

# 1 A personal and population-based Egyptian genome reference

2  
3 Inken Wohlers, Axel Künstner, Matthias Munz, Michael Olbrich, Anke Fähnrich, Caixia Ma,  
4 Misa Hirose, Shaaban El-Mosallamy, Mohamed Salama, Hauke Busch & Saleh Ibrahim

## 7 Abstract

8  
9 The human genome is composed of 23 chromosomal DNA sequences of bases A, C, G and T  
10 -- the blueprint to implement the molecular functions that are at the basis of every individual's  
11 life. Deciphering the first human genome was a consortium effort that took more than a decade  
12 and cost about 3 billion dollars. With latest technological advances, determining an individual's  
13 entire personal genome at manageable cost and effort comes into reach. Although the benefit  
14 of all-encompassing genetic information that entire genomes provide is widely noted, so far  
15 only a small number of *de novo* assembled human genomes have been reported. Even less have  
16 been characterized and complemented with respect to population-specific variation. Here we  
17 combine long- and short-read whole genome next-generation sequencing data together with the  
18 recent assembly approaches for the first *de novo* assembly of the genome of an Egyptian  
19 individual, which we merged with Egyptian variant data into a population reference genome.  
20 The resulting genome assembly demonstrates overall well-balanced quality metrics and comes  
21 along with high quality variant phasing into maternal and paternal haplotypes. Further, we  
22 assayed population-specific variations genome-wide within a representative cohort of more  
23 than 100 Egyptian individuals. By annotation of these genetic data and integration with public  
24 databases we showcase genetic variants that alter protein sequence and that are linked to allelic  
25 gene expression. This is one of a handful of studies that comprehensively describe a population  
26 reference genome based on a high-quality personal genome and which highlights population-  
27 specific variants of interest. It is a proof-of-concept to be considered by the many national  
28 genome initiatives underway. And, more importantly, we anticipate that the Egyptian reference  
29 genome will be a valuable resource for precision medicine initiatives targeting the Egyptian  
30 population and beyond.

31  
32 *All summary data of the Egyptian genome reference is available at [www.egyptian-genome.org](http://www.egyptian-genome.org).*  
33 *The Egyptian genome reference will be publicly available upon journal publication.*

## 36 Main

37  
38 In the last years, several high-quality *de novo* human genome assemblies (1–3) and, more  
39 recently, pan-genomes (4) extended human sequence information and improved the *de facto*  
40 reference genome. At present, many national genome initiatives are established which aim to  
41 genetically characterize human populations (5).

42  
43 Population-specific genetic variation as part of an individual's personal genetic variation is  
44 indispensable for precision medicine (PM). Currently, genomics-based PM compares the  
45 patients' genetic make-up to a reference genome, a genome model inferred from people of  
46 mostly European descent, to detect risk mutations that are related to disease. However, genetic  
47 and epidemiologic studies have long recognized the importance of ancestral origin in conferring  
48 risk genes for disease. Risk alleles and structural variants (6) can be missing from the reference  
49 genome or can have different population frequencies such that alternative pathways become  
50 disease related in patients of different ancestral origin, which motivates to establish national  
51 genome projects. At present, there are several population-based sequencing efforts that aim at

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

60  
61 In this study we have generated a phased *de novo* assembly of an Egyptian individual and used  
62 it as a basis to identify single-nucleotide variants (SNVs) and structural variants (SVs) from an  
63 additional 109 Egyptian individuals obtained from short-read sequencing. Those were  
64 integrated to generate a consensus reference Egyptian genome. We anticipate that an Egyptian  
65 population reference genome will strengthen precision medicine efforts that may eventually  
66 benefit nearly 100 million Egyptians. Likewise, our genome will be of universal value for  
67 research purposes, since it contains both European and African variant features, and could thus  
68 be used to investigate the validity of genetic disease risk transfer across populations. As most  
69 genetic association studies are performed in Europeans (18), an Egyptian genome will be well  
70 suited to identify (i) genetic loci with shared or with distinct disease susceptibility across  
71 populations (ii) haplotypes that influence gene expression and (iii) variants that are likely  
72 protein-damaging and putatively related to disease.

73  
74 Our Egyptian genome is based on a high-quality human *de novo* assembly for one Egyptian  
75 individual (see workflow in Suppl. Figure 1). This assembly was generated from PacBio, 10x  
76 Genomics and Illumina paired-end sequencing data at overall 270x genome coverage (Suppl.  
77 Table 1). For this personal genome, we constructed two draft assemblies, one based on long-  
78 read assembly by an established assembler, FALCON (19), and another one based on the  
79 assembly by a novel assembler, WTDBG2 (20), that has a much lower runtime at comparable  
80 accuracy (cf. Suppl. Fig. 1). Both assemblies were polished using short-reads and various  
81 polishing tools. For the FALCON-based assembly, scaffolding was performed, whereas we  
82 found that the WTDBG2-based assembly was of comparable accuracy without scaffolding (cf.  
83 dotplots in Suppl. Figs. 2-3). We compared our two draft assemblies to the publicly available  
84 assemblies of a Korean (1) and a Yoruba individual (GeneBank assembly accession  
85 GCA\_001524155.4, unpublished) with respect to various quality control (QC) measures using  
86 QUAST-LG (21) (Table 1). The WTDBG2-based assembly was selected as base, because it  
87 performs comparable or better concerning various QC measures (Suppl. Table 2).

88  
89 Where larger gaps outside centromere regions occurred, we complemented this assembly with  
90 sequence from the FALCON-based assembly (Suppl. Table 3) to obtain a final Egyptian meta  
91 assembly, denoted as EGYPT (for overall assembly strategy, see Suppl. Figure 1). The  
92 comparative assembly statistics are summarized in Table 1. Suppl. Figure 2 compares the  
93 assemblies NA-values and Suppl. Figures 3-7 show dot plots of alignment with reference  
94 GRCh38. We performed repeat annotation and repeat masking for all assemblies (Suppl. Table  
95 4).

96  
97 The meta assembly was complemented with high-quality phasing information (Suppl. Table 5).  
98 Variants and small insertions and deletions (indels) called using short-read sequencing data  
99 were phased using high-converge linked-read sequencing data. This resulted in 98.99% of  
100 variants being phased. Further, nearly all (99.41%) of genes with length less than 100kb and  
101 more than one heterozygous SNP were phased into a single phase-block.

102

103 Based on the personal Egyptian genome, we constructed an Egyptian population genome by  
104 considering genome-wide SNV allele frequencies in 109 additional Egyptians (Suppl. Table 6).  
105 This enabled the characterization of the major allele (i.e. the allele with highest allele frequency)  
106 in the given Egyptian cohort. For this, we called variants using short-read data of 12 Egyptians  
107 sequenced at high coverage and 97 Egyptians sequenced at low coverage. Although sequence  
108 coverage affects variant-based statistics (Suppl. Fig. 8), due to combined genotyping most  
109 variants could also be called reliably in low coverage samples (Suppl. Fig. 9). Altogether, we  
110 called a total of 19,758,992 SNVs and small indels (Suppl. Fig. 10) in all 110 Egyptian  
111 individuals (Table 2). The number of called variants per individual varied between 2,901,883  
112 to 3,934,367 and was correlated with sequencing depth (see Suppl. Figs. 8-9). This relation was  
113 particularly pronounced for low coverage samples. The majority of variants was intergenic  
114 (53.5%) or intronic (37.2%) (Suppl. Fig. 11). Only about 0.7% of variants were located within  
115 coding exons, of which 54.4% were non-synonymous and thus have an impact on protein  
116 structure (Suppl. Fig. 12).

117  
118 Using short-read sequencing data of 110 Egyptians, we called 121,141 structural variants,  
119 which were mostly deletions (Suppl. Fig. 13), but also inversions, duplications, insertions and  
120 translocations of various orders of magnitudes (Table 2, Suppl. Fig. 14). Similar to SNVs, also  
121 SV calls vary between individuals (Suppl. Fig. 15) and are slightly affected by coverage (Suppl.  
122 Fig. 16). After merging overlapping SV calls we obtained on average 2,773 SVs per Egyptian  
123 individual (Suppl. Table 7, Suppl. Figs. 17-19).

124  
125 To characterize the Egyptian population with respect to European and African populations  
126 which have been genotyped within the 1000 Genomes Project (22) (Suppl. Table 8), we used  
127 SNVs and short indels for a genotype-based principal component analysis. According to this  
128 analysis, Egyptians are a genetically homogenous population compared to other populations,  
129 sharing genetic variants with both Europeans and Sub-Saharan Africans (see Fig. 1 and Suppl.  
130 Figs. 20-32). So far, there are no North-African populations with high-quality genome-wide  
131 genotype data available, and from the European and Sub-Saharan African populations reported  
132 by the 1000 Genomes Project, Egyptians are closest to the European Tuscany population (see  
133 Fig. 1 and Suppl. Figs. 20-32), which has been previously proposed through the genetic studies  
134 of ancient Egyptian mummies (23).

135  
136 The mixed European and African ancestry of Egyptians is further supported by mitochondrial  
137 haplogroup assessment from literature (17) and our own analyses. We found that Egyptians  
138 have haplogroups most frequent in Europeans (e.g. H,V,T,J etc.; more than 60%), but many  
139 also had African (e.g. L with 24.8%) or Asian/East Asian haplogroups (e.g. M with 6.7%),  
140 indicating that the Egyptian genome contains genetic variations from various major human  
141 population (Suppl. Fig. 33).

142  
143 In total we identified 2,270,642 common Egyptian SNVs (MAF > 5%) of which 26,564 are  
144 population-specific, i.e., they are rare (MAF < 1%) to non-existent in all other continental  
145 populations according to the 1000 Genomes data (Table 2). This is comparable to population-  
146 specific variant numbers reported previously for 1000 Genomes populations (24). Additionally,  
147 we found 4,807 African, 2 Ad Mixed American, 11 East Asian, 3 European and 77 South Asian  
148 SNVs that are population-specific in the Egyptian cohort and the respective continental  
149 population (Figure 1). These numbers clearly indicate an insufficient coverage of the genetic  
150 heterogeneity of the world's population for precision medicine and thus the need for local  
151 reference genomes.

152

153 To detect a putative genetic predisposition of Egyptian population-specific SNPs towards  
154 molecular pathways, phenotypes or disease, we selected all genes having a Combined  
155 Annotation-Dependent Depletion (CADD) phred score  $> 15$  (25). This resulted in 361  
156 associated genes out of which we discarded 159 non-protein coding or anti-sense genes. The  
157 resulting 202 genes were uploaded to Enrichr, a gene list enrichment tool incorporating 153  
158 gene set and pathway databases (26). Among the most enriched pathways we found 4 out of 23  
159 body fat percentage related genes from the GWAS catalogue 2019 (adj. p-value = 0.038; Genes  
160 CRT1; IGF2BP1; WDR41; SULT1A2) as well as Glycolysis in humans from the 2016  
161 Panther database (adj.p-val = 0.017, 3 out of 17 Genes: TPI1;BPGM;GAPDH), which was  
162 confirmed by the HumanCyc 2016 database. There, we found the terms glycolysis,  
163 gluconeogenesis and superpathway of conversion of glucose to acetyl CoA (pathway IDs PWY-  
164 6313, PWY66-400, PWY66-407) significant (adj. p-value=0.013; Genes TPI1; SUCLG2;  
165 BPGM; GAPDH). Lastly, there are 7 out of 103 genes frequently mutated which are related to  
166 obesity according to the DISEASES resource (27) (adj. p-value=0.019; Genes: PKHD1;  
167 ANKDD1B; SV2C; NRXN3; CDH12; ZNF248; SLC30A10). These results might hint at  
168 population-specific metabolism regulation that is linked to body weight.

169  
170 Variants that are not protein-coding may have a regulatory effect on gene and eventually protein  
171 expression. Using blood expression data obtained from RNA sequencing for the assembly  
172 individual in conjunction with the phased variant data, we identified genes whose expression  
173 differs between maternal and paternal haplotype (see Suppl. Fig. 34 for the analysis overview  
174 and Suppl. Figs. 35-36 for results). We report 1,180 such genes (see Suppl. Table 9).

175  
176 Through our analysis it will be possible to perform integrated genome and transcriptome  
177 comparisons for Egyptian individuals based on our reference genome, which might shed light  
178 on personal as well population-wide common genetic variants. Figure 2 depicts an example for  
179 such an integrated analysis. Here we use the DNA repair associated gene BRCA2, which is  
180 linked to breast and other cancer types, if mutated. The figure depicts the sample coverage  
181 based on different PacBio, 10x Genomics and Illumina whole genome short-read sequencing  
182 for a personal genome together with previously identified risk loci and common Egyptian SNPs.  
183 The bottom compares the identified SNVs and Indels from the Korean and Yoruba reference  
184 genome with our *de novo* EGYPT assembly. Visual inspection already yields significantly  
185 different variants. Furthermore, note the three significant GWAS SNPs between position  
186 32,390 and 32,400kb. These examples support the need for whole genome sequencing analysis  
187 to shed light on both mutations and structural variations on the personal and population-based  
188 genome level.

189  
190 In conclusion, we have constructed the first Egyptian reference genome, which is a hitherto  
191 unprecedented substantial step towards compiling a comprehensive, genome-wide knowledge  
192 base of personal and population-specific genetic variation. The wealth of information it  
193 provides can be immediately utilized to evaluate, on a genome-wide scale, whether a genetic  
194 region of interest is affected by personal or population-specific variation. A comprehensive  
195 annotation of these variations indicates their impact on molecular phenotypes such as RNA  
196 abundance or protein structure and therefore their potential relevance in disease and will pave  
197 the way towards a better understanding of the genomic landscape of the Egyptian population  
198 for precision medicine.

199

## 200 **Methods**

201

### 202 **Sample acquisition**

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

209

### 210 **PacBio data generation**

211 For Pacbio library preparation, the SMRTbell DNA libraries were constructed following the  
212 manufacturer's instructions (Pacific Bioscience, [www.pacb.com](http://www.pacb.com)). The SMRTbell DNA  
213 libraries were sequenced on the PacBio Sequel and generated 298.2GB of data.

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

215

### 216 **Illumina short-read data generation**

217 For 350bp library construction, the genomic DNA was sheared, and fragments with sizes  
218 around 350bp were purified from agarose gels. The fragments were ligated to adaptors and PCR  
219 amplified. The generated libraries were then sequenced on the Illumina HiSeq X Ten using  
220 PE150 and generated 312.8GB of data.

221 For the assembly individual, sequencing data from five libraries was generated at 90x genome  
222 coverage. For nine additional individuals, one library each was generated amounting to overall  
223 305x coverage of sequencing data. For the 100 individuals of Pagani et al (17), three were  
224 sequenced at high coverage (30x) and 97 at low coverage (8x). Average coverage over SNV  
225 positions for all 110 samples is provided in Supplementary Table 6.

226

### 227 **RNA sequencing data generation**

228 For RNA sequencing, ribosomal RNA was removed from total RNA, and double-stranded  
229 cDNA were synthesized, and then adaptors were ligated. The second strand of cDNA was  
230 then degraded to generate a directional library. The generated libraries with insert size of 250-  
231 300 bp were selected and amplified, and then sequenced on the Illumina HiSeq using PE150.  
232 Overall, 64,875,631 150-bp paired-end sequencing reads were generated.

233

### 234 **10x sequencing data generation**

235 For 10x genomic sequencing, the Chromium Controller was used for DNA indexing and  
236 barcoding according to the manufacturer's instructions (10x Genomics,  
237 [www.10xgenomics.com](http://www.10xgenomics.com)). The generated fragments were sheared, and then adaptors were  
238 ligated. The generated libraries were sequenced on the Illumina HiSeq X Ten using PE150 and  
239 generated 272.7 GB of data.

240 Sequencing data from four 10x libraries was generated at 80x genome coverage.

241

### 242 **Construction of draft *de novo* assemblies and meta assembly**

243 We used WTDBG2 (20) for human *de novo* assembly followed by its accompanying polishing  
244 tool WTPOA-CNS with PacBio reads and in a subsequent polishing run with Illumina short-  
245 reads. This assembly was further polished using pilon with short-read data (cf. Suppl. Methods:  
246 *WTDBG2-based assembly*).

247 An alternative assembly was generated by using FALCON, QUIVER, SSPACE-LONGREAD  
248 (28), PBJELLY (29), FRAGSCAFF (30) and PILON (31) (cf. Suppl. Methods: *FALCON-*  
249 *based assembly*).

250 Proceeding from the WTDBG2-based assembly, we constructed a meta assembly. Regions larger  
251 than 800kb that were not covered by this base assembly and were not located within centromere  
252 regions were extracted from the alternative FALCON-based assembly (Suppl. Table 3). See

253 Supplementary Figure 1 for an overview of our assembly strategy including meta-assembly  
254 construction (cf. Suppl. Methods: *Meta assembly construction*).  
255 Assembly quality and characteristics were assessed with QUAST-LG (cf. Suppl. Methods:  
256 *Assembly comparison and QC*). The extraction of coordinates for meta-assembly construction  
257 was performed using QUAST-LG output.

258

### 259 **Repeatmasking**

260 Repeatmasking was performed by using REPEATMASKER (32) with RepBase version 3.0  
261 (Repeatmasker Edition 20181026) and Dfam\_consensus (<http://www.dfam-consensus.org>) (cf.  
262 Suppl. Methods: *Repeat annotation*).

263

### 264 **Phasing**

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

270

### 271 **SNVs and small indels**

272 Calling of SNVs and small indels was performed with GATK 3.8 (33) using the parameters of  
273 the best practice workflow. Reads in each read group were trimmed using Trimmomatic (34)  
274 and mapped against reference genome hg38 using BWA, subsequently. Then the alignments for  
275 all read groups were merged sample-wise and marked for duplicates. After the base  
276 recalibration, we run the variant calling using HaplotypeCaller to obtain GVCF files.  
277 These files were then combined into batches and inputted into GenotypeGVCFs to perform  
278 joint genotyping. Lastly, the variants in the outputted VCF file were recalibrated and only  
279 considered only those variants that were flagged as "PASS" were kept for further analyses. We  
280 used FastQC (35), Picard Tools (36) and verifyBamId (37) for QC (cf. Suppl.  
281 Methods: *Small variant QC*).

282

### 283 **Variant annotation**

284 Variant annotation was performed using ANNOVAR (38) and VEP (39) (cf. Suppl. Methods:  
285 *Small variant annotation*)

286

### 287 **Structural variants**

288 Structural variants were called using DELLY2 (40) with default parameters as described on the  
289 DELLY2 website for germline SV calling (<https://github.com/dellytools/delly>) (cf. Suppl.  
290 Methods: *Structural variant QC*). Overlapping SV calls in the same individual were collapsed  
291 by the use of custom scripts. See Supplementary Methods *Collapsing structural variants* for  
292 details.

293

### 294 **Genotype principal components**

295 1000 Genomes phase 3 variant data was obtained for all European and African individuals and  
296 merged with the Egyptian variant data. Variants were excluded if their minor allele frequency  
297 was less than 5% in 1000 Genomes individuals, they violate Hardy-Weinberg-Equilibrium, are  
298 multi-allelic or within regions of high LD and/or known inversions. LD pruning was performed  
299 and remaining SNPs passed on to the SMARTPCA program (41) of the EIGENSOFT package  
300 for PC computation. See Supplementary Methods *Genotype principal components* for details.

301

### 302 **Mitochondrial haplogroups**

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

306

### 307 **Population-specific variants**

308 SNVs that are common in the 110 Egyptians and otherwise rare in the 1000 Genomes  
309 populations were considered Egyptian-specific. We considered a variant common if it has a  
310 minor allele frequency of at least 5% and as rare if it has a minor allele frequency of less than  
311 1%.

312

### 313 **Haplotypic expression analysis**

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

318

### 319 **Integrative genomics view**

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

325

### 326 **Ethics statement**

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

330

331

### 332 **References**

- 333 1. Seo J-S, Rhie A, Kim J, Lee S, Sohn M-H, Kim C-U, et al. De novo assembly and  
334 phasing of a Korean human genome. *Nature*. 2016 Oct 13;538(7624):243–7.
- 335 2. Cho YS, Kim H, Kim H-M, Jho S, Jun J, Lee YJ, et al. An ethnically relevant consensus  
336 Korean reference genome is a step towards personal reference genomes. *Nat Commun*.  
337 2016 24;7:13637.
- 338 3. Shi L, Guo Y, Dong C, Huddleston J, Yang H, Han X, et al. Long-read sequencing and  
339 *de novo* assembly of a Chinese genome. *Nature Communications*. 2016 Jun 30;7:12065.
- 340 4. Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, Rafaels N, et al. Assembly of a  
341 pan-genome from deep sequencing of 910 humans of African descent. *Nature Genetics*.  
342 2019 Jan;51(1):30.
- 343 5. Stark Z, Dolman L, Manolio TA, Ozenberger B, Hill SL, Caulfield MJ, et al. Integrating  
344 Genomics into Healthcare: A Global Responsibility. *Am J Hum Genet*. 2019 Jan  
345 3;104(1):13–20.
- 346 6. Levy-Sakin M, Pastor S, Mostovoy Y, Li L, Leung AKY, McCaffrey J, et al. Genome  
347 maps across 26 human populations reveal population-specific patterns of structural  
348 variation. *Nature Communications*. 2019 Mar 4;10(1):1025.

- 349 7. GenomeAsia 100k [Internet]. GenomeAsia 100k. [cited 2019 Mar 4]. Available from:  
350 <http://www.genomeasia100k.com/>
- 351 8. Turnbull C, Scott RH, Thomas E, Jones L, Murugaesu N, Pretty FB, et al. The 100 000  
352 Genomes Project: bringing whole genome sequencing to the NHS. *BMJ*. 2018 Apr  
353 24;361:k1687.
- 354 9. Schneider VA, Lindsay TG, Howe K, Bouk N, Chen H-C, Kitts PA, et al. Evaluation of  
355 GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of  
356 the reference assembly. *bioRxiv*. 2016 Aug 30;072116.
- 357 10. Maretty L, Jensen JM, Petersen B, Sibbesen JA, Liu S, Villesen P, et al. Sequencing and  
358 de novo assembly of 150 genomes from Denmark as a population reference. *Nature*.  
359 2017 03;548(7665):87–91.
- 360 11. Henn BM, Botigué LR, Gravel S, Wang W, Brisbin A, Byrnes JK, et al. Genomic  
361 Ancestry of North Africans Supports Back-to-Africa Migrations. *PLOS Genetics*. 2012  
362 Jan 12;8(1):e1002397.
- 363 12. Scott EM, Halees A, Itan Y, Spencer EG, He Y, Azab MA, et al. Characterization of  
364 Greater Middle Eastern genetic variation for enhanced disease gene discovery. *Nature*  
365 *Genetics*. 2016 Sep;48(9):1071.
- 366 13. Bai H, Guo X, Narisu N, Lan T, Wu Q, Xing Y, et al. Whole-genome sequencing of 175  
367 Mongolians uncovers population-specific genetic architecture and gene flow throughout  
368 North and East Asia. *Nat Genet*. 2018 Nov 5;
- 369 14. Chiang CWK, Mangul S, Robles C, Sankararaman S. A Comprehensive Map of Genetic  
370 Variation in the World's Largest Ethnic Group-Han Chinese. *Mol Biol Evol*. 2018 Nov  
371 1;35(11):2736–50.
- 372 15. ElHefnawi M, Jeon S, Bhak Y, ElFiky A, Horaiz A, Jun J, et al. Whole genome  
373 sequencing and bioinformatics analysis of two Egyptian genomes. *Gene*. 2018 Aug  
374 20;668:129–34.
- 375 16. Choudhury A, Ramsay M, Hazelhurst S, Aron S, Bardien S, Botha G, et al. Whole-  
376 genome sequencing for an enhanced understanding of genetic variation among South  
377 Africans. *Nat Commun*. 2017 Dec 12;8(1):2062.
- 378 17. Pagani L, Schiffels S, Gurdasani D, Danecek P, Scally A, Chen Y, et al. Tracing the  
379 Route of Modern Humans out of Africa by Using 225 Human Genome Sequences from  
380 Ethiopians and Egyptians. *Am J Hum Genet*. 2015 Jun 4;96(6):986–91.
- 381 18. Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic Studies.  
382 *Cell*. 2019 Mar 21;177(1):26–31.
- 383 19. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased  
384 diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*.  
385 2016 Dec;13(12):1050–4.
- 386 20. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. *bioRxiv*. 2019 Jan  
387 26;530972.



- 388 21. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome  
389 assembly evaluation with QUAST-LG. *Bioinformatics*. 2018 01;34(13):i142–50.
- 390 22. The 1000 Genomes Project Consortium. A global reference for human genetic variation.  
391 *Nature*. 2015 Oct;526(7571):68–74.
- 392 23. Schuenemann VJ, Peltzer A, Welte B, Pelt WP van, Molak M, Wang C-C, et al. Ancient  
393 Egyptian mummy genomes suggest an increase of Sub-Saharan African ancestry in post-  
394 Roman periods. *Nature Communications*. 2017 May 30;8:ncomms15694.
- 395 24. Choudhury A, Hazelhurst S, Meintjes A, Achinike-Oduaran O, Aron S, Gamielidien J, et  
396 al. Population-specific common SNPs reflect demographic histories and highlight  
397 regions of genomic plasticity with functional relevance. *BMC Genomics* [Internet]. 2014  
398 Jun 6 [cited 2019 Jun 20];15(1). Available from:  
399 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4092225/>
- 400 25. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the  
401 deleteriousness of variants throughout the human genome. *Nucleic Acids Res*. 2019 Jan  
402 8;47(D1):D886–94.
- 403 26. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al.  
404 Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic  
405 Acids Res*. 2016 08;44(W1):W90-97.
- 406 27. Pletscher-Frankild S, Pallejà A, Tsafou K, Binder JX, Jensen LJ. DISEASES: Text  
407 mining and data integration of disease–gene associations. *Methods*. 2015 Mar 1;74:83–  
408 9.
- 409 28. Boetzer M, Pirovano W. SSPACE-LongRead: scaffolding bacterial draft genomes using  
410 long read sequence information. *BMC Bioinformatics* [Internet]. 2014 Dec [cited 2019  
411 Jun 24];15(1). Available from:  
412 <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-15-211>
- 413 29. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, et al. Mind the Gap: Upgrading  
414 Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. Liu Z,  
415 editor. *PLoS ONE*. 2012 Nov 21;7(11):e47768.
- 416 30. Adey A, Kitzman JO, Burton JN, Daza R, Kumar A, Christiansen L, et al. In vitro, long-  
417 range sequence information for de novo genome assembly via transposase contiguity.  
418 *Genome Research*. 2014 Dec;24(12):2041–9.
- 419 31. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an  
420 integrated tool for comprehensive microbial variant detection and genome assembly  
421 improvement. *PLoS ONE*. 2014;9(11):e112963.
- 422 32. SMIT AFA. Repeat-Masker Open-3.0. <http://www.repeatmasker.org> [Internet]. 2004  
423 [cited 2019 Jun 21]; Available from: <https://ci.nii.ac.jp/naid/10029514778/>
- 424 33. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytzky A, et al. The  
425 Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation  
426 DNA sequencing data. *Genome Research*. 2010 Sep 1;20(9):1297–303.

- 427 34. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence  
428 data. *Bioinformatics*. 2014 Aug 1;30(15):2114–20.
- 429 35. Andrews S. FASTQC - A quality control tool for high throughput sequence data.  
430 [Internet]. Available from: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- 431 36. Picard Toolkit [Internet]. Available from: <http://broadinstitute.github.io/picard/>
- 432 37. Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis GR, et al. Detecting  
433 and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based  
434 Genotype Data. *The American Journal of Human Genetics*. 2012 Nov;91(5):839–48.
- 435 38. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants  
436 from high-throughput sequencing data. *Nucleic Acids Res*. 2010 Sep 1;38(16):e164–  
437 e164.
- 438 39. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl  
439 Variant Effect Predictor. *Genome Biology*. 2016 Jun 6;17(1):122.
- 440 40. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural  
441 variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*. 2012  
442 Sep 15;28(18):i333–9.
- 443 41. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet*.  
444 2006 Dec;2(12):e190.
- 445 42. Kloss-Brandstätter A, Pacher D, Schönherr S, Weissensteiner H, Binna R, Specht G, et  
446 al. HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial  
447 DNA haplogroups. *Hum Mutat*. 2011 Jan;32(1):25–32.
- 448 43. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast  
449 universal RNA-seq aligner. *Bioinformatics*. 2012 Oct 25;bts635.
- 450 44. Castel SE, Mohammadi P, Chung WK, Shen Y, Lappalainen T. Rare variant phasing and  
451 haplotypic expression from RNA sequencing with phASER. *Nat Commun*. 2016  
452 08;7:12817.
- 453 45. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al.  
454 Integrative genomics viewer. *Nat Biotechnol*. 2011 Jan;29(1):24–6.
- 455 46. Michailidou K, Lindström S, Dennis J, Beesley J, Hui S, Kar S, et al. Association  
456 analysis identifies 65 new breast cancer risk loci. *Nature*. 2017 Nov;551(7678):92–4.

457

## 458 **Supplementary Information**

459

## 460 **Acknowledgements**

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

463

## 464 **Author information**

465

466 *Medical Systems Biology Division, Lübeck Institute of Experimental Dermatology and Institute*  
467 *for Cardiogenetics, University of Lübeck, Lübeck, Germany*

468 Inken Wohlers, Axel Kunstner, Matthias Munz, Michael Olbrich, Anke Fähnrich & Hauke  
469 Busch

470

471 *Novogene (UK) Company Limited, Babraham Research Campus, Cambridge, United Kingdom*

472 Caixa Ma

473

474 *Medical Experimental Research Institute, Mansoura University and the American University*  
475 *in Cairo, Egypt*

476 Mohamed Salama & Shaaban El-Mosallamy

477

478 *Genetics Division, Lübeck Institute of Experimental Dermatology, University of Lübeck,*  
479 *Lübeck, Germany*

480 Misa Hirose and Saleh Ibrahim

481

#### 482 **Contributions**

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

487

#### 488 **Competing interests**

489 The authors declare no competing interests

490

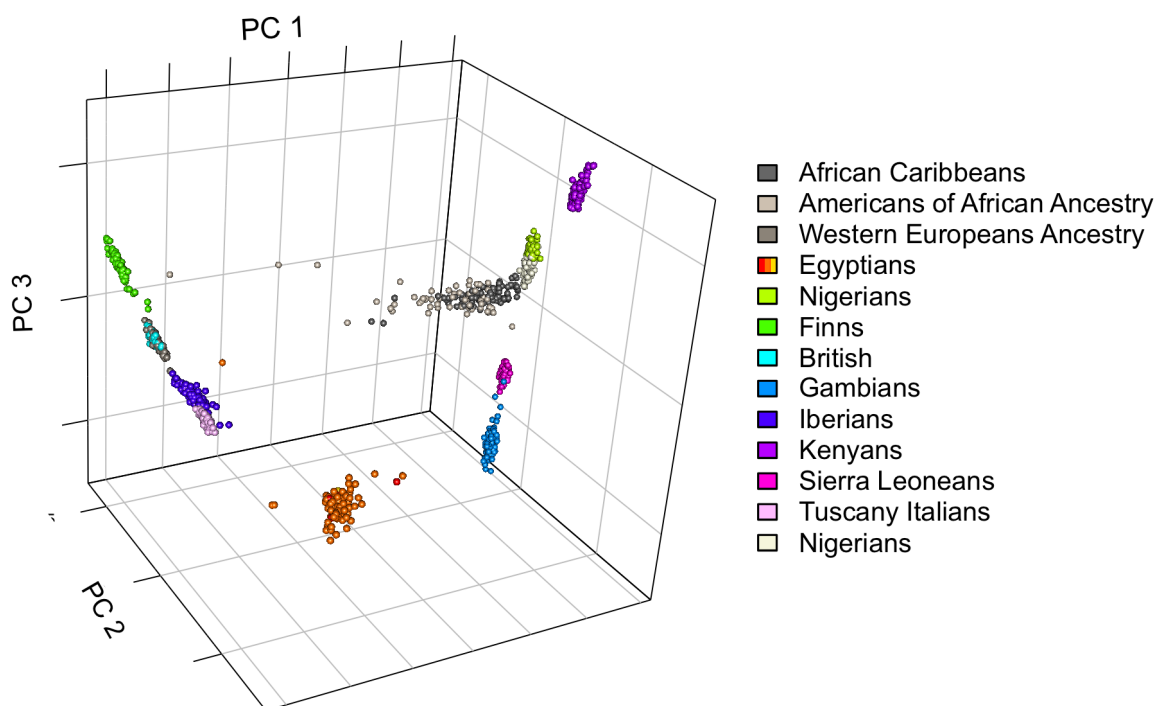
#### 491 **Corresponding authors**

492 Correspondence to Hauke Busch or Saleh Ibrahim

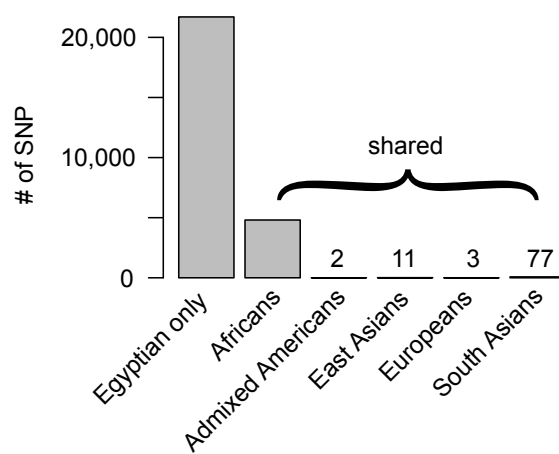
493

494

495 **Figures**  
496

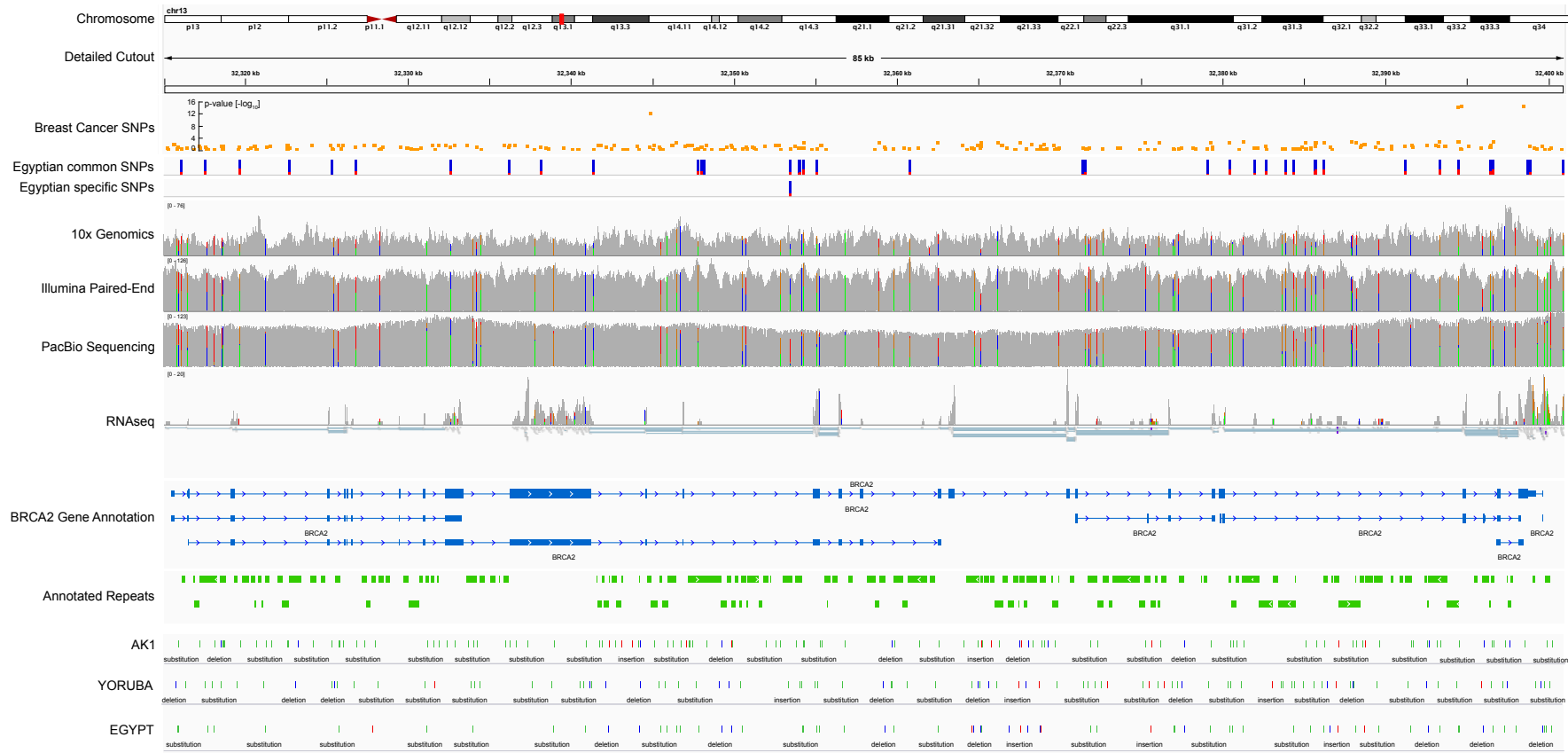


**Population-specific SNPs**



498  
499 *Figure 1: Top: PCA plot of different populations from the 1000 Genomes Project and 110 Egyptian genomes from Pagani et*  
500 *al. as well as from our own study. Bottom: Egyptian population-specific SNPs and SNPs that are common in Egyptians and*  
501 *specific to a single continental population.*

502  
503



504  
505  
506  
507  
508  
509

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

510 **Tables**

511

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

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

514

515

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

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

518