Cell Genomics

CellPress
OPEN ACCESS

Preview

Uncovering complex trait heritability hidden in the repeatome

Po-Ru Loh[1,2,3,*]
[1]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA
[2]Center for Data Sciences, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA
[3]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
*Correspondence: poruloh@broadinstitute.org
https://doi.org/10.1016/j.xgen.2023.100461

Short tandem repeats (STRs) account for a substantial fraction of human genetic variation, but their contribution to complex human phenotypes is largely unknown. Margoliash et al. perform detailed genome-wide association analysis and fine-mapping of STRs in UK Biobank, identifying many STRs likely to influence variation in blood and serum traits.

The human genome contains over a million short tandem repeats (STRs), defined as segments of DNA in which a 1–6-base pair (bp) motif is repeated multiple times. STRs are among the most mutable sites in the genome and comprise a considerable fraction of genetic variation across individuals (~10%, depending on precisely how variation is quantified). However, despite numerous examples of pathogenic STRs identified to date,[1] the contribution of STRs to common human phenotypes is largely uncharacterized, in large part because STRs have been left out of the analytical pipelines used by genome-wide association studies (GWASs). As such, STRs represent a key class of polymorphisms at which "missing heritability" undiscovered from GWASs thus far might yet be found.[2] This hypothesis is highly plausible for multiple reasons, including (1) the sheer number of STRs in the genome, (2) the propensity of STRs to be multiallelic and poorly tagged by biallelic SNPs and indels currently considered in GWASs, and (3) recent work identifying effects of STRs on gene expression[3] and of slightly larger variable number tandem repeats (VNTRs) on complex traits and diseases.[4,5] In this issue of *Cell Genomics*, Margoliash et al.[6] thoroughly explore the influences of STR polymorphisms on 44 blood and biomarker phenotypes in UK Biobank (UKB),[7] demonstrating that STRs indeed contribute a sizable proportion of complex trait heritability and identifying several compelling examples of STRs that appear to underlie top GWAS signals.

A major technical barrier that has impeded inclusion of STRs in genetic association studies is the difficulty of genotyping STRs in large-scale genetic datasets. Direct genotyping of STRs requires whole-genome sequencing (WGS) data, which has been prohibitively expensive to generate for most GWAS cohorts. Moreover, even when short-read WGS is available, accurately genotyping STRs is still a formidable task because sequencing errors can mimic STR variation and STR alleles can exceed the lengths of sequencing reads. To overcome this challenge, Margoliash et al. instead used statistical imputation to estimate STR alleles present on SNP haplotypes of UKB participants, leveraging an SNP+STR haplotype reference panel they previously generated for common STRs.[8] This approach accurately imputed most STRs in the panel (as predicted based on previous benchmarks[8] and confirmed using WGS data available for a subset of UKB participants). The high accuracy of STR imputation achieved strongly motivates further development of computational tools expanding standard SNP-imputation pipelines to include STRs: doing so will enable STR analysis in all GWAS cohorts—enabling reuse of extensive prior investment in genotyping data resources—and will benefit from continuing efforts to generate larger, more comprehensive STR reference panels,[9] eventually from long-read sequencing.

The imputed STR dataset enabled Margoliash et al. to test STRs for association with 44 highly polygenic blood and biomarker phenotypes in UKB and—importantly—fine-map significant associations to identify a subset of STRs likely to causally influence phenotypes. Performing statistical fine-mapping to evaluate whether genetic associations reflect causal effects (rather than linkage disequilibrium with causal variants nearby) is essential for interpreting GWAS results, even when analyzing multiallelic variants. Margoliash et al. identified 119 STR-trait associations with high confidence of causality by developing an impressively rigorous fine-mapping pipeline (combining two complementary fine-mapping methods and running each method using several alternative settings). Notably, fine-mapping results exhibited some discordance between the two methods and across runs, which will be important for future studies to consider. Beyond prioritizing high-confidence STR-trait associations for follow up, the fine-mapping data (together with simulation results) further enabled an estimate that 5.2%–7.6% of causal variants identifiable from GWASs on the 44 blood and biomarker traits can be attributed to STRs. This estimate, which is broadly consistent with the overall contribution of STRs to human genetic variation, indicates the promise of incorporating STRs into standard GWAS pipelines.

The high-confidence fine-mapped STR-trait associations included several novel examples of STRs that appear to explain some of the strongest GWAS hits for blood cell traits. A strikingly polymorphic

CGG repeat in the promoter of *CBL* generated two very strong, statistically independent associations with platelet indices. These associations appeared to be driven by STR length and a common imperfection within the STR, which together completely accounted for all genetic association signal at the locus. Follow-up analyses of gene expression data from the Genotype-Tissue Expression (GTEx) project suggested a causal pathway in which longer repeat alleles decrease *CBL* expression, leading to increased platelet production. This hypothesis is consistent with previous literature implicating CBL in degradation of the thrombopoietin receptor[10] and motivates further investigation of the precise molecular mechanism by which STR variation in the promoter of *CBL* regulates its expression. Other noteworthy findings included a poly-A repeat within an intronic DNase I hypersensitivity site in *TAOK1* that associated strongly with mean platelet volume and a CCG repeat in the 5′ UTR of *BCL2L11* that associated strongly with eosinophil indices.

Overall, Margoliash et al.'s findings demonstrate the importance of considering STRs (and other oft-neglected types of genetic variation) in genetic association studies—and the feasibility of doing so by incorporating STRs into genotype imputation pipelines. The work is a technical *tour de force*—applying imputation, association testing, statistical fine-mapping, and follow-up analysis across populations and in gene expression data sets—that simultaneously shows what is achievable using existing methods and motivates further development of methods and resources for STR analysis. The high-confidence STR-trait associations identified by Margoliash et al. are sure to represent only the beginning of a stream of discoveries to come as STR association studies expand to consider more quantitative and disease traits, more genetic ancestries, more STRs (e.g., STRs too mutable to accurately impute or too difficult to genotype from short-read sequencing), and more complex STR-phenotype models (e.g., nonlinear effects of STR length and effects of within-repeat variation). We can look forward to unearthing more long-hidden heritability in the coming years.

ACKNOWLEDGMENTS

P.-R.L. was supported by US NIH grants R56 HG012698 and R01 HG013110, a Burroughs Wellcome Fund Career Award at the Scientific Interfaces, and the Next Generation Fund at the Broad Institute of MIT and Harvard.

DECLARATION OF INTERESTS

The author declares no competing interests.

REFERENCES

1. Hannan, A.J. (2018). Tandem repeats mediating genetic plasticity in health and disease. Nat. Rev. Genet. *19*, 286–298.
2. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.
3. Fotsing, S.F., Margoliash, J., Wang, C., Saini, S., Yanicky, R., Shleizer-Burko, S., Goren, A., and Gymrek, M. (2019). The impact of short tandem repeat variation on gene expression. Nat. Genet. *51*, 1652–1659.
4. Mukamel, R.E., Handsaker, R.E., Sherman, M.A., Barton, A.R., Zheng, Y., McCarroll, S.A., and Loh, P.R. (2021). Protein-coding repeat polymorphisms strongly shape diverse human phenotypes. Science *373*, 1499–1505.
5. Mukamel, R.E., Handsaker, R.E., Sherman, M.A., Barton, A.R., Hujoel, M.L.A., McCarroll, S.A., and Loh, P.R. (2023). Repeat polymorphisms underlie top genetic risk loci for glaucoma and colorectal cancer. Cell *186*, 3659–3673.e23.
6. Margoliash, J., Fuchs, S., Li, Y., Zhang, X., Massarat, A., Goren, A., and Gymrek, M. (2023). Polymorphic short tandem repeats make widespread contributions to blood and serum traits. Cell Genomics *3*, 100458.
7. 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.
8. Saini, S., Mitra, I., Mousavi, N., Fotsing, S.F., and Gymrek, M. (2018). A reference haplotype panel for genome-wide imputation of short tandem repeats. Nat. Commun. *9*, 4397.
9. 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.
10. Saur, S.J., Sangkhae, V., Geddis, A.E., Kaushansky, K., and Hitchcock, I.S. (2010). Ubiquitination and degradation of the thrombopoietin receptor c-Mpl. Blood *115*, 1254–1263.