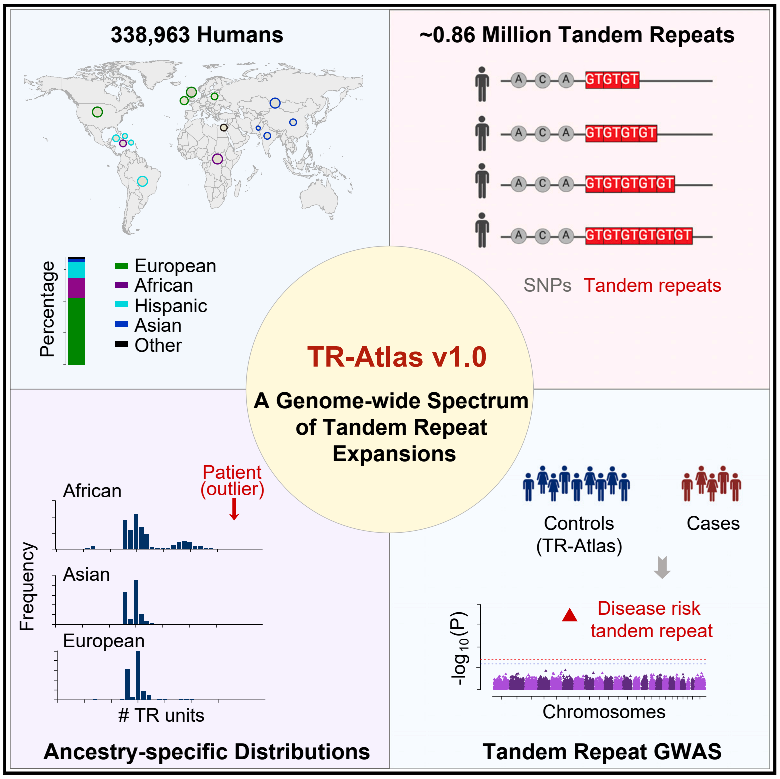
Cell

Resource

A genome-wide spectrum of tandem repeat expansions in 338,963 humans

Graphical abstract



Highlights

- A biobank-scale reference of 0.86 million TRs derived from 338,963 humans
- A TR reference map for diverse ancestries, including 39.5% non-European samples
- Critical insights into the prevalence of ancestry-specific TR disorders
- An invaluable resource for interpreting TR expansions in diseases

Authors

Ya Cui, Wenbin Ye, Jason Sheng Li, Jingyi Jessica Li, Eric Vilain, Tamer Sallam, Wei Li

Correspondence

yac7@uci.edu (Y.C.),
wei.li@uci.edu (W.L.)

In brief

The Tandem Repeat Aggregation Atlas (TR-Atlas) is a biobank-scale reference of 0.86 million TRs derived from 338,963 whole-genome sequencing (WGS) samples of diverse ancestries.

Cui et al., 2024, Cell *187*, 2336–2341
April 25, 2024 © 2024 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.cell.2024.03.004

CellPress

Resource

A genome-wide spectrum of tandem repeat expansions in 338,963 humans

Ya Cui,[1,6,*] Wenbin Ye,[1,6] Jason Sheng Li,[1] Jingyi Jessica Li,[2] Eric Vilain,[3,4] Tamer Sallam,[5] and Wei Li[1,7,*]
[1]Division of Computational Biomedicine, Department of Biological Chemistry, School of Medicine, University of California, Irvine, Irvine, CA 92697, USA
[2]Department of Statistics, University of California, Los Angeles, Los Angeles, CA 90095, USA
[3]Institute for Clinical and Translational Science, University of California, Irvine, Irvine, CA 92697, USA
[4]Department of Pediatrics, University of California, Irvine, Irvine, CA 92697, USA
[5]Division of Cardiology, Department of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA
[6]These authors contributed equally
[7]Lead contact
*Correspondence: yac7@uci.edu (Y.C.), wei.li@uci.edu (W.L.)
https://doi.org/10.1016/j.cell.2024.03.004

SUMMARY

The Genome Aggregation Database (gnomAD), widely recognized as the gold-standard reference map of human genetic variation, has largely overlooked tandem repeat (TR) expansions, despite the fact that TRs constitute ∼6% of our genome and are linked to over 50 human diseases. Here, we introduce the TR-Atlas (https://wlcb.oit.uci.edu/TRatlas/), a biobank-scale reference of 0.86 million TRs derived from 338,963 whole-genome sequencing (WGS) samples of diverse ancestries (39.5% non-European samples). TR-Atlas offers critical insights into ancestry-specific disease prevalence using disparities in TR unit number frequencies among ancestries. Moreover, TR-Atlas is able to differentiate between common, presumably benign TR expansions, which are prevalent in TR-Atlas, from those potentially pathogenic TR expansions, which are found more frequently in disease groups than within TR-Atlas. Together, TR-Atlas is an invaluable resource for researchers and physicians to interpret TR expansions in individuals with genetic diseases.

INTRODUCTION

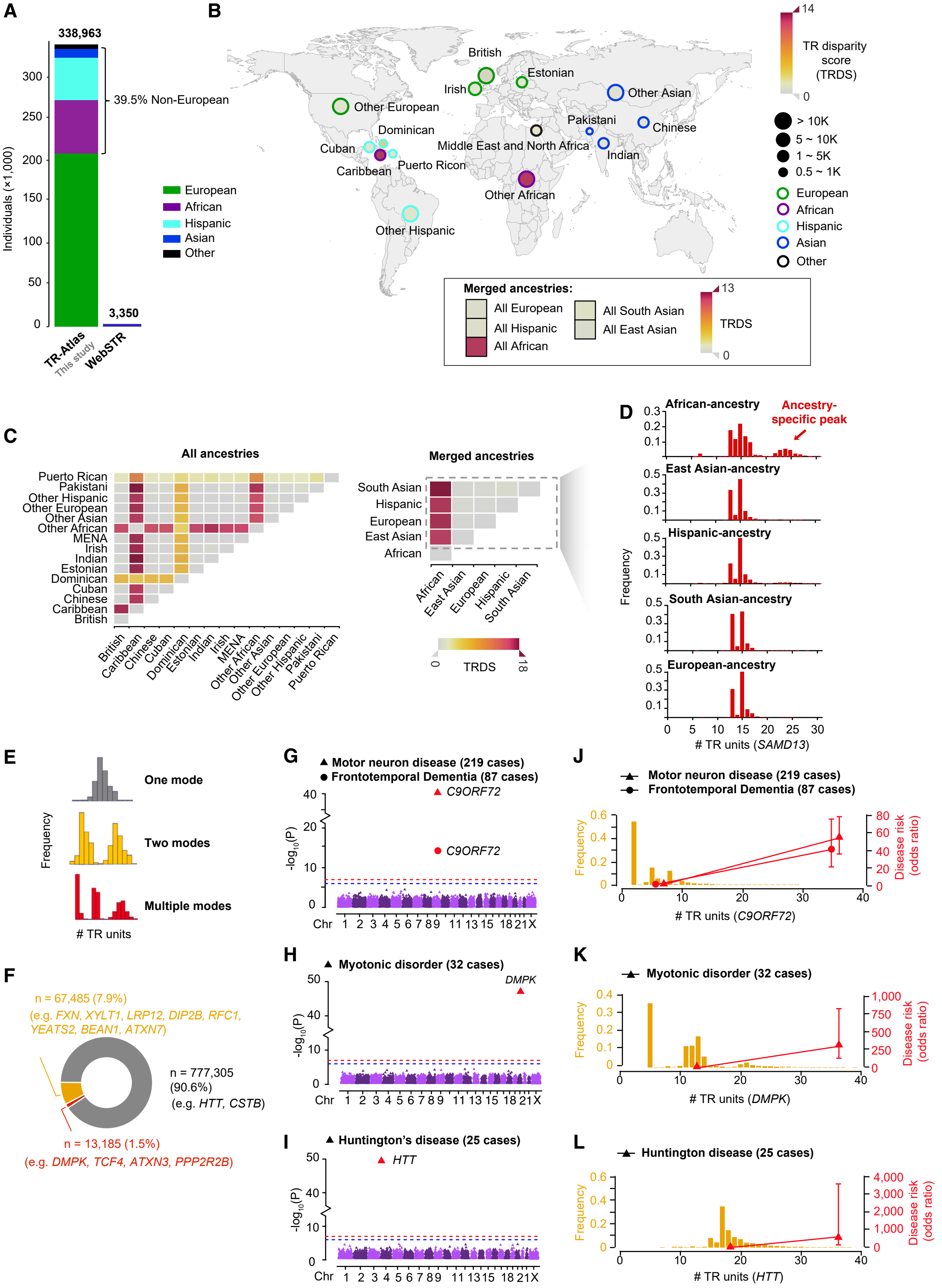
The Genome Aggregation Database (gnomAD),[1] widely recognized as the gold-standard reference map of human genetic variation, has served as an essential tool for interpreting single-nucleotide variants (SNVs)[1] and structural variants (SVs)[2] identified in disease-association studies and clinical genetic diagnostic tests. However, there is no biobank-scale reference map for tandem repeat (TR) expansions, where an expansion refers to an increase in the number of "TR units" (i.e., consecutive repeated DNA sequences). To date, the only reference maps for TR expansions were constructed from limited samples,[3,4] e.g., 3,350 individuals from the 1000 Genomes Project and H3Africa cohorts.[3] This limitation clearly underscores the urgent need to develop a more comprehensive reference map of TR expansions, thereby offering a complete understanding of human genetic variation.

Millions of TRs, constituting ∼6% of the human genome, have profound impacts on evolution and human disease.[5] Compared with SNVs and SVs, TRs have mutation rates that are several orders of magnitude higher.[6] Importantly, TR expansions are implicated in more than 50 lethal human diseases, including amyotrophic lateral sclerosis (ALS),[5] Huntington's disease,[5] and multiple cancers.[7] Furthermore, TR expansions captured by whole-genome sequencing (WGS) have been widely used in the diagnosis of rare diseases.[8] Surprisingly, current biobank-scale resources utilizing WGS data, such as gnomAD[1] and the Trans-Omics for Precision Medicine (TOPMed) program,[9] have largely overlooked TR expansions. This oversight means that, despite the substantial contribution of TRs to monogenic and complex genetic diseases,[5] our understanding of TR expansion spectrum at a biobank-scale remains strikingly limited.

RESULTS

A biobank-scale reference of TR expansions in 338,963 humans

Here, we introduce the TR-Atlas at the University of California, Irvine (UCI TR-Atlas or TR-Atlas: https://wlcb.oit.uci.edu/TRatlas/), a biobank-scale reference of TR expansions derived from 338,963 human genomes. These genomes represent diverse ancestries (39.5% non-European), with an average WGS data coverage of 33× (Figures 1A and 1B; Table S1). We used two accurate and widely used TR genotypers, ExpansionHunter[10] and GangSTR,[11] to increase the coverage of TR genotyping. This strategy enabled us to genotype 0.91 million TRs per individual. After removing low-quality TR alleles using TRTools,[12] a total of 0.86 million TRs were retained for further analysis (STAR



(legend on next page)

Methods). We found that 30.5% of TRs have at least two common (frequency ≥ 0.05) alleles, e.g., 15 TR units and 24 TR units in the African ancestry in Figure 1D. The extremely large sample size and diverse ancestries make TR-Atlas as a pivotal reference for TR expansions.

TR-Atlas enables users to ascertain the prevalence or rarity of a specific TR expansion within its respective ancestry (Figures 1B–1D). Moreover, TR-Atlas offers critical insights into ancestry-specific disease prevalence using disparities in TR unit number frequencies among ancestries (Figures S1A and S1B). Using the 2-Wasserstein distance—a classical metric in mathematics that quantifies the difference between two distributions—we developed the TR disparity score (TRDS) to quantify the difference between the TR unit number frequency distributions of two groups of human samples. The TR-Atlas browser displays these TRDSs through an interactive world map (Figure 1B), comparing each ancestry to all TR-Atlas samples. Additionally, the browser also features heatmaps of TRDSs for direct pairwise comparisons between ancestries (Figure 1C). We found 4.0% of TRs with high TRDSs (>5.0) in at least one ancestry. For example, an intergenic TR located 32.8 kb upstream of the *SAMD13* gene had a distinct TR unit number distribution in the African ancestry compared with other ancestries (Figures 1C and 1D). Furthermore, we found that over 9.6% of TRs have multiple modes (peaks) in their TR unit number distributions, and many of these TRs were found to be pathogenic (Figures 1E and 1F). The increased frequency of expanded TR units, specific to certain ancestries, could directly influence the ancestry-specific prevalence of the TR-associated diseases or traits. For example, the prevalence of expanded CAG TR units in the *DMPK* gene, responsible for myotonic dystrophy type 1 (DM1), is notably less common in individuals of African ancestry than in European ancestry (Figure S1A). This pattern corresponds very well with the observed rarity of DM1 in Africa.[13] Similarly, the prevalence of expanded GAA TR units in *FXN*, linked to Friedreich ataxia (FRDA), is less prevalent in East Asian samples than in other ancestries (Figure S1B). This observation again mirrors the low frequency of reported cases of FRDA in Japan.[13] Furthermore, our analysis revealed a much smaller TRDS between samples of the same ancestry (European) from different cohorts (UK Biobank and TOPMed) compared with those observed between samples of different ancestries (Figures S1C and S1D). This indicates that combining multiple cohorts has minimal impact on TRDS.

TR-Atlas may serve as a control cohort for interpretating pathogenic TRs

TR-Atlas may be able to differentiate between common (presumably benign) TR expansions, which are prevalent in TR-Atlas, from those potentially pathogenic TR expansions, which are found more frequently in disease groups than within TR-Atlas. We evaluated the capacity of TR-Atlas to serve as a control cohort for the interpretation of known clinically pathogenic TRs. We genotyped 0.17 million TRs using ExpansionHunter[10] in European samples related to four TR-associated rare diseases from the UK Biobank, with case numbers ranging from 25 to 219. Due to privacy constraints preventing access to real TR unit numbers at the individual level, for each TR locus, we compared its unit numbers within a disease group to those randomly sampled (n = 18,000) from individuals of matched ancestry in TR-Atlas. Please note that we strongly recommend using matched ancestry in such analysis. A recent extensive clinical study[14] provides compelling evidence that short-read WGS, when applied with a predefined risk threshold, can successfully separate normal and risk TR groups. This approach achieves 97.3% sensitivity and 99.6% specificity when compared with the gold-standard diagnostic method, PCR testing, even for disease TR loci with large repeat expansions (larger than the short-read length 150 bp). Motivated by this study, we stratified TRs into two groups for subsequent analyses based on a predefined risk threshold of 99.5% percentiles. For example, in our test case of CAG repeat pathogenicity at the *DMPK* locus (Figures 1H and 1K), TR extensions were stratified into two groups based on a 99.5% percentile threshold of 30 CAG repeat copies, corresponding to a length threshold of 90 bp. According to this threshold approach, individuals with TR expansions of greater than 90 bp were classified as being in the risk group, whereas individuals with TR expansions of 90 bp or less were considered normal. Both TR-based genome-wide association studies (TR-based GWASs) and the chi-squared tests confirmed the associations of all four known pathogenic TRs with their related diseases (Figures 1G–1L). For example, the risk of motor neuron

Figure 1. A biobank-scale tandem repeat expansion spectrum quantified from 338,963 whole-genome sequencing genomes in diverse human ancestries

(A) The size and diversity of WGS samples in the TR-Atlas and WebSTR.[3]

(B) The TR-Atlas browser displays the TR disparity scores (TRDSs) between specific ancestries and all TR-Atlas samples for an intergenic TR at 32.8 kb upstream of the *SAMD13* gene. Node size represents sample size.

(C) Heatmaps of TRDSs in the TR-Atlas browser for an intergenic TR at 32.8 kb upstream of the *SAMD13* gene. MENA represents the Middle East and North Africa.

(D) Histograms in the TR-Atlas browser show the repeat unit number distributions for an intergenic TR at 32.8 kb upstream of the *SAMD13* gene. In each panel, the x axis denotes the number of TR unit, and the y axis denotes the frequency of each TR unit number. Each panel shows a merged ancestry.

(E) Three types of TR unit number distributions.

(F) The number and percentage of TRs in three types of TR unit number distributions. TR-related Mendelian disease genes in different types are shown as examples.

(G–I) Manhattan plots of TR-based genome-wide association studies for MND (triangle) (G), FTD (circle) (G), myotonic disorder (H), and Huntington's disease (I).

(J–L) TR unit number distribution (histograms, left axis) and disease odds ratios (lines, right axis) of *C9ORF72* TR units (J), *DMPK* TR units (K), and *HTT* TR units (L). In (G)–(L), *C9ORF72* TR units were stratified into two groups for phenotype analyses: short (≤25 repeat units) and long (>25 repeat units). *DMPK* TR units were stratified into two groups for phenotype analyses: short (≤ 30 repeat units) and long (>30 repeat units). *HTT* TR units stratified into two groups for phenotype analyses: short (≤ 30 repeat units) and long (>30 repeat units). Error bars indicate 95% confidence intervals (CIs).

See also Figure S1.

disease (MND) and frontotemporal dementia (FTD) increased with a TR expansion (>25 copies) in the first intron of *C9ORF72* (odds ratio [OR] = 54.2, two-sided p = 1.2e−44, χ^2, for MND; OR = 41.5, two-sided p = 2.4e−14, χ^2, for FTD) (Figure 1J). Likewise, individuals with a 3′ UTR TR expansion (>30 copies) in *DMPK* were at a higher risk for myotonic disorder (OR = 323.4, two-sided p = 2.2e−43, χ^2) (Figure 1K). Additionally, carriers with a coding TR expansion (>30 copies) in *HTT* demonstrated a higher risk for Huntington's disease (OR = 639.8, two-sided p = 3.2e−39, χ^2) (Figure 1L). Collectively, these results underscore the potential utility of TR-Atlas as a tool to improve clinical diagnosis.

DISCUSSION

Looking ahead, we intend to analyze more WGS data from diverse backgrounds, with a particular focus on underrepresented ancestries, including those within the All of Us Research Program. In the next phase of TR-Atlas project, we plan to genotype more TRs, prioritizing those recently identified as disease-associated by long-read sequencing.[15] In summary, freely available to the public, TR-Atlas stands as an invaluable resource for researchers, physicians, and genetic counselors to interpret TR expansions in individuals with genetic diseases.

Limitations of the study

We acknowledge several limitations in our study. First, although our TR-Atlas provides TR expansion reference maps for 11 sub populations (Caribbean, Chinese, Indian, Pakistani, British, Estonian, Irish, Cuban, Dominican, Puerto Rican, and Middle Eastern/North African [MENA]), it does not yet cover certain underrepresented ancestries, such as Australian, Pacific Islander, and Mongolian. The TR-Atlas team is committed to incorporating these underrepresented ancestries as additional WGS data become available. Second, although the current version of TR-Atlas has successfully genotyped ∼0.86 million TRs, this number still represents a fraction of the total number of TRs in the human genome. Future phases of TR-Atlas will prioritize the integration of a greater number of high-quality TRs. Third, TR-Atlas has not yet been used to identify unknown disease risk repeat expansions or to define risk thresholds for known pathogenic TRs. Further analyses involving patient cohorts potentially affected by TRs will address these questions. Users should proceed with caution to avoid case-control mismatch and ascertainment bias when using TR-Atlas as a control cohort in these applications. Finally, the current version of TR-Atlas is limited to short-read WGS data, for which there are large numbers of human sequence samples available. While short-read WGS is increasingly used as a first-line test for human genetic disorders,[14] it may underestimate the allele lengths of large expansions (i.e., >150 bp). However, the determination of TR pathogenicity does not rely on the exact count of such large expansions, as evidenced by recent studies.[12] Nonetheless, to enhance the quality of TR-Atlas, we plan to include long-read sequencing data, such as those from the "All of Us" project, in future versions of TR-Atlas as soon as a large-scale long-read sequencing dataset becomes available.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - ○ Human Participants
- METHOD DETAILS
  - ○ WGS samples in TR-Atlas
  - ○ TR genotyping and processing in TR-Atlas
  - ○ TR disparity score
  - ○ Comparison of TRDS from different cohorts
  - ○ TR genotyping in disease cohorts
  - ○ TR-based genome-wide association analyses
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2024.03.004.

ACKNOWLEDGMENTS

We thank members of the Li lab for helpful discussions. We thank Hoda Anton-Culver for her assistance in analyzing the All of Us data. We thank Qin Yang, Changrui Xiao, Seth Berger, Fabiola Quintero-Rivera, Peter Kaiser, Suzanne Sandmeyer, Ahmed Mohyeldin, Albert La Spada, Leslie Thompson, and Ivan Marazzi for helpful discussions. We also thank Philip Papadopoulos and Imam Toufique for their computing support. This research was supported by the US National Institutes of Health to W.L. (R01HG007538, R01CA193466, R01CA228140, and U01HG011745, part of the GREGoR consortium) and to E.V. (5UL1TR001414 and U01HG011745, part of the GREGoR consortium). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program were supported by the National Heart, Lung, and Blood Institute (NHLBI). Genome sequencing for "NHLBI TOPMed: Whole Genome Sequencing and Related Phenotypes in the Framingham Heart Study" (dbGaP: phs000974.v1.p1) was performed at the Broad Institute Genomics Platform (3R01HL092577-06S1, 3U54HG003067-12S2). Genome sequencing for "NHLBI TOPMed: The Jackson Heart Study" (dbGaP: phs000964.v1.p1) was performed at the Northwest Genomics Center (HHSN268201100037C). Core support, including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering, were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support, including phenotype harmonization, data management, sample-identity QC, and general program coordination, was provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). This research has been conducted using the UK Biobank Resource under application number 78898. The All of Us Research Program is supported by the National Institutes of Health, Office of the Director: Regional Medical Centers: 1 OT2 OD026549; 1 OT2 OD026554; 1 OT2 OD026557; 1 OT2 OD026556; 1 OT2 OD026550; 1 OT2 OD 026552; 1 OT2 OD026553; 1 OT2 OD026548; 1 OT2 OD026551; 1 OT2 OD026555; IAA #: AOD 16037; Federally Qualified Health Centers: HHSN 263201600085U; Data and Research Center: 5 U2C OD023196; Biobank: 1 U24 OD023121; The Participant Center: U24 OD023176; Participant Technology Systems Center: 1 U24 OD023163; Communications and Engagement: 3 OT2 OD023205; 3 OT2 OD023206; and Community Partners: 1 OT2 OD025277; 3 OT2 OD025315; 1

OT2 OD025337; and 1 OT2 OD025276. In addition, the All of Us Research Program would not be possible without the partnership of its participants. We thank the Broad Institute for generating high-quality sequence data supported by NHGRI funds (grant # U54 HG003067) with Eric Lander as the principal investigator. The Estonian Biobank WGS data used in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number, dbGaP: phs001230. We gratefully acknowledge the studies and participants who provided biological samples and data for the TOPMed, UK Biobank, Estonian Biobank, 1000 Genome Project, and All of Us Research Program.

AUTHOR CONTRIBUTIONS

Y.C. and W.L. conceived and designed the project. Y.C., W.Y., and J.S.L. performed the bioinformatics analysis. Y.C., W.L., W.Y., J.S.L., J.J.L., E.V., and T.S. interpreted analytical results. Y.C. and W.L. drafted the initial manuscript. All authors reviewed and edited the manuscript. All authors approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 8, 2023
Revised: January 23, 2024
Accepted: March 5, 2024
Published: April 5, 2024; Corrected online: October 3, 2024

REFERENCES

1. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443. https://doi.org/10.1038/s41586-020-2308-7.
2. Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., Alföldi, J., Francioli, L.C., Khera, A.V., Lowther, C., Gauthier, L.D., Wang, H., et al. (2020). A structural variation reference for medical and population genetics. Nature *581*, 444–451. https://doi.org/10.1038/s41586-020-2287-8.
3. Ziaei Jam, H., Li, Y., DeVito, R., Mousavi, N., Ma, N., Lujumba, I., Adam, Y., Maksimov, M., Huang, B., Dolzhenko, E., et al. (2023). A deep population reference panel of tandem repeat variation. Nat. Commun. *14*, 6711. https://doi.org/10.1038/s41467-023-42278-3.
4. Shi, Y., Niu, Y., Zhang, P., Luo, H., Liu, S., Zhang, S., Wang, J., Li, Y., Liu, X., Song, T., et al. (2023). Characterization of genome-wide STR variation in 6487 human genomes. Nat. Commun. *14*, 2092. https://doi.org/10.1038/s41467-023-37690-8.
5. Hannan, A.J. (2018). Tandem repeats mediating genetic plasticity in health and disease. Nat. Rev. Genet. *19*, 286–298. https://doi.org/10.1038/nrg.2017.115.
6. Sun, J.X., Helgason, A., Masson, G., Ebenesersdóttir, S.S., Li, H., Mallick, S., Gnerre, S., Patterson, N., Kong, A., Reich, D., et al. (2012). A direct characterization of human mutation based on microsatellites. Nat. Genet. *44*, 1161–1165. https://doi.org/10.1038/ng.2398.
7. Erwin, G.S., Gürsoy, G., Al-Abri, R., Suriyaprakash, A., Dolzhenko, E., Zhu, K., Hoerner, C.R., White, S.M., Ramirez, L., Vadlakonda, A., et al. (2023). Recurrent repeat expansions in human cancer genomes. Nature *613*, 96–102. https://doi.org/10.1038/s41586-022-05515-1.
8. Stranneheim, H., Lagerstedt-Robinson, K., Magnusson, M., Kvarnung, M., Nilsson, D., Lesko, N., Engvall, M., Anderlid, B.M., Arnell, H., Johansson, C.B., et al. (2021). Integration of whole genome sequencing into a healthcare setting: high diagnostic rates across multiple clinical entities in 3219 rare disease patients. Genome Med. *13*, 40. https://doi.org/10.1186/s13073-021-00855-5.
9. Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Taliun, S.A.G., Corvelo, A., Gogarten, S.M., Kang, H.M., et al. (2021). Sequencing of 53,831 diverse genomes from the Nhlbi TOPMed Program. Nature *590*, 290–299. https://doi.org/10.1038/s41586-021-03205-y.
10. Dolzhenko, E., Deshpande, V., Schlesinger, F., Krusche, P., Petrovski, R., Chen, S., Emig-Agius, D., Gross, A., Narzisi, G., Bowman, B., et al. (2019). ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. Bioinformatics *35*, 4754–4756. https://doi.org/10.1093/bioinformatics/btz431.
11. Mousavi, N., Shleizer-Burko, S., Yanicky, R., and Gymrek, M. (2019). Profiling the genome-wide landscape of tandem repeat expansions. Nucleic Acids Res. *47*, e90. https://doi.org/10.1093/nar/gkz501.
12. Mousavi, N., Margoliash, J., Pusarla, N., Saini, S., Yanicky, R., and Gymrek, M. (2021). TRTools: a toolkit for genome-wide analysis of tandem repeats. Bioinformatics *37*, 731–733. https://doi.org/10.1093/bioinformatics/btaa736.
13. Depienne, C., and Mandel, J.L. (2021). 30 years of repeat expansion disorders: What have we learned and what are the remaining challenges? Am. J. Hum. Genet. *108*, 764–785. https://doi.org/10.1016/j.ajhg.2021.03.011.
14. Ibañez, K., Polke, J., Hagelstrom, R.T., Dolzhenko, E., Pasko, D., Thomas, E.R.A., Daugherty, L.C., Kasperaviciute, D., Smith, K.R., et al.; WGS for Neurological Diseases Group (2022). Whole genome sequencing for the diagnosis of neurological repeat expansion disorders in the UK: a retrospective diagnostic accuracy and prospective clinical validation study. Lancet Neurol. *21*, 234–245. https://doi.org/10.1016/S1474-4422(21)00462-2.
15. Course, M.M., Gudsnuk, K., Smukowski, S.N., Winston, K., Desai, N., Ross, J.P., Sulovari, A., Bourassa, C.V., Spiegelman, D., Couthouis, J., et al. (2020). Evolution of a Human-Specific Tandem Repeat Associated with ALS. Am. J. Hum. Genet. *107*, 445–460. https://doi.org/10.1016/j.ajhg.2020.07.004.
16. Backman, J.D., Li, A.H., Marcketta, A., Sun, D., Mbatchou, J., Kessler, M.D., Benner, C., Liu, D., Locke, A.E., Balasubramanian, S., et al. (2021). Exome sequencing and analysis of 454,787 UK Biobank participants. Nature *599*, 628–634. https://doi.org/10.1038/s41586-021-04103-z.
17. Halldorsson, B.V., Eggertsson, H.P., Moore, K.H.S., Hauswedell, H., Eiriksson, O., Ulfarsson, M.O., Palsson, G., Hardarson, M.T., Oddsson, A., Jensson, B.O., et al. (2022). The sequences of 150,119 genomes in the UK Biobank. Nature *607*, 732–740. https://doi.org/10.1038/s41586-022-04965-x.
18. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209. https://doi.org/10.1038/s41586-018-0579-z.
19. Ramirez, A.H., Sulieman, L., Schlueter, D.J., Halvorson, A., Qian, J., Ratsimbazafy, F., Loperena, R., Mayo, K., Basford, M., Deflaux, N., et al. (2022). The All of Us Research Program: Data quality, utility, and diversity. Patterns (N Y) *3*, 100570. https://doi.org/10.1016/j.patter.2022.100570.
20. Byrska-Bishop, M., Evani, U.S., Zhao, X., Basile, A.O., Abel, H.J., Regier, A.A., Corvelo, A., Clarke, W.E., Musunuri, R., Nagulapalli, K., et al. (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell *185*, 3426–3440.e19. https://doi.org/10.1016/j.cell.2022.08.004.
21. Schefzik, R., Flesch, J., and Goncalves, A. (2021). Fast identification of differential distributions in single-cell RNA-sequencing data with waddR. Bioinformatics *37*, 3204–3211. https://doi.org/10.1093/bioinformatics/btab226.
22. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

23. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

24. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. *89*, 82–93. https://doi.org/10.1016/j.ajhg.2011.05.029.

25. Aragon, T.J., Fay, M., and Wollschlaeger, D. (2020). epitools: Epidemiology Tools. R package version 0.5-10.1 (CRAN). https://cran.r-project.org/web/packages/epitools/epitools.pdf.

26. Yin, L., Zhang, H., Tang, Z., Xu, J., Yin, D., Zhang, Z., Yuan, X., Zhu, M., Zhao, S., Li, X., et al. (2021). rMVP: A Memory-efficient, Visualization-enhanced, and Parallel-accelerated Tool for Genome-wide Association Study. Genomics Proteomics Bioinformatics *19*, 619–628. https://doi.org/10.1016/j.gpb.2020.10.007.

27. Leitsalu, L., Haller, T., Esko, T., Tammesoo, M.L., Alavere, H., Snieder, H., Perola, M., Ng, P.C., Mägi, R., Milani, L., et al. (2015). Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. Int. J. Epidemiol. *44*, 1137–1147. https://doi.org/10.1093/ije/dyt268.

28. Navarro Gonzalez, J., Zweig, A.S., Speir, M.L., Schmelter, D., Rosenbloom, K.R., Raney, B.J., Powell, C.C., Nassar, L.R., Maulding, N.D., Lee, C.M., et al. (2021). The UCSC Genome Browser database: 2021 update. Nucleic Acids Res. *49*, D1046–D1057. https://doi.org/10.1093/nar/gkaa1070.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
TR frequency distributions and TR disparity scores in TR-Atlas	This paper	https://doi.org/10.5281/zenodo.10806727
UK Biobank WGS data	Backman et al.[16]; Halldorsson et al.[17]; Bycroft et al.[18]	http://www.ukbiobank.ac.uk
All of Us WGS data	Ramirez et al.[19]	https://allofus.nih.gov/
TOPMed WGS data	Taliun et al.[9]	https://topmed.nhlbi.nih.gov/
1000 Genomes Project WGS data	Byrska-Bishop et al.[20]	https://www.internationalgenome.org/data-portal/data-collection/30x-grch38
Estonian Biobank WGS data	dbGaP	Accession phs001230
ExpansionHunter reference TR catalog	Dolzhenko et al.[10]	https://github.com/Illumina/RepeatCatalogs
GangSTR reference TR catalog	Mousavi et al.[11]	https://github.com/gymreklab/GangSTR
Software and algorithms		
Code of TR genotyping and TR-based GWAS	This paper	https://doi.org/10.5281/zenodo.10662989
ExpansionHunter	Dolzhenko et al.[10]	https://github.com/Illumina/ExpansionHunter
GangSTR	Mousavi et al.[11]	https://github.com/gymreklab/GangSTR
TRTools	Mousavi et al.[12]	https://github.com/gymreklab/TRTools
waddR	Schefzik et al.[21]	https://bioconductor.org/packages/release/bioc/html/waddR.html
Samtools	Li et al.[22]	http://samtools.sourceforge.net
Bedtools	Quinlan and Hall[23]	https://github.com/arq5x/bedtools2
SKAT	Wu et al.[24]	https://cran.r-project.org/web/packages/SKAT
epitools	Aragon et al.[25]	https://CRAN.R-project.org/package=epitools
CMplot	Yin et al.[26]	https://github.com/YinLiLin/CMplot

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Wei Li (wei.li@uci.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyzes existing, publicly available WGS data from TOPMed,[9] UK Biobank,[16–18] Estonian Biobank,[27] 1000 Genome Project[20] and All of Us Research Program.[19] The links and accession numbers for these datasets are listed in the key resources table. The ExpansionHunter and GangSTR reference TR catalogs used in this paper have been deposited at GitHub. The links for these datasets are listed in the key resources table. In addition, TR frequency distributions and TR disparity scores in this study have been deposited at Zenodo and are publicly available as of the date of publication. The DOI is listed in the key resources table.
- All original code has been deposited at Zenodo and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Human Participants

Participants used in this study were included from five existing, publicly available cohorts: TOPMed, UK Biobank, Estonian Biobank, 1000 Genome Project and All of Us Research Program. Detailed information on the participant numbers from each cohort was included in Table S1. More detailed information of these participants, such as age, sex, ancestry, could be found in each original cohort. Sample size was determined based on the availability of existing samples in participating cohorts. No statistical methods were used to predetermine sample size. All participants provided written informed consent and the project was approved by each institution's ethical committee.

METHOD DETAILS

WGS samples in TR-Atlas

In this study, we analyzed 338,963 WGS samples from five publicly available cohorts: TOPMed (n = 43,069),[9] UK Biobank (n = 188,326),[16–18] Estonian Biobank (n = 2,279),[27] 1000 Genome Project (n = 3,202)[20] and All of Us Research Program (n = 102,087).[19] Among the UK Biobank and All of Us Research Program data, samples with unknown ancestry information or mixed ancestries were excluded. For the All of Us Research Program data, all European samples were also excluded in the analysis to save computational cost. After filtering, 102,087 WGS samples from the All of Us Research Program and 188,326 samples from the UK Biobank were retained. Detailed information on the WGS sample numbers of each ancestry from each cohort in TR-Atlas is included in Table S1. We only used ancestry information provided by the participating cohorts in TR-Atlas.

TR genotyping and processing in TR-Atlas

Using two accurate and widely-used TR genotyping tools, ExpansionHunter[10] and GangSTR,[11] we genotyped 0.91 million TRs for each WGS sample. The TR genotyping and filtering approaches were largely concordant with methodologies established in a previous study.[3] ExpansionHunter v5.0.0[10] was run separately on each individual using the latest hg38 reference TR catalog (174,293 TR loci) in the RepeatCatalogs (https://github.com/Illumina/RepeatCatalogs). GangSTR v2.5[11] was performed individually for each sample with non-default parameters –grid-threshold 250, –bam-samps and –samp-sex. The GangSTR reference TR catalog ("hg38_ver17.bed.gz") was downloaded from https://github.com/gymreklab/GangSTR, and there were 739,010 TR loci retained after excluding homopolymer TRs (e.g., AAAA) and sites overlap with the ExpansionHunter reference TR catalog using Bedtools v2.29.2.[23] MergeSTR in TRTools toolkit v4.0.1[12] with default parameters was performed to merge the VCF files of each individual into a single VCF file. Samtools v1.10[22] was used to compress and index VCF files. Both call-level and locus-level filtering were performed in VCF files from each tool using dumpSTR[12] in TRTools toolkit v4.0.1[12] with the following parameters: –min-locus-hwep 0.000001, –min-locus-callrate 0.4, –filter-regions hg38_segdup.sorted.bed.gz and –filter-regions-names SEGDUP. dumpSTR[12] removed TRs with low call rate, TRs whose genotypes do not follow Hardy-Weinberg Equilibrium and TRs overlapping segmental duplications downloaded from the UCSC Genome Browser.[28] We additionally used the parameter –eh-min-call-LC 10 to remove low quality calls for ExpansionHunter. For GangSTR, we additionally used parameters –max-call-DP 1000, –min-call-DP 10, –gangstr-filter-spanbound-only and –gangstr-filter-badCI to remove low quality TRs. To achieve harmonization across different cohorts, we applied the same TR genotyping and quality control methodology pipelines for each cohort. After the TR genotyping and filtering, we adopted the approach outlined in the previous published paper,[3] combining TR genotypes of the same ancestry from different cohorts for subsequent analyses.

TR disparity score

The 2-Wasserstein distance is a classical metric in mathematics that quantifies the difference between two distributions. In this study, we developed the TR disparity score (TRDS) to quantify the difference between the TR unit number frequency distributions of two groups of human samples using the 2-Wasserstein distance. The 2-Wasserstein distance (TRDS) in TR-Atlas was calculated using waddR v1.16.0.[21]

Comparison of TRDS from different cohorts

To demonstrate that samples from different cohorts do not impact the ancestry related TRDS differences, we conducted the TRDS comparison analysis using European samples from UK Biobank (n = 181,350 samples) and TOPMed (n = 20,854 samples). Only cohorts with large European sample size (n > 10,000) were considered. To make TRDS from different TRs comparable, TRDSs were normalized from 0 to 1 in Figure S1C.

TR genotyping in disease cohorts

To evaluate the capacity of TR-Atlas to serve as a control cohort for the interpretation of known pathogenic TRs, we obtained WGS data of European individuals with TR-associated rare diseases from the UK Biobank. In total, 25 Huntington's disease cases, 32 Myotonic disorder cases, 87 Frontotemporal dementia cases and 291 Motor neuron disease cases were used in the analysis. ExpansionHunter v5.0.0[10] was run separately on each individual using the latest hg38 reference TR catalog (174,293 TR loci) in

RepeatCatalogs (https://github.com/Illumina/RepeatCatalogs). Both call-level and locus-level filtering were performed in VCF files from each tool using dumpSTR[12] in TRTools toolkit v4.0.1[12] with the following parameters: –eh-min-call-LC 10, –min-locus-callrate 0.4, –filter-regions hg38_segdup.sorted.bed.gz and –filter-regions-names SEGDUP. dumpSTR[12] removed TRs with low quality, low call rate and TRs overlapping segmental duplications downloaded from the UCSC Genome Browser.[28]

TR-based genome-wide association analyses

We performed TR-based genome-wide association analyses (TR-based GWAS) for four diseases (Huntington's disease, Myotonic disorder, Frontotemporal dementia and Motor neuron disease) to show the capacity of TR-Atlas to serve as a control cohort for interpreting known pathogenic TRs. Due to privacy constraints preventing access to real TR unit numbers at the individual level in TR-Atlas, for each TR locus, we compared its unit numbers within a disease group to those randomly sampled (n = 18,000) from individuals of matched ancestry in TR-Atlas. Motivated by a recent extensive clinical study,[14] we stratified TRs into two groups for subsequent analyses based on a predefined risk threshold of 99.5% percentiles. For example, in our test case of CAG repeat pathogenicity at the *DMPK* locus, TR extensions were stratified into two groups based on a 99.5% percentile threshold of 30 CAG repeat copies, corresponding to a length threshold of 90bp. According to this threshold approach, individuals with TR expansions of greater than 90bp were classified as the being in the risk group, whereas individuals with TR expansions of 90bp or less were considered normal. We then performed TR-based GWAS by SKAT[24] to identify TR expansions which were enriched in the disease cohorts. We performed the analysis using the "SKATBinary_Robust" function with parameter "Burden" in SKAT[24] to reduce false positive results when case-control ratios were extremely unbalanced. We further performed Chi-Squared tests using epitools v0.5-10.1[25] to confirm the enrichment of significant TRs (TR-based GWAS P-value $\leq$ 1.0E-8) in patients. Disease risk odds ratios were also computed by epitools v0.5-10.1.[25] Manhattan plots were created using CMplot.[26]

QUANTIFICATION AND STATISTICAL ANALYSIS

Chi-Squared tests and disease risk odds ratios between TR expansion risk and normal groups were analyzed using statistical R package epitools v0.5-10.1.[25] Two-sided P-values of < 0.05 were considered statistically significant. Error bars, 95% CIs.

Supplemental figures

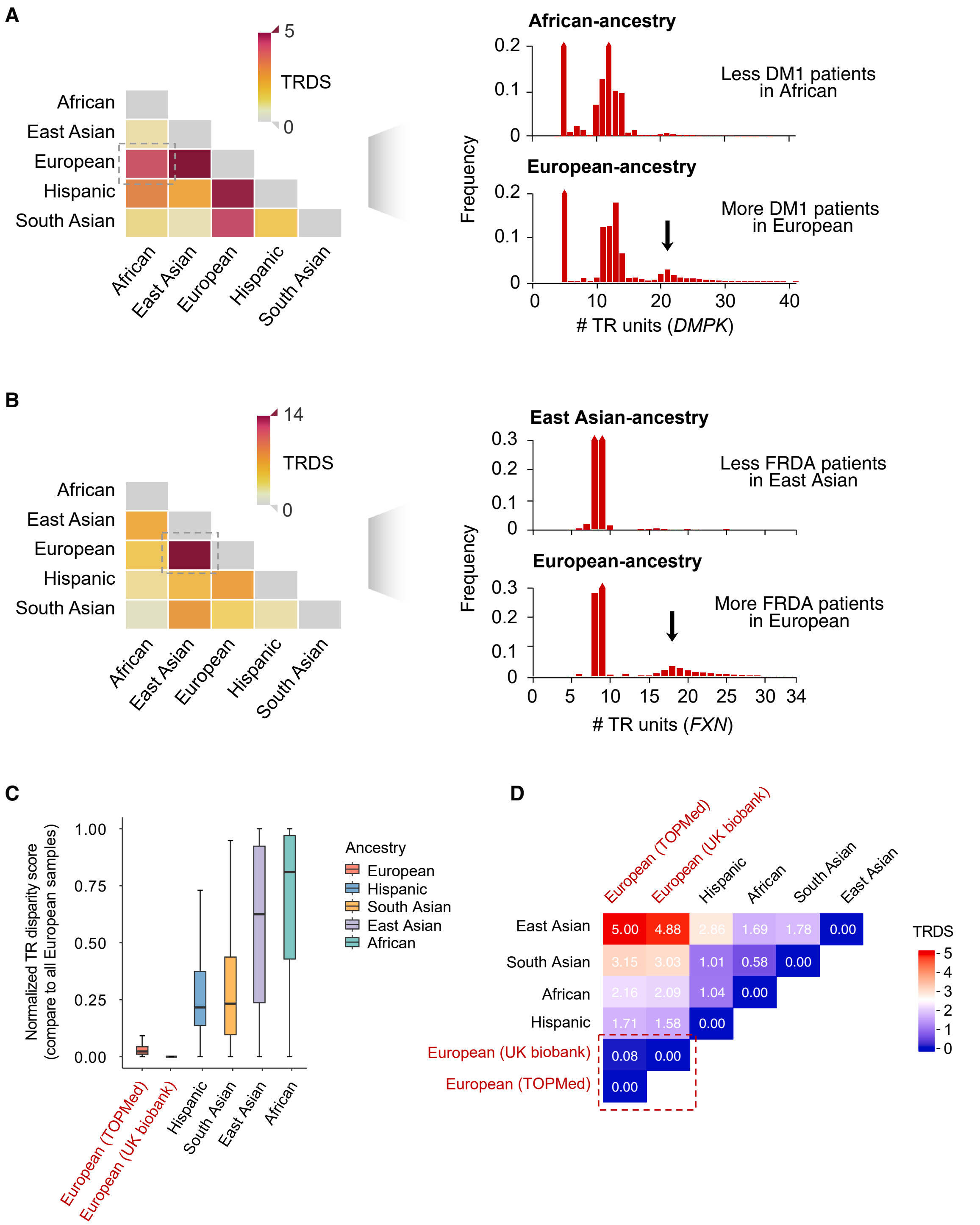


Figure S1. Example TR repeat unit number distributions and TR disparity scores between different samples, related to Figure 1

(A) The heatmap of TRDS and the histogram of TR unit number distribution for a CAG TR in the 3′ UTR of *DMPK*. The prevalence of expanded CAG TR units in the *DMPK* gene, responsible for myotonic dystrophy type 1 (DM1), is notably less common in individuals of African ancestry than in individuals of European ancestry.

(legend continued on next page)

(B) The heatmap of TRDS and the histogram of TR unit number distribution for a GAA TR in the intron of *FXN*. The prevalence of expanded GAA TR units in *FXN*, linked to Friedreich ataxia (FRDA), is lower in East Asian samples than in other ancestries. The histogram x axis denotes the number of TR unit, and the y axis denotes the frequency of each TR unit number.

(C) The boxplot shows much smaller TRDSs between samples of the same ancestry (European) from different cohorts (UK Biobank and TOPMed) than those observed between samples of different ancestries. TRDSs are normalized from 0 to 1. The center horizontal lines within the plot represent the median values, and the boxes are bounded by the 25th and 75th percentiles.

(D) The heatmap of TRDSs for an example TR (GGC repeat at the 5′ UTR of *DIP2B*) across different ancestries and cohorts.