes. *Nature* **453**, 56–64 (2008).
574 39. Fan, X., Chaisson, M., Nakhleh, L. & Chen, K. HySA: a Hybrid Structural variant
575 Assembly approach using next-generation and single-molecule sequencing technologies.
576 *Genome Res.* **27**, 793–800 (2017).
577 40. Kehr, B. *et al.* Diversity in non-repetitive human sequences not found in the reference
578 genome. *Nat. Genet.* **49**, 588–593 (2017).
579 41. Bergström, A. *et al.* Insights into human genetic variation and population history from
580 929 diverse genomes. *Science* **367**, (2020).
581 42. Rodriguez-Flores, J. L. *et al.* Indigenous Arabs are descendants of the earliest split from
582 ancient Eurasian populations. *Genome Res.* **26**, 151–162 (2016).
583 43. Busby, G. Genotype data for a set of 163 worldwide populations. **3**, (2020).
584 44. Hodgson, J. A., Mulligan, C. J., Al-Meeri, A. & Raaum, R. L. Early back-to-Africa
585 migration into the Horn of Africa. *PLoS Genet.* **10**, e1004393 (2014).
586 45. Serra-Vidal, G. *et al.* Heterogeneity in Palaeolithic Population Continuity and Neolithic
587 Expansion in North Africa. *Curr. Biol.* **29**, 3953-3959.e4 (2019).
588 46. Choudhury, A. *et al.* Population-specific common SNPs reflect demographic histories and
589 highlight regions of genomic plasticity with functional relevance. *BMC Genomics* **15**, 437
590 (2014).

591 47. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting
592 the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**,
593 D886–D894 (2019).
594 48. Abusibaa, W. A. *et al.* Expression of the GBGT1 Gene and the Forssman Antigen in Red
595 Blood Cells in a Palestinian Population. *Transfusion Medicine and Hemotherapy* (2019)
596 doi:10.1159/000497288.
597 49. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting
598 evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
599 50. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**,
600 308–311 (2001).
601 51. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biology* **17**, 122
602 (2016).
603 52. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server
604 2016 update. *Nucleic Acids Res.* **44**, W90-97 (2016).
605 53. Munz, M. *et al.* Qtlizer: comprehensive QTL annotation of GWAS results. *bioRxiv*
606 495903 (2019) doi:10.1101/495903.
607 54. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide
608 association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**,
609 D1005–D1012 (2019).
610 55. Bejaoui, Y. *et al.* Genome-wide association study of psoriasis in an Egyptian population.
611 *Exp. Dermatol.* **28**, 623–627 (2019).
612 56. Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse
613 human populations. *Nat Commun* **10**, 3328 (2019).
614 57. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
615 58. Michailidou, K. *et al.* Association analysis identifies 65 new breast cancer risk loci.
616 *Nature* **551**, 92–94 (2017).

617 59. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection
618 and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
619 60. Chin, C.-S. *et al.* Phased diploid genome assembly with single-molecule real-time
620 sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
621 61. Chin, C.-S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read
622 SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
623 62. Boetzer, M. & Pirovano, W. SSPACE-LongRead: scaffolding bacterial draft genomes
624 using long read sequence information. *BMC Bioinformatics* **15**, (2014).
625 63. English, A. C. *et al.* Mind the Gap: Upgrading Genomes with Pacific Biosciences RS
626 Long-Read Sequencing Technology. *PLoS ONE* **7**, e47768 (2012).
627 64. Adey, A. *et al.* In vitro, long-range sequence information for de novo genome assembly
628 via transposase contiguity. *Genome Research* **24**, 2041–2049 (2014).
629 65. Porubsky, D. *et al.* A fully phased accurate assembly of an individual human genome.
630 *bioRxiv* 855049 (2019) doi:10.1101/855049.
631 66. SMIT, A. F. A. Repeat-Masker Open-3.0. *http://www.repeatmasker.org* (2004).
632 67. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for
633 analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297–1303
634 (2010).
635 68. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina
636 sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
637 69. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler
638 transform. *Bioinformatics* **26**, 589–595 (2010).
639 70. Andrews, S. FASTQC - A quality control tool for high throughput sequence data.
640 http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
641 71. Picard Toolkit. http://broadinstitute.github.io/picard/.

642 72. Jun, G. *et al.* Detecting and Estimating Contamination of Human DNA Samples in
643 Sequencing and Array-Based Genotype Data. *The American Journal of Human Genetics*
644 **91**, 839–848 (2012).
645 73. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic
646 variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164–e164 (2010).
647 74. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-
648 read analysis. *Bioinformatics* **28**, i333–i339 (2012).
649 75. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS*
650 *Genet.* **2**, e190 (2006).
651 76. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in
652 unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
653 77. Kloss-Brandstätter, A. *et al.* HaploGrep: a fast and reliable algorithm for automatic
654 classification of mitochondrial DNA haplogroups. *Hum. Mutat.* **32**, 25–32 (2011).
655 78. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* bts635 (2012)
656 doi:10.1093/bioinformatics/bts635.
657 79. Castel, S. E., Mohammadi, P., Chung, W. K., Shen, Y. & Lappalainen, T. Rare variant
658 phasing and haplotypic expression from RNA sequencing with phASER. *Nat Commun* **7**,
659 12817 (2016).
660

661 **Supplementary information**

662 *Supplementary Tables 1-17:* An_Egyptian_genome_reference_supplementary_tables.xlsx
663 *Supplementary Methods and Supplementary Figures 1-48:*
664 An_Egyptian_genome_reference_supplement.pdf
665

666 **Acknowledgements**

667 We acknowledge support on coordination of the project and assembly work w.r.t the FALCON-
668 based assembly through Ms. Lu Wang from the Novogene (UK) Company Limited. IW, HB
669 and SI acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German
670 Research Foundation) under Germany`s Excellence Strategy – EXC 22167-390884018.
671 Verónica Calonga-Solís was supported by a DAAD scholarship. All authors acknowledge
672 computational support from the OMICS compute cluster at the University of Lübeck.
673

674 **Author information**

675

676 *Medical Systems Biology Division, Lübeck Institute of Experimental Dermatology and Institute*
677 *for Cardiogenetics, University of Lübeck, Lübeck, Germany*
678 Inken Wohlers, Axel Künstner, Matthias Munz, Michael Olbrich, Anke Fähnrich, Verónica
679 Calonga-Solís & Hauke Busch
680

681 *Department of Genetics, Federal University of Paraná (UFPR), Curitiba, Brazil*
682 Verónica Calonga-Solís
683

684 *Novogene (UK) Company Limited, Babraham Research Campus, Cambridge, United Kingdom*
685 Caixia Ma
686

687 *Medical Experimental Research Center (MERC), Mansoura University, Mansoura, Egypt*
688 Mohamed Salama & Shaaban El-Mosallamy
689

690 *Institute of Global Health and Human Ecology, The American University in Cairo, Cairo,*
691 *Egypt*
692 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, A.F. and V. C.-S. 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`, where also variant allele frequencies can be queried online. Raw sequencing data and variant data are available at EGA under study ID EGAS00001004303. De novo assemblies are available at NCBI under BioProject ID PRJNA613239.

Code availability

Computational tools used are specified in the Supplementary Methods. Workflows use Snakemake and Conda (especially Bioconda) for reproducible data analysis and are provided on request.

Corresponding authors

Correspondence to Hauke Busch or Saleh Ibrahim.

722

723 *Table 1: Default assembly quality measures according to* `QUAST-LG`*. The extended* `QUAST-LG` *report is*

724 *provided in Suppl. Table 2. Yoruba is a chromosome-level assembly. Best quality for every measure is denoted in*

725 *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

726

727

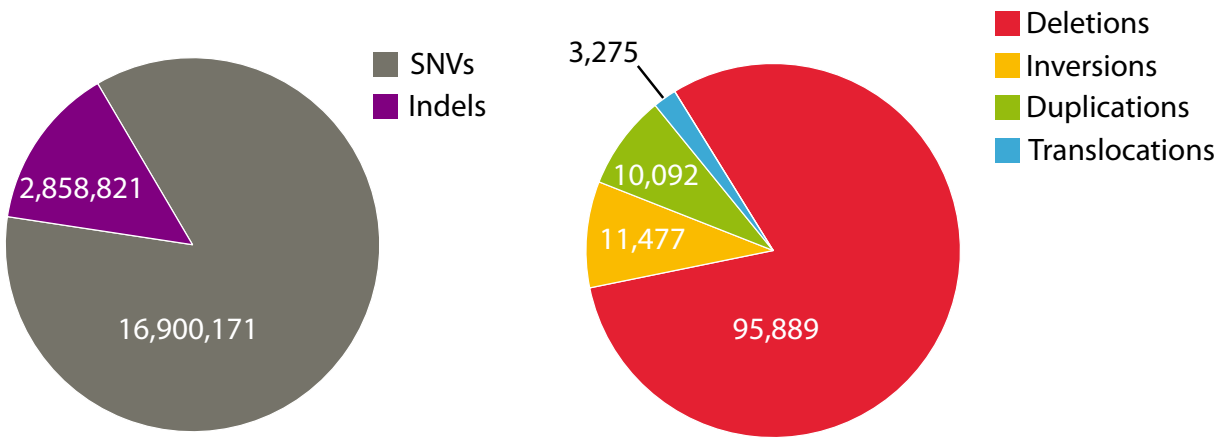
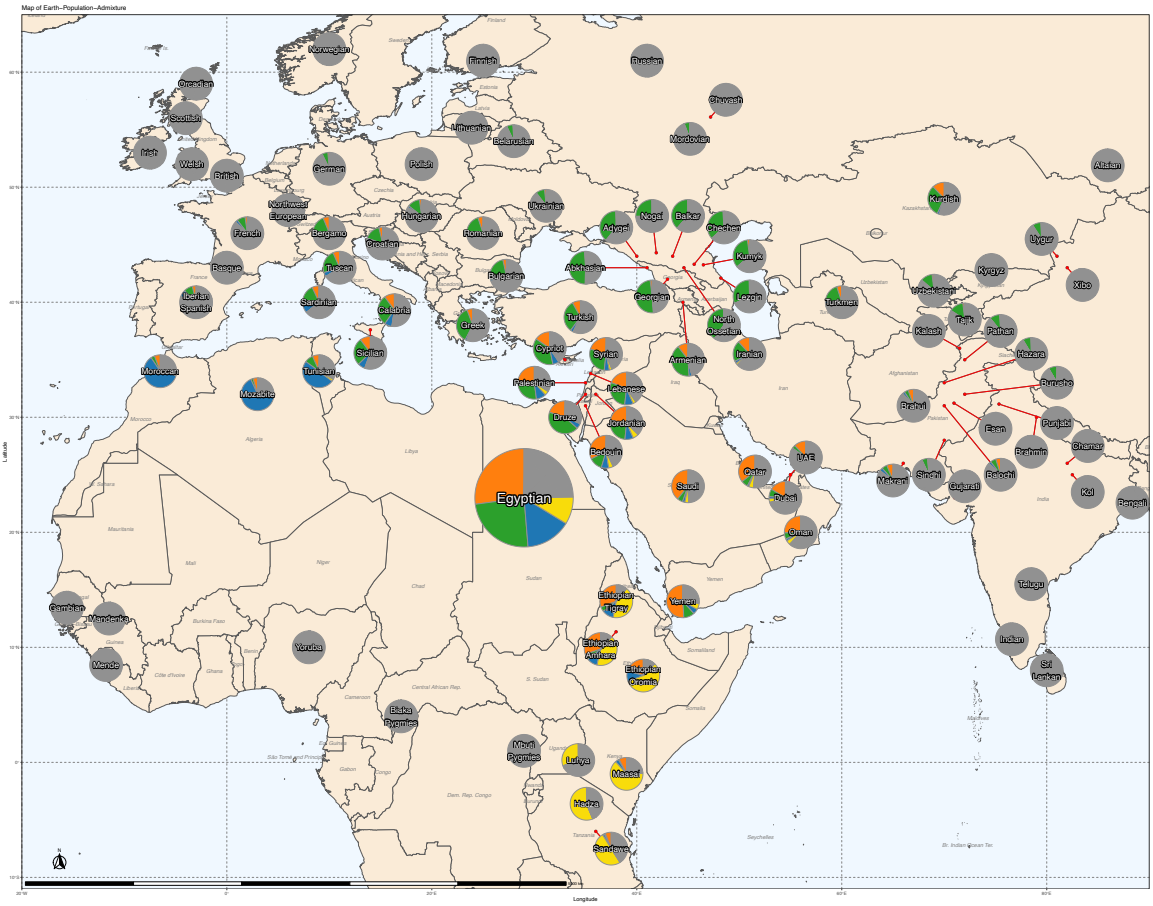
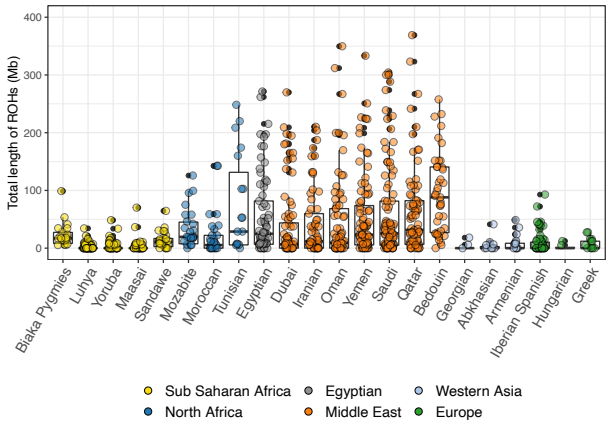


Figure 1: Number of various genetic variant types identified in the Egyptian cohort. Left: The number of SNVs and indels. Right: The number of SV calls: deletions, inversions, duplications and translocations. Additionally, 408 insertions have been called.

a)



b)



c)

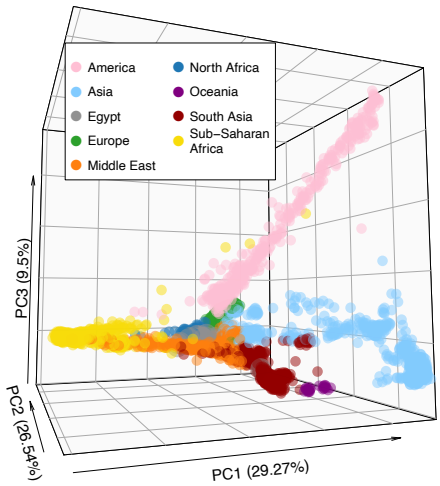
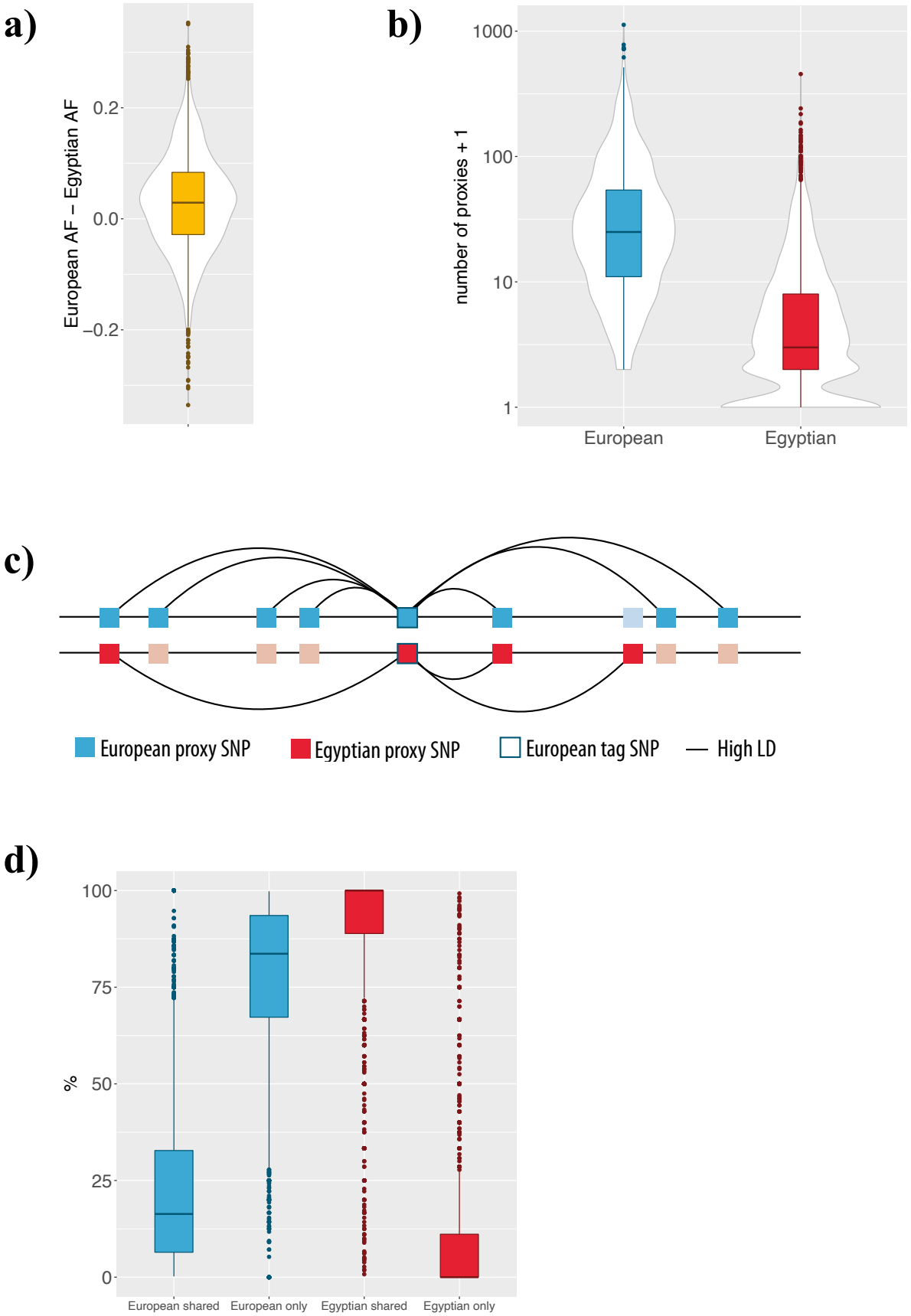


Figure 2: Population genetic characterization of the Egyptian population a) The four largest admixture components in the Egyptian population for African, European and Western Asian populations. b) Box plots for total length of runs of homozygosity for the Egyptians and several populations from relevant world regions (one Qatari not shown). c) Principal component analysis with individuals from populations world-wide.

739


a)
European AF – Egyptian AF
0.2
0.0
−0.2
b)
number of proxies + 1
1000
100
10
1
European
Egyptian
c)
European proxy SNP
Egyptian proxy SNP
European tag SNP
High LD
d)
%
100
75
50
25
0
European shared
European only
Egyptian shared
Egyptian only

740

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

750

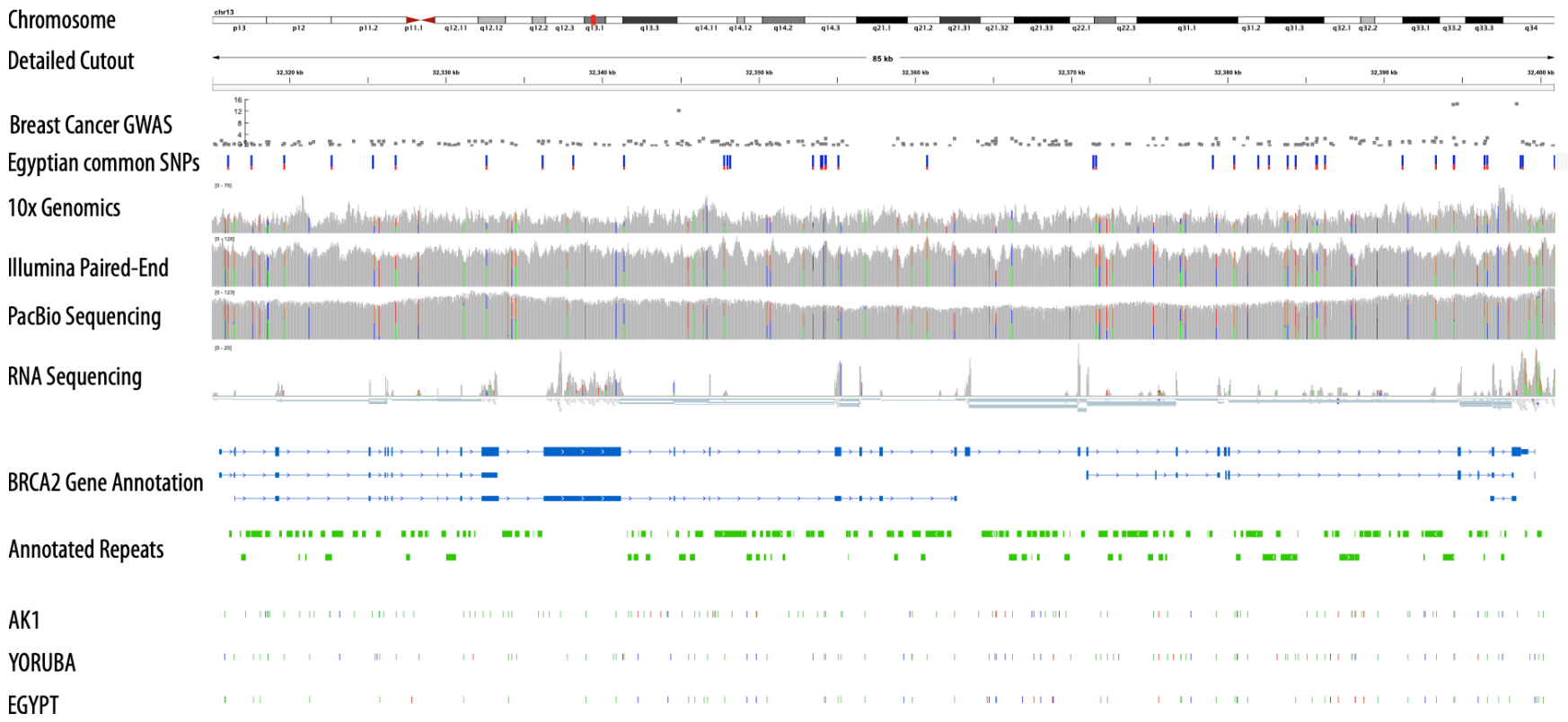


Figure 4: Integrative view of Egyptian genome reference data for the gene BRCA2, which is associated with breast cancer. The rows denote from top to bottom: Genome location on chromosome 13 of the magnified region for BRCA2 (first and second row); GWAS data for breast cancer risk [58]; Variants that 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). The corresponding variant phasing for the EGYPT individual is displayed in Suppl. Fig. 48.*