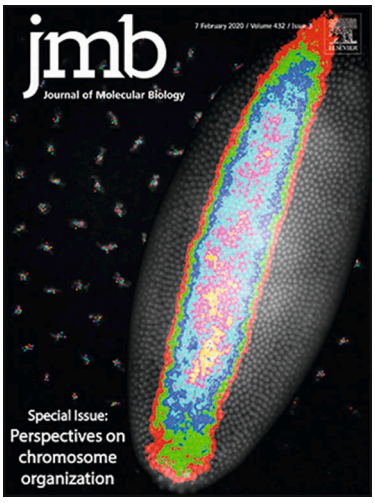
Journal Pre-proofs



Database

WebSTR: a population-wide database of short tandem repeat variation in humans

Oxana Sachenkova Lundström, Max Verbiest, Feifei Xia, Helyaneh Ziaei jam, Inti Zlobec, Maria Anisimova, Melissa Gymrek

PII: S0022-2836(23)00371-6
DOI: https://doi.org/10.1016/j.jmb.2023.168260
Reference: YJMBI 168260

To appear in: *Journal of Molecular Biology*

Received Date: 27 March 2023
Revised Date: 29 August 2023
Accepted Date: 29 August 2023

Please cite this article as: O. Sachenkova Lundström, M. Verbiest, F. Xia, H. Ziaei jam, I. Zlobec, M. Anisimova, M. Gymrek, WebSTR: a population-wide database of short tandem repeat variation in humans, *Journal of Molecular Biology* (2023), doi: https://doi.org/10.1016/j.jmb.2023.168260

WebSTR: a population-wide database of short tandem repeat variation in humans

List of authors and institutions

Oxana (Sachenkova) Lundström[1,2,3],

Max Verbiest[3,4,5]

Feifei Xia[3,4,5]

Helyaneh Ziaei jam[7],

Inti Zlobec[6],

Maria Anisimova[3,4 *]

Melissa Gymrek[7, 8 *]

1. Department of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden
2. Vildly AB, Kalmar, Sweden
3. Institute of Computational Life Sciences, School of Life Sciences and Facility Management, Zürich University of Applied Sciences (ZHAW), Waedenswil, Switzerland
4. Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland
5. Department of Molecular Life Sciences, University of Zurich, Zurich, Switzerland
6. Institute of Pathology, University of Bern, Switzerland
7. Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, USA
8. Department of Medicine, University of California San Diego, La Jolla, CA, USA

*Correspondence should be addressed to anis@zhaw.ch or mgymrek@ucsd.edu.

Abstract

Short tandem repeats (STRs) are consecutive repetitions of one to six nucleotide motifs. They are hypervariable due to the high prevalence of repeat unit insertions or deletions primarily caused by polymerase slippage during replication. Genetic variation at STRs has been shown to influence a range of traits in humans, including gene expression, cancer risk, and autism. Until recently STRs have been poorly studied since they pose significant challenges to bioinformatics analyses. Moreover, genome-wide analysis of STR variation in population-scale cohorts requires large amounts of data and

computational resources. However, the recent advent of genome-wide analysis tools has resulted in multiple large genome-wide datasets of STR variation spanning nearly two million genomic loci in thousands of individuals from diverse populations.

Here we present WebSTR, a database of genetic variation and other characteristics of genome-wide STRs across human populations. WebSTR is based on reference panels of more than 1.7 million human STRs created with state of the art repeat annotation methods and can easily be extended to include additional cohorts or species. It currently contains data based on STR genotypes for individuals from the 1000 Genomes Project, H3Africa, the Genotype-Tissue Expression (GTEx) Project and colorectal cancer patients from the TCGA dataset.

WebSTR is implemented as a relational database with programmatic access available through an API and a web portal for browsing data. The web portal is publicly available at http://webstr.ucsd.edu.

Keywords

next-generation sequencing, human genetic variation, disease-associated variants, short tandem repeats, gene regulatory regions, transcription factor binding sites, genotyping, reference panel, genome-wide association studies, WebSTR, allele frequencies, mutation rates, web portal, database, API

Introduction

Advances in next-generation sequencing (NGS) have enabled generation of deep catalogs of human genetic variation, and resulted in the discovery of an extensive set of disease-associated variants. The majority of NGS applications have focused on single nucleotide polymorphisms (SNPs) or short insertions and deletions (indels). Tandem repeats are an additional rich source of genetic variability that have been largely overlooked due to technical difficulties in obtaining accurate genotypes. Here we focus primarily on short tandem repeats (STRs) with repeat unit lengths of 1-6 bp. Collectively, STRs span around 3% of the human genome, more than the entire protein coding exome [1]. STRs are enriched in gene regulatory regions ([2],[3]), and variation in repeat copy number can impact gene regulation through a variety of mechanisms, including modifying transcription factor binding sites, altering DNA methylation patterns [4], or other means. Larger expansions of the number of repeated units in STRs are implicated in dozens of disorders [5], such as Huntington's Disease [6] and Fragile X Syndrome [7], and more modest stepwise changes have been implicated in complex traits including blood and lipid biomarkers ([8], [9]). STRs have also served as genetic markers for diagnostics in cancer research and play a role in many cancers, including colorectal [10] and breast cancers [11].

Due to the highly polymorphic nature [1] of STRs and high rates of sequencing errors in these regions, short-read NGS pipelines struggle with STRs [12] and routinely filter them out from further analysis [13]. However, recent advances in both sequencing technologies and computational approaches enable accurate genotyping of STRs from sequencing data. Multiple tools are now available for genotyping a range of STRs. For example, GangSTR [14,15] and ExpansionHunter [14] incorporate multiple properties of paired-end reads into models capable of genotyping both normal length and expanded repeats. HipSTR [16] is a haplotype-based method that takes into account the repeat sequence, as well as the length, of each allele. HipSTR has shown higher accuracy than GangSTR [17] but only considers repeats fully enclosed within the read length. Importantly, these tools require a reference panel of known STR locations. Repeat annotation remains a difficult task, with multiple repeat detector programs often producing contradictory results [18]. TRAL [19,20] introduces a statistical framework to help address this challenge using a multi-step tandem repeat annotation workflow. Applications of the methods described above have revealed novel impacts of STRs on gene expression [21],[22],[23], and protein functions in neurodevelopmental conditions [24], [25].

Performing STR annotation and genotyping in sufficiently large cohorts for downstream applications such as genome-wide association studies (GWAS) or expression quantitative trait loci (eQTL) analysis presents a significant computational burden. Further, implementing these pipelines and interpreting the results can require domain expertise and programming knowledge. Web interfaces, such as the recently published gnomAD browser [26], can provide informative statistics and visualizations of genotyping summary statistics that are easily accessible and interpretable even by researchers without bioinformatics expertise. While gnomAD includes data for a small number of known pathogenic STRs, no existing web browser presents genome-wide variation data for a large set of STRs.

We present WebSTR, the first comprehensive resource for genome-wide STR variation and other characteristics of STRs. WebSTR is based on reference panels of more than 1.7 million human STRs created with state of the art repeat annotation methods. It currently combines data from five different studies as well as new data from the Sinergia-CRC cohort described here, and can be easily extended to incorporate additional cohorts or STR loci. WebSTR provides a browsable web interface for exploring summary level data at each STR, including allele frequencies, mutation rates, and trait associations. It also supports programmatic access to the data to enable further analysis or integration with third party tools. We envision WebSTR will serve as a valuable community resource and facilitate future studies of genome-wide STR variation.

Results

WebSTR database

WebSTR (http://webstr.ucsd.edu), aggregates data from several studies on STRs that collectively annotate over 1.7 million loci in the human genome. These studies present

data from multiple cohorts, including data on colorectal cancer (CRC) from TCGA [27], the 1000 Genomes Project [28], H3Africa [29], and the Genotype-Tissue Expression Project [30] (**Table 1**). For each STR, multiple metrics are reported, including average repeat length, population-specific allele frequencies, associations with expression levels of nearby genes, mutation rates, and imputation metrics. Reference genome coordinates for each STR as well as the repeated motif are also stored and displayed. It is important to note that for complex STRs we currently report only the consensus motif based on the reference genome, this is due to the differences in how repeat motifs are handled by different genotyping tools.

These datasets are stored in a PostgreSQL database, which is accessed by the web portal through an application programming interface (API; **Figures S1-2**).

WebSTR interface overview

Users can access the data through a web portal by specifying a reference genome build and searching for a gene or genomic region of interest (**Figure 1A.1**). A valid search takes the user to *the region-level page*. The top of this page displays the exon/intron structure of genes in the region as well as the genomic location of each STR in the region, shown as dots color-coded by the repeat unit length. This page additionally displays a table of all STRs identified in the region (**Figure 1A.2**) which includes the coordinates, repeat unit sequence, and length of the repeat in the reference genome.

From the region-level page, users can select a specific STR of interest to view the *locus-level* page. This page displays the STR sequence and its genomic context based on a reference genome and contains histograms to visualize population-specific allele frequencies at the locus. It additionally shows locus-level statistics collected from various studies, including imputation quality metrics, estimated mutation rate, and trait associations. Results from different studies can be displayed or hidden by clicking on the respective title boxes.

Programmatic access to the WebSTR database

For accessing data on repeat locations and variation for a larger genomic region or for several genes at once, we made annotations using the hg38 assembly version available through an API hosted at http://webstr-api.ucsd.edu/. A set of endpoints allows users to access gene annotation features that are used for WebSTR visualizations. Further, all repeats from different reference panels can be queried by Ensembl gene identifiers, gene names or genomic coordinates. Variation data is available on a cohort level on the repeats endpoint or can be queried by a repeat ID if it is available on a locus level. By default, the WebSTR-API returns results in JSON format but an option for streaming downloaded

data as a csv file is also available. Documentation for the available endpoints is available on the main page of the WebSTR-API (http://webstr-api.ucsd.edu/).

Although WebSTR is publicly hosted and currently includes datasets listed in Table 1, users may also deploy a local version of the WebSTR browser and API to explore data from their own cohort of interest by extending existing Python modules. Instructions for doing so are provided on the WebSTR github page (https://github.com/gymrek-lab/webstr). Additional cohorts will be added to the public WebSTR version as they become available and based on community requests.

Case study using the API

To illustrate the utility of the WebSTR API to enable custom applications, we set up a comprehensive case study to investigate and visualize the associations between STR variations, gene expression and somatic mutation in colorectal cancer (CRC) patients (**Figure 1B**). Data on STR locations and locus variation obtained from the WebSTR database was integrated with controlled variation data on the individual patient level, as well as the clinical and transcriptomics data available from the TCGA Consortium [27]. This case study was performed on data from 28 normal and 377 tumor samples of CRC patients from TCGA projects COAD and READ. The dashboard allows users to input a gene of interest and displays an overview of all nearby STRs and their length variations for the targeted gene in all tumor samples (**Figure 1B.1**). By selecting an STR, users can explore the relationships of STR length variation and gene expression in both normal and tumor samples (**Figure S2**). When normal samples were available, comparisons between paired normal and tumor samples were provided. For tumor samples, users can select different cancer subtypes to refine visualization and analysis. In addition, the STR length distribution and STR-gene expression graphs were generated based on the targeted gene mutation status in tumor samples (**Figure 1B.1**). This tool is being used to explore potentially clinically relevant STR loci in colorectal cancer.

Materials and Methods

Platform architecture

An overview of the WebSTR architecture along with implementation choices is available in **Figure S3**. All reference STR panels, corresponding gene annotations and variation data for each cohort are stored in a relational database (PostgreSQL [31]) using the database schema shown in **Figure S4**. The backend of WebSTR, a RESTful (REpresentational State Transfer) API to access the data, is implemented using the FastAPI Python framework (v.0.68.0) [32]. SQLAlchemy (v. 1.4.23) [33] and SQLModel (v.0.0.4) [34] are used as the Object Relational Mapper (ORM) between PostgreSQL and

Python. API documentation was auto generated using Redocly [35]. The front-end of the interactive web application is implemented using Python/Flask, together with Bootstrap and Javascript. The Plotly package (v.2.16.1) [36] is used for interactive data visualizations.

Genome-wide reference panels for Human STRs

EnsembleTR Panel and older studies available through WebSTR

The ensembletr_hg38 panel is based on the GRCh38 reference assembly and contains 1.7 million unique autosomal STRs based on a combined set of TRs genotyped by four separate methods (HipSTR [16], GangSTR [14,15], ExpansionHunter [14], and AdVNTR [37]). Detailed procedures on how this panel was constructed are described in [38]. Previous panels for hg19 are based on the published HipSTR hg19 reference (https://github.com/HipSTR-Tool/HipSTR-references/) and include STR data compiled from multiple sources described in **Table 1**.

Sinergia-CRC Panel

The Sinergia-CRC STR panel (gangstr_crc_hg38 in **Table 1**) is based on the GRCh38 reference assembly and is newly presented in this study. This panel is an annotation of all 19,814 protein coding genes included in GENCODE v22 [39] (the same version used by TCGA), plus 5 kb upstream of transcription start sites. It contains 1,548,993 STRs with repeat units of 1-6bp, where the copy number for repeat units sizes 1, 2, 3, 4, 5 and 6 is at least 9, 4, 4, 3, 3 and 3 units in the reference genome, respectively.

Repeat detection was performed using the tandem repeat annotation library (TRAL) [20]. TRAL enables running various repeat detection algorithms, integrating their outputs and processing the results. STRs were detected with PHOBOS [40], XSTREAM [41] and TRF [42] using their default parameter settings in TRAL. To filter out false positives, STRs with divergence >= 0.1 or p-value >= 0.05 were discarded. The panel was made non-redundant and subsequently refined using circular-profile hidden Markov models [18]. More details are available in **Supplementary Material**.

A generalized pipeline for STR genotyping from NGS data is illustrated in **Figure 2**. The Sinergia-CRC repeats were genotyped using GangSTR (v.2.5) (default parameters and a --nonuniform flag due to exome data) [15] on TCGA-412 (**Figure S1**), a subset of genomes from patients with colorectal cancer available through the TCGA consortium [27]. This subset of data has been formed for other studies performed within the Sinergia study; the logic behind its creation is described in [43].

The resulting genotyping VCF files was filtered using TRTools (v5.0.1) [44], the dumpSTR utility to filter on call coverage and remove all calls that are not spanning or bounding or where ML estimate is outside of the confidence interval (parameters: --gangstr-min-call-

DP 20, --gangstr-max-call-DP 1000, ---gangstr-filter-spanbound-only, --gangstr-filter-badCI).

Patient-level length variation data of each STR is available only upon approval from the TCGA consortium as it can potentially be used to identify the patient. Cohort-level summary statistics describing locus-level patterns of genetic variation is stored in the WebSTR database and is publicly available for each STR. The following parameters can be accessed: the average amount of repeat units difference that was found for this locus between healthy and tumor sample (avg_size_diff), the total number of patients where the locus was called in both the healthy and tumor sample (total_calls), the fraction of patients where a difference in repeat size was observed between healthy and tumor for this locus (frac_variable).

Code and data availability

Reference panels created for studies stored in WebSTR are freely accessible at the WebSTR downloads page (http://webstr.ucsd.edu/downloads). Additional data needed to fully set up the project locally is available upon request. Code and instructions to set up WebSTR database and API locally are available on https://github.com/acg-team/webSTR-API/. Code and instructions for a local set up of the web portal are available on https://github.com/gymrek-lab/webstr

Authors' contributions

Oxana Lundström: Conceptualization, Methodology, Software, Writing- Original draft preparation, Project administration. **Max Verbiest**: Data curation, Software, Writing - Review & Editing. **Feifei Xia**: Visualization. **Helyaneh Ziaei Jam**: Data curation, Software. **Inti Zlobec**: Validation, Supervision. **Maria Anisimova:** Writing- Reviewing and Editing, Supervision, Funding acquisition. **Melissa Gymrek**: Conceptualization, Supervision, Funding acquisition, Writing- Reviewing and Editing.

Acknowledgements

O.L, M.V., and F.X. were supported by the SNSF Sinergia grant CRSII5_193832 to M.A. and I.Z. and the EU Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 823886. Hosting, maintenance and development of WebSTR was also partially funded by the NIH/NHGRI grants R01HG010885 and R01HG010149 to M.G. We also acknowledge Richard Yanicky, who helped in the initial design of the WebSTR website.

Figure legends

Figure 1. A) Overview of WebSTR web browser features. Top=Screen capture of a result on the region-level page for an example search of WebSTR using a gene name query; bottom=example allele frequency visualizations displayed on the locus-level page. Each color denotes a different population. B) Example application of the WebSTR API to enable visualization of STR variation in the TCGA CRC dataset. Using an input field for a gene of interest (1) users can explore STRs in this region and their length variations (number of repeat units) in tumor samples. (2) Interactive visualization tool that allows users to explore the relationship between STR length variation and somatic mutations. The violin plot shows the length distribution of selected STRs in tumor samples with (+) and without (-) point mutations in the selected gene. The split violin plot further shows the relationship of STR length and gene expression across mutated and unmutated tumor samples.

Figure 2. STR genotyping pipeline. The first step in this workflow is creating a reference panel of STRs, a set of repeats found in the reference genome of interest. At this stage various tandem repeat prediction tools can be used. STR panel and sequencing data (whole genome or whole exome sequencing) will then serve as inputs to the STR Genotyping tools. The output of this step is usually variant calls in the vcf format that can be further filtered and processed, for easier access and analysis they are then imported to WebSTR database using a set of Python scripts.

References

[1] T. Willems, M. Gymrek, G. Highnam, 1000 Genomes Project Consortium, D. Mittelman, Y. Erlich, The landscape of human STR variation, Genome Res. 24 (2014) 1894–1904. https://doi.org/10.1101/gr.177774.114.

[2] M.D. Vinces, M. Legendre, M. Caldara, M. Hagihara, K.J. Verstrepen, Unstable tandem repeats in promoters confer transcriptional evolvability, Science. 324 (2009) 1213–1216. https://doi.org/10.1126/science.1170097.

[3] S. Sawaya, A. Bagshaw, E. Buschiazzo, P. Kumar, S. Chowdhury, M.A. Black, N. Gemmell, Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements, PLoS One. 8 (2013) e54710. https://doi.org/10.1371/journal.pone.0054710.

[4] R. Gemayel, M.D. Vinces, M. Legendre, K.J. Verstrepen, Variable tandem repeats accelerate evolution of coding and regulatory sequences, Annu. Rev. Genet. 44

(2010) 445–477. https://doi.org/10.1146/annurev-genet-072610-155046.

[5] S.M. Mirkin, Expandable DNA repeats and human disease, Nature. 447 (2007) 932–940. https://doi.org/10.1038/nature05977.

[6] O.W.J. Quarrell, H.M. Brewer, F. Squitieri, R.A. Barker, Juvenile Huntington's Disease: And Other Trinucleotide Repeat Disorders, Oxford University Press, 2009. https://books.google.com/books/about/Juvenile_Huntington_s_Disease.html?hl=&id=NlUeXtQeHw4C.

[7] R. Willemsen, F. Kooy, Fragile X Syndrome: From Genetics to Targeted Treatment, Academic Press, 2017. https://play.google.com/store/books/details?id=ItB1DQAAQBAJ.

[8] J. Margoliash, S. Fuchs, Y. Li, A. Massarat, A. Goren, M. Gymrek, Polymorphic short tandem repeats make widespread contributions to blood and serum traits, bioRxiv. (2022). https://doi.org/10.1101/2022.08.01.502370.

[9] M. Verbiest, M. Maksimov, Y. Jin, M. Anisimova, M. Gymrek, T. Bilgin Sonay, Mutation and selection processes regulating short tandem repeats give rise to genetic and phenotypic diversity across species, J. Evol. Biol. 36 (2023) 321–336. https://doi.org/10.1111/jeb.14106.

[10] S. Popat, R. Hubner, R.S. Houlston, Systematic Review of Microsatellite Instability and Colorectal Cancer Prognosis, Journal of Clinical Oncology. 23 (2005) 609–618. https://doi.org/10.1200/jco.2005.01.086.

[11] W. Zhang, Y.Y. Yu, Polymorphisms of short tandem repeat of genes and breast cancer susceptibility, Eur. J. Surg. Oncol. 33 (2007) 529–534. https://doi.org/10.1016/j.ejso.2006.11.027.

[12] O.K. Tørresen, B. Star, P. Mier, M.A. Andrade-Navarro, A. Bateman, P. Jarnot, A. Gruca, M. Grynberg, A.V. Kajava, V.J. Promponas, M. Anisimova, K.S. Jakobsen, D. Linke, Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases, Nucleic Acids Res. 47 (2019) 10994–11006. https://doi.org/10.1093/nar/gkz841.

[13] M.D. Cao, S. Balasubramanian, M. Bodén, Sequencing technologies and tools for short tandem repeat variation detection, Brief. Bioinform. 16 (2015) 193–204. https://doi.org/10.1093/bib/bbu001.

[14] E. Dolzhenko, M.F. Bennett, P.A. Richmond, B. Trost, S. Chen, J.J.F.A. van Vugt, C. Nguyen, G. Narzisi, V.G. Gainullin, A.M. Gross, B.R. Lajoie, R.J. Taft, W.W. Wasserman, S.W. Scherer, J.H. Veldink, D.R. Bentley, R.K.C. Yuen, M. Bahlo, M.A. Eberle, ExpansionHunter Denovo: a computational method for locating known and novel repeat expansions in short-read sequencing data, Genome Biol. 21 (2020) 102. https://doi.org/10.1186/s13059-020-02017-z.

[15] N. Mousavi, S. Shleizer-Burko, R. Yanicky, M. Gymrek, Profiling the genome-wide landscape of tandem repeat expansions, Nucleic Acids Res. 47 (2019) e90. https://doi.org/10.1093/nar/gkz501.

[16] T. Willems, D. Zielinski, J. Yuan, A. Gordon, M. Gymrek, Y. Erlich, Genome-wide profiling of heritable and de novo STR variations, Nat. Methods. 14 (2017) 590–592. https://doi.org/10.1038/nmeth.4267.

[17] A. Halman, A. Oshlack, Accuracy of short tandem repeats genotyping tools in whole exome sequencing data, F1000Res. 9 (2020) 200. https://doi.org/10.12688/f1000research.22639.1.

[18] E. Schaper, A.V. Kajava, A. Hauser, M. Anisimova, Repeat or not repeat?—Statistical validation of tandem repeat prediction in genomic sequences, Nucleic Acids Research. 40 (2012) 10005–10017. https://doi.org/10.1093/nar/gks726.

[19] E. Schaper, A. Korsunsky, J. Pečerska, A. Messina, R. Murri, H. Stockinger, S. Zoller, I. Xenarios, M. Anisimova, TRAL: tandem repeat annotation library, Bioinformatics. 31 (2015) 3051–3053. https://doi.org/10.1093/bioinformatics/btv306.

[20] M. Delucchi, P. Näf, S. Bliven, M. Anisimova, TRAL 2.0: Tandem Repeat Detection With Circular Profile Hidden Markov Models and Evolutionary Aligner, Front Bioinform. 1 (2021) 691865. https://doi.org/10.3389/fbinf.2021.691865.

[21] T. Bilgin Sonay, T. Carvalho, M.D. Robinson, M.P. Greminger, M. Krützen, D. Comas, G. Highnam, D. Mittelman, A. Sharp, T. Marques-Bonet, A. Wagner, Tandem repeat variation in human and great ape populations and its impact on gene expression divergence, Genome Res. 25 (2015) 1591–1599. https://doi.org/10.1101/gr.190868.115.

[22] M. Gymrek, T. Willems, A. Guilmatre, H. Zeng, B. Markus, S. Georgiev, M.J. Daly, A.L. Price, J.K. Pritchard, A.J. Sharp, Y. Erlich, Abundant contribution of short tandem repeats to gene expression variation in humans, Nat. Genet. 48 (2016) 22–29. https://doi.org/10.1038/ng.3461.

[23] J. Quilez, A. Guilmatre, P. Garg, G. Highnam, M. Gymrek, Y. Erlich, R.S. Joshi, D. Mittelman, A.J. Sharp, Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans, Nucleic Acids Res. 44 (2016) 3750–3762. https://doi.org/10.1093/nar/gkw219.

[24] I. Mitra, B. Huang, N. Mousavi, N. Ma, M. Lamkin, R. Yanicky, S. Shleizer-Burko, K.E. Lohmueller, M. Gymrek, Patterns of de novo tandem repeat mutations and their role in autism, Nature. 589 (2021) 246–250. https://doi.org/10.1038/s41586-020-03078-7.

[25] J. Wen, B. Trost, W. Engchuan, M. Halvorsen, L.M. Pallotto, A. Mitina, N. Ancalade, M. Farrell, I. Backstrom, K. Guo, G. Pellecchia, B. Thiruvahindrapuram, P. Giusti-Rodriguez, J.D. Rosen, Y. Li, H. Won, P.K.E. Magnusson, U. Gyllensten,

A.S. Bassett, C.M. Hultman, P.F. Sullivan, R.K.C. Yuen, J.P. Szatkiewicz, Rare tandem repeat expansions associate with genes involved in synaptic and neuronal signaling functions in schizophrenia, Mol. Psychiatry. 28 (2023) 475–482. https://doi.org/10.1038/s41380-022-01857-4.

[26] S. Gudmundsson, M. Singer-Berk, N.A. Watts, W. Phu, J.K. Goodrich, M. Solomonson, Genome Aggregation Database Consortium, H.L. Rehm, D.G. MacArthur, A. O'Donnell-Luria, Variant interpretation using population databases: Lessons from gnomAD, Hum. Mutat. 43 (2022) 1012–1030. https://doi.org/10.1002/humu.24309.

[27] The Cancer Genome Atlas Network, Comprehensive molecular characterization of human colon and rectal cancer, Nature. 487 (2012) 330–337. https://doi.org/10.1038/nature11252.

[28] 1000 Genomes Project Consortium, A. Auton, L.D. Brooks, R.M. Durbin, E.P. Garrison, H.M. Kang, J.O. Korbel, J.L. Marchini, S. McCarthy, G.A. McVean, G.R. Abecasis, A global reference for human genetic variation, Nature. 526 (2015) 68–74. https://doi.org/10.1038/nature15393.

[29] N. Mulder, A. 'le Abimiku, S.N. Adebamowo, J. de Vries, A. Matimba, P. Olowoyo, M. Ramsay, M. Skelton, D.J. Stein, H3Africa: current perspectives, Pharmgenomics. Pers. Med. 11 (2018) 59–66. https://doi.org/10.2147/PGPM.S141546.

[30] J. Lonsdale, The Genotype-Tissue Expression (GTEx) project, Nat. Genet. 45 (2013) 580–585. https://doi.org/10.1038/ng.2653.

[31] PostgreSQL, PostgreSQL v. 10. (2022). https://www.postgresql.org/ (accessed December 10, 2022).

[32] S. Ramírez, FastAPI, FastAPI v. 0.88. (2022). https://fastapi.tiangolo.com (accessed 2022).

[33] A. Brown, G. Wilson, The Architecture of Open Source Applications: Elegance, Evolution, and a Few Fearless Hacks, Lulu.com, 2011.

[34] S. Ramírez, SQLModel, SQLModel v 0.0.8. (2022). https://sqlmodel.tiangolo.com/ (accessed December 10, 2022).

[35] The Best API Documentation Tool, Redocly. (2022). https://redocly.com/ (accessed December 12, 2022).

[36] Plotly Technologies Inc., Plotly: Collaborative data science, (2015). https://plot.ly (accessed 2022).

[37] M. Bakhtiari, S. Shleizer-Burko, M. Gymrek, V. Bansal, V. Bafna, Targeted genotyping of variable number tandem repeats with adVNTR, Genome Res. 28

(2018) 1709–1719. https://doi.org/10.1101/gr.235119.118.

[38] H.Z. Jam, Y. Li, R. DeVito, N. Mousavi, N. Ma, I. Lujumba, Y. Adam, M. Maksimov, B. Huang, E. Dolzhenko, Y. Qiu, F.E. Kakembo, H. Joseph, B. Onyido, J. Adeyemi, M. Bakhtiari, J. Park, S. Javadzadeh, D. Jjingo, E. Adebiyi, V. Bafna, M. Gymrek, A deep population reference panel of tandem repeat variation, bioRxiv. (2023) 2023.03.09.531600. https://doi.org/10.1101/2023.03.09.531600.

[39] J. Harrow, F. Denoeud, A. Frankish, A. Reymond, C.-K. Chen, J. Chrast, J. Lagarde, J.G.R. Gilbert, R. Storey, D. Swarbreck, C. Rossier, C. Ucla, T. Hubbard, S.E. Antonarakis, R. Guigo, GENCODE: producing a reference annotation for ENCODE, Genome Biol. 7 Suppl 1 (2006) S4.1–9. https://doi.org/10.1186/gb-2006-7-s1-s4.

[40] C. Mayer, F. Leese, R. Tollrian, Genome-wide analysis of tandem repeats in Daphnia pulex--a comparative approach, BMC Genomics. 11 (2010) 277. https://doi.org/10.1186/1471-2164-11-277.

[41] A.M. Newman, J.B. Cooper, XSTREAM: A practical algorithm for identification and architecture modeling of tandem repeats in protein sequences, BMC Bioinformatics. 8 (2007) 382. https://doi.org/10.1186/1471-2105-8-382.

[42] G. Benson, Tandem repeats finder: a program to analyze DNA sequences, Nucleic Acids Res. 27 (1999) 573–580. https://doi.org/10.1093/nar/27.2.573.

[43] H.-G. Nguyen, O. Lundström, A. Blank, H. Dawson, A. Lugli, M. Anisimova, I. Zlobec, Image-based assessment of extracellular mucin-to-tumor area predicts consensus molecular subtypes (CMS) in colorectal cancer, Mod. Pathol. 35 (2022) 240–248. https://doi.org/10.1038/s41379-021-00894-8.

[44] N. Mousavi, J. Margoliash, N. Pusarla, S. Saini, R. Yanicky, M. Gymrek, TRTools: a toolkit for genome-wide analysis of tandem repeats, Bioinformatics. 37 (2021) 731–733. https://doi.org/10.1093/bioinformatics/btaa736.

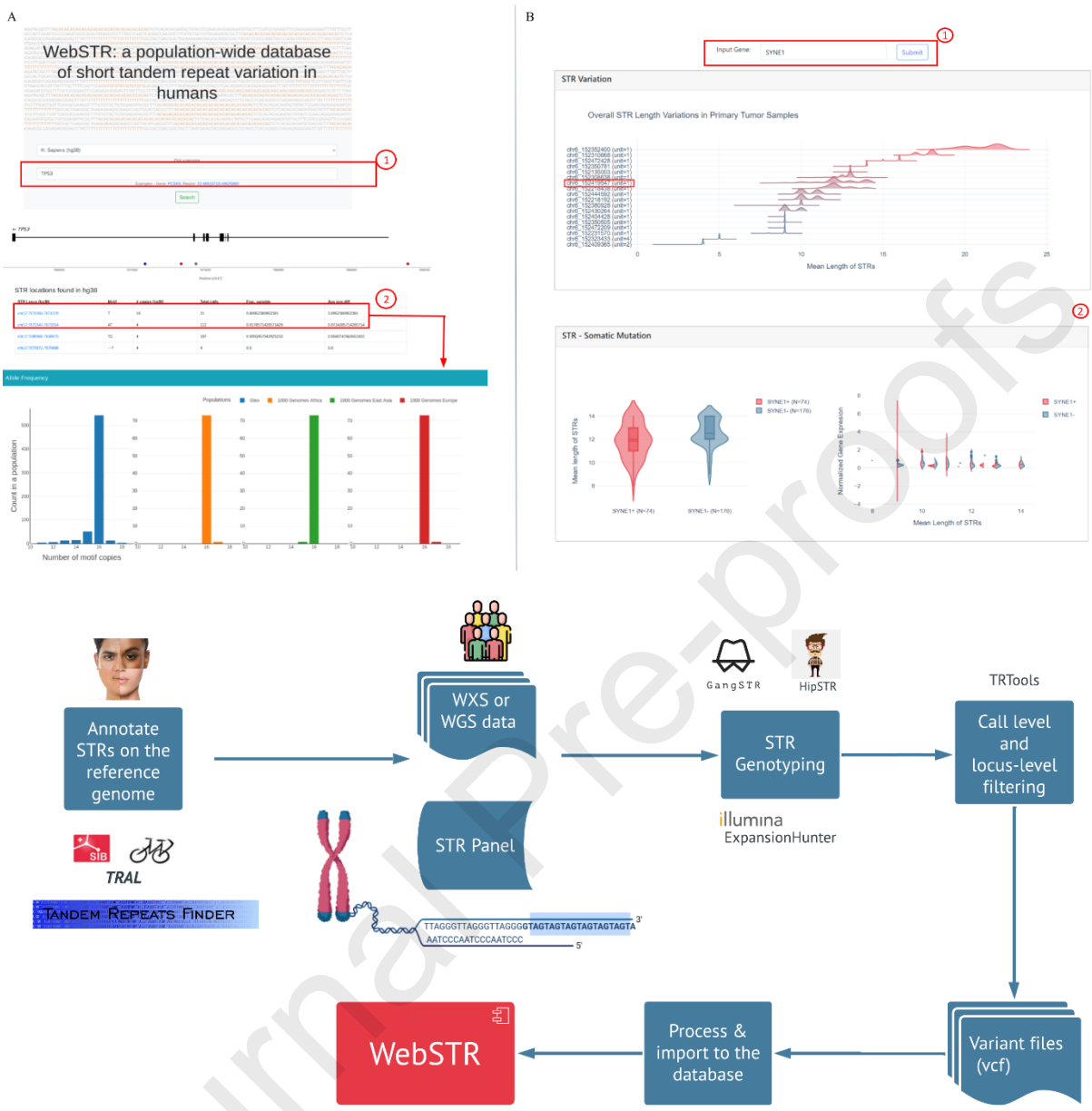
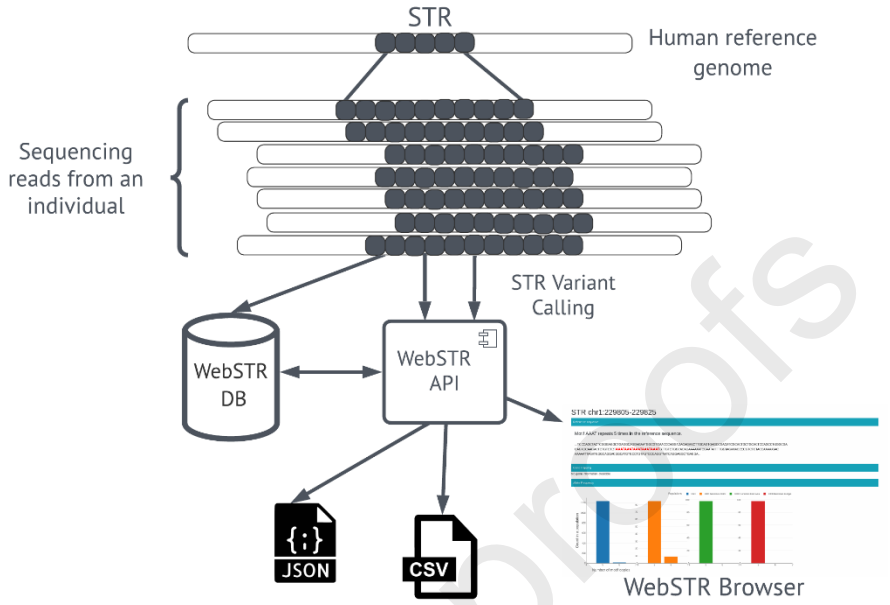


Table 1: Summary of genome-wide reference panels for Human STRs and other datasets available through WebSTR. WebSTR database contains genotyped data from many different sources that are organized as different study cohorts. Reference panels for Human Reference genome hg38 have been analyzed on colorectal cancer (CRC) patients data from TCGA [44], the 1000 Genomes Project [1] and H3Africa [25]. Reference panel for hg19 [10,32] has been used on data from the Genotype-Tissue Expression Project (GTEx) [17], The Simons Simplex Collection (SSC) [10,32], the 1000 Genomes project and the Simons Genome Diversity Project (SGDP) [18]

Panel alias	Human Genome Assembly Version	Annotation method	Genotyping methods	STRs	Cohorts	Samples	Available data
gangstr_crc_hg38	GRCh38.p2	TRAL	GangSTR	1,548,993 Reference STRs	Sinergia-CRC	412	Average variation of the number of repeat units in the dataset
				219,394 were found to have at least 1 variation call in the dataset			
				81,262 STRs have reliable variation data			
ensembletr_hg38	GRCh38.p11	TRF	EnsembleTR	1,710,833 total reference panel with allele frequency data available	1000Genomes H3Africa	3,550	Allele frequencies
			ExpansionHunter	35,998			

			GangSTR	1,133,225			
			HipSTR	1,331,280			
hipstr_hg19	GRCh37	TRF	HipSTR	~1.6 million	GTEx SSC 1000Genomes SGDP	652 1,916 150 300	Allele frequencies, trait associations Imputation metrics Mutation parameters

Graphical abstract



WebSTR Highlights:

- STRs have often been overlooked in human variation studies due to challenges in annotating and genotyping them, as well as by limited and low quality sequencing data. Although bioinformatics tools have enabled profiling STR variation genome-wide at population scale, existing large databases of human genetic variation do not explicitly handle STRs.

- WebSTR is the first comprehensive resource for genome-wide STR variation in humans, and currently contains data for approximately 1.7 million unique STRs. It enables users to easily view summary statistics including population-specific allele frequencies, mutation rates, and phenotype associations for specific STRs of interest. Users may additionally browse for STRs by gene or genomic region.

- WebSTR can be extended with results from other genotyping studies and is coupled with programmatic access to download the data.

- In making STR variation data easily accessible to the wider scientific community, we hope to increase awareness of the important regulatory roles these genetic elements play and to facilitate use of existing large STR genotype datasets by the broader genomics community

Authors' contributions

Oxana Lundström: Conceptualization, Methodology, Software, Writing- Original draft preparation, Project administration. **Max Verbiest**: Data curation, Software, Writing - Review & Editing. **Feifei Xia**: Visualization. **Helyaneh Ziaei Jam**: Data curation, Software. **Inti Zlobec**: Validation, Supervision. **Maria Anisimova:** Writing- Reviewing and Editing, Supervision, Funding acquisition. **Melissa Gymrek**: Conceptualization, Supervision, Funding acquisition, Writing- Reviewing and Editing.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: