bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

De novo mutation in ancestral generations evolves haplotypes contributing to disease

Zeynep Coban-Akdemir [1†], Xiaofei Song[1†], Davut Pehlivan[1,2], Ender Karaca[1], Yavuz Bayram[1],Tomasz Gambin[3], Shalini N. Jhangiani[4], Donna M. Muzny[4], Richard A. Lewis[1,7], Baylor Hopkins Center for Mendelian Genomics, Pengfei Liu[1], Eric Boerwinkle[4,5], Ada Hamosh[6], Richard A. Gibbs[1,4], V. Reid Sutton[1,7], Nara Sobreira[6], Claudia M. B. Carvalho[1], Jennifer E. Posey[1], Chad A. Shaw[1,8], David Valle[6], James R. Lupski[1,4,7,9*]

1. Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA

2. Section of Neurology, Department of Pediatrics, Baylor College of Medicine, Houston, TX 77030, USA

3. Institute of Computer Science, Warsaw University of Technology, Warsaw 00-665, Poland

4. Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA

5. Human Genetics Center, University of Texas Health Science Center at Houston, TX 77030, USA

6. McKusick-Nathans Department of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

7. Texas Children's Hospital, Houston, TX 77030, USA

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

8. Baylor Genetics, Houston, TX 77021, USA

9. Department of Pediatrics, Baylor College of Medicine, Houston, TX 77030, USA

* Correspondence: jlupski@bcm.edu

† These two authors contributed equally

Summary

We investigated the influences of admixture and consanguinity on the genetic architecture of disease by generating a database of variants derived from exome sequencing (ES) of 853 unrelated Turkish (TK) individuals with different disease phenotypes. We observed that TK genomes are more similar to Europeans with 69.3% of the unique variants (N = 356,613) not present in the Greater Middle Eastern variome. We found higher inbreeding coefficient values in the TK cohort correlating with a larger median span of long-sized (>1.606 Mb) runs of homozygosity (ROH). We show that long-sized ROHs arose from recently configured haplotypes and are enriched for rare homozygous deleterious variants. Such haplotypes, and the combinatorial effect of their embedded ultra-rare variants, provide the most explanatory molecular diagnoses for the TK individuals' observed disease traits. Such haplotype evolution results in homozygosity of disease associated haplotypes due to identity-by-descent in a family or extended clan.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

The transmission of traits, genes and variant alleles from one generation to the next, may result in identity-by-descent (IBD) at a locus in a population characterized by consanguinity or founder effect due to a historical population bottleneck or geographic isolation. Experimentally, evidence for IBD in an individual genome is suggested by the presence of runs of homozygosity (ROH) not accompanied by copy-number variation; i.e. copy number neutral gene dosage at a locus. ROH regions in an individual's genome can be detected by genome-wide SNP arrays or can be calculated from unphased exome sequencing (ES) data that may reveal an absence of heterozygosity for specific genomic intervals.

Genome-wide analysis of ROH regions in individuals from various populations can inform population-level differences (McQuillan et al., 2008) and may provide clues about the demographic and evolutionary history of a population, clan, or nuclear family (Lupski et al., 2011; Nothnagel et al., 2010). For instance, when considering the number of genomic intervals and size (measured genetically in cM and physically in Mb) of ROH in an individual genome, a higher number of loci showing ROH ('distributed ROH') suggests a population bottleneck (Auton et al., 2009), while longer ROH regions ('contiguous ROH') suggests a high rate of consanguinity (Li et al., 2006). On the other hand, fewer and shorter stretches of ROH regions may indicate shared ancestry over tens or hundreds of generations (Kirin et al., 2010).

For clinical genetics, analysis of ROH regions in an individual genome can be used to prioritize pathogenic variations at a gene locus and may unveil potential genetic susceptibility underlying disease (Lohmueller et al., 2008; Tennessen et al., 2012;

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Torkamani et al., 2012). Therefore, analysis of ROH regions can be an adjuvant analytical tool to address the genetic architecture of disease. For instance, homozygosity mapping has been a central and robust approach for the identification of biallelic variants causing autosomal recessive (AR) disease traits. For the identified "disease gene" the locus resides in ROH blocks shared among affected individuals (Broman and Weber, 1999; Lander and Botstein, 1987; Morton, 1991; Seelow et al., 2009; Smith, 1953). This approach was successfully used to map disease genes in consanguineous families with heterogeneous neurological disorders (Alazami et al., 2015), multisystem disorders (Alazami et al., 2008; Alazami et al., 2009), and inborn errors of metabolism (Alkuraya, 2010a, b). Furthermore, analysis of rare homozygous deleterious variants in ROH regions can elucidate the genetic heterogeneity and molecular mechanisms underlying Mendelian traits due to either multi-locus variation or oligogenic/polygenic recessive effects (Ceballos et al., 2018; Kaiser et al., 2015; Karaca et al., 2018; Kirin et al., 2010; McQuillan et al., 2008; Posey et al., 2017).

Both the frequency and the degree of relatedness for consanguineous marriages within populations vary in different geographic regions and countries around the globe; the highest documented frequency is found in Pakistan, reported as 60-76% (Hashmi, 1997) of unions. Although consanguineous families facilitate genetic locus mapping and disease gene discovery, many consanguineous populations, such as the Turkish (TK) population (Bittles, 2001; Bittles and Black, 2010), have not been deeply investigated with genomic studies, including a paucity of clinical genomics studies. The population of Turkey presents peculiar features compared to other Middle East countries. Besides having a high reported rate of consanguinity, 20.1% (Tuncbilek and Koc, 1994), Turkey

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

is also often described as a geographic and social 'bridge' between Asia and Europe, an important hub of both ancient and contemporary population migration. Previous studies of healthy unrelated individuals from the Greater Middle East have demonstrated that the TK peninsula population is the most admixed among the other populations within that region (Scott et al., 2016). Therefore, combined clinical and population genetic variation and substructure studies in the TK population can potentially provide insights about the effects of high admixture and a relatively increased rate of consanguinity in a single population to impact the genomic architecture of disease. Such populations are expected to be presented with a richer gene pool, containing more homozygous deleterious variants in comparison to isolated populations (Erzurumluoglu et al., 2016).

Population structure is relevant to disease because of individual pathogenic alleles confined and enriched within specific populations; however, its impact on rare (MAF <= .05), ultra-rare (<= 0.0001) (Hansen et al., 2019), or even private variants, or combinations of alleles at a genetic locus, to disease traits remains under explored. To investigate the influences of both admixture and consanguinity on population genetic variation and the genetic architecture of disease, we studied a sizeable TK cohort (~1000 genomes) with a considerable amount of consanguinity and admixture. We performed ES and family-based genomics on 1,053 TK individuals from 558 families affected with a wide variety of suspected Mendelian disease traits. We performed an unbiased exome variant analysis that would enable a rare variant family-based genomics approach to elucidate the molecular etiology and define genes and their variant alleles potentially contributing to the clinically observed disease traits. We further

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

carried out systematic genomic and phenotypic analysis of this population that might reveal key features of the TK population variome and substructure that impact the genetic architecture of disease in this TK cohort.

We hypothesized that long-sized ROH regions that are more likely be formed on recently configured haplotypes in populations with a high coefficient of consanguinity and a considerable amount of admixture could be enriched for combinations of rare deleterious variants per locus that are most explanatory for the clinical phenotypic features. Large haplotype blocks resulting from multiple consanguineous unions over a short time span in a family or clan may result from the limited generational time for shuffling haplotype blocks through meiotic recombination and selection to eliminate deleterious haplotypes from the population. Multiple rare pathogenic alleles present in the same haplotype block can be brought to homozygosity in such a context.

Indeed, our data and analyses reveal that long-sized ROH regions are enriched for rare homozygous deleterious variants in the individual genomes of TK affecteds compared to TK-unaffecteds and non-TK affecteds as controls ($P = 5.4e-10$ and $< 2.2e-16$). Analysis of genotype-phenotype correlations using the complete haplotype in long-sized (>1.606 Mb), but not short-sized (40-515 Kb) or medium-sized (0.515-1.606 Mb) ROH, provided the most comprehensive set of molecular diagnoses for the observed disease traits using a systematic analysis of HPO (Human Phenotype Ontology) terms. This unique combination of rare homozygous variants brought together due to IBD in recently configured long-sized ROH regions in each individual genome characteristic of their recent family lineage, pedigree, or clan may explain their clinical phenotypic features and thereby their pathogenicity. This provides further support for the Clan

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Genomics hypothesis (Lupski et al., 2011), the notion that novel rare variant alleles arising within a patient, or the recent past generations of a family or more extended clan, significantly contribute to disease in populations. These data underscore the value of studying alternative population substructures that present with high levels of both admixture and consanguinity to elucidate the molecular mechanisms and genetic and genomic architecture underlying disease in populations.

Results

The Turkish (TK) variome

After removing related individuals from the Baylor Hopkins Center for Mendelian Genomics (BHCMG) database, exomes from 4,933 unrelated individuals remained for further analyses. From this data set, we used principal component analysis (PCA) to investigate the population structure between our TK and non-TK cohorts in comparison to the African, Asian, and European population samples from the 1000 Genomes Project. The first main principal component axis (PC1) separated the African samples from the Asian, the European populations and the TK and non-TK cohorts of the BHCMG samples (Figure 1A). The second main principal component axis (PC2) further separated the Asian samples from the European and the TK and non-TK groups of the BHCMG cohort. These studies showed that the TK genomes were distinct from the African and Asian populations, and more similar to the variomes of European samples compared to the non-TK samples that spread out across different populations (Figure 1A and Supplementary Figure 1).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Regarding the Turkish variome, we found that 23.4%, 17.8%, and 69.3 % of the unique variants (N = 356,613) in the TK individuals were not present in ExAC, the genome aggregation database (gnomAD) and the Greater Middle Eastern (GME) variome, respectively, underscoring the necessity of population-matched control databases to analyze rare variants for potential variant pathogenicity (Figure 1B). The systematic analysis of the TK population variation at each genetic locus produced a TK variome database (publicly available at https://turkishvariomedb.shinyapps.io/tvdb/).

Phenotypic and population structure characterization of a mixed disease cohort

The 1,053 individuals who are of TK origin in the BHCMG samples, includes 695 with disease phenotypes (collectively referred to as 'affecteds'), 351 unaffected family members, and 7 individuals of unknown clinical status (Figure 2A). To determine if there is any ascertainment bias in the recruitment of the TK cohort compared to the non-TK cohort, we compared the affected status of the 1,053 TK subjects to the subjects of a non-TK origin, including 2,987 affecteds, 1,732 unaffecteds, and 799 individuals of unknown clinical status (affected or unaffected status unknown) whose samples had genomic studies by ES using the same experimental methods and analytical pipeline (Figure 2A). Based on the family structure documented in PhenoDB (Hamosh et al., 2013; Sobreira et al., 2015), 482 out of 1,053 TK (45.8%) samples were available as a family-based trio structure (proband(s) + parents, or a trio/quad structure with additional affected/unaffected siblings or other family members), compared to 1,336 out of 3,399 non-TK subjects (39.3%) (Figure 2B), corresponding to 152 out of 558 TK families (27.2%) and 413 out of 1,911 (21.6%) non-TK families.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The TK subjects were recruited into the BHCMG because of suspected Mendelian disorders and presented with a wide variety of clinical phenotypes. To define the range and spectrum of clinical phenotypes observed in these individuals, we generated a phenotypic similarity score using the R package ontologySimilarity (James et al., 2016; Liu et al., 2019b; Posey et al., 2017) between each pair of individuals using the HPO terms recorded in PhenoDB. We performed an unsupervised clustering of those subjects based on this phenotypic similarity score, revealing four major 'disease phenotype groups' in the TK affected individuals. Clusters 1, 3, and 4 consist mainly of individuals who could be grouped in a defined clinical cohort. These include the hypergonadotropic hypogonadism (HH) cohort (Jolly et al., 2019), the arthrogryposis cohort (Bayram et al., 2016; Pehlivan et al., 2019), and the neurological disorders cohort (Karaca et al., 2015), respectively. Cluster 2 consists of individuals with diverse phenotypic features which do not fit into a 'singular disease category or group', i.e. higher order HPO term or generalizable disease state or clinical diagnosis, represented as Figure 2C.

Observation of higher estimated inbreeding coefficient values in the TK subjects

A previous study reported the parental consanguinity level of the TK population as 20.1% (Tuncbilek and Koc, 1994), nearly ~12 fold higher than that of the African, Asian, and European populations from the 1000 Genomes project (Scott et al., 2016). The self-reported consanguinity level is also elevated in the BHCMG TK cohort: 26.3% (147/558 families) compared to 3.0% of the non-TK cohort (57/1,911 families, Figure 3A) (Fisher's exact test P-value=1.62e-56). This comparison could be biased due to

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ascertainment by clinical phenotype or in some cases potentially skewed due to missing information.

To obtain a more objective measure of the consanguinity level, we estimated the inbreeding coefficient (F) values from ES data of an unrelated set of 4,933 individuals available in the BHCMG database. These analyses revealed that the measured F values in the TK affecteds on average were significantly higher than the affecteds of non-TK origin [0.06 (s.d. 0.12) versus 0.03 (s.d. 0.09), respectively; Wilcoxon test one-tailed, P-value=1.5e-12]. Similarly, the F values of TK unaffecteds were increased significantly compared to non-TK unaffecteds [0.06 (s.d. 0.1) versus 0.04 (s.d. 0.08)), respectively; Wilcoxon test one-tailed, P-value 0.00014] (Figure 3B). We further observed that the estimated F values were higher in the individuals with a historical self-report of consanguinity, 0.08, as compared to those without a report of consanguinity available in PhenoDB, 0.04 (Supplementary Figure 2A, Wilcoxon test one-tailed P-value < 2.2e-16).

Comparison of the TK cohort variant spectrum to control variant databases

We examined the characteristics of the distinct 356, 613 variants found in our TK variome as compared to control variant databases including the ExAC, gnomAD and the Greater Middle Eastern (GME) variome database. As shown above, the overall comparison of all distinct SNVs found in the TK genomes (356,613 variants) revealed that 23.4% (83,355 variants), 17.8% (63,425 variants) and 69.3% (246,585) were not represented in ExAC (Lek et al., 2016), gnomAD (https://gnomad.broadinstitute.org) or GME variome (Scott et al., 2016) database, respectively (Figure 1B). To identify how much of this underrepresentation of the TK subjects' variants in those control databases

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

are due to homozygous variants carried in the TK genomes, we especially examined the presence of those variants in the gnomAD. We found that exomes of TK affecteds and unaffecteds have an average number of 5.35 and 3.64 homozygous variants respectively, not represented in the gnomAD database. This number was lower in the non-TK cohort: non-TK affecteds and unaffecteds on average harbor 3.01 and 2.80 homozygous variants not found in the gnomAD database (Figure 3C).

Enrichment of long-sized ROH genomic regions in TK cohort individuals

To test our hypothesis that higher F values measured in the TK cohort correlate with an increased length and number of long-sized ROH regions, we first identified ROH regions from ES data using an informatics tool, BafCalculator (https://github.com/ BCM-Lupskilab/BafCalculator) (Karaca et al., 2018), that calculates genomic intervals with absence of heterozygosity, AOH, from unphased ES data as a surrogate measure of ROH. To obtain copy-number neutral ROH regions from the calculated ROH intervals, we excluded a subset of the apparent homozygous regions caused by CNV deletions. As expected, we observed that the self-reported consanguinity level captured for the BHCMG cohort samples in PhenoDB correlates significantly with the total length of long-sized ROH segments (P-value 5.22e-10, Supplementary Figure 2B).

We then examined and compared the features of ROH segments in each size category (long, medium, and short, see Methods) and the total genome-wide aggregate ROH burden in these four subject groups: TK unaffecteds, non-TK unaffecteds, TK affecteds, and non-TK affecteds (Figure 4 and Supplementary Figures 3A, B and C). The TK affecteds were found to carry long-sized ROH regions greater in total length (median=113.7 Mb), median length (median=4 Mb), and a higher number of ROH

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

region counts (median=18) compared to the non-TK affecteds (Wilcoxon test one-tailed P-values < 2.2e-16 in each of these comparisons) (Figure 4). Concordantly, the same trend was observed among unaffecteds: the TK unaffecteds were found to have long-sized ROH regions greater in total length (median=30.2 Mb), median length (median=2.8 Mb), and a higher count (median=8) compared to the non-TK unaffecteds (Wilcoxon test one-tailed P-values < 2.2e-16, <2.2e-16, and 5e-12, Figure 4).

However, when comparing affected to unaffected subject samples, we observed that only the TK cohort affecteds displayed an enrichment of long-sized ROH regions in terms of total length, median length, and total count compared to the unaffecteds, while this trend was not observed in the non-TK BHCMG cohort (Figure 4). Similar to the long-sized ROH category, when all the ROH segments regardless of their size category were compared, the same pattern was observed among the four subject groups for both total and medium length (Supplementary Figures 3A, B and C, right panels). On the other hand, the total number of ROH segments regardless of their size category did not differ significantly between the TK and non-TK populations, in both affecteds and unaffecteds (Wilcoxon test one-tailed P-values 0.33 and 1). However, the total number of ROH segments was significantly higher in affecteds compared to unaffecteds in both the TK and non-TK cohorts (Wilcoxon test one-tailed P-values < 2.2e-16 and 4.9e-6). In summary, affected subjects were shown to carry more ROH regions in their genomes than unaffected individuals, regardless of their population background (Supplementary Figure 3C, lower right panel), although for the non-TK cohort that does not reflect long-sized ROH segments.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We then examined the characteristics of other ROH size categories. The total length of short-sized ROH regions, which result from short homozygous blocks on ancient haplotypes, did not correlate with self-reported consanguinity levels historically obtained for the BHCMG cohort samples in PhenoDB (Supplementary Figure 2B, P-value 0.215). Concordant with this result, the short-sized ROH regions in the affecteds were greater in size and higher in number than those of unaffecteds regardless of TK- or non-TK origin (Supplementary Figure 3A and C, lower left panel). In contrast, medium-sized ROH regions (505-1606 Kb) that mostly arise due to background relatedness (McQuillan et al., 2008; Pemberton et al., 2012) show a significantly greater total length and a higher count only in the TK-affecteds compared to TK-unaffecteds, but not in non-TK affecteds compared to non-TK unaffecteds (Wilcoxon test one-tailed P-values 0.015 and 0.013) (Supplementary Figure 3A and C, upper right panel).

In aggregate, these ROH analyses support the notion that long-, medium-, and short-sized ROH regions result from individual or personal genome population history and dynamics such as consanguinity level, or background relatedness (McQuillan et al., 2008; Pemberton et al., 2012) in nuclear families and clans. Of note, the long-, medium-, and short-sized ROH regions also show an increased genome-wide burden in either a population-specific or a disease-related manner.

Enrichment of predicted deleterious variants in long-sized ROH regions

We hypothesized that long-sized ROH regions, due to recent parental relatedness, would be enriched for rare homozygous deleterious variants because time, and here we refer to the genetic measure of time, i.e. generations, would not be sufficient for selection to eliminate such detrimental alleles from a population. To test this, we first

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

identified rare homozygous variants (MAF ≤ 0.05) predicted to be potentially deleterious, by selecting variants above a CADD PHRED-scaled score of >=15 and by utilizing a prediction tool algorithm, NMDescPredictor, to predict potential LoF variants (Coban-Akdemir et al., 2018). We tested for an overall increase in rare homozygous deleterious genotype burden in the TK genomes. The TK affecteds and unaffecteds displayed a higher number of rare homozygous deleterious genotypes (median=17 and 9) compared to the non-TK affecteds and unaffecteds (both with median=7), respectively (Wilcoxon test one-tailed P-values < 2.2e-16 and 2.1e-5). In addition, the calculated rare homozygous deleterious genotype number showed a significant elevation in affecteds in comparison to unaffecteds in the TK cohort (Wilcoxon test one-tailed P-value < 2.2e-16) (Supplementary Figure 4A).

To account for variable mutation rates across different samples and genomic regions, we computed a variant density metric by normalizing the count of rare homozygous deleterious variants to the count of rare homozygous synonymous variants. We found a significantly greater value of rare homozygous deleterious variant density in the TK affected and unaffected individuals as compared to the non-TK affecteds and unaffecteds, respectively (P-values < 2.2e-16 and 0.00012, Figure 5A). Our analysis also revealed that such density in the TK affecteds was significantly elevated compared to the TK unaffecteds with P-values < 2.2e-16. However, this disease-related genome-wide rare homozygous deleterious variant burden was population-specific and present only in the TK cohort, not in the non-TK cohort.

Since ROH regions are likely to enable deleterious variation to exist in a homozygous form, we then investigated the contribution of ROH regions to the

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

mutational burden in the BHCMG cohort, by grouping rare homozygous deleterious variants into five groups based on their genomic location: variants within ROH regions regardless of size (total ROH), or outside of an ROH block (non-ROH), or ROH regions in three different size categories (long-, medium- and short-sized). As expected, the rare homozygous deleterious variant burden observed in the TK cohort compared to the non-TK cohort mainly resulted from ROH regions (total ROH, Figure 5B, P-values < 2.2e-16 and 2.4e-12). On the other hand, this burden was not present in non-ROH regions (Figure 5C).

An increased level of mutational burden existing in ROH regions in the TK cohort compared to the non-TK cohort in both affecteds and unaffecteds was more striking in the long-sized ROH regions, which are more likely to be shaped by young haplotype blocks including deleterious variation (Figure 5D, P-values < 2.2e-16 in each comparison). The deleterious variant burden was also represented in medium-sized ROH regions, but to a lesser extent: the TK affecteds presented with a significantly higher level of deleterious variant burden in medium-sized ROH regions compared to the non-TK affecteds and TK-unaffecteds (Supplementary Figure 4A, P-values = 7.4e-9 and 0.00062). On the other hand, despite the fact that the total length and count of short-sized ROH regions differ significantly between affecteds and unaffecteds regardless of population background (Supplementary Figures 3A and C), they do not seem to contribute to the overall homozygous deleterious variant burden presented in ROH regions (Supplementary Figure 4B). In summary, this analysis documents an increased level of rare homozygous deleterious variant burden in both long-sized and

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

medium-sized ROH regions, indicating a contribution of both recent parental relatedness and founder events to mutational burden in the TK population.

Genotype-phenotype analyses revealing ROH-associated genetic architecture of diseases

We then tested to what extent the mutational burden caused by rare homozygous deleterious variants in long-sized ROH regions explains clinical phenotypic features of the individuals in the BHCMG cohort. To achieve this, as shown as Figure 6A, we first selected rare homozygous deleterious variants in genes that have been associated with a disease phenotype in OMIM (https://www.omim.org). We then linked those disease genes to their related HPO terms according to HPO annotation resource databases. For each ROH category, we combined all those HPO terms related to the disease genes harboring rare homozygous deleterious variants located in those regions. Next, for the clinical phenotypic features recorded in PhenoDB, we compared the associated HPO term sets to the merged HPO term sets defined for each ROH category, with a normalized semantic similarity score metric (James et al., 2016; Liu et al., 2019b; Posey et al., 2017). Last, we evaluated the contribution of disease genes in each ROH category to individual patient phenotypes compared to all genes presented with rare homozygous deleterious variants throughout the genome.

These analyses showed that, in the TK patients, genes with rare homozygous deleterious variants in long-sized ROH regions are most informative of clinical phenotypic features objectively as listed in the information submitted to PhenoDB (Figure 6B, left panel). However, in the non-TK cohort, genes in long-sized ROH regions with rare homozygous deleterious variants show no significant differences in their ability

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

to explain phenotypic features as compared to those in genes in the non-ROH regions (Figure 6B, right panel). This result independently confirms the previous analysis showing that long-sized ROH regions did not have a significant difference in rare homozygous deleterious variant burden between the non-TK affecteds and unaffecteds (P-value = 0.059, Figure 5D). In summary, our results provide evidence that an increased level of consanguinity observed in the TK population is correlated with a higher number of long-sized ROH blocks, thereby an elevated mutational burden that likely contributes to the disease trait phenotype observed in the individuals studied.

Discussion

Applications of ES and genome sequencing (GS) to unravel rare and pathogenic variation in the genome that contributes to Mendelian disease traits have enabled molecular diagnoses for thousands of patients and also facilitated novel disease gene discovery (Chong et al., 2015; de Ligt et al., 2012; Dharmadhikari et al., 2019a; Eldomery et al., 2017; Lee et al., 2014; Liu et al., 2019b; Lupski et al., 2010; Worthey et al., 2011; Yang et al., 2013; Yang et al., 2014), especially through the CMG program (Posey et al., 2019) and other national and international organized efforts (Deciphering Developmental Disorders, 2015; Sawyer et al., 2016) . Comprehensive catalogues of common and rare human genetic variation stored in population variant databases, such as ExAC (Lek et al., 2016), gnomAD (https://gnomad.broadinstitute.org), and ARIC (Coban-Akdemir et al., 2018; Gambin et al., 2015), have proven to be powerful resources and interpretive tools to identify rare and pathogenic variation that influences expression of Mendelian disease traits. However, there remain many populations for

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

which genetic variation is underrepresented or entirely absent in such databases, including populations with more prevalent consanguinity, such as the TK population (28). Our studies revealed that 23.4%, 17.8%, and 69.3 % of the unique variants (N = 356,613) found in an unrelated set of 853 TK genomes were not present in ExAC, gnomAD, and the GME variome, respectively, underscoring the necessity of population-matched control databases for interpreting potential variant pathogenicity. Although GME studies generated a variome of 1,111 unrelated subjects from the Greater Middle East, only 140 of them were of TK ancestry (Scott et al., 2016). Our findings suggest that a TK variome database may provide an important resource to potentially gain further understanding of genetic diversity and “disease gene” biology in the TK population.

Beyond the mapping of disease genes, analysis and characterization of molecular features in personal genomes can potentially provide some insights into the genetic architecture contributing to disease traits in a population. Using a TK population cohort, we show that investigation of the molecular features of ROH regions culled from personal unphased ES data through BafCalculator (https://github.com/ BCM-Lupskilab/BafCalculator) (Karaca et al., 2018) can provide insights into the genetic architecture of disease. Long-sized ROH (>1.606 Mb) regions that arose on recently configured haplotype blocks were greater in size (both total and median length) in the genomes of the TK cohort. This finding is concordant with higher F values and high level of consanguinity due to recent shared ancestors in the TK population (Scott et al., 2016). Characterization of ROH genomic intervals also revealed that long-sized ROH regions on newly derived haplotypes are particularly enriched with rare homozygous

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

deleterious variants specifically in TK affected personal genomes compared to TK unaffected and non-TK affected (P =5.4-10 and < 2.2e-16) genomes. To test the contribution of those rare variant combinations to Mendelian phenotypes, we performed a large-scale analysis of genotype-phenotype correlations in the BHCMG cohort from their clinical features captured as structured human phenotype ontology (HPO) terms (Kohler et al., 2014) (https://hpo.jax.org/app/download/annotation) recorded in PhenoDB (Hamosh et al., 2013; Sobreira et al., 2015) . Next, we compared patients' associated HPO term sets to the merged HPO term sets defined for OMIM disease genes presented in each ROH length category, (i.e long, medium, small) with a semantic similarity score metric that controls for the number of genes and ROH block size using a permutation approach. This systematic integrative quantitative analysis revealed that the combinatorial phenotypic effect of ultra-rare variants embedded within long-size ROH regions provided the most comprehensive set of molecular diagnoses for the observed disease traits in the TK cohort. That these variants were ultra-rare within the TK population itself suggests that they may represent new mutation events within the clan, with population substructure ultimately driving formation of these newly configured – and unique -- haplotypes that can be rapidly brought to homozygosity through IBD (Narasimhan et al., 2017).

This phenomenon of new mutations as ultra-rare variants on a newly derived haplotype may provide important insights into the molecular etiology of disease traits, and further amplify several prior observations including (1) the role of multilocus variation underlying some cases of apparent phenotypic expansion (Karaca et al., 2018; Liu et al., 2019b; Pehlivan et al., 2019; Posey et al., 2017); (2) homozygosity of ultra-

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

rare pathogenic variation in *PRUNE* in sect-driven population isolates (Karaca et al., 2015; Zollo et al., 2017) in stark contrast with the contribution of rare (but not ultra-rare) *CLP1* founder alleles to microcephaly and neurodevelopmental disease (Karaca et al., 2014; Schaffer et al., 2014); and (3) the seemingly now-common identification of genes for which biallelic variation can lead to disease traits previously categorized as strictly 'dominant Mendelian loci'-traits (Monies et al., 2019; Rainger et al., 2014; Yuan et al., 2015).

Our findings also highlight the advantage of merging per-locus variation with genomics (inherited vs. *de novo* mutations) for gleaning insights into the genetic architecture contributing to disease in a population. Multiple genetic changes could be brought together per locus to generate unique haplotype blocks. This phenomenon could be exemplified by the observation of uniparental isodisomy (UPD) manifest by a long tract of homozygosity on the entirety of chromosome 7 explaining both the short stature phenotype in addition to the expected clinical features of cystic fibrosis (CF, OMIM #219700) in a child (Spence et al., 1988). Viewed from such a perpective, disease phenotypes resulting from UPD, and epigenetic/imprinting diseases may benefit from haplotype phased genomes and long read technologies that differentiate methylated W-C bases (Carvalho et al., 2019). Another example of per-locus genetic studies is the modulation of disease risk through gene expression and dosage effects of regulatory common variant (expression quantitative trait loci (eQTLs) haplotype configurations of coding pathogenic variants and CNVs (Castel et al., 2018; Liu et al., 2019a; Ren et al., 2019; Wu et al., 2015; Yang et al., 2019).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

To apply a 'merging' of genetics (per locus variation) and genomics (inherited 'variome' and *de novo* mutations) thinking in the clinic, our data emphasize identifying and systematically analyzing the ROH genomic intervals culled from the patient's personal unphased ES data using BafCalculator (https://github.com/ BCM-Lupskilab/BafCalculator) (Karaca et al., 2018). If the degree of parental relatedness as judged by the ROH size culled from unphased ES data suggests that the patient may come from a clan with a high value for the coefficient of consanguinity, the rare homozygous variants mapping within the ROH regions and their combinatorial effect should be prioritized in molecular diagnoses (Karaca et al., 2018; Liu et al., 2019b; Pehlivan et al., 2019; Posey et al., 2017). Culling ROH regions from unphased clinical ES data using BafCalculator (https://github.com/ BCM-Lupskilab/BafCalculator) (Karaca et al., 2018) may also enhance the discovery of pathogenic homozygous or hemizygous exonic CNV deletions arising on newly derived structural variant (SV) haplotypes in a clan homozygosed by IBD (Dharmadhikari et al., 2019b; Gambin et al., 2017).

In summary, we provide evidence suggesting that ultra-rare (recently arisen) genomic variants in a clan evolve newly derived haplotypes that could potentilly become homozygosed due to IBD in a given population. Our data demonstrate that such haplotypes, and the combinatorial effect of their embedded ultra-rare variants, are the most explanatory for the TK individuals' observed disease traits. Such haplotype evolution, when placed in the population substructure context of IBD, results in homozygosity of disease haplotypes.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Acknowledgments

We thank all patients and their families and their referring physicians who submitted samples for genomic studies. No additional compensation was offered for these studies. Funding: this work was funded in part by the US National Human Genome Research Institute (NHGRI)/ National Heart Lung and Blood Institute (NHLBI) grant number UM1HG006542 to the Baylor Hopkins Center for Mendelian Genomics (BHCMG) (DV and JRL), the National Institute of Neurological Disorders and Stroke (NINDS) R35NS105078 (JRL), and the National Human Genome Research Institute U54-HG003273 (RAG). JEP was supported by NHGRI K08 HG008986.

Author Contributions

Z.C.A., X.S., and J.R.L. designed the study, analyzed the data, and drafted the manuscript. D.P., E.K., and Y.B. evaluated the clinical studies and the ascertainment and collection of samples from Turkey. A.H, V.R.S, R.A.L. and N.S. supervised the clinical data generation of PhenoDB entries of TK and non-TK subjects. J.R.L. and D.V. supervised the study. S.N.J., T.G., C.M.B.C., D.M.M., P.L, E.B., R.A.G., C.A.S., J.E.P. generated ES data and advised the sequencing and variant data analyses. All contributing co-authors have read, edited, and approved the final manuscript.

Declaration of Interests

J.R.L. has stock ownership in 23andMe, is a paid consultant for Regeneron Pharmaceuticals and Novartis, and is a co-inventor on multiple United States and European patents related to molecular diagnostics for inherited neuropathies, eye

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

diseases, and bacterial genomic fingerprinting. C.A.S and P.L. are employees of Baylor College of Medicine and derive support through a professional services agreement with the Baylor Genetics (BG). The Department of Molecular and Human Genetics at Baylor College of Medicine derives revenue from molecular genetic testing offered at BG (http://www.bcm.edu/geneticlabs/). JRL serves on the SAB of BG. The other authors declare no competing financial interests.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

Alazami, A.M., Al-Saif, A., Al-Semari, A., Bohlega, S., Zlitni, S., Alzahrani, F., Bavi, P., Kaya, N., Colak, D., Khalak, H.*, et al.* (2008). Mutations in *C2orf37*, encoding a nucleolar protein, cause hypogonadism, alopecia, diabetes mellitus, mental retardation, and extrapyramidal syndrome. Am J Hum Genet *83*, 684-691.

Alazami, A.M., Patel, N., Shamseldin, H.E., Anazi, S., Al-Dosari, M.S., Alzahrani, F., Hijazi, H., Alshammari, M., Aldahmesh, M.A., Salih, M.A.*, et al.* (2015). Accelerating novel candidate gene discovery in neurogenetic disorders via whole-exome sequencing of prescreened multiplex consanguineous families. Cell Rep *10*, 148-161.

Alazami, A.M., Shaheen, R., Alzahrani, F., Snape, K., Saggar, A., Brinkmann, B., Bavi, P., Al-Gazali, L.I., and Alkuraya, F.S. (2009). *FREM1* mutations cause bifid nose, renal agenesis, and anorectal malformations syndrome. Am J Hum Genet *85*, 414-418.

Alkuraya, F.S. (2010a). Autozygome decoded. Genet Med *12*, 765-771.

Alkuraya, F.S. (2010b). Homozygosity mapping: one more tool in the clinical geneticist's toolbox. Genet Med *12*, 236-239.

Auton, A., Bryc, K., Boyko, A.R., Lohmueller, K.E., Novembre, J., Reynolds, A., Indap, A., Wright, M.H., Degenhardt, J.D., Gutenkunst, R.N.*, et al.* (2009). Global distribution of genomic diversity underscores rich complex history of continental human populations. Genome Res *19*, 795-803.

Bayram, Y., Karaca, E., Coban Akdemir, Z., Yilmaz, E.O., Tayfun, G.A., Aydin, H., Torun, D., Bozdogan, S.T., Gezdirici, A., Isikay, S.*, et al.* (2016). Molecular etiology of arthrogryposis in multiple families of mostly Turkish origin. J Clin Invest *126*, 762-778.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Bittles, A. (2001). Consanguinity and its relevance to clinical genetics. Clin Genet *60*, 89-98.

Bittles, A.H., and Black, M.L. (2010). Evolution in health and medicine Sackler colloquium: Consanguinity, human evolution, and complex diseases. Proc Natl Acad Sci U S A *107 Suppl 1*, 1779-1786.

Broman, K.W., and Weber, J.L. (1999). Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain. Am J Hum Genet *65*, 1493-1500.

Carvalho, C.M.B., Coban-Akdemir, Z., Hijazi, H., Yuan, B., Pendleton, M., Harrington, E., Beaulaurier, J., Juul, S., Turner, D.J., Kanchi, R.S.*, et al.* (2019). Interchromosomal template-switching as a novel molecular mechanism for imprinting perturbations associated with Temple syndrome. Genome Med *11*, 25.

Castel, S.E., Cervera, A., Mohammadi, P., Aguet, F., Reverter, F., Wolman, A., Guigo, R., Iossifov, I., Vasileva, A., and Lappalainen, T. (2018). Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. Nat Genet *50*, 1327-1334.

Ceballos, 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.

Challis, D., Yu, J., Evani, U.S., Jackson, A.R., Paithankar, S., Coarfa, C., Milosavljevic, A., Gibbs, R.A., and Yu, F. (2012). An integrative variant analysis suite for whole exome next-generation sequencing data. BMC Bioinformatics *13*, 8.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Chong, J.X., Buckingham, K.J., Jhangiani, S.N., Boehm, C., Sobreira, N., Smith, J.D., Harrell, T.M., McMillin, M.J., Wiszniewski, W., Gambin, T.*, et al.* (2015). The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. Am J Hum Genet *97*, 199-215.

Coban-Akdemir, Z., White, J.J., Song, X., Jhangiani, S.N., Fatih, J.M., Gambin, T., Bayram, Y., Chinn, I.K., Karaca, E., Punetha, J.*, et al.* (2018). Identifying Genes Whose Mutant Transcripts Cause Dominant Disease Traits by Potential Gain-of-Function Alleles. Am J Hum Genet *103*, 171-187.

de Ligt, J., Willemsen, M.H., van Bon, B.W., Kleefstra, T., Yntema, H.G., Kroes, T., Vulto-van Silfhout, A.T., Koolen, D.A., de Vries, P., Gilissen, C.*, et al.* (2012). Diagnostic exome sequencing in persons with severe intellectual disability. N Engl J Med *367*, 1921-1929.

Deciphering Developmental Disorders, S. (2015). Large-scale discovery of novel genetic causes of developmental disorders. Nature *519*, 223-228.

Dharmadhikari, A., Ghosh, R., Yuan, B., Liu, P., Dai, H., Masri, S., Scull, J., Posey, J., Jiang, A., He, W.*, et al.* (2019a). Copy number variant and runs of homozygosity detection by microarrays enabled more precise molecular diagnoses in 11,020 clinical exome cases. In press Genome Med.

Dharmadhikari, A.V., Ghosh, R., Yuan, B., Liu, P., Dai, H., Al Masri, S., Scull, J., Posey, J.E., Jiang, A.H., He, W.*, et al.* (2019b). Copy number variant and runs of homozygosity detection by microarrays enabled more precise molecular diagnoses in 11,020 clinical exome cases. Genome Med *11*, 30.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Eldomery, M.K., Coban-Akdemir, Z., Harel, T., Rosenfeld, J.A., Gambin, T., Stray-Pedersen, A., Kury, S., Mercier, S., Lessel, D., Denecke, J.*, et al.* (2017). Lessons learned from additional research analyses of unsolved clinical exome cases. Genome Med *9*, 26.

Erzurumluoglu, A.M., Shihab, H.A., Rodriguez, S., Gaunt, T.R., and Day, I.N. (2016). Importance of Genetic Studies in Consanguineous Populations for the Characterization of Novel Human Gene Functions. Ann Hum Genet *80*, 187-196.

Fromer, M., and Purcell, S.M. (2014). Using XHMM Software to Detect Copy Number Variation in Whole-Exome Sequencing Data. Curr Protoc Hum Genet *81*, 7 23 21-21.

Gambin, T., Akdemir, Z.C., Yuan, B., Gu, S., Chiang, T., Carvalho, C.M.B., Shaw, C., Jhangiani, S., Boone, P.M., Eldomery, M.K.*, et al.* (2017). Homozygous and hemizygous CNV detection from exome sequencing data in a Mendelian disease cohort. Nucleic Acids Res *45*, 1633-1648.

Gambin, T., Jhangiani, S.N., Below, J.E., Campbell, I.M., Wiszniewski, W., Muzny, D.M., Staples, J., Morrison, A.C., Bainbridge, M.N., Penney, S.*, et al.* (2015). Secondary findings and carrier test frequencies in a large multiethnic sample. Genome Med *7*, 54.

Hamosh, A., Sobreira, N., Hoover-Fong, J., Sutton, V.R., Boehm, C., Schiettecatte, F., and Valle, D. (2013). PhenoDB: a new web-based tool for the collection, storage, and analysis of phenotypic features. Hum Mutat *34*, 566-571.

Hansen, A.W., Murugan, M., Li, H., Khayat, M.M., Wang, L., Rosenfeld, J., Andrews, B.K., Jhangiani, S.N., Coban Akdemir, Z.H., Sedlazeck, F.J.*, et al.* (2019). A Genocentric Approach to Discovery of Mendelian Disorders. Am J Hum Genet *105*, 974-986.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Hashmi, M.A. (1997). Frequency of consanguinity and its effect on congenital malformation--a hospital based study. J Pak Med Assoc *47*, 75-78.

Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B.S., Bravo, H.C., Davis, S., Gatto, L., Girke, T.*, et al.* (2015). Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods *12*, 115-121.

James, R.A., Campbell, I.M., Chen, E.S., Boone, P.M., Rao, M.A., Bainbridge, M.N., Lupski, J.R., Yang, Y., Eng, C.M., Posey, J.E.*, et al.* (2016). A visual and curatorial approach to clinical variant prioritization and disease gene discovery in genome-wide diagnostics. Genome Med *8*, 13.

Jolly, A., Bayram, Y., Turan, S., Aycan, Z., Tos, T., Abali, Z.Y., Hacihamdioglu, B., Coban Akdemir, Z.H., Hijazi, H., Bas, S.*, et al.* (2019). Exome Sequencing of a Primary Ovarian Insufficiency Cohort Reveals Common Molecular Etiologies for a Spectrum of Disease. J Clin Endocrinol Metab *104*, 3049-3067.

Kaiser, V.B., Svinti, V., Prendergast, J.G., Chau, Y.Y., Campbell, A., Patarcic, I., Barroso, I., Joshi, P.K., Hastie, N.D., Miljkovic, A.*, et al.* (2015). Homozygous loss-of-function variants in European cosmopolitan and isolate populations. Hum Mol Genet *24*, 5464-5474.

Karaca, E., Harel, T., Pehlivan, D., Jhangiani, S.N., Gambin, T., Coban Akdemir, Z., Gonzaga-Jauregui, C., Erdin, S., Bayram, Y., Campbell, I.M.*, et al.* (2015). Genes that Affect Brain Structure and Function Identified by Rare Variant Analyses of Mendelian Neurologic Disease. Neuron *88*, 499-513.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Karaca, E., Posey, J.E., Coban Akdemir, Z., Pehlivan, D., Harel, T., Jhangiani, S.N., Bayram, Y., Song, X., Bahrambeigi, V., Yuregir, O.O.*, et al.* (2018). Phenotypic expansion illuminates multilocus pathogenic variation. Genet Med *20*, 1528-1537.

Karaca, E., Weitzer, S., Pehlivan, D., Shiraishi, H., Gogakos, T., Hanada, T., Jhangiani, S.N., Wiszniewski, W., Withers, M., Campbell, I.M.*, et al.* (2014). Human *CLP1* mutations alter tRNA biogenesis, affecting both peripheral and central nervous system function. Cell *157*, 636-650.

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

Kohler, S., Doelken, S.C., Mungall, C.J., Bauer, S., Firth, H.V., Bailleul-Forestier, I., Black, G.C., Brown, D.L., Brudno, M., Campbell, J.*, et al.* (2014). The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res *42*, D966-974.

Lander, E.S., and Botstein, D. (1987). Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science *236*, 1567-1570.

Lee, H., Deignan, J.L., Dorrani, N., Strom, S.P., Kantarci, S., Quintero-Rivera, F., Das, K., Toy, T., Harry, B., Yourshaw, M.*, et al.* (2014). Clinical exome sequencing for genetic identification of rare Mendelian disorders. JAMA *312*, 1880-1887.

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.* (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285-291.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Li, L.H., Ho, S.F., Chen, C.H., Wei, C.Y., Wong, W.C., Li, L.Y., Hung, S.I., Chung, W.H., Pan, W.H., Lee, M.T.*, et al.* (2006). Long contiguous stretches of homozygosity in the human genome. Hum Mutat *27*, 1115-1121.

Liu, J., Wu, N., Deciphering Disorders Involving, S., study, C.O., Yang, N., Takeda, K., Chen, W., Li, W., Du, R., Liu, S.*, et al.* (2019a). *TBX6*-associated congenital scoliosis (TACS) as a clinically distinguishable subtype of congenital scoliosis: further evidence supporting the compound inheritance and *TBX6* gene dosage model. Genet Med *21*, 1548-1558.

Liu, P., Meng, L., Normand, E.A., Xia, F., Song, X., Ghazi, A., Rosenfeld, J., Magoulas, P.L., Braxton, A., Ward, P.*, et al.* (2019b). Reanalysis of Clinical Exome Sequencing Data. N Engl J Med *380*, 2478-2480.

Lohmueller, K.E., Indap, A.R., Schmidt, S., Boyko, A.R., Hernandez, R.D., Hubisz, M.J., Sninsky, J.J., White, T.J., Sunyaev, S.R., Nielsen, R.*, et al.* (2008). Proportionally more deleterious genetic variation in European than in African populations. Nature *451*, 994-997.

Lupski, J.R., Belmont, J.W., Boerwinkle, E., and Gibbs, R.A. (2011). Clan genomics and the complex architecture of human disease. Cell *147*, 32-43.

Lupski, J.R., Reid, J.G., Gonzaga-Jauregui, C., Rio Deiros, D., Chen, D.C., Nazareth, L., Bainbridge, M., Dinh, H., Jing, C., Wheeler, D.A.*, et al.* (2010). Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. N Engl J Med *362*, 1181-1191.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

McQuillan, R., Leutenegger, A.L., Abdel-Rahman, R., Franklin, C.S., Pericic, M., Barac-Lauc, L., Smolej-Narancic, N., Janicijevic, B., Polasek, O., Tenesa, A.*, et al.* (2008). Runs of homozygosity in European populations. Am J Hum Genet *83*, 359-372.

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.

Morton, N.E. (1991). Genetic epidemiology of hearing impairment. Ann N Y Acad Sci *630*, 16-31.

Narasimhan, V.M., Rahbari, R., Scally, A., Wuster, A., Mason, D., Xue, Y., Wright, J., Trembath, R.C., Maher, E.R., van Heel, D.A.*, et al.* (2017). Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes. Nat Commun *8*, 303.

Nothnagel, M., Lu, T.T., Kayser, M., and Krawczak, M. (2010). Genomic and geographic distribution of SNP-defined runs of homozygosity in Europeans. Hum Mol Genet *19*, 2927-2935.

Olshen, A.B., Venkatraman, E.S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics *5*, 557-572.

Pehlivan, D., Bayram, Y., Gunes, N., Coban Akdemir, Z., Shukla, A., Bierhals, T., Tabakci, B., Sahin, Y., Gezdirici, A., Fatih, J.M.*, et al.* (2019). The Genomics of Arthrogryposis, a Complex Trait: Candidate Genes and Further Evidence for Oligogenic Inheritance. Am J Hum Genet *105*, 132-150.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Pemberton, T.J., Absher, D., Feldman, M.W., Myers, R.M., Rosenberg, N.A., and Li, J.Z. (2012). Genomic patterns of homozygosity in worldwide human populations. Am J Hum Genet *91*, 275-292.

Posey, J.E., Harel, T., Liu, P., Rosenfeld, J.A., James, R.A., Coban Akdemir, Z.H., Walkiewicz, M., Bi, W., Xiao, R., Ding, Y.*, et al.* (2017). Resolution of Disease Phenotypes Resulting from Multilocus Genomic Variation. N Engl J Med *376*, 21-31.

Posey, J.E., O'Donnell-Luria, A.H., Chong, J.X., Harel, T., Jhangiani, S.N., Coban Akdemir, Z.H., Buyske, S., Pehlivan, D., Carvalho, C.M.B., Baxter, S.*, et al.* (2019). Insights into genetics, human biology and disease gleaned from family based genomic studies. Genet Med.

Quinlan, A.R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr Protoc Bioinformatics *47*, 11 12 11-34.

Rainger, J., Pehlivan, D., Johansson, S., Bengani, H., Sanchez-Pulido, L., Williamson, K.A., Ture, M., Barker, H., Rosendahl, K., Spranger, J.*, et al.* (2014). Monoallelic and biallelic mutations in *MAB21L2* cause a spectrum of major eye malformations. Am J Hum Genet *94*, 915-923.

Reid, J.G., Carroll, A., Veeraraghavan, N., Dahdouli, M., Sundquist, A., English, A., Bainbridge, M., White, S., Salerno, W., Buhay, C.*, et al.* (2014). Launching genomics into the cloud: deployment of Mercury, a next generation sequence analysis pipeline. BMC Bioinformatics *15*, 30.

Ren, X., Yang, N., Wu, N., Xu, X., Chen, W., Zhang, L., Li, Y., Du, R.Q., Dong, S., Zhao, S.*, et al.* (2019). Increased *TBX6* gene dosages induce congenital cervical vertebral malformations in humans and mice. J Med Genet.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Sawyer, S.L., Hartley, T., Dyment, D.A., Beaulieu, C.L., Schwartzentruber, J., Smith, A., Bedford, H.M., Bernard, G., Bernier, F.P., Brais, B.*, et al.* (2016). Utility of whole-exome sequencing for those near the end of the diagnostic odyssey: time to address gaps in care. Clin Genet *89*, 275-284.

Schaffer, A.E., Eggens, V.R., Caglayan, A.O., Reuter, M.S., Scott, E., Coufal, N.G., Silhavy, J.L., Xue, Y., Kayserili, H., Yasuno, K.*, et al.* (2014). *CLP1* founder mutation links tRNA splicing and maturation to cerebellar development and neurodegeneration. Cell *157*, 651-663.

Scott, E.M., Halees, A., Itan, Y., Spencer, E.G., He, Y., Azab, M.A., Gabriel, S.B., Belkadi, A., Boisson, B., Abel, L.*, et al.* (2016). Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. Nat Genet *48*, 1071-1076.

Seelow, D., Schuelke, M., Hildebrandt, F., and Nurnberg, P. (2009). HomozygosityMapper--an interactive approach to homozygosity mapping. Nucleic Acids Res *37*, W593-599.

Smith, C.A.B. (1953). The Detection of Linkage in Human Genetics. Journal of the Royal Statistical Society Series B (Methodological) *15*, 153-192.

Sobreira, N., Schiettecatte, F., Boehm, C., Valle, D., and Hamosh, A. (2015). New tools for Mendelian disease gene identification: PhenoDB variant analysis module; and GeneMatcher, a web-based tool for linking investigators with an interest in the same gene. Hum Mutat *36*, 425-431.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Spence, J.E., Perciaccante, R.G., Greig, G.M., Willard, H.F., Ledbetter, D.H., Hejtmancik, J.F., Pollack, M.S., O'Brien, W.E., and Beaudet, A.L. (1988). Uniparental disomy as a mechanism for human genetic disease. Am J Hum Genet *42*, 217-226.

Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G.*, et al.* (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science *337*, 64-69.

Torkamani, A., Pham, P., Libiger, O., Bansal, V., Zhang, G., Scott-Van Zeeland, A.A., Tewhey, R., Topol, E.J., and Schork, N.J. (2012). Clinical implications of human population differences in genome-wide rates of functional genotypes. Front Genet *3*, 211.

Tuncbilek, E., and Koc, I. (1994). Consanguineous marriage in Turkey and its impact on fertility and mortality. Ann Hum Genet *58*, 321-329.

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

Worthey, E.A., Mayer, A.N., Syverson, G.D., Helbling, D., Bonacci, B.B., Decker, B., Serpe, J.M., Dasu, T., Tschannen, M.R., Veith, R.L.*, et al.* (2011). Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. Genet Med *13*, 255-262.

Wu, N., Ming, X., Xiao, J., Wu, Z., Chen, X., Shinawi, M., Shen, Y., Yu, G., Liu, J., Xie, H.*, et al.* (2015). *TBX6* null variants and a common hypomorphic allele in congenital scoliosis. N Engl J Med *372*, 341-350.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Yang, N., Wu, N., Zhang, L., Zhao, Y., Liu, J., Liang, X., Ren, X., Li, W., Chen, W., Dong, S.*, et al.* (2019). *TBX6* compound inheritance leads to congenital vertebral malformations in humans and mice. Hum Mol Genet *28*, 539-547.

Yang, Y., Muzny, D.M., Reid, J.G., Bainbridge, M.N., Willis, A., Ward, P.A., Braxton, A., Beuten, J., Xia, F., Niu, Z.*, et al.* (2013). Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N Engl J Med *369*, 1502-1511.

Yang, Y., Muzny, D.M., Xia, F., Niu, Z., Person, R., Ding, Y., Ward, P., Braxton, A., Wang, M., Buhay, C.*, et al.* (2014). Molecular findings among patients referred for clinical whole-exome sequencing. JAMA *312*, 1870-1879.

Yuan, B., Pehlivan, D., Karaca, E., Patel, N., Charng, W.L., Gambin, T., Gonzaga-Jauregui, C., Sutton, V.R., Yesil, G., Bozdogan, S.T.*, et al.* (2015). Global transcriptional disturbances underlie Cornelia de Lange syndrome and related phenotypes. J Clin Invest *125*, 636-651.

Zollo, M., Ahmed, M., Ferrucci, V., Salpietro, V., Asadzadeh, F., Carotenuto, M., Maroofian, R., Al-Amri, A., Singh, R., Scognamiglio, I.*, et al.* (2017). *PRUNE* is crucial for normal brain development and mutated in microcephaly with neurodevelopmental impairment. Brain *140*, 940-952.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure Legends

Figure 1 - Comparison of the TK cohort variome to control databases. A) Principal component analysis (PCA) compares the population structure among the TK and non-TK cohorts of the BHCMG cohort, the African, Asian, and European population samples from the 1000 Genomes Project. **B)** We observed that 23.4%, 17.8% and 69.1% of the unique variants (N=356,613) in the TK genomes were not present in ExAC, gnomAD and the GME variomes, respectively

Figure 2- Phenotypic description of the study cohort. A) The affection status of a suspected Mendelian disease trait observed in Turkish (green) and non-Turkish (orange) subjects. **B)** The family structure of each sequenced individual. **C)** TK patients were clustered into four general phenotypic categories or groups by calculating a pairwise phenotypic similarity score.

Figure 3- Estimation of consanguinity levels from both self-reported and proband genotype data. A) The column plot shows the number of samples in the Turkish (TK) and non-Turkish (non-TK) cohort at different degrees of self-reported consanguinity levels with records available in the PhenoDB database; the TK cohort revealed 26.3% (147/448 families); whereas, the non-TK cohort, consisting of mostly North American families was 3% (57/1,911 families). **B)** The box plots report the estimated inbreeding coefficient levels (F) calculated from ES data. Both of the TK affecteds and unaffecteds showed higher F values on average compared to the non-TK affecteds (P=1.5e-12) and non-TK unaffecteds (0.00014). On the other hand, there was no significant difference noted between F values of the TK affecteds and unaffecteds

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(P=0.58) and the non-TK affecteds and unaffecteds (P=1). **C)** Box plots show the number of homozygous variants not represented at all in the gnomAD database (y-axis). Exomes of TK affecteds and unaffecteds carry an average number of 5.35 and 3.64 homozygous variants respectively, not represented at all in the gnomAD database; whereas, non-TK affecteds and unaffecteds on average harbor 3.01 and 2.80 homozygous variants not found in the gnomAD database.

Figure 4- The enrichment of long-sized ROH regions in Turkish affected individuals. We characterized the features of long-sized ROH regions on **A)** the total length of long-sized ROH regions, **B)** the median length of long-sized ROH regions, and **C)** the number of long-sized ROH blocks. In each panel, we showed the difference among the four individual groups: Turkish (TK) unaffected, non-Turkish (non-TK) unaffected, TK affected, and non-TK affected individuals. We performed a one-sided Wilcoxon rank sum test to compare TK versus non-TK individuals, either affected or unaffected, and marked the P-value significance level on the top of each pair of groups compared (*<0.05, **<0.01, ***<0.001, ****<0.0001). Outliers are not shown in the boxplots.

Figure 5- A mutational burden of rare homozygous deleterious variants maps to long-sized ROH regions. We calculated a variant density metric by normalizing the count of rare homozygous deleterious variants to the count of rare homozygous synonymous variants. We compared the density of variants **A)** that are located in ROH regions, **B)** not in ROH regions, **C)** in long-sized ROH regions. **D)** The subject group classification and comparison method are same as described in Figure 4. Outliers are not shown in the boxplots.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 6- The contribution of variants that are located in each size class of ROH blocks to the clinical phenotypes of patients. A) A diagram depicting the analytical workflow. **B)** The comparison between the Turkish and non-Turkish population with regard to the performance of variants located in each ROH size group in contribution to disease traits. The y-axis describes the ratios of z-scores calculated for variants located in a specific ROH size group, e.g., long-sized ROH group, versus z-scores calculated for all variants of interest.

STAR Methods

CONTACT FOR REAGENT AND RESOURCE SHARING

Requests for further information on raw data, genomic and phenotypic analyses, DNA samples may be directed to, and will be fulfilled by Lead Contact James R. Lupski (jlupski@bcm.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

We recruited 4,933 unrelated TK and non-TK individuals in the BHCMG cohort (data freeze: July 2018) After all relevant family members provided written informed consent for the use of their DNA in identification of disease variants and for broad data sharing. Peripheral blood was collected from affected individuals, parents and unaffected relatives if available. Genomic DNA was extracted from blood leukocytes according to standard procedures. All genomic studies were performed on DNA samples isolated from blood. This study was approved by the Institutional Review Board at Baylor College of Medicine (IRB protocol # H-29697).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

METHOD DETAILS

ES and annotation

ES was performed at the Human Genome Sequencing Center (HGSC) at Baylor College of Medicine through the BHHopkins Center for Mendelian Genomics (BHCMG) initiative. With 1ug of DNA an Illumina paired-end pre-capture library was constructed according to the manufacturer's protocol (Illumina Multiplexing_SamplePrep_Guide_1005361_D) with modifications as described in the BCM-HGSC Illumina Barcoded Paired-End Capture Library Preparation protocol. Pre-capture libraries were captured into 4-plex library pools and hybridized in solution to the HGSC-designed Core capture reagent (52Mb, NimbleGen) or 6-plex library pools with the custom VCRome 2.1 capture reagent (42Mb, NimbleGen) according to the manufacturer's protocol (NimbleGen SeqCap EZ Exome Library SR User's Guide) with minor revisions. The sequencing run was performed in paired-end mode with the Illumina HiSeq 2000 platform, with sequencing-by-synthesis reactions extended for 101 cycles from each end and an additional 7 cycles for the index read. With a sequencing yield of 11 Gb, the sample achieved 92% of the targeted exome bases covered to a depth of 20X or greater. Illumina sequence analysis was performed with the HGSC Mercury analysis pipeline (https://www.hgsc.bcm.edu/software/mercury) (Challis et al., 2012; Reid et al., 2014), which moves data through various analysis tools from the initial sequence generation to annotated variant calls (SNPs and intra-read

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

insertion/deletions; i.e. indels). Variants were called with the ATLAS2 variant calling method and the Sequence Alignment/Map (SAMtools) suites and annotated with an in-house-developed Cassandra annotation pipeline that uses Annotation of Genetic Variants (ANNOVAR) (Wang et al., 2010) and additional tools and databases including ExAC (http://exac.broadinstitute.org), gnomAD (https://gnomad.broadinstitute.org) and *the GME variome* (http://igm.ucsd.edu/gme/) and the ARIC database (http://drupal.cscc.unc.edu/aric/).

Variant discovery in the BHCMG cohort

The analysis was implemented on 72.3 terabases of ES data from 6,571 individual genome samples available in the BHCMG database (data freeze: July 2018). The study participants presented with either detailed phenotypic features recorded in PhenoDB (Hamosh et al., 2013; Sobreira et al., 2015) or affected status information captured from the internal laboratory database. Accordingly, 6,571 samples originated from 3,682 subjects with diverse clinical phenotypes (defined as affecteds), 2,083 unaffecteds (no apparent clinical disease), and 806 samples with an unknown affection status. These data were generated with two different capture technologies, the HGSC-designed Core capture reagent (3,016 samples) and the custom VCRome 2.1 capture reagent (3,555 samples). Thus, we parsed the data and retrieved the variants from the intersection of both designs resulting in a final dataset of 1,340,496 distinct single-nucleotide variants (SNVs).

There were no evident differences of capture reagent- and affected status-specifics in the number of unfiltered SNVs extracted from VCF files with a median number of 13,817

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

heterozygous SNVs and 8,015 homozygous SNVs (Supplementary Figure 5), or with the median number of homozygous vs. heterozygous SNV ratio as 0.58 and transition (Ti) to transversion (Tv) SNV ratio as 3.01 per exome, which was within the expected range of 3-3.3 from previous studies (12) (Supplementary Figure 6).

Phenotypic characterization of the BHCMG cohort

For this BHCMG project study, we utilized the PhenoDB database to collect and store the information of clinical features, pedigree structures, and self-reported consanguinity levels based on referring physicians' clinical evaluations of research subjects and their unaffected family members. Computational analyses of phenotyping data was performed using HPO terms and essentially as described in (James et al., 2016; Liu et al., 2019b; Posey et al., 2017).

Obtaining a final set of unrelated individuals in the BHCMG cohort

To minimize overrepresentation of individuals from the same family, we calculated the kinship coefficient of relatedness between each pair of individuals in the study cohort. Removing individuals with a close relationship from the analysis by a threshold of kinship coefficient as 0.044 (third-degree relationship or more) resulted in a final set of 4,933 unrelated individuals: 3,121 unrelated affecteds (556 unrelated TK affecteds), and 1,812 unrelated unaffecteds (297 unrelated TK unaffecteds), who presented with 1,172,271 distinct SNVs in total (Supplementary Figure 7).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Principal Component Analysis (PCA)

We investigated the population structure of TK cohort along with non-TK cohort in the BHCMG and the African, Asian, and European population samples from the 1000 Genomes Project phase 3 release data including 2,504 individuals' genotypes in total (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/). First, genotypes for 31 individuals who have blood relationship with the 2504 samples were removed from the analysis. Variants falling out of the HGSC-designed Core capture design and VCRome 2.1 capture design in the 1000 Genomes Project data were further filtered out from the analysis. A pruned subset of the remaining polymorphic SNVs that is in approximate linkage equilibrium of each other (N=58,860) was used for the PCA. PCA was performed using the SNPRelate R package.

Estimation of inbreeding coefficient values

The coefficient of inbreeding of an individual represents the probability that two alleles at any randomly chosen locus in an individual are identical-by-descent.

The inbreeding coefficient values of the BHCMG cohort were estimated from the ES data using PC-Relate method provided in the GENESIS R Package.

Identifying and analyzing ROH segments from ES data

We detected ROH regions from unphased exome sequence (ES) data with BafCalculator (https://github.com/ BCM-Lupskilab/BafCalculator) (Karaca et al., 2018). BafCalculator calculates genomic intervals with AOH from unphased ES data as a surrogate measure of ROH. VCF files were used to identify regions of AOH using the

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

following algorithm: first, from all SNVs that passed quality filters in the single VCF, we extracted a B-allele frequency (i.e., variant reads/total reads ratio); next, we transformed this ratio by subtracting 0.5 and taking the absolute value for each data point. After such a transformation, values > 0.45 were considered indicative of homozygous or hemizygous variants (expected value is 0.5) corresponding to either alternative or reference alleles, whereas lower values likely indicate heterozygous alleles. Transformed B-allele frequency data were then processed using Circular Binary Segmentation (CBS) implemented in the DNAcopy R Bioconductor package (Huber et al., 2015; Olshen et al., 2004). In summary, segments with the mean signal > 0.45 and size >1 Kb were classified as AOH regions. The calculated AOH intervals from BafCalculator could represent individual genomic/gene loci resulting in ROH for diploid alleles that can occur by: i) IBD, ii) UPD (Spence et al., 1988), or iii) a large deletion CNV. To exclude the ROH blocks that could be caused by genomic overlapping of a deletion CNV, we first identified deletion CNVs through XHMM (Fromer and Purcell, 2014). We further intersected ROH segments and potential deletion CNVs with BEDTools (Quinlan, 2014), and retained only ROH regions overlapping less than 50% of their size with a variant deletion CNV. We grouped ROH regions (N = 1,469,298) into three size categories using the same parameters as the published values (6): long genomic intervals or ROH blocks (> 1,606 Kb), medium ROH blocks (515 Kb-1,606 Kb), and short ROH blocks (40 Kb-515 Kb).

To control for the variable mutation rates across different genomic regions and among different individuals, for each individual i and ROH region category $r, r \in \{total -$

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$AOH, non-AOH, long-sized\ AOH, medium-sized\ AOH, short-sized\ AOH\}$, we computed a variant density $f_{i,r}$

$$f_{i,r} = \frac{N_{i,r}^{d}}{N_{i,r}^{s}}$$

in which $N_{i,r}^{d}$ is the count of rare homozygous deleterious or likely damaging variant alleles (variants above a certain CADD score (>=15) or by utilizing a prediction tool algorithm, NMDescPredictor (Coban-Akdemir et al., 2018)) that are located within a region r, while $N_{i,r}^{s}$ is the count of synonymous variants in a region r.

Generation of an objective score for phenotypic similarity comparison

We applied two R packages, OntologyIndex and ontologySimilarity, to measure the phenotypic similarity between two sets of human phenotype ontology (HPO) terms; each set of terms was associated with a patient's clinical features recorded in PhenoDB (https://phenodb.org) (Hamosh et al., 2013; Sobreira et al., 2015). To assess the ability of a disease gene to explain a patient's clinical features, we applied a previously published method, in which first a MICA (most informative common ancestor) matrix was calculated for each pair of HPO terms and then a Resnik score was calculated for two sets of HPO terms (James et al., 2016; Liu et al., 2019b; Posey et al., 2017). First, we calculated the information content (IC) for HPO term t by

$$IC(t) = -\log(f(t))$$

where $f(t)$ is the frequency of term t observed in all of the OMIM entries. The similarity of term i and term j is calculated with a Resnik method:

$$R_{ij} = IC(MICA(t_i, t_j))$$

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

where the Resnik score, R_{ij}, is determined by the IC of the most information common ancestor (MICA) of term i and term j. Next, we defined the phenotypic similarity score Sim for two HPO sets $l_1$and l_2 as follows:

$$l_1 = \{t_{11}, t_{12}, \dots, t_{1m}\}$$

$$l_2 = \{t_{21}, t_{22}, \dots, t_{2n}\}$$

$$Sim(l_1, l_2) = \frac{1}{2}\left(\frac{1}{m}\sum_{i=1}^{m} \max_{1 \le j \le n} R_{ij}, \frac{1}{n}\sum_{j=1}^{n} \max_{1 \le i \le n} R_{ij}\right)$$

The main features of known human disease genes were summarized in OMIM (https://www.omim.org/) in the format of both plain texts and clinical synopses, and the associated HPO terms for each OMIM entry were manually annotated by the HPO-team (https://hpo.jax.org/app/download/annotation). We further adapted this method to measure the phenotypic similarity between a patient's phenotypes and a list of disease genes. For each list of the tested disease genes, we calculated a z-score performing 1,000 simulations; in each simulation we computed a similarity score between the patient's clinical features and the associated HPO terms of a randomly selected disease gene list, which has a same number of genes as the tested disease gene list. Furthermore, to compare the contribution of disease gene lists across different genomic regions to the explanation of a patient's clinical phenotypic features, we computed a ratio of the similarity score calculated for a subset of disease genes, e.g., genes located in long-sized ROH regions to the similarity score calculated for all of the associated disease genes in a patient's genome.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Statistical Analyses

We performed the statistical analyses with R version 3.3.3. We compared pairwise differences in the average values of estimated inbreeding coefficient values (F), homozygous rare deleterious variant burden (count and density) and total, median length and count of ROHs in 4 subject groups TK-unaffecteds, non-TK unaffecteds, TK-affecteds and non-TK affecteds using Wilcoxon-rank sum one-tailed test. The self-reported consanguinity levels were compared between the TK and non-TK cohort using Fisher's exact test. Box plots and violin plots were generated by ggplot2 data visualization R package and stat_compare_means function was used to add P-values and significance levels to those plots. Ddply function in plyr CRAN R package was used to report the summary statistics of estimated inbreeding coefficient values (F), homozygous rare deleterious variant burden (count and density) and total, median length and count of ROHs in 4 subject groups TK-unaffecteds, non-TK unaffecteds, TK-affecteds and non-TK affecteds.

DATA AND CODE AVAILABILITY

TK variome database is now available as a research and molecular diagnostic community resource (https://turkishvariomedb.shinyapps.io/tvdb/).

Supplemental Figure Titles and Legends

Figure S1. Principal component analysis compares the population structure among the TK and non-TK cohorts of the BHCMG cohort, the African, Asian, and European population samples from the 1000 Genomes Project. Plots show all combinations of PC1, PC2, PC3, and PC4 and percentages of variance explained.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure S2. The estimated F values were higher in the individuals with a historical self-report of consanguinity. The inbreeding coefficients estimated from genotype data of the BHCMG cohort samples were shown in accordance with their PhenoDB records**. A)** Violin plots show the estimated F values were higher in the individuals with a historical self-report of consanguinity compared to the ones without a report of consanguinity available in PhenoDB in the BHCMG cohort (Wilcoxon test one-tailed, $P < 2.2e^{-16}$). **B)** The correlation between the self-reported consanguinity level and the total length of ROH segments in each size category (long-sized, medium-sized, short-sized and total ROH). The self-reported consanguinity level captured for the BHCMG cohort samples in PhenoDB significantly correlates only with the total length of long-sized ROH segments (P-value 5.22e-10) but not correlated with the total length of medium-sized and short-sized ROH segments (P-values 0.138 and P-values 0.215).

Figure S3. **Characterization of ROH features in each size category. A)** Box plots show the total length of ROH segments (Mb) in each size category (long-sized, medium-sized, short-sized and total ROH) in 4 subject groups TK-unaffecteds, non-TK unaffecteds, TK-affecteds and non-TK affecteds. **B)** Box plots show the medium length of ROH segments (Mb) in each size category (long-sized, medium-sized, short-sized and total ROH) in 4 subject groups TK-unaffecteds, non-TK unaffecteds, TK-affecteds and non-TK affecteds. **C)** Box plots show the count of ROH segments in each size category (long-sized, medium-sized, short-sized and total ROH) in 4 subject groups TK-unaffecteds, non-TK unaffecteds, TK-affecteds and non-TK affecteds.

Figure S4. **Rare homozygous variant number and density the TK cohort individuals compared to the non-TK cohort individuals. A)** Violin plots display rare

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.27.064824; this version posted April 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(MAF <=0.05) homozygous deleterious (CADD PHRED-scaled score >=15) variant number/exome in TK-unaffecteds, non-TK unaffecteds, TK-affecteds and non-TK affecteds. Higher rare homozygous deleterious variant number/exome in the TK cohort were observed compared to the non-TK cohort in both affected and unaffected subgroups with P-values < 2.2e-16 and 2.2e-16 (Wilcoxon test one-tailed). **B)** The correlation between estimated *F* values and rare (MAF <=0.05) homozygous variant number/exome in the four groups of subjects. We calculated a variant density metric by normalizing the count of rare homozygous deleterious variants to the count of rare homozygous synonymous variants. We calculated the density of variants in medium-sized **C)** and short-sized ROH regions **D)** The subject group classification and comparison method are same as described in Figure 4. Outliers are not shown in the boxplots.

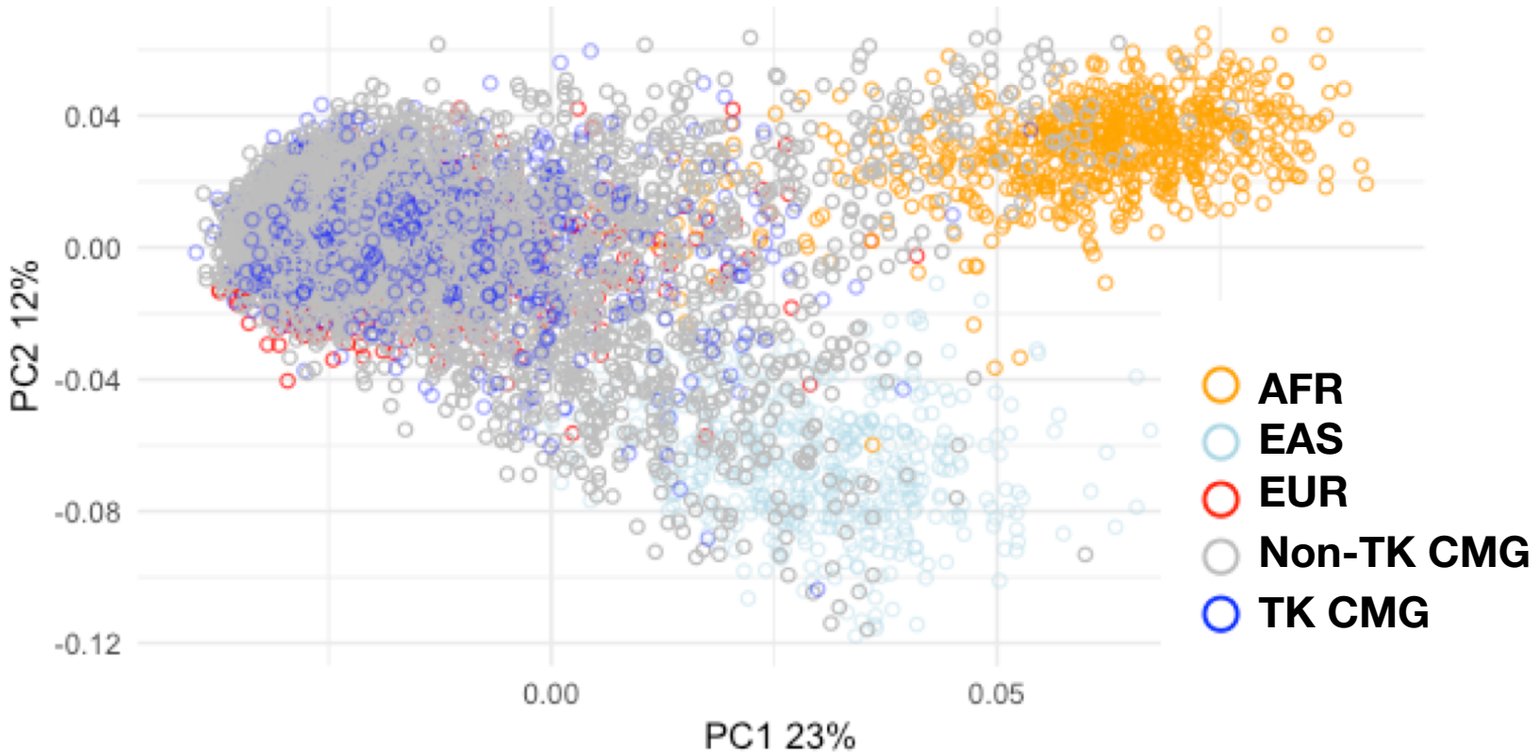
Figure S5. No evident differences of capture reagent- and affected status-specifics in the median number of heterozygous and homozygous SNVs in the BHCMG cohort extracted from unfiltered variant call format (VCF) files.

Figure S6. No evident differences of capture reagent- and affected status-specifics in the median number of heterozygous vs. homozygous SNV ratio and transition (Ti) to transversion (Tv) SNV ratio in the BHCMG cohort extracted from unfiltered variant call format (VCF) files.

Figure S7. The correspondence among the kinship coefficient of relatedness and familial relationship based on PhenoDB records between each pair of individuals in the BHCMG cohort.

Figure 1

A


0.04
0.00
-0.04
-0.08
-0.12
PC2 12%
0.00
0.05
PC1 23%
AFR
EAS
EUR
Non-TK CMG
TK CMG

B

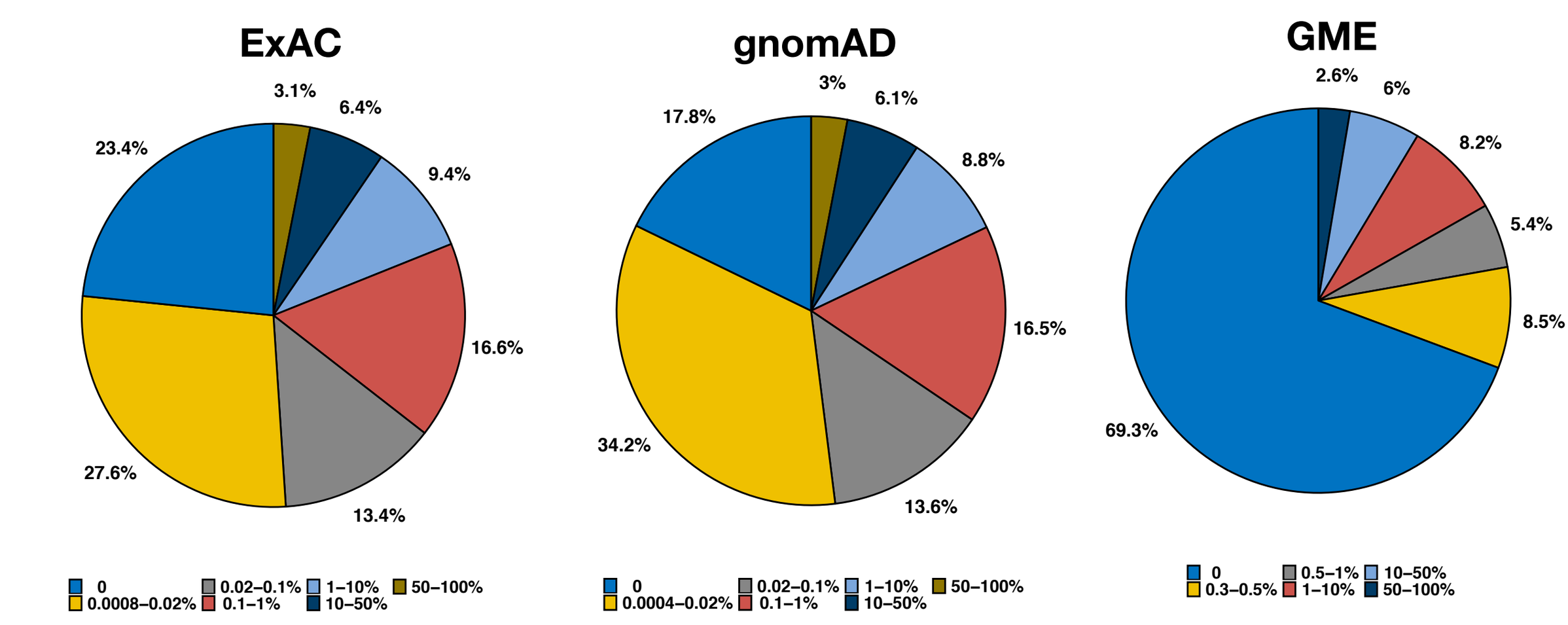
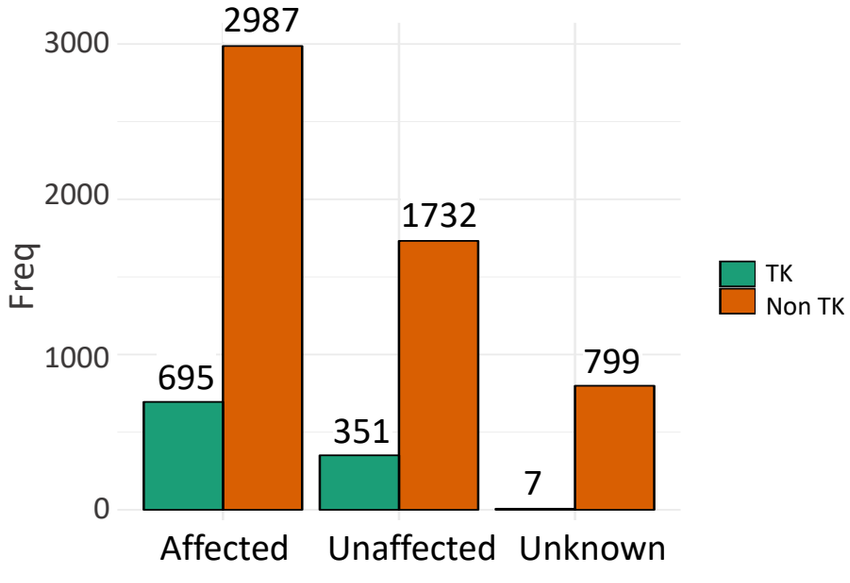
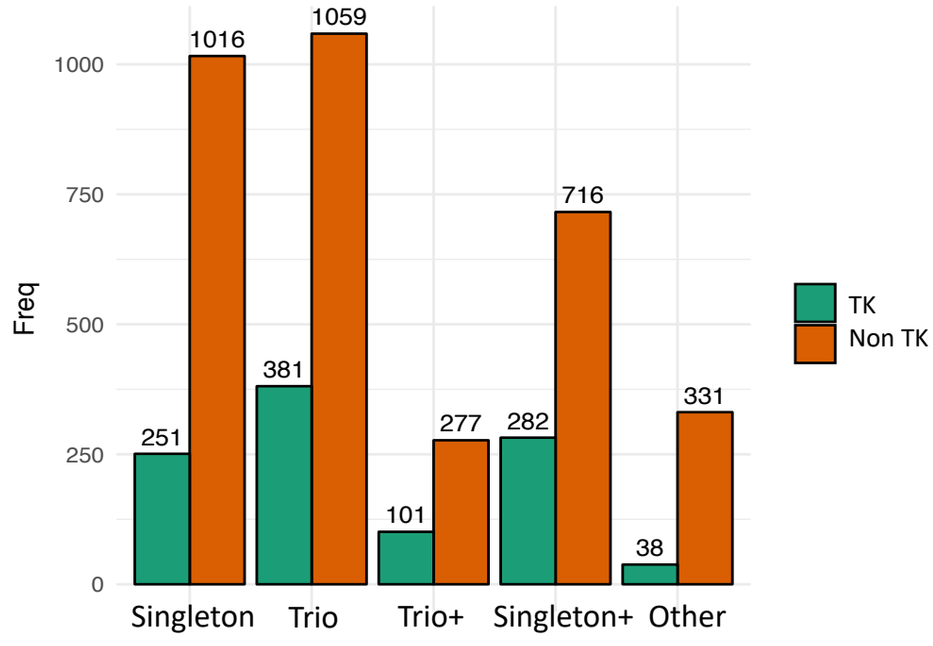

ExAC
3.1%
6.4%
9.4%
16.6%
13.4%
27.6%
23.4%
0
0.0008–0.02%
0.02–0.1%
0.1–1%
1–10%
10–50%
50–100%
gnomAD
3%
6.1%
8.8%
16.5%
13.6%
34.2%
17.8%
0
0.0004–0.02%
0.02–0.1%
0.1–1%
1–10%
10–50%
50–100%
GME
2.6%
6%
8.2%
5.4%
8.5%
69.3%
0
0.3–0.5%
0.5–1%
1–10%
10–50%
50–100%

Figure 2

A


3000
2000
1000
0
Freq
2987
1732
695
351
7
799
Affected
Unaffected
Unknown
TK
Non TK

B


1000
750
500
250
0
Freq
1016
1059
716
251
381
277
282
101
331
38
Singleton
Trio
Trio+
Singleton+
Other
TK
Non TK

C

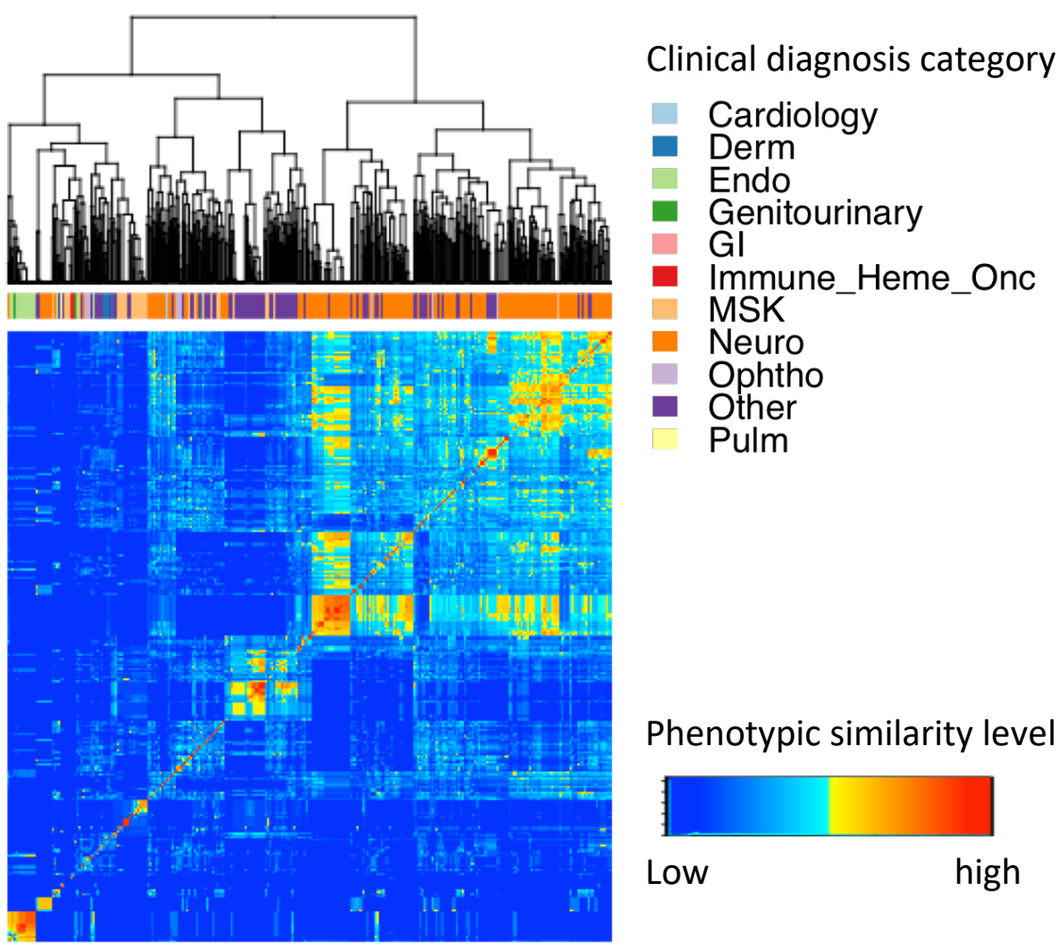
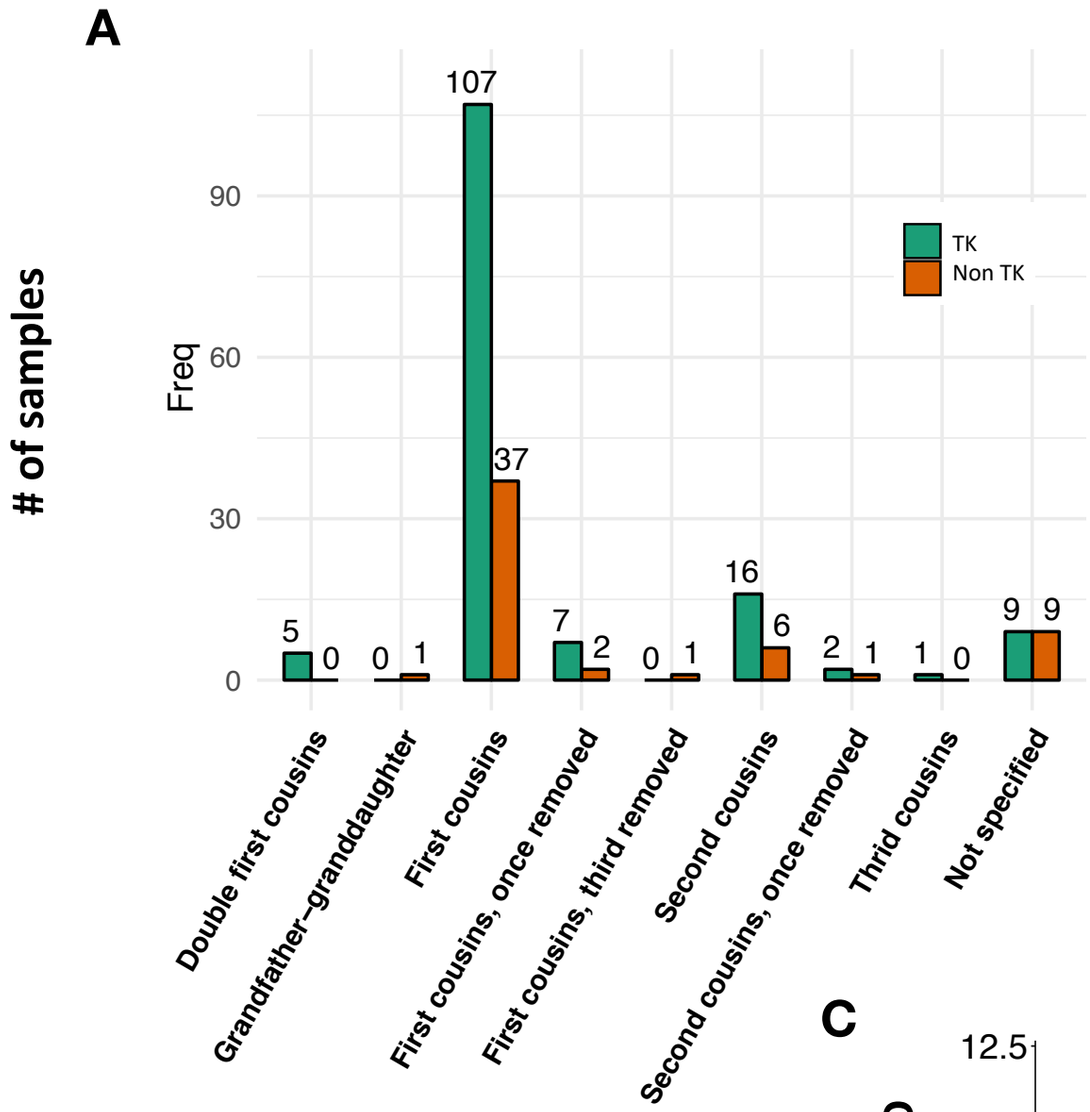
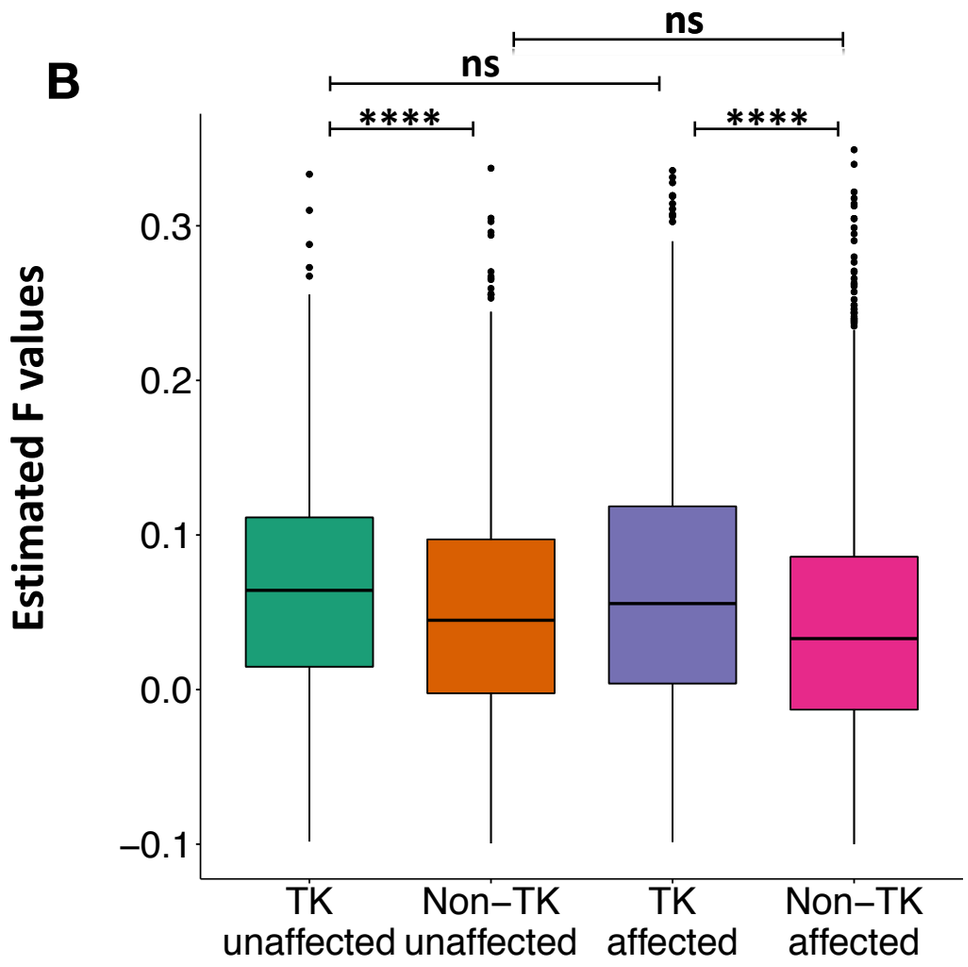

Clinical diagnosis category
Cardiology
Derm
Endo
Genitourinary
GI
Immune_Heme_Onc
MSK
Neuro
Ophtho
Other
Pulm
Phenotypic similarity level
Low
high

Figure 3


A
# of samples
Freq
90
60
30
0
107
37
5
0
0
1
7
2
0
1
16
6
2
1
1
0
9
9
TK
Non TK
Double first cousins
Grandfather–granddaughter
First cousins
First cousins, once removed
First cousins, third removed
Second cousins
Second cousins, once removed
Thrid cousins
Not specified


B
Estimated F values
ns
ns
****
****
0.3
0.2
0.1
0.0
−0.1
TK unaffected
Non−TK unaffected
TK affected
Non−TK affected

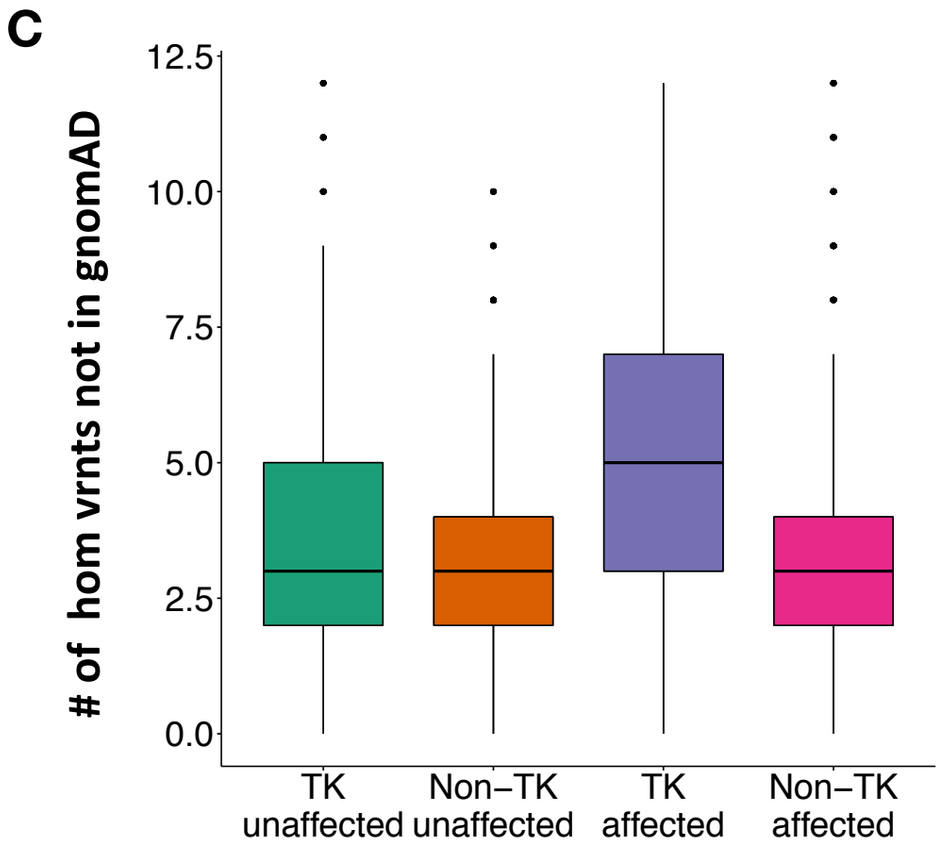
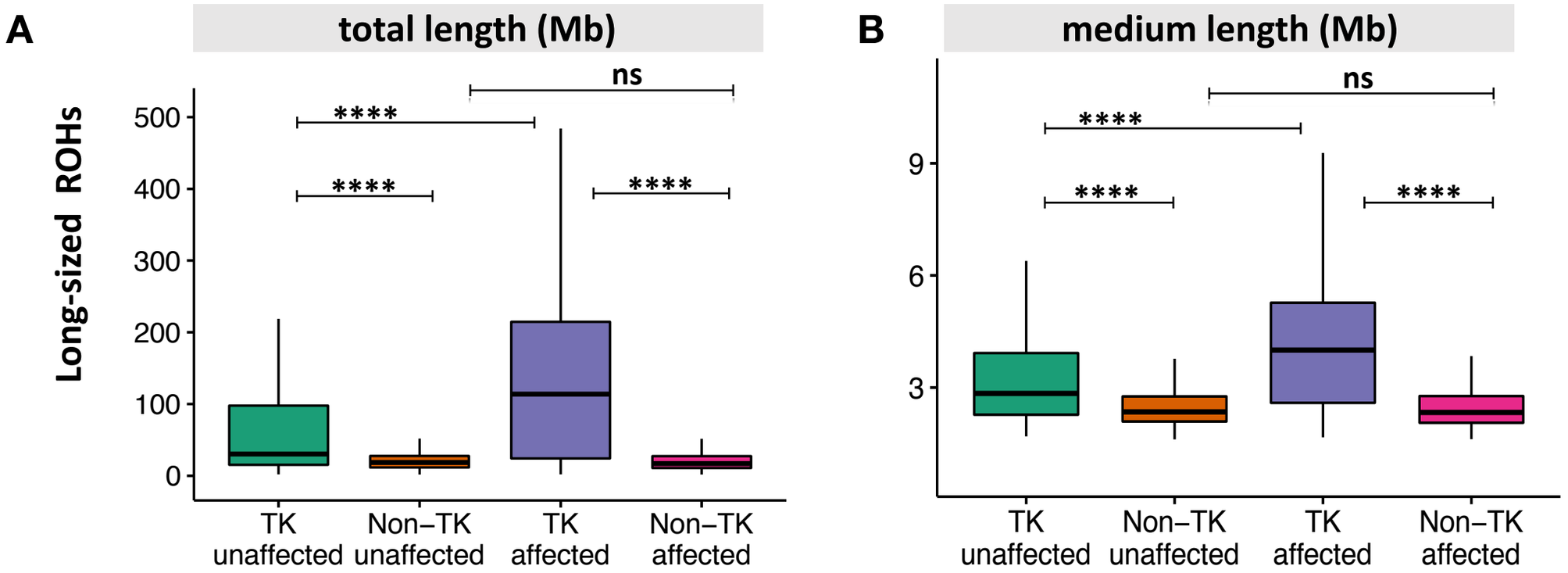

C
# of hom vrnts not in gnomAD
12.5
10.0
7.5
5.0
2.5
0.0
TK unaffected
Non−TK unaffected
TK affected
Non−TK affected

Figure 4


A
total length (Mb)
Long-sized ROHs
500
400
300
200
100
0
****
****
ns
****
TK unaffected
Non−TK unaffected
TK affected
Non−TK affected
B
medium length (Mb)
9
6
3
****
****
ns
****
TK unaffected
Non−TK unaffected
TK affected
Non−TK affected

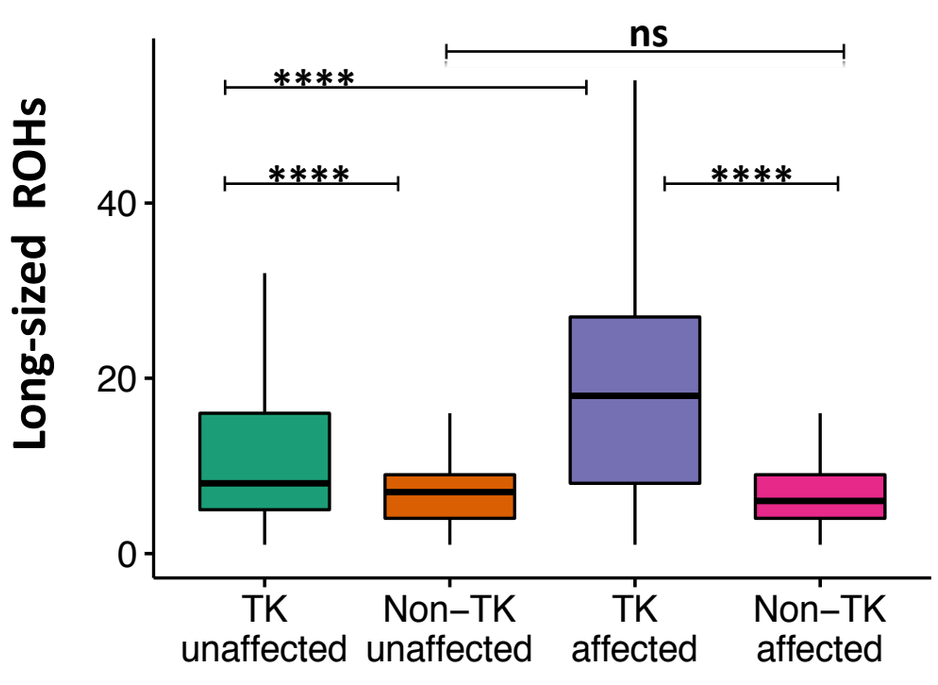

C
count
Long-sized ROHs
40
20
0
****
****
ns
****
TK unaffected
Non−TK unaffected
TK affected
Non−TK affected

Figure 5

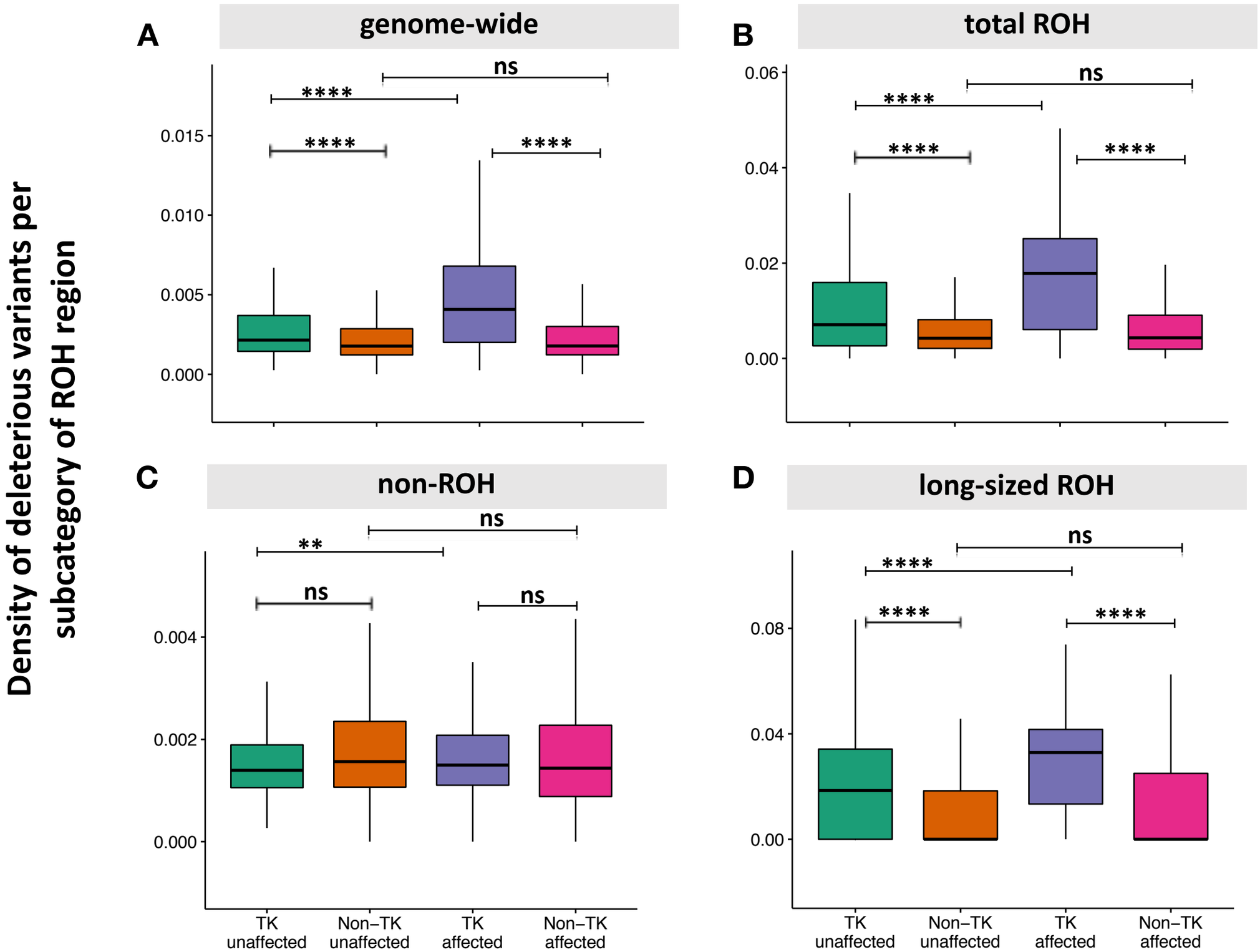
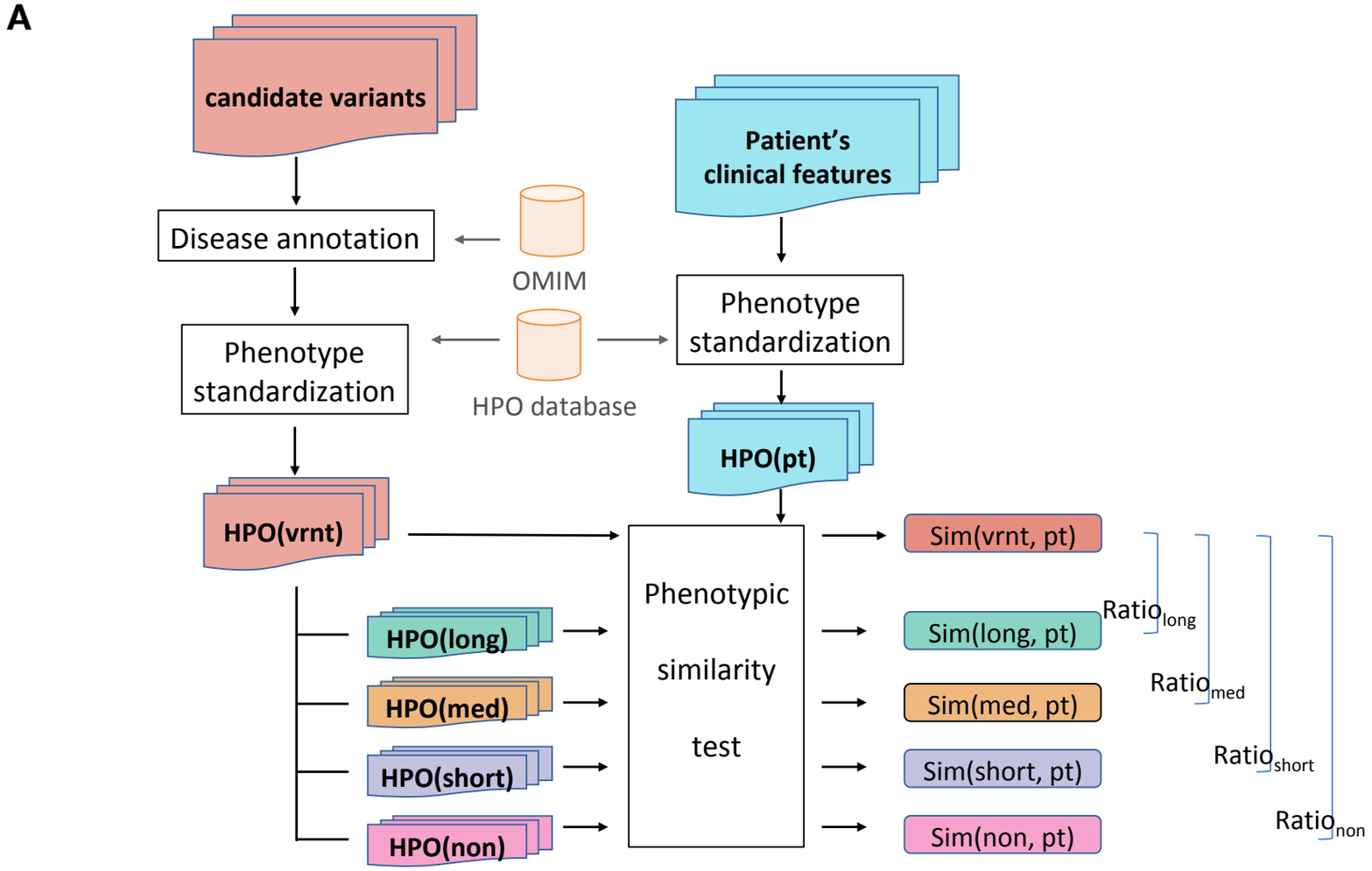
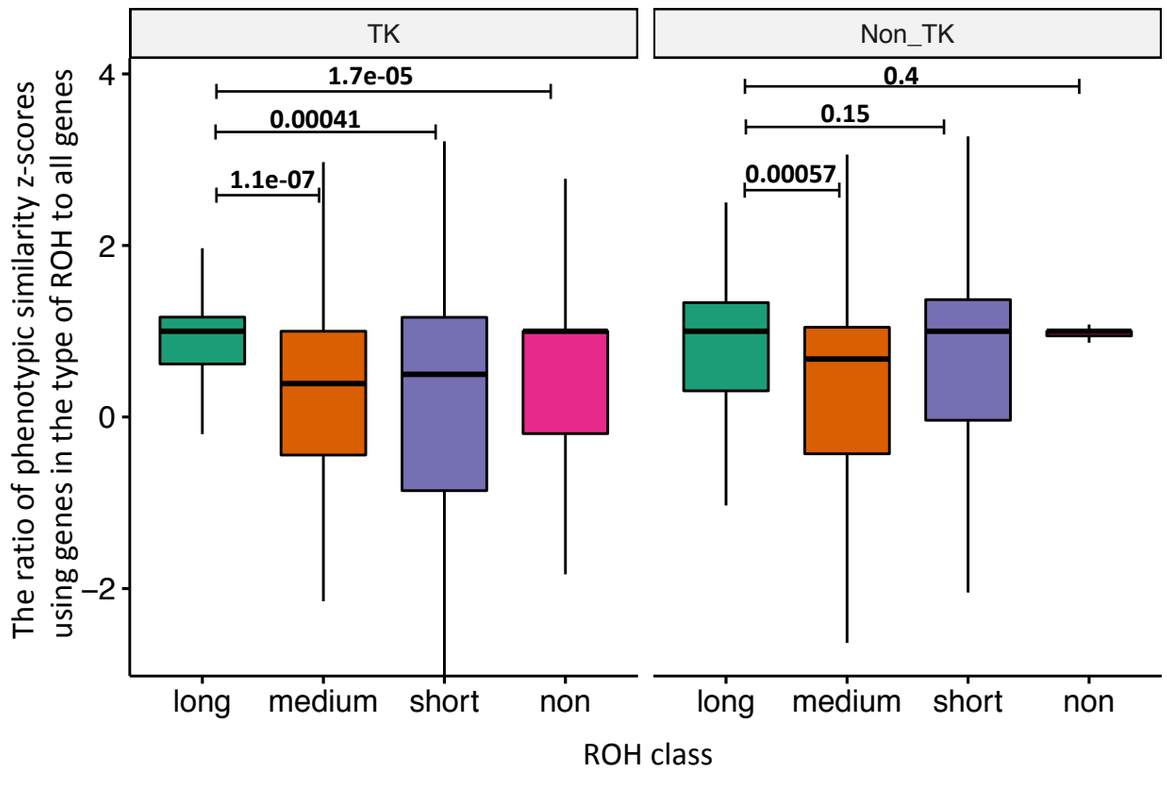

A
genome-wide
B
total ROH
C
non-ROH
D
long-sized ROH
Density of deleterious variants per subcategory of ROH region
ns
****
**
0.015
0.010
0.005
0.000
0.06
0.04
0.02
0.00
0.004
0.002
0.08
TK unaffected
Non−TK unaffected
TK affected
Non−TK affected

Figure 6

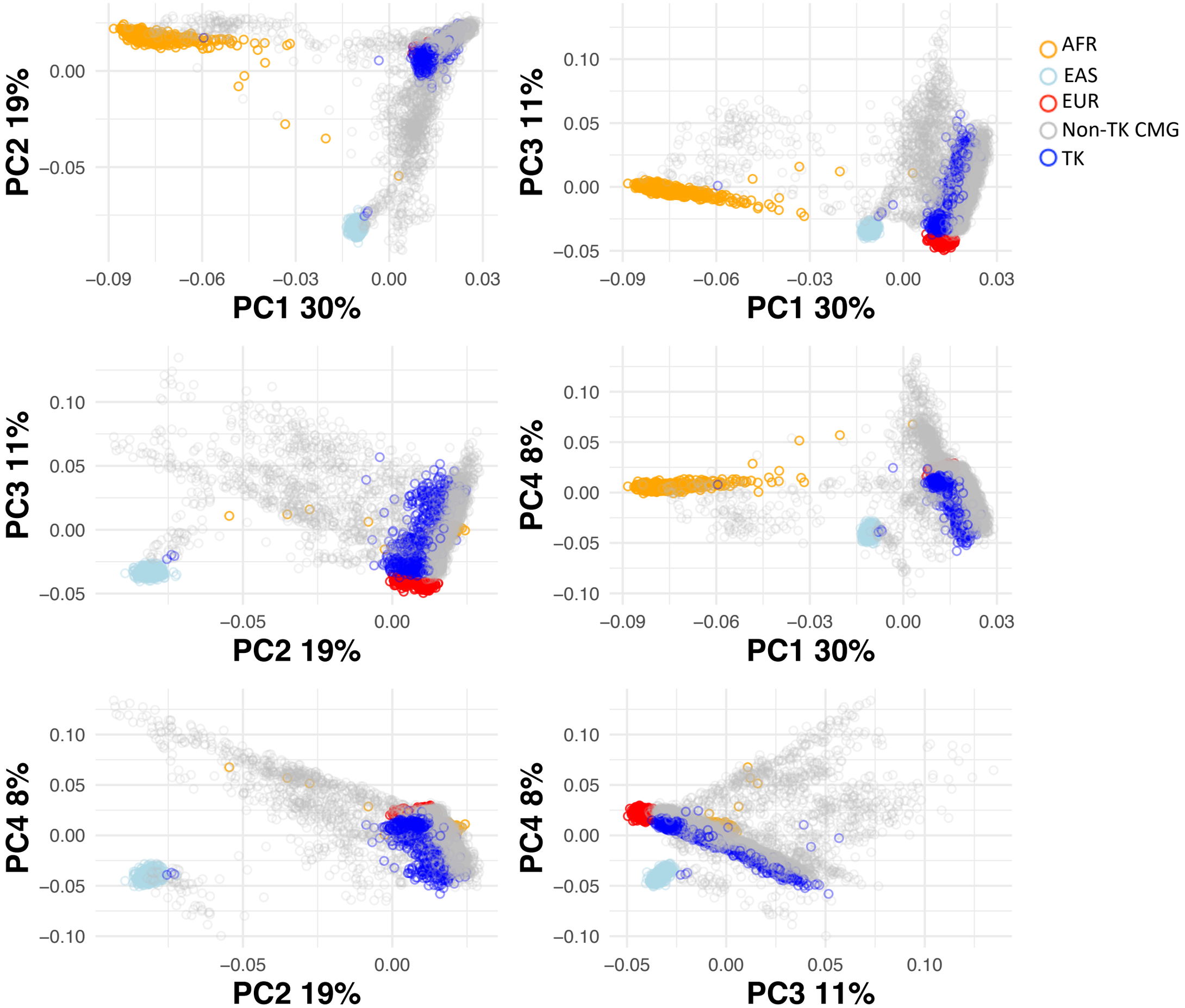
A


candidate variants
Patient's clinical features
Disease annotation
OMIM
Phenotype standardization
Phenotype standardization
HPO database
HPO(pt)
HPO(vrnt)
HPO(long)
HPO(med)
HPO(short)
HPO(non)
Phenotypic similarity test
Sim(vrnt, pt)
Sim(long, pt)
Sim(med, pt)
Sim(short, pt)
Sim(non, pt)
Ratio long
Ratio med
Ratio short
Ratio non

B

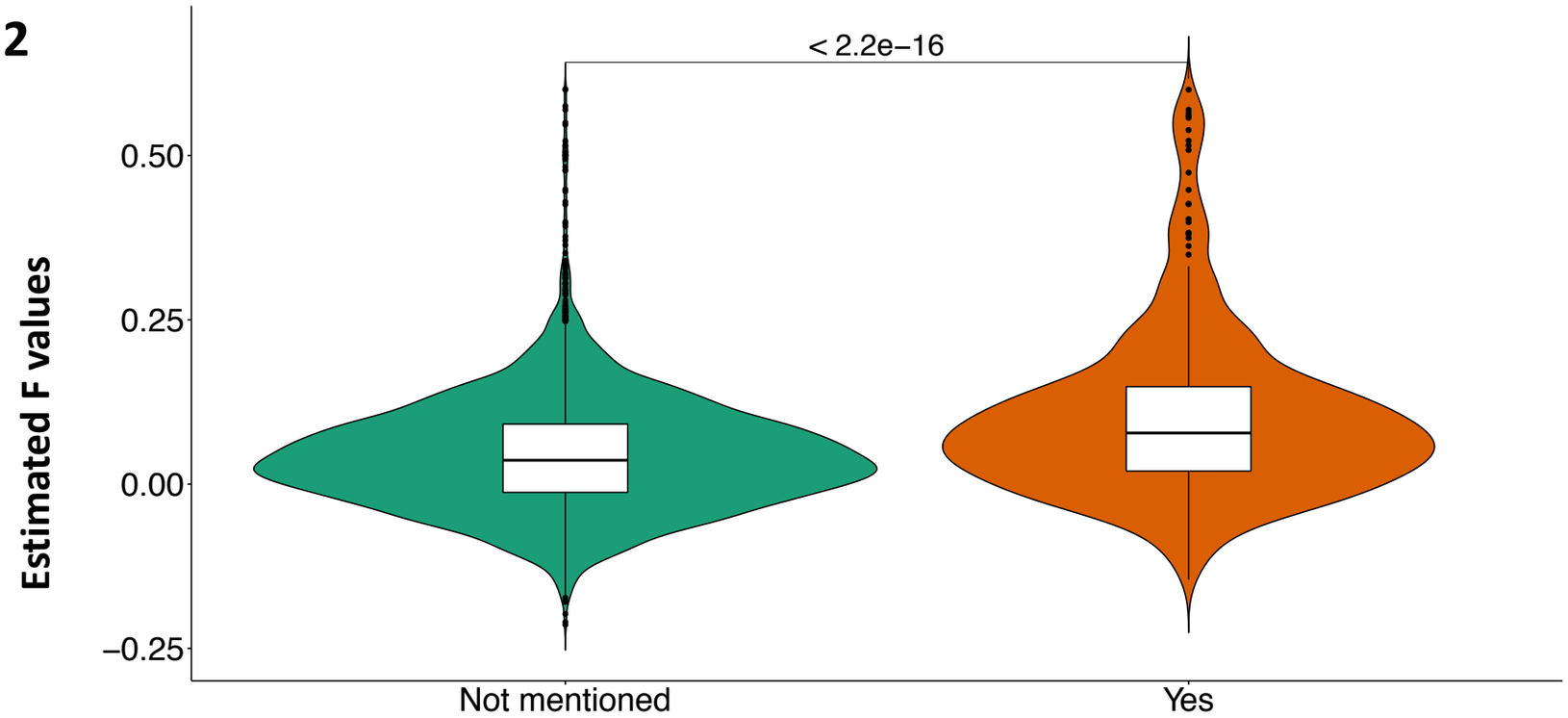

TK
Non_TK
1.7e-05
0.00041
1.1e-07
0.4
0.15
0.00057
The ratio of phenotypic similarity z-scores using genes in the type of ROH to all genes
4
2
0
−2
long
medium
short
non
ROH class

Supplementary Figure 1

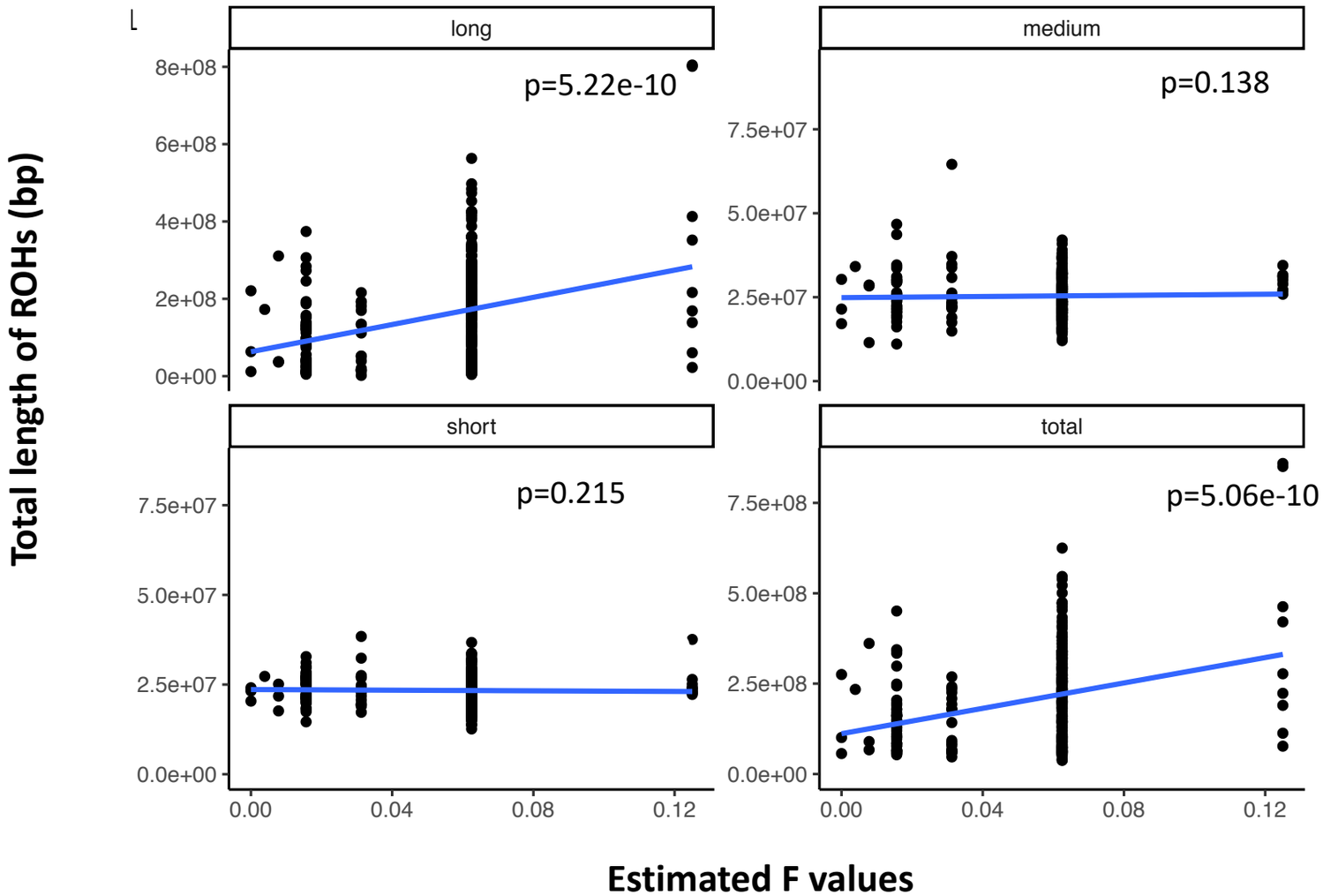


Supplementary Figure 2

A

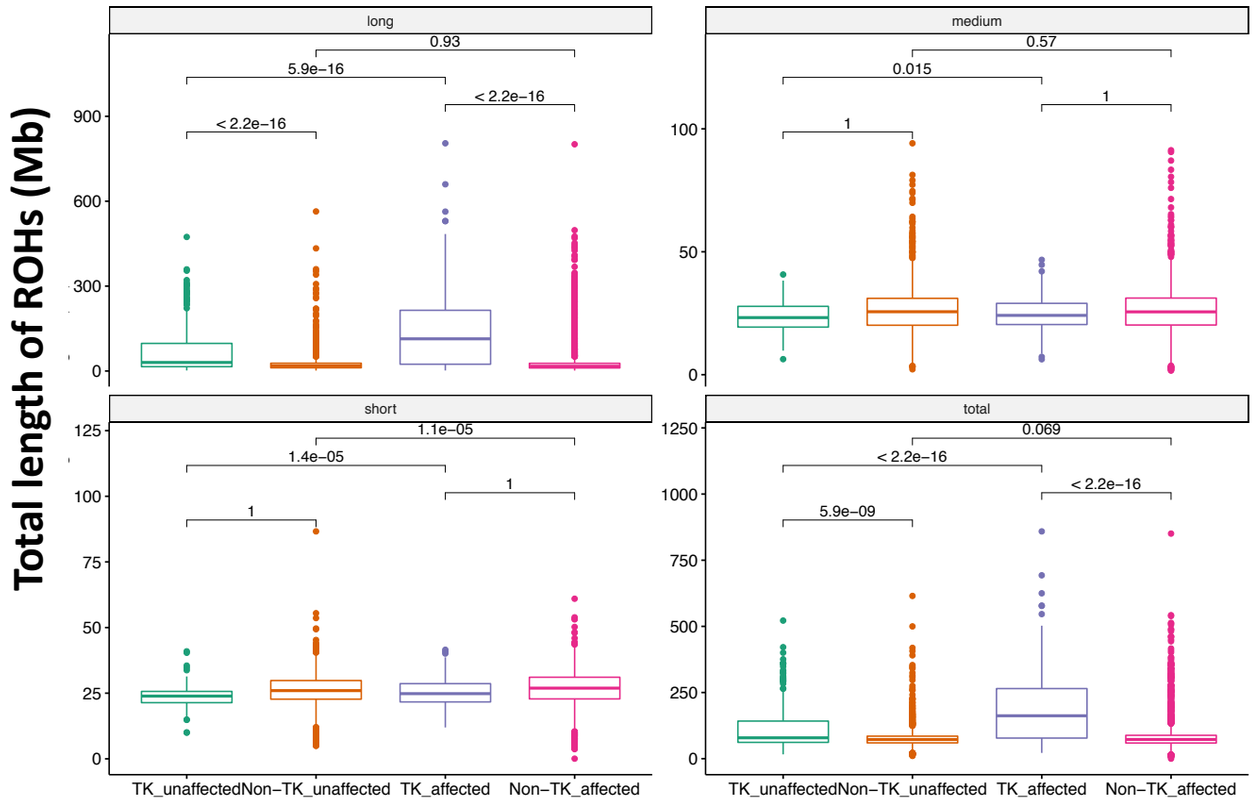


B

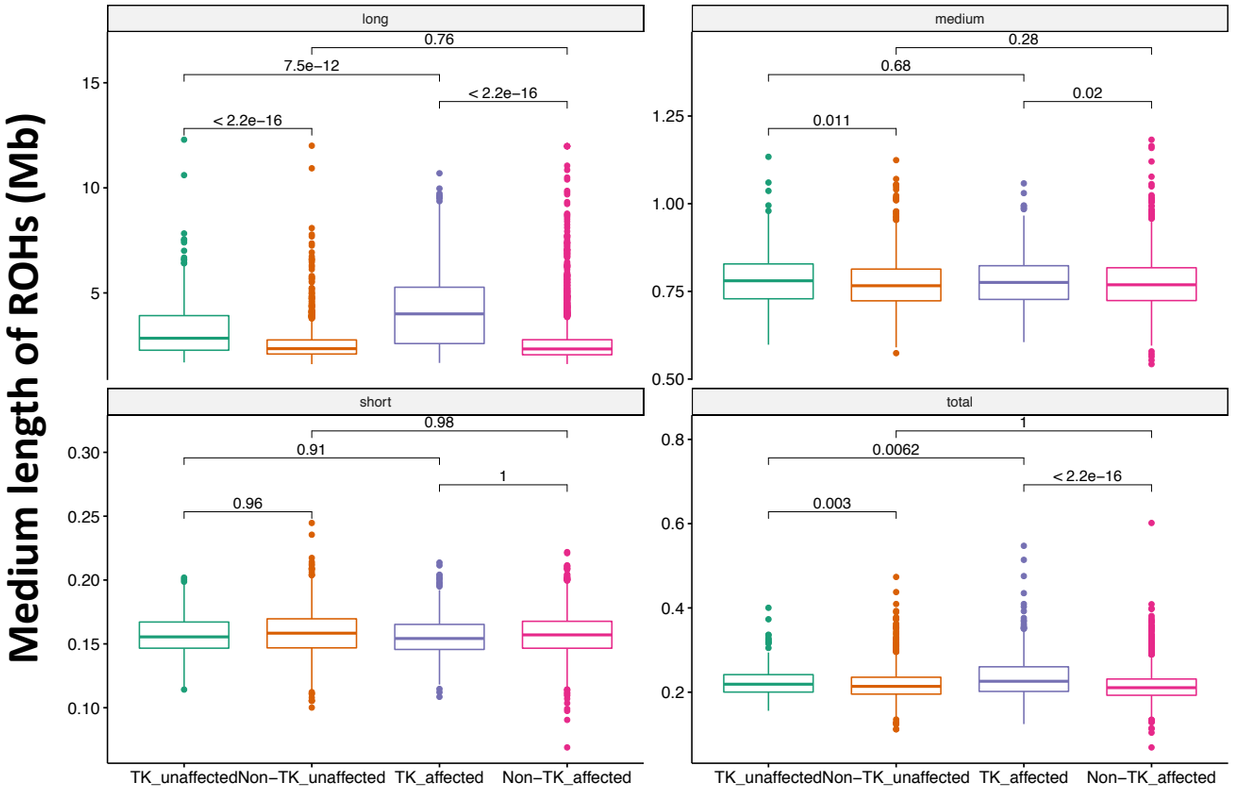


Supplementary Figure 3

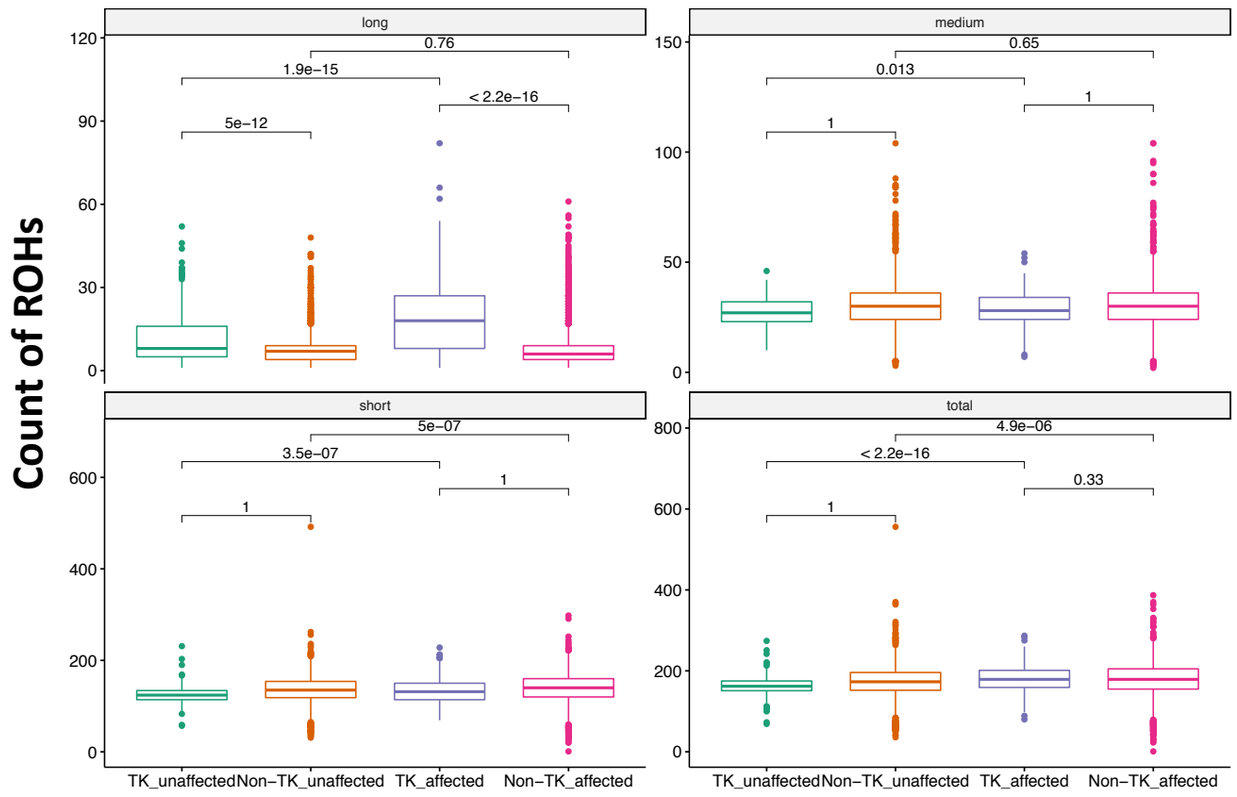
A



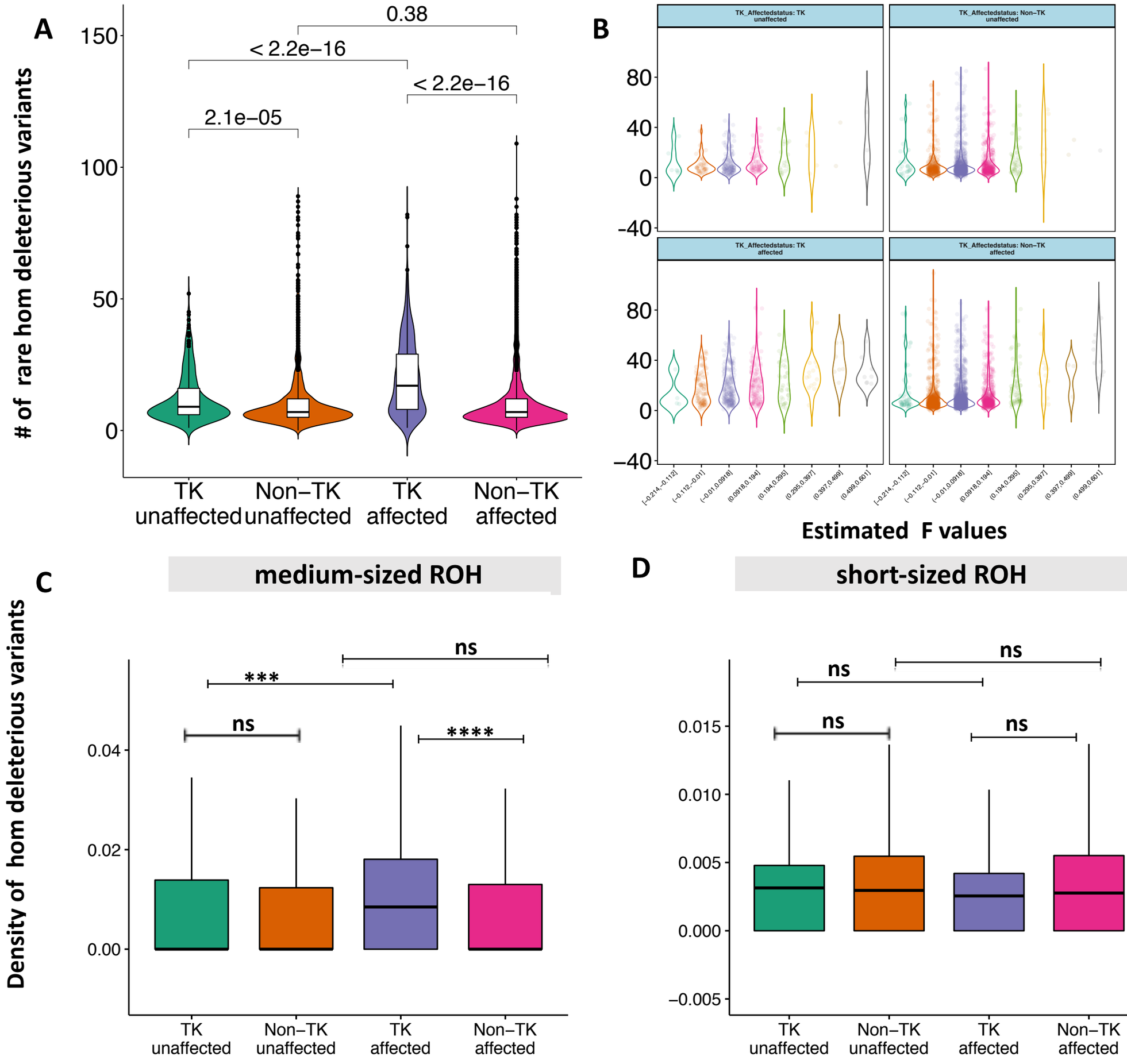
B



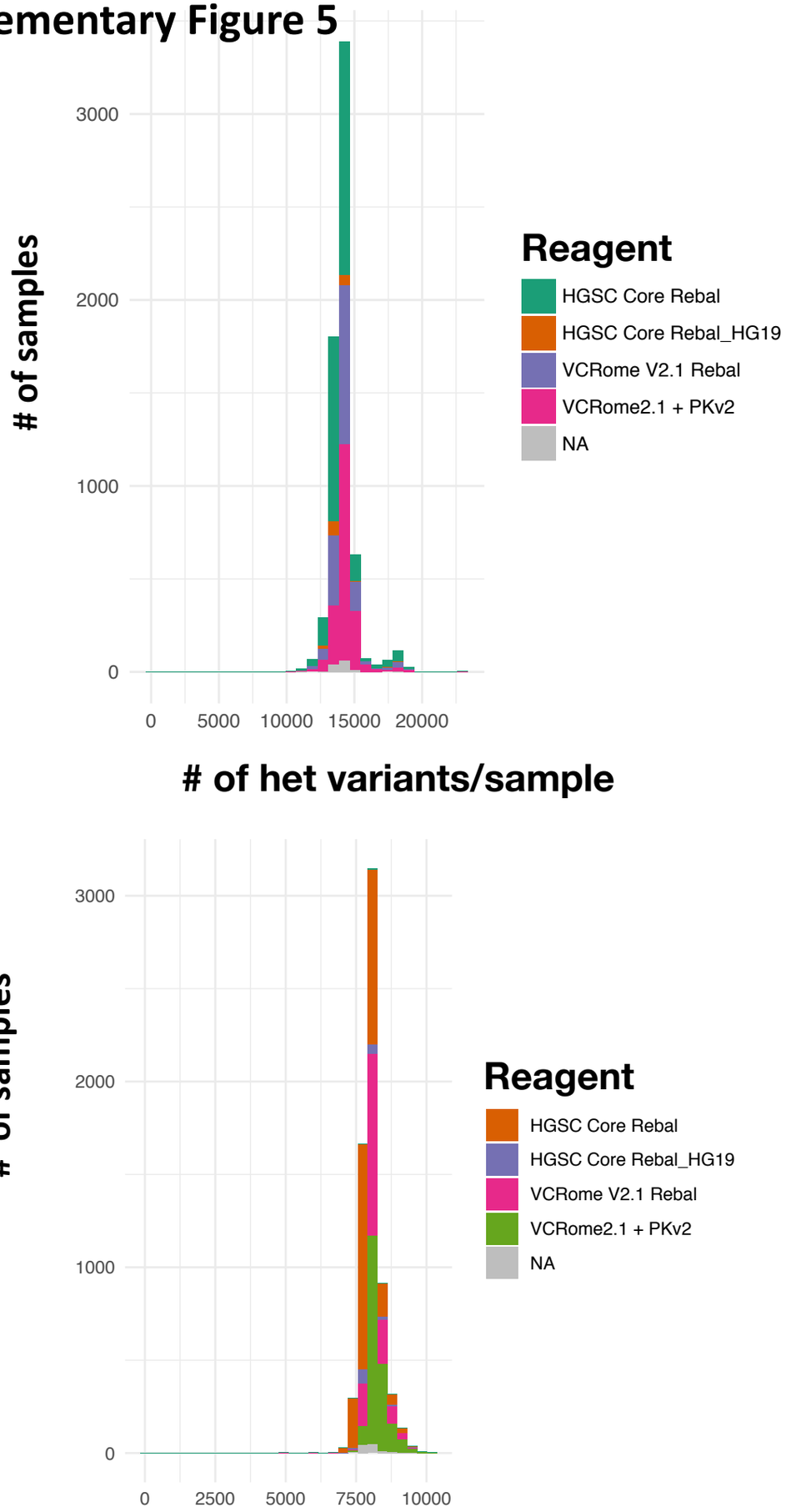
C

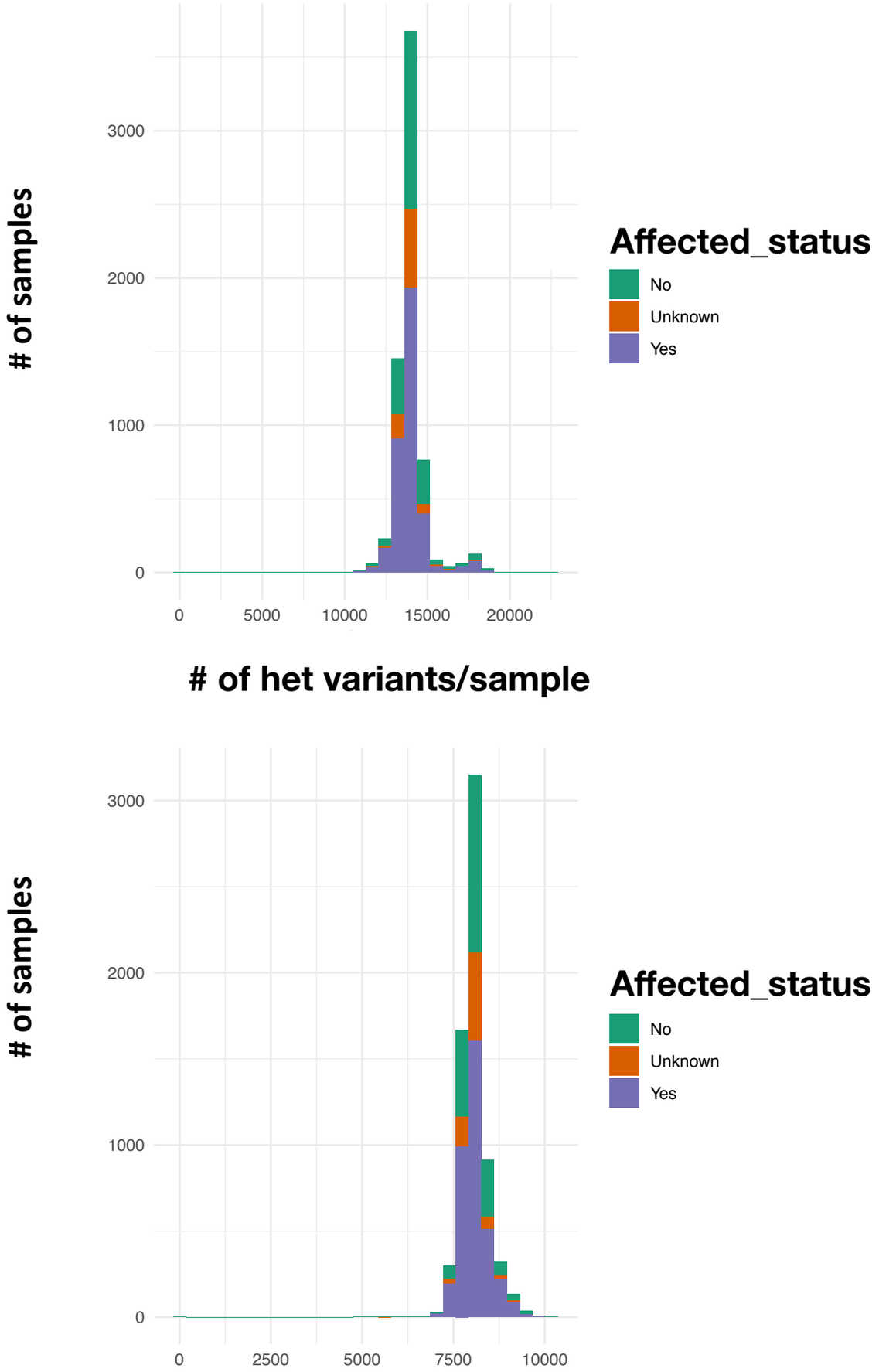


Supplementary Figure 4

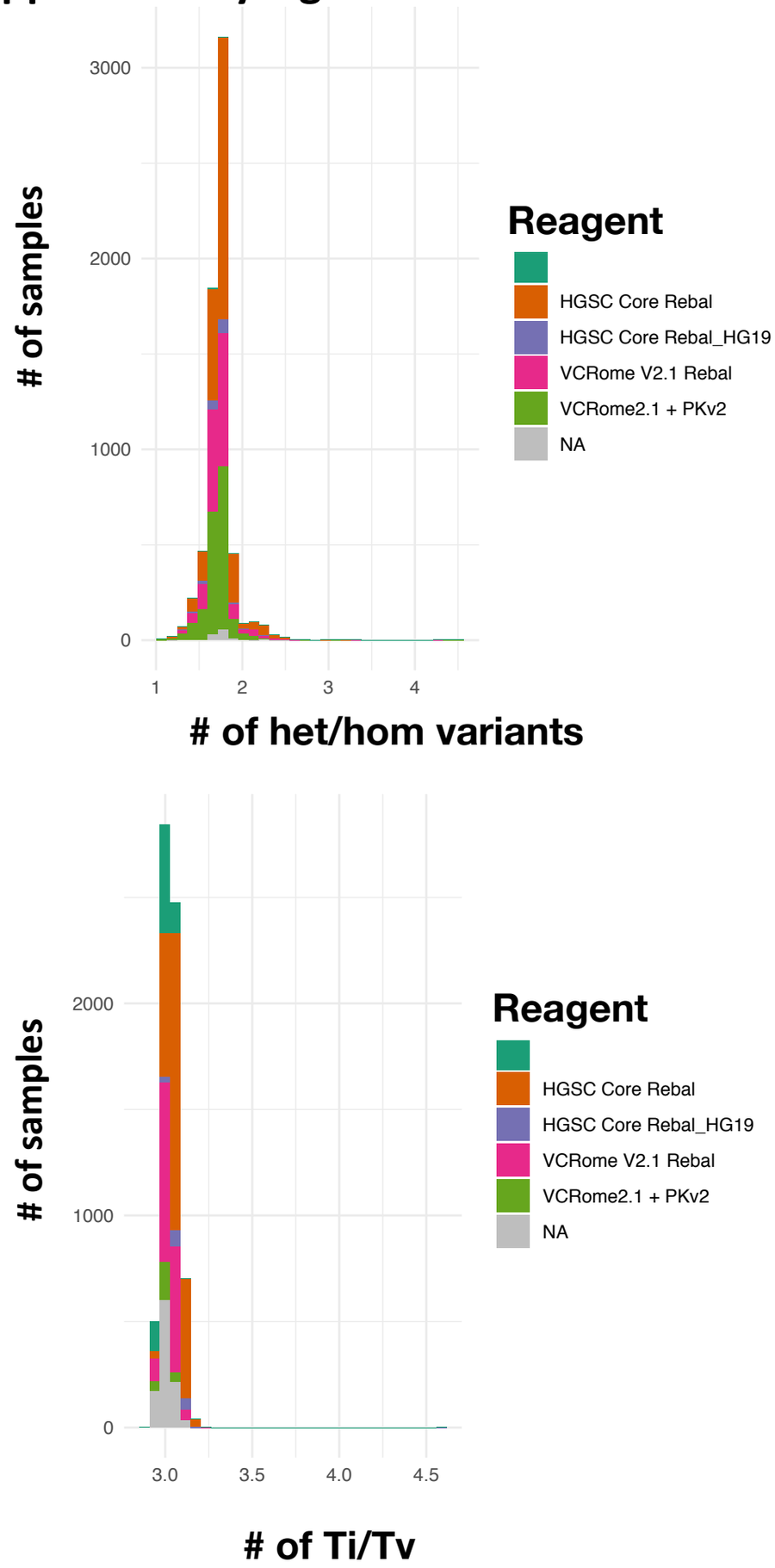
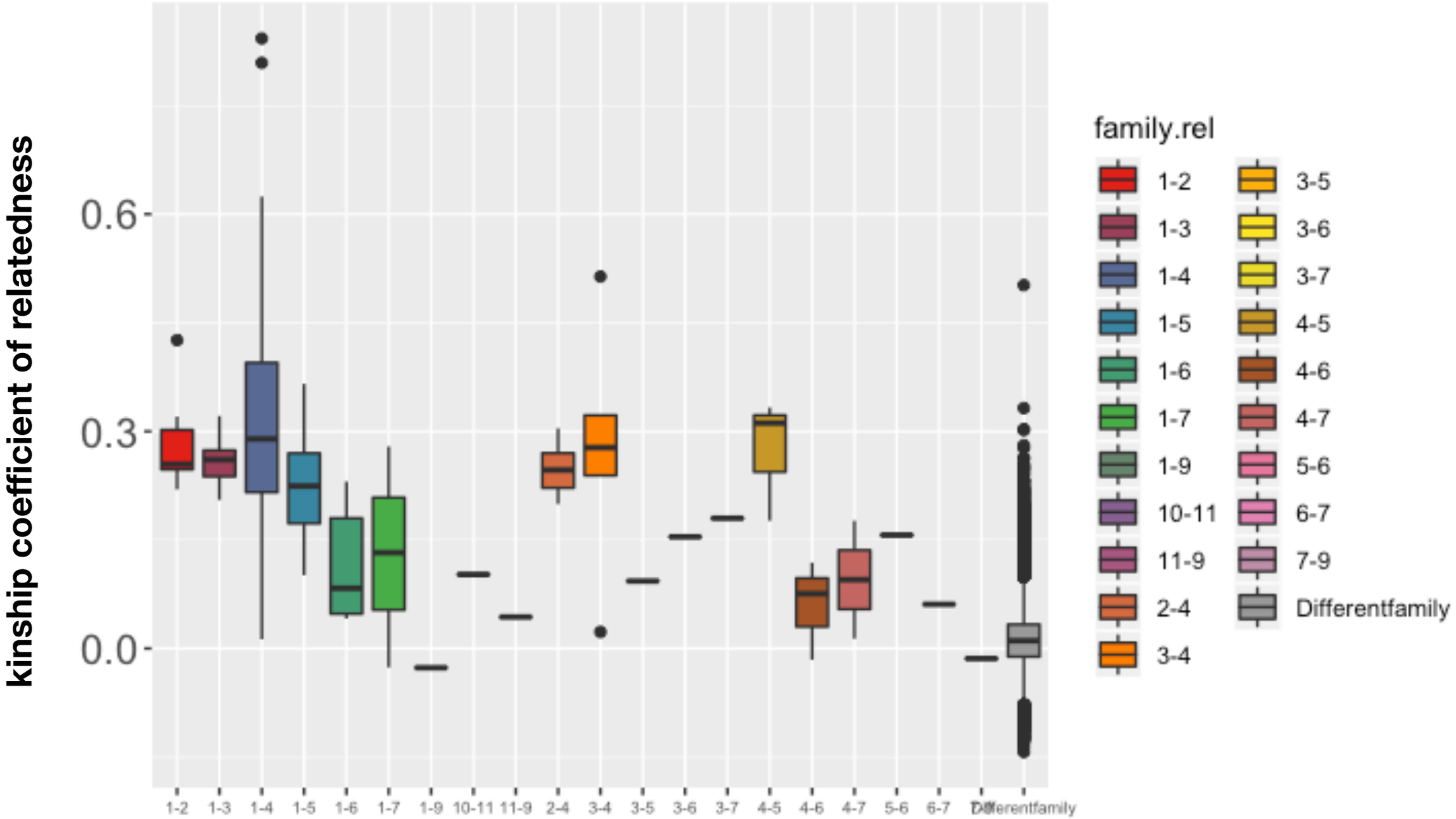


Supplementary Figure 5





Supplementary Figure 6



Supplementary Figure 7