An integrated personal and population-based Egyptian genome reference

Supplement

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

* These authors contributed equally to this work

Contents

Supplementary Methods **3**
- `WTDBG2`-based assembly . . . 3
- `FALCON`-based assembly . . . 3
- Meta assembly construction . . . 3
- Assembly comparison and QC . . . 4
- Repeat annotation . . . 4
- Variant phasing . . . 4
- Small variant QC . . . 5
- Small variant annotation . . . 5
- Structural variant QC . . . 5
- Collapsing structural variants . . . 6
- Genotype-based principal components . . . 6
- Mitochondrial haplogroups . . . 6
- Haplotypic expression . . . 7
- Data integration with the GWAS catalog . . . 7
- Sequencing read mapping to GRCh38 . . . 8
- Alignment to GRCh38 . . . 8
- Assembly-based variant identification . . . 9
- Gene-centric integrative data views . . . 9

Supplementary Figures **10**

Supplementary References **44**

Supplementary Methods

WTDBG2-based assembly

The WTDBG2-based assembly was constructed with WTDBG2 [1] using preset option "-x sq", which applies default options for PacBio sequencing data-based assembly and "-g 3g" to specify an approximate genome size of 3 Gb. WTPOA-CNS, a consensuser for WTDBG2, was applied as described on the WTDBG2 website. Briefly, WTPOA-CNS was applied once for obtaining a draft assembly, then for polishing with PacBio data (after mapping reads to the draft assembly with MINIMAP2 [2] and option "-x map-pb") and finally for polishing with Illumina short-read data using preset "-x sam-sr" for using default options for short read polishing (after mapping to the PacBio-polished assembly with BWA-MEM [3]). We remapped the short reads to this polished assembly using again BWA-MEM to apply another round of polishing using PILON [4], an established short-read polishing tool. This slightly improved the already very good overall assembly quality.

FALCON-based assembly

The FALCON-based assembly was generated with FALCON (version 0.7), followed by consensusing using QUIVER (version smrtlink_5.0.1). Program SSPACE-LONGREAD (version 1-1) was used to assemble to scaffolds and PBJELLY (version 15.8.24) to fill gaps. For scaffolding using 10x data, FRAGSCAFF (version 140324) was used. The assembly has been polished using PILON (version 1.22) [4].

Meta assembly construction

Centromere regions have been obtained from the UCSC Genome browser (URL https://bit.ly/2IYfpgW) by selecting "group: all tables" and "table: centromeres". GRCh38 coordinates of regions greater than 800 kb of the GRCh38 reference genome which were not covered by the WTDBG2-based assembly and were not located within centromere regions were identified and coordinates of aligned segments of the FALCON-based assembly which overlap these gaps were obtained. The aligned regions of the FALCON-based assembly were extracted and added to the WTDBG2-based assembly as individual, additional contigs. The considered assembly gaps and corresponding novel FALCON-based contigs which were added in the meta

assembly are given in Suppl. Table 3. By filling 31 gaps, which affect 10 chromosomes, this approach slightly improves the GRCh38-covered genome fraction and the k-mer-based completeness of the meta assembly in comparison to the WTDBG2-based assembly (see also Suppl. Table 2).

Assembly comparison and QC

We compared the assembly with the latest version of the reference genome assembly GRCh38, which we obtained from ftp://ftp.ensembl.org/pub/release-93/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz. In the following, we refer to this reference assembly version as GRCh38.

For assembly quality assessment, we compared our two assemblies, EGYPT_wtdbg2 and EGYPT_falcon and the final meta assembly EGYPT with the published assembly of a Korean individual [5] obtained from https://www.ncbi.nlm.nih.gov/assembly/GCA_001750385.2 as well as a chromosome-level assembly of a Yoruba individual (1000 Genomes ID NA19420) downloaded from https://www.ncbi.nlm.nih.gov/assembly/GCA_001524155.4. We assessed all quality measures available by QUAST_LG [6] using default options and, as suggested for large genomes, the options "—large" and "—memory-efficient". We also computed k-mer based statistics using QUAST_LG with the option "—k-mer-stats". For QUAST_LG analysis concerning the number of genes contained in the assembly, we used gene annotation from Ensembl version 95 obtained from ftp://ftp.ensembl.org/pub/release-95/gff3/homo_sapiens/Homo_sapiens.GRCh38.95.gff3.gz. Interactive overview of QUAST_LG statistics as well as the ICARUS [7] contig viewer displaying all assemblies is provided at www.egyptian-genome.org.

Repeat annotation

We performed repeat annotation with the tool REPEATMASKER version 4.0.7 (Smit *et al.* (2015)) using human repeat databases from Repbase (RepBaseRepeatMaskerEdition-20181026.tar.gz obtained from www.girinst.org) and Dfam_consensus (http://www.dfam-consensus.org, distributed with REPEATMASKER). REPEATMASKER was run with option "-q" for increased speed at slightly lower sensitivity and "-xsmall", which returns a file with sequences masked such that repetitive regions are denoted in lower case letters and non-repetitive regions in capital letters.

Variant phasing

Variant phasing was performed based on 10x sequencing data from four libraries by using LONGRANGER WGS (version 2.2.2) with the GRCh38 reference files provided by 10x genomics at http://cf.10xgenomics.com/supp/genome/refdata-GRCh38-2.1.0.tar.gz. We have four 10x libraries available and barcodes may be shared between libraries, but denote different molecules. Thus, we adjusted the configuration MRO file of the 10x pipeline management

framework Martian as described on the 10x genomics web site. This results in a suffix being appended to barcodes indicating the source library number.

Small variant QC

We summarized and QC'ed SNV and small indels with respect to number of SNV, percentage of missing SNV calls, mean depth and heterozygosity per individual (Suppl. Fig. 9) to detect potential outliers. Sequencing depth did not substantially influence variant calling, but the aforementioned numbers are correlated with (and likely driven by) sequencing depth (Suppl. Fig. 8). Small indels are most frequent, and insertions of a given size have been called about as often as deletions, which is denoted by a symmetric size distribution (Suppl. Fig. 10).

Small variant annotation

SNVs and small indels have been annotated using `ANNOVAR` [8] as well as `VEP` [9] and classified according to variant location (Suppl. Fig. 12). Exonic variants have further been classified according to variant consequence types (Suppl. Fig. 12). Deleterious effects of protein-coding variants were predicted using `CADD` [10], `POLYPHEN` [11] and `SIFT` [12] scores provided by the annotation tools. The stand-alone `VEP` tool used specifies the internally used data sources as VEP="v95" time="2019-04-12 09:14:39" ensembl-io=95.78ccac5 ensembl=95.4f83453 ensembl-funcgen=95.94439f4 ensembl-variation=95.858de3e 1000genomes="phase3" COSMIC="86" ClinVar="201810" ESP="V2-SSA137" HGMD-PUBLIC="20174" assembly="GRCh38.p12" dbSNP="151" gencode="GENCODE 29" genebuild="2014-07" gnomAD="170228" polyphen="2.2.2" refseq="2018-07-10 14:50:52 - GCF_000001405.38_GRCh38.p12_genomic.gff" regbuild="1.0" sift="sift5.2.2". Allele frequency information has been obtained via the Ensembl API (accessed 09/2019). The subset of Egyptian common variants, which has not been assigned an rsID has been annotated additionally with the newest available VEP version via the Ensembl web interface (https://www.ensembl.org/Tools/VEP). The data sources here are "v98" time="2019-10-24 12:37:40" cache="/nfs/public/release/ensweb-data/latest/tools/www/e98/vep/cache/homo_sapiens/98_GRCh38" db="homo_sapiens_core_98_38@hh-mysql-ens-species-web-1" 1000genomes="phase3" COSMIC="89" ClinVar="201907" ESP="V2-SSA137" HGMD-PUBLIC="20184" assembly="GRCh38.p13" dbSNP="152" gencode="GENCODE 32" genebuild="2014-07" gnomAD="r2.1" polyphen="2.2.2" regbuild="1.0" sift="sift5.2.2". This reduces the number of common Egyptian SNPs without rsID to 48 (cf. Suppl. Table 12).

Structural variant QC

We computed for the 135,819 SVs called by `DELLY2` [13] in the cohort of 110 Egyptians the number of SVs, percentage of missing SV calls and heterozygosity per individual (see

Suppl. Fig. 15). These numbers are correlated (Suppl. Fig. 16). Most of 121,141 DELLY filter passing SV calls thereof are deletions (n=95,889), but also inversions (n=11,477), duplications (n=10,092), translocations (n=3,275) and insertions (n=408) have been called. Corresponding SV lengths are up to 10^8 with most SV calls affecting sequence up to 10kB (Suppl. Fig. 14).

Collapsing structural variants

Structural variant calls of type deletion, insertion, inversion or duplication were collapsed per individual by dividing variants calls into groups with respect to their chromosomal region (chromosome, start position, end position). That is, each group only contains variant calls that are overlapping with each other and with none of the variant calls in any other group. Translocations were collapsed per individual by merging variants with the same original chromosomal position and the same new chromosomal position. The number of SVs per individual after collapsing and a boxplot of the corresponding SV sizes are provided in Suppl. Figs 17,18 and 19 for deletions, inversions and duplications.

Genotype-based principal components

We performed genotype-based principal component analysis using SMARTPCA [14] from the EIGENSOFT software package. We used 1000 Genomes genotypes from ftp/release/20130502/supporting/GRCh38_positions/ALL.chr*_GRCh38.genotypes.20170504.vcf.gz of all available African and European populations summarized in Suppl. Table 8. Variants were filtered using PLINK [15] version 1.9 (www.cog-genomics.org/plink/1.9). We kept biallelic variants which in the 1000 Genomes individuals had MAF greater 5% and were not violating Hardy-Weinberg equilibrium ($p = 10^{-6}$). Corresponding Egyptian genotypes were extracted and only variants without missing genotypes kept. Variants in high LD regions or known inversions have been removed (chr6:25 Mb–33.5 Mb, chr5:44 Mb–51.5 Mb, chr8:8 Mb–12 Mb, chr11:45 Mb–57 Mb) [16]. LD pruning was performed using PLINK with "–indep-pairwise 1000 10 0.2". Statistical significance of principal components was computed using the Tracy-Widom statistic available in the EIGENSOFT package.

Mitochondrial haplogroups

Genomic DNA samples were processed for library preparation, as previously described in the Human mtDNA Genome protocol for Illumina Sequencing Platform (http://emea.support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_legacy/human-mtdna-genome-guide-15037958-01.pdf). In brief, two primer sets [MTL-F1 (AAA GCA CAT ACC AAG GCC AC) and MTL-R1 (TTG GCT CTC CTT GCA AAG TT); MTL-F2 (TAT CCG CCA TCC CAT ACA TT) and MTL-R2 (AAT GTT GAG CCG TAG ATG CC)] were used to amplify the mtDNA by long-range

PCR. Library preparation was performed using a Nextera XT DNA Library Preparation Kit (Illumina Inc., CA, USA), and the 10-pM library was sequenced on the Illumina MiSeq sequencing platform (2x150 bp paired-end reads) (Illumina Inc.). Haplogroup assignment was performed using `HaploGrep 2` [17]. In brief, `HaploGrep` weights each polymorphism present in PhyloTree17 (a phylogenetic tree of worldwide human mitochondrial DNA variation) based on its informativeness to define haplogroups. The set of SNPs in the input file are classified as informative or remaining (not informative). A score is given based on the weights of the "informative SNPs" but it is "penalized" by the number of remaining SNPs. For quality control, our recomputed haplogroup frequencies of 100 Egyptian individuals from Pagani *et al.* [18] have been compared with frequencies reported in the corresponding publication. Recomputed and previously reported frequencies were identical.

Haplotypic expression

The tool `PHASER` [19] (version 1.1.1) was used to calculate haplotypic expression for the Egyptian assembly individual using our 10x-phased variants as gold-standard phasing, i.e. without the option to perform phasing from the RNA sequencing data. Suppl. Fig. 34 provides an overview of the analysis. RNA sequencing data obtained from blood was pre-processed with `FASTP` [20]. `STAR` version 2.6.1.c [21] was used to align reads to GRCh38 using Ensembl version 95 annotation. RNA-Seq QC was performed with `QUALIMAP` [22]. In total, expression of 58,738 genes was quantified. In a two-step process, `PHASER` was first used to calculate haplotypic counts which resulted in 16,566 genes with non-zero counts being analyzed (Suppl. Fig. 35). Of those, we excluded genes with less than 30 reads mapped to both haplotypes. In a second step, for the remaining 7,202 genes, allelic fold change was computed and significance tested using a Binomial test (Suppl. Fig. 36). Both steps were performed as described in the HowTo provided by the `PHASER` authors (`https://stephanecastel.wordpress.com/2017/02/15/how-to-generate-ase-data-with-phaser`). Multiple testing correction using FDR was performed. Allelic expression results for 1,180 significant genes at FDR of 5% are provided in Suppl. Table 12.

Data integration with the GWAS catalog

We downloaded associations and ancestry information from the GWAS catalog (https://www.ebi.ac.uk/gwas/, version 2019/08/27), a curated database of published genome-wide association studies. We split associations into 3,064 trait-specific data sets according to the mapped terms from the Experimental Factor Ontology (EFO). For every disease, only associations from studies of individuals from European descent according to column "BROAD ANCESTRAL CATEGORY" were kept. In a next screen we kept only associations of variants for which there is another variant within 1 MB that is associated with the same trait. In order to select a set of common, high-quality European GWAS SNPs we used the genome-wide genotypic data of 503 individuals of European descent from the state-of-the-art

1000 Genomes data and keep bi-allelic variants with more than 5% minor allele frequency, not violating Hardy-Weinberg Equilibrium (p=0.000001), from which we selected variants with rsIDs matching associations in the GWAS catalog. For every trait we computed all-versus-all linkage disequilibrium of the remaining associated SNPs and output all pairs of SNPs within 1 MB with LD of $R^2 \geq 0.8$. Associations of SNPs with at least one other associated SNP in LD were kept and can be considered a replicated association signal. At this step we thus exclude associations reported in multiple studies, but always for the exactly same SNP and not for at least one other SNP in LD. For 585 trait ontology terms we kept a minimum of 2 associations. In a next step, we kept one associated SNP per trait and locus, where we define a locus as a region of 1 MB with a tag SNP. We automatically selected the tag SNP based on the number of studies reporting the rsID position for the respective trait. Note that we may have exclude additional, independent association signals that are closer than 1 MB from a SNP selected as tag SNP. The resulting tag SNPs were combined over all traits; there are 4,008 such SNPs. Of these, 261 (6.5%) have not been called in the Egyptian cohort. We investigated these closer and 42 are located within the MHC locus (chr6:28510020-33480577) and very few have other variants (largely indels) with start position +/-15 BP, possibly compromising the variant calling. For the tag SNPs we computed proxy SNPs that are in LD ($R^2 \geq 0.8$) within 1 MB using genotypes of the European and Egyptian cohort, respectively. In visualizations we display data for 3,959 tag SNPs with missing genotypes in at most 10% of Egyptian individuals.

Sequencing read mapping to GRCh38

We performed quality control on raw FASTQ files using `FASTQC` (`https://www.bioinformatics.babraham.ac.uk/projects/fastqc`). All NGS data was mapped to GRCh38. For PacBio long reads, we used `MINIMAP2` [2] with option "-x map-pb" for mapping PacBio genomic reads. For Illumina short read data, we used `BWA-MEM` [3] with default options. QC for mapped read data was performed using `SAMTOOLS STATS` [23]. Linked 10x genomics reads were processed by `LONGRANGER BASIC` (version 2.2.2) to generate barcoded FASTQ files, which were subsequently processed by `LONGRANGER WGS`, which outputs also a BAM mapping file of linked reads.

Alignment to GRCh38

We aligned all assemblies with GRCh38 using `NUCMER` from the recently updated `MUMMER4` suite [24] with default options, except additionally using "—mum" which denotes that alignment anchor matches need to be unique in both GRCh38 and assembly. We filtered alignments using `MUMMER4`'s delta-filter with option "-q", which maps each assembly position onto the best hit in GRCh38, allowing for GRCh38 overlaps, as well as option "-1", which performs 1-to-1 alignment allowing for rearrangements. Alignments from 1-to-1 mapping have been further used for assembly-based variant detection using `NUCDIFF` [25]. `MUMMER4`'s mummerplot

script has been used for dotplot visualization (Suppl. Figs. 3-7).

Assembly-based variant identification

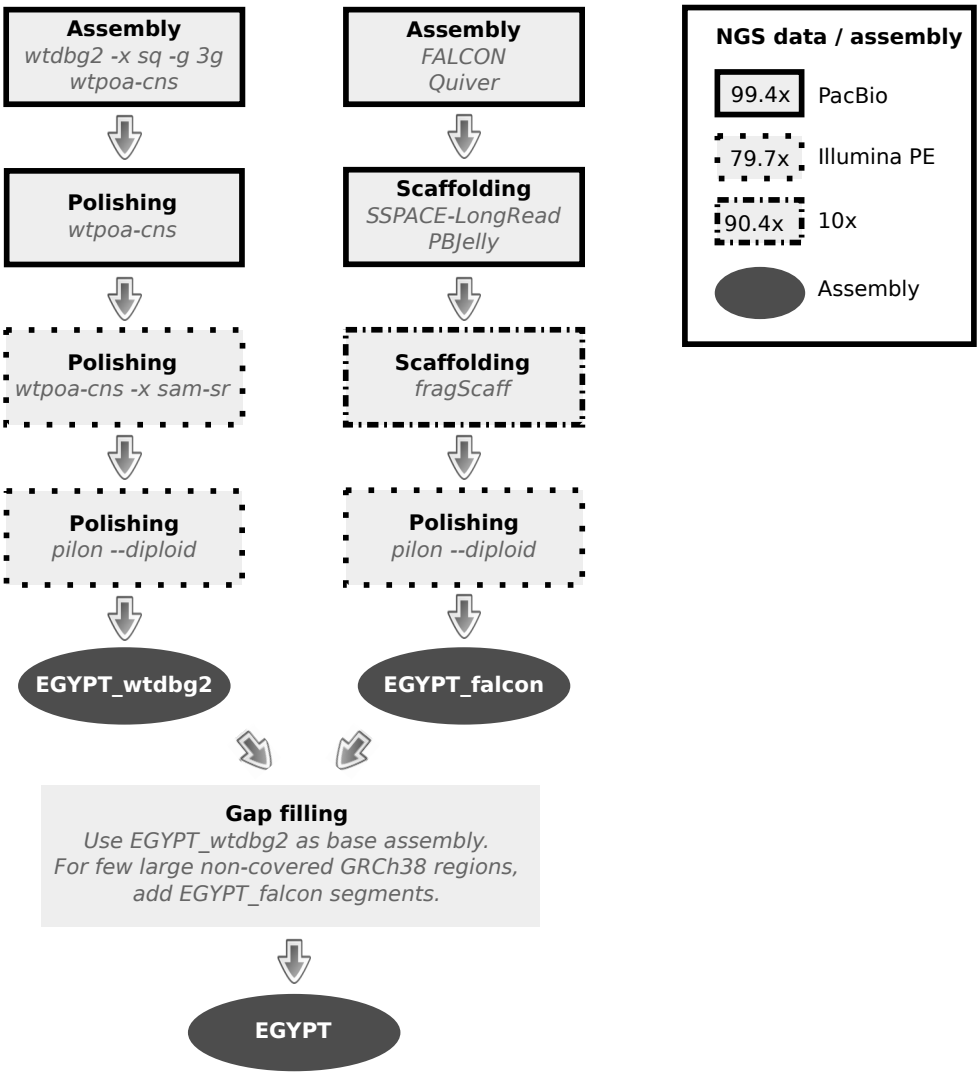
We identified SNVs, indels and structural variants present in the assembly by aligning it with the human reference genome GRCh38 and detecting sequence differences. This was achieved by using the tool `NUCDIFF` [25] on selected genomic regions utilizing genome-wide EGYPT-to-GRCh38 alignments from 1-to-1 mapping (see last section).

Gene-centric integrative data views

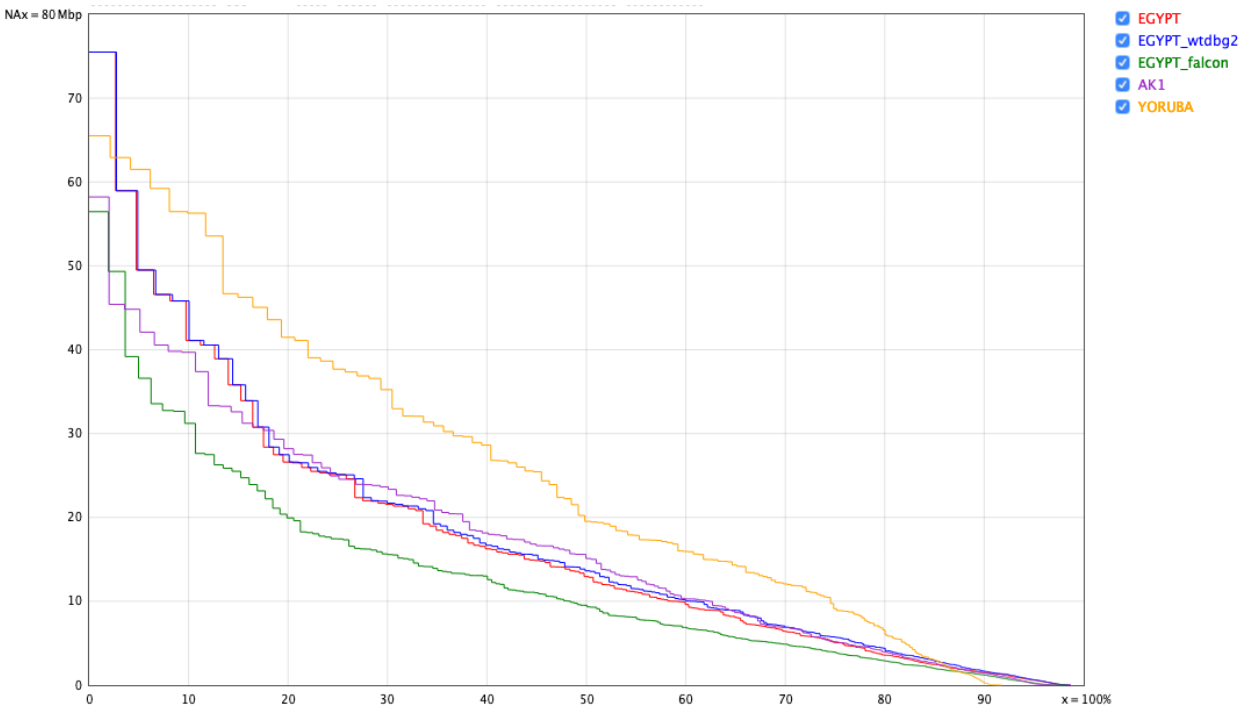
Most users of the Egyptian genome reference will be interested in a specific genetic region or a specific gene and would like to investigate, if and where this region or gene is affected by personal or population-specific variation. To facilitate such analysis, we implemented a workflow to extract all relevant Egyptian data within a specified region centered around a gene of interest (by default +/- 100 kB). The resulting files can be viewed in the Integrative Genomics Viewer (IGV) [26]. The gene-centric data contains

- EGYPT PacBio long reads mapped to GRCh38 (BAM)
- EGYPT Illumina short reads mapped to GRCh38 (BAM)
- EGYPT 10x linked reads mapped to GRCh38 (BAM)
- EGYPT blood RNA-Seq reads mapped to mapped to Ensembl genes version 95 given with GRCh38 coordinates (BAM)
- EGYPT assembly differences to GRCh38 (BED)
- AK1 assembly differences to GRCh38 (BED)
- YORUBA assembly differences to GRCh38 (BED)
- All, common and population-specific small variants of 110 Egyptian individuals including EGYPT (VCF)
- VEP annotations of small variants (flat file)
- Small variants from all 1000 Genomes phase 3 individuals with genotypes (VCF)
- Ensembl gene annotation version 95 (GTF)
- Variant data from dbSNP (VCF)

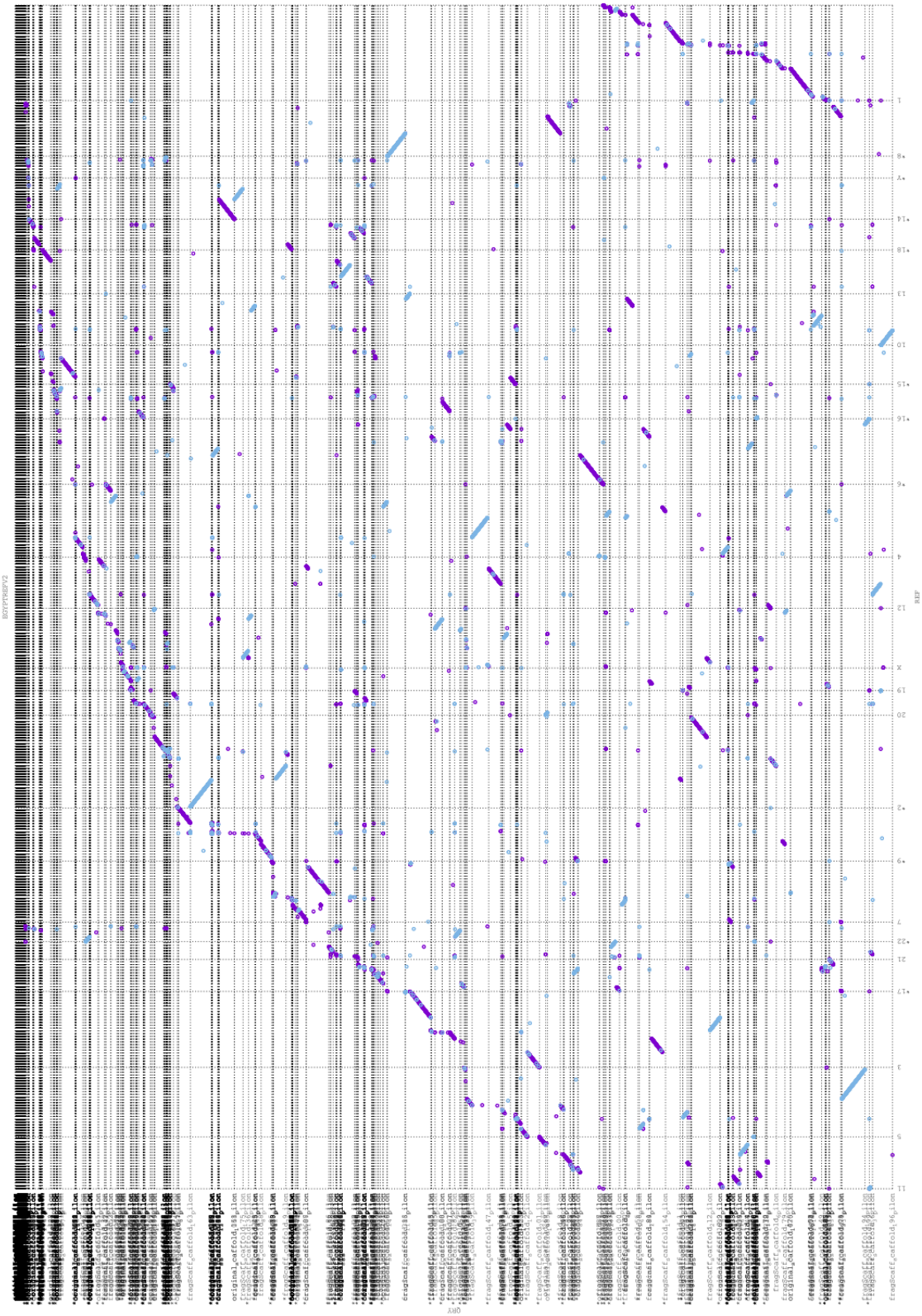
Supplementary Figures



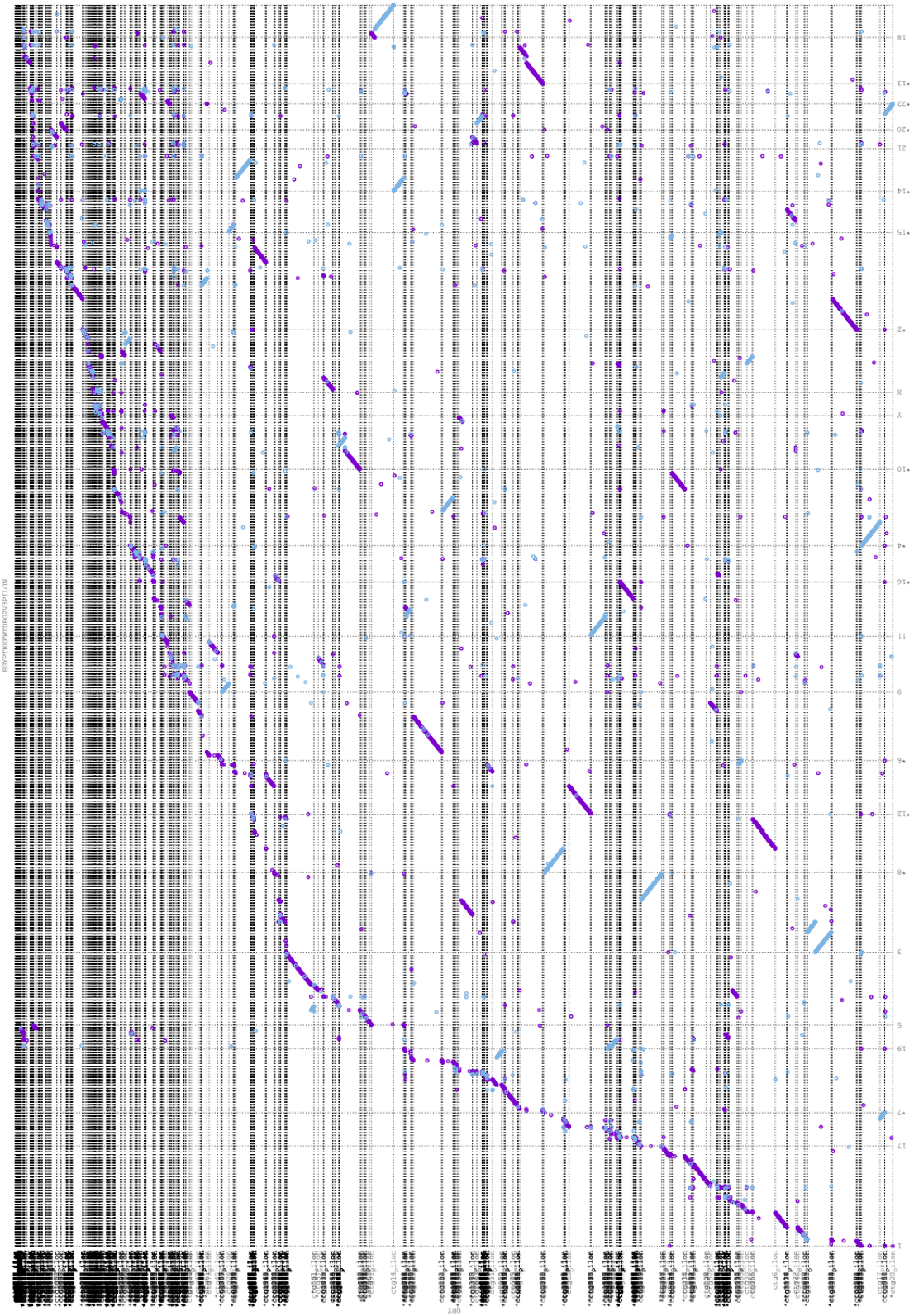
Supplementary Figure 1: Overview of the *de novo* assembly strategy with the NGS data types used in every step. EGYPT_wtdbg2 and EGYPT_falcon are two individual, alternative high quality *de novo* assemblies, which we combined into a final meta assembly that we refer to as EGYPT.



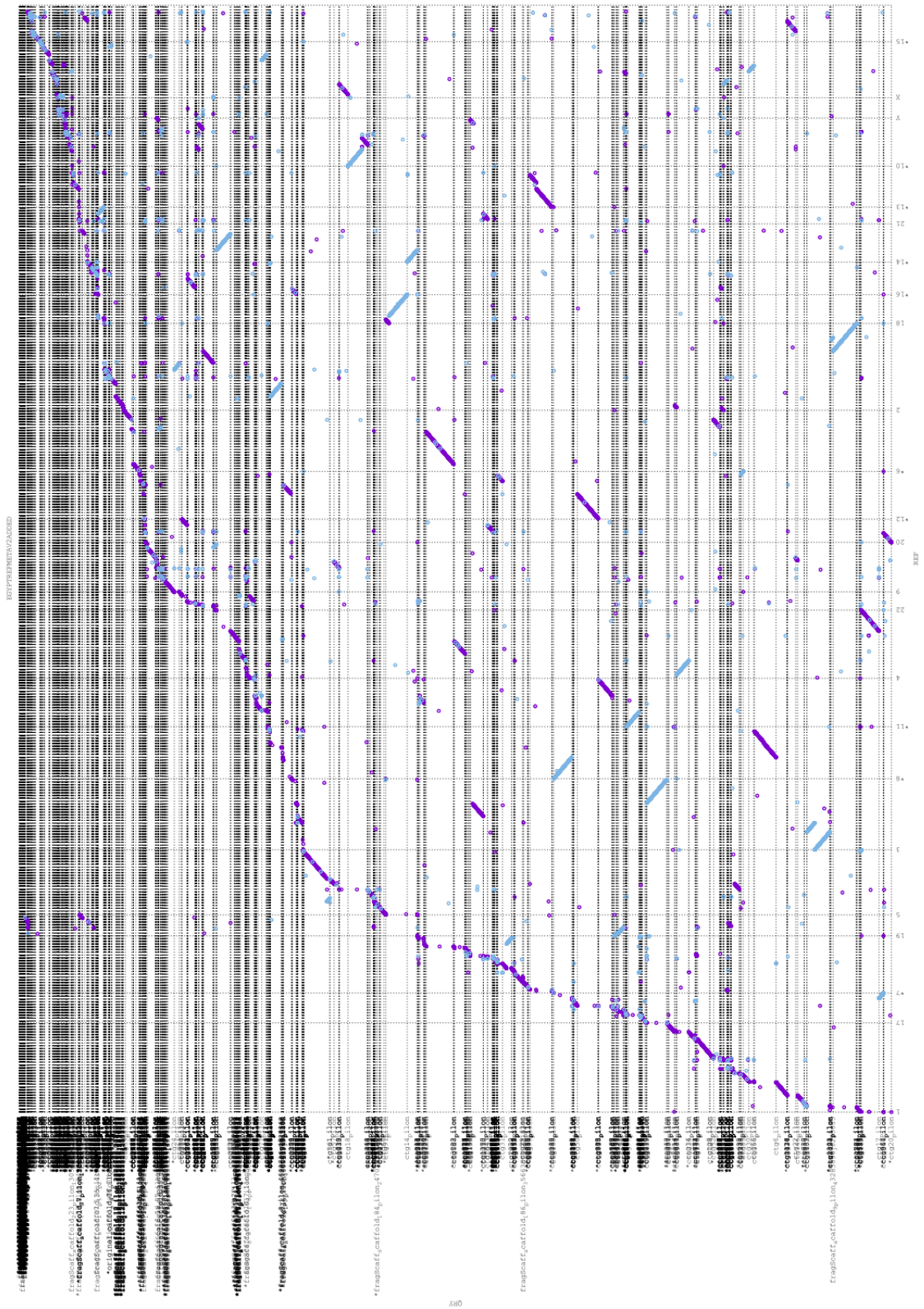
Supplementary Figure 2: NAx display of the N-values of the five assemblies; NA-values are like N-values, but contigs are split at misassembly sites.



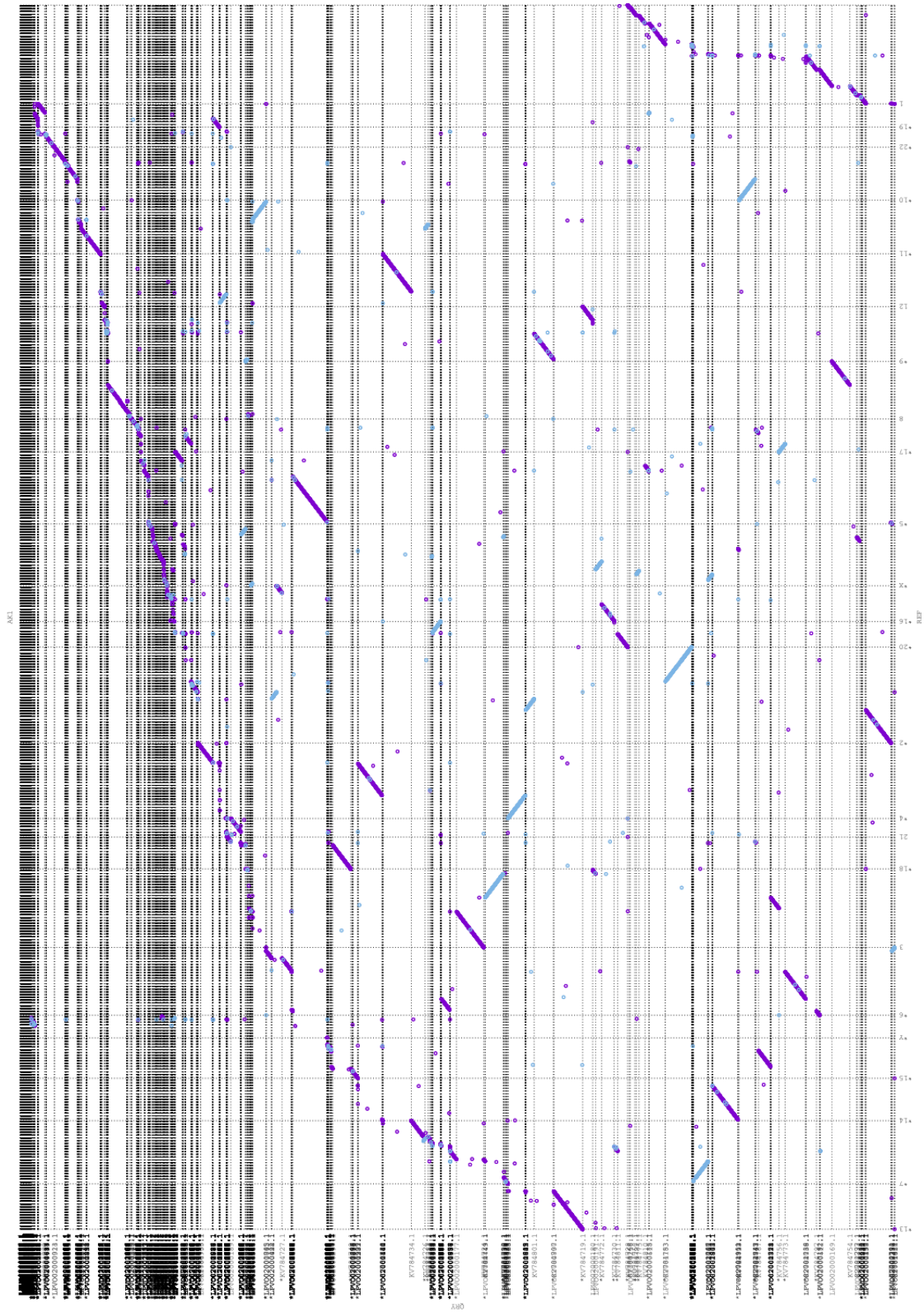
Supplementary Figure 3: EGYPT_falcon assembly dotplot for 1-to-1 alignment with GRCh38.



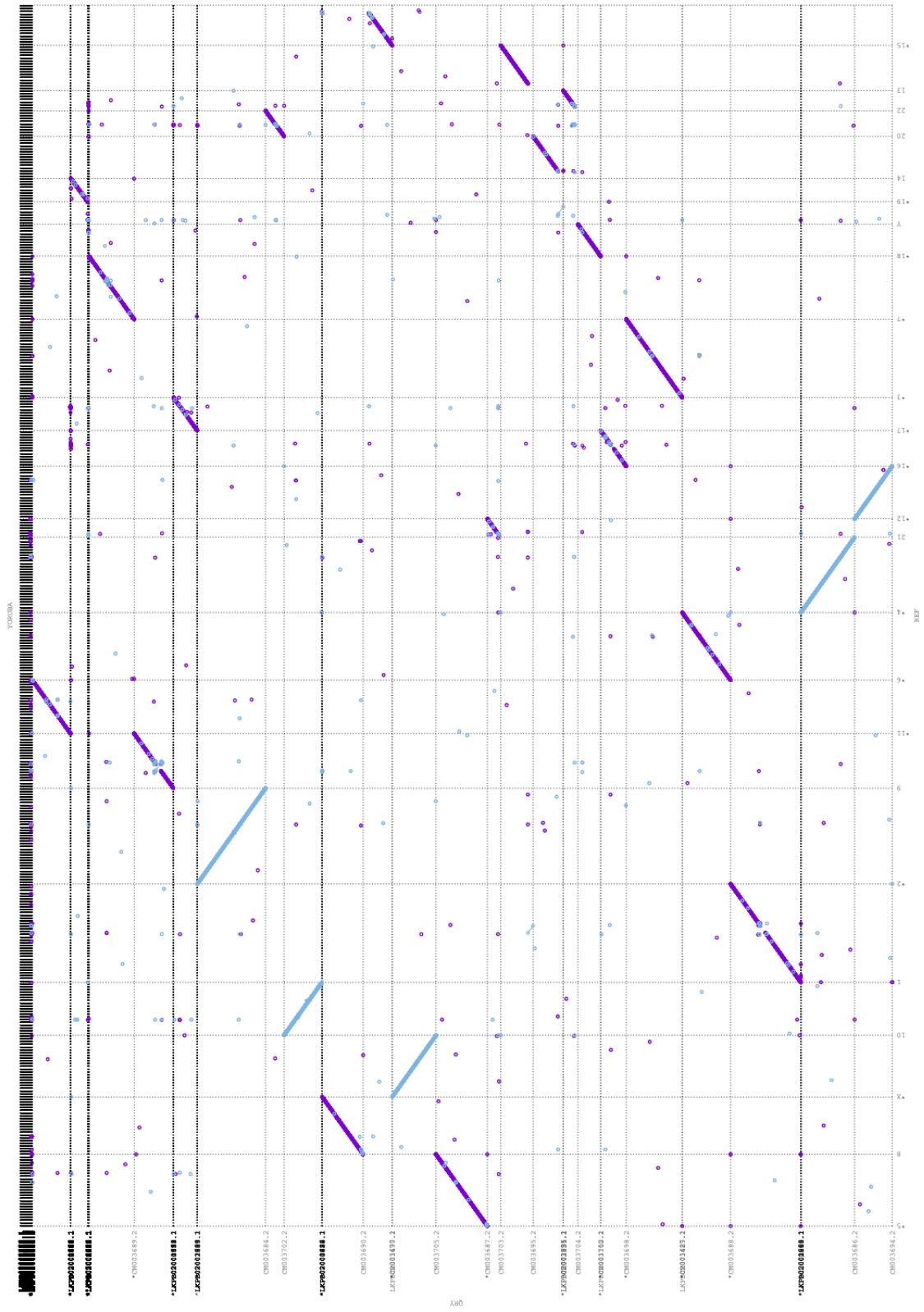
Supplementary Figure 4: EGYPT_wtdbg2 assembly dotplot for 1-to-1 alignment with GRCh38.



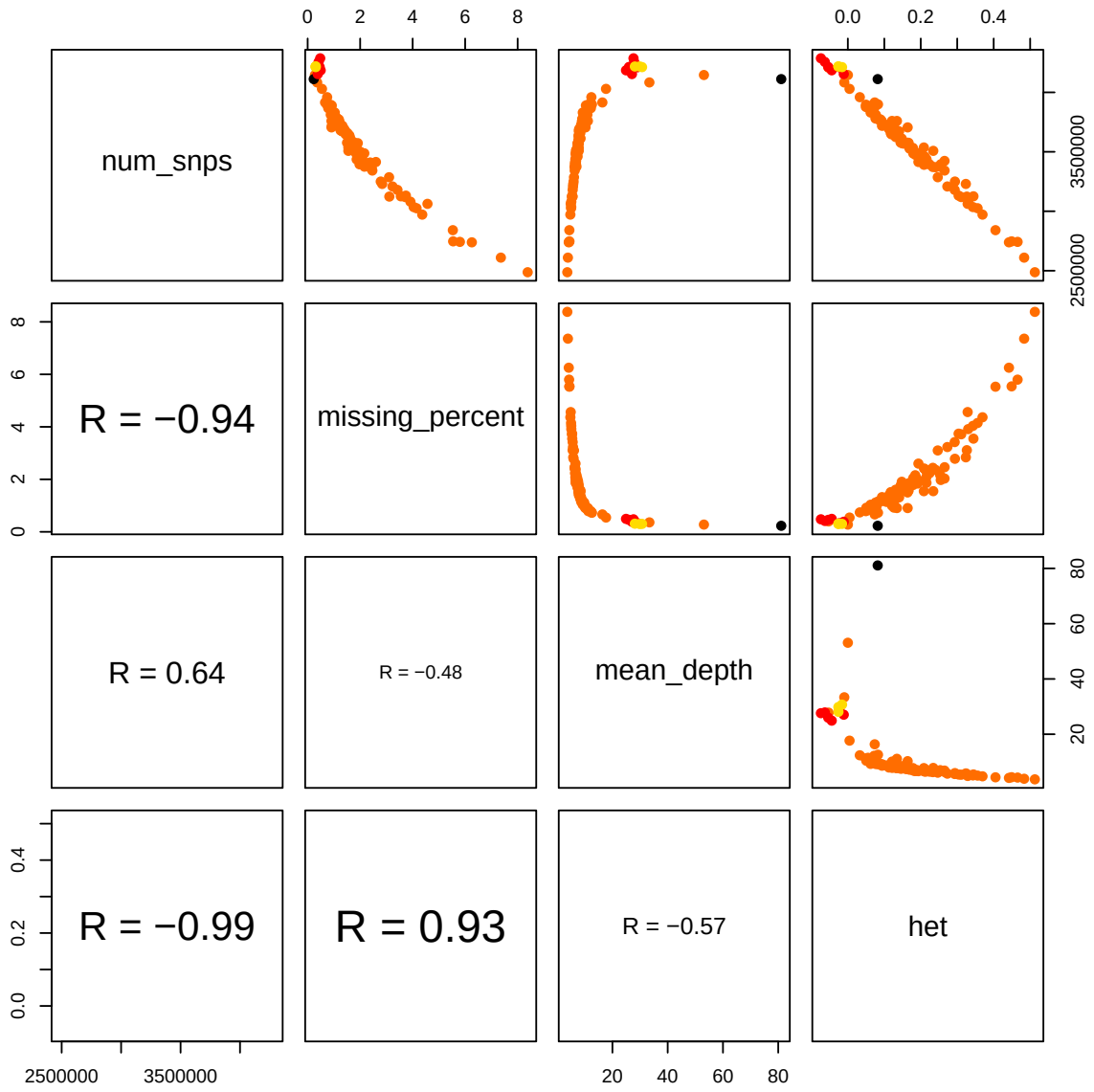
Supplementary Figure 5: EGYPT assembly dotplot for 1-to-1 alignment with GRCh38.



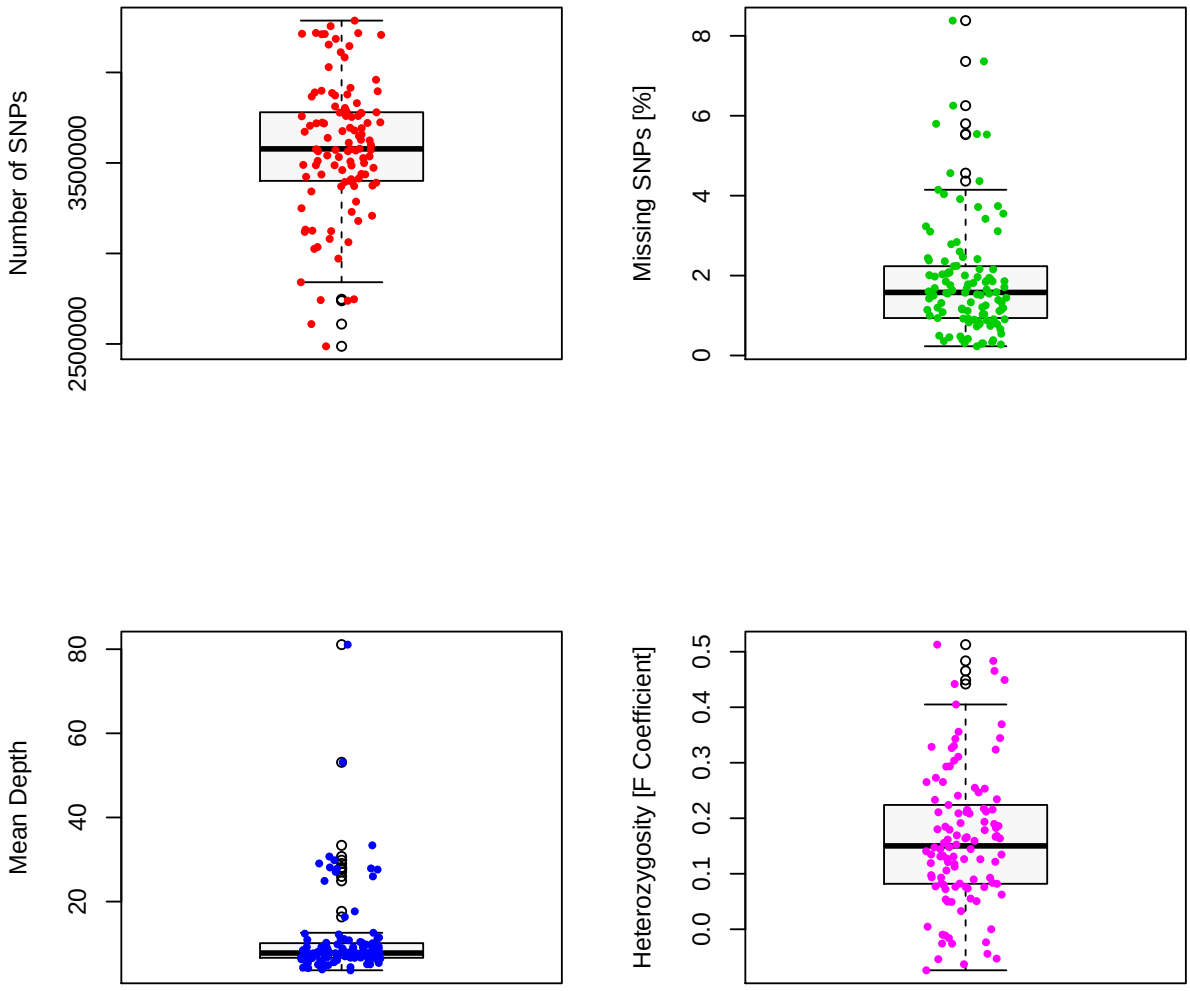
Supplementary Figure 6: AK1 assembly dotplot for 1-to-1 alignment with GRCh38.



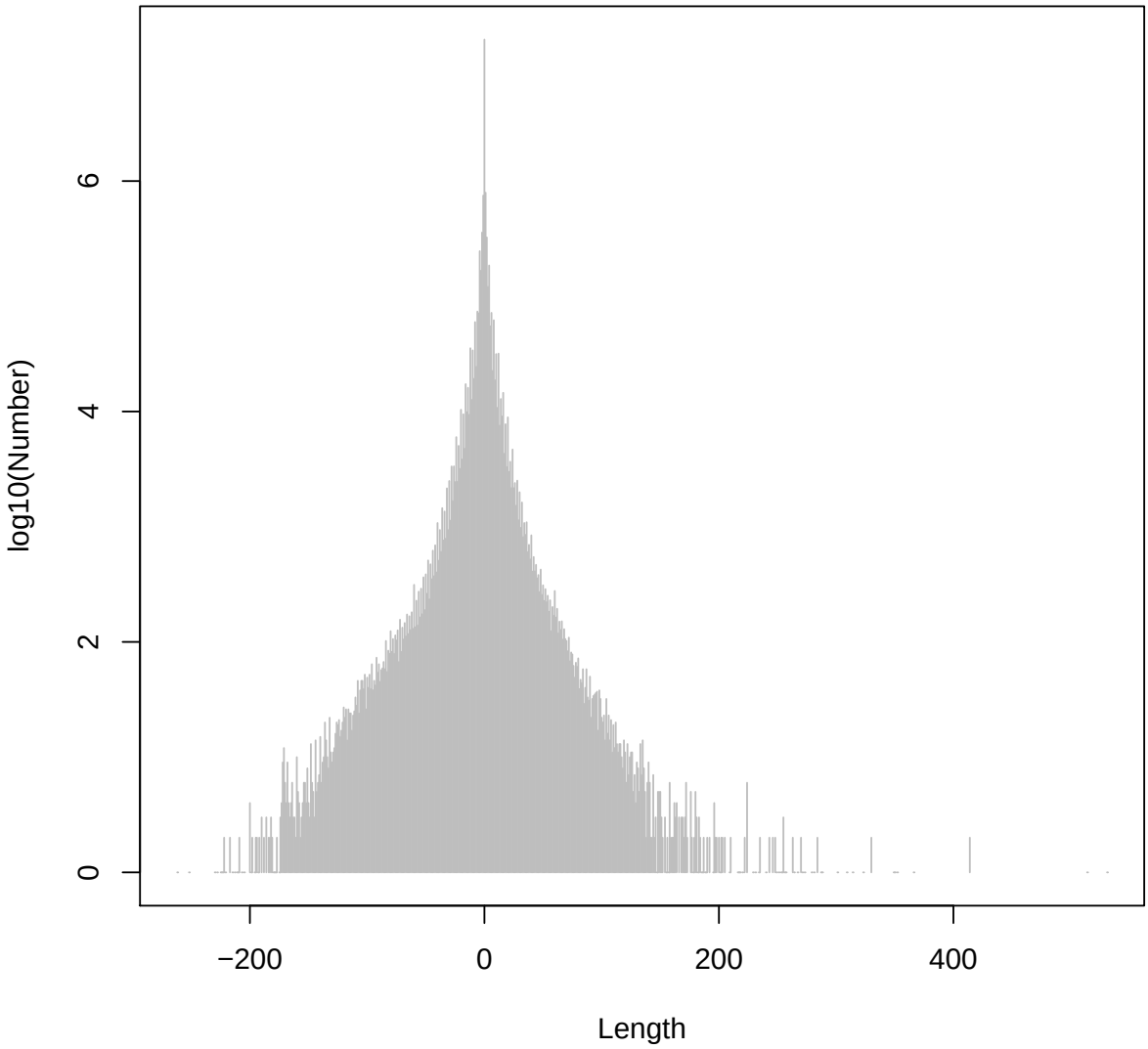
Supplementary Figure 7: YORUBA assembly dotplot for 1-to-1 alignment with GRCh38.



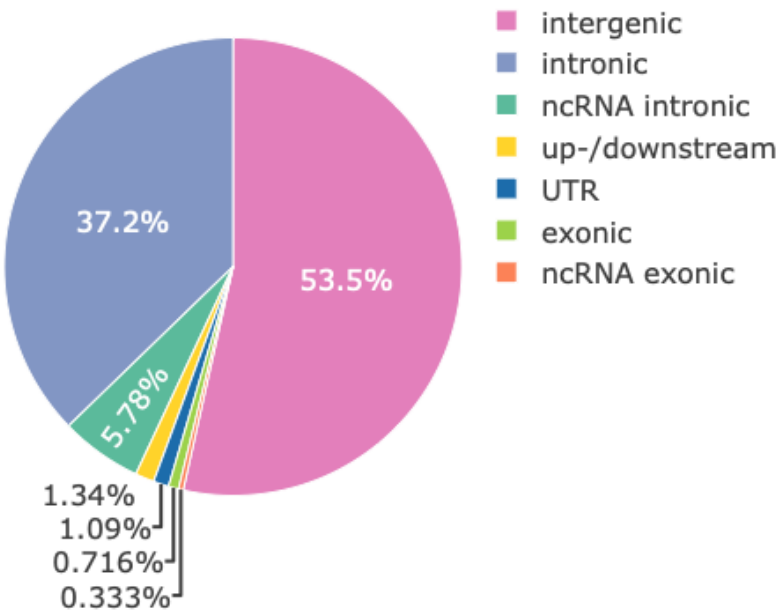
Supplementary Figure 8: Correlations between variant statistics. Pagani *et al.* individuals are orange, Egyptians from Delta red, Upper Egyptians yellow and the EGYPT individual black.



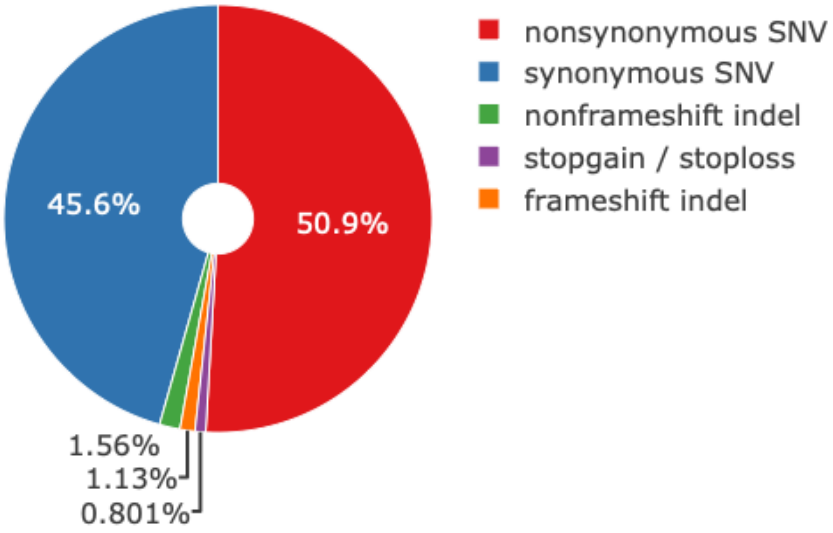
Supplementary Figure 9: Boxplots of number of SNV, percentage of missing SNV calls, mean depth and heterozygosity, each for the cohort of 110 Egyptians.



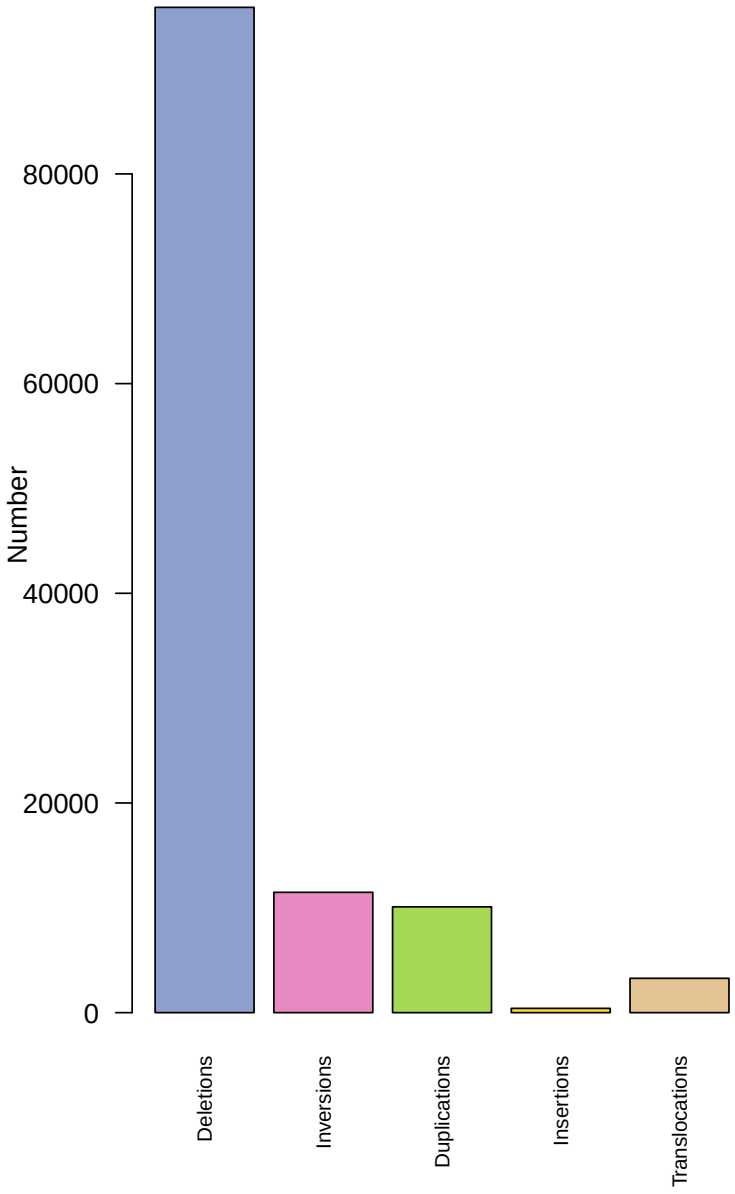
Supplementary Figure 10: Histogram of indel sizes: negative length refers to deletions, positive length to insertions



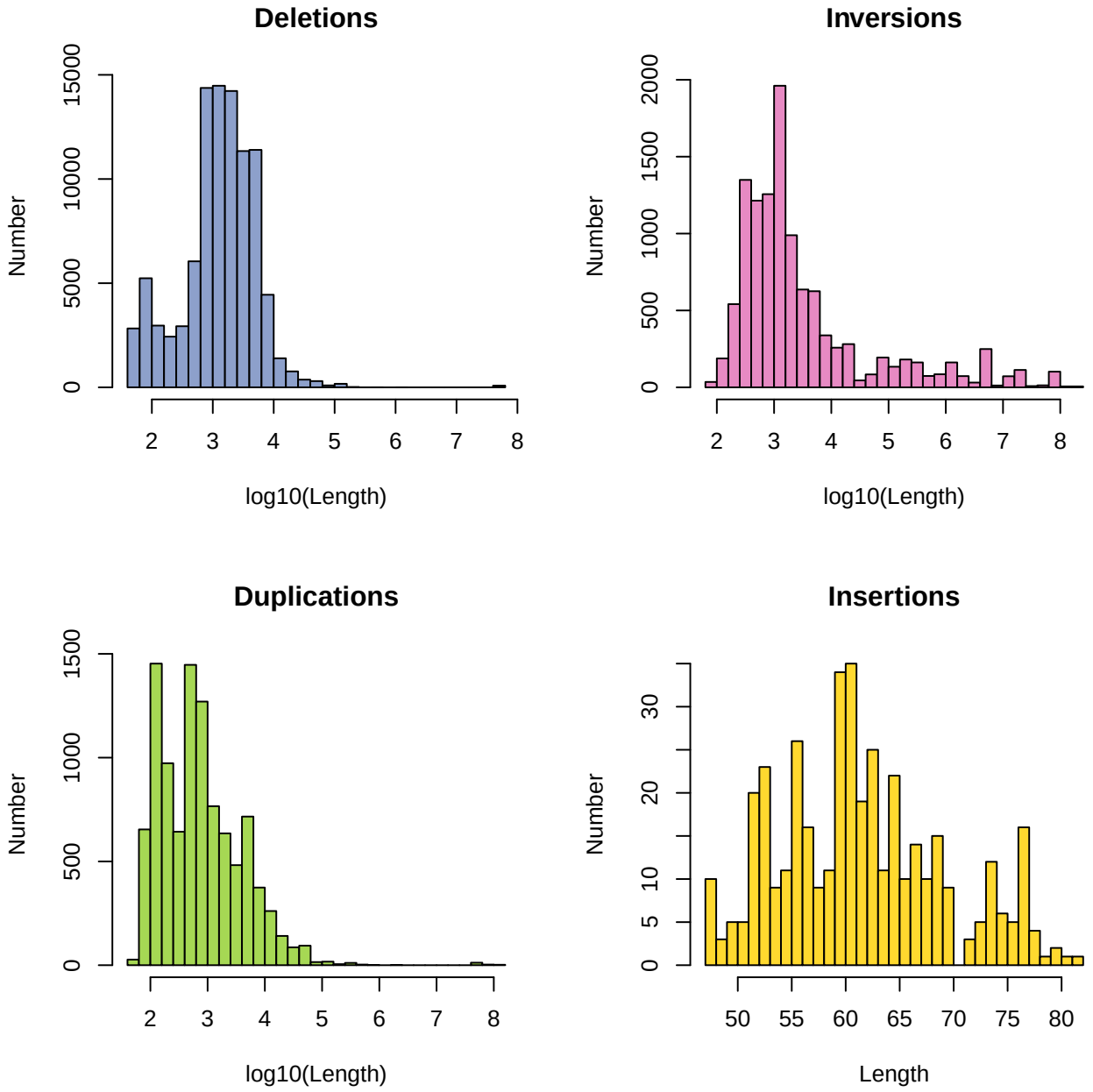
Supplementary Figure 11: Classification of small variants according to genomic location according to `Annovar` annotations.



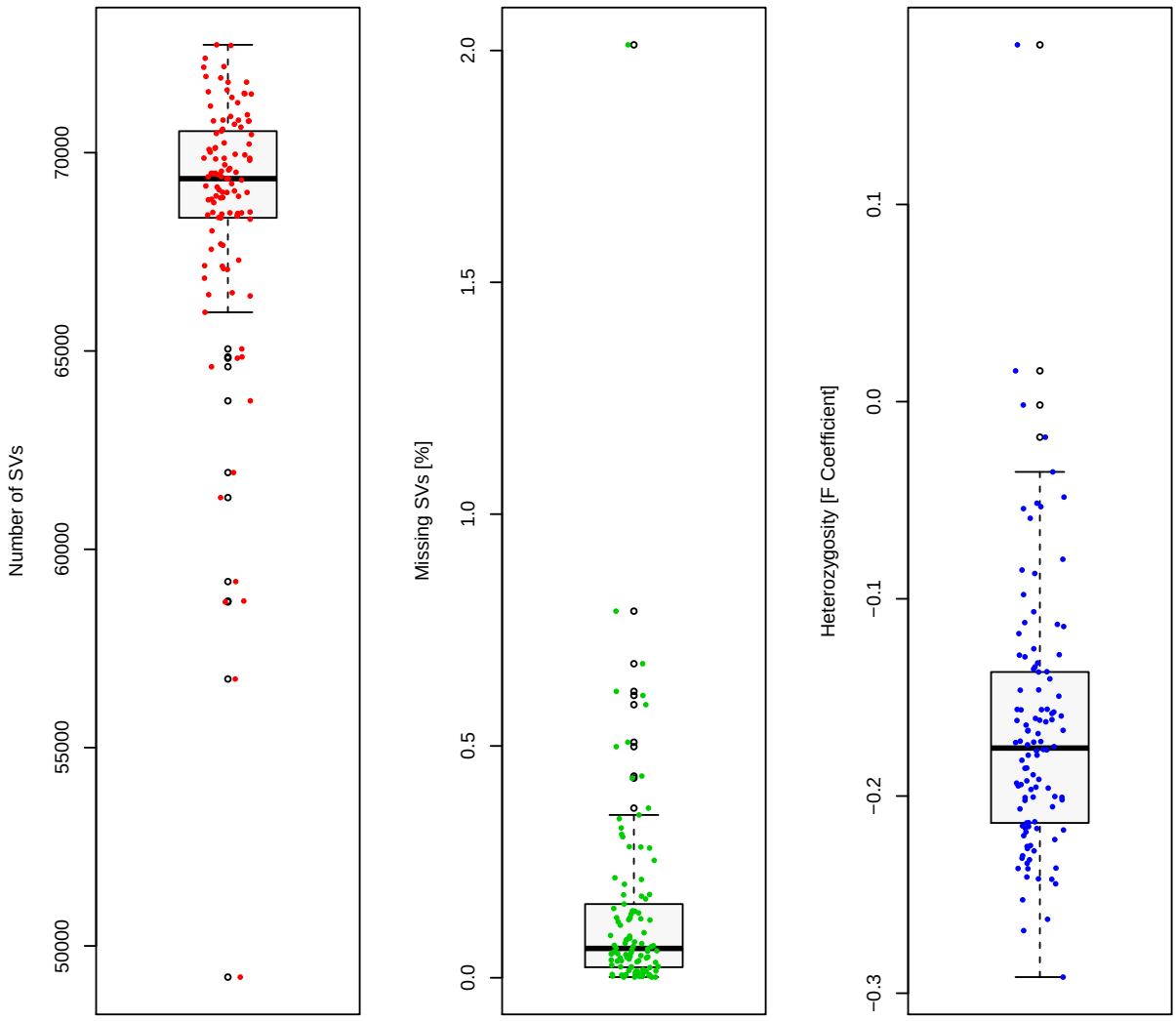
Supplementary Figure 12: Classification of small variants according to exonic variant consequences.



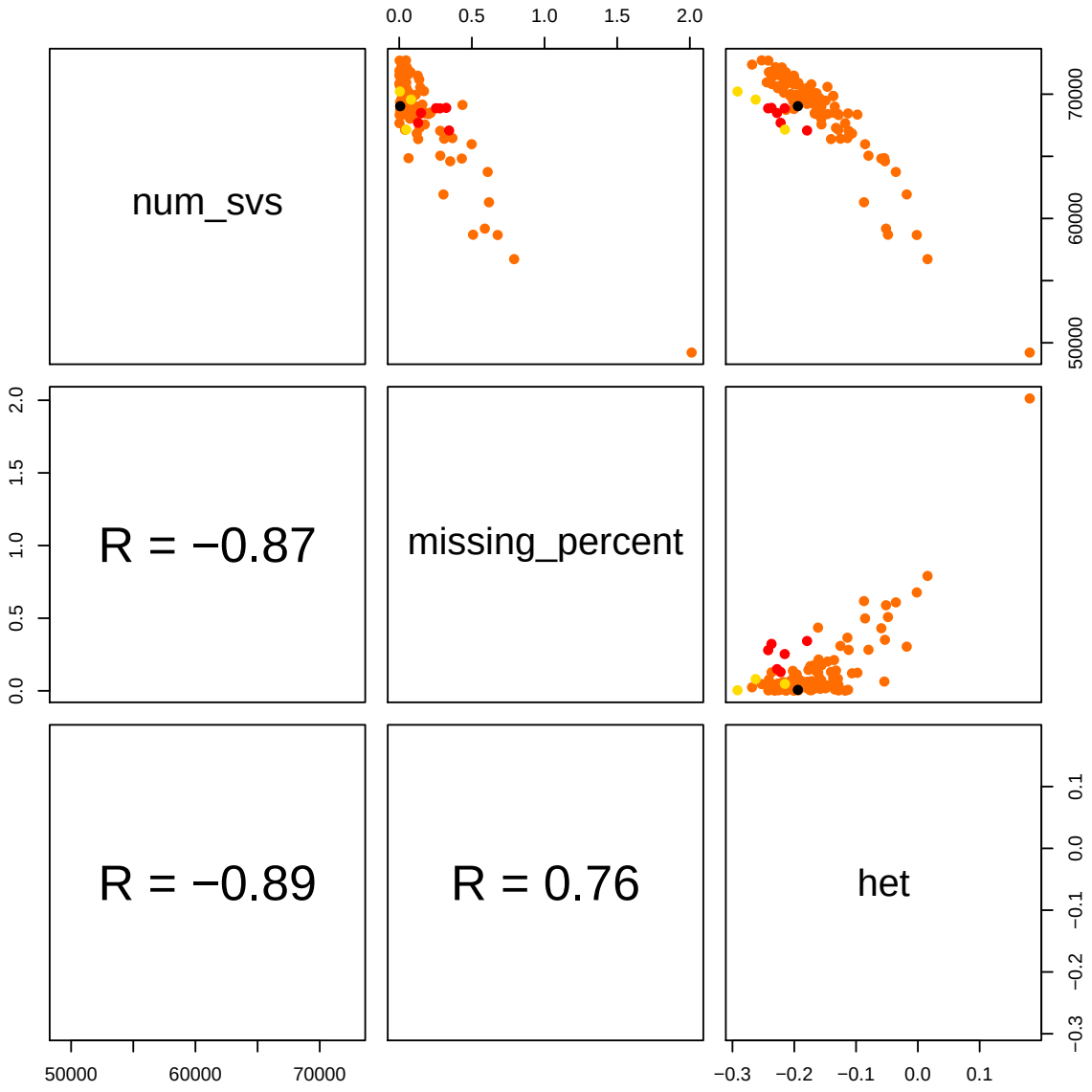
Supplementary Figure 13: Histogram of types of SV calls.



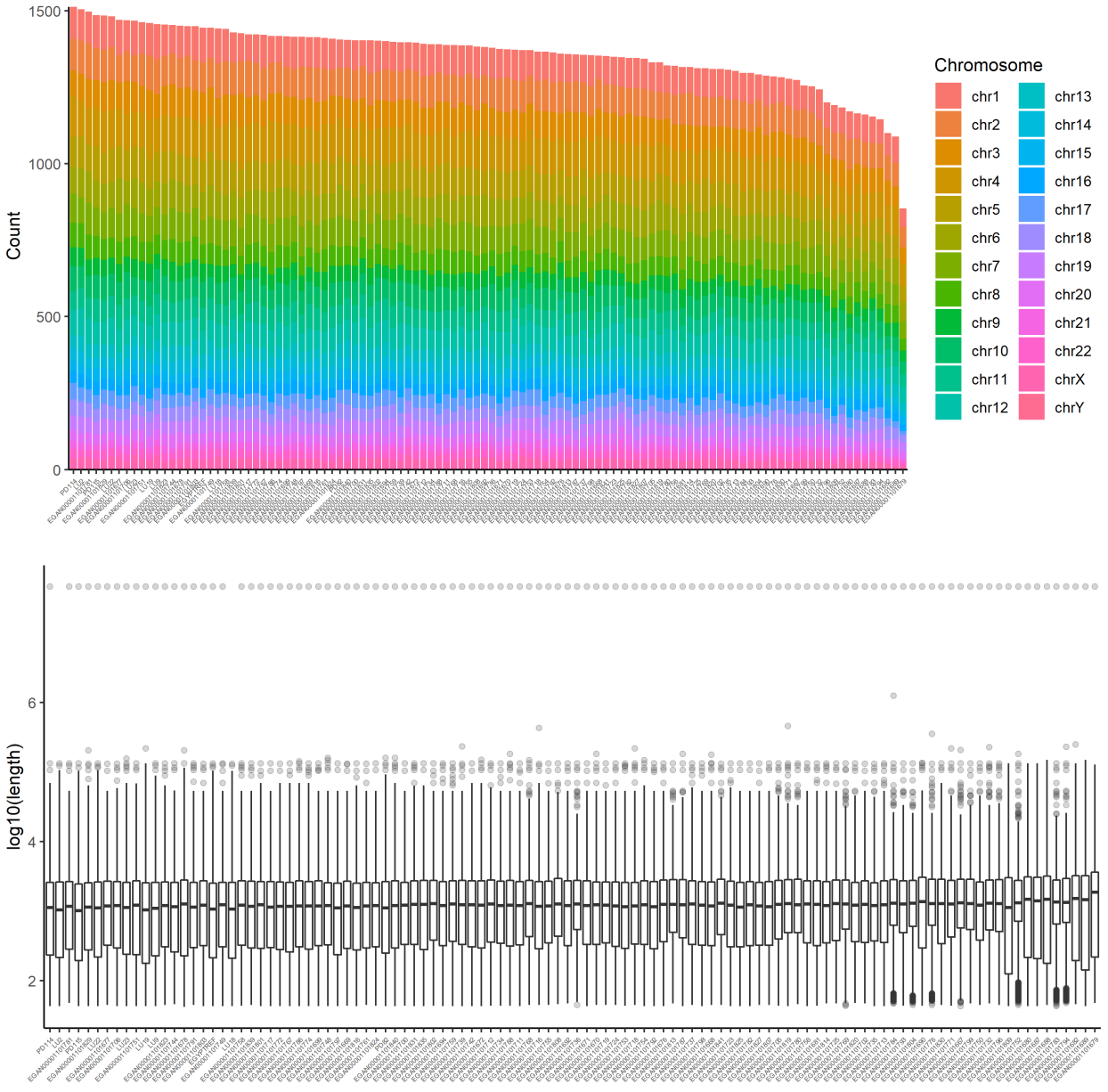
Supplementary Figure 14: Histogram of lengths of deletions, inversions, duplications and insertions passing `DELLY` filter.



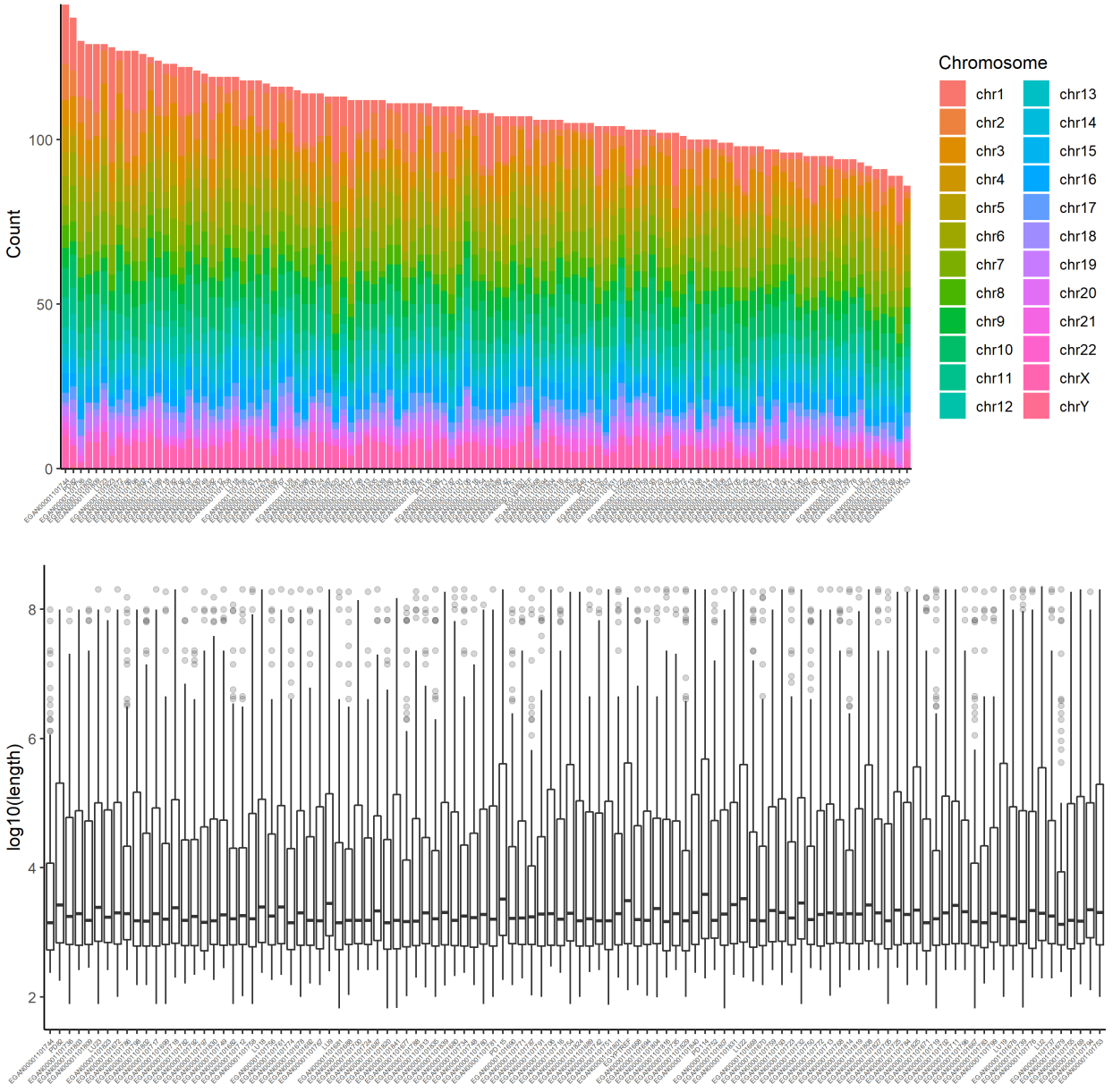
Supplementary Figure 15: Boxplots of number of SV calls, missing SV calls and SV-based heterozygosity for 110 Egyptian individuals.



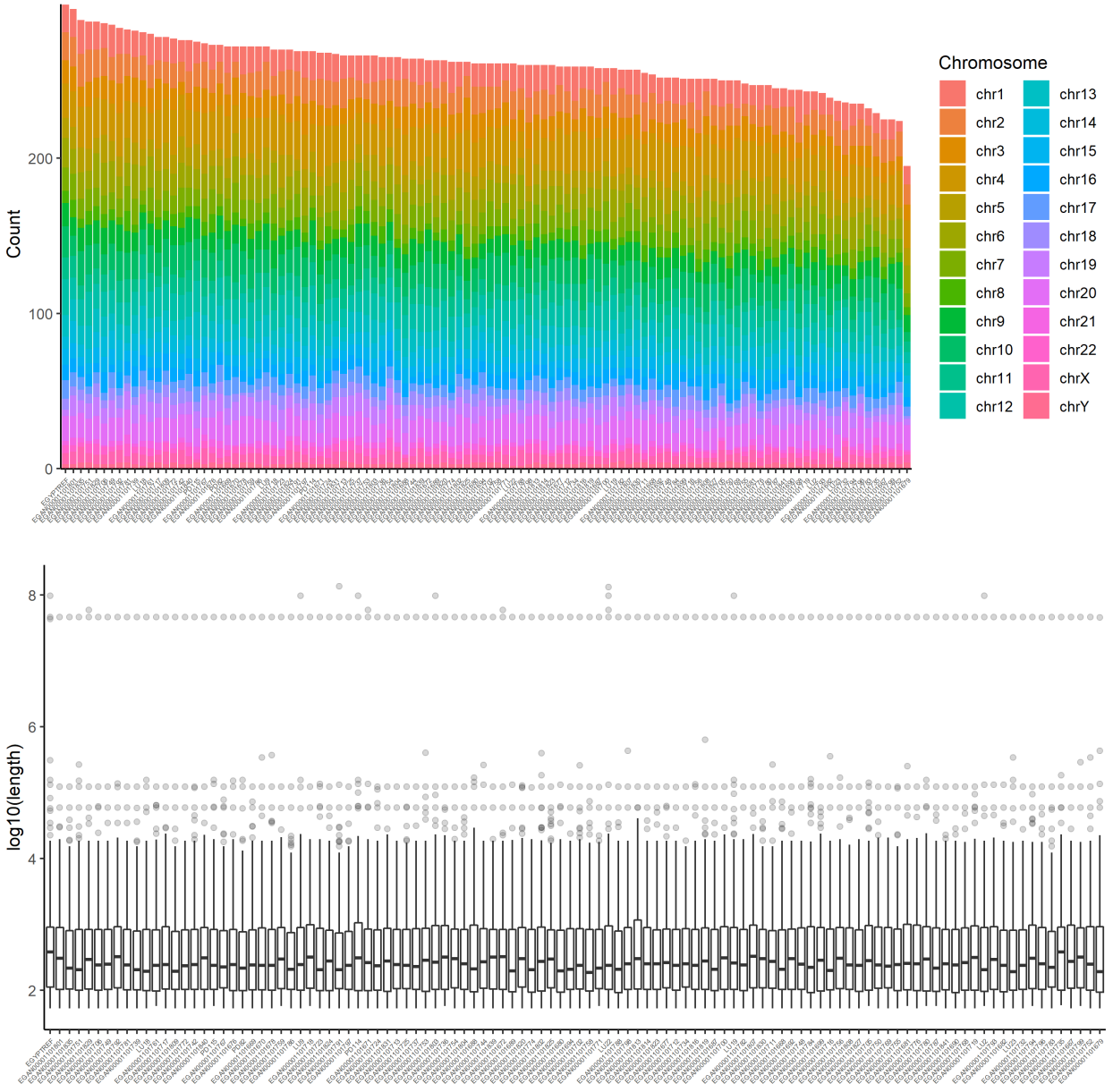
Supplementary Figure 16: Scatterplots and correlation of SV call numbers, missing SV calls and heterozygosity. Pagani *et al.* individuals are orange, Egyptians from Delta red, Upper Egyptians yellow and the EGYPT individual black.



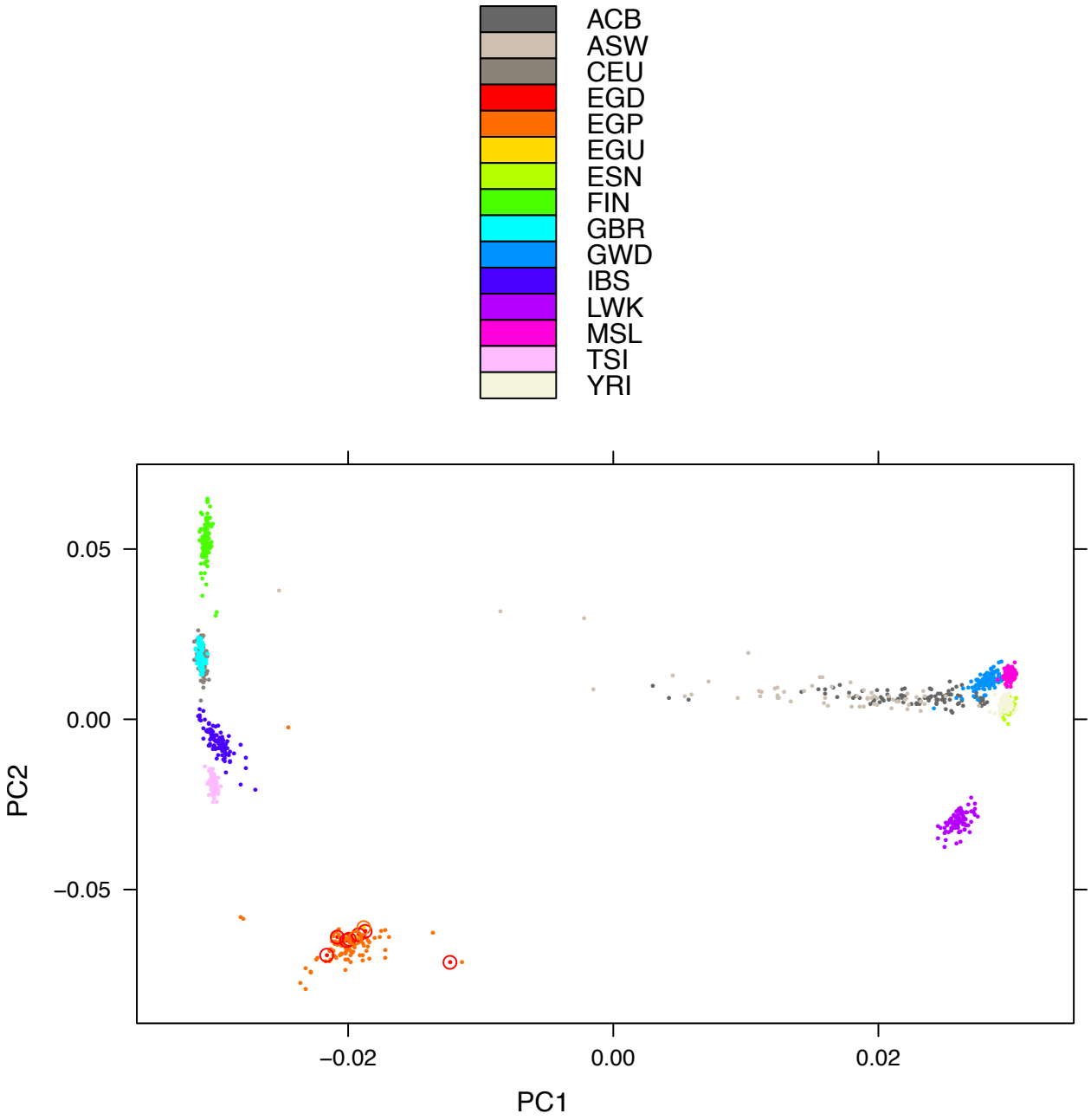
Supplementary Figure 17: Number of deletions and boxplots of deletion lengths per individual after collapsing SVs.



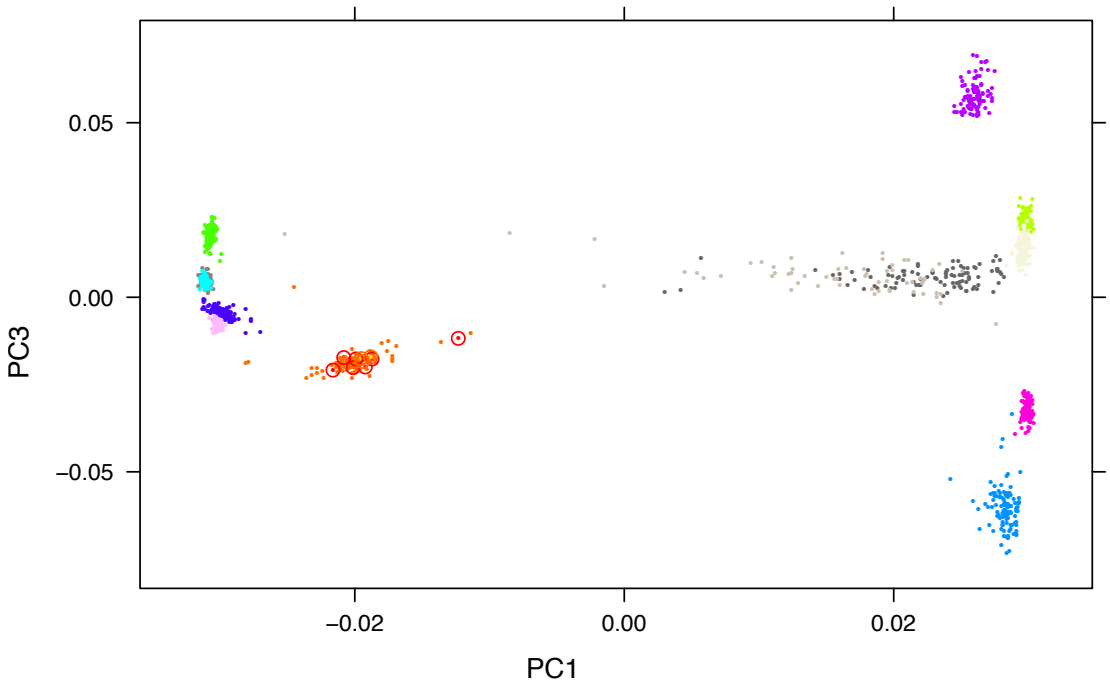
Supplementary Figure 18: Number of inversions and boxplots of inversion lengths per individual after collapsing SVs.



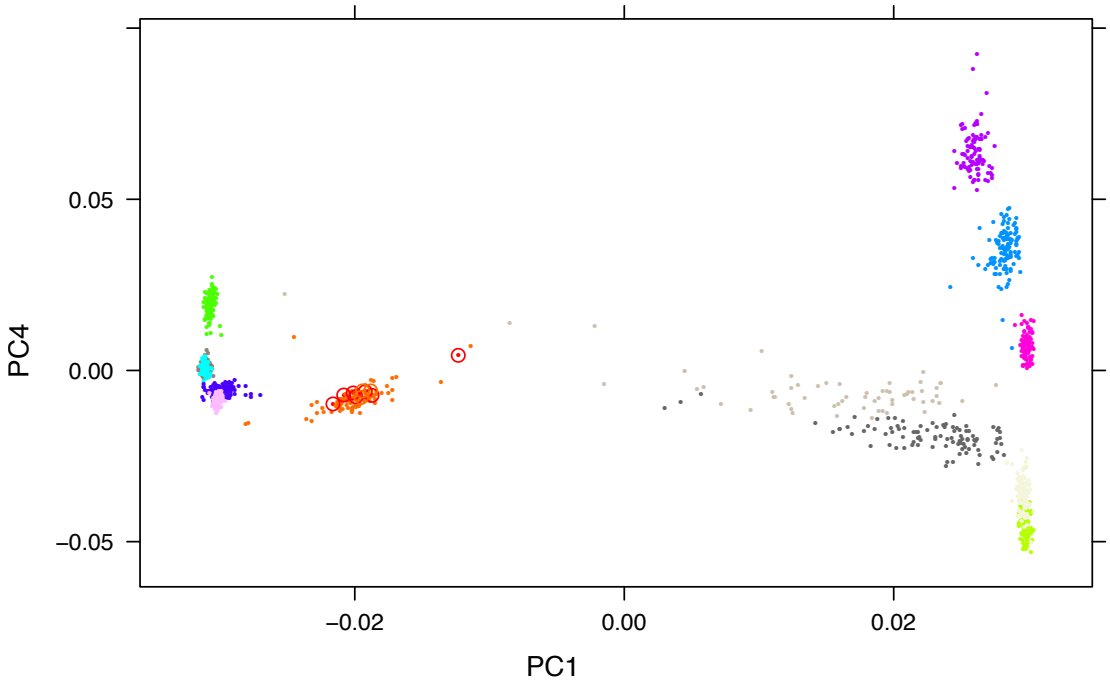
Supplementary Figure 19: Number of duplications and boxplots of duplication lengths per individual after collapsing SVs.



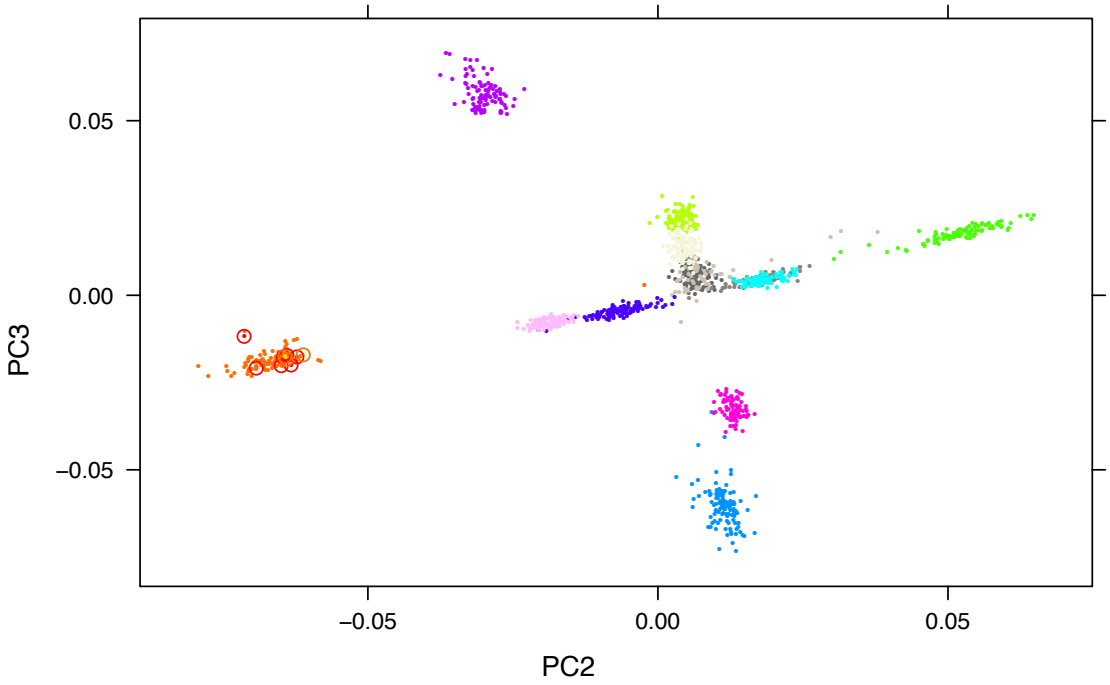
Supplementary Figure 20: Genotype principal components 1 versus 2. For population codes see Suppl. Table 8. EGD: Egyptian - Nile Delta; EGU: Egyptian - Upper Egypt; EGP: Egyptian from Pagani *et al.*



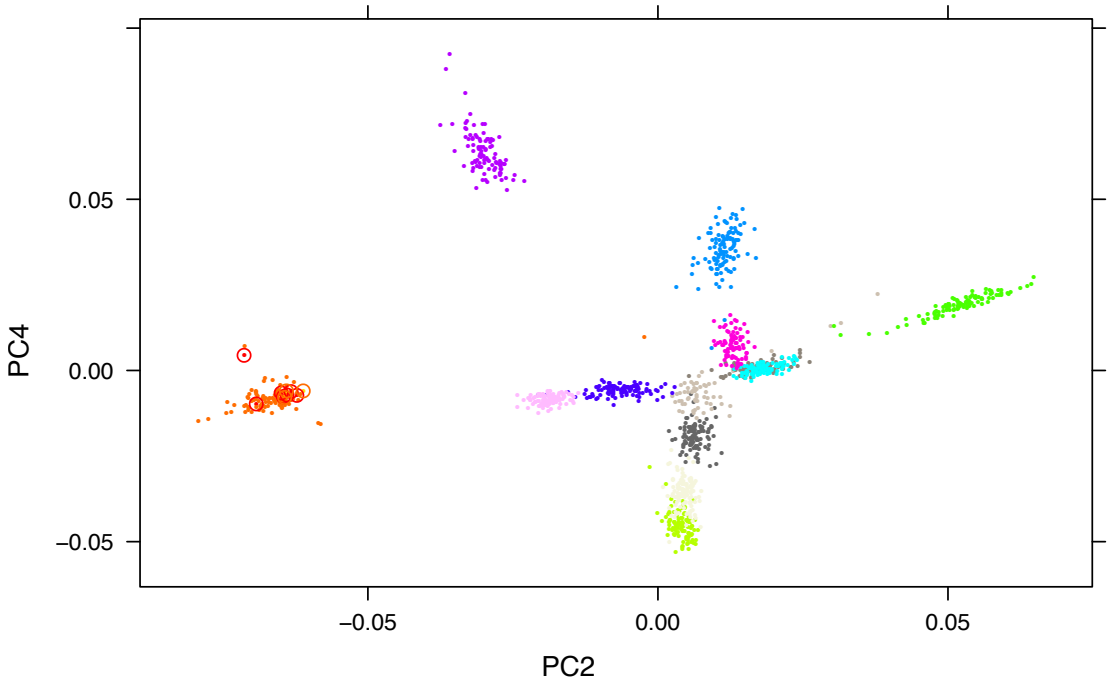
Supplementary Figure 21: Genotype principal components 1 versus 3. For population codes see Suppl. Table 8. For color code see Fig. 20.



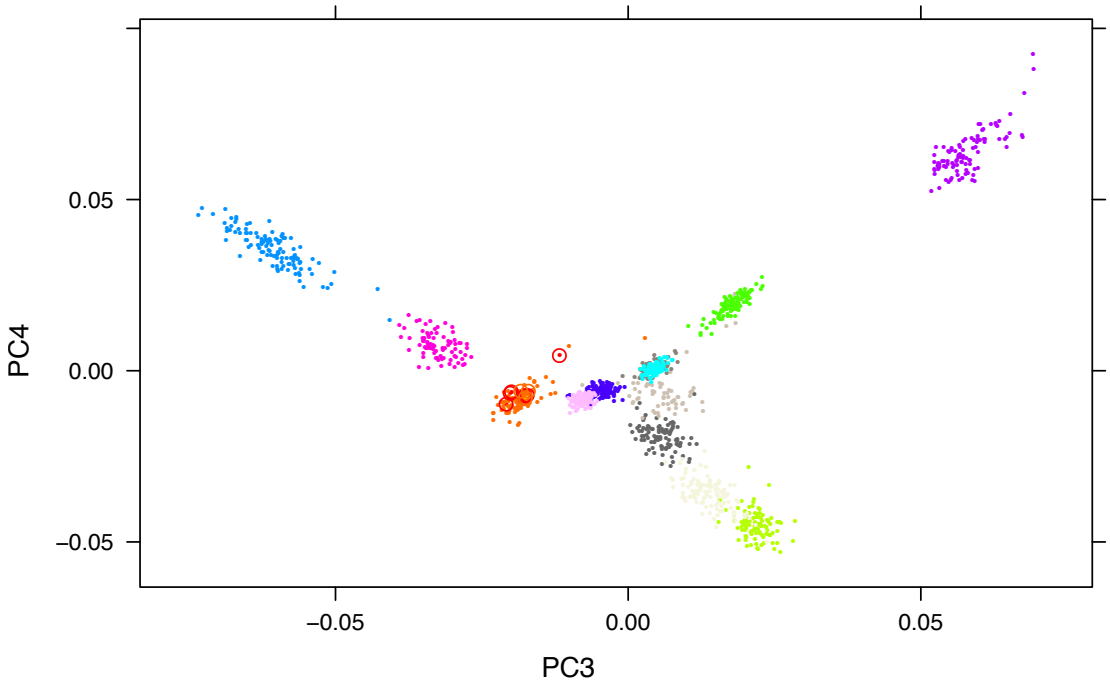
Supplementary Figure 22: Genotype principal components 1 versus 4. For population codes see Suppl. Table 8. For color code see Fig. 20



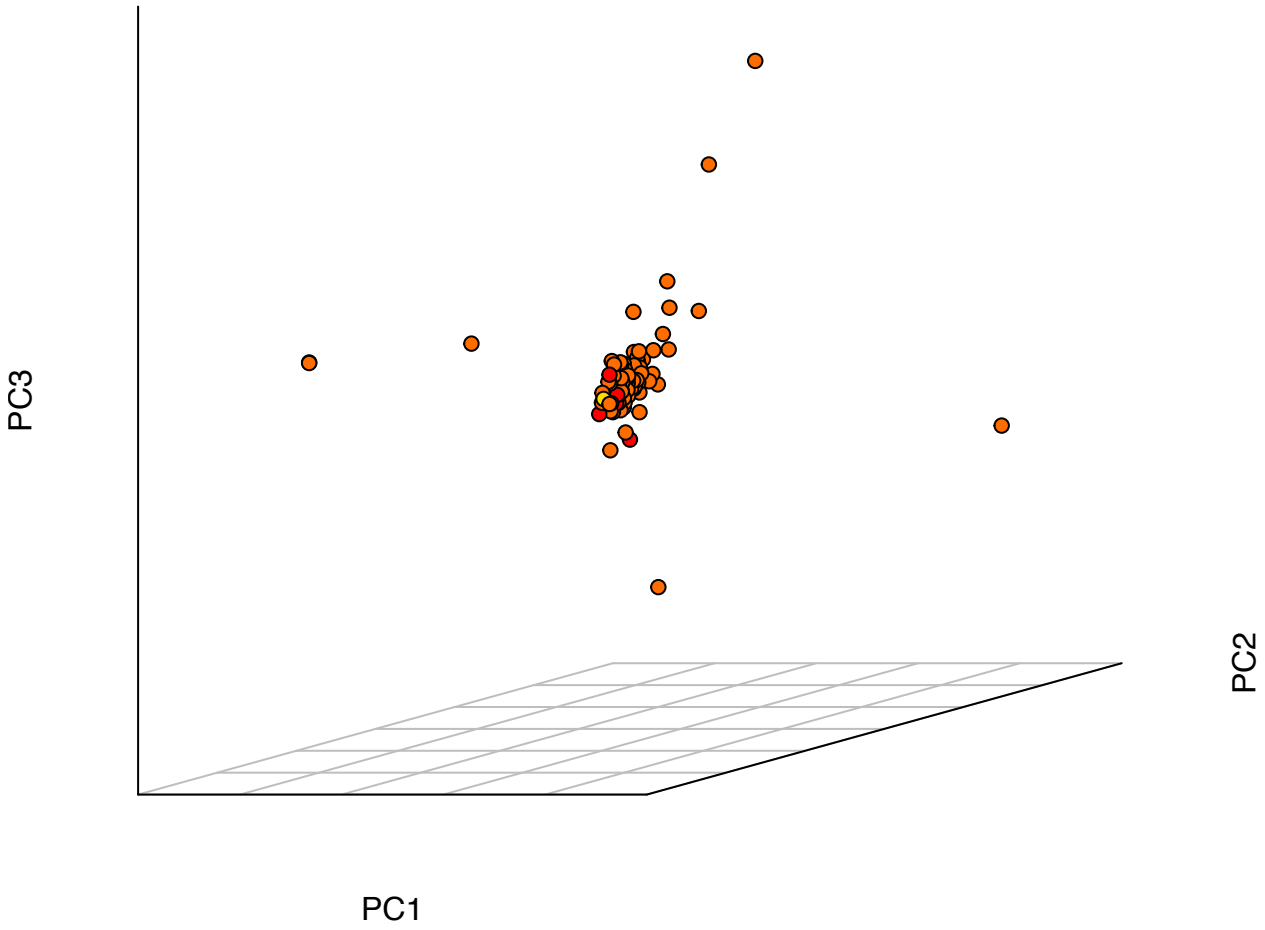
Supplementary Figure 23: Genotype principal components 2 versus 3. For population codes see Suppl. Table 8. For color code see Fig. 20



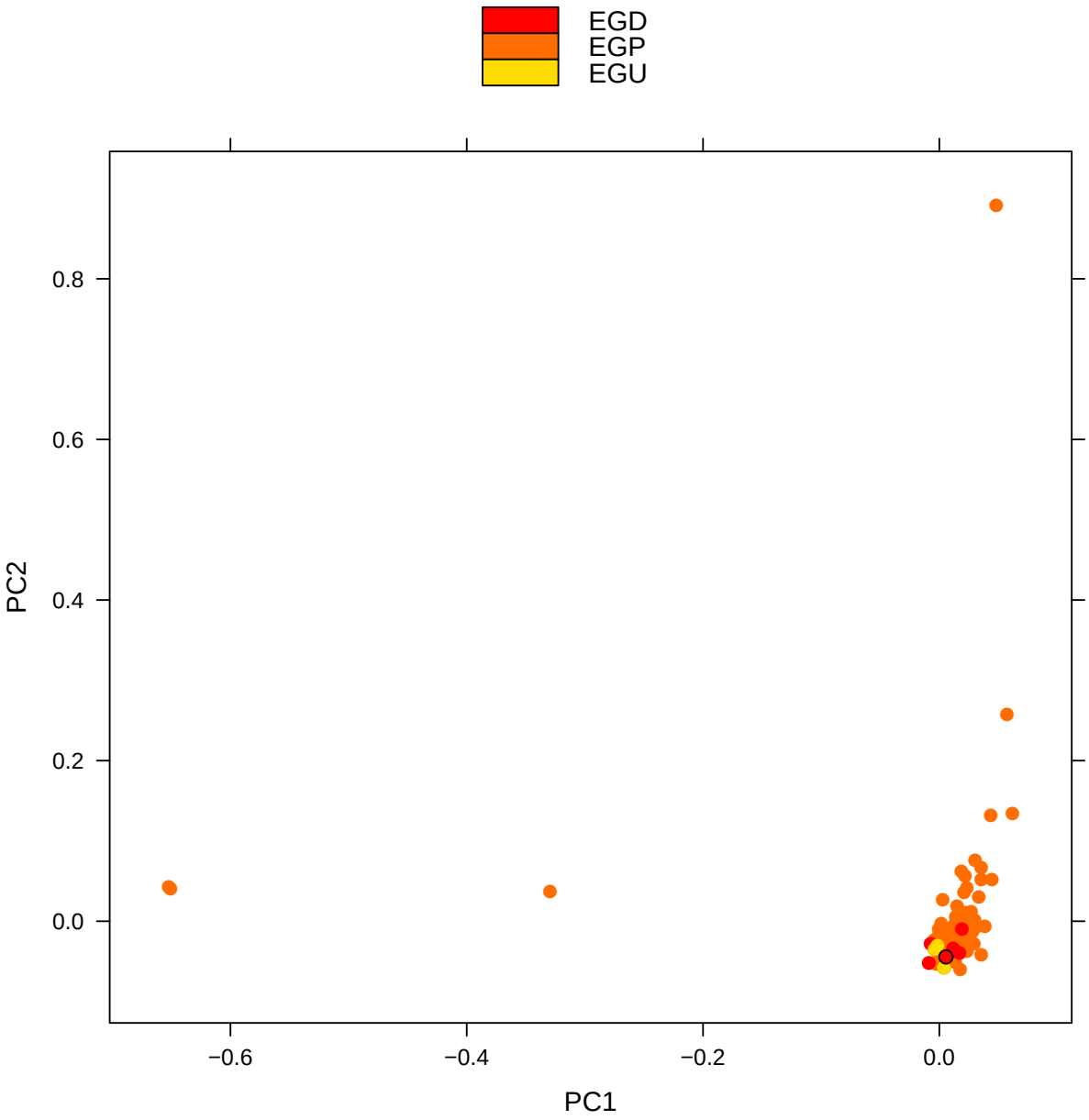
Supplementary Figure 24: Genotype principal components 2 versus 4. For population codes see Suppl. Table 8. For color code see Fig. 20



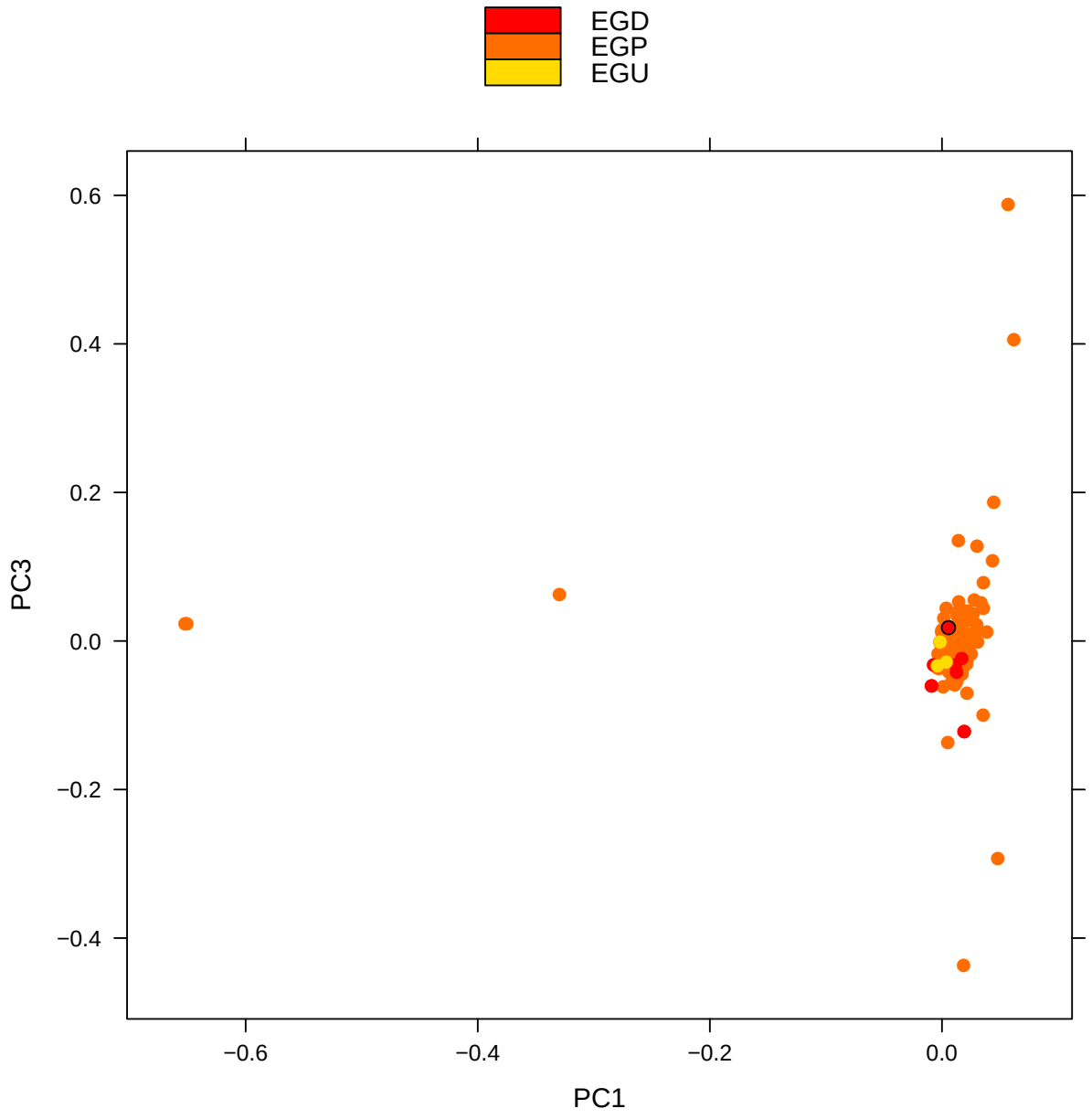
Supplementary Figure 25: Genotype principal components 3 versus 4. For population codes see Suppl. Table 8. For color code see Fig. 20



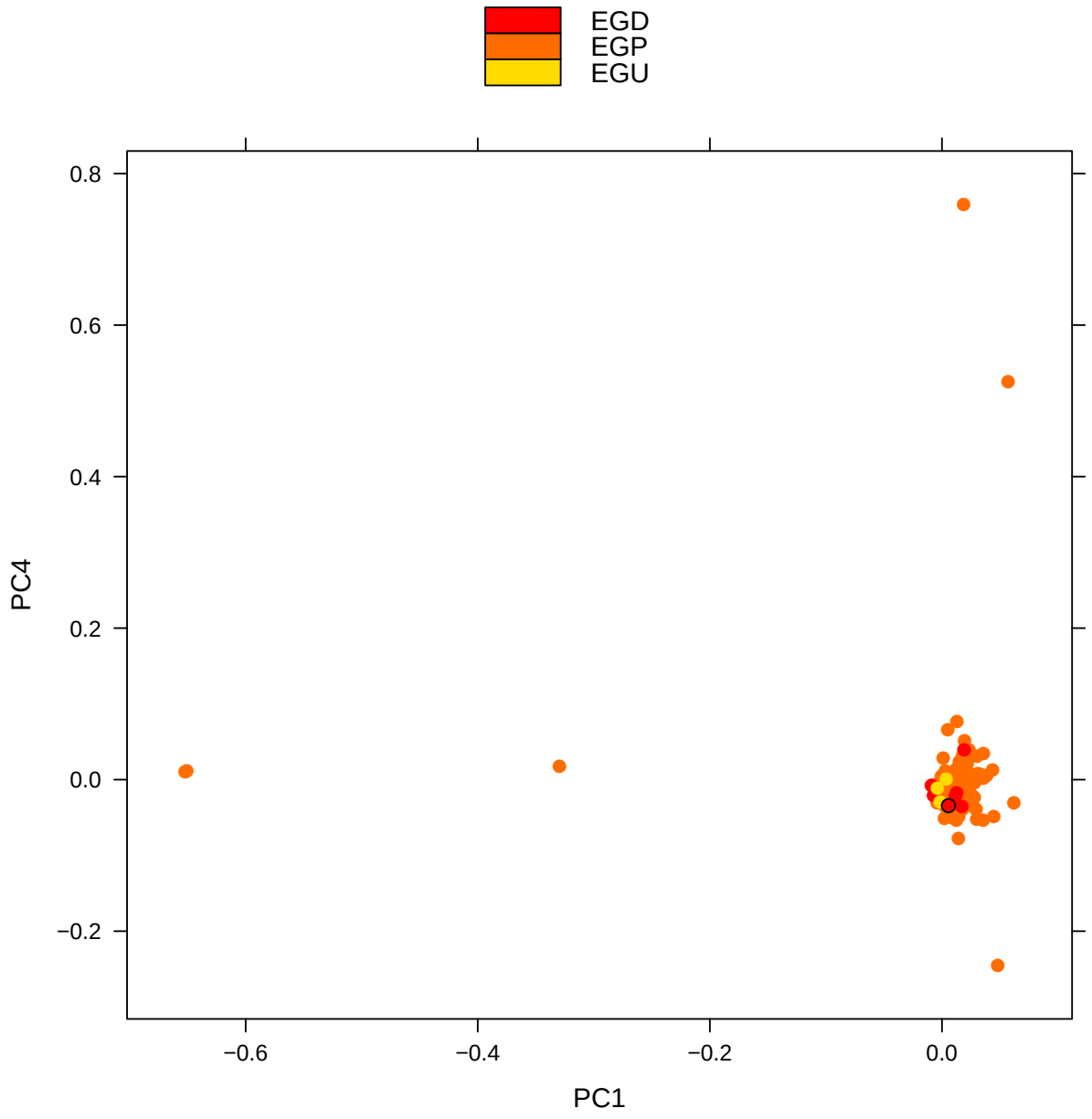
Supplementary Figure 26: See legend of Fig. 27 Genotype principal components 1, 2 and 3 from Egyptian-only PCA. Red: Egyptian - Nile Delta; Yellow: Egyptian - Upper Egypt; Orange: Egyptian from Pagani *et al.*



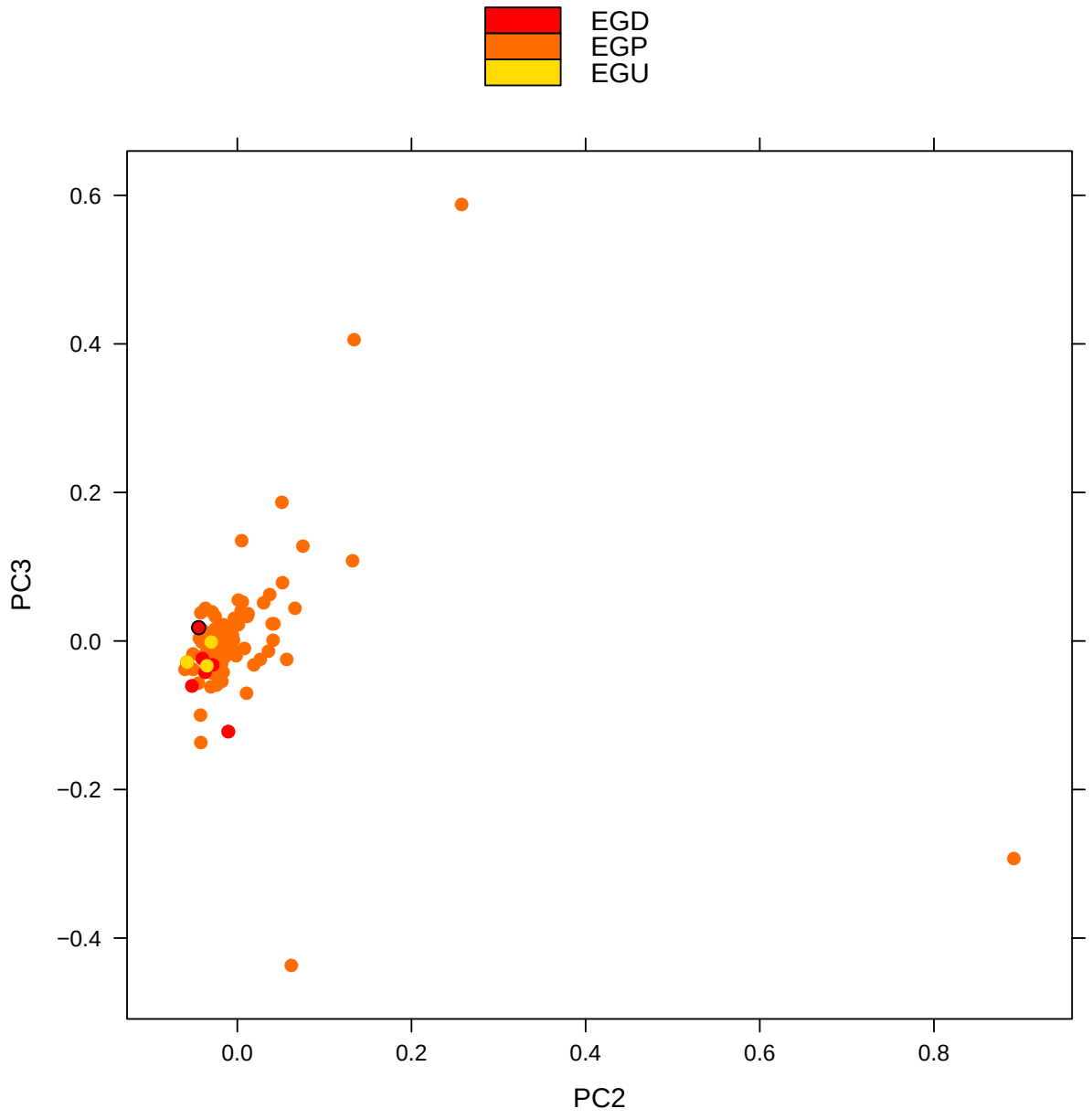
Supplementary Figure 27: Genotype principal components 1 versus 2 from Egyptian-only PCA. EGD: Egyptian - Nile Delta; EGU: Egyptian - Upper Egypt; EGP: Egyptian from Pagani *et al.* A black circle denotes the assembly individual EGYPT.



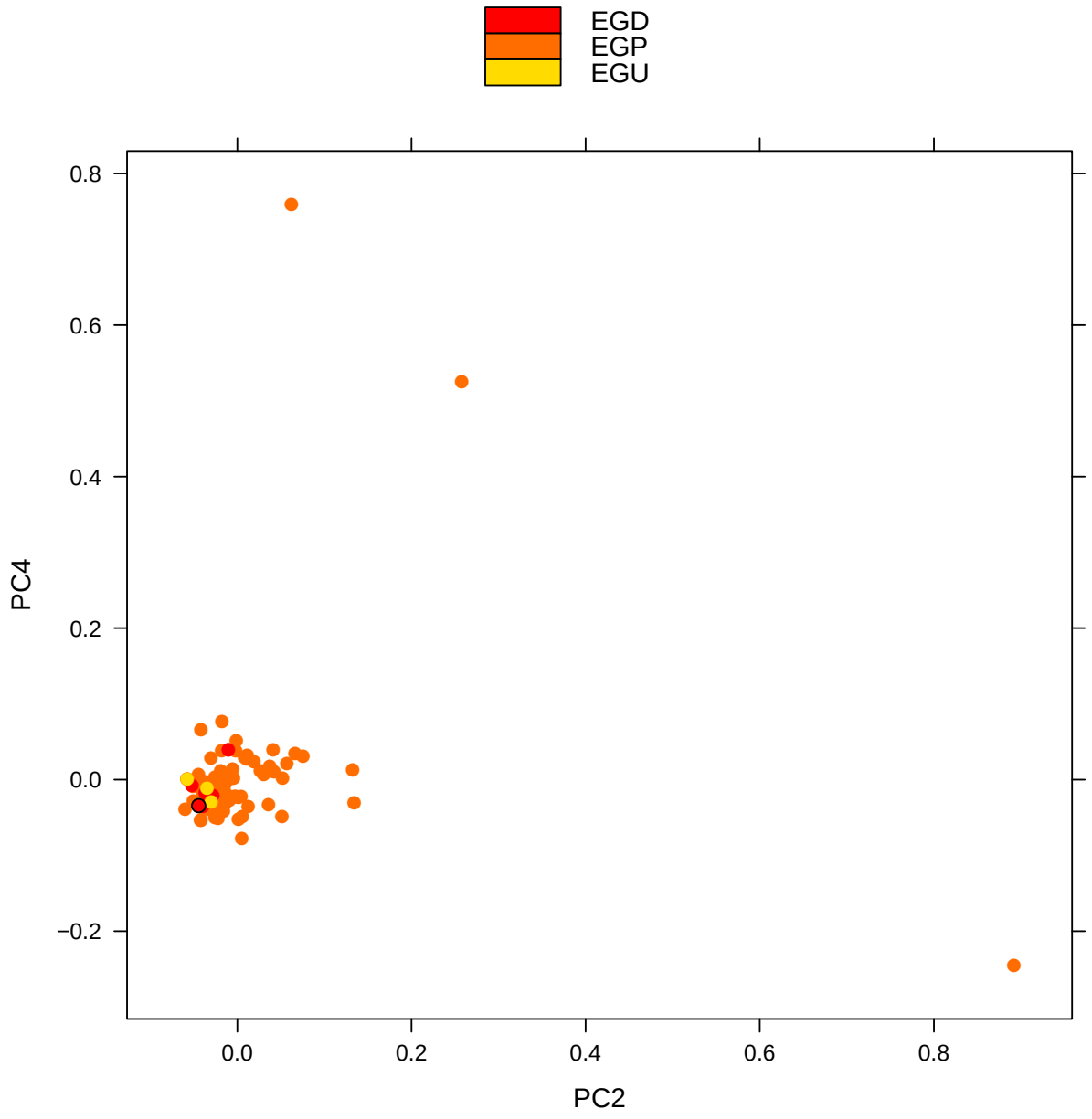
Supplementary Figure 28: Genotype principal components 1 versus 3 from Egyptian-only PCA. For label description see Fig. 27



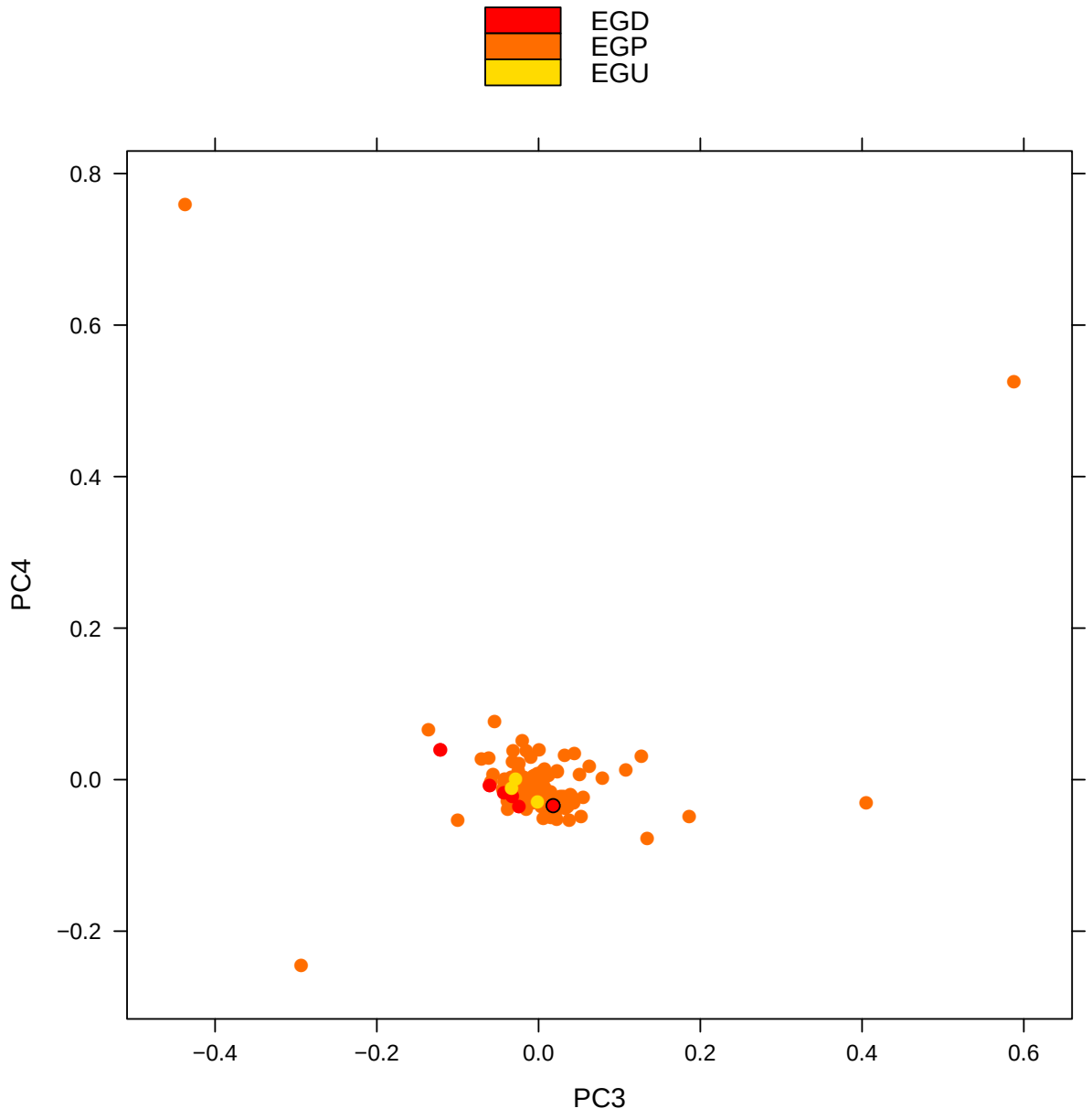
Supplementary Figure 29: Genotype principal components 1 versus 4 from Egyptian-only PCA. For label description see Fig. 27



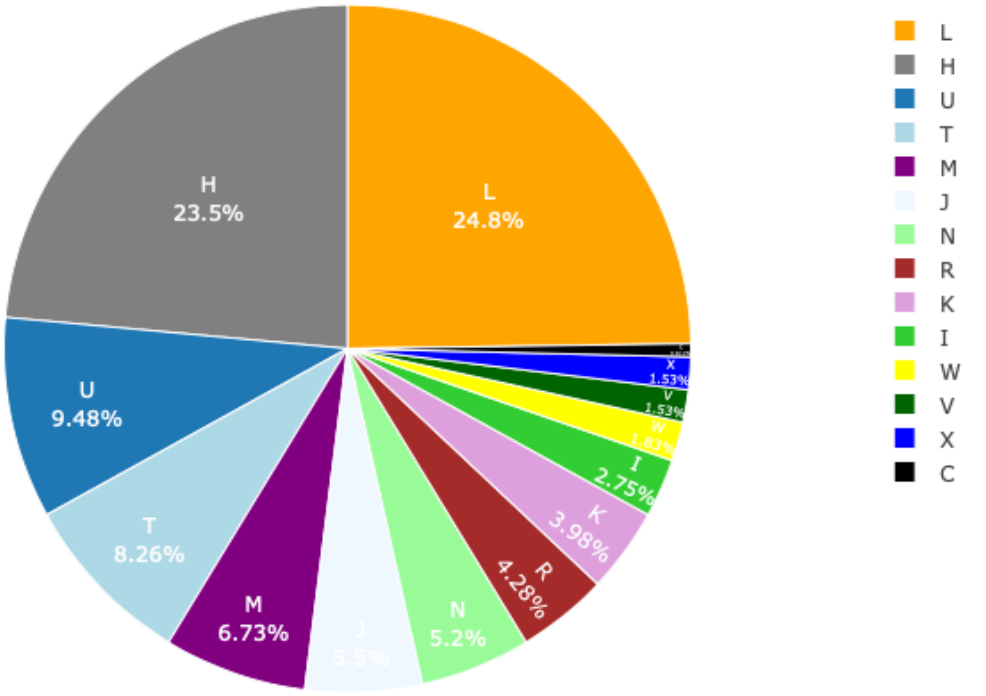
Supplementary Figure 30: Genotype principal components 2 versus 3 from Egyptian-only PCA. For label description see Fig. 27



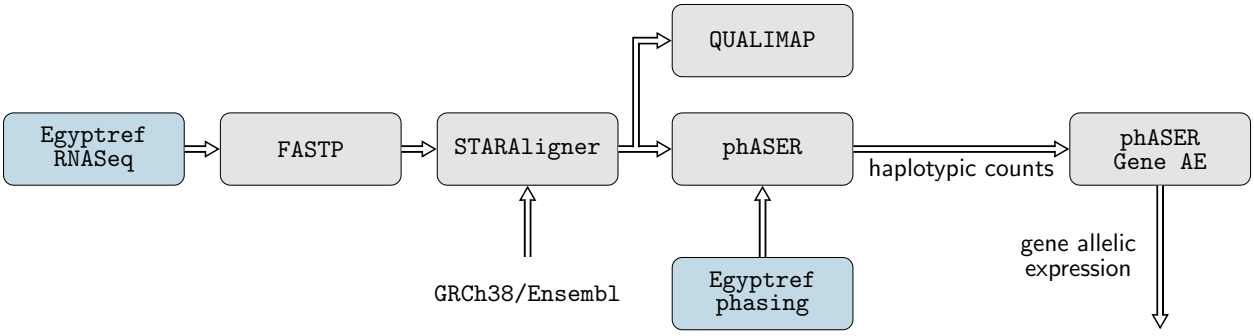
Supplementary Figure 31: Genotype principal components 2 versus 4 from Egyptian-only PCA. For label description see Fig. 27



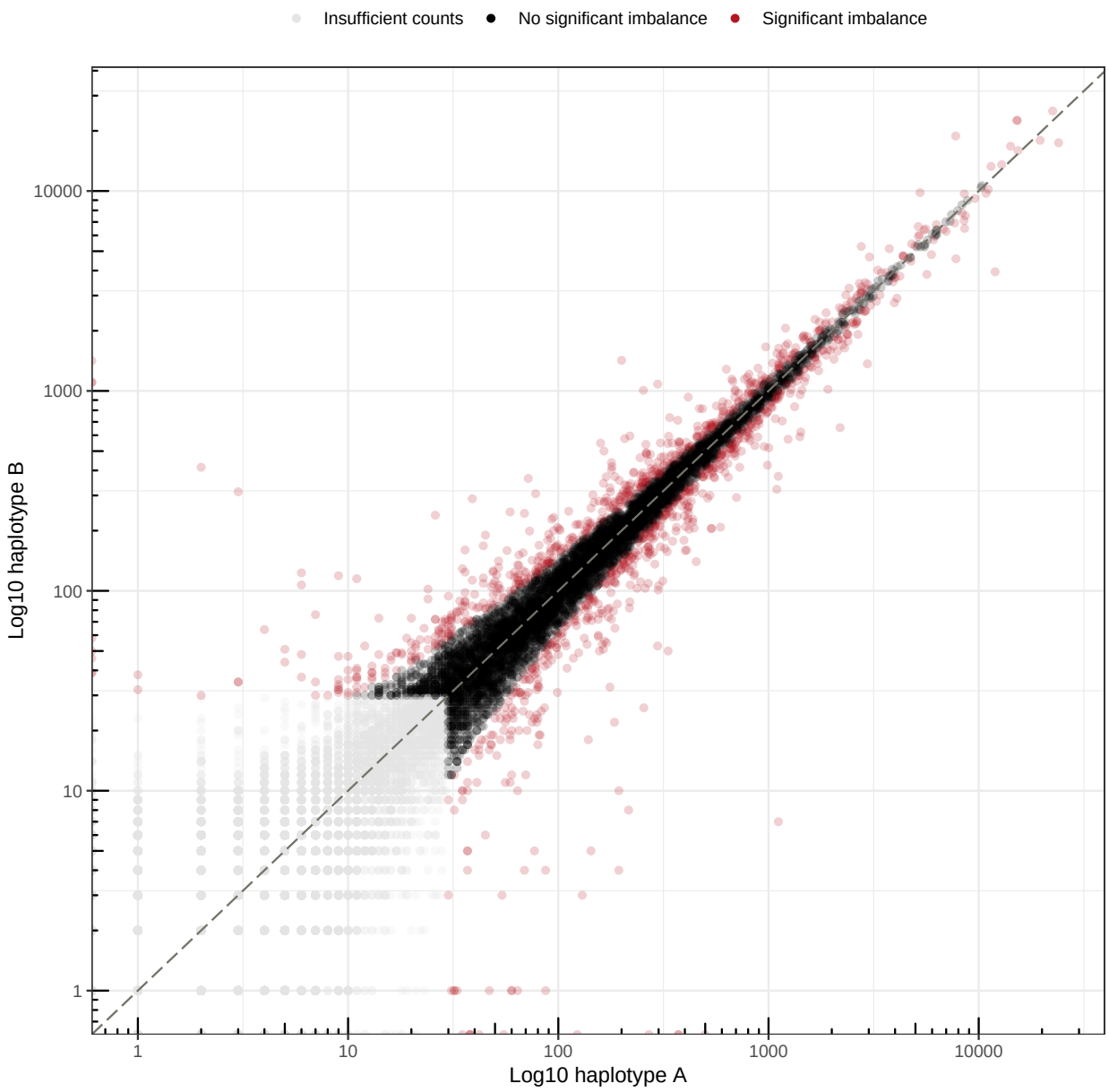
Supplementary Figure 32: Genotype principal components 3 versus 4 from Egyptian-only PCA. For label description see Fig. 27



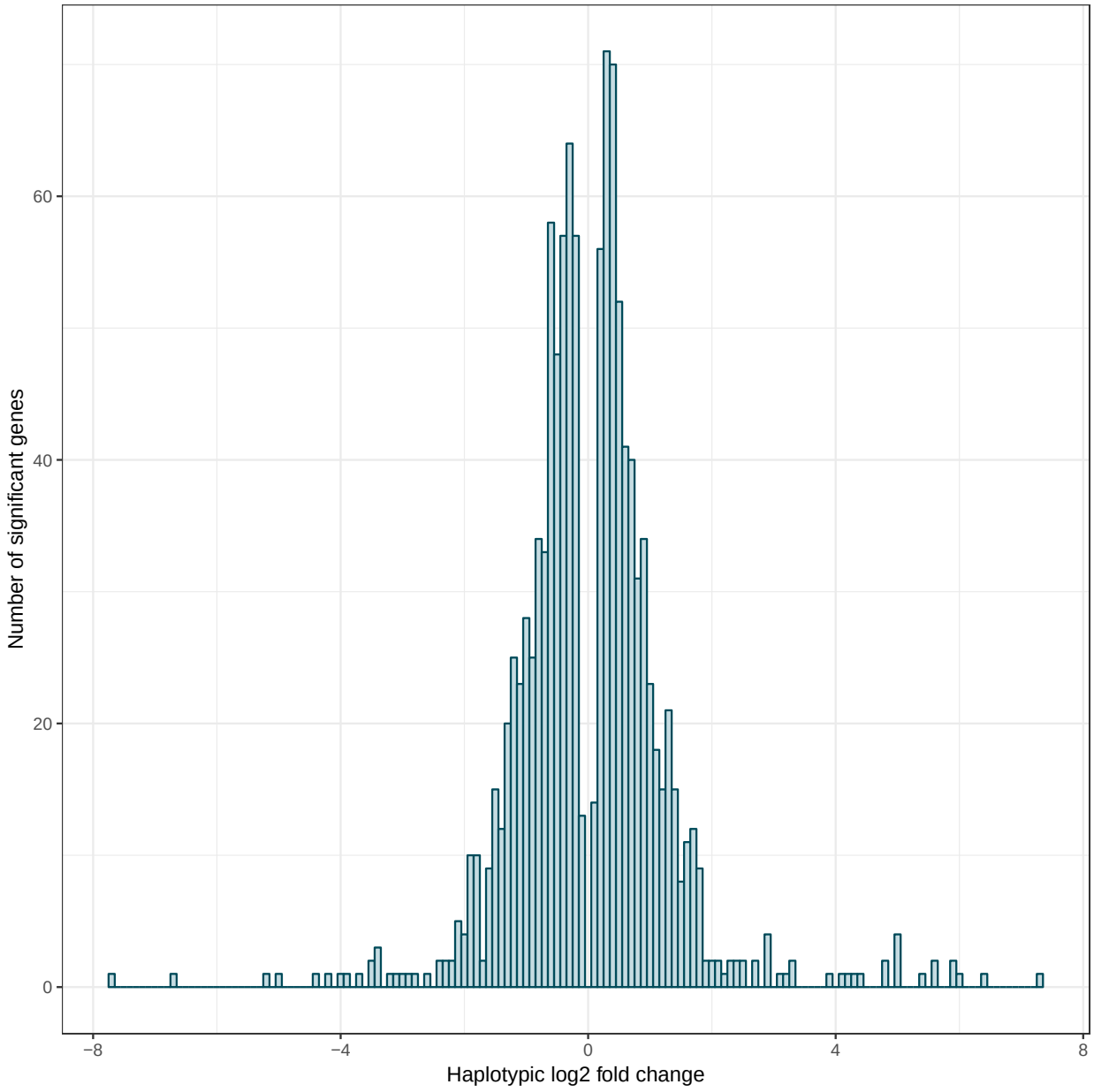
Supplementary Figure 33: Pie chart of mitochondrial haplogroups of 327 Egyptian individuals.



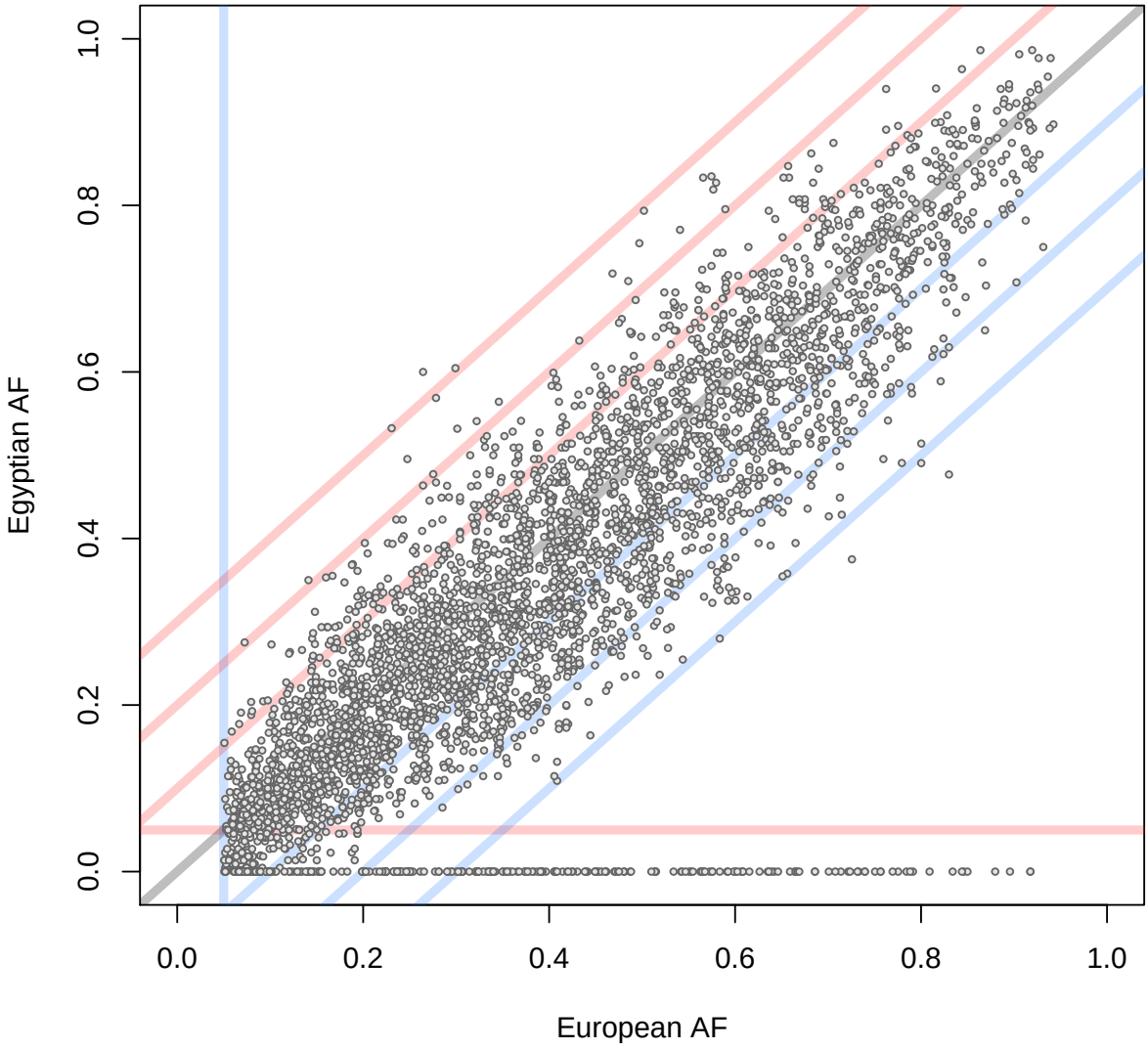
Supplementary Figure 34: Overview of haplotypic expression analysis using PHASER.



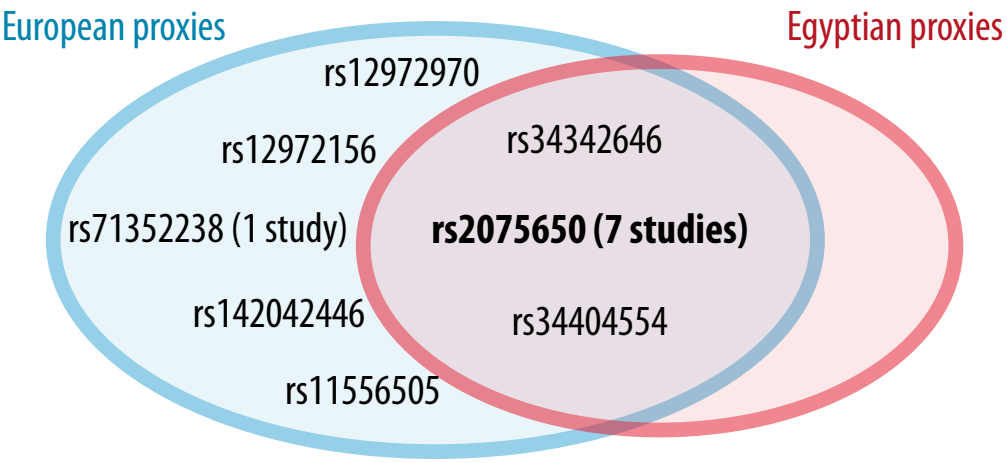
Supplementary Figure 35: Haplotypic counts for 16,566 genes. Genes with insufficient number of reads are displayed light gray and have not been used in allelic expression analysis. Significant genes are displayed red.



Supplementary Figure 36: Histogram of haplotypic fold changes for 1,180 genes with significant haplotypic expression.



Supplementary Figure 37: Scatterplot of Egyptian AF versus European AF for GWAS tag SNPs which have genotypes for more than 100 Egyptian individuals. The blue vertical line denotes 5% European MAF. The red horizontal line denotes 5% Egyptian MAF. Diagonal lines denote 10, 20 and 30% AF difference. Note that there is a number of variants that are not present in the Egyptian data, i.e. have AF of 0%.



Supplementary Figure 38: Proxy SNP comparison for an Alzheimer's disease locus sometimes attributed to gene TOMM40. In seven Alzheimer's diesease GWAS, SNP rs207650 has been reported as tag SNP according to GWAS catalog. For this SNP, there are two shared proxy variants, which are in LD with rs207650 in Europeans as well as Egyptians ($R^2 \geq 0.8$). Further, there are five European-only proxies with Alzheimer's disease tag SNP rs207650. One variant, rs71352238, has also been reported in a study, illustrating that results of GWAS performed with European individuals may not be transferred to the Egyptian population because of LD differences.

Supplementary References

[1] Jue Ruan and Heng Li. Fast and accurate long-read assembly with wtdbg2. *bioRxiv*, page 530972, January 2019.

[2] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics (Oxford, England)*, 34(18):3094–3100, 2018.

[3] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–1760, July 2009.

[4] Bruce J. Walker, Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A. Cuomo, Qiandong Zeng, Jennifer Wortman, Sarah K. Young, and Ashlee M. Earl. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS One*, 9(11):e112963, 2014.

[5] Jeong-Sun Seo, Arang Rhie, Junsoo Kim, Sangjin Lee, Min-Hwan Sohn, Chang-Uk Kim, Alex Hastie, Han Cao, Ji-Young Yun, Jihye Kim, Junho Kuk, Gun Hwa Park, Juhyeok Kim, Hanna Ryu, Jongbum Kim, Mira Roh, Jeonghun Baek, Michael W. Hunkapiller, Jonas Korlach, Jong-Yeon Shin, and Changhoon Kim. De novo assembly and phasing of a Korean human genome. *Nature*, 538(7624):243–247, October 2016.

[6] Alla Mikheenko, Andrey Prjibelski, Vladislav Saveliev, Dmitry Antipov, and Alexey Gurevich. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics (Oxford, England)*, 34(13):i142–i150, 2018.

[7] Alla Mikheenko, Gleb Valin, Andrey Prjibelski, Vladislav Saveliev, and Alexey Gurevich. Icarus: visualizer for de novo assembly evaluation. *Bioinformatics (Oxford, England)*, 32(21):3321–3323, 2016.

[8] Kai Wang, Mingyao Li, and Hakon Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16):e164, September 2010.

[9] William McLaren, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The Ensembl Variant Effect Predictor. *Genome Biology*, 17(1):122, 2016.

[10] Philipp Rentzsch, Daniela Witten, Gregory M. Cooper, Jay Shendure, and Martin Kircher. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, 47(D1):D886–D894, January 2019.

[11] Ivan Adzhubei, Daniel M. Jordan, and Shamil R. Sunyaev. Predicting functional effect of human missense mutations using PolyPhen-2. *Current Protocols in Human Genetics*, Chapter 7:Unit7.20, January 2013.

[12] Robert Vaser, Swarnaseetha Adusumalli, Sim Ngak Leng, Mile Sikic, and Pauline C. Ng. SIFT missense predictions for genomes. *Nature Protocols*, 11(1):1–9, January 2016.

[13] Tobias Rausch, Thomas Zichner, Andreas Schlattl, Adrian M. Stütz, Vladimir Benes, and Jan O. Korbel. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics (Oxford, England)*, 28(18):i333–i339, September 2012.

[14] Nick Patterson, Alkes L. Price, and David Reich. Population structure and eigenanalysis. *PLoS genetics*, 2(12):e190, December 2006.

[15] Christopher C. Chang, Carson C. Chow, Laurent Cam Tellier, Shashaank Vattikuti, Shaun M. Purcell, and James J. Lee. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4:7, 2015.

[16] Gad Abraham and Michael Inouye. Fast Principal Component Analysis of Large-Scale Genome-Wide Data. *PLoS ONE*, 9(4), April 2014.

[17] Anita Kloss-Brandstätter, Dominic Pacher, Sebastian Schönherr, Hansi Weissensteiner, Robert Binna, Günther Specht, and Florian Kronenberg. HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Human Mutation*, 32(1):25–32, January 2011.

[18] Luca Pagani, Stephan Schiffels, Deepti Gurdasani, Petr Danecek, Aylwyn Scally, Yuan Chen, Yali Xue, Marc Haber, Rosemary Ekong, Tamiru Oljira, Ephrem Mekonnen, Donata Luiselli, Neil Bradman, Endashaw Bekele, Pierre Zalloua, Richard Durbin, Toomas Kivisild, and Chris Tyler-Smith. Tracing the route of modern humans out of Africa by using 225 human genome sequences from Ethiopians and Egyptians. *American Journal of Human Genetics*, 96(6):986–991, June 2015.

[19] Stephane E. Castel, Pejman Mohammadi, Wendy K. Chung, Yufeng Shen, and Tuuli Lappalainen. Rare variant phasing and haplotypic expression from RNA sequencing with phASER. *Nature Communications*, 7:12817, 2016.

[20] Shifu Chen, Yanqing Zhou, Yaru Chen, and Jia Gu. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics (Oxford, England)*, 34(17):i884–i890, 2018.

[21] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, 29(1):15–21, January 2013.

[22] Konstantin Okonechnikov, Ana Conesa, and Fernando García-Alcalde. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics (Oxford, England)*, 32(2):292–294, January 2016.

[23] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–2079, August 2009.

[24] Guillaume Marçais, Arthur L. Delcher, Adam M. Phillippy, Rachel Coston, Steven L. Salzberg, and Aleksey Zimin. MUMmer4: A fast and versatile genome alignment system. *PLoS computational biology*, 14(1):e1005944, 2018.

[25] Ksenia Khelik, Karin Lagesen, Geir Kjetil Sandve, Torbjørn Rognes, and Alexander Johan Nederbragt. NucDiff: in-depth characterization and annotation of differences between two sets of DNA sequences. *BMC bioinformatics*, 18(1):338, July 2017.

[26] James T. Robinson, Helga Thorvaldsdóttir, Aaron M. Wenger, Ahmet Zehir, and Jill P. Mesirov. Variant Review with the Integrative Genomics Viewer. *Cancer Research*, 77(21):e31–e34, 2017.