Characterization of Arabian Peninsula whole exomes: Contributing to the catalogue of human diversity

Highlights
- Characterization of 90 Arabian WES enriched for UTRs identified 17,000 new variants
- The AP WES have a high burden in potentially pathogenic variants
- These variants were in 764 genes associated with neurologic and congenital diseases
- This burden was significantly and positively correlated with the consanguinity level
Characterization of Arabian Peninsula whole exomes: Contributing to the catalogue of human diversity

Joana C. Ferreira,1,2,3 Farida Alshamali,4 Luisa Pereira,1,2 and Veronica Fernandes1,2,5,*

SUMMARY

The cataloging of human genomic variation is a challenging task given the size and diversity of the global human population. Several geographic regions remain under-characterized, such as the Arabian Peninsula (AP), which is unique in being at the cross-roads between Africa, Europe and Asia, and hence is a continuous hot-spot for admixture, counteracted by the worldwide highest levels of consanguinity. We conducted whole exome sequencing enriched for untranslated regions (WES + UTRs) on 90 Arabians, and identified a considerable amount of new variants (~17,000 out of ~145,000). By applying pathogenic predicting tools, we demonstrated that AP WES have a high burden in potentially deleterious variants, especially in nonsynonymous and UTR variants, and that these are located in genes associated with neurologic diseases and congenital malformations. This burden was significantly and positively correlated with the consanguinity level. These results testify the importance of surveying consanguineous populations where pathogenic variants are not efficiently eliminated by genetic drift.

INTRODUCTION

The technological developments introduced with the next-generation sequencing methodology are allowing the high throughput characterization of millions of polymorphisms, enhancing the knowledge on genetic diversity between and within populations, and enabling new lens to explore the genetic basis of diseases (DePristo et al., 2011; Jeroncic et al., 2016). Directed sequencing of all exons (known as whole exome sequencing, WES) is being increasingly used in clinical genetics (Steyaert et al., 2018) because these regions are probable locations of candidate alleles conferring susceptibility to monogenic and complex diseases. International consortia have been collecting big exome data and making them publicly available for mining of variants (International HapMap Consortium, 2003; Cox, 2018), such as the Exome Sequencing Project (ESP) Exome Variant Server (EVS) (Fu et al., 2013) representing more than 200,000 individuals from multiple ESP cohorts, and the Exome Aggregation Consortium (ExAC) database (Lek et al., 2016) spanning 60,706 unrelated individuals sequenced as part of various disease-specific and population genetic studies. The ExAC creators launched afterward the Genome Aggregation Database (gnomAD) (Karczewski et al., 2020). This big data is allowing us to predict the functional effects of variants on diverse traits in different populations, and guiding inference of disease risk at an individual basis. However, WES catalogs are still extremely biased in terms of ancestry, being largely based on data from European and East Asian populations (Buniello et al., 2019; International HapMap Consortium, 2003), which limits the transferability of findings to other population groups (Pereira et al., 2021).

A region particularly understudied in large-scale sequencing projects is the Middle East and more specifically the Arabian Peninsula (AP), with the exception of Qatar (Fakhro et al., 2016; Razali et al., 2021). AP populations are amongst the most non-African ancestral populations in the globe, as this region was the first outpost of the successful out-of-Africa (OOA) migration, at around 60,000 ago (ka) (Fernandes et al., 2012; Soares et al., 2012). Concordantly, signatures of ancient ancestry were detected in extant AP populations, especially in the Arabo-Persian Gulf region: (1) as relic maternal lineages from the west-Eurasian haplogroups N1, N2 and X, branching directly from the root of macro-haplogroup N (Fernandes et al., 2012); (2) in the autosomal genome, the component known as basal Eurasian, identifiable when ancient DNA information is integrated in the analysis (Ferreira et al., 2021). There were clear continuaus
throughout time linking western Arabia with the Levant and Africa, and eastern Arabia with Iran and the Caucasus, making AP an important migration nexus between the main human population groups (Fernandes et al., 2015, 2019). This high admixture enriched the Arabian genomes with non-autochthonous selected variants, mainly for malaria protection (higher in the west), lactose tolerance (European/South Asian-derived allele in the eastern AP) and other immune system defences (throughout the Peninsula) (Almarri et al., 2021; Fernandes et al., 2019).

Despite the high ancestry admixture in AP, Arabs have traditionally a high level of consanguinity, reaching an overall prevalence of 56% in Saudi Arabia, from which 33.6% are first-cousin consanguineous marriages (Almazroua et al., 2020). In addition to a high consanguinity rate, Arabs are also characterized by large family sizes and advanced paternal and/or maternal age (Al-Gazali et al., 2006), increasing the risk of congenital anomalies in their offspring (Almazroua et al., 2020). These factors are associated with a high number of congenital and genetic disorders (Al-Gazali et al., 2006), such as impaired hearing (3.5 times higher in consanguineous than in non-consanguineous mating) (Almazroua et al., 2020) and Down syndrome (Arab countries exceed the 1.2–1.7 per 1000 typical for industrialized countries; (AbdulAzeez et al., 2019; Al-Gazali et al., 2006)). Other complex disorders with a genetic component are also common throughout the Arab world, including haemoglobinopathies, glucose-6-phosphate dehydrogenase deficiency and metabolic diseases (obesity, type 2 diabetes and dyslipidemia), and all have been associated with the high level of consanguinity in this region (Almazroua et al., 2020). The reason for this is that high level of consanguinity increases frequencies of rare variants and extends stretches of homozygous chromosomal fragments (long runs of homozygosity across all sites, ROH) (Kirin et al., 2010). Thus, the study of ROH length and burden of deleterious variation can provide important insights into the human demographic history and clinical applications (Ceballos et al., 2018; Szpiech et al., 2013). Most of the WES available for Arabs were obtained in patient cohorts (Almaghlouth et al., 2021; Monies et al., 2019; Project Team, 2015; Wahabi et al., 2017). Fewer WES are available for the general population: one obtained in Qatari populations identified eight hematologic variants, five metabolic, four eye-related, three inflammatory, three cardiovascular and three neurologic as the most common disease-related variants in those consanguineous cohorts (Rodriguez-Flores et al., 2014); an important cohort, the Greater Middle East (GME) Variome Project containing 2,497 individuals from 19 Arab and non-Arab Muslim countries (Scott et al., 2016) detected large and rare homozygous blocks, compatible with recent consanguineous matings, rendering easy to identify genes harboring putatively high-impact homozygous variants. More recently, Razali et al. (2021) (Razali et al., 2021) made available 6,218 whole genomes from Qatar, attempting to build a reference panel for genotype imputation of Arabs and related ethnicities. The authors saw longer ROH fragments (>80 MB) in Qataris than previously published for exome/array data, and many of those ROHs contained disease genes, with some Qataris showing more than 4000 ROH-overlapping disease genes.

Because the effective size of the WES catalog for AP general population remains low, in this work we performed a characterization of Arabian WES by randomly selecting 25 individuals from each AP country (Saudi Arabia, Yemen, United Arab Emirates, and Oman). We enriched these WES with the sequencing of the untranslated regions 5 and 3 (UTR5 and UTR3). In fact, it is unfortunate that the UTRs are usually not included in the WES panels, as those regions play important functional roles (Steri et al., 2018), in the modulation of mRNA transcription, secondary structure, stability, localization, translation, and access to regulators like microRNAs and RNA-binding proteins. We aimed to identify new and rare variants (especially in UTRs), to infer their possible functional impact, and to relate them with the consanguinity levels of the AP populations.

RESULTS

Diversity of the AP WES (enriched for UTRs)

The 90 AP WES presented 145,630 variants (Table S1), around half the level of diversity displayed by the European (270,913 variants) and East Asian (254,527 variants) regions, and one-third by sub-Saharan Africans (443,767 variants) (1000 Genomes populations, for an equal number of individuals and matching genomic segments). The transition-to-transversion ratio (Ti/Tv) was 2.59 for the entire dataset, and 3.23 when considering only the exonic variants as this latter value is more comparable with other high quality WES datasets not enriched for UTRs (DePristo et al., 2011; Jeroncic et al., 2016).

Of the total 145,630 variants, 17,060 (11.6%) were new when compared against several public databases (Figure 1A). Among the known variants (Figure 1B; Table S1), identical proportions (around a quarter)
were shared between nonsynonymous, intronic and UTR3 classes, followed by synonymous (~16%) class, and rare proportions of all remaining classes (nonsense, UTRS, ncRNA, upstream/downstream, intergenic and splicing). Of interest, for the new variants (Figure 1C), almost half of these were located in UTR3, followed still by 29% of intronic, ~15% of synonymous and rare instances of the other classes of new variants. It is not surprising the high proportion of new variants in UTR3, and that the value of new variants in UTR5 is also twice the value for this class in known variants, as UTRs are rarely covered in WES screenings. We detected very few new nonsynonymous variants (34 - <1% of total new variants), showing that saturation for this class of variants is almost reached for populations of mainly Eurasian ancestry.

Interestingly, a substantial amount of AP variants (6866) was only shared with clinically/functionally focused databases, namely ClinVar (Landrum et al., 2014) and dbNSFP (Liu et al., 2016). A careful inspection of these variants revealed they were mostly included in the dbNSFP dataset (88.54% nonsynonymous, 6.95% nonsense and 2.19% splicing variants), which prioritizes these type of variants from UK10K and ExAC datasets. These variants are of possible functional relevance, as we will see in detail in the next section.

**Functional constraint features of AP WES**

The conservation score GERP (Figures 2A and 2B, Table S1) allowed to confirm that AP had 83.9% of variants classified either as neutral (66.2% for GERP<2) or as slightly deleterious (17.7% for 2 > GERP> 4). These variants were broadly distributed by class of variants, testifying that most of the genome of AP is functionally neutral as expected. Comparatively, in the two most deleterious categories (4 > GERP> 6 and 4 > GERP> 6) there was an increase of nonsense, nonsynonymous and splicing variants, and a decrease of UTR5 and UTR3. The results from the Phylop algorithm, another conservation score, were consistent with GERP (Figure S1). When comparing with the other worldwide regions (Figure 2C), AP displayed the higher amount of moderate and extremely deleterious variants (15.31 and 0.81% respectively, compensated by the lower amount of neutral variants), followed ex aequo by Europe (11.22 and 0.37% respectively) and East Asia (11.49 and 0.39% respectively), and then Africa (10.18 and 0.30% respectively) with the lower values.

A similar pattern was observed for the integrative scaled C-scores, but now with better resolution (Figures 2D and 2E). Nonsense variants are the most extreme in the harmful scale, preceded by the splicing variants. For nonsynonymous variants, a high percentage of them (45.89%) displayed values equal or higher...
to 20, so more prone to have functional impact. In the other extreme, variants that are of lower functional impact were the synonymous, intronic, UTRs, and the rarer ncRNA, upstream/downstream and intergenic variants. Comparing with the other geographical regions (Figure 2F), again AP had a higher proportion of

Figure 2. Results for the conservation GERP and the integrative scaled C scores (Table S1)
(A) Proportion of AP variants from each class of variants in each GERP category. The line represents the total number of variants in each score category (same meaning for B, D and E).
(B) Proportion of AP variants after normalizing by the total number of variants in each class, observed in that GERP category.
(C) Proportion of variants from each GERP category in AP and other worldwide regions. GERP values were classified into four groups: neutral (2 < GERP), slightly deleterious (2 <= GERP < 4), moderate (4 <= GERP < 6), and extremely deleterious (GERP > 6).
(D) Proportion of AP variants from each class of variants along the scaled C-scores.
(E) Proportion of AP variants after normalizing by the total number of variants in each class, along the scaled C-scores.
(F) Proportion of variants with scaled C-scores below and above 20 (benign and harmful variants, respectively) in AP and other worldwide regions.
potentially harmful variants when inferred through the scaled C-scores (10.86%) than the rest of the globe (East Asia – 8.89%; Europe – 8.52%; Africa – 7.38%).

Focusing on the predicted pathogenicity for AP nonsynonymous variants, provided by SIFT and PolyPhen metrics (Figure 3A), a proportion of 15.54% (4,673 out of 30,072 nonsynonymous variants for which values for both metrics were available) were inferred as pathogenic by SIFT and PolyPhen (“deleterious” and “probably damaging”, respectively). This proportion is higher in AP compared with the other regions of the globe (Figure 3B).

We then analyzed the diseases associated with the pathogenic burden in AP (Table S2). We began by focusing on the pathogenic nonsynonymous variants (by both SIFT and PolyPhen; with a VDA score>0.2 in DisGeNET), and identified 76 known of these variants associated mainly with congenital, neurologic and neoplasm disease (Figure 4A). If we include all of these pathogenic nonsynonymous variants (by both SIFT and PolyPhen), and focus on the genes they are located, and search now for the diseases associated with these genes, we end up with 744 genes (with a GDA score>0.2 in DisGeNET) associated with diseases of various types (Figure 4B), again mainly congenital malformation, neurological diseases and neoplasms. Only two of these inferred as pathogenic nonsynonymous variants were new, and they were located in the genes: CYFIP1, associated with congenital malformation and neurologic diseases, and ITGA10, associated with oncologic, and ear and eye diseases. Taking advantage of our AP WES having information for the UTR regions, we verified that most of the UTR variants are non-conserved, with the proportion of UTRs with C-scores above 20 being 1.76% of UTR5 and 0.31% of UTR3 variants (47 out of 2671 UTR5 and 138 out of 43,868 UTR3). These variants were distributed in 151 genes, and 29 of these genes (with a GDA score>0.2 in DisGeNET) have been associated with various types of diseases (Figure 4C), again in almost the same order of predominance: neurologic, neoplasm, and congenital. There was an overlap of nine genes with potentially pathogenic nonsynonymous and UTR variants.

Inbreeding features inferred from the AP WES

The inference of relatedness between the 90 AP WES, based on the genomic kinship coefficient, revealed that all the individuals were classified as unrelated (up 3th-degree), as expected for a cohort representative of the general population. We then estimated ROHs longer than 1Mb in AP and other worldwide
populations, to compare inbreeding features (Table S1). In concordance with the known higher consanguinity of AP populations, there were more ROHs of all sizes in Arabia than in the other regions of the globe (Figure 5A), mainly for the ROH size category 2-4Mb. The largest ROHs detected in each AP population were: Oman-70.4Mb; Yemen-59.8Mb; Saudi Arabia-41.2Mb; and UAE-26Mb.

Not only AP individuals had larger ROHs, but also the total sum of ROHs (SROHs) per AP genome was in mean twice the values observed in the individuals from other worldwide populations (Figure 5B): 1180Mb in AP; 730Mb in East Asians, 700Mb in Europeans; and 650Mb in Africans. This distribution is compatible with different population histories (Ceballos et al., 2018, 2019): the lower in the graph the more panmictic is the population; the higher in the graph, and especially the larger the ROHs (SROHs) the more consanguineous is the population. Within AP, the highest values were observed in Saudi Arabia and the lowest in Oman. In terms of SROH distribution within each population group, there was a high heterogeneity between values from AP individuals (Figure S2), whereas individuals from other regions were more homogeneous in this metric. This highest SROHs values in each AP country were: Saudi Arabia-1969Mb; Yemen-1898Mb; UAE-1805Mb; and Oman-1690MB.

Of interest, within AP the burden of pathogenic variants increases with the SROHs (Figure 5C), both for predicted pathogenic nonsynonymous variants (inferred for both SIFT and PolyPhen algorithms; $r^2 = 0.7573$; $p$ value = 8.512e-29) and variants with C-score $\geq$ 20 ($r^2 = 0.7772$; $p$ value = 1.954e-20). Most of these variants are low frequent in AP (Figure S3), and, as expected, they are absent in comparable-size samples from the other regions of the globe.

## DISCUSSION

In the present study, we have performed a WES exome capture enriched for UTRs on AP populations, still poorly represented on international WES catalogs. We were able to identify 17,060 new variants (11.6% of the total 145,630), that were not cataloged in other public databases, and almost half of these were located in UTR3. We further confirmed that 185 of the UTR variants could be deleterious (inference based on the scaled C-score), being located in genes associated with various types of diseases, especially neurologic and congenital malformation, testifying the importance of screening UTR variants (Steri et al., 2018; Chatterjee and Pal, 2009). Our results match a previous report (Steri et al., 2018) focused on the UTR variants from the NHGRI GWAS Catalog (MacArthur et al., 2017), which were shown to be mostly associated with immunologic, neoplastic and neurological pathologies. As the fine-mapping of UTR regions, especially in terms of their functional impact, is lacking behind the coding regions, there are currently no better pathogenic inferences for these variants than metrics based on conservation. As we improve our knowledge on the functional impact of UTRs, we will refine inferences about their role in diseases. These UTR variants add up to the substantial proportion in the AP WES of 4,673 nonsynonymous variants (15.56% out of 30,033 total) inferred as pathogenic by SIFT and PolyPhen metrics, and that were distributed in 3,285 genes,
23% of which have been associated also with neurologic and congenital malformation diseases, as well as other complex disorders.

We have demonstrated that this high burden on pathogenic variants in a relatively low diverse WES cohort (half the level of European and East Asian, and one-third of sub-Saharan cohorts) can be explained by the demonstrated considerably high proportion of ROHs because of consanguinity practices. It is known that the strong bottleneck in the OOA dispersal led to a higher burden of pathogenic variants in Europeans and East Asians relative to sub-Saharan Africans (Henn et al., 2016), but the values for AP are impressive. The low diversity in the AP WES seems contra-intuitive to the identified high admixture in AP populations of sub-Saharan African and South Asian ancestries (Fernandes et al., 2015, 2019; Ferreira et al., 2021), but the high consanguinity is strong enough to oppose the enrichment with diversity from other population groups.

We must reinforce that we were careful in pre-selecting individuals with a main Arabian/NearEast background for WES screening, avoiding Arabian individuals with recent events of admixture. So, values of SROHs would still be more heterogeneous for AP populations if we had included these Arabians with more recent events of admixture, as can be seen in the heterogeneous and large Qatari cohort from Razali et al. (2021). As it is, the main Arabian/NearEast background WES studied here contained several large and homozygous blocks, especially so in Saudi Arabia, where first-cousin marriages are common (Almazroua et al., 2020). Two Saudi Arabian genomes had the highest sum of ROHs, as high as 2,000Mb. For contextualization, the diploid human genome has 6,200Mb, so those Saudi individuals had ROHs across 32% of

Figure 5. Inbreeding features (Table S1)

(A) Mean number of ROHs, NROHs ≥ 1 Mb per population in different length categories, for AP and other globe regions.
(B) NROHs versus the mean sum ROH length (SROH) for each region: Africa in red (ESN-Esan in Nigeria, GWD-Mandinka in Gambia, LWK-Luhya in Kenya, MSL-Mende in Sierra Leone, and YRI-Yoruba in Nigeria); AP in green; East Asia in blue (CDX-Dai in China, CHB-Han in China, CHS-Han in South China, JPT-Japanese in Japan, and KHV-Kinh in Vietnam), and Europe in purple (GBR-British in UK, IBS-Iberians in Spain, and TSI-Toscani in Italia). The diagonal (gray line) was obtained by regressing both variables for the 1000 Genome populations.
(C) Linear regression between SROH and predicted pathogenic nonsynonymous variants (for both SIFT and PolyPhen algorithms), and between SROH and variants with C-score ≥ 20, in individuals from AP.
their genomes. In mean, the Arabian individuals had ROHs across 19% of their genomes, against 12% in East Asians, 11% in Europeans and 10% in sub-Saharan Africans. And as we are sequencing coding and regulatory regions of the genome, it is not surprising the high amount of predicted pathogenic variants we found.

Our findings highlight the importance of continuing to catalog WES in general population cohorts, and in regions of the globe poorly represented in international consortia. The gains definitely pay off the efforts. A contribution of near 17,000 new variants from sequencing around 2% of the genome in 90 AP individuals is substantial. The pursuit of this cataloging in populations with high consanguinity is advisable. As we have seen here, the higher the consanguinity the higher the burden of potentially pathogenic variants. These variants are harder to detect in panmitic populations because of their removal by genetic drift, which is opposed by consanguinity. In conclusion, we must enrich WES catalogs in ethnic groups and in populations with diverse breeding features, to increase the power to robustly identify disease-associated variants in the human species.

Limitations of the study
As we have demonstrated, the inclusion of the regulatory UTRs revealed a high amount of new variants. Other important regulatory regions, such as promoter and enhancers, were not included in our sequencing screening. These extra regions are potential sources of new regulatory variants of pathogenic importance. Their characterization, through a whole genome sequencing strategy, is a valid solution in the near future as prices per sample continue to decrease.

STAR METHODS
Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Sample collection and WES capture
- METHOD DETAILS
  - Variant calling, filtering and annotation
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Functional constraint analysis
  - Inbreeding features
  - Mining of diseases associated with highlighted genes
  - Graphs and statistical tests

SUPPLEMENTAL INFORMATION
Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.105336.

ACKNOWLEDGMENTS
This work was financed by FEDER-Fundo Europeu de Desenvolvimento Regional funds through COMPETE 2020-Operacional Program for Competitiveness and Internationalisation (POCI), Portugal 2020, by Portuguese funds through FCT-Fundação para a Ciência e a Tecnologia, Ministério da Ciência, Tecnologia e Inovação in the framework of the project “Biomedical anthropological study in Arabian Peninsula based on high throughput genomics” (POCI-01-0145-FEDER-016609). VF has a postdoc grant through FCT (SFRH/ BPD/114927/2016). i3S is financed by FEDER-COMPETE 2020, Portugal 2020 and by Portuguese funds through FCT in the framework of the project ‘Institute for Research and Innovation in Health Sciences’ (POCI-01-0145-FEDER-007274).

AUTHOR CONTRIBUTIONS
V.F. and L.P. designed the study. J.F. performed the analyses. All authors participated in the interpretation of the results and writing the manuscript, and approved the submission.
DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 6, 2022
Revised: August 1, 2022
Accepted: October 10, 2022
Published: November 18, 2022

REFERENCES

Abdulazeez, S., Al Qahtani, N.H., Almandil, N.B., Al-Amodi, A.M., Aldakeel, S.A., Ghahem, N.Z., Alkuroud, D.N., Alturki, A., Alqattan, Q.A., Alghamdi, A., et al. (2019). Genetic disorder prenatal diagnosis and pregnancy termination practices among high consanguinity population, Saudi Arabia. Sci. Rep. 9, 17248.

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods 7, 248–249.

Al-Gazali, L., Hamamy, H., and Al-Arrayad, S. (2006). Genetic disorders in the Arab world. BMJ 333, 831–834.

Almaghouth, I.A., Hassen, L.M., Almahairi, H.S., Bedaiwi, A., Albarrak, R., Daghhestani, M., Alqurtas, E., Alkhalaf, A., Bedaiwi, M., Omar, M., et al. (2021). National systemic lupus erythematosus prospective cohort in Saudi Arabia: a study protocol. Medicine 100, e26704.

Almarni, M.A., Haber, M., Lootah, R.A., Hallast, P., Al Turki, S., Martin, H.C., Xue, Y., and Tyler-Smith, C. (2021). The genomic history of the Middle East. Cell 184, 4612–4625.e14.

Almarri, M.A., Almazroua, A.M., Alsughayer, L., Ababtain, R., Almarni, M.A., Haber, M., Lootah, R.A., Hallast, P., Al Turki, S., Martin, H.C., Xue, Y., and Tyler-Smith, C. (2021). The genomic history of the Middle East. Cell 184, 4612–4625.e14.

Auer, P.L., Reiner, A.P., Wang, G., Kang, H.M., Abecasis, G.R., Alshuleh, D., Bamshad, M.J., Nickerson, D.A., Tracy, R.P., Rich, S.S.; NHLBI GO Exome Sequencing Project, and Leal, S.M. (2016). Guidelines for large-scale sequence-based complex trait association studies: lessons learned from the NHLBI exome sequencing project. Am. J. Hum. Genet. 99, 791–801.

Buniello, A., Macarthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., Mcmahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 47, D1005–D1012.

Caballos, F.C., Hazelhurst, S., and Ramsay, M. (2019). Runs of homozygosity in sub-Saharan African populations provide insights into complex demographic histories. Hum. Genet. 138, 1123–1142.

Caballos, F.C., Joshi, P.K., Clark, D.W., Ramsay, M., and Wilson, J.F. (2018). Runs of homozygosity: windows into population history and trait architecture. Nat. Rev. Genet. 19, 220–234.

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 4, 7.

Chatterjee, S., and Pal, J.K. (2009). Role of 5′- and 3′-untranslated regions of miRNAs in human diseases. Biol. Cell 101, 251–262.

Clarke, L., Fairley, S., Zheng-Bradley, X., Streeter, I., Perry, E., Lowy, E., Tasse, A.M., and Flicek, P. (2017). The international Genome sample resource (IGSR): a worldwide collection of genome variation incorporating the 1000 Genomes Project data. Nucleic Acids Res. 45, D854–D859.

Cox, N. (2018). UK Biobank shares the promise of big data. Nature 562, 194–195.

Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput. Biol. 6, e1001025.

Depristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., Del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43, 491–498.

Fakhro, K.A., Staudt, M.R., Ramstetter, M.D., Robay, A., Malek, J.A., Badri, R., Al-Marri, A.A.N., Abi Khalil, C., Al-Shakaki, A., Chidiac, C., et al. (2016). The Qatar genome: a population-specific tool for precision medicine in the Middle East. Hum. Genome Var. 3, 16016.

Fernandes, V., Alshamali, F., Alves, M., Costa, M.D., Pereira, J.B., Silva, N.M., Cherni, L., Harich, N., Cerny, N., Soares, P., et al. (2012). The Arabian cradle: mitochondrial relics of the first steps along the southern route out of Africa. Am. J. Hum. Genet. 90, 347–355.

Fernandes, V., Brucato, N., Ferreira, J.C., Pedro, N., Cavadas, B., Ricart, F.X., Alshamali, F., and Pereira, L. (2019). Genome-Wide characterization of Arabian Peninsula populations: shedding light on the history of afundamental bridge between continents. Mol. Biol. Evol. 36, 575–586.

Fernandes, V., Triska, P., Pereira, J.B., Alshamali, F., Rito, T., Machado, A., Font-Porterias, N., Caro-Consuegra, R., Quintana-Murci, L., and Comas, D. (2021). The counteracting effects of demography on functional genomic variation: the roma paradigm. Mol. Biol. Evol. 38, 2805–2817.

Fu, W., O’Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Rieder, M.J., Altshuler, D., Shendure, J., et al.; NHLBI Exome Sequencing Project (2013). Analysis of 6, 515 exomes reveals the recent origin of most human protein-coding variants. Nature 493, 216–220.

Henn, B.M., Botiguè, L.R., Peischl, S., Dumanlop, I., Lipatov, M., Maples, B.K., Martin, A.R., Musharoff, S., Cann, H., Snyder, M.P., et al. (2016). Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. Proc. Natl. Acad. Sci. USA 113, E440–E449.

International HapMap Consortium (2003). The international HapMap project. Nature 426, 789–796.

Jeronic, A., Memari, Y., Ritchie, G.R., Hendricks, A.E., Kola-Kokocinski, A., Matchan, A., Vitart, V., Hayward, C., Kolici, I., Glodzik, D., et al. (2016). Whole-exome sequencing in an isolated population from the Dalmatian island of Vis. Eur. J. Hum. Genet. 24, 1479–1487.

Kancheva, D., Atkinson, D., De Rijk, P., Zimon, M., Chavmova, T., Mitev, V., Yaramis, A., Maria Fabrizi, G., Topaloglu, H., Tournev, I., et al. (2016). Novel mutations in genes causing hereditary spastic paraplegia and Charcot-Marie-Tooth neuropathy identified by an optimized protocol for homozygosity mapping based on whole-exome sequencing. Genet. Med. 18, 600–607.

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Affoldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al.; Genome Aggregation Database Consortium (2020). The mutational constraint spectrum quantified from variation in 141, 456 humans. Nature 581, 434–443.

Kircher, M., Witten, D.M., Jain, P., O’roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. 46, 310–315.

Kirin, M., Mcquillan, R., Franklin, C.S., Campbell, H., Mckiege, P.M., and Wilson, J.F. (2010). Genomic runs of homozygosity record population history and consanguinity. PLoS One 5, e13996.

Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous
variants on protein function using the SIFT algorithm. Nat. Protoc. 4, 1073–1081.

Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipirala, S., Gu, B., Hart, J., Hoffman, D., Hoovер, J., et al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 44, D982–D988.

Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and Maglott, D.R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 42, D983–D985.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O’donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 265–291.

Li, H. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics 30, 2843–2851.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R., 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMTools. Bioinformatics 25, 2078–2087.

Liu, X., Wu, C., Li, C., and Boerwinkle, E. (2016). dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. Hum. Mutat. 37, 235–241.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hasting, E., Junkins, H., Mcmahon, A., Milan, A., Morales, J., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 45, D896–D901.

Mamakani, K., Myrvold, W., and Ruskey, F. (2011). Generating all simple convexly-drawable polar symmetric 6-venn diagrams. In International Workshop on Combinatorial Algorithms, 7056, pp. 275–286.

Manchaitul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics 26, 2867–2873.

Mccarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., et al. Haptype Reference Consortium (2016). A reference panel of 64,976 haplotypes for genome-wide association studies. Nat. Genet. 48, 1279–1283.

Monies, D., Abouelhoda, M., Assoum, M., Moghrabi, N., Rafiullah, R., Almontashiri, N., Alowain, M., Alzaidan, H., Alsayed, M., Subhani, S., et al. (2019). Lessons learned from large-scale, first-tier clinical exome sequencing in a highly consanguineous population. Am. J. Hum. Genet. 104, 1182–1201.

Overbey, E.G., Da Silva, V.A., Stanbouly, S., Nishiyama, N.C., Roque-Torres, G.D., Pecaut, M.J., Zawieja, D.C., Wang, C., Willey, J.S., Delp, M.D., et al. (2019). Spaceflight influences gene expression, photoreceptor integrity, and oxidative stress-related damage in the murine retina. Sci. Rep. 9, 13304.

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

Pereira, L., Mutesa, L., Tindana, P., and Ramsay, M. (2021). Pinpoint making and adaptation inform a precision medicine agenda. Nat. Rev. Genet. 22, 284–306.

Perscheid, C., Grassnich, B., and Ullacker, M. (2018). Integrative gene selection on gene expression data: providing biological context to traditional approaches. J. Integr. Bioinform. 16.

Pitro, J., Ramirez-Anguita, M.J., Sauch-Pitchar, J., Ronzano, F., Centeno, E., Sanz, F., and Furlong, L.I. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res. 48, D845–D855.

Pollard, K.S., Hubisz, M.J., Rosenberg, L.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 20, 110–121.

Poplin, R., Ruano-Rubio, V., Depristo, M.A., Fennell, T.J., Carnero, M.O., Van Der Auwera, G.A., Kling, D.E., Gautier, L.D., Levy-Moonshine, A., Roazen, D., et al. (2018). Scalling accurate genetic variant discovery to tens of thousands of samples. Preprint at bioRxiv https://doi.org/10.1101/2017178.

Project Team, S.G. (2015). The Saudi Human Genome Program: an oasis in the desert of Arab medicine is providing clues to genetic disease. IEEE Pulse 6, 22–26.

R Core Team (2018). R: A language and environment for statistical computing. In Statistical Computing). R Core Team (2018). R: a language and environment for statistical computing. In Statistical Computing). R Core Team (2018). R: a language and environment for statistical computing. In Statistical Computing).

Rozowsky, J.S., Hill, A.J., Cummings, B.B., et al. Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 265–291.

Rutgers University (2007). The sequence alignment/map format and SAMTools. Bioinformatics 25, 2078–2087.

Stein, A.G., Kling, D.E., Gauthier, L.D., Levy-Moonshine, A., Roazen, D., et al. (2018). Scalling accurate genetic variant discovery to tens of thousands of samples. Preprint at bioRxiv https://doi.org/10.1101/2017178.

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

Wang, Q., Pierce-Hoffman, E., Cummings, B.B., Affolter, J., Francisco, L.C., Gautier, L.D., Hill, A.J., O’donnell-Luria, A.H., Genome Aggregation Database Production Team; Genome Aggregation Database Consortium, Karczewski, K.J., MacArthur, D.G., et al. (2018). Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes. Nat. Commun. 11, 2539.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag).

Wingett, S.W., and Andrews, S. (2018). FastQ Screen: a tool for multi-genome mapping and quality control. F1000Res. 7, 1338.
# STAR★METHODS

## KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---------------------|--------|------------|
| **Biological samples** | | |
| DNA of 94 individuals from Arabian Peninsula | This paper | |
| **Deposited data** | | |
| WES captured using the SureSelect Human All Exon V5 + UTRs | European Genome-Phenome Archive; accession number: Study EGAS000001006487; Dataset EGAD00001009162; upon request. | https://ega-archive.org/datasets |
| **Software and algorithms** | | |
| Burrows-Wheeler Aligner (BWA v0.7.16a) | (Li et al., 2009) | https://github.com/lh3/bwa |
| Samtools (v1.5) | (Li et al., 2009) | http://www.htslib.org/ |
| FastQC | (Wingett and Andrews, 2018) | https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ |
| Alfred tools | (Rausch et al., 2019) | https://github.com/tobiasrausch/alfred |
| Picard (v1.139) | Picard Toolkit 2018 | https://github.com/broadinstitute/picard/releases |
| GATK (v4.2.0.0) | (Poplin et al., 2018) | https://github.com/broadinstitute/gatk/releases |
| bcftools (v2.26.0) | (Li, 2014) | https://github.com/samtools/bcftools |
| EiGENSTRAT (v6.1.3) | (Patterson et al., 2006) | https://github.com/DReichLab/EIG |
| ANNOVAR tool (downloaded on 15th July 2022) | (Wang et al., 2010) | https://annovar.openbioinformatics.org/en/latest/ |
| CADD tool (v2.2.7) | (Rentzsch et al., 2021) | https://cadd.gs.washington.edu/snv |
| KING tool (v2.2.7) | (Manichaikul et al., 2010) | https://www.kingrelatedness.com/Download.shtml |
| PLINK tool (version 1.9) | (Chang et al., 2015) | https://www.cog-genomics.org/plink/ |
| Venn and ggplot2 package in R | (Mamakani et al., 2011; Ruskey and Weston, 1997) | https://webhome.cs.uvic.ca/~ruskey/Publications/Venn11/Venn11.html; https://r-graph-gallery.com/ggplot2-package.html |
| **Other** | | |
| WES + UTRs of 94 individuals from Arabian Peninsula | This paper | https://ega-archive.org/datasets |
| The 1000 Genomes | (Clarke et al., 2017) | https://www.internationalgenome.org/data/ |
| human genome GRCh37 | NCBI | https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.1_GRCh37/ |
| GnomAD_WGS (v2.1.1) | (Lek et al., 2016) | https://gnomad.broadinstitute.org/ |
| avsnp150 | (Sherry et al., 2001) | https://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=avsnp150 |
| Haplotype Reference Consortium (HRC) | (McCarthy et al., 2016) | http://www.haplotype-reference-consortium.org/ |
| GnomAD_exome (v2.1.1) | (Wang et al., 2020) | https://gnomad.broadinstitute.org/ |
| NHLBI-ESP (exp6500sv2_all) | (Auer et al., 2016) | https://evs.gs.washington.edu/EVS/ |

(Continued on next page)
RESOURCE AVAILABILITY

Lead contact
Further information and requests should be directed to and will be fulfilled by the lead contact, Dra. Veronica Fernandes (vfernandes@ipatimup.pt).

Materials availability
This study did not generate new unique reagents.

Data and code availability
This paper analyses new and existing, publicly available data. The new whole exome data analysed in this study can be accessed from the EGA repository (European Genome-Phenome Archive; Study: EGAS00001006487; Dataset: EGAD00001009162), upon request. The accession numbers for the existing, publicly available datasets are listed in the key resources table.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Sample collection and WES capture
We conducted WES of 94 individuals from AP (23 from Saudi Arabia, 24 from Yemen, 25 from Oman and 22 from UAE; Table S3). These samples were part of a larger cohort of 420 Dubai residents who were born across the AP. As this cohort was previously analysed with the Illumina Human Omni Express Bead Chip containing 741,000 SNPs (Fernandes et al., 2019; Ferreira et al., 2021), it allowed us to select the 94 individuals for WES based on the genomic information that they were non-related and non-recent migrants from sub-Saharan Africa. This study obtained the ethical approval from the Ethics Committee of the University of Porto, Portugal (17/CEUP/2012).

WES was performed at two companies: Macrogen Inc. (Seoul, South Korea) for 50 samples with a 100x average depth coverage; and STAB VIDA (Caparica, Portugal) for 44 samples with a 30x average depth coverage. The WES were captured using the SureSelect Human All Exon V5 + UTRs (50 Mb) target enrichment kit. Sequencing was performed with 2 × 100 bp paired end reads on Illumina HiSeq platform (Illumina, San Diego, CA, USA) according to manufacturer’s protocol.

In order to have comparable values for other worldwide populations, the 1000 Genomes (Clarke et al., 2017) WGS from African, European and East Asian populations (90 individuals randomly selected from each of these regions; Table S3) were extracted, and variants located in the genomic regions covered by our WES (including regulatory untranslated regions UTRs) were considered for further analyses in this manuscript.

METHOD DETAILS

Variant calling, filtering and annotation
Paired-end reads were trimmed using the trimmomatic tool and aligned to human reference genome NCBI Build 37 using the Burrows-Wheeler Aligner (BWA v0.7.16a) algorithm (Li and Durbin, 2009). The paired-read alignments were sorted and stored in BAM format using samtools (v1.5; Li et al., 2009). All the samples passed the quality control performed with FastQC (Wingett and Andrews, 2018) and Alfred tools (Rausch et al.,...
Duplicates were marked and eliminated with Picard (v1.139), and local re-alignment and Base Quality Score Recalibration were carried out using the Genome Analysis Toolkit (GATK; v4.2.0.0; Poplin et al., 2018). Variants were called using the Genome Analysis Toolkit (GATK v4.2.0.0) Haplotypecaller. A minimum value of 10% was set for missing genotype, leading to exclusion of three samples from further analyses. Variants were initially filtered to have a minimum depth of 7, Phred quality score >30, and genotype quality score >20, using the bcftools (v2.26.0; Li, 2014). In the case of heterozygous callings for which the ratio of the less covered allele (reference or derived) over the total calls was <25%, the genotyping was corrected to homozygous of the most frequent allele, by using an in-house script. Finally, we excluded from the analysis multi-allelic variants, indels and positions for which more than 5% of the genotypes were missing. A principal component analysis (PCA) was performed using the SmartPCA tool from the EIGENSTRAT software package (v6.1.3; Patterson et al., 2006), to identify potential batch effects between laboratories, or outliers. After the PCA, one sample was eliminated as it was a clear outlier. The final dataset contained 90 samples (23 from Saudi Arabia, 24 from Yemen, 24 from Oman and 19 from UAE). To further assess the quality of the calling, we evaluated in the bcftools stats the aggregate transition-to-transversion (Ti/Tv) ratio for all variants and exonic variants.

The ANNOVAR tool (version available in 2019-10-24; Wang et al., 2010) was used for the functional annotation of the called variants, and to verify if they were previously described in the following publicly available databases (downloaded on 15th July 2022): WGS group, which included the GnomAD_WGS V2.1.1 (18th March 2019; Lek et al., 2016)), dbSNP_avsnp150 (29th September 2017; Sherry et al., 2001), Haplotype Reference Consortium (HRC; 3rd December 2015; McCarthy et al., 2016), and 1000 Genomes (24th August 2015); WES group, consisting in GnomAD_exome V2.1.1 (18th March 2019; Wang et al., 2020), NHLBI-ESP (esp6500siv2_all; 22nd December 2014; (esp6500siv2_all; 22nd December 2014; Auer et al., 2016)) and ExAC (29th November 2015); ClinVar (clinvar_20220320; 30th March 2022; Landrum et al., 2016) and dbnsfp42c group (10th July 2021; Liu et al., 2016); and finally, Greater Middle East group (GME; 24th October 2016; Scott et al., 2016).

QUANTIFICATION AND STATISTICAL ANALYSIS

Functional constraint analysis
The online CADD tool (Rentzsch et al., 2021) was used to evaluate the evolutionary conservation of AP (and comparable 1000 Genomes datasets) variants, through several metrics. The first type of metrics considered only alignment-based conservation values: (1) GERP or Genomic Evolutionary Rate Profiling (Davydov et al., 2010), which evaluates non-neutral rates of substitution from multiple mammalian species alignments, and categorizes mutations by their predicted deleterious effect in neutral (−2 < GERP < 2), slightly deleterious (2 < GERP < 4), moderate (4 < GERP < 6), and extremely deleterious (GERP > 6) groups (categories according to (Font-Porterias et al., 2021; Henn et al., 2016)); (2) Phylop (Pollard et al., 2010), also based on the alignment of mammalian species, for which a negative value indicates faster-than-expected evolution, while positive values imply conservation.

Secondly, the metrics SIFT (Sorting Intolerant From Tolerant; (Sim et al., 2012)) and PolyPhen (Adzhubei et al., 2010) that predict whether nonsynonymous substitutions are likely to have a deleterious effect on the protein function were investigated. A nonsynonymous variant with a SIFT score <0.05 will be classified as ‘deleterious’ while others are called ‘tolerated’ (benign) (Kumar et al., 2009). In contrast, PolyPhen2 calculates the probability that a given variant will be ‘benign’ for scores less than or equal to 0.446, ‘possibly damaging’ for scores greater than 0.446 and less than or equal to 0.908, and ‘probably damaging’ for scores greater than 0.908.

Finally, the integrative “scaled C-score” was considered. This score provides a ranking of variants more likely to be deleterious by integrating multiple annotations including conservation and functional information into one metric (Kircher et al., 2014). We applied a cutoff of 20, below which the variants were classified as benign and otherwise harmful, as suggested by the authors (Kircher et al., 2014).

Inbreeding features
To infer the degree of relatedness between the 90 AP WES we used the KING tool (v2.2.7; Manichaimkul et al., 2010) by estimating kinship coefficients and inferring IBD segments for all pairwise relationships. Unrelated pairs can be precisely separated from close relatives with accuracy up 3th-degree.
To assess the individual runs of homozygosity (ROHs), we applied a MAF of 0.01 and estimated ROHs using the PLINK tool (version 1.9; Chang et al., 2015) following these published (Kancheva et al., 2016) parameters: a size threshold (kb) to call an ROH (homozyg-kb) of 1000 kb; a SNP number threshold to call an ROH (Homozyg-snp) of 10 SNPs; a sliding window size in SNPs (Homozyg-window-snp) of 20 SNPs; allowing 10 missing SNPs (Homozyg-window-missing); with a proportion of homozygous windows threshold (Homozyg-window-threshold) of 0.05; a minimum SNP density of 200 kb to call an ROH (Homozyg-density); allowing a maximum gap (Homozyg-gap) of 4000 kb; and allowing only 1 heterozygous SNP (Homozyg-window-het). For each AP individual included in the analysis, the sum of the length of ROHs (SROH) was calculated. These AP SROHs were used for testing the linear regression with the burden of predicted pathogenic non-synonymous variants (inferred for both SIFT and PolyPhen algorithms) and well as the burden of variants with C-score ≥ 20, and a f-statistic test was applied.

**Mining of diseases associated with highlighted genes**

Information for disease associated with variants and genes for which AP individuals presented predicted pathogenic non-synonymous and conserved UTR variants was collected from the DisGeNET database (15th March 2022; Table S2; Piñero et al., 2020): a database that integrates information of human gene-disease associations and variant-disease associations from various repositories (https://www.disgenet.org/), for classification in broad categories. We applied variant-disease association (VDA) and human gene-disease association (GDA) scores greater than 0.2 following the same protocol as (Overbey et al., 2019; Perscheid et al., 2018) (Table S2).

**Graphs and statistical tests**

The plots were built with the venn (Mamakani et al., 2011; Ruskey and Weston, 1997) and ggplot2 package (Wickham, 2016) in R (R Core Team, 2018). The calculations for those plots were performed through in-house R-scripts.