Science Bulletin 69 (2024) 3633–3636

Contents lists available at ScienceDirect

Science Bulletin

journal homepage: www.elsevier.com/locate/scib

News & Views

Advancements and future perspectives of human tandem repeats

Wenbin Ye [1,*], Jason Sheng Li [1], Wei Li [*], Ya Cui [*]

Department of Biological Chemistry, School of Medicine, University of California, Irvine, Irvine, CA 92697, USA

Tandem repeats (TRs) are DNA sequences where specific nucleotide patterns are repeated consecutively along the DNA strand. TRs (~8.1% of the genome) constitute a substantial source of genetic diversity within the human genome [1], influencing complex traits and disease susceptibilities across populations [2,3]. Over the past 30 years, TR expansions have been linked to over 60 human diseases, predominantly neurological, including Huntington's disease, fragile X syndrome, and frontotemporal dementia [3–5]. The length of TRs varies, ranging from a few base pairs (bp) to thousands. They are typically categorized into two main types based on motif length: short tandem repeats (STRs), which consist of repeat units of 1–6 bp; and variable number tandem repeats, which have longer repeat units (≥7 bp) [1,4,5]. TRs exhibit high levels of polymorphism, with mutation rates decreasing as motif length increases [6–8]. TR variation can manifest in both coding and non-coding regions; notably, more than 90% of polymorphic TR loci are concentrated within intronic and intergenic regions [6]. Although TRs significantly contribute to genetic variation, they have largely been overlooked in genomic research due to technical challenges in sequencing and analysis, a historical focus on single nucleotide variants (SNVs), limitations of bioinformatic tools, biased database representation, and underappreciated biological significance [2–6,9,10]. This paper aims to describe the advancement of TR research in humans, focusing on large-scale characterizations, current resources, and limitations.

Understanding the historical context and technical evolution of TR profiling tools is essential to a comprehensive view of both progress and challenges in this field. In the 1970s, Southern blotting was first effectively used to identify STR loci and genotypes, and it remains in use today [1,2,5]. Later, technologies such as the polymerase chain reaction, optical genome mapping, and fluorescent fragment analysis were developed, enabling extensive validation of disease-associated TR loci [5]. For example, myotonic dystrophy type 1 (DM1) results from a CTG repeat expansion in the *DMPK* gene [11]. Recently, Yoon et al. [12] utilized Southern blotting to identify a long CTG repeat expansion of over a thousand TR units, diagnosing a previously undiscovered congenital DM1 case. These methods are precise but complex, time-consuming, and labor-intensive, limiting their use in genome-wide studies. Subsequently, the advent of whole-genome sequencing (WGS) revolutionized the field. Currently, short-read WGS (typically 100–150 bp) dominates genomic diversity research and clinical diagnostics due to its low base error rates, high coverage, and low cost. However, it struggles to accurately genotype large and complex TR expansions, potentially overlooking disease-associated TRs that exceed the read length [2,5]. Alternatively, long-read WGS (e.g., Oxford Nanopore Technologies (ONT) and PacBio sequencing) is capable of detecting long TR alleles (>10 kb), can effectively resolve large and complex TR expansions, and reliably identifies novel pathogenic loci [1–3,5]. However, long-read platforms generally suffer from lower throughput, and higher costs, making them more suitable for specialized studies. Consequently, most large-scale TR efforts continue to utilize short-read WGS.

Aside from advances in sequencing technologies, two other major factors have enabled the rapid growth of TR research: (1) the widespread availability of genomic data; (2) the development of improved TR genotyping tools. The completion of the Human Genome Project and the subsequent explosive increase of WGS data in the public domain greatly accelerated the exploration of human genetic variation. However, previous studies mainly focused on SNVs, indels, and structural variants, with TRs remaining largely uncharacterized due to methodological challenges and the need for specialized bioinformatics approaches [1]. Thus, despite the fact that data capable of generating novel TR insights became widely available, there was a lack of computational tools ready to genotype TRs within this data. In 2012, LobSTR became the pioneering tool for genotyping TRs, opening the doors of population-level TR research; however, it fails to capture alleles longer than the read length [5]. Even more recent methods such as popSTR and HipSTR fail to address this limitation [6,7]. In contemporary TR research, specialized tools such as GangSTR, STRetch, exSTRa, ExpansionHunter, and TREDPARSE (reviewed by Tanudisastro et al. [5]) can now detect expanded allele sizes larger than read lengths. However, these tools cannot reliably quantify TR unit numbers for these large expansions due to the read length of short-read WGS. Additionally, these different methods exhibit their own advantages and limitations. For example, while HipSTR cannot call large repeat expansions, it can extract the repeat sequence and haplotype. Both STRetch and exSTR perform well in determining pathogenic expansions, but are limited to known TR loci. Therefore,

* Corresponding authors.

E-mail addresses: wenbiy1@uci.edu (W. Ye), wei.li@uci.edu (W. Li), yac7@uci.edu (Y. Cui).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.scib.2024.08.007

the selection of a method should be guided by specific research objectives and informed by the limitations of each tool.

Despite the shortcomings of TR profiling tools, the evolution of TR research within the context of large-scale population data analytics has witnessed remarkable strides, providing unprecedented insights into human genetic variation at population scales. Leveraging high-throughput WGS data, particularly through short-read sequencing technologies, has catalyzed substantial progress in this domain. The progressive development of population-scale TR studies, reference maps, and resources can be traced over the last decade (Fig. 1a):

(1) In 2014, the profiling of approximately 0.7 million STR loci across more than 1000 individuals from Phase 1 of the 1000 Genomes Project (1 KGP) marked a pivotal milestone in the characterization of human STR variation [8]. This represented the first genome-wide, population-scale study of its kind. However, it was limited by low coverage (~5×), resulting in diminished accuracy and high rates of missing genotypes.

(2) In 2018, the first reference panel combining single nucleotide polymorphisms (SNPs) and STRs was developed using high-coverage (~30×) WGS data from nearly 2,000 individuals [10]. The aim was to provide a new haplotype panel resource to better explain missing heritability links in human diseases.

(3) In 2020, Trost et al. [11] identified a robust association between TR expansions and ASD, genotyping 0.32 million TRs and 2,588 TR loci linked to ASD from the largest ASD cohorts analyzed to date (n = 17,231, ~30× coverage). The detected TR expansions accounted for 2.6% of inherited ASD risk and showcased the importance of TR variation in human traits.

(4) The Genome Aggregation Database (gnomAD, https://gnomad.broadinstitute.org/) is a critical repository of human genetic

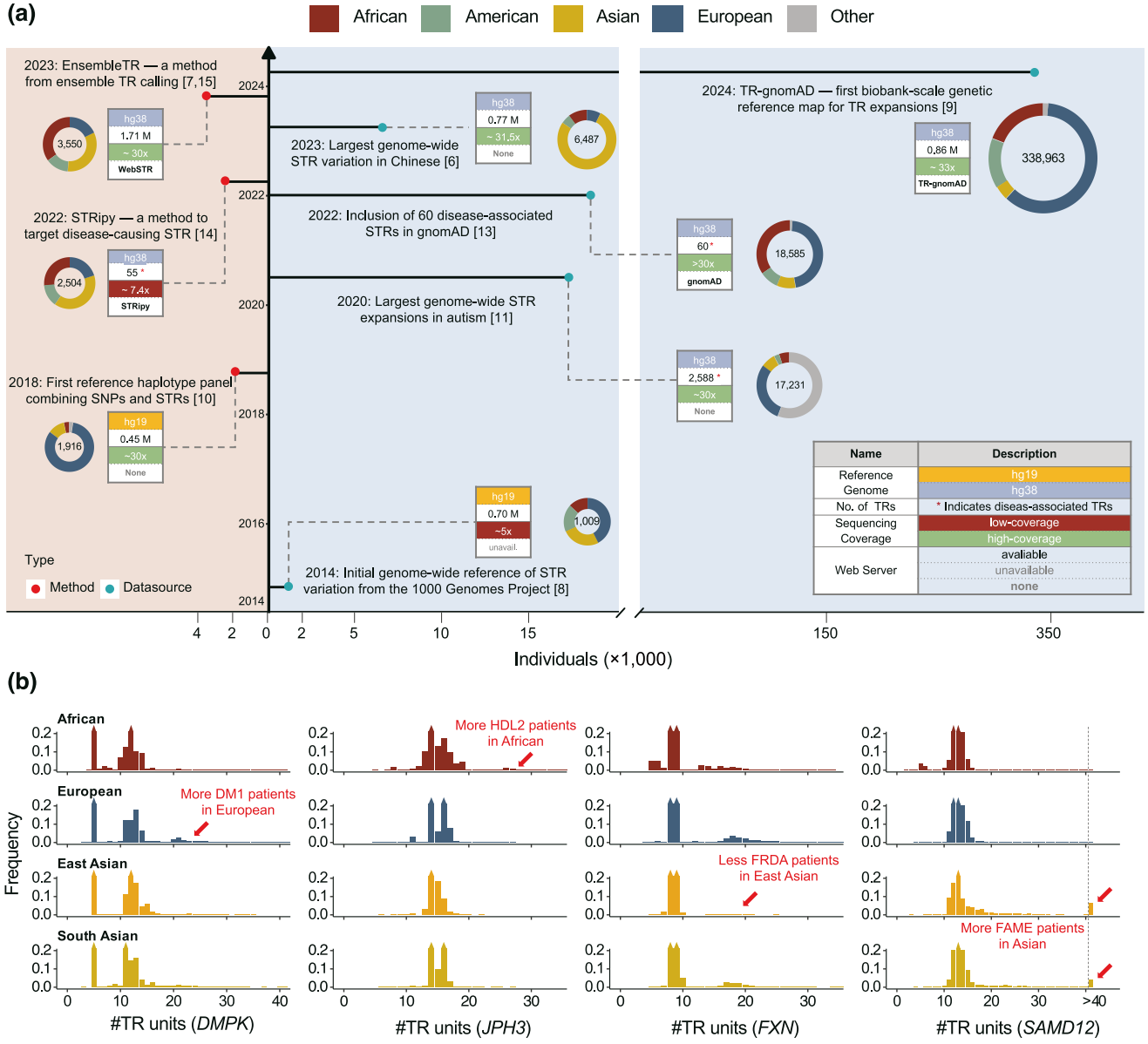


Fig. 1. Ancestry representation and timeline of TR profiling in large population-scale WGS studies up to May 2024. (a) Timeline of TR profiling in population-scale data analytics. The donut plot illustrates the proportions of ancestral components. Circle size represents the total number of samples, with the sample size indicated in the center. "M" in the table denotes million. (b) Ancestry-specific TR expansions. Histograms depicting the distribution of TR units are based on the TR-gnomAD database. The x-axis represents the number of TR units, and the y-axis shows the frequency of each TR unit number. DM1: Myotonic Dystrophy Type 1; HDL2: Huntington's Disease-Like 2; FRDA: Friedreich Ataxia; FAME: Familial Adult Myoclonic Epilepsy.

variation, serving as the gold standard reference map of human variation since its inception in 2017 [13]. In 2022, 60 disease-associated TRs were added to gnomAD. TRs were genotyped using ExpansionHunter on 18,511 high-quality WGS samples.

(5) Later in 2022, a new software, STRipy, was released to simplify the detection of STR expansions and the discovery of pathogenic STR loci. The inventors of STRipy cataloged 55 disease-associated STR loci from a cohort of 2,504 individuals from the 1KGP and provided an online tool for STR genotyping and pathogenicity assessment (https://stripy.org/) [14].

(6) The Han Chinese, despite being the largest ethnic group globally, have been underrepresented in genomic research. Recently, Shi et al. [6] endeavored to redress this imbalance by conducting a genome-wide investigation of STR variation in the Chinese populace, leveraging datasets from NwuWa (n = 3,983, ~31.5×) and 1 KGP (n = 2,504, ~33.3×). This study identified 366,013 polymorphic STRs, with 3,273 associated with gene expression modulation.

(7) EnsembleTR, launched in late 2023, uses an ensemble approach to derive high-quality genotypes, yielding over 1.7 million TR loci from the 1KGP and H3African datasets (~30× coverage) [7]. This consensus TR dataset, spanning 3,550 individuals, formed the foundation of WebSTR (https://webstr.ucsd.edu/), the most comprehensive human genome-wide TR variation resource developed up to this point [15].

(8) TR-gnomAD (https://wlcb.oit.uci.edu/TRgnomAD/), launched in 2024, offers a biobank-scale reference map for TR expansions, covering 338,963 individuals from diverse ancestries [9]. It represents a substantial stride forward in mapping human TR variation, encompassing ~100 times more individuals than the closest comparable resource, WebSTR. Importantly, it incorporated one of the most diverse cohorts to date, containing ~40% non-European individuals. Also of note, TR-gnomAD introduces a "TR disparity score" to identify ancestry-specific TRs and serves as a valuable resource for discovering disease-linked TR loci, Numerous known disease-associated TRs (e.g., *C9orf72* and *DMPK*) have been recapitulated through TR-gnomAD, highlighting its capabilities.

From these studies, it becomes clear that TRs exhibit significant human genetic variation, with allele length, expansion frequency, and disease association differing widely between populations. Historically, most studies have focused on European ancestry, underscoring the need for future TR resources to include and profile diverse and historically underrepresented ancestries. Diseases associated with TR expansions often show varied distribution by ancestry [3,6,7,9]. Some widely recognized examples include the previously mentioned DM1 disease, caused by an expanded CAG repeat in the *DMPK* gene; both the expansion and the disease are notably less common in individuals of African compared to those of European descent. In the Japanese population, a low prevalence of GAA expansions in the *FXN* gene corresponds with low rates of Friedreich's ataxia compared to other populations. Familial adult myoclonic epilepsy 1 is characterized by a TTTA repeat expansion in the *SAMD12* gene and preferentially occurs in patients of East and South Asian descent, reflecting the higher frequency of expansions in families from Japan, China, India, and Sri Lanka. Huntington disease-like 2 (HDL2) is caused by an inherited CAG/CTG repeat expansion in the *JPH3* gene; to date, HDL2 and its associated TR expansion have only ever been reported in families of African ancestry. Using the TR-gnomAD, we observed consistent ancestry-specific TR expansions in line with reported disparities in disease prevalence (Fig. 1b).

Besides improving the inclusion of underrepresented populations, there is still significant progress to be made in TR research. Importantly, despite recent advances in TR cataloging, only a portion of TRs within the human genome have been profiled. For example, TR-gnomAD genotypes approximately 0.86 million TRs, but still fails to capture the majority of TRs in the human genome. Most tools attempt to call TRs using short-read WGS, which struggles to accurately profile TRs larger than the read length. This limitation leads to the underestimation of allele lengths for large TRs (>150 bp), affecting the accuracy of *de novo* TR genotyping and the identification of pathogenic TR expansions. In the previously mentioned report of a long expansion in DM1 disease, ExpansionHunter estimated a repeat count of 64 compared to Southern blotting, which more precisely identified the expansion to have a count of 1,171 [12]. Additionally, long-read sequencing is far better suited for genotyping TR loci with complex and expanded repeat structures or in low-complexity, GC-rich, or repetitive regions that are currently poorly profiled by short-read sequencing. For example, the 1000 Genomes Project ONT Sequencing Consortium (1KGP-ONT) has been leveraging ONT long-read sequencing to enhance our understanding of human variation. Recent advances in long-read sequencing technologies, such as the development of more accurate base-calling algorithms and improved library preparation techniques, have further increased the resolution and reliability of TR detection (reviewed by Tanudisastro et al. [5]). In the future, large long-read sequencing datasets will be crucial for validating current TR tools and genotypes.

Taken together, TR research is an emerging field with substantial room for improvement, both in the quality and breadth of datasets and in the development of more accurate TR genotyping tools. As population-scale resources expand to include more individuals and incorporate additional TRs as they are discovered, we move closer to fitting another piece of the human genetic variation puzzle.

Conflict of interest

The authors declare they have no conflict of interest.

Acknowledgments

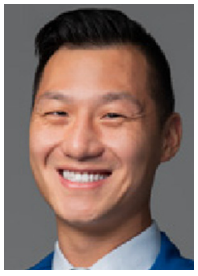
We thank Xueyi Teng, Zhuoxin Wu and Chaorong Chen for helpful discussions.

References

[1] English AC, Dolzhenko E, Ziaei Jam H, et al. Analysis and benchmarking of small and large genomic variants across tandem repeats. Nat Biotechnol 2024. https://doi.org/10.1038/s41587-024-02225-z.
[2] Leitão E, Schröder C, Depienne DC. Identification and characterization of repeat expansions in neurological disorders: Methodologies, tools, and strategies. Rev Neurol 2024;180:383–92.
[3] Depienne C, Mandel JL. 30 years of repeat expansion disorders: What have we learned and what are the remaining challenges? Am J Hum Genet 2021;108:764–85.
[4] Hannan AJ. Tandem repeats mediating genetic plasticity in health and disease. Nat Rev Genet 2018;19:286–98.
[5] Tanudisastro HA, Deveson IW, Dashnow H, et al. Sequencing and characterizing short tandem repeats in the human genome. Nat Rev Genet 2024;25:460–75.
[6] Shi Y, Niu Y, Zhang P, et al. Characterization of genome-wide STR variation in 6487 human genomes. Nat Commun 2023;14:2092.
[7] Jam HZ, Li Y, DeVito R, et al. A deep population reference panel of tandem repeat variation. Nat Commun 2023;14:6711.
[8] Willems T, Gymrek M, Highnam G, et al. The landscape of human STR variation. Genome Res 2014;24:1894–904.
[9] Cui Y, Ye W, Li JS, et al. A genome-wide spectrum of tandem repeat expansions in 338,963 humans. Cell 2024;187:2336–41.
[10] Saini S, Mitra I, Mousavi N, et al. A reference haplotype panel for genome-wide imputation of short tandem repeats. Nat Commun 2018;9:4397.
[11] Trost B, Engchuan W, Nguyen CM, et al. Genome-wide detection of tandem DNA repeats that are expanded in autism. Nature 2020;586:80–6.
[12] Yoon JG, Lee S, Cho J, et al. Diagnostic uplift through the implementation of short tandem repeat analysis using exome sequencing. Eur J Hum Genet 2024;32:584–7.
[13] Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 2020;581:434–43.

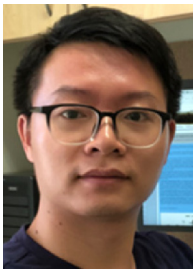
[14] Halman A, Dolzhenko E, Oshlack A. STRipy: A graphical application for enhanced genotyping of pathogenic short tandem repeats in sequencing data. Hum Mutat 2022;43:859–68.
[15] Lundström O, Adriaan Verbiest M, Xia F, et al. WebSTR: A population-wide database of short tandem repeat variation in humans. J Mol Biol 2023;435:168260.

Wenbin Ye is a postdoctoral scholar in the Department of Biological Chemistry at the UCI School of Medicine. She earned her Ph.D. in systems engineering from Xiamen University. Her research primarily focuses on developing novel bioinformatics algorithms to elucidate the molecular mechanisms and functional consequences of alternative polyadenylation in various biological processes.



Jason Sheng Li is a Ph.D. candidate in the Department of Biological Chemistry at the UCI School of Medicine. His research centers on epigenetic biomarkers of disease and the use of epigenome-wide information to understand the molecular basis of oncogenesis.

Wei Li is the Grace B. Bell Endowed Chair and Professor of Bioinformatics in the Division of Computational Biomedicine and the Department of Biological Chemistry at the UCI School of Medicine. His research bridges computational biology, epigenetics, RNA regulation, liquid biopsy, and human genetics, focusing on transforming genomics data into actionable medical insights. His work has yielded significant discoveries of novel epigenetic mechanisms, biomarkers, and therapeutic targets for various human diseases.



Ya Cui is a research assistant professor of computational biology and human genetics at the UCI School of Medicine. He received his Ph.D. in computational biology from the Institute of Biophysics, Chinese Academy of Sciences. His research focused on understanding the genetic basis of coding and noncoding variants in human complex traits and diseases.