UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Identifying and characterizing de novo tandem repeat mutations and their contribution to autism spectrum disorders

Permalink

https://escholarship.org/uc/item/8w72368c

Author

Mitra, Ileena

Publication Date

2021

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA SAN DIEGO

Identifying and characterizing *de novo* tandem repeat mutations and their contribution to autism spectrum disorders

A dissertation submitted in partial satisfaction of the requirements for the degree
Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Ileena Mitra

Committee in charge:

Professor Melissa Gymrek, Chair
Professor Vineet Bafna
Professor Joseph Gleeson
Professor Abraham Palmer
Professor Jonathan Sebat

2021

Copyright

Ileena Mitra, 2021

All rights reserved.

The dissertation of Ileena Mitra is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego

2021

DEDICATION

To my parents,

for your unconditional love and support.

your dreams are mine, and

my success is yours.

EPIGRAPH

When I had all the answers,

the questions changed.

Paulo Coelho

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I thank my advisor Professor Melissa Gymrek for her invaluable support, guidance, and encouragement throughout my graduate career. Her brilliant and creative enthusiasm for science captivated me instantly to join her new lab, and has inspired me to throughout my tenure to perform my absolute best. Her kindness, empathy, and affability has led me to successfully overcome obstacles that are inevitable in a PhD. Her positive and practical teaching and mentoring style helped me learn new skills and concepts, and become the scientist I am today.

I express my gratitude to my dissertation committee, Professor Jonathan Sebat, Professor Joe Gleeson, Professor Abe Palmer, and Professor Vineet Bafna, for their input, advice and recommendations on the direction of my research. I also sincerely thank Professor Alon Goren and Professor Lauren Weiss for their support and guidance on many occasions throughout my academic career.

Most importantly, I thank my family for believing in me and supporting me throughout my life. I also thank Team Gymrek, especially Nima, Shubham, and An for their encouragement and friendship as we grew as scientists together. In addition, I thank my friends Isaac and Kevin in the BISB program for being the best bioinformatics buddies and making my years in San Diego fun. Lastly, I acknowledge my fluffy co-author, Rosie, for co-working with me through lots of coding and writing sessions and Taylor Swift parties.

Chapters 2-6 is an adapted reprint of the material "Patterns of *de novo* tandem repeat mutations and their role in autism" by Ileena Mitra, Bonnie Huang, Nima Mousavi, Nichole Ma, Michael Lamkin, Richard Yanicky, Sharona Shleizer-Burko, Kirk E. Lohmueller, and Melissa Gymrek published in *Nature* 2021. The dissertation author was the primary investigator and author of this paper.

VITA

2009 - 2013 University of California Santa Cruz

Bachelor of Science, Biomolecular Engineering, cum laude with honors

2016 - 2021 University of California San Diego

Doctor of Philosophy, Bioinformatics and Systems Biology

PUBLICATIONS

Mitra, I., Huang, B., Mousavi, N., Ma, N., Lamkin, M., Yanicky, R., Shleizer-Burko, S., Lohmueller, K.E. & Gymrek, M. Patterns of *de novo* tandem repeat mutations and their role in autism. *Nature*. 2021.

Breuss, M.W., Antaki, D., George, R.D., Kleiber, M., James, K.N., Ball, L.L., Hong, O., **Mitra, I**., Yang, X., Wirth, S.A., Gu, J., Garcia, C.A.B., Gujral, M., Brandler, W.M., Musaev, D., Nguyen, A., McEvoy-Venneri, J., Knox, R., Sticca, E., Botello, M.C.C., Uribe, J.F., Pérez, M.C., Arranz, M., Moffitt, A.B., Wang, Z., Hervás, A., Devinsky, O., Gymrek, M., Sebat, J. & Gleeson, J.G.. Autism risk in offspring can be assessed through quantification of male sperm mosaicism. *Nature Medicine*. 2020. 26 (1):143-150.

Saini, S., **Mitra, I**., Mousavi, N., Fotsing, S.F. & Gymrek, M. A reference haplotype panel for genome-wide imputation of short tandem repeats. *Nature communications.* 2018. 9 (1):4397.

Mitra, I., Lavillaureix, A., Yeh, E., Traglia, M., Tsang, K., Bearden, C.E., Rauen, K.A. & Weiss, L.A. Reverse Pathway Genetic Approach Identifies Epistasis in Autism Spectrum Disorders. *PLoS Genetics*. 2017. 13 (1):e1006516.

Mitra, I, Tsang K, Ladd-Acosta C, Croen LA, Aldinger KA, Hendren RL, Traglia M, Lavillaureix A, Zaitlen N, Oldham MC, Levitt P, Nelson S, Amaral DG, Hertz-Picciotto I, Fallin MD, Weiss LA. Pleiotropic Mechanisms Indicated for Sex Differences in Autism. *PLoS Genetics*. 2016. 12 (11):e1006425.

Troll, C.J., Adhikary, S., Cueff, M., **Mitra, I**., Eichman, B.F. & Camps, M. Interplay between base excision repair activity and toxicity of 3-methyladenine DNA glycosylases in an E. coli complementation system. *Mutation Research*. 2014. 763-764:64-73.

ABSTRACT OF THE DISSERTATION

Identifying and characterizing *de novo* tandem repeat mutations and their contribution to autism spectrum disorders

by

Ileena Mitra

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2021

Professor Melissa Gymrek, Chair

Genetic factors are known to make a large contribution to the risk of Autism Spectrum Disorders (ASD) (1). The heritability of ASD is estimated to be over 50%, and it is estimated that *de novo* rare variants contribute in about 30% of simplex autism-affected cases (2,3). To date, population sequencing studies have been limited to analyzing single nucleotide variants (SNVs), small insertions and deletions (indels), or copy number variants (CNVs) (4).

This dissertation expands genetic research to further identify potential genomic regions and pathogenic mutations associated with ASD. Tandem repeats (TRs) are a class of repetitive structural variants composed of 1-20 base pair repeating units (5). TRs exhibit mutation rates that are orders of magnitude higher than SNPs, indels, or CNVs (6), and thus represent one of the largest sources of human genomic variability (4,5). TRs are often associated with diseases characterized by neurological and developmental symptoms (7–9). for example, Fragile X Syndrome, the most prevalent genetic cause of ASD (10). To date, direct studies of *de novo* TR mutations have been limited in population genetic studies.

In this dissertation, I present a framework for population-scale characterization of genome-wide *de novo* TR mutations and their contribution to the genetic etiology of ASD. In my first chapter, I present my bioinformatics pipeline using MonSTR to analyze whole genome sequencing data to identify high-confidence, germline *de novo* TRs within parent-offspring trios. MonSTR, a novel statistical method, takes genotype likelihood values reported by a TR variant caller as input and estimates the posterior probability of a mutation resulting in a repeat copy number change at each TR loci in each child.

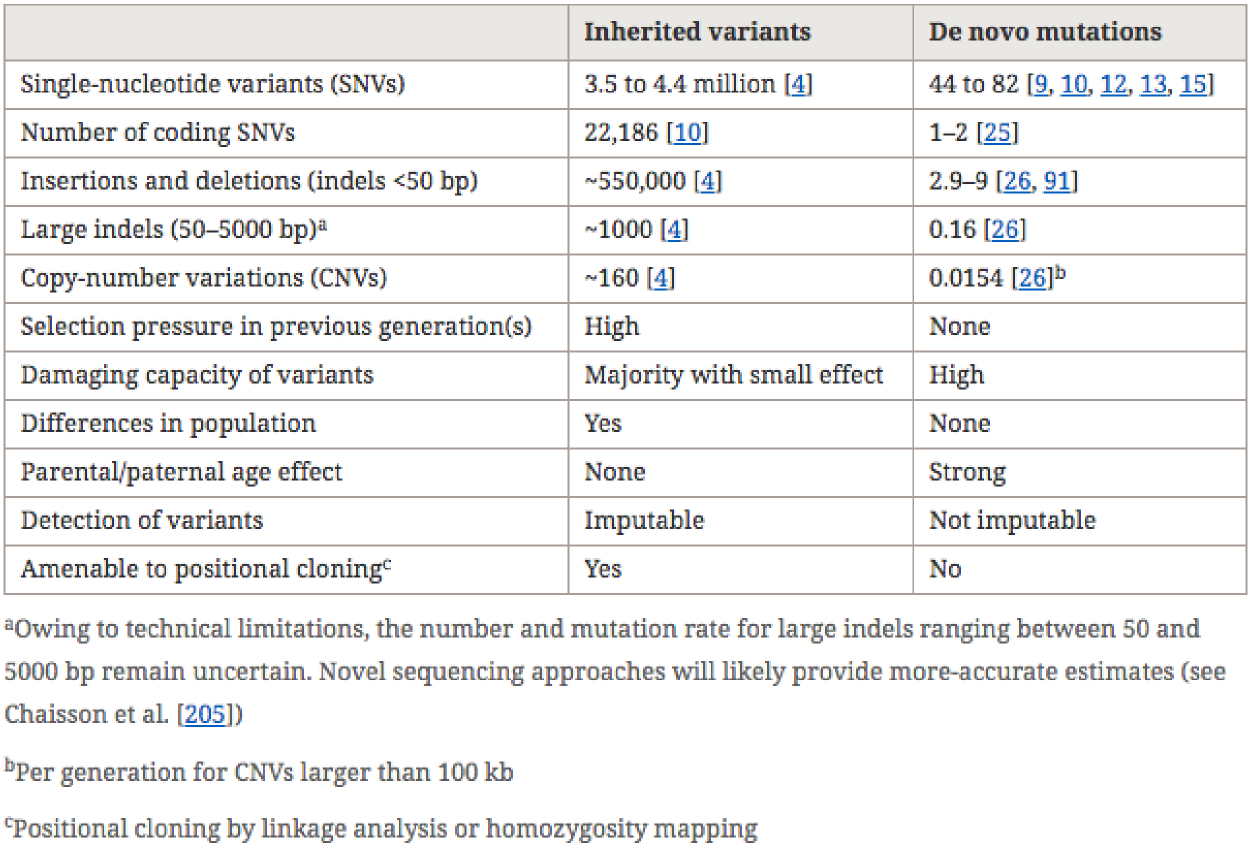
In the following chapters, I present the results from identifying *de novo* TR mutations in autism-affected and unaffected children. I characterize patterns of TR mutational mechanism in the general population, in which I found an average of 54 *de novo* TRs per individual. I show that ASD affected individuals have a higher number of *de novo* TR mutations, specifically in regulatory regions and brain-related genes, as well as larger sized mutations, compared to matched unaffected siblings. Lastly, I applied a novel natural selection-based method (SISTR) to identify deleterious *de novo* TR mutations, and show that autism probands are enriched for rare and pathogenic TR mutations. Overall, this dissertation presents and applies a novel framework for identifying and prioritizing *de novo* TR mutations in order to better understand TR mutational mechanisms and the genetic etiology of ASD.

CHAPTER 1: INTRODUCTION

1.1 Tandem Repeats

Repetitive variants constitute an estimated 30% percent of the human genome, and are found in both coding and non-coding regions (11,12). Repetitive variants fall into various classes based on their structure and mode of duplication. I focus on tandem repeats (TRs), which I define as composed of short tandem repeats (STRs) and variable number tandem repeats (VNTRs). STRs, also called microsatellites, are distinguished as one to six base-pair (bp) motifs and VNTRs consist of motif lengths of over six bp. TRs consist of motifs repeated consecutively in sequence for a variable number of times. The genotype for TR variants is represented by the number of repeats of a motif at a given locus. Most TRs are highly multi-allelic due to their variable nature. TRs gain and lose repeat units at high rates due to polymerase slippage during DNA replication (13). Due to this error prone replication process, STRs have been reported to have a genome-wide average mutation rate of 10^{-8} to 10^{-2} mutations/locus/generation (6) and VNTR mutation rates are estimated to be around 10^{-5} (14), which is higher than most other types of *de novo* variations (**Table 1**) (4).

Table 1: Comparison of inherited and *de novo* variants by Rocio Acuna-Hidalgo, et al. (2016). (4)



	Inherited variants	De novo mutations
Single-nucleotide variants (SNVs)	3.5 to 4.4 million [4]	44 to 82 [9, 10, 12, 13, 15]
Number of coding SNVs	22,186 [10]	1–2 [25]
Insertions and deletions (indels <50 bp)	~550,000 [4]	2.9–9 [26, 91]
Large indels (50–5000 bp)[a]	~1000 [4]	0.16 [26]
Copy-number variations (CNVs)	~160 [4]	0.0154 [26][b]
Selection pressure in previous generation(s)	High	None
Damaging capacity of variants	Majority with small effect	High
Differences in population	Yes	None
Parental/paternal age effect	None	Strong
Detection of variants	Imputable	Not imputable
Amenable to positional cloning[c]	Yes	No

[a] Owing to technical limitations, the number and mutation rate for large indels ranging between 50 and 5000 bp remain uncertain. Novel sequencing approaches will likely provide more-accurate estimates (see Chaisson et al. [205])

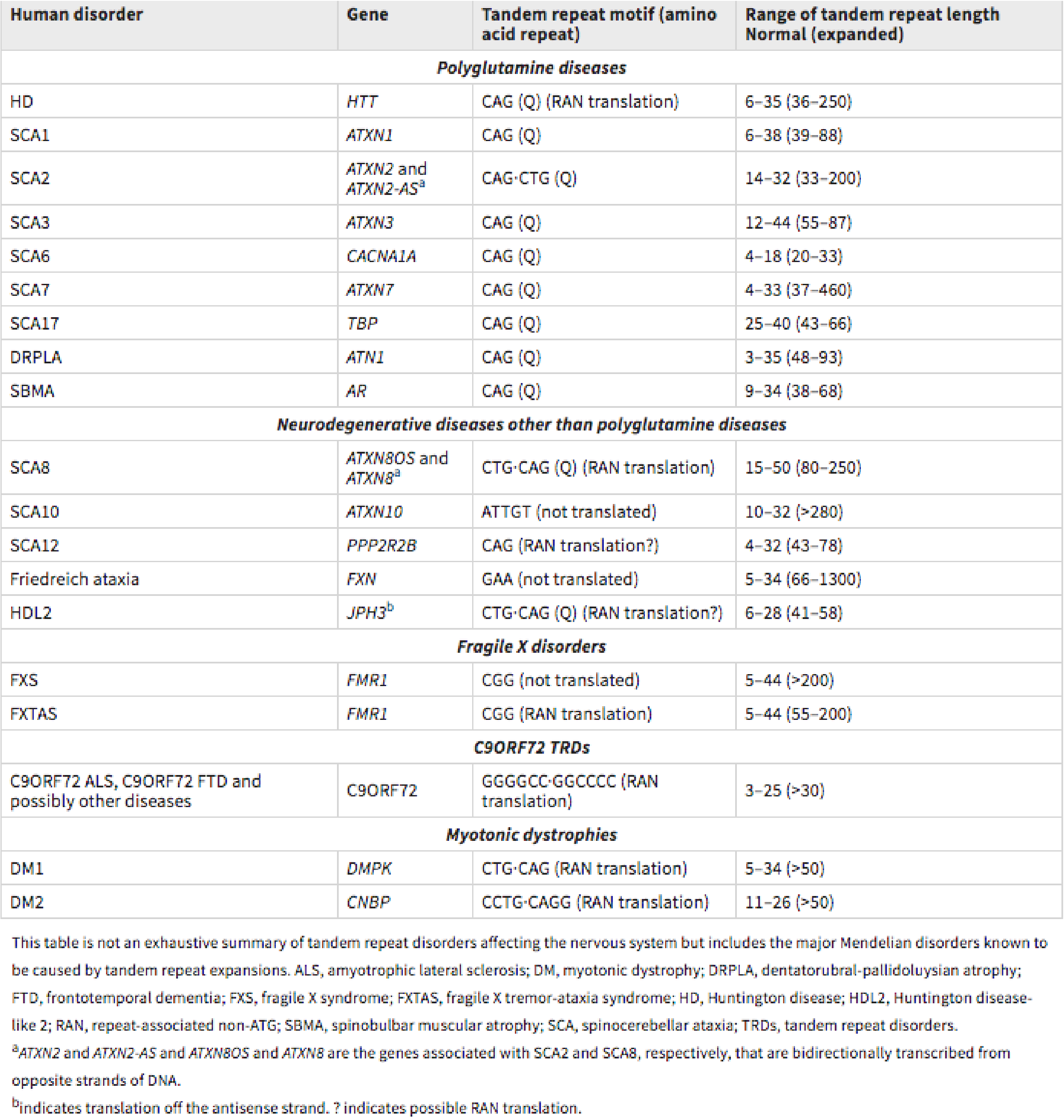
[b] Per generation for CNVs larger than 100 kb

[c] Positional cloning by linkage analysis or homozygosity mapping

Due to their repetitive structure, TRs present unique challenges in sequencing, genotyping and mutation calling on a genome-wide population-level (5). First, PCR preparation for sequencing will cause replication errors during the amplification stage, therefore PCR-free protocols must be used. Second, there are many algorithmic difficulties during read-pair alignment (5) due to normal variation in TR lengths, where insertions or deletions deviate from the reference genome. Third, the ability to genotype (infer the number of repeats present) is complicated due to the fact that TRs are much longer than the standard Illumina read lengths of 100-150 bp. Due to these technical challenges in the interpretation of TRs, most previous sequencing studies have deliberately removed repetitive regions of the genome in analysis.

A number of genetic studies have demonstrated TRs to have widespread effects on a range of complex phenotypes in humans (5,7). Many Mendelian diseases with neurological or psychological symptoms are due to TR expansions (**Table 2**) (7). In addition, there is growing evidence that these repetitive variants are likely to contribute to biological differences (15) and polygenic diseases (7). A particularly interesting case is Fragile X Syndrome which accounts for about 2-6% of all ASD cases (16). It is due to an expansion of more than 200 repeats of a “CGG” STR upstream of *FMR1* (17,18). TRs are a class of genetic variation likely contributing to risk for ASD and other neuropsychiatric disorders and remains to be uncovered.

Table 2: Tandem repeat disorders affecting the nervous system by Anthony J. Hannan (2018). (7)



Human disorder	Gene	Tandem repeat motif (amino acid repeat)	Range of tandem repeat length Normal (expanded)
Polyglutamine diseases			
HD	*HTT*	CAG (Q) (RAN translation)	6–35 (36–250)
SCA1	*ATXN1*	CAG (Q)	6–38 (39–88)
SCA2	*ATXN2* and *ATXN2-AS*[a]	CAG·CTG (Q)	14–32 (33–200)
SCA3	*ATXN3*	CAG (Q)	12–44 (55–87)
SCA6	*CACNA1A*	CAG (Q)	4–18 (20–33)
SCA7	*ATXN7*	CAG (Q)	4–33 (37–460)
SCA17	*TBP*	CAG (Q)	25–40 (43–66)
DRPLA	*ATN1*	CAG (Q)	3–35 (48–93)
SBMA	*AR*	CAG (Q)	9–34 (38–68)
Neurodegenerative diseases other than polyglutamine diseases			
SCA8	*ATXN8OS* and *ATXN8*[a]	CTG·CAG (Q) (RAN translation)	15–50 (80–250)
SCA10	*ATXN10*	ATTGT (not translated)	10–32 (>280)
SCA12	*PPP2R2B*	CAG (RAN translation?)	4–32 (43–78)
Friedreich ataxia	*FXN*	GAA (not translated)	5–34 (66–1300)
HDL2	*JPH3*[b]	CTG·CAG (Q) (RAN translation?)	6–28 (41–58)
Fragile X disorders			
FXS	*FMR1*	CGG (not translated)	5–44 (>200)
FXTAS	*FMR1*	CGG (RAN translation)	5–44 (55–200)
C9ORF72 TRDs			
C9ORF72 ALS, C9ORF72 FTD and possibly other diseases	C9ORF72	GGGGCC·GGCCCC (RAN translation)	3–25 (>30)
Myotonic dystrophies			
DM1	*DMPK*	CTG·CAG (RAN translation)	5–34 (>50)
DM2	*CNBP*	CCTG·CAGG (RAN translation)	11–26 (>50)

This table is not an exhaustive summary of tandem repeat disorders affecting the nervous system but includes the major Mendelian disorders known to be caused by tandem repeat expansions. ALS, amyotrophic lateral sclerosis; DM, myotonic dystrophy; DRPLA, dentatorubral-pallidoluysian atrophy; FTD, frontotemporal dementia; FXS, fragile X syndrome; FXTAS, fragile X tremor-ataxia syndrome; HD, Huntington disease; HDL2, Huntington disease-like 2; RAN, repeat-associated non-ATG; SBMA, spinobulbar muscular atrophy; SCA, spinocerebellar ataxia; TRDs, tandem repeat disorders.

[a] *ATXN2* and *ATXN2-AS* and *ATXN8OS* and *ATXN8* are the genes associated with SCA2 and SCA8, respectively, that are bidirectionally transcribed from opposite strands of DNA.

[b] indicates translation off the antisense strand. ? indicates possible RAN translation.

1.2 Genetic architecture of Autism Spectrum Disorders

Autism Spectrum Disorders (ASD) is an early onset developmental disorder characterized by symptoms of deficit in communication and social interaction and restrictive and repetitive behaviors (1). Family studies demonstrate a significant genetic basis for susceptibility of ASD (19) and the genetic SNP-based heritability has been estimated between 17% to 52% (2).

Numerous family whole-exome and whole-genome sequencing (WGS) studies have found a strong contribution of *de novo* mutations to ASD (20–23). Germline *de novo* mutations are alleles that do not follow Mendelian inheritance, meaning they are observed in the offspring's germline DNA but are not present in the germline DNA of the parents. The novel nature of germline *de novo* variation excludes them from evolutionary selection and purification, and thus, these mutations are more likely to have negative fitness consequences compared to inherited genetic variation (24).

It is estimated that *de novo* point mutations and *de novo* copy number variants (CNVs) contribute to an estimated 30% of all simplex ASD (3). In addition, our genome-wide study of mosaic *de novo* single nucleotide variants (SNVs), structural variants (SVs), and STRs in paternal sperm show a role of these variants in ASD risk recurrence (25). Studies have found over 102 genes have been found to be associated with ASD (26), yet these associated genes only explain a small percentage of ASD cases (17,27). Previous studies assessing *de novo* mutations contributing to ASD have been limited to SNVs, small insertion-deletions (indels), and CNVs and have excluded repetitive variants.

Information from repetitive variants remains missing in most population sequencing studies due to sequencing errors and algorithmic difficulties in genotype interpretation. Given the rapid mutation rates of repetitive variants, they contribute a large number of *de novo* mutations per generation (5). Thus, TRs present rich and unexplored source of *de novo* mutations that may contribute to ASD and other neuropsychiatric disorders. Assessment of repetitive variants will add more information to understanding the genetic etiology of ASD and may potentially explain a fraction ASD cases.

1.3 Outline

Chapter 2 describes the statistical method and bioinformatics pipeline created to identify *de novo* TR mutations in WGS data of parent-offspring trios.

Chapter 3 and 4 examine the patterns of *de novo* TR mutations in healthy and ASD affected individuals.

Chapter 5 describes the interpreting pathogenic *de novo* TR mutations for ASD risk.

CHAPTER 2: IDENTIFYING TANDEM REPEAT MUTATIONS

2.1 Introduction

Human genetic studies have focused on understanding how genetic variation and mutations affect phenotypes. Population sequencing studies primarily analyze SNV mutations using standardized pipelines. At present, there are no standardized procedures for discovering repeat mutations in WGS data, such as exists for SNV mutations (28). Therefore, my goal was to develop a streamlined and efficient pipeline to process raw WGS or whole exome sequencing (WES) data into a format readily usable for analyses of TR mutations. This would mitigate numerous technical challenges and allow the human genetics community to have a standardized and reproducible methodology for TR population analyses.

In this chapter, I present the novel statistical method MonSTR, a novel bioinformatics algorithm for detecting TR mutations within parent-offspring trios (**Section 2.2**). MonSTR can be applied genome-wide to large population whole-genome sequencing (WGS) datasets. MonSTR features a likelihood-based method for detecting mutation events based on individual genotype likelihood values, transmission rate, and prior locus mutation probability. MonSTR further attempts to determine the maternal versus paternal phase for unambiguous mutation events.

In addition, I describe building a comprehensive, high-throughput pipeline that includes all necessary steps starting from TR variant calling in WGS to ending with a list of high confidence *de novo* TR mutations identified within a parents-offspring family (trio) **(Fig. 1, Section 2.3).** The experimental and statistical validation results presented below (**Section 2.4 and 2.5**) show that this pipeline can robustly identify genome-wide *de novo* TR mutations.

2.2 MonSTR model and method details

2.2.1 MonSTR statistical model

MonSTR's likelihood model is based on a previously published model for identifying *de novo* point mutations (28). It considers a single trio (father, mother, and child) at a time. Let $G_m = (g_{m1}, g_{m2})$ be the diploid genotype (allele lengths) of the mother, $G_f = (g_{f1}, g_{f2})$ the diploid genotype of the father, and $G_c = (g_{c1}, g_{c2})$ the diploid genotype of the child in a trio. Here we will assume that the child's genotype is ordered, with allele g_{c1} derived from the mother and g_{c2} from the father. The likelihood of a particular genotype configuration with no mutation can be defined as:

$$L(G_m, G_f, G_c \mid D) \propto M(G_m|D)F(G_f|D)C(G_c|D) \times T(G_c| \, G_m, G_f) \times P(G_m, G_f)$$

Where D denotes available WGS data for the family at that locus, $M(G_m|D)$ denotes the likelihood of maternal genotype G_m given the data, $F(G_f|D)$ denotes the likelihood of paternal genotype G_f given the data, $C(G_c|D)$ denotes the likelihood of child genotype G_c given the data, $T(G_c| \, G_m, G_f)$ is a transition probability, and $P(G_m, G_f)$ is the prior probability of observing the parent genotypes. Genotype likelihoods M, F, and C are obtained directly from the GGL (genotype likelihood) field output by GangSTR (29). MonSTR is also compatible with the GL field (genotype likelihood) output by HipSTR (30). We assume a uniform prior on all genotypes, so the prior term is dropped. For simplicity below we drop the D term.

The likelihood the observed data with no mutation is computed as:

$$L(nomut) = \sum_{G_m} \sum_{G_f} M(G_m)F(G_f) \sum_{G_c \in I} T(G_c| \, G_m, G_f) \times C(G_c) \qquad \text{(Eq. 1)}$$

Where I represents genotypes that could result from all possible transmission scenarios, i.e., $G_c = (g_{m1}, g_{f1}),\ (g_{m1}, g_{f2}),\ (g_{m2}, g_{f1})$, or (g_{m2}, g_{f2}). The transmission probability $T(G_c|\ G_m, G_f)$ is simply equal to 0.25 for any particular transmission.

The likelihood of a mutation occurring is computed as:

$$L(mut\ ) = \sum_{G_m} \sum_{G_f} M(G_m)F\left(G_f\right) \sum_{G_c \in I} T(G_c|\ G_m, G_f)\, [\sum_k R(g_{c1}, k)C(g_{c1} + k, g_{c2}) + R(g_{c2}, k)C(g_{c1}, g_{c2} + k)] \text{ (Eq. 2)}$$

Where $R(g, k)$ gives the transition probability of allele g mutating by k units. The above equation considers all four possible transition scenarios (similar to the likelihood for no mutation), followed by a mutation either to the maternally inherited allele (g_{c1}) or to the paternally inherited allele (g_{c2}).

Finally, MonSTR computes the posterior probability of a mutation as:

$$P(mut|D)\ = \frac{L(mut|D)P(mut)}{L(mut|D)P(mut) + L(nomut|D)(1 - P(mut))}$$

Where $P(mut)$ gives the prior probability of mutation.

MonSTR allows for specifying locus-specific mutation parameters based on the mutation model presented below. For this study we used a naive mutation transition probability treating each mutation size as equally likely and a constant prior probability of mutation across all TRs. In future use cases, mutation rate priors and transition probabilities could be more accurately set based on observed *de novo* mutations from this or other studies.

2.2.2 MonSTR statistical model for chromosome X

The model presented above (2.2.1) applies to autosomal TRs. MonSTR implements a modified version of the model for identifying mutations on chromosome X.

For chromosome X loci, males have haploid genotype calls. Thus, $G_f = (g_f)$ rather than $G_f = (g_{f1}, g_{f2})$. For female children, Equations 1 and 2 above have only slight modifications:

- I, which represents the set of child genotypes that could result from all possible transmission scenarios, includes $G_c = (g_{m1}, g_f)$ or (g_{m2}, g_f).
- The transition probability $T(G_c |\, G_m, G_f)$ is set to 0.5 for each case.

For male children, $G_c = (g_c)$, and the following further modifications are made:

- I includes child genotype possibilities $G_c = (g_{m1})$ or (g_{m2}).
- The transition probability only depends on the maternal genotype and $T(G_c |\, G_m)$ is set to 0.5 for each possible child genotype.
- Equation 2 (likelihood of mutation) is modified to only consider mutations from the mother:

$$L(mut\ ) = \sum_{G_m} M(G_m) \sum_{G_c \in I} T(G_c |\, G_m)\,[\sum_k R(g_{c1}, k) C(g_{c1} + k)\ ]$$

2.2.3 Naïve mutation model

In addition to the model-based mutation detection method described above (2.2.1, 2.2.2), MonSTR also implements two naïve methods:

If the --naïve option is specified, MonSTR will simply check if the child genotype call can be explained by Mendelian inheritance from genotypes called in the parents. Thus, it does not consider uncertainty from genotype likelihoods and outputs a binary result for mutation/no mutation rather than a posterior probability.

If the --naïve-expansions-frr <int1,int2> option is specified, MonSTR will output candidate expansion mutations that have either <int1> fully repetitive reads (FRRs. described previously (29,31). Fully repetitive reads at a locus indicate a potential repeat expansion) in a child and 0 in parents, or <int2> flanking reads supporting an allele length greater than the largest allele observed in parents.

2.2.4 Inferring parent of origin

For each mutation identified, we attempted to infer the likely parent of origin based on the child and parent genotypes. Let $C = (c_1, c_2)$ be the child genotype, where c_i is the number of repeats called in the *i*th allele. Similarly, let $F = (f_1, f_2)$ and $M = (m_1, m_2)$ denote the father and mother genotypes, respectively. If $c_1 \in M$ and $c_1 \notin F$, we determine c_2 to be the new allele and phase the mutation to the father. If $c_2 \in M$ and $c_2 \notin F$, we determine $c_1$to be the new allele and phase the mutation to the father. Similarly, if $c_1 \in F$ and $c_1 \notin M$ or $c_2 \in F$ and $c_2 \notin M$, the new allele is determined to be c_2 or c_1, respectively, and the mutation is phased to the mother. All mutations not meeting the above conditions were labeled with phase as unknown.

2.2.5 MonSTR implementation details

MonSTR is implemented as a standalone command-line program written in C++. To improve speed and integration with other tools, it uses standard file formats and leverages the htslib library (www.htslib.org) and previously written VCF parsing libraries implemented in HipSTR (30).

MonSTR takes as input a multi-sample VCF with genotype likelihoods for each sample generated by HipSTR (30) or GangSTR (29) and a pedigree file (in plink.fam (32) format https://www.cog-genomics.org/plink2/formats#fam). It outputs a tab-delimited file describing mutation posterior probabilities, mutation sizes, and parent of origin inferences, for each trio at each TR.

MonSTR can call mutations using either "model-based" (default) or "naïve" mode (options --naïve or --naïve-expansions-frr described above). In the "model-based" option, users can choose either a global prior probability of mutation (default to 10^{-3} mutations per locus per generation) or input a file with per-locus mutation rates to use as priors. Further, the transition probabilities $R(g, k)$ can be set either to a uniform probability to mutate to any allele, or users may input values of parameters β (length dependent direction bias), p (mutation step size geometric distribution parameter), and the central allele, which are described in detail below, to implement more detailed step size distribution models.

MonSTR also offers many options to discard individual calls based on properties including their coverage, genotype quality score, or noise in estimated repeat copy number from individual reads. It will only process a family if all members of the trio have genotypes remaining after filtering.

The MonSTR implementation extends code originally included in the HipSTR software (https://github.com/tfwillems/HipSTR) written by Thomas Willems (30). Full documentation of these and other options can be found at https://github.com/gymreklab/STRDenovoTools.

2.2.6 MonSTR limitations and relationship to other tools

Our study presents MonSTR, a tool for identifying *de novo* mutations at TRs from family-based TR genotypes. MonSTR uses a statistical model incorporating genotype likelihoods and mutation parameters to estimate the posterior probability of mutation at each TR in each child. Compared to a naïve method of simply identifying conflicting genotypes between parents and children, MonSTR achieves lower false positive rates on simulated data, especially for loci with high genotype uncertainty as is the case for genotype calls with low coverage (**Fig. 3**).

To our knowledge, no similar tool exists for identifying *de novo* TR mutations from genotype data. A recent study by Trost, *et al.* (33), applied ExpansionHunter Denovo (34) to identify candidate TR expansions enriched in ASD. However, this method does not actually detect *de novo* mutations and is not directly comparable to MonSTR. The *de novo* in their method name instead refers to the fact that it does not rely on a pre-specified annotation of the locations of TRs in the reference genome, but rather takes a reference-free approach to identify potential repeat expansions. Further, by design ExpansionHunter Denovo (34) only considers large repeat expansions beyond the sequencing read length, whereas our pipeline (GangSTR/MonSTR) focuses on stepwise changes at shorter TRs. While there is some overlap, the set of TRs analyzed by these two methods is largely orthogonal.

While MonSTR is capable of utilizing detailed locus-specific mutation parameters as prior information, our study used a uniform prior mutation rate and naïve transition probabilities across all TRs for several reasons. (i) At the outset of this study, per-locus mutation parameters were only available for a subset of genome-wide loci. Most of these were inferred indirectly by MUTEA (35) and have high standard errors. (ii) A major goal of our study was to characterize genome-wide properties of TR mutations. Incorporating information about mutation rates or transition probabilities learned from smaller previous studies could bias our results toward learning trends that have already been reported. By using inputs to MonSTR that are relatively agnostic to known trends, we are more confident in our findings such as differences in relative mutation rates between classes of TRs and biases in step sizes. (iii) Finally, genotyping TRs from NGS is still noisy, and genotype likelihoods output by GangSTR (29) or other tools are imperfect. We found that applying default MonSTR parameters to GangSTR genotypes identified a high number of false positive mutations (~50% validation rate). Inspection of these TRs identified that GangSTR occasionally assigned high likelihood to erroneous genotypes at some loci prone to read misalignments at TR regions. In these cases, tuning parameters such as mutation rate priors or posterior thresholds is insufficient to remove false positive calls. Thus, we needed to apply additional heuristic filters (e.g., requiring the *de novo* allele to be supported by at least 3 reads) to achieve reasonable validation rates (~90%). These filters currently have more influence than tuning mutation rate priors and reduce sensitivity, but are necessary to achieve low false positive rates. In future work, with more robust TR genotype likelihoods from GangSTR or other tools and locus-specific prior mutation parameters inferred from this or other studies, MonSTR's model-based method is likely to be of further benefit.

2.3 TR genotyping and *de novo* mutation discovery pipeline

2.3.1 Simplex autism family WGS data

The Simons Simplex Collection (SSC) dataset, collected by the Simons Foundation Autism Research Initiative (SFARI) (36), used in this study consists of 1,637 quad families (**Table 3**). CRAM files containing WGS reads aligned to the hg38 reference genome and phenotype information for phases 1-3 were obtained from SFARI base (https://base.sfari.org/). SFARI recruited families where neither proband nor sibling were known to have: (1) a confirmed rare and exonic *de novo* CNV (≤1% population frequency). (2) an inherited CNV that is rare and encompasses ≥ 10 genes. and/or (3) a known, rare likely gene disrupting mutation within the exome (20). The selected families were also selected to minimize birth-order effects due to paternal age (20). The sporadic autism diagnosis in only a single child in the simplex families suggests that *de novo* mutations are likely to contribute to the ASD phenotype. For each individual, 1μg of DNA was extracted from blood samples. The data was generated by Illumina Hiseq paired-end WGS with PCR-free library preparation protocol. Genome sequencing coverage consisted of 34.8 ± 5.3x and a median library insert size of 417.8 ± 111.5 bp (20).

Table 3: Description of SSC datasets used for analysis of *de novo* TR mutations.

Table shows the number of families from the Simons Simplex Collection (36) included in our study.

	# Families Available	**# Families Analyzed**	**# Families After All QC**
Phase 1	500 quads	500	488
Phase 2	598 quads	597	578
Phase 3-1	449 quads	449	438
Phase 3-2	91 quads	91	89
Total	1638 families	1637	1593

2.3.2 Bioinformatics pipeline architecture

We performed a genome-wide analysis of *de novo* TR mutations (**Fig. 1**) using WGS available for 1,637 quad simplex families sequenced to 35× coverage as part of the Simons Simplex Collection (SSC) (36) (**Table 3**), which have been ascertained to enrich for probands likely to harbor previously uncharacterized pathogenic *de novo* mutations (20). We used GangSTR (29) to estimate diploid repeat lengths in each sample at 1,189,198 TRs with repeat unit lengths of 1–20 base pairs (bp) and median total TR lengths of 12 bp in the hg38 reference human genome assembly. TR genotype results were used as input to MonSTR to identify mutations in each child.

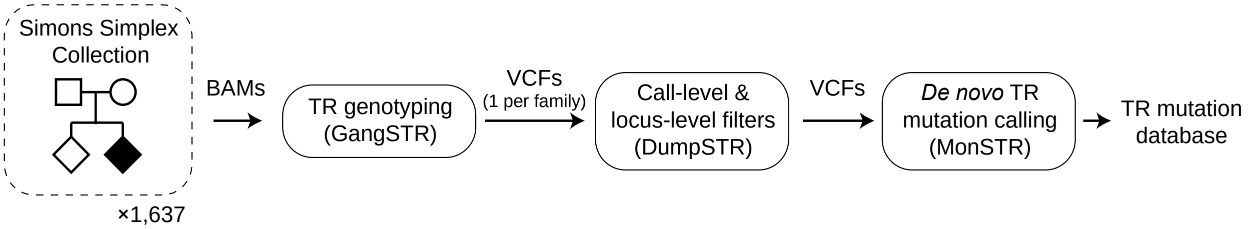

Simons Simplex Collection
×1,637
BAMs
TR genotyping (GangSTR)
VCFs (1 per family)
Call-level & locus-level filters (DumpSTR)
VCFs
De novo TR mutation calling (MonSTR)
TR mutation database

Figure 1: Study Design.

We analyzed *de novo* TR mutations from WGS data for quad families from the Simons Simplex Collection. BAM, binary sequence alignment/map format. VCF, variant call format.

2.3.3 Genome-wide TR genotyping

CRAM files were processed on Amazon Web Services (AWS) using the AWS Batch service. Genotyping of autosomal TRs was performed with GangSTR v2.4.2 (29) using the reference TR file hg38_ver16.bed.gz available on the GangSTR website (https://github.com/gymreklab/GangSTR) and with the option --include-ggl to enable outputting detailed genotype likelihood information. Chromosome X TRs were genotyped using GangSTR v2.4.4 with additional options --bam-samps and --samp-sex to interpret sample sex for chromosome X. A separate GangSTR job was run for each family on each chromosome resulting in separate VCF files for each.

Genotypes were then subject to call-level filtering using dumpSTR, which is included in the TRTools toolkit v1.0.0 (37). DumpSTR was applied separately to each VCF with parameters --min-call-DP 20 --max-call-DP 1000 --filter-spanbound-only --filter-badCI --require-support 2 --readlen 150. Male chromosome X genotypes were filtered separately using the same parameters except with --min-call-DP 10. These options remove genotypes with too low or too high coverage, with only spanning or flanking reads identified indicating poor alignment, and with maximum likelihood genotypes falling outside 95%

confidence intervals reported by GangSTR (29). After call-level filtering, each sample was examined for call-level missingness. All samples had >90% call rate and no outliers were identified.

Filtered VCFs from each phase were then merged using mergeSTR (TRTools v1.0.0) (37) with default parameters. The merged VCF was then used as input to dumpSTR to compute locus-level filters using the parameters --min-locus-hwep 10^{-5} --min-locus-callrate 0.8 --filter-regions GRCh38GenomicSuperDup.sorted.gz --filter-regions-names SEGDUP to remove genotypes overlapping segmental duplications. The file GRCh38GenomicSuperDup.sorted.gz was obtained using the UCSC Table Browser (38) (hg38.genomicSuperDups table). For chromosome X, the Hardy-Weinberg Equilibrium filter was applied only to females. Filters obtained from analyzing each phase were combined and any TRs failing locus-level filters in any phase were removed from further analysis.

2.3.4 Identifying de novo TR mutations

MonSTR v1.0.0 was called separately on each family after applying call-level and locus-level genotype filters described above. MonSTR was called with non-default parameters --max-num-alleles 100 --include-invariant --gangstr --require-all-children --output-all-loci --min-num-encl-child 3 --max-perc-encl-parent 0.05 --min-encl-match 0.9 --min-total-encl 10 --posterior-threshold 0.5. Autosomes were run with the --default-prior -3 and chromosome X was run with the --naive option. These options remove TRs with too many alleles which are more likely to be error-prone, process all TRs even if no variation was observed, indicate to use GangSTR-output likelihoods (29) (rather than HipSTR (30)), only output loci if both children in the quad were analyzed, output all loci even if no mutation was observed, apply a constant prior of per-locus mutation rate of 10^{-3}, require *de novo* mutation alleles to be supported by at least 3 enclosing reads, require *de novo* mutation alleles to be supported by fewer than 5% of parent enclosing reads, require 90% of enclosing reads in each sample to match the genotype call, require a minimum of 10 enclosing reads per sample in the family, and label calls with posterior probability ≥0.5 as mutations.

Resulting mutation lists output by MonSTR were subject to further quality control. We filtered families with likely sample contamination evidenced by extreme mutation counts (7 families, number of mutations > 1000), outlier mutation rates (16 families with number of mutations <20 and >241), mutations for which both children in the family were identified as having mutations at the same TR (n=43,239), and TRs with more than 25 mutations identified (n=15) as these are likely error-prone loci. We further filtered: calls for which the child was homozygous for the new allele (n=214,639), loci with a strong bias toward only observing contractions or expansions (n=179, two-sided binomial p<0.0001). We initially observed that mutations for which the parent of origin was homozygous often appeared to be erroneous due to drop out of one allele at heterozygous parents. This was most apparent for large mutations (± ≥ 5 repeat units) involving longer alleles difficult to span with short reads. We thus further required the new alleles to be supported by at least 6 enclosing reads in the child when the parent was called as homozygous.

Our stringent filtering of input genotypes and resulting mutations is unlikely to capture large repeat expansions, which are often not supported by enclosing reads because the resulting alleles are longer than Illumina read lengths. Thus, genotype likelihoods are more spread out and posterior estimates at these loci are lower and they will fail many of the QC options specified above. To additionally identify candidate expansions, we called MonSTR again on each family using the non-default parameter --naive-expansions-frr 3,8 which looks for TRs for which either: (1) the child has at least three fully repetitive reads and both parents have none or (2) the child has at least 8 flanking reads supporting an allele longer than any allele supported in either parent. We filtered candidate expansions identified in more than 3 samples, as we expect expansions to be rare. A total of 78 candidate expansions were identified across all families (**Appendix, Table 9)**. These were merged with the total list of mutations for downstream analysis.

2.4 Evaluation of mutations with capillary electrophoresis (CE) fragment analysis

To directly assess the quality of genotype and mutation calls within families, we performed fragment analysis using capillary electrophoresis on 49 TR mutations across 5 SSC quad families (**Appendix, Table 6**). Tested mutations show a validation rate of 90% (44 out of 49), an improvement over validation rates previously reported for *de novo* indels (39) (**Appendix, Table 7**).

Whole blood-derived genomic DNA for 5 SSC quad families was obtained through SFARI Base to validate a subset of TR mutation calls. For each candidate TR, we designed primers to amplify the TR and surrounding region (**Appendix, Table 6**). A universal M13(-21) sequence (5'-TGTAAAACGACGGCCAGT-3') was appended to each forward primer. We then amplified each TR using a three-primer reaction previously described (40) consisting of the forward primer with the M13(-21) sequence, the reverse primer, and a third primer consisting of the M13(-21) sequence labeled with a fluorophore.

The forward (with M13(-21) sequence) and reverse primers for each TR were purchased through IDT. The labeled M13 primers were obtained through ThermoFisher (#450007) with fluorescent labels added to the 5' ends (either FAM, VIC, NED, or PET). TRs were amplified using the forward and reverse primers plus an M13 primer with one of the four fluorophores with GoTaq polymerase (Promega #PRM7123) using PCR program: 94°C for 5 minutes, followed by 30 cycles of 94°C for 30 seconds, 58°C for 45 seconds, 72°C for 45 seconds, followed by 8 cycles of 94°C for 30 seconds, 53°C for 45 seconds, 72°C for 45 seconds, followed by 72°C for 30 minutes.

The CGG repeat at chr7:103989357 in the 5’UTR of *RELN* could not be amplified using the three-primer method and was genotyped using published primers (41) (forward: 5′-FAM-CGCCTTCTTCTCGCCTTCTC-3′ and reverse: 5′-CGAAAAGCGGGGGTAATAGC-3′). The TR was amplified with HotStarTaq Polymerase (Qiagen #203203) using PCR program: 95°C for 15 minutes, followed by 35 cycles of 94°C for 45 seconds, 58°C for 60 seconds, 72°C for 60 seconds, followed by 72°C for 30 minutes.

Fragment analysis of PCR products was performed on a ThermoFisher SeqStudio instrument using the GSLIZ1200 ladder, G5 (DS-33) dye set, and long fragment analysis options. Resulting .fsa files were analyzed by manual review in GeneMapper (ThermoFisher # 4475073).

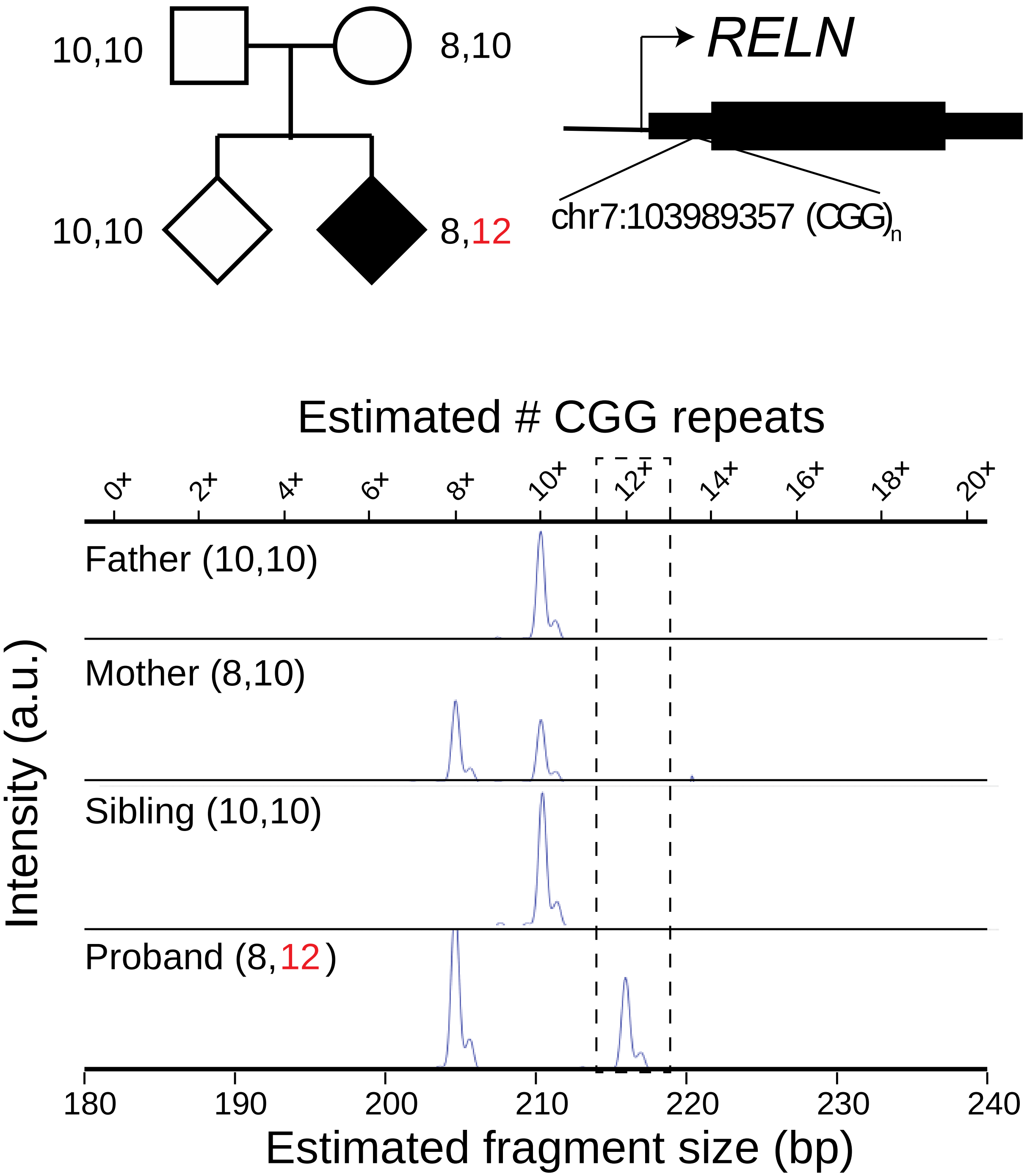


Figure 2: Example TR mutation validated by capillary electrophoresis.

A mutation resulting in two additional copies of CGG in the 5'UTR of the gene RELN (chr7:103989357. hg38) was identified based on WGS analysis (top). Alleles with 11 or more copies of CGG at this TR were previously implicated in ASD and have been shown to reduce RELN expression (42,43) . The region was amplified by PCR and the mutation was confirmed by capillary electrophoresis. Estimated fragment sizes for each sample (bottom x-axis) and the corresponding repeat numbers (top x- axis) are annotated. The fragment length corresponding to the *de novo* allele (12×) is denoted by the dashed gray box.

2.5 Statistical validation of pipeline

2.5.1 Evaluating MonSTR on simulated WGS data

We tested our framework on simulated WGS data, which demonstrated high sensitivity to detect *de novo* TR mutations resulting in changes of up to 10 repeat copies and low false-positive rate (<1%) compared with a naive method in most settings (**Fig. 3**).

We created 78 quad families with 100 TR loci randomly selected from TRs passing all filters described above in the SSC cohort. One simulated quad family consists of the father, mother, child with known mutation (proband), and child with no mutation (control). We tested the ability of our entire pipeline to genotype TRs with GangSTR and call *de novo* mutations with MonSTR. To test the effect of depth of coverage, we generated datasets with 1-50x mean coverage with a mutation size of +1 or -1 repeat unit changes in the proband. To test the effect of TR mutation size, we generated WGS data with 40x coverage and mutations in probands ranging from -10 to 30 repeat unit changes. Contraction mutations that would have resulted in negative repeat copy numbers were excluded. For both tests, we simulated data under three scenarios: (1) both parents with homozygous reference TR genotypes, (2) one parent heterozygous, (3) both parents heterozygous (**Fig. 3**).

WGS data were simulated using ART_illumina v2.5.8 (44) with non-default parameters -ss HS25 (HiSeq 2500 simulation profile), -l 150 (150b reads), -p (paired-end reads), -f coverage (coverage was set as described above), -m 500 (mean fragment size) and -s 100 (standard deviation of fragment size). ART_illumina was applied to fasta files generated from 10Kb windows surrounding each TR locus, applying any mutations as described above. The resulting fastq files were aligned to the hg38 reference genome using bwa mem v0.7.12-r1039 (45) with non-default parameter -R "@RG\tID:sample_id\tSM:sample_id", which sets the read group tag ID and sample name to sample_id for each simulated sample. TRs were genotyped from aligned reads jointly across all members of the same family with GangSTR using identical settings to those applied to SSC data.

We tested three mutation calling settings: a naïve mutation calling method based on hard genotype calls, MonSTR using default parameters, and MonSTR using an identical set of filters as applied to SSC data. We found overall all methods perform similarly well above 30x coverage. At lower coverage, MonSTR's model-based method achieves reduced sensitivity but greater specificity compared to a naïve mutation calling pipeline (**Fig 3**).

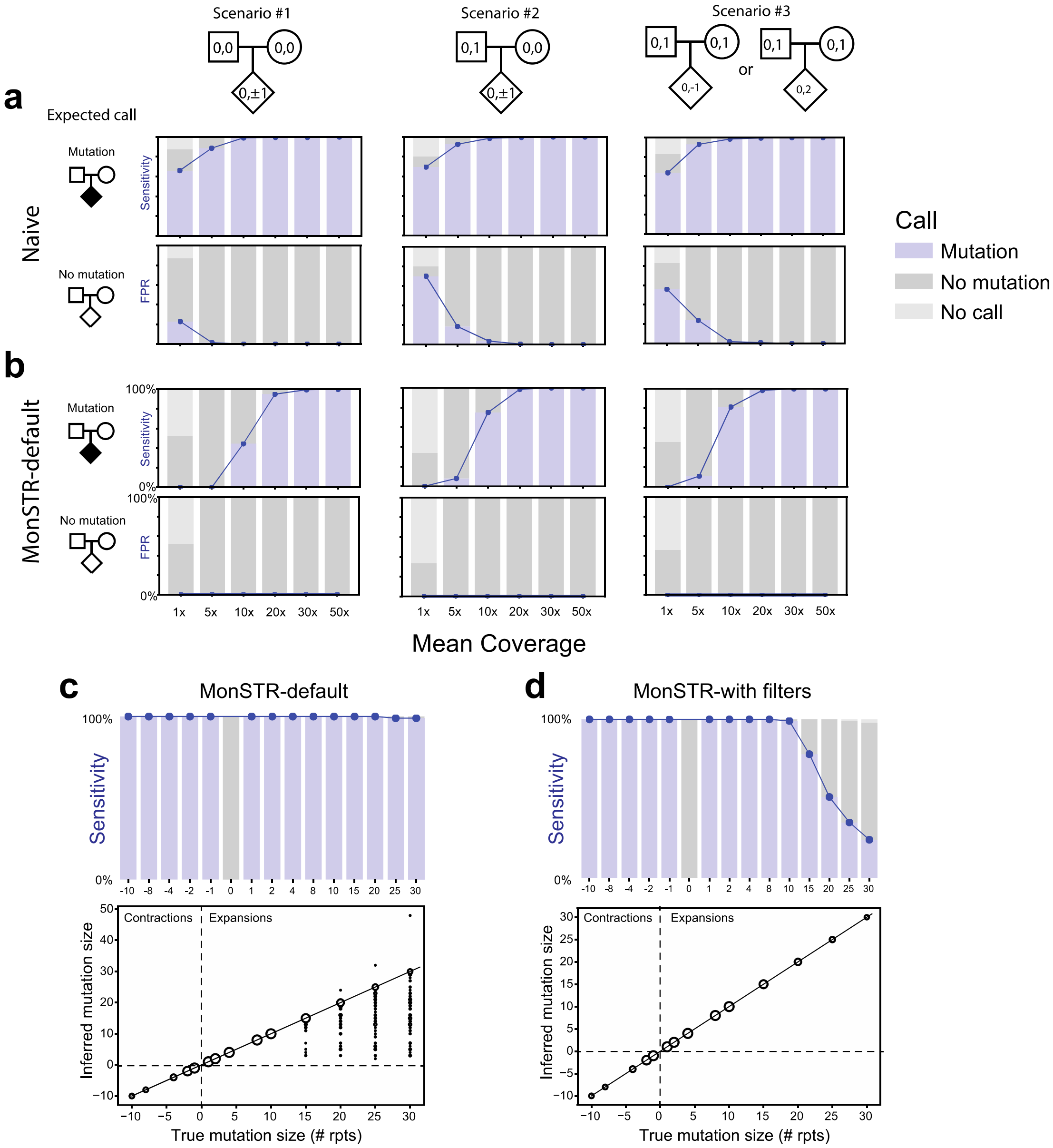


Figure 3: Evaluation of MonSTR using simulated data.

a, Evaluation of a naive TR mutation-calling method. WGS was simulated for probands with mutations and controls with no mutation under three different scenarios for a range of mean sequencing coverages. Top plots show the sensitivity (blue line). Bottom plots show the false positive rate (FPR). Shaded bars show the percent of transmissions called as mutation (blue), no mutation (dark grey), or no call (light ray). b, Evaluation of MonSTR's default model-based method. Plots are the same as in a. but based on MonSTR's default model. Note FPR lines are not visible because all are at 0%. c, Evaluation of TR mutation calling using default model-based MonSTR settings as a function of mutation size. The top plot is the same as in a, b, and shows the sensitivity to detect mutations as a function of their size. The bottom plot compares the estimated called mutation size (y-axis) compared to the true simulated mutation size (x-axis). Bubble sizes show the number of mutation calls represented at each point. d, Evaluation of TR mutation calling as a function of mutation size after quality filtering. Plots are same as in c, but using the stringent quality filters in MonSTR applied to analyze the SSC cohort. Compared to default settings, sensitivity is decreased especially for larger expansions, but inferred mutation sizes are unbiased. All plots are based on simulation of 100 randomly chosen TR loci. c, d, show results for scenario #1.

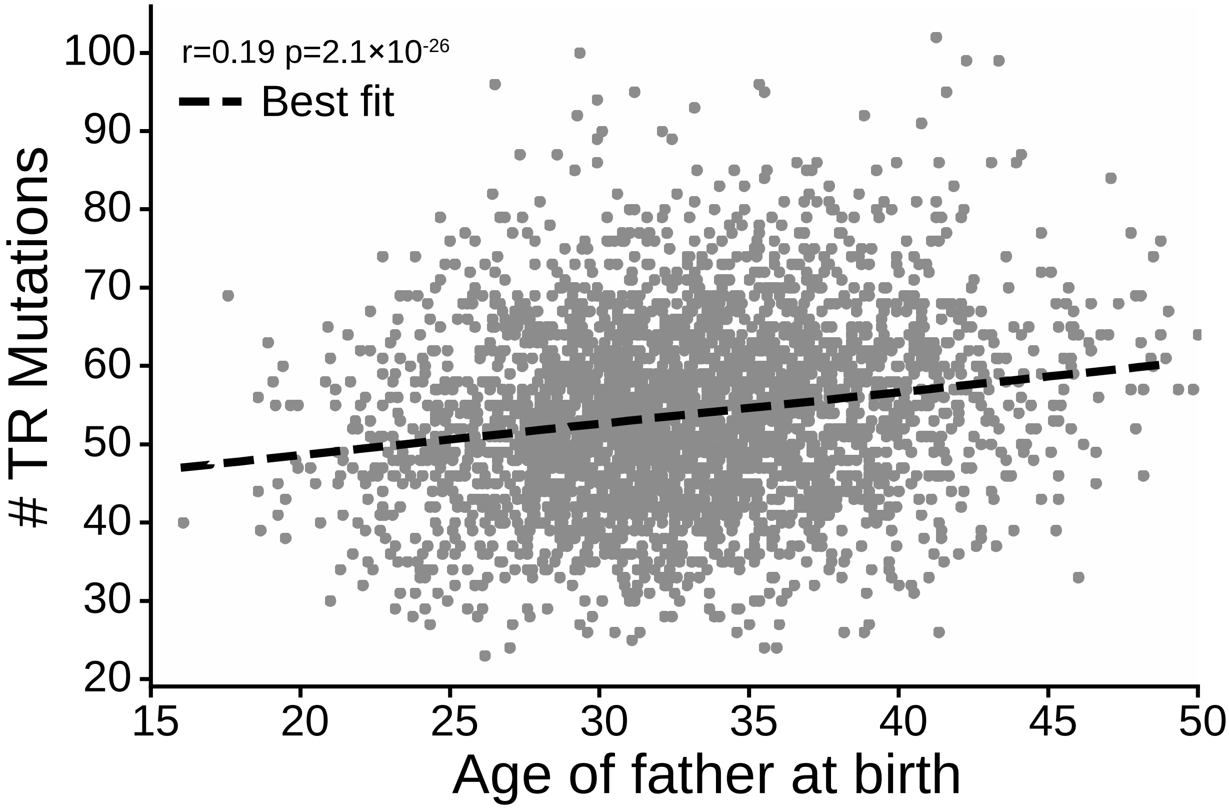


Figure 4: Correlation of mutation rate with paternal age per child.

The scatter plot shows the father's age at birth (*x*-axis) versus the number of autosomal *de novo* TR mutations identified (*y*-axis). Each point represents one child (*n* = 3,186). The dashed black line gives the best fit line.

2.5.2 Comparison to previously reported mutation rates

We next compared our results to known TR mutation trends (**Fig. 5**). Similar to previous studies (6,35,46), estimated mutation rates are highest for TRs with shorter repeat units (**Fig. 5a**) and are positively related to total length (in bp) of the reference TR (**Fig. 5b**).

Following *de novo* SNVs genetic studies (20,47), autosomal TR mutation rates are correlated with paternal age (Pearson's $r = 0.19$. two-sided $P = 2.1 \times 10^{-26}$. $n = 3{,}186$. **Fig. 4**). At TR mutations (excluding homopolymers) for which the parent of origin could be inferred, 74% were phased to the father, which is similar to previous reports for *de novo* SNVs (48,49).

Mutation counts in SSC are concordant with published mutation rates for CODIS forensics TRs (**Fig. 5c**), and are significantly correlated with genome-wide rates estimated by our MUTEA (35) method on an orthogonal set of unrelated individuals (Pearson $r = 0.26$. two-sided $P < 10^{-200}$. $n = 548{,}724$. **Fig. 5d**).

Mutation rates for CODIS markers were obtained from the National Institute of Standards and Technology (NIST) website (https://strbase.nist.gov/mutation.htm). 95% confidence intervals on the estimated number of mutations that should be observed in SSC were obtained by drawing mutation counts from a binomial distribution with n=the total number of children genotyped at each locus and p=the NIST estimated mutation rate. Intervals were obtained based on 1,000 simulations.

Genome-wide autosomal TR mutation rates and constraint scores estimated using MUTEA (35) were obtained from https://s3-us-west-2.amazonaws.com/strconstraint/Gymrek_etal_SupplementalData1_v2.bed.gz (columns est_logmu_ml and zscore_2). TRs were converted from hg19 to hg38 coordinates using the liftOver tool available from the UCSC Genome Browser (38) Store free for academic use (https://genome-store.ucsc.edu/). We intersected the lifted over coordinates with the GangSTR reference using the intersectBed tool included in BEDTools v2.28.0 (50). Only TRs overlapping GangSTR TRs by at least 50% (-f 0.5) and with the same repeat unit length in each set were used for analysis.

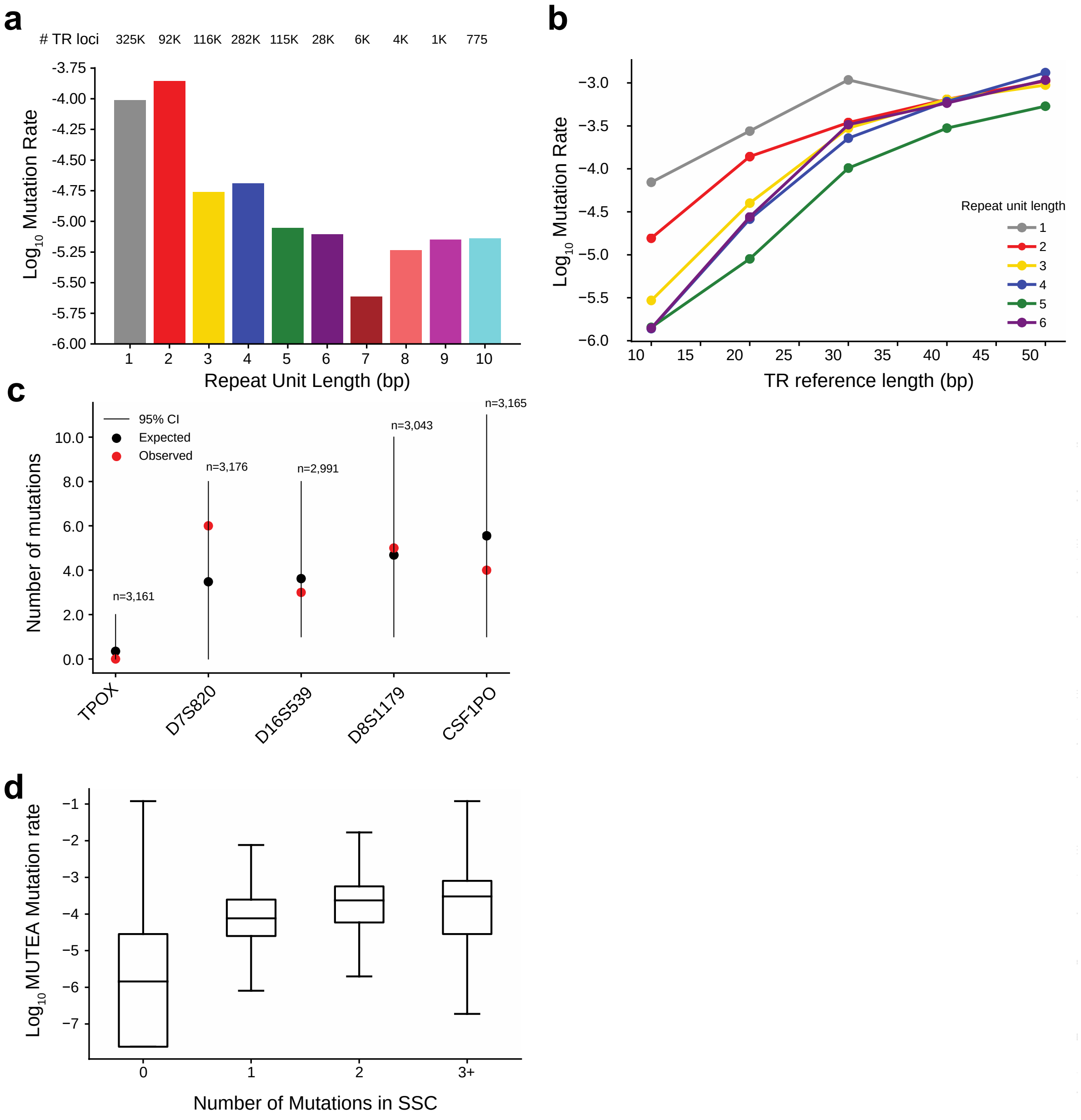


Figure 5: Genome-wide *de novo* TR mutation rate patterns.

a, Distribution of average TR mutation rates by period. For each repeat unit length (x-axis), bars give the genome-wide estimated TR mutation rate (y-axis, log10 scale). Average mutation rates were computed as the total number of mutations divided by the total number of children analyzed. The numbers of TRs considered (rounded to the nearest 1,000) in each category are annotated. b, TR mutation rate vs. length. The x-axis shows the TR reference length (hg38), and the y-axis shows the log10 mutation rate estimated across all TRs with each reference length. Colors denote different repeat unit lengths. c, Number of TR mutations observed for CODIS markers. Red dots show observed mutation counts. Black dots show expected mutation counts and lines give 95% confidence intervals based on mutation rates reported by NIST. Each x-axis category denotes a separate CODIS marker. The total number of children analyzed is annotated above each marker d, Observed TR mutation counts concordant with MUTEA (35). Boxes show the distribution of log10 mutation rates estimated by MUTEA (y-axis) at each TR with a given number of mutations observed in SSC children (x-axis). Black middle lines give medians and boxes span from the 25th percentile (Q1) to the 75th percentile (Q3). Whiskers extend to Q1-1.5*IQR (minima) and Q3+1.5*IQR (maxima), where IQR gives the interquartile range (Q3-Q1). Data are shown for n = 548,724 TRs for which MUTEA estimates were available.

2.6 Acknowledgements

Chapter 2 is an adapted reprint of the material "Patterns of *de novo* tandem repeat mutations and their role in autism" by Ileena Mitra, Bonnie Huang, Nima Mousavi, Nichole Ma, Michael Lamkin, Richard Yanicky, Sharona Shleizer-Burko, Kirk E. Lohmueller, and Melissa Gymrek published in *Nature* 2021. The dissertation author was the primary investigator and author of this paper.

CHAPTER 3: PATTERNS OF TANDEM REPEAT MUTATIONS IN THE GENERAL POPULATION

3.1 Introduction

Studying genomic mutational mechanisms is necessary to the understanding of evolution and basic DNA biology. Population genetics studies analyzing genomic mutations have focused on SNVs, indels, and CNVs (**Table 4**) (4). To date, studies on TR mutational processes have been limited in the number of TR loci analyzed due to the availability applicable technologies (6). Therefore, the genome-wide mutational patterns of *de novo* repetitive mutations remained to be understood.

We applied MonSTR and the bioinformatics pipeline (**Chapter 2, Fig. 1**) to analyze *de novo* TR mutations first in the SSC 1,593 unaffected children (**Table 3**). We characterized genome-wide mutational properties of autosomal and chromosome X TR loci in the general population. Seeing that TR motif classes have different mutational patterns, we examined each motif class separately, as well, as together. We sought to assess the contribution of external factors (e.g., sex, parental age, parent of origin) and intrinsic sequence features (e.g., GC content, recombination rate, replication timing, etc.) to modulate mutation rate for each TR class. Our bioinformatics pipeline (**Fig. 1**) allowed for the analyzes of over of over 1 million TRs genome-wide and varying motif sizes up to 20 bp. Our analysis of precise TR mutation changes and their sizes enables the first genome-wide characterization of TR mutation properties. Our results below highlight the importance of including TRs in human genetic studies because *de novo* TRs double the known average mutation burden per child (**Table 5, Fig. 6**). Importantly, this analyses allowed us to gain a better understanding of TR mutational mechanisms, such as, new insights on the influence of parental origin (**Section 3.3.2**) and length-bias in mutation sizes (**Section 3.3.3**).

3.2 Methods

3.2.1 Determinants of TR mutation rates

Genomic and epigenomic features for each TR (**Fig. 13**) were compiled from a variety of resources. BEDTools intersect v2.28.0 (50) was used to overlap GangSTR reference TRs for each annotation after using liftOver to convert each to hg38 coordinates. Recombination rates (51) were obtained from https://github.com/cbherer/Bherer_etal_SexualDimorphismRecombination/blob/master/Refined_genetic_map_b37.tar.gz. Rates were $\log_{10}$ transformed with a pseudo count of 0.0001 to avoid infinite values. TRs with scaled recombination rates ≤-3 were filtered. PhastCons (52) annotations were obtained from http://hgdownload.cse.ucsc.edu/goldenpath/hg38/phastCons100way/. Values were $\log_{10}$ transformed with a pseudo count of 0.0001 to avoid infinite values. TRs with transformed scores ≤-4 were filtered. Nucleosome occupancy scores were obtained from http://hgdownload.soe.ucsc.edu/goldenPath/hg18/database/uwNucOccMec.txt.gz (Mec (53)) and http://hgdownload.soe.ucsc.edu/goldenPath/hg18/database/uwNucOccDennis.txt.gz (Dennis (54,55)). Notably these two annotations were scored in opposite directions and should be anti-correlated. TRs with "Mec" annotations ≤-1 were filtered. TRs with "Dennis" annotations ≤-4 or ≥2, respectively, were filtered. Conserved promoter annotations were obtained from Table S9 of An *et al* (39). DNaseI hypersensitivity peaks based on 125 cell types profiled by the ENCODE Project (56) were obtained from the UCSC genome browser (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeRegDnaseClustered/) and were treated as binary features. Histone modification peaks for embryonic stem cells (H1) were obtained from the Encode Project website (https://www.encodeproject.org) and were treated as binary features (accessions ENCFF180RPI, ENCFF720LVE, ENCFF835TGA, ENCFF219TGT, ENCFF483GVK, ENCFF695ZZV, ENCFF781GRI, ENCFF067WBB, ENCFF714VTU, ENCFF073WSF). GC content for

TR motifs and for varying size windows around each TR were computed using a custom script based on the hg38 reference genome.

We fit a Poisson regression model to predict mutation counts in unaffected individuals at each TR based on these features. A separate model was fit for each feature and each repeat unit length, in each case using TR reference length (in bp) as a covariate. In each model the exposure was set to the number of observed transmissions in unaffected individuals. Models were fit using the discrete.discrete_model.Poisson module from the Python statsmodels library v0.10.1 (https://www.statsmodels.org/).

3.2.2 Analysis of mutation directionality bias

The observed bias of longer alleles to contract and shorter alleles to expand (**Fig. 11**) could potentially be explained by genotyping errors at heterozygous loci due to "heterozygote dropout" of long alleles, leading to erroneous homozygous genotype calls. To reduce the potential impact of heterozygote dropout on apparent mutation directionality, we restricted this analysis to mutations with an absolute size of ≤5 units. When analyzing mutations from heterozygous vs. homozygous parents (**Fig. 12**), we further restricted to mutations consisting of a single unit and for which the child had at least 10 enclosing reads supporting the *de novo* allele, indicating the allele could be easily spanned and would be less prone to dropout.

3.3 Genome-wide patterns of *de novo* TR mutations

After applying the bioinformatics pipeline in about 1600 simplex ASD families (**Chapter 2**), we identified a total of 175,291 high-confidence TR mutations across 94,616 distinct loci in 1,593 families (average 53.9 autosomal mutations per child. **Fig. 6**) corresponding to an average mutation rate of 5.6×10^{-5} mutations per locus per generation.

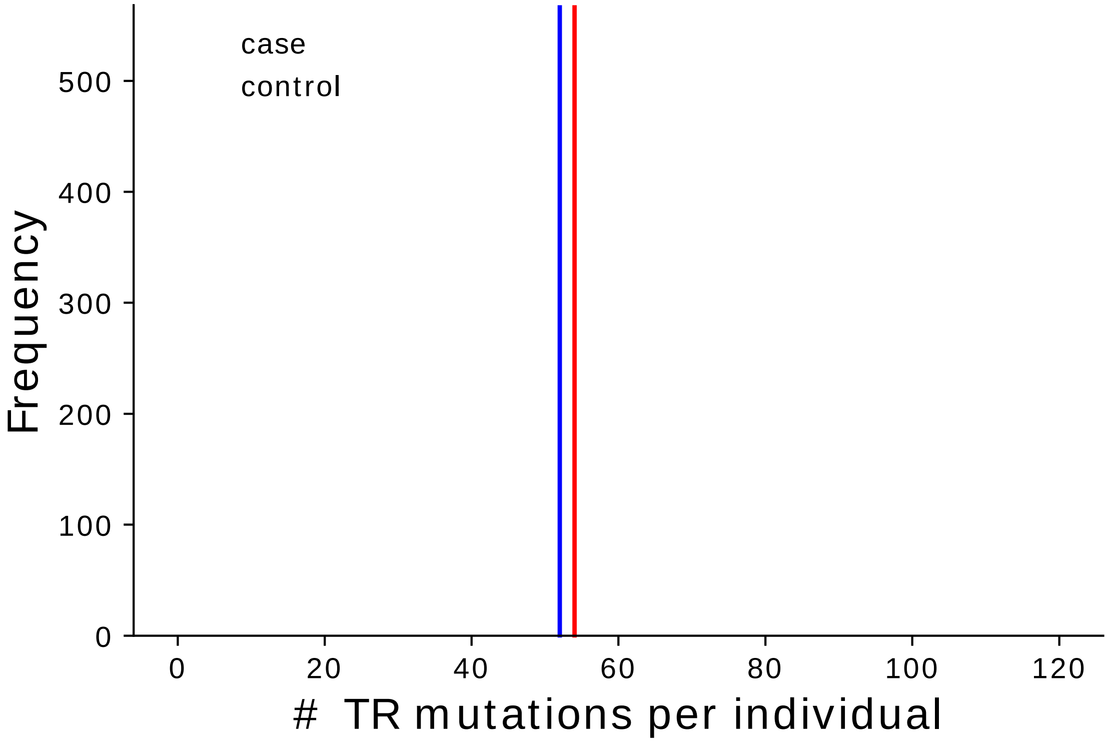


Figure 6: Distribution of the number of autosomal *de novo* TR mutations.

TR mutation counts are shown for non-ASD siblings (blue) and probands (red).

3.3.1 Expansions versus contractions TR mutations

The majority of mutations observed result from expansions or contractions by a single repeat unit, with a smaller proportion of larger mutations (**Fig. 7**), although this trend varies by repeat unit length

(6,57–59) (**Table 4, Fig. 8**). Overall, mutations show a bias toward expansions (71%) vs. contractions (29%). When excluding error-prone homopolymer TRs, only 56% of mutations are expansions, still significantly more than the 50% expected by chance (binomial two-sided P=4.8×10^{-249}. n=71,822).

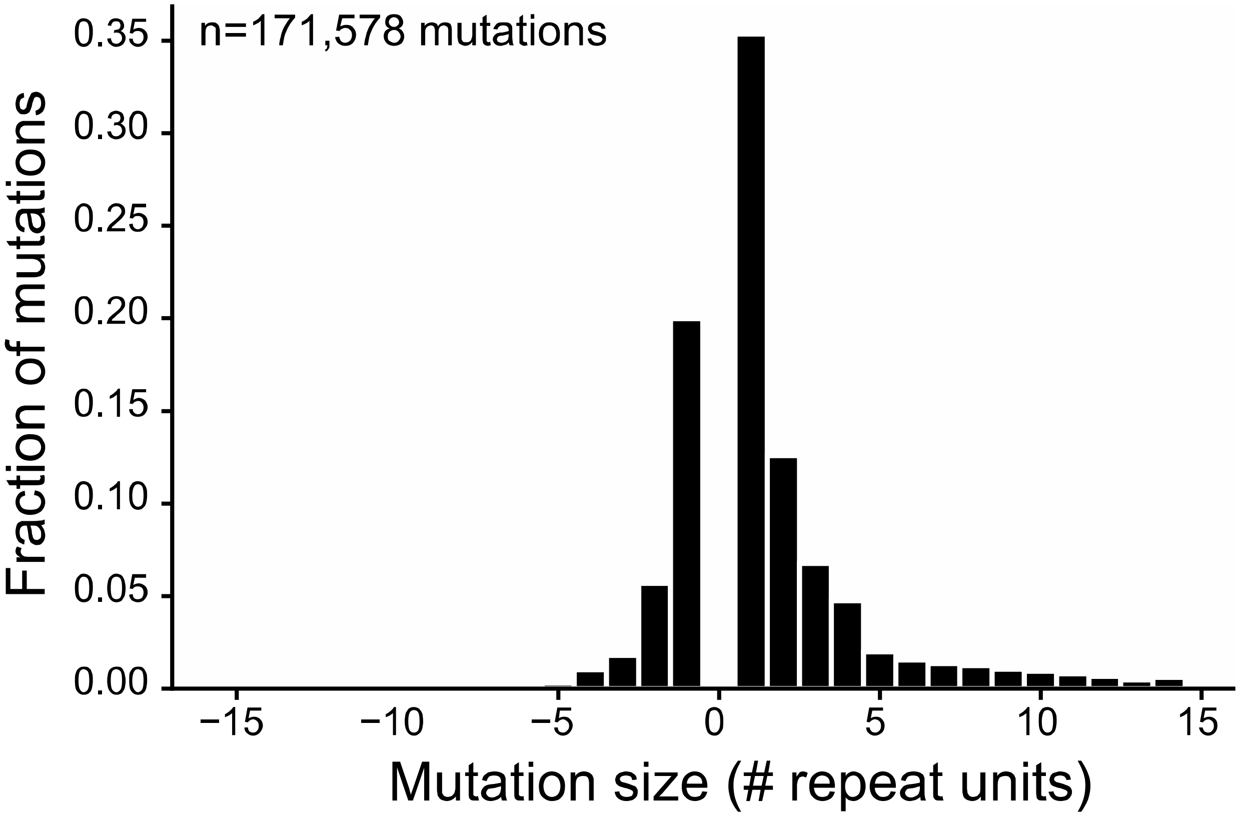


Figure 7: Size distribution of TR mutations.

Sizes are expressed in terms of repeat units, where >0 represents expansions and <0 represents contractions.

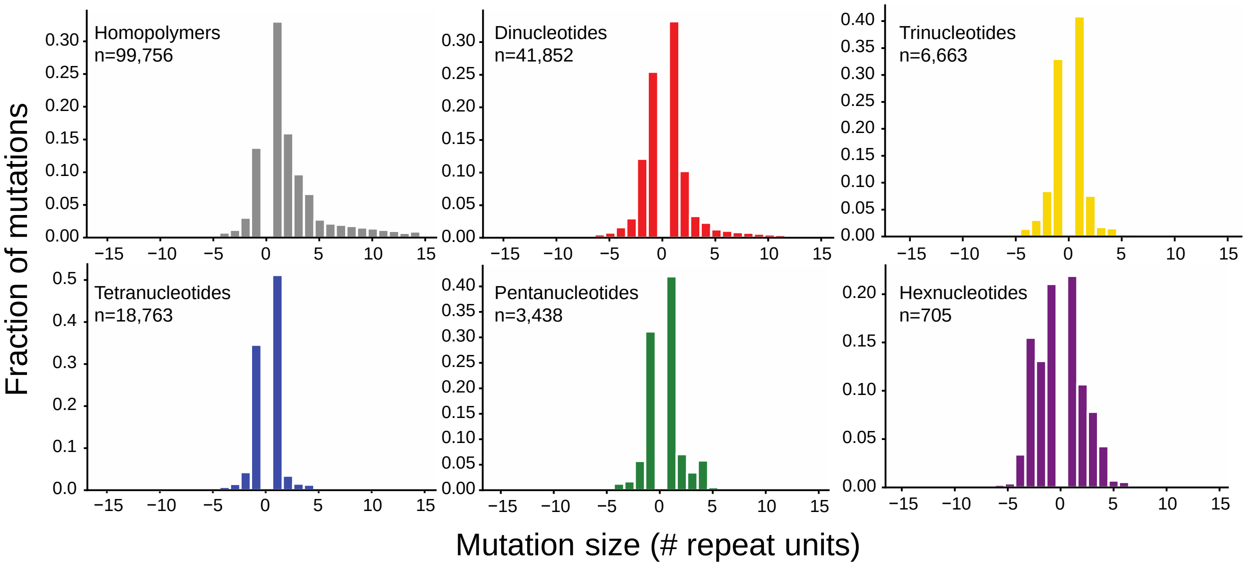


Figure 8: Mutation size distributions by repeat unit length.

Histograms show the distribution (*y*-axis, fraction of total) of *de novo* TR mutation sizes for each repeat unit length (*x*-axis, number of repeat units). Mutations <0 denote contractions and >0 denote expansions. Colors denote different repeat unit lengths (grey = homopolymers. red = dinucleotides. gold = trinucleotides. blue = tetranucleotides. green = pentanucleotides. purple = hexanucleotides).

Table 4: Comparison to previously reported TR mutation parameters.

Mutation rates were computed as the total number of mutations divided by the total number of transmission events observed at each class of TRs. Only healthy individuals were included in this analysis. p denotes the probability that a mutation is a single unit.

Study	Repeat unit (bp)	# TRs	# Mutations	Mutation rate	Step size (p)
Present study	1	325,079	49,106	9.8 × 10^-5	0.47
	2	92,425	20,677	1.4 × 10^-4	0.59
	3	116,185	3,253	1.8 × 10^-5	0.74
	4	282,332	9,253	2.1 × 10^-5	0.86
	5	115,465	1,641	9.0 × 10^-6	0.73
	6	28,166	356	8.0 × 10^-6	0.43
Weber and Wong	2,4	28	24*	1.2 × 10^-3	0.87
Ellegren	2 4	52	102	-	0.85 0.92
Huang et al	2	362	97	1.9 × 10^−4	0.37
Ballantyne et al (Y-STRs)	3-6	186	924	3.78×10^−4 to 7.44×10^−2	0.96
Sun et al	2 4	2,477	2,058	10.01 × 10^−4 (tetra) 2.73 × 10^−4 (di)	0.68 0.99

3.3.2 TR mutations show distinct patterns based on parent of origin

We find significant biases in TR mutation characteristics arising in maternal vs. paternal germlines which provide insights into general biological mechanisms of TR mutation. We examined mutation sizes separately for the subset of mutations phased to either the maternal or the paternal germline. The bias towards expansions versus contractions (excluding homopolymers) is significant for maternal phased mutations (57% expansions. binomial two-sided $P = 3.7 \times 10^{-39}$. $n = 9{,}190$) but not for paternal phased mutations (50% expansions. $P = 0.71$. $n = 26{,}550$) (**Fig. 10**), suggesting that the overall expansion bias observed is primarily driven by maternally derived mutations. Further, maternal phased mutations result in significantly larger changes in repeat unit copy number (Mann–Whitney one-sided

P < 10^{-200}). This trend is recapitulated across all repeat unit lengths (**Fig. 9**), with the strongest effect at dinucleotide TRs.

Given that we find that distinct parent of origin patterns on mutation frequency and size, we believe distinct prezygotic biological mechanism are involved. Strand slippage during DNA replication is widely considered the predominant driver of TR mutations (60). However, previous studies have reported that other mechanisms, including non-homologous end joining (NHEJ) of DNA double-stranded breaks (61) and recombination-mediated processes such as meiotic gene conversion (62), may also play a role. Intriguingly, we find that mutations derived from maternal germlines are significantly larger and more prone to expansion than those from paternal germlines. Whereas spermatogonia undergo more frequent mitosis events which may lead to a higher rate of slippage events, oocytes lie dormant for decades and can accumulate DNA damage that must be repaired by error-prone processes such as homologous recombination or NHEJ (63). Further, oocytes have been reported to have crossover frequencies 1.7x of that of spermatocytes (64), providing increased potential for recombination-mediated mutations. Our results are consistent with a stronger influence of slippage resulting in smaller mutations in the paternal germline and of alternative TR mutation processes leading to larger mutations in the maternal germline.

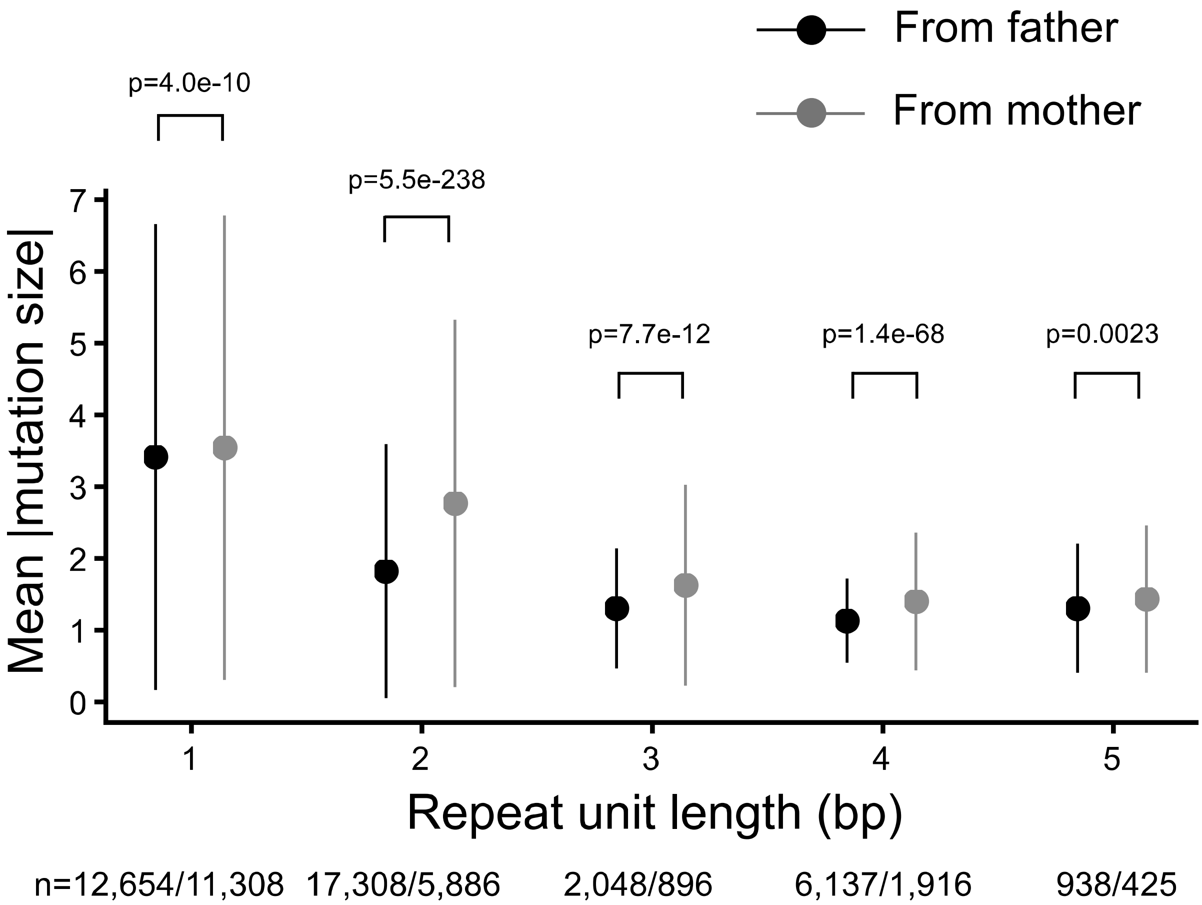


Figure 9: Mean absolute mutation size by parental origin.

Dots show the mean absolute mutation size for mutations phased towards the paternal (black) and the maternal (grey) germlines. Data are mean ± s.d. One-sided *P*-values were computed using a Mann–Whitney test.

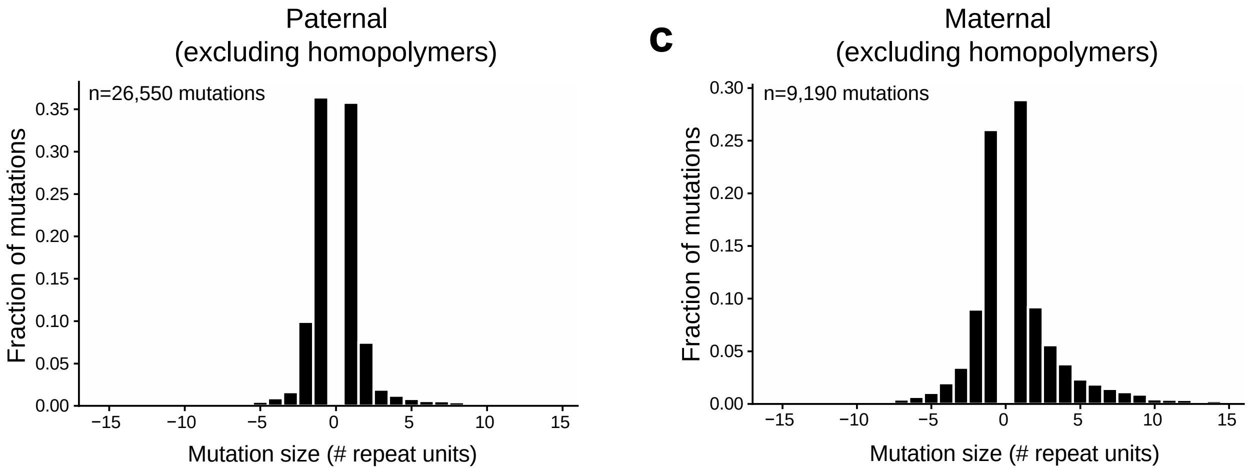


Figure 10: Mutation size distributions by parental origin.

Histograms show the distribution of *de novo* TR mutation sizes for mutations arising in the paternal (b) and maternal (c) germlines (homopolymers excluded).

3.3.3 Allele directionality bias

Previous studies assessing TR mutational patterns reported a directionality bias in mutations, with longer alleles more likely to experience contractions and shorter alleles more likely to experience expansions (6,35,65). We observe a similar bias (**Fig. 11**). We find that the directionality bias is notably stronger for mutations originating from parents heterozygous for two different allele lengths (**Fig. 12**), whereas little bias is observed for mutations from homozygous parents. This suggests that the observed trend could be driven in part by interaction between parent alleles, which has been previously hypothesized (65).

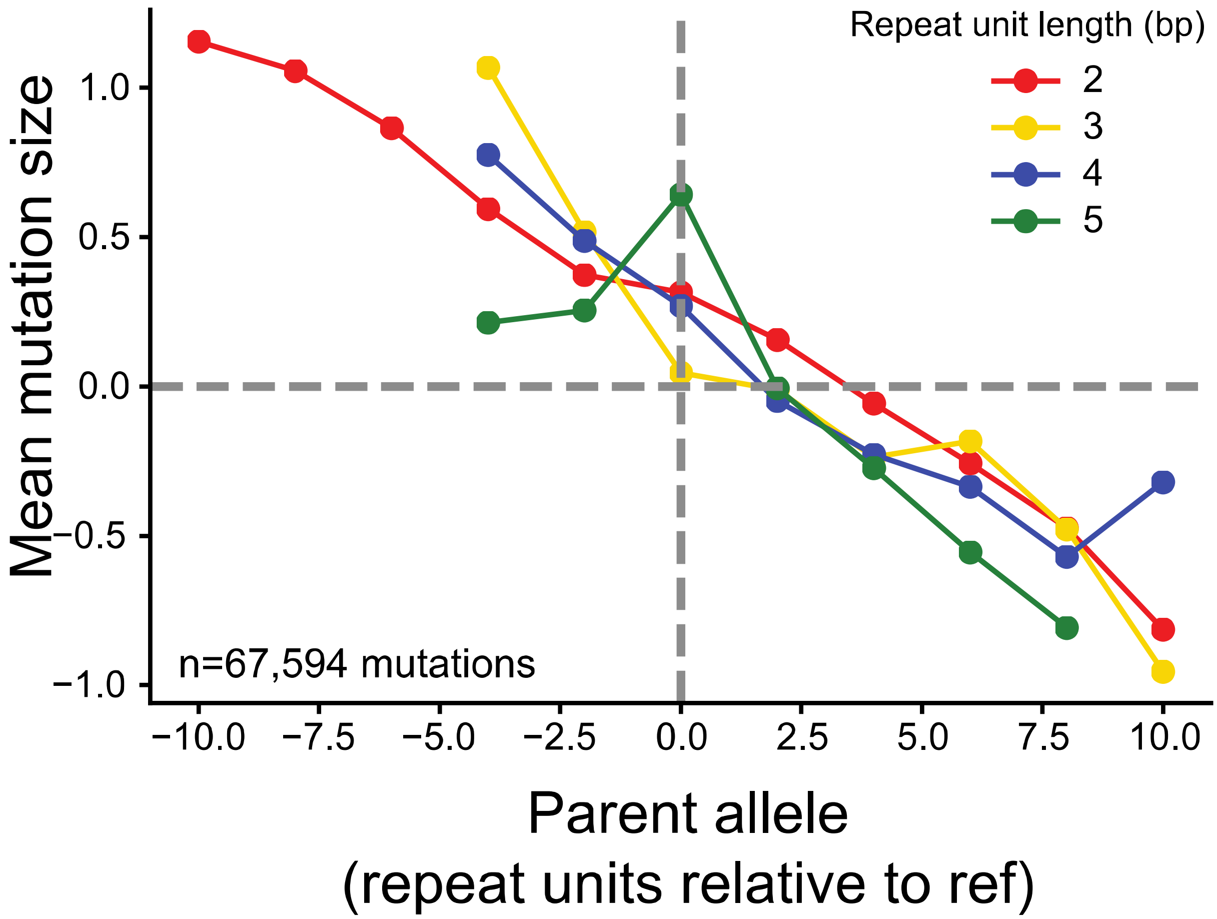


Figure 11: Directionality bias in TR mutation size.

The *x*-axis gives the size of the parent allele relative to the hg38 reference human genome.

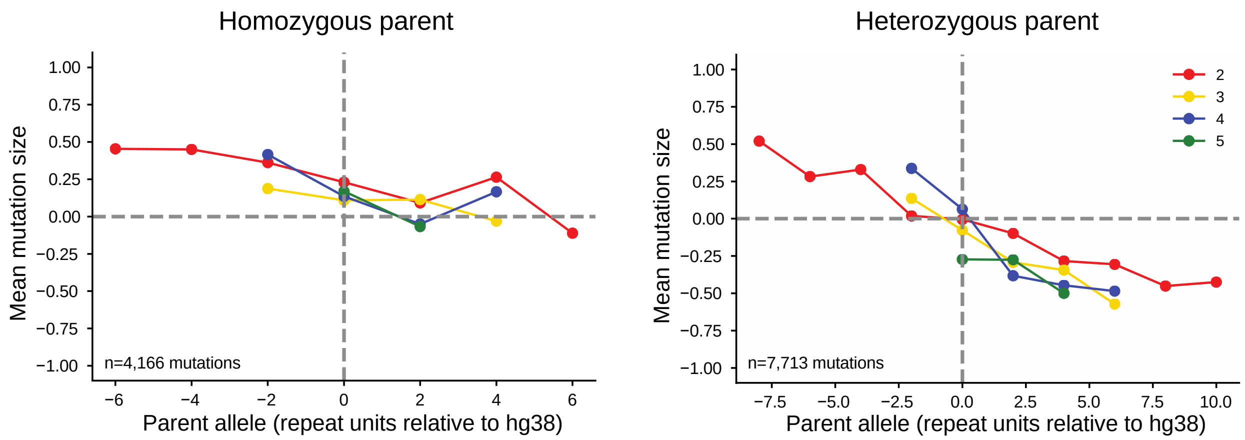


Figure 12: Mutation directionality bias in homozygous vs. heterozygous parents.

In each plot, the *x*-axis gives the size of the parent allele relative to the reference genome (hg38). The *y*-axis gives the mean mutation size in terms of number of repeat units across all mutations with a given parent allele length. A separate colored line is shown for each repeat unit length (red = dinucleotides. gold = trinucleotides. blue = tetranucleotides. green = pentanucleotides). Plots are restricted to mutations that were successfully phased to either the mother or the father for which the parent of origin was homozygous (b) or heterozygous (c). To restrict to highest confidence mutations, these plots are based only on mutations with step size of ±1 and for which the child had more than 10 enclosing reads supporting the *de novo* allele.

3.3.4 Influence of local genomic and epigenomic features on TR mutation rates

We investigated determinants of TR mutation rates and found that local genomic features are only modestly predictive of TR mutation rates, similar to previous reports (**Fig 13**). We investigated relationships between TR mutation rates and genomic or epigenomic features. We fit a separate Poisson regression for each repeat unit length relating observed mutation counts to each feature. As expected, reference TR length is the strongest predictor of mutation rates across all TRs (**Fig 13**). Several features show similar patterns across all TRs, including the presence of active chromatin marks (negative effect) and recombination rate (positive effect, which has been previously suggested (66)). Other features, such as GC content, show distinct patterns across different TR classes. These results suggest TR variation is driven by a variety of mutational mechanisms that may be unique to each TR unit class.

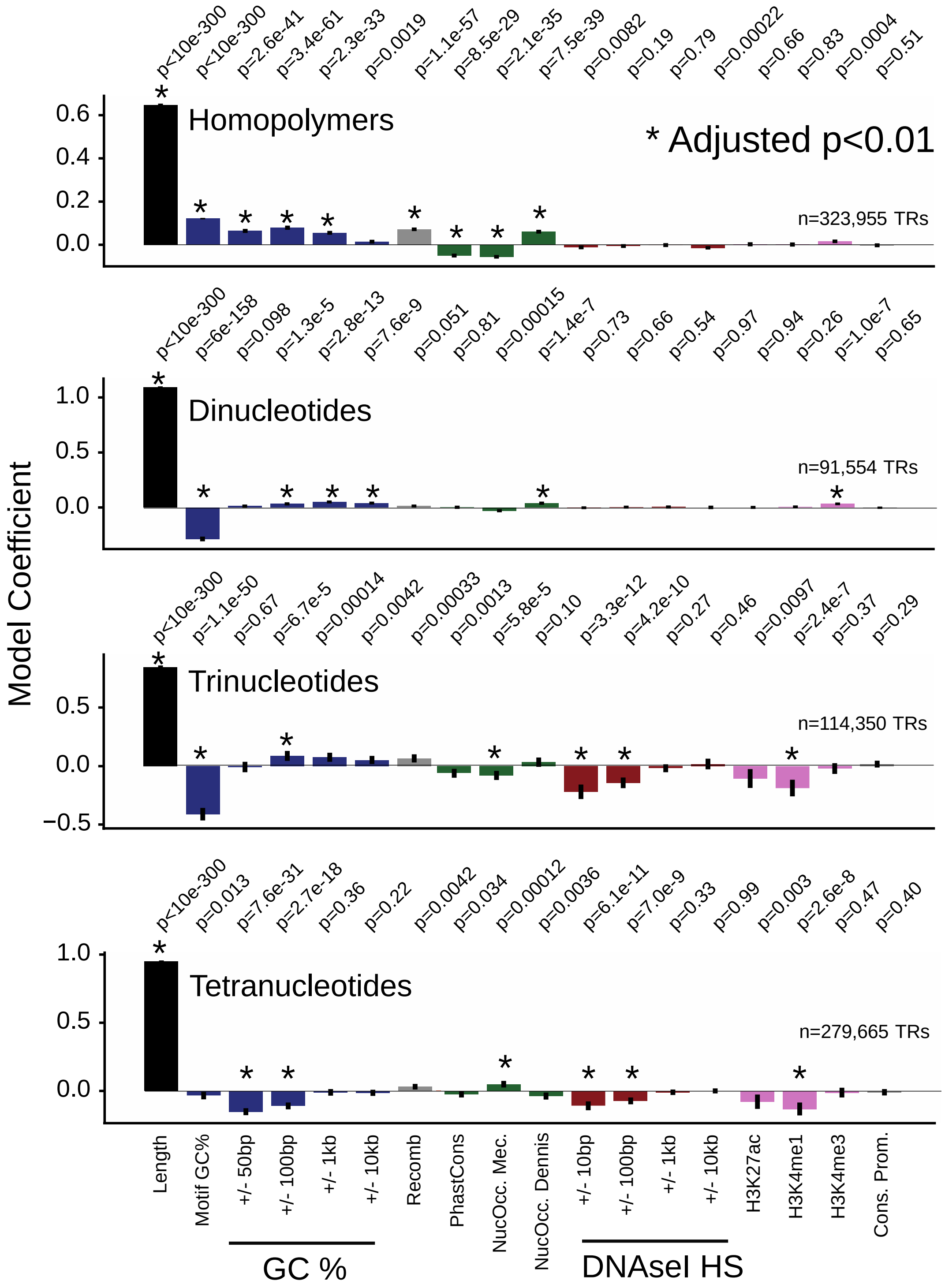


Figure 13: Determinants of TR mutation rates.

The Poisson regression coefficient is shown for each feature in models trained separately for each repeat unit length. Features marked with an asterisk denote significant effects (two-sided *P* < 0.01 after Bonferroni correction for the number of features tested across all models). Nominal *P*-values are annotated above each plot. Error bars give 95% confidence intervals.

3.4 Acknowledgements

Chapter 3 is an adapted reprint of the material “Patterns of *de novo* tandem repeat mutations and their role in autism” by Ileena Mitra, Bonnie Huang, Nima Mousavi, Nichole Ma, Michael Lamkin, Richard Yanicky, Sharona Shleizer-Burko, Kirk E. Lohmueller, and Melissa Gymrek published in *Nature* 2021. The dissertation author was the primary investigator and author of this paper.

CHAPTER 4: PATTERNS OF DE NOVO TANDEM REPEAT MUTATIONS IN AUTISM

4.1 Introduction

Autism spectrum disorders (ASDs) are due to complex genetic factors, including common polygenic variation, rare variants, and epistatic effects (2). Genetic studies have been unraveling the complex genetic architecture of ASD for decades. Many studies have found a significant contribution of *de novo* mutations to ASD (21–23,67). These studies, and ours, aim to better understand the pathology of ASD.

To better understand the contribution of *de novo* TR mutations to ASD, we applied MonSTR and the bioinformatics pipeline (**Chapter 2, Fig. 1**) to the SSC dataset, which consists of simplex families (**Table 3**). The sporadic autism diagnosis in only a single child in the simplex families suggests that *de novo* mutations are more likely to contribute to the ASDs phenotype (20). We compared the TR mutational burden in ASD children to their matched unaffected siblings in order to identify genetic patterns associated with ASD. We sought to assess the mutational burden in genomic regions (e.g., coding, promoter, intron, etc.) to identify disrupted regions relevant to ASD. Our analyses reveled the ways in which *de novo* TR mutations differ in ASD, such as, by number, size, and other factors (**Section 4.3**). We also assessed the functional consequences of TR mutations on gene expression (**Section 4.5**), highlighting the importance to study expression TRs in the context of diseases (15). This analyses allowed us to gain a better understanding of how TRs play a role in in the genetic etiology of ASD.

4.2 Methods

4.2.1 Mutation burden statistical testing

Mutation excess in probands vs. non-ASD siblings was tested using a paired t-test as implemented in the Python SciPy library v1.3.1 (https://docs.scipy.org/doc/scipy/reference/index.html) function scipy.stats.ttest_rel. We compared a vector of counts of mutations in probands to a vector of counts in mutations in non-ASD siblings, ordered by family ID.

Comparison of TR mutation burden in probands vs non-ASD siblings was also computed after adjusting for the father's age at birth. We used the Python statsmodels ordinary least squares regression module to regress unaffected mutation counts on paternal age. We then used this model to compute residual mutation counts in each sample after regressing on paternal age.

Relative risk was computed as the ratio of the mean number of mutations in probands vs. non-ASD siblings. Relative risk of greater than 1 indicates a higher burden in the probands. We estimated a 95% confidence interval on the fraction of mutations $p = \frac{n_p}{n_p + n_s}$ in each category that are in probands vs. siblings based on a binomial distribution ($SE(p) = \sqrt{\frac{p(1-p)}{n_p + n_s}}$) where n_p and n_s are the number of mutations observed in probands and siblings, respectively. We then used the upper and lower bounds on the fraction of mutations in probands $p_{low} = p - 1.96SE(p)$. $p_{high} = p + 1.96SE(p)$ to compute the corresponding 95% confidence intervals for relative risk as $\left(\frac{t_s p_{low}}{t_p(1-p_{low})}, \frac{t_s p_{high}}{t_p(1-p_{high})}\right)$, where t_s and t_p are the total number of sibling and proband samples considered, respectively.

Gene annotations were obtained from the UCSC Table Browser (38) using the hg38 reference genome. Fetal brain promoter and enhancer annotations were obtained from fetal brain male ChromHMM

(68) annotations available on the ENCODE Project website (https://www.encodeproject.org/. accession ENCSR770CMJ).

The contribution of *de novo* TR mutations to ASD risk was calculated by taking the difference in total autosomal mutations identified in probands vs. siblings divided by the number of probands, as was done in a previous study of non-coding mutations in ASD (27).

4.2.2 Enrichment of common variant risk

GWAS SNP associations were downloaded from GWAS catalog (69) (ASD [EFO_0003756] n=637 SNPs. SCZ [EFO_0000692] n=3,476. EA [EFO_0004784] n=3,966). We tested whether TR mutations falling within 50kb of autosomal GWAS SNPs for each trait showed increased burden in probands vs. siblings by performing a Mann-Whitney test (Python function scipy.stats.mannwhitneyu) comparing mutation counts in probands vs. non-ASD siblings. We performed an additional test excluding mutations resulting in alleles with AF<0.05.

4.2.3 Gene-set enrichment analysis using MAGMA

For gene-set enrichment analysis, the autism gene set was defined as genes with coding or promoter TR mutations (transcription start site +/- 5kb) in probands only (n=268 genes). We similarly defined a control gene set with coding or promoter mutations only in unaffected siblings (n=242 genes). Genes that could not be mapped to Entrez IDs and SNPs required for MAGMA analyses were excluded. These are the same gene sets as used in **Fig. 21** and **Fig. 22a**. We performed an additional test with genes limited to predicted pathogenic mutations in proband (n=17 genes were successfully mapped) vs unaffected (n=5 genes were successfully mapped) (**Appendix, Table 10**). As input we used SNP summary statistics available from GWAS for ASD (70), schizophrenia (SCZ) (71), and educational attainment (EA) (72). NCBI gene definitions were used with an upstream and downstream window of 10kb, and gene-level SNP *P*-values and gene-set enrichment *P*-values were obtained using the default settings in MAGMA.

4.2.4 Gene expression analysis

The Developmental Transcriptome dataset containing RNA-seq normalized gene expression values and meta-data for developmental brain tissue regions was downloaded from the BrainSpan Atlas of the Developing Human Brain (73) (https://www.brainspan.org/static/download.html). Expression values were log-transformed before analysis, adding a pseudo count of 0.01 to avoid 0 values. We excluded brain structures "CB", "LGE", "CGE", "URL", "DTH", "M1C-S1c", "Ocx", "MGE", "PCx", and "TCx" since those structures only had data for male samples at 3 or fewer time points. We used a one-sided Mann-Whitney test (scipy.stats.mannwhitneyu) to compare the distribution of expression in genes with only proband mutations vs. genes with only unaffected sibling mutations separately for each tissue. Meta-analysis across all brain regions was performed using Fisher's method to combine *P*-values. The following abbreviations are used for brain structures: A1C=primary auditory cortex. AMY= amygdaloid complex. CBC=cerebellar cortex. DFC=dorsolateral prefrontal cortex. HIP=hippocampus. IPC=posteroventral (inferior) parietal cortex. ITC=inferolateral temporal cortex. M1C=primary motor cortex. MD=mediodorsal nucleus of thalamus. MFC=anterior cingulate cortex. OFC=orbital frontal cortex. S1C=primary somatosensory cortex. STC=posterior superior temporal cortex. STR=striatum. V1C=primary visual cortex. VFC=ventrolateral prefrontal cortex. Expression STR summary statistics were obtained from Supplementary Data 2 of Fotsing, *et al.* (15).

4.3 TR mutation burden in ASD

The total number of *de novo* autosomal TR mutations observed genome-wide is significantly higher in probands (mean=54.65 mutations) vs. non-ASD siblings (mean=53.05 mutations) (**Fig. 6 and 14,** paired t-test two-sided P=9.4×10^{-7}. n=1,593. relative risk [RR] = 1.03). This trend remains after adjusting mutation counts for paternal age (P=1.08×10^{-5}), excluding homopolymers (P=0.0071 after paternal age adjustment), and is consistently observed across each SSC phase (**Table 5**). Autosomal mutations in probands result in significantly larger repeat copy number changes (Mann-Whitney one-sided P=0.017. **Fig. 15**). We analyzed chromosome X mutations separately and observed a moderate excess in male probands vs. male non-ASD siblings (Mann-Whitney two-sided P=0.01) but no difference in females (P=0.73).

Our study is underpowered to detect specific TR loci enriched for mutations in probands vs. siblings at genome-wide significance (**Fig. 16**). Instead, we evaluated whether TRs within particular genomic annotations show an excess of mutations in probands vs. non-ASD siblings (**Fig. 14**). Mutations in coding regions have the highest magnitude of excess in probands vs. non-ASD siblings, but the excess is not statistically significant (RR=1.67. paired t-test two-sided P=0.16) likely due to the small number of autosomal coding mutations (n=32, **Appendix, Table 8**). We observe significant enrichment for *de novo* TR mutations falling within annotated fetal brain promoters (**Fig. 14**. RR=1.20. paired t-test two-sided P=0.0013. significant after Bonferroni correction for 7 tests), which was observed previously for non-coding point mutations[10].

The observed genome-wide excess of TR mutations in probands is modest (RR=1.03), suggesting that only a subset of mutations are pathogenic. Indeed, the majority (84%) of TR mutations result in alleles that are already common (allele frequency [AF] ≥1%) in unaffected SSC parents, and thus, are likely benign. When we stratify our mutation burden analysis by the frequency of the mutant allele (**Fig. 18**), we find that the mutation excess in probands increases for mutations resulting in rarer alleles, with the strongest effect at alleles unobserved (AF=0) in SSC parents (RR=1.10. paired t-test two-sided *P*=0.021. **Fig. 17**). This pattern remains after excluding error-prone homopolymer TRs (**Fig. 19**).

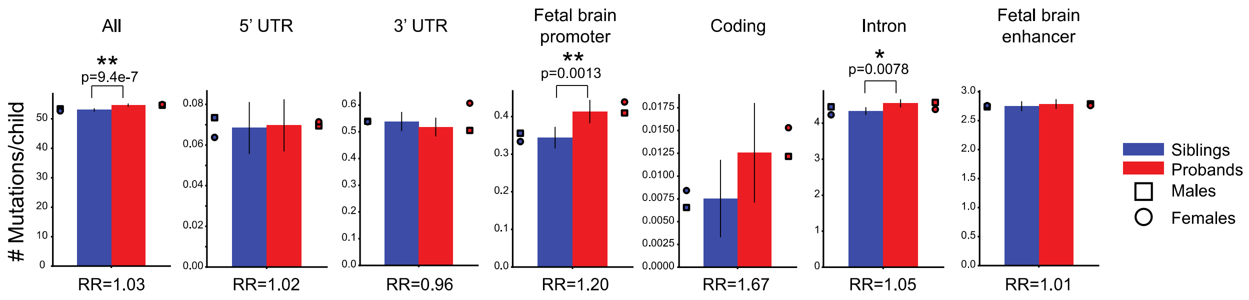


Figure 14: Mean mutation counts by gene annotation.

Bars denote the mean number of mutations in non-ASD siblings (blue) and probands (red). Error bars give 95% confidence intervals. Circles and squares show counts for females and males, respectively. UTR, untranslated region.

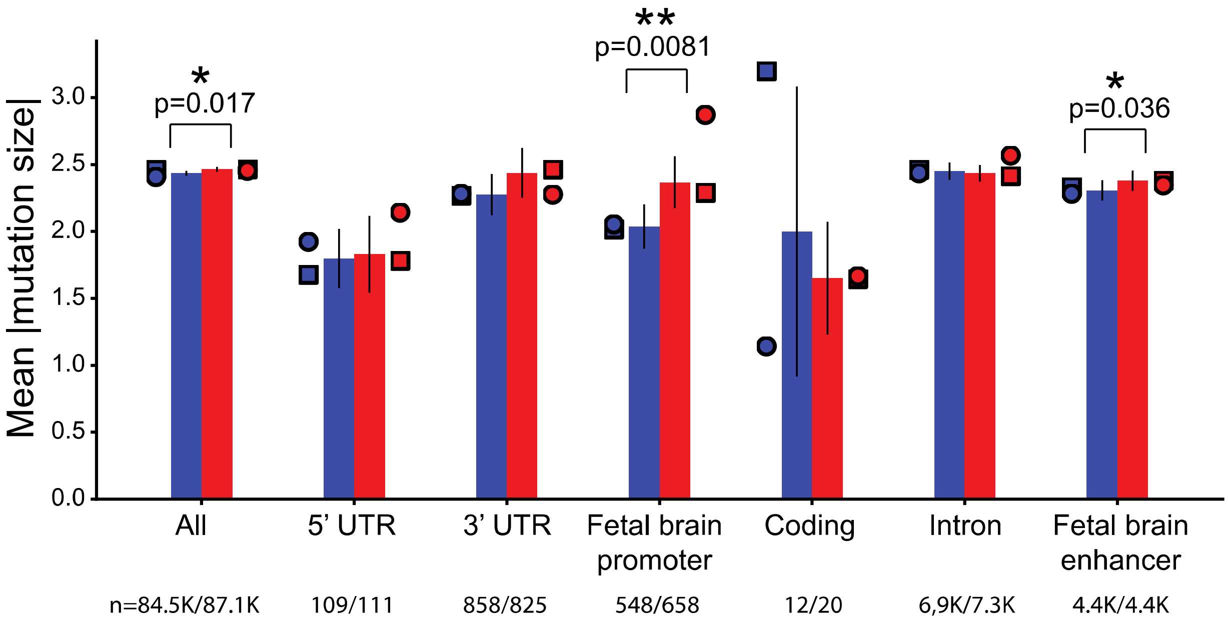


Figure 15: Mean mutation sizes in probands versus non-ASD siblings

Bars show mean mutation size ± 95% CI (in number of repeat units). The number of mutations in each category is annotated in the figure. In a, b, single and double asterisks denote significant increases ($P < 0.05$) before and after Bonferroni correction, respectively.

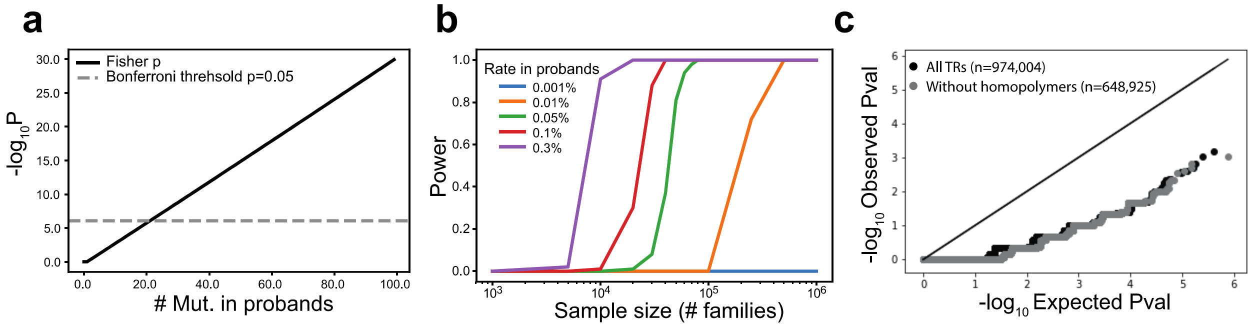
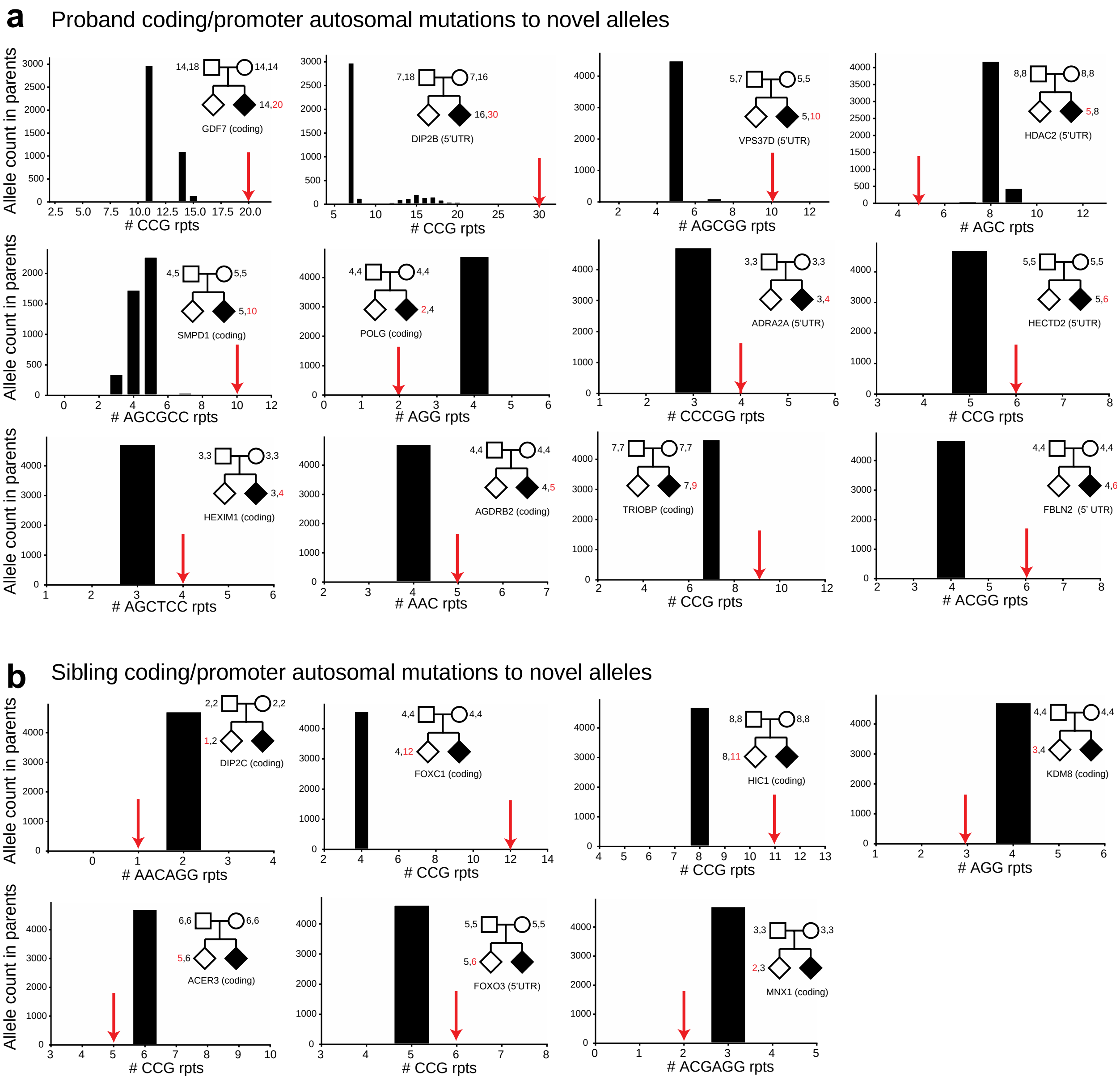


Figure 16: Power to detect per-locus TR mutation enrichments.

a, Number of recurrent mutations required to reach genome-wide significance. We performed a Fisher's exact test to test for an excess of mutations in probands (*n* = 1,593) vs. non-ASD siblings (*n* = 1,593), for a different number of hypothetical mutation counts in probands (*x*-axis) and assuming 0 mutations observed in non-ASD siblings. The black line shows the two-sided *P*-value ($\log_{10}$ scale) obtained for each test. The grey dashed line denotes the *P*-value required to meet a genome-wide significance of *P* < 0.05 with Bonferroni multiple testing correction. b, Sample sizes required to identify genome-wide significant TRs. The *x*-axis shows sample size ($\log_{10}$ scale) in terms of the number of quad families analyzed. Each line represents a different rate of mutation at a particular TR in probands, assuming 0 mutations at that TR in siblings (blue = 0.001%. orange = 0.01%. green = 0.05%. red = 0.1%. purple = 0.3%). The *y*-axis shows the power to detect a specific TR at genome-wide significance for each rate. c, Quantile-Quantile plots for per-locus TR mutation burden testing. For each TR we performed a Fisher's exact test to test for an excess of mutations in probands vs. siblings. The *x*-axis gives expected $-\log_{10}$ *P*-values under a null (uniform) distribution. The *y*-axis gives observed $-\log_{10}$ *P*-values from burden tests. Each dot represents a single TR. Black = all TRs. Gray = homopolymers excluded.



a Proband coding/promoter autosomal mutations to novel alleles

b Sibling coding/promoter autosomal mutations to novel alleles

Figure 17: All coding and 5′UTR mutations to novel alleles.

a, Mutations in probands at coding or 5′UTR TRs to unobserved alleles. Each panel shows a *de novo* TR mutation observed in ASD probands to an allele (x-axis, repeat copy number) not observed in SSC parents. Black histograms give the allele counts in parents. Red arrows denote the allele resulting from each specified *de novo* TR mutation. Pedigrees show genotypes of parents and the child with the mutation (probands = black diamonds. non-ASD siblings = white diamonds). The text below pedigrees gives the gene and region in which the mutation occurred. b, Mutations in non-ASD siblings at coding or 5′UTR TRs to unobserved alleles. Plots are the same as in a. except show mutations in non-ASD siblings.

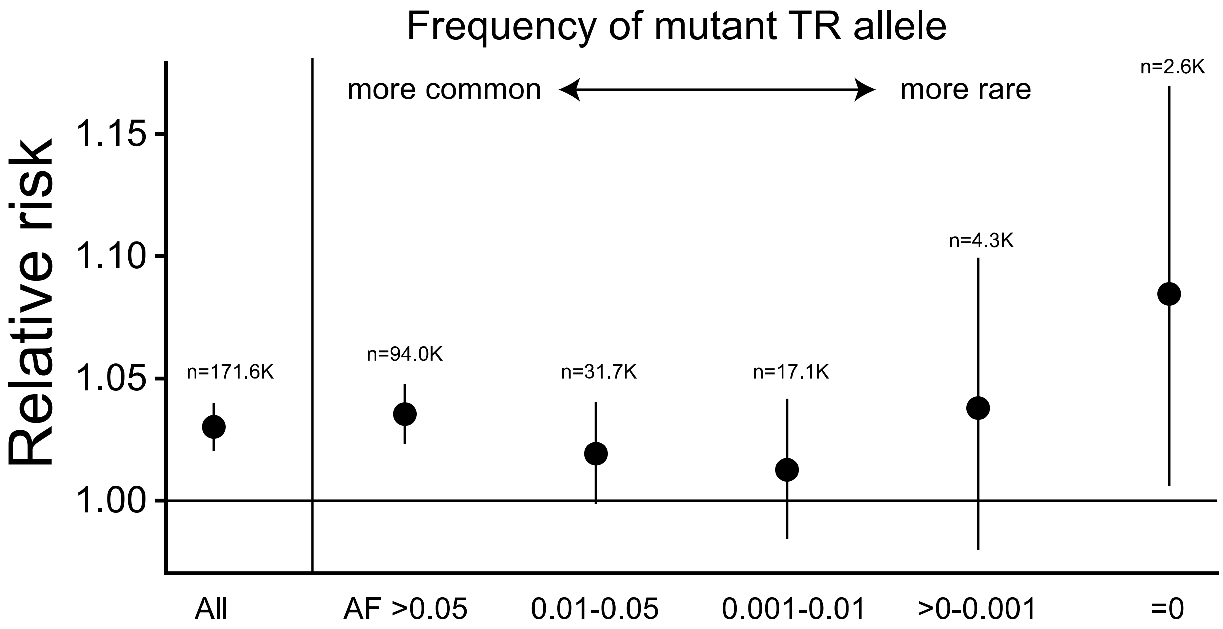


Figure 18: Mutation burden by AF.

The *x*-axis stratifies mutations on the basis of non-overlapping bins of the frequency of the mutant allele in parents in the SSC. Data are mean ± 95% CI. The number of mutations in each category is annotated in the figure. 'All' includes all mutations. For other bins, only TRs for which precise copy numbers could be inferred in at least 80% of SSC parents are included. a, b, d are based on mutations in *n* = 1,593 probands and *n* = 1,593 siblings.

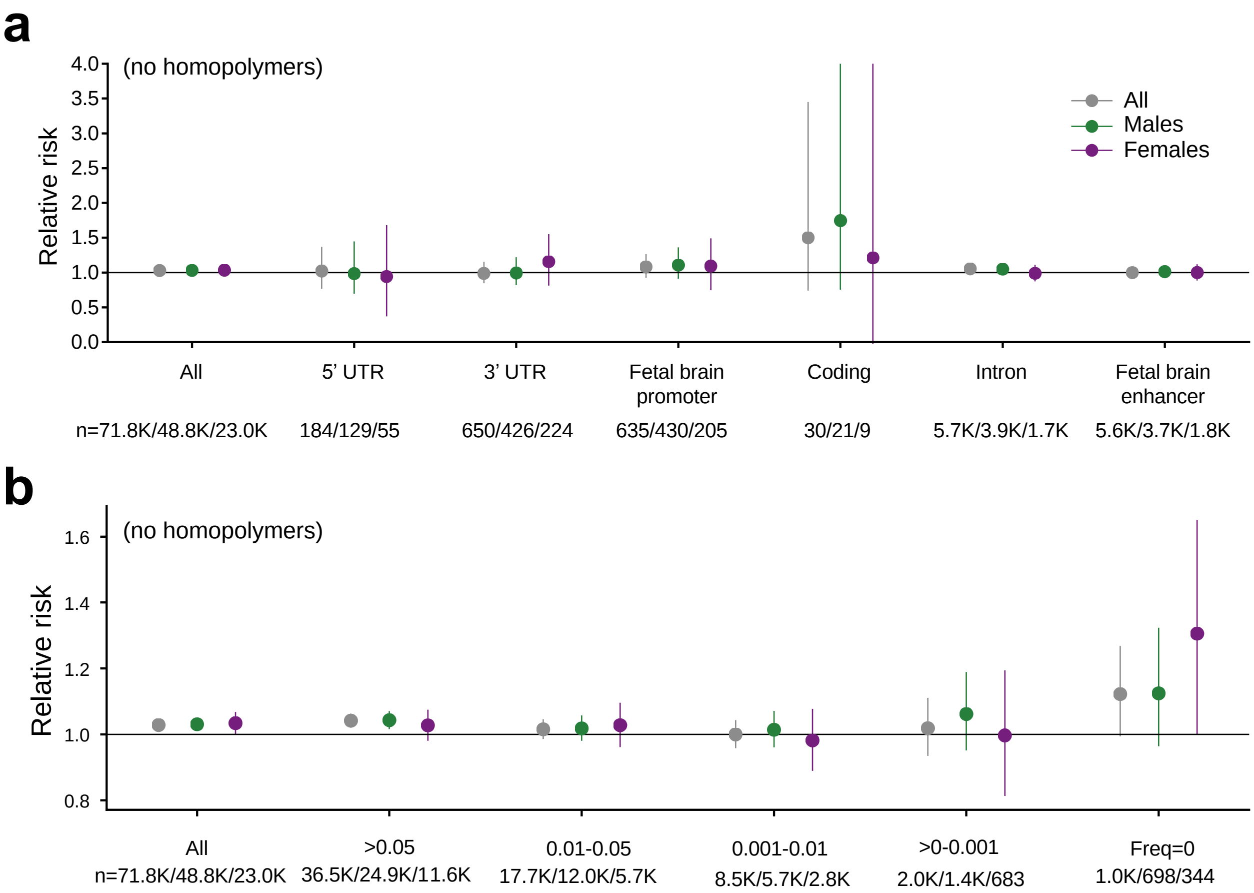


Figure 19:TR mutation burden in ASD excluding homopolymers.

a, Mutation burden by gene annotation. b, Mutation burden by frequency of the allele arising by *de novo* mutation. The *x*-axis stratifies mutations based on non-overlapping bins of the frequency of the *de novo* allele in healthy controls (SSC parents). “All” includes all mutations. For other allele frequency bins, only TRs for which precise copy numbers could be inferred in at least 80% of SSC parents are included. AF = allele frequency. In both plots, the *y*-axis gives RR in probands vs. non-ASD siblings. Dots show estimated relative risk and lines give 95% confidence intervals. Gray = all samples. green = males only. purple = females only. Both plots show only TRs with repeat unit length >1bp.

Table 5: De novo TR mutation burden stratified by SSC datasets.

Unadjusted p-values are based on a two-sided paired t-test comparing counts of each mutation in probands vs. unaffected siblings. Adjusted p-values are based on counts adjusted for paternal age.

Phase	Mean # Mutations - probands	Mean # Mutations - controls	P-value (unadjusted)	P-value (adjusted)
Phase 1	53.4	51.9	0.026	0.069
Phase 2	54.3	53.1	0.027	0.048
Phase 3_1	55.9	54.1	0.00045	0.0011
Phase 3_2	57.8	54.0	0.008	0.008

4.4 Common variant enrichment analyses

We repeated our mutation burden analysis restricting to TRs within 50kb of SNPs previously associated with ASD or related traits through genome-wide association studies (GWAS). We observed no significant mutation excess for TRs within 50kb of ASD genome-wide association study (GWAS) signals, but observe nominally significant increased mutation burden in ASD probands for mutations near GWAS signals for schizophrenia (SCZ) and educational attainment (EA), which are positively genetically correlated with ASD (70) (**Fig. 20a**). The burden is strongest and significant for EA after multiple hypothesis correction (Mann Whitney two-sided *P*=0.0073) when only considering mutations resulting in common alleles (frequency >0.05. **Fig. 20b**), suggesting counter-intuitively that some *de novo* TR mutations may result in ASD risk alleles that are common in the population and may be in linkage disequilibrium with signals identified by SNP-based GWAS. The observation of stronger enrichment for SCZ and EA is consistent with previous analyses of *de novo* point mutations (26) and may be in part due to higher-powered GWAS for those traits compared to ASD.

We additionally performed gene-set enrichment analyses using MAGMA (74) to test whether genes identified by our TR analysis in ASD are enriched in common variants associated with EA, SCZ, or ASD. We used two methods to construct gene sets from TR mutations, most of which are non-coding and cannot be directly assigned to a gene. First, as in **Fig. 21**, we defined a proband gene set consisting of genes with proband TR mutations in coding or promoter regions and no mutations identified in unaffected siblings. Second, we defined a second target list consisting only of genes with predicted severe mutations in probands (**Appendix, Table 10**). For both gene sets, we defined a similar set of control genes as those with only mutations in unaffected siblings. No significant enrichments were observed for control gene sets. We observed nominally significant enrichment of the first set in ASD GWAS (*P*=0.0309) and the second set in SCZ GWAS (*P*=0.0283). However, after correcting for multiple hypothesis testing, none of these enrichments remains significant.

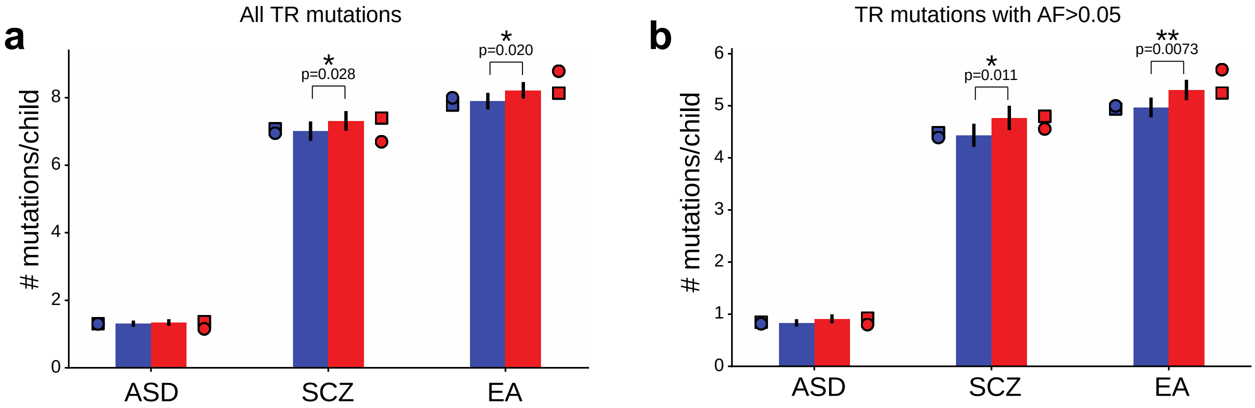


Figure 20: TR mutation burden near SNPs associated with ASD and related traits.

a, b, Bars show mean TR mutation counts in probands (red) vs. non-ASD siblings (blue) for TRs within 50kb of published GWAS associated SNPs (ASD = autism spectrum disorder. SCZ = schizophrenia. EA = educational attainment) considering (a) all TR mutations (ASD *n* = 4,213. SCZ *n* = 22,811. SCZ *n* = 25,668 TR mutations) or (b) mutant allele frequency is >5% in controls (SSC parents) (ASD *n* = 2,774. SCZ *n* = 14,661. SCZ *n* = 16,364 TR mutations). Error bars give 95% confidence intervals around the mean. Single asterisks denote nominally significant increases (Mann–Whitney one-sided *P* < 0.05). Double asterisks denote trends that are significant after Bonferroni correction for the six categories tested. Circles and squares show counts for females and males, respectively.

4.5 Proband mutations predicted to alter gene expression

Based on the observed enrichment in fetal brain promoters and previously demonstrated role of TRs in regulating gene expression (15), we hypothesized that *de novo* TR mutations in probands may act in part by altering gene expression during brain development. We examined expression of genes with coding or promoter mutations using the BrainSpan Atlas of the Developing Human Brain (73) resource. We found that genes with TR mutations only observed in ASD probands (proband gene set) show significantly higher prenatal expression compared to genes with only mutations found in unaffected siblings (control gene set) (Mann-Whitney one-sided P=6.3×10^{-15} at 13 post-conceptional weeks [pcw], meta-analysis across 16 brain structures. **Fig. 21**). Median expression of the proband gene set is higher across all time points with the strongest effects at prenatal periods (**Fig. 22a**) in all brain structures analyzed except cerebellar cortex (CBC) and mediodorsal nucleus of thalamus (MD). We additionally tested whether proband mutations are predicted to alter brain expression of nearby genes based on our previous genome-wide analysis of effects of TR variation on gene expression (15). We found that predicted effects of proband mutations are significantly stronger than for TRs with only mutations in unaffected siblings in Brain-Caudate (**Fig. 22b**. Mann-Whitney two-sided P=0.037), but not for Brain-Cerebellum or the 15 other non-brain tissues analyzed in that study. Proband mutations are predicted to more significantly alter expression of nearby genes in the brain compared to control mutations (**Fig. 22b**).

We identified specific TR mutations in coding or promoter regions resulting in alleles unobserved in unaffected parents. One example such proband mutation shown in **Fig. 17a** is a deletion of 3 copies of CAG in the 5'UTR of *HDAC2*, which results in a previously unobserved allele of 5 copies. In our previous genome-wide analysis of effects of TRs on gene expression (15), we identified a negative association between CAG copy number and expression of *HDAC2*. The *de novo* allele of 5 copies in the proband is thus predicted to increase expression of *HDAC2*, which is highly expressed prenatally in the dorsolateral prefrontal cortex and other brain regions in the BrainSpan database. Notably, a recent study found that *HDAC2* is upregulated in the prefrontal cortex of a Shank3-deficient mouse model of ASD (75) and that down-regulating *HDAC2* rescues social deficits. These results are consistent with the hypothesis that deletion of CAG copies, resulting in increased *HDAC2* expression, could contribute to an ASD phenotype.

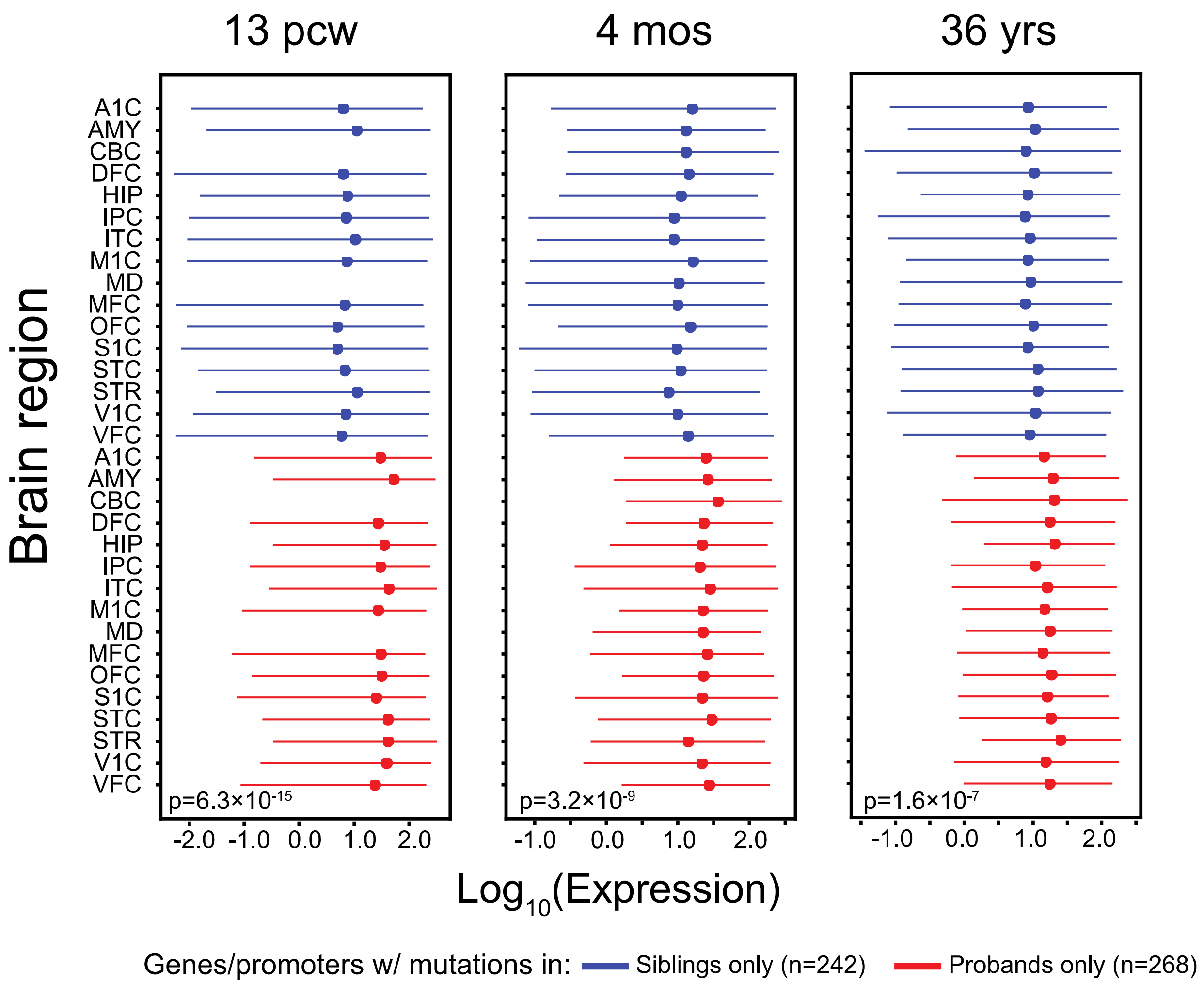


Figure 21: Expression of genes with *de novo* TR mutations in brain.

Red and blue lines show the distribution of expression of genes with only proband (*n* = 268 genes) or sibling mutations (*n* = 242 genes), respectively. Dots give medians and lines extend from the 25th to 75th percentiles of expression across all genes in each set. A1C, primary auditory cortex. AMY, amygdaloid complex. CBC, cerebellar cortex. DFC, dorsolateral prefrontal cortex. HIP, hippocampus. IPC, posteroventral (inferior) parietal cortex. ITC, inferolateral temporal cortex. M1C, primary motor cortex. MD, mediodorsal nucleus of thalamus. MFC, anterior cingulate cortex. OFC, orbital frontal cortex. S1C, primary somatosensory cortex. STC, posterior superior temporal cortex. STR, striatum. V1C = primary visual cortex. VFC = ventrolateral prefrontal cortex.

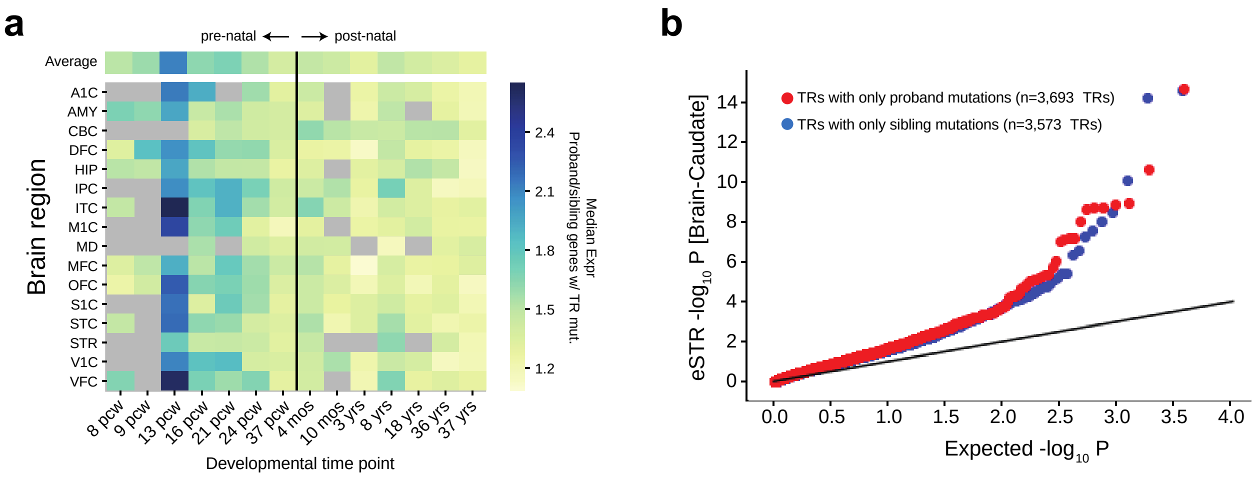


Figure 22: Proband *de novo* TR mutations enriched in brain-expressed genes.

a, Ratio of median expression in proband-only genes to control-only genes across time points. The heatmap shows the ratio of the median expression of genes with only proband mutations (n = 268 genes) to that of genes with only mutations in non-ASD siblings (n = 242 genes). Each row shows a different brain structure from the BrainSpan dataset. Each column shows a different developmental time point. The black vertical line separates pre-natal from post-natal time points. Gray boxes indicate no data was available for that time point. Brain structure acronyms are defined in 4.2.4. b, Proband TR mutations enriched for brain expression STRs. The quantile-quantile plot shows the distribution of expression STR (eSTR) unadjusted P-values based on associating TR length with gene expression in Brain-Caudate samples in the GTEx cohort46. eSTR association P-values are two-sided and are based on t-statistics computed using linear regression analyses performed previously. Each point represents a TR by gene association test using a linear regression model42. The x-axis gives expected -log10 P-values and the y-axis gives observed -log10 P-values. Red points show TRs with at least one *de novo* mutation in probands and 0 in controls. Blue points show TRs with at least one *de novo* mutation in controls and 0 in probands. We found no significant difference in either Brain-Cerebellum or the other 15 non-brain tissues analyzed in that study, which we expected should not be relevant to ASD (not shown).

4.6 Acknowledgements

Chapter 4 is an adapted reprint of the material "Patterns of *de novo* tandem repeat mutations and their role in autism" by Ileena Mitra, Bonnie Huang, Nima Mousavi, Nichole Ma, Michael Lamkin, Richard Yanicky, Sharona Shleizer-Burko, Kirk E. Lohmueller, and Melissa Gymrek published in *Nature* 2021. The dissertation author was the primary investigator and author of this paper.

CHAPTER 5: PRIORITIZING DISEASE RELEVANT TR MUTATIONS IN AUTISM

5.1 Introduction

A major barrier in genomic analyses of TRs is interpreting their consequence for disease risk. Determining likely causal subsets of TR mutations presents a major challenge because traditional SNV-based annotations (e.g. protein-truncating, missense) are not applicable to TRs (76,77). Therefore, the genomics field requires a rigorous statistical framework for identifying likely pathogenic TRs, similar to metrics, such as, gene-level constraint or pLI scores used for SNVs (76,77).

We aimed to create a novel framework to prioritize *de novo* TR mutations most contributing to ASD risk. To assess the severity of TR mutations, we categorized mutations based on allele frequency and predicted deleteriousness. We applied SISTR (Selection Inference on Short Tandem Repeats), a novel method to measure negative selection at short TRs (STRs) (**Section 5.2**). This method is used to predict the pathogenicity of specific TR alleles arising from *de novo* mutations. SISTR currently only supports short TRs with repeat lengths of 2-4bp, as these repeats are abundant and can be genotyped relatively accurately, and our mutation models are most accurate for these loci (35). Importantly, SISTR allowed us to identify an abundance of rare and pathogenic TR mutations in ASD children that may be clinically relevant (**Section 5.3**). By uncovering pathogenic TR mutations contribution to ASD, we add a novel layer of information to the genetic etiology of ASDs, uncover additional disease susceptibility loci, and capture new gene targets for functional and therapeutic research.

5.2 Inferring selection coefficients at TRs using SISTR

We developed SISTR (Selection Inference at Short TRs), a population genetics framework for inferring selection coefficients at individual TR loci. SISTR fits an evolutionary model of TR variation that includes mutation, genetic drift, and negative natural selection to available empirical allele frequencies to infer the posterior distribution of selection coefficients. Our mutation model is based on a modified version of the generalized stepwise mutation model (GSM) (78). To model negative selection, we assume the central allele at each TR has optimal fitness (*w*=1), and that the fitness of other alleles is based on their difference in size from the optimal allele.

SISTR applies approximate Bayesian computation (ABC) based on a previously described forward simulation technique (79) to infer per-locus selection coefficients by fitting allele frequencies for one TR at a time given a predefined optimal allele length and fixed set of mutation parameters. Our method outputs the median posterior estimate of *s* and computes a likelihood ratio test comparing the likelihood of the inferred *s* value to the likelihood of *s*=0.

For each TR with a repeat unit length of 2-4bp, we used SISTR to estimate selection coefficients based on allele frequencies in SSC parents. We set the optimal allele length at each TR to the modal allele and used mutation parameters described in the above as input. We excluded TRs with repeat lengths in hg38 <11 units for dinucleotides, <5 units for trinucleotides, and <7 repeats for tetranucleotides, since those repeats are typically not polymorphic. We included only TRs for which precise copy numbers could be inferred in at least 80% of SSC parents. We further filtered TRs at which the 95% confidence interval on our estimate for s was greater than 0.3, indicating we could not estimate *s* precisely. After filtering, 62,941 STRs remained for analysis.

We used the Benjamini-Hochberg procedure (80) to adjust *P*-values for multiple hypothesis testing. To identify TRs under significant selection, we chose TRs with adjusted *P*-value <0.01, corresponding to a false discovery rate of 1%. Allele-specific selection coefficients, which can be interpreted as pathogenicity scores, were computed as $(|a - opt|)s$, where a is the number of repeat copies for the *de novo* allele, opt is the optimum (modal) repeat and s is the selection coefficient for the TR inferred using SISTR.

5.3 Prioritizing pathogenic TR mutations

We sought to further prioritize TR mutations based on their predicted deleterious effects. Metrics commonly used to annotate SNV mutations (76,81,82) are not applicable to TRs, which tend to be multi-allelic and result in either non-coding mutations or in-frame indels. To overcome this challenge, we developed a novel population genetics framework, Selection Inference at Short TRs (SISTR) to measure negative selection against individual TR alleles. SISTR fits an evolutionary model of TR variation that includes mutation, genetic drift, and negative natural selection to empirical allele frequency data (per-locus frequencies of each allele length) to infer the posterior distribution of selection coefficients (s) at individual TRs (**Fig. 23**). SISTR is agnostic to gene annotations and analyzes both coding and non-coding TRs. Parameter s can be interpreted as the decrease in reproductive fitness impact for each repeat unit copy number away from the population modal allele at a given TR. Testing our method on simulated datasets capturing a range of mutation and selection models, SISTR accurately recovers simulated values down to s=10^{-4}, corresponding to strong or moderate selection, for most settings (**Fig. 24a-b. Fig. 25**).

We applied SISTR to estimate selection coefficients at genome-wide TRs based on allele frequencies observed in unaffected SSC parents. Notably, SISTR currently only handles TRs with repeat unit lengths 2-4bp. Of those, SISTR could not fit models at 4.4% of TRs, potentially indicating inaccurate model assumptions for those loci. After filtering TRs where s could not be reliably inferred 62,941 TRs remained for analysis. We found that the overall distribution of selection coefficients is robust to input choices including demographic model and prior distribution on *s* (**Fig. 24c**). As expected, TRs with significant predicted selection coefficients have significantly stronger MUTEA (35) constraint scores (Mann-Whitney one-sided $p<10^{-200}$. **Fig. 25b**). Further, protein-coding TRs under strongest negative selection tend to be in genes less tolerant of missense mutations (83) (Mann-Whitney one-sided P=0.00028. **Fig. 24d**), or loss of function SNV mutations (82) (Mann-Whitney one-sided P=0.00067. **Fig. 24e**), compared to coding STRs not inferred to be under negative selection (s=0).

We next tested for an enrichment of evolutionarily deleterious TRs in probands compared to non-ASD siblings. When restricting to TR loci predicted to be under selection (s>0 with false discovery rate [FDR]<1%), we find an increased mutational burden in probands (**Fig. 25c**), which is most notable for mutations resulting in rare mutant alleles. Stratifying mutations based on allele-specific selection coefficients results in a further increased mutational burden (**Fig. 25d**). *De novo* TR mutations with rare or unobserved allele frequencies and estimated to be the most deleterious (top 1% of *s* scores) show the strongest relative risk (RR=1.34 [95% CI 1.05-1.73. one-sided P=0.010] for rare [AF<0.01] alleles and RR=2.50 [95% CI 1.30-6.35. one-sided P=0.0056] for unobserved low fitness alleles, compared to RR=1.03 [95% CI 1.02-1.04. one-sided $P=4.7\times10^{-7}$] genome-wide). We identified 35 mutations, of which 25 are in probands, resulting in previously unobserved alleles predicted to be strongly deleterious (top 1% of *s* scores). Of these, multiple proband mutations are in genes with point mutations previously implicated in ASD (*e.g., PDCD1, KCNB1, AGO1, CACNA2D3, FOXP1, RFX3, MED13L)* or related phenotypes, whereas only two rare mutations are found in siblings to be related to ASD genes (**Appendix, Table 10**). Overall, these results suggest that the subset of TR mutations resulting in rare alleles under strongest selection are most pathogenic for ASD risk.

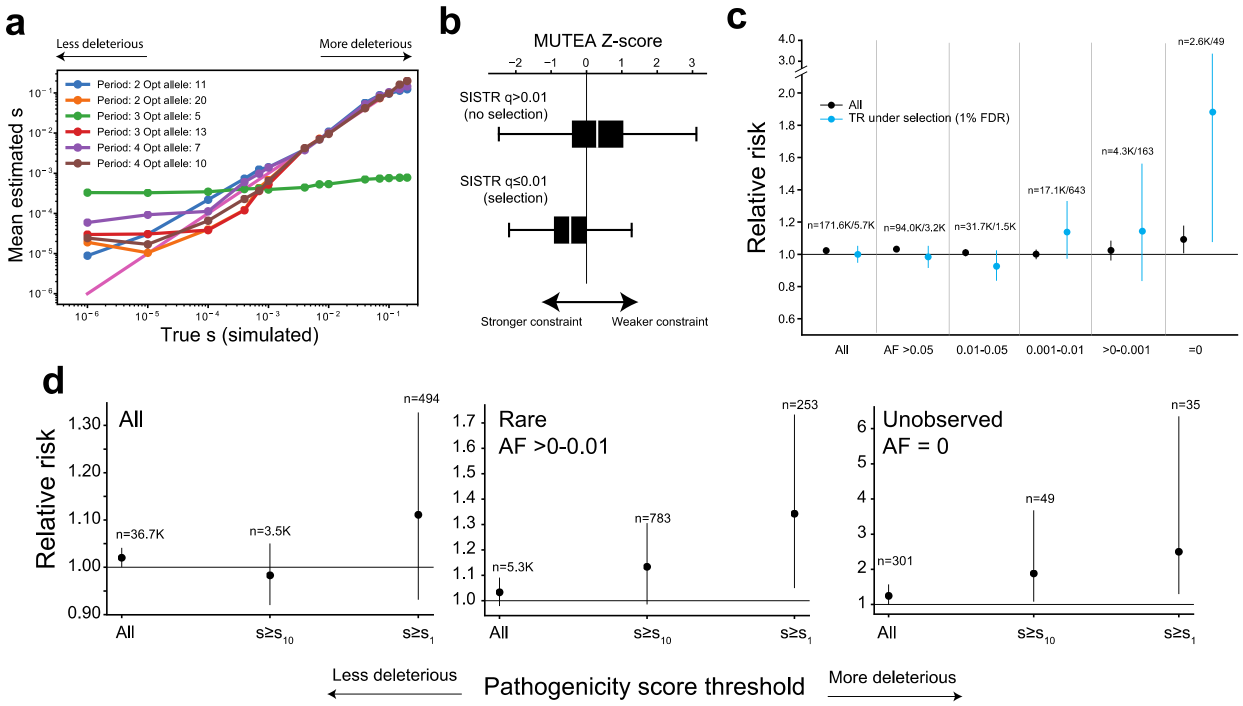


Figure 23: Prioritizing TR mutations by fitness effects.

a, Comparison of true versus inferred per-locus selection coefficients. The x-axis shows the true simulated value of s, and the y-axis shows the mean s value inferred by SISTR across 200 simulation replicates. Each color denotes a separate mutation model based on the repeat unit length (period) and optimal allele. b, Comparison of SISTR and MUTEA (35) . Boxes show the distribution of MUTEA constraint scores for TRs inferred to have non-significant (top. n = 43,672 TRs) or significant (bottom. n = 6,251 TRs) selection coefficients (FDR <1%). White middle lines show the medians and boxes span from the 25th percentile (Q1) to the 75th percentile (Q3). Whiskers extend to Q1 − 1.5× interquartile range (IQR) (minima) and Q3 + 1.5× IQR (maxima). c, Mutation burden at TR loci under negative selection. The x-axis stratifies mutations on the basis of the same allele frequency categories as in **Fig. 18**. Blue dots show relative risk considering only TRs inferred to be under the strongest negative selection (FDR <1%). Data are mean ± 95% CI. d, Per-allele selection coefficients stratify mutation burden within allele frequency bins. Larger s values denote a mutation resulting in an allele predicted to be more deleterious. s10 and s1 correspond to the top 10% and top 1% of pathogenicity scores, respectively. Data are mean ± 95% CI.

5.4 Acknowledgements

Chapter 5 is an adapted reprint of the material “Patterns of *de novo* tandem repeat mutations and their role in autism” by Ileena Mitra, Bonnie Huang, Nima Mousavi, Nichole Ma, Michael Lamkin, Richard Yanicky, Sharona Shleizer-Burko, Kirk E. Lohmueller, and Melissa Gymrek published in *Nature* 2021. The dissertation author was the primary investigator and author of this paper.

CHAPTER 6: CONCLUSION

6.1 Conclusion

We present a framework for the identification and prioritization of *de novo* TR mutations. We find on average 54 autosomal TR mutations per individual. The true number of mutations is probably underestimated owing to the stringent filtering applied to candidate mutations. Overall, our results identify novel patterns of TR mutation and suggest that the burden of *de novo* TR mutations is similar in magnitude to the total number of *de novo* point mutations per child (20,84).

We find a significant genome-wide excess of *de novo* TR mutations in probands compared with non-ASD siblings. On the basis of this excess, we estimate that these mutations contribute to approximately 1.6% of simplex idiopathic ASD probands. A recent study analyzing an orthogonal set of variants estimated that larger complex TR expansions contribute to 2.6% of simplex cases (33) .Taken together, these results suggest TRs may account for around 4% of simplex ASD cases, comparable in magnitude to non-coding point mutations (27).

Notably, only a subset of *de novo* TR mutations is likely to contribute to ASD risk or have deleterious effects. We find that mutations resulting in mutant alleles that are very rare (AF <0.001) or estimated to be under strong negative selection show the greatest signal of excess mutations in probands. The relative risk observed for these most severe mutations (RR = 2.50), which are all non-coding, is similar in magnitude to previously reported relative risks for protein-truncating variants (3). We estimate the overall contribution to simplex ASD to be highest for mutations resulting in common alleles (of the 1.6% estimated above, 1.1% is attributed to mutations with AF >0.05). The impact of these mutations is not obvious and is the subject of future study.

Our study faced several limitations: (1) identification of TR mutations remains challenging and requires stringent filtering to achieve high validation rates. (2) our results exclude important TR mutation classes, such as sequence interruptions (85), somatic variation (25), and complex repeat expansions which have been recently studied elsewhere (33). and (3) we do not currently have power to implicate specific TRs at genome-wide significance (**Fig. 16**). Future methods improvements and increasing sample sizes are likely to pinpoint the specific TR mutations most relevant to ASD. The framework developed in our study will serve as a valuable resource for further characterizing TR mutations and their role in ASD and other diseases.

6.2 Acknowledgements

Chapter 6 is an adapted reprint of the material "Patterns of *de novo* tandem repeat mutations and their role in autism" by Ileena Mitra, Bonnie Huang, Nima Mousavi, Nichole Ma, Michael Lamkin, Richard Yanicky, Sharona Shleizer-Burko, Kirk E. Lohmueller, and Melissa Gymrek published in *Nature* 2021. The dissertation author was the primary investigator and author of this paper.

I.M. performed TR genotyping, identification of *de novo* mutations, downstream analyses, and helped write the manuscript. B.H. developed SISTR and performed analysis of TR selection scores in the SSC cohort. Nima M. helped design GangSTR analysis and filtering settings, and analyses to evaluate MonSTR. Nichole M. performed capillary electrophoresis validation experiments. M.L. performed TR annotation for identification of determinants of TR mutation rates. R.Y. designed AWS cloud analysis pipelines. S.S.-B. helped design and set up validation experiments. K.E.L. conceived the SISTR method, supervised analysis of TR selection scores, and drafted the manuscript. M.G. conceived the study, designed and performed analyses, and drafted the manuscript. All authors have read and approved the final manuscript.

This study was supported by the Simons Foundation Autism Research Initiative (SFARI Grant #630705). I.M. was additionally supported by a predoctoral fellowship from the Autism Science

Foundation. M.G. was additionally supported in part by the Office of The Director, National Institutes of Health under Award Number DP5OD024577 and NIH/NHGRI grants R01HG010149 and R21HG010070. K.E.L. was supported by the National Institutes of Health grant R35GM119856. We thank J. Gleeson, J. Sebat, A. Palmer and A. Goren for helpful comments on this study.

REFERENCES

1. Rosti RO, Sadek AA, Vaux KK, Gleeson JG. The genetic landscape of autism spectrum disorders. Dev Med Child Neurol. 2014. 56(1):12–8.

2. Iakoucheva LM, Muotri AR, Sebat J. Getting to the Cores of Autism. Cell. 2019. 178(6):1287–98.

3. Iossifov I, O'Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, et al. The contribution of de novo coding mutations to autism spectrum disorder. Nature. 2014 Nov 13. 515(7526).

4. Acuna-Hidalgo R, Veltman JA, Hoischen A. New insights into the generation and role of de novo mutations in health and disease. Genome Biol. 2016. 17(1):241.

5. Gymrek M. A genomic view of short tandem repeats. Curr Opin Genet Dev. 2017. 44:9–16.

6. Sun JX, Helgason A, Masson G, Ebenesersdóttir SSSS, Li H, Mallick S, et al. A direct characterization of human mutation based on microsatellites. Nat Genet. 2012 Oct. 44(10):1161–5.

7. Hannan AJ. Tandem repeats mediating genetic plasticity in health and disease. Nat Rev Genet. 2018 Feb 5. 19(5):286–98.

8. Song J, Lowe CB, Kingsley DM. Characterization of a human-specific tandem repeat associated with bipolar disorder and schizophrenia. bioRxiv. 2018. 311795.

9. Rubinsztein DC. Lessons from animal models of Huntington's disease. Trends Genet. 2002 Apr 1. 18(4):202–9.

10. Marco EJ, Skuse DH. Autism-lessons from the X chromosome. Soc Cogn Affect Neurosci. 2006 Dec 1. 1(3):183–93.

11. Subramanian S, Mishra RK, Singh L. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. Genome Biol. 2003 Jan 23. 4(2):R13.

12. Pelley JW. Organization, Synthesis, and Repair of DNA. In: Elsevier's Integrated Biochemistry. Elsevier. 2007. p. 123–33.

13. Dieringer D, Schlötterer C. Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. Genome Res. 2003 Oct [cited 2018 May 20]. 13(10):2242–51. Available from: http://www.ncbi.nlm.nih.gov/pubmed/14525926

14. Bakhtiari M, Shleizer-Burko S, Gymrek M, Bansal V, Bafna V. Targeted genotyping of variable number tandem repeats with adVNTR. Genome Res. 2018 Nov 1. 28(11):1709–19.

15. Fotsing SF, Margoliash J, Wang C, Saini S, Yanicky R, Shleizer-Burko S, et al. The impact of short tandem repeat variation on gene expression. Nat Genet. 2019 Nov 1. 51(11):1652–9.

16. Niu M, Han Y, Dy ABC, Du J, Jin H, Qin J, et al. Autism Symptoms in Fragile X Syndrome. J Child Neurol. 2017 Sep 15. 32(10):903–9.

17. Cristina A, Luiza A, Pereira-Nascimento P. Genetic Etiology of Autism. In: Recent Advances in Autism Spectrum Disorders - Volume I. InTech. 2013.

18. Sutcliffe JS, Nelson DL, Zhang F, Pieretti M, Caskey CT, Saxe D, et al. DNA methylation represses *FMR-1* transcription in fragile X syndrome. Hum Mol Genet. 1992. 1(6):397–400. 7

19. He X, Sanders SJ, Liu L, De Rubeis S, Lim ET, Sutcliffe JS, et al. Integrated Model of De Novo and Inherited Genetic Variants Yields Greater Power to Identify Risk Genes. PLoS Genet. 2013 Aug. 9(8).

20. Turner TN, Coe BP, Dickel DE, Hoekzema K, Nelson BJ, Zody MC, et al. Genomic Patterns of De Novo Mutation in Simplex Autism. Cell. 2017 Oct 19. 171(3):710-722.e12.

21. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, et al. Strong association of de novo copy number mutations with autism. Science. 2007 Apr 20. 316(5823):445–9.

22. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature. 2012 Apr 4. 485(7397):242–5.

23. Sanders SJ, He X, Willsey AJ, Devlin B, Roeder K, State MW, et al. Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. Neuron. 2015. 87:1215–33.

24. Veltman JA, Brunner HG. De novo mutations in human genetic disease. Nat Rev Genet. 2012. 13(8):565–75.

25. Breuss MW, Antaki D, George RD, Kleiber M, James KN, Ball LL, et al. Autism risk in offspring can be assessed through quantification of male sperm mosaicism. Nat Med. 2019.

26. Kyle Satterstrom F, Kosmicki JA, Wang J, Breen MS, Rubeis D, An J-Y, et al. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. Cell. 2020. 25(10):18.

27. Zhou J, Park CY, Theesfeld CL, Wong AK, Yuan Y, Scheckel C, et al. Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. Nat Genet. 2019 Jun 1. 51(6):973–80.

28. Ramu A, Noordam MJ, Schwartz RS, Wuster A, Hurles ME, Cartwright RA, et al. DeNovoGear: De novo indel and point mutation discovery and phasing. Nat Methods. 2013 Oct 25. 10(10):985–7.

29. Mousavi N, Shleizer-Burko S, Yanicky R, Gymrek M. Profiling the genome-wide landscape of tandem repeat expansions. Nucleic Acids Res. 2019 Sep 5. 47(15):e90–e90.

30. Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y. Genome-wide profiling of heritable and de novo STR variations. Nat Methods. 2017 Jun. 14(6):590–2.

31. Dolzhenko E, van Vugt JJFA, Shaw RJ, Bekritsky MA, Van Blitterswijk M, Narzisi G, et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. Genome Res.

2017.

32. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007 Sep. 81(3):559–75.

33. Trost B, Engchuan W, Nguyen CM, Thiruvahindrapuram B, Dolzhenko E, Backstrom I, et al. Genome-wide detection of tandem DNA repeats that are expanded in autism. Nature. 2020.

34. Dolzhenko E, Bennett MF, Richmond PA, Trost B, Chen S, Van Vugt JJFA, et al. ExpansionHunter Denovo: A computational method for locating known and novel repeat expansions in short-read sequencing data. Genome Biol. 2020.

35. Gymrek MA, Willems T, Reich D, Erlich Y. Interpreting short tandem repeat variations in humans using mutational constraint. Nat Genet. 2016 Sep 11 [cited 2018 May 20]. 49(September):8–10.

36. Fischbach GD, Lord C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. Neuron. 2010 Oct 21 [cited 2015 Aug 17]. 68(2):192–5.

37. Mousavi N, Margoliash J, Pusarla N, Saini S, Yanicky R, Gymrek M. TRTools: a toolkit for genome-wide analysis of tandem repeats. Bioinformatics. 2020.

38. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser at UCSC. Genome Res. 2002.

39. An JY, Lin K, Zhu L, Werling DM, Dong S, Brand H, et al. Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. Science (80-). 2018 Dec 14. 362(6420).

40. Schuelke M. An economic method for the fluorescent labeling of PCR fragments. Nat Biotechnol. 2000.

41. Krebs MO, Betancur C, Leroy S, Bourdel MC, Gillberg C, Leboyer M, et al. Absence of association between a polymorphic GGC repeat in the 5’ untranslated region of the reelin gene and autism. Mol Psychiatry. 2002.

42. Zhang H, Liu X, Zhang C, Mundo E, Macciardi F, Grayson DR, et al. Reelin gene alleles and susceptibility to autism spectrum disorders. Mol Psychiatry. 2002.

43. Lammert DB, Howell BW. RELN mutations in autism spectrum disorder. Frontiers in Cellular Neuroscience. 2016.

44. Huang W, Li L, Myers JR, Marth GT. ART: A next-generation sequencing read simulator. Bioinformatics. 2012.

45. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013 Mar 16 [cited 2017 Jun 13]. Available from: http://arxiv.org/abs/1303.3997

46. Payseur BA, Jing P, Haasl RJ. A genomic portrait of human microsatellite variation. Mol Biol

Evol. 2011.

47. Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, et al. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. Cell. 2012. 151(7):1431–42.

48. O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, et al. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature. 2012 Apr 4 [cited 2018 May 20]. 485(7397):246–50. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22495309

49. Rahbari R, Wuster A, Lindsay SJ, Hardwick RJ, Alexandrov LB, Al Turki S, et al. Timing, rates and spectra of human germline mutation. Nat Genet. 2016.

50. Quinlan AR. BEDTools: The Swiss-Army tool for genome feature analysis. Curr Protoc Bioinforma. 2014.

51. Bherer C, Campbell CL, Auton A. Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. Nat Commun. 2017.

52. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010.

53. Ozsolak F, Song JS, Liu XS, Fisher DE. High-throughput mapping of the chromatin structure of human promoters. Nat Biotechnol. 2007.

54. Dennis JH, Fan HY, Reynolds SM, Yuan G, Meldrim JC, Richter DJ, et al. Independent and complementary methods for large-scale structural analysis of mammalian chromatin. Genome Res. 2007.

55. Gupta S, Dennis J, Thurman RE, Kingston R, Stamatoyannopoulos JA, Noble WS. Predicting human nucleosome occupancy from primary sequence. PLoS Comput Biol. 2008.

56. Consortium EP, Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis C a, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2013.

57. Ellegren H. Heterogeneous mutation processes in human microsatellite DNA sequences. Nat Genet. 2000.

58. Huang QY, Xu FH, Shen H, Deng HY, Liu YJ, Liu YZ, et al. Mutation patterns at dinucleotide microsatellite loci in humans. Am J Hum Genet. 2002.

59. Weber JL, Wong C. Mutation of human short tandem repeats. Hum Mol Genet. 1993.

60. Ellegren H. Microsatellites: simple sequences with complex evolution. Nat Rev Genet. 2004 Jun. (6):435–45.

61. Leclercq SB, Rivals E, Jarne P. DNA slippage occurs at microsatellite loci without minimal threshold length in humans: A comparative genomic approach. Genome Biol Evol. 2010.

62. Jakupciak JP, Wells RD. Genetic instabilities of triplet repeat sequences by recombination. IUBMB Life. 2000.

63. Carroll J, Marangos P. The DNA damage response in mammalian oocytes. Frontiers in Genetics. 2013.

64. Broman KW, Murray JC, Sheffield VC, White RL, Weber JL. Comprehensive human genetic maps: Individual and sex-specific variation in recombination. Am J Hum Genet. 1998.

65. Eriksson A, Amos W, Kosanovic D, Kosanović D, Eriksson A. Inter-allelic interactions play a major role in microsatellite evolution. Proc R Soc B Biol Sci. 2015.

66. Payseur BA, Nachman MW. Microsatellite variation and recombination rate in the human genome. Genetics. 2000.

67. Iossifov I, Ronemus M, Levy D, Wang Z, Hakker I, Rosenbaum J, et al. De novo gene disruptions in children on the autistic spectrum. Neuron. 2012. 74(2):285–99.

68. Ernst J, Kellis M. ChromHMM: Automating chromatin-state discovery and characterization. Nature Methods. 2012.

69. Buniello A, Macarthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019.

70. Grove J, Ripke S, Als TD, Mattheisen M, Walters RK, Won H, et al. Identification of common genetic risk variants for autism spectrum disorder. Nat Genet. 2019.

71. Ripke S, Neale BM, Corvin A, Walters JTR, Farh K-H, Holmans PA, et al. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014 Jul 22. 511(7510):421–7.

72. Lee JJ, Wedow R, Okbay A, Kong E, Maghzian O, Zacher M, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nat Genet. 2018.

73. Miller JA, Ding SL, Sunkin SM, Smith KA, Ng L, Szafer A, et al. Transcriptional landscape of the prenatal human brain. Nature. 2014.

74. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. PLoS Comput Biol. 2015.

75. Qin L, Ma K, Wang ZJ, Hu Z, Matas E, Wei J, et al. Social deficits in Shank3-deficient mouse models of autism are rescued by histone deacetylase (HDAC) inhibition. Nat Neurosci. 2018.

76. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: Predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019.

77. Fuller ZL, Berg JJ, Mostafavi H, Sella G, Przeworski M. Measuring intolerance to mutation in human genetics. Nat Genet. 2019 Apr 8.

78. Fu YX, Chakraborty R. Simultaneous estimation of all the parameters of a stepwise mutation model. Genetics. 1998.

79. Haasl RJ, Payseur BA. Microsatellites as targets of natural selection. Mol Biol Evol. 2013.

80. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B. 1995.

81. Davydov E V., Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. 2010.

82. Lek M, Karczewski KJ, Minikel E V., Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016.

83. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, et al. A framework for the interpretation of de novo mutation in human disease. Nat Genet. 2014 Sep.

84. Werling DM, Brand H, An J-YY, Stone MR, Zhu L, Glessner JT, et al. An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. Nat Genet. 2018 May 26 [cited 2018 Oct 22]. 50(5):727–36.

85. Grünewald TGP, Bernard V, Gilardi-Hebenstreit P, Raynal V, Surdez D, Aynaud MM, et al. Chimeric EWSR1-FLI1 regulates the Ewing sarcoma susceptibility gene EGR2 via a GGAA microsatellite. Nat Genet. 2015.

APPENDIX

Abbreviations:

ASD, Autism Spectrum Disorders

bp, base pair

CE, capillary electrophoresis

CNV, copy number variant

indels, small insertions and deletions

STR, short tandem repeats

SNV, single nucleotide variants

TR, tandem repeats

VNTR, variable number tandem repeats

Table 6: Primers used for capillary electrophoresis validation experiments.

The table lists primers for each TR mutation validated. Each forward primer had an M13(-21) universal adapter sequence appended, shown in blue. The CGG TR at chr7:103989357 could not be amplified using the three primer method and was genotyped separately using published primers from Krebs et al. 2007.

Chromosome	TR start position (hg38)	Repeat unit	Forward primer	Reverse primer
chr1	106950468	GT	TGTAAAACGACGGCCAGTCATGCACTCTGGCAACCTAA	TCGTGTAGACGGTAGGCACA
chr1	217937145	AAAT	TGTAAAACGACGGCCAGTGCTGAGCTCTCCTTTGCTTC	TGTCAGGAAACAATGCCAAA
chr1	242233155	AATG	TGTAAAACGACGGCCAGTCAGTCTGGGTGACAGAGCAA	AAAACCCTGGGTCTCACCTT
chr1	6733191	CA	TGTAAAACGACGGCCAGTGGTGAGCTATGATTGCACCA	CAGGCCTCTGAAGCAGAAAG
chr1	76108862	GT	TGTAAAACGACGGCCAGTAATCCCTCGATCGAAACAAA	TAAGGCACCCCAAAGGAAAC
chr10	57337770	CA	TGTAAAACGACGGCCAGTCAGGATCCCTGAACTCAAGC	ATGACAGTTGCCATTGCTGT
chr10	72427734	AT	TGTAAAACGACGGCCAGTCAAGACCAGCTTCAGCATCA	GGCATGGTTAGGACCTCAAA
chr11	2442377	AC	TGTAAAACGACGGCCAGTTGGTTCCAAAGGAATTTAGCA	GCTTCAGCTGTGCTGTGGTA
chr11	36246191	AATA	TGTAAAACGACGGCCAGTCTGGGCAACAGAGTGAGATG	TCTCTTACCAGAGGGGCTGA
chr11	63798227	AC	TGTAAAACGACGGCCAGTCACACCTGGCACTGTCTCAT	GGGAATAGCGGAGGAAGGTA
chr11	85908552	TAGA	TGTAAAACGACGGCCAGTCTGGGCAACACAGCATAGAC	CGGGGTTTCATCATGTTGGT
chr12	4096182	TTCC	TGTAAAACGACGGCCAGTCACACAAATGACCCCAACTG	GAGATGAGATCGCGTCACTG
chr12	75962280	T	TGTAAAACGACGGCCAGTTACAACCATTGTGCCTGGAA	TGGGAGGCTGAGGTAGAGAA
chr12	131901040	T	TGTAAAACGACGGCCAGTTGTAACTCCCCATCCCAGAG	CCCAGTCTCATCCCATTGTT
chr12	70011632	AG	TGTAAAACGACGGCCAGTTGAGGTGGTGGTTACAGCAG	CCATGCAGAGACTCTTGCTC
chr12	92269453	T	TGTAAAACGACGGCCAGTCCAGGCTGGAATACAGTGGTA	TACTTTGGGAGGTCGAGGTG
chr13	102648430	GT	TGTAAAACGACGGCCAGTCCAGTTAACAGCCACTGCAC	CATGGGTCCCTCAGAGACAT
chr13	75404064	AC	TGTAAAACGACGGCCAGTCACAATGCTAGAGAAAGTTCAAGG	TTCTTACTGCGCCATCTTTTT
chr13	81527591	CTAT	TGTAAAACGACGGCCAGTTTGAACAGCAAGTGAACCTTT	TTTTTCTGCTATTTTTGGTATTTTCA
chr14	41606868	GCC	TGTAAAACGACGGCCAGTCTTTGGGAAGCCCAGCTC	ACACGCGCACACACATACAT
chr15	53480481	AC	TGTAAAACGACGGCCAGTGCATTTCTTTTCATTGCATTTT	CACCCACACATTCATTCCAC
chr15	72443969	T	TGTAAAACGACGGCCAGTACATTCTGGCCTCGTACCTG	AGTGAGGCCCCATCTCTTTT
chr15	80839290	GT	TGTAAAACGACGGCCAGTGGAGTGAAGGCTGTGGAGTC	CTCCCTCAGAAGCTGGTGTC
chr16	62573638	AATA	TGTAAAACGACGGCCAGTCCCAGGAGTTTGAGGCTACA	TGATGATCTACAGCCCTTTGC
chr16	58599051	A	TGTAAAACGACGGCCAGTGGCAGGAGAACTGCTTGAA	CAGCAGGAAATACAGCATGTAAG
chr16	67644335	T	TGTAAAACGACGGCCAGTACCTTAGCCCCAGGCTTC	GAGACCCTGTCTCTGCAAAA

chr17	8128309	TAGA	TGTAAAACGACGGCCAGTGGCCAACAGACCCAGACTC	CCGCACGTTAAGCAAATACA
chr18	38020886	TCTA	TGTAAAACGACGGCCAGTTTATGGCAGGAAGAGGTTGG	TTGGTGATAGAAACAAATAGACGA
chr18	38020886	TCTA	TGTAAAACGACGGCCAGTTTATGGCAGGAAGAGGTTGG	TGTCTTCTTGGATTATTTAGGATCTTT
chr2	1273636	TA	TGTAAAACGACGGCCAGTATGCTGGGATAATTGGATGC	TTCATGGTTTGTGCTTCTGG
chr20	18095896	AC	TGTAAAACGACGGCCAGTTGAGAGGACAACTGGGAGGA	AGGACAAAAGCAACCTGGAA
chr20	42508128	TG	TGTAAAACGACGGCCAGTTCCCAGGCCTGAAATAACAA	CAGGCGCTCCTAGAAACAAA
chr3	54501407	T	TGTAAAACGACGGCCAGTTCCTCCCTCCGGTTTCTTAT	TTGAGGCTGCAGTGAGTCAT
chr3	46654755	TG	TGTAAAACGACGGCCAGTTGTCAAAACCCATAGAATGAACA	AGATGCCCCACTGCACTC
chr4	123573491	TTAT	TGTAAAACGACGGCCAGTTCCCAGGTTTAAAGCCACTG	AAACGGCAAAGACAAATTGC
chr4	136965932	AT	TGTAAAACGACGGCCAGTGGCAGGTCTTTCTTGAGCTG	AGCCTGGCTTTTTACTGGTAG
chr4	176019003	T	TGTAAAACGACGGCCAGTCCCTGGCCACACTTACCTTA	CATGTGCCTGTAATCCCAGA
chr4	46950717	AAAT	TGTAAAACGACGGCCAGTTTCCACCTCTTTAAAAGCCATT	GCTTGAAGCTTCTTGCTTCC
chr5	18768193	GT	TGTAAAACGACGGCCAGTTCAGGAGCTTTGTTGAAGGTG	TGGAACAAAGCAGAGAATCC
chr6	12322154	AC	TGTAAAACGACGGCCAGTGGTGGAAGTAATGGTTTCTTGCT	TGTCCCCTGGAAAGAAAAATC
chr6	50207587	TCTA	TGTAAAACGACGGCCAGTCCCAAACCTTGGATCCTTTT	TGGGTGGGTGGAGAGATAGA
chr6	84874972	AC	TGTAAAACGACGGCCAGTGTCCCAATGCCTCTACTGGA	CCGGGGTGTTGTTCATATTC
chr6	89959098	CA	TGTAAAACGACGGCCAGTGAAGCTGGCCCTGTCAATAA	GTGGGCTGACCATGTTTTTC
chr7	18349308	AC	TGTAAAACGACGGCCAGTAAGGCTTTTGCATTTGTTGG	AAATAAGCCAGCAAGGAGGA
chr7	27264534	AC	TGTAAAACGACGGCCAGTCCCAGCTACTTGGGAAACTG	CCATGCAATAGCTTGGGTTT
chr7	103989357	CCG	FAM-CGCCTTCTTCTCGCCTTCTC	CGAAAAGCGGGGGTAATAGC
chr8	50595021	TG	TGTAAAACGACGGCCAGTCCCAACCCCTCTCTTTTCTC	CATTCCCCAAAAATAAAGACCA
chr9	36061394	AC	TGTAAAACGACGGCCAGTTGCTTGTACCCAGCATCCTT	TCCAGTGGCCTCTTAGAACA
chr9	6685998	TTA	TGTAAAACGACGGCCAGTCCCAGGTACAAGCGATTCTG	GGGTGACAGAGCAAGAACCT

Table 7: Comparison of GangSTR vs. capillary genotypes at candidate TR mutations.

The table lists 49 TR candidate mutations in 5 families. Columns include: family identifier, family member (fa=father, mo=mother, s1=sibling, p1=proband), TR chromosome, TR start position (hg 38), genotype called by GangSTR, genotype called by capillary electrophoresis, genotype validated between both methods (Y=yes, N=no), mutation validated between both methods (Y=yes, N=no, NA=low sequencing quality). The row for corresponding to the child with the inferred mutation (proband or sibling) at each locus is in bold. All genotypes are given in terms of the number of copies of the repeat unit.

Family	Member	Chrom	TR start position (hg38)	GangSTR Genotype (A1)	GangSTR Genotype (A2)	Capillary Genotype (A1)	Capillary Genotype (A2)	Genotype Validated	Mutation Validated
1	fa	chr6	50207587	12	12	12	12	Y	Y
1	mo	chr6	50207587	11	12	11	12	Y	
1	p1	chr6	50207587	12	12	12	12	Y	
1	s1	chr6	50207587	12	13	12	13	Y	
1	fa	chr1	6733191	16	16	16	16	Y	N
1	mo	chr1	6733191	16	16	16	25	N	
1	p1	chr1	6733191	16	16	16	16	Y	
1	s1	chr1	6733191	16	25	16	25	Y	
1	fa	chr1	76108862	17	17	17	17	Y	Y
1	mo	chr1	76108862	15	15	15	15	Y	
1	p1	chr1	76108862	15	17	15	17	Y	
1	s1	chr1	76108862	14	17	14	17	Y	
1	fa	chr10	72427734	16	23	16	23	Y	Y
1	mo	chr10	72427734	8	26	8	26	Y	
1	p1	chr10	72427734	8	17	8	17	Y	
1	s1	chr10	72427734	8	23	8	23	Y	
1	fa	chr11	85908552	16	16	16	16	Y	Y
1	mo	chr11	85908552	13	15	13	15	Y	
1	p1	chr11	85908552	13	16	13	16	Y	
1	s1	chr11	85908552	13	15	13	15	Y	
1	fa	chr12	70011632	15	23	15	23	Y	Y
1	mo	chr12	70011632	12	12	12	12	Y	
1	p1	chr12	70011632	12	15	12	15	Y	
1	s1	chr12	70011632	12	24	12	24	Y	
1	fa	chr12	92269453	14	27	14	14	N	N
1	mo	chr12	92269453	14	14	14	14	Y	
1	p1	chr12	92269453	14	28	14	14	N	
1	s1	chr12	92269453	14	27	14	14	N	
1	fa	chr13	81527591	11	13	11	13	Y	Y
1	mo	chr13	81527591	10	12	10	12	Y	
1	p1	chr13	81527591	12	14	12	14	Y	
1	s1	chr13	81527591	10	13	10	13	Y	
1	fa	chr15	53480481	18	18	18	18	Y	Y
1	mo	chr15	53480481	21	24	21	24	Y	
1	p1	chr15	53480481	18	20	18	20	Y	
1	s1	chr15	53480481	18	21	18	21	Y	
1	fa	chr15	72443969	14	14	14	14	Y	Y
1	mo	chr15	72443969	14	14	14	14	Y	
1	p1	chr15	72443969	14	15	14	15	Y	
1	s1	chr15	72443969	14	14	14	14	Y	
1	fa	chr16	58599051	19	20	19	20	Y	Y
1	mo	chr16	58599051	20	20	20	21	N	
1	p1	chr16	58599051	19	23	19	23	Y	
1	s1	chr16	58599051	19	20	19	20	Y	
1	fa	chr16	67644335	18	18	18	18	Y	Y
1	mo	chr16	67644335	18	18	18	18	Y	
1	p1	chr16	67644335	18	18	18	18	Y	
1	s1	chr16	67644335	18	19	18	19	Y	

1	fa	chr17	8128309	14	16	14	16	Y	Y
1	mo	chr17	8128309	13	16	13	16	Y	
1	p1	chr17	8128309	13	13	13	13	Y	
1	s1	chr17	8128309	14	16	14	16	Y	
1	fa	chr18	38020886	9	12	9	12	Y	Y
1	mo	chr18	38020886	10	10	10	10	Y	
1	p1	chr18	38020886	10	12	10	12	Y	
1	s1	chr18	38020886	11	12	11	12	Y	
1	fa	chr3	46654755	10	11	10	11	Y	Y
1	mo	chr3	46654755	10	11	10	11	Y	
1	p1	chr3	46654755	11	12	11	12	Y	
1	s1	chr3	46654755	11	11	11	11	Y	
1	fa	chr8	50595021	16	16	16	16	Y	Y
1	mo	chr8	50595021	16	18	16	18	Y	
1	p1	chr8	50595021	16	16	16	16	Y	
1	s1	chr8	50595021	15	16	15	16	Y	
1	fa	chr9	36061394	22	26	22	26	Y	Y
1	mo	chr9	36061394	22	29	22	29	Y	
1	p1	chr9	36061394	26	37	26	29	N	
1	s1	chr9	36061394	27	29	27	29	Y	
1	fa	chr9	6685998	15	18	15	18	Y	Y
1	mo	chr9	6685998	9	15	9	15	Y	
1	p1	chr9	6685998	9	17	9	17	Y	
1	s1	chr9	6685998	15	18	15	18	Y	
1	mo	chr4	46950717	8	14	8	14	Y	Y
1	p1	chr4	46950717	8	15	8	15	Y	
1	s1	chr4	46950717	8	8	8	8	Y	
1	fa	chr4	46950717	8	13	8	13	Y	
1	fa	chr6	84874972	24	25	24	25	Y	Y
1	mo	chr6	84874972	18	18	18	18	Y	
1	p1	chr6	84874972	18	25	18	25	Y	
1	s1	chr6	84874972	18	22	18	22	Y	
1	fa	chr6	89959098	20	28	20	28	Y	Y
1	mo	chr6	89959098	24	25	24	25	Y	
1	p1	chr6	89959098	25	26	25	26	Y	
1	s1	chr6	89959098	20	24	20	24	Y	
2	fa	chr10	57337770	22	26	22	26	Y	Y
2	mo	chr10	57337770	22	23	22	23	Y	
2	s1	chr10	57337770	23	26	23	26	Y	
2	p1	chr10	57337770	22	24	22	24	Y	
2	fa	chr11	63798227	16	24	16	24	Y	Y
2	mo	chr11	63798227	16	16	16	16	Y	
2	s1	chr11	63798227	16	23	16	23	Y	
2	p1	chr11	63798227	16	24	16	24	Y	
2	fa	chr12	131901040	15	15	15	15	Y	Y
2	mo	chr12	131901040	15	15	15	15	Y	
2	s1	chr12	131901040	15	15	15	15	Y	
2	p1	chr12	131901040	15	16	15	16	Y	
2	fa	chr13	75404064	12	12	12	12	Y	N
2	mo	chr13	75404064	12	19	12	19	Y	
2	s1	chr13	75404064	10	12	12	12	N	
2	p1	chr13	75404064	12	19	12	19	Y	
3	fa	chr14	41606868	7	8	7	8	Y	Y
3	mo	chr14	41606868	7	8	7	8	Y	
3	s1	chr14	41606868	7	8	7	8	Y	
3	p1	chr14	41606868	8	10	8	10	Y	
2	fa	chr15	80839290	23	23	23	28	N	N
2	mo	chr15	80839290	23	29	23	29	Y	

2	s1	chr15	80839290	23	28	23	28	Y	
2	p1	chr15	80839290	23	30	23	29	N	
2	fa	chr7	103989357	10	10	10	10	Y	Y
2	mo	chr7	103989357	8	10	8	10	Y	
2	s1	chr7	103989357	10	10	10	10	Y	
2	p1	chr7	103989357	8	12	8	12	Y	
2	fa	chr1	242233155	9	9	9	9	Y	Y
2	mo	chr1	242233155	9	11	9	11	Y	
2	s1	chr1	242233155	9	10	9	10	Y	
2	p1	chr1	242233155	9	9	NA	NA	NA	
4	fa	chr11	36246191	9	12	9	12	Y	Y
4	mo	chr11	36246191	12	13	12	13	Y	
4	s1	chr11	36246191	9	12	9	12	Y	
4	p1	chr11	36246191	11	13	11	13	Y	
4	fa	chr16	62573638	11	14	11	14	Y	Y
4	mo	chr16	62573638	10	11	10	11	Y	
4	s1	chr16	62573638	11	11	11	11	Y	
4	p1	chr16	62573638	11	13	11	13	Y	
5	fa	chr1	106950468	22	26	22	26	Y	Y
5	mo	chr1	106950468	19	19	19	19	Y	
5	p1	chr1	106950468	19	25	19	25	Y	
5	s1	chr1	106950468	19	22	19	22	Y	
5	fa	chr1	217937145	7	11	7	11	Y	Y
5	mo	chr1	217937145	7	11	7	11	Y	
5	p1	chr1	217937145	11	13	11	13	Y	
5	s1	chr1	217937145	11	11	11	11	Y	
5	fa	chr2	1273636	8	8	8	8	Y	Y
5	mo	chr2	1273636	8	8	8	8	Y	
5	p1	chr2	1273636	8	8	8	8	Y	
5	s1	chr2	1273636	6	8	6	8	Y	
5	fa	chr3	54501407	14	16	14	16	Y	Y
5	mo	chr3	54501407	14	16	14	16	Y	
5	p1	chr3	54501407	16	17	16	17	Y	
5	s1	chr3	54501407	14	16	14	16	Y	
5	fa	chr4	123573491	3	3	3	3	Y	Y
5	mo	chr4	123573491	3	3	3	3	Y	
5	p1	chr4	123573491	2	3	2	3	Y	
5	s1	chr4	123573491	3	3	3	3	Y	
5	fa	chr4	136965932	12	13	12	13	Y	N
5	mo	chr4	136965932	7	7	7	15	N	

5	p1	chr4	136965932	7	12	7	12	Y	
5	s1	chr4	136965932	12	15	12	15	Y	
5	fa	chr4	176019003	19	23	19	23	Y	Y
5	mo	chr4	176019003	19	23	19	23	Y	
5	p1	chr4	176019003	22	23	22	23	Y	
5	s1	chr4	176019003	19	23	19	23	Y	
5	fa	chr5	18768193	19	24	19	24	Y	Y
5	mo	chr5	18768193	19	24	19	24	Y	
5	p1	chr5	18768193	24	24	24	24	Y	
5	s1	chr5	18768193	22	24	22	24	Y	
5	fa	chr6	12322154	16	17	16	17	Y	Y
5	mo	chr6	12322154	13	15	13	15	Y	
5	p1	chr6	12322154	11	17	11	17	Y	
5	s1	chr6	12322154	13	17	13	17	Y	
5	fa	chr7	18349308	25	26	25	26	Y	Y
5	mo	chr7	18349308	20	25	20	25	Y	
5	p1	chr7	18349308	25	25	25	25	Y	
5	s1	chr7	18349308	20	28	20	28	Y	
5	fa	chr7	27264534	23	28	23	28	Y	Y
5	mo	chr7	27264534	22	23	22	23	Y	
5	p1	chr7	27264534	23	25	23	25	Y	
5	s1	chr7	27264534	22	23	22	23	Y	
5	fa	chr11	2442377	19	21	19	21	Y	Y
5	mo	chr11	2442377	20	23	20	23	Y	
5	p1	chr11	2442377	20	21	20	21	Y	
5	s1	chr11	2442377	18	20	18	20	Y	
5	fa	chr12	4096182	13	17	13	17	Y	Y
5	mo	chr12	4096182	10	13	10	13	Y	
5	p1	chr12	4096182	13	14	13	14	Y	
5	s1	chr12	4096182	10	13	10	13	Y	
5	fa	chr12	75962280	20	23	20	23	Y	Y
5	mo	chr12	75962280	21	21	21	21	Y	
5	p1	chr12	75962280	21	24	21	24	Y	
5	s1	chr12	75962280	21	23	21	23	Y	
5	fa	chr13	102648430	12	17	12	17	Y	Y
5	mo	chr13	102648430	12	19	12	19	Y	
5	p1	chr13	102648430	12	12	12	12	Y	
5	s1	chr13	102648430	16	19	16	19	Y	
5	fa	chr18	38020886	6	10	6	10	Y	Y
5	mo	chr18	38020886	10	11	10	11	Y	
5	p1	chr18	38020886	6	12	6	12	Y	
5	s1	chr18	38020886	10	11	10	11	Y	
5	fa	chr20	18095896	22	25	22	25	Y	Y
5	mo	chr20	18095896	22	28	22	28	Y	
5	p1	chr20	18095896	21	28	21	28	Y	
5	s1	chr20	18095896	22	22	22	22	Y	
5	fa	chr20	42508128	11	11	11	11	Y	Y
5	mo	chr20	42508128	11	11	11	11	Y	
5	p1	chr20	42508128	11	12	11	12	Y	
5	s1	chr20	42508128	11	11	11	11	Y	

Table 8: All *de novo* TR Mutations in coding regions.

The table lists 33 *de novo* TR mutations in protein-coding regions. Columns include: Phenotype (1=unaffected sibling, 2=ASD proband), TR chromosome, TR start position (hg 38 reference), mutation unit size (number of repeats), repeat motif, frequency of *de novo* allele in SSC parents, and gene name.

Phenotype (2=proband, 1=control)	Chromosome	Position (hg38)	Mutation size (# units)	Repeat motif	Frequency of new allele	Gene
2	1	31756261	1	AAC	0	ADGRB2
2	1	50419102	-2	CCG	0.172459	DMRTA2
2	1	154869724	2	AGC	0.002472	KCNN3
2	2	20667363	2	CCG	0	GDF7
2	2	199819527	1	A	0.112661	FTCDNL1
2	3	40462030	1	AGC	0.0136	RPL14
2	3	48927631	-1	AAG	0.003322	ARIH2
2	3	49312475	1	A	0.000644	USP4
2	6	108561445	-2	CCG	0.011272	FOXO3
2	7	91265145	1	CCG	0.406848	FZD1
2	9	12775888	3	AGC	0.001498	LURAP1L
2	11	6390707	5	AGCGCC	0	SMPD1
2	11	62727008	-1	CCG	0.00043	HNRNPUL2
2	12	6667905	-2	AGC	0.001076	ZNF384
2	12	102958394	1	AGC	0.01428	ASCL1
2	15	89326710	-2	AGG	0	POLG
2	17	45150159	1	AGCTCC	0	HEXIM1
2	17	67959646	1	AGCCCCTCC	0.174449	BPTF
2	19	40512815	1	AGCGGGCGC	0.006607	SPTBN4
2	22	37746313	2	CCG	0	TRIOBP
2	X	136502889	-1	AAG	0	HTATSF1
1	1	154869724	-1	AGC	0.111236	KCNN3
1	1	154869724	1	AGC	0.11573	KCNN3
1	3	63912686	-2	AGC	0.000665	ATXN7
1	6	1611317	8	CCG	0	FOXC1
1	7	157005635	-1	ACGAGG	0	MNX1
1	9	97854419	2	CCG	0.308129	FOXE1
1	10	356496	-1	AACAGG	0	DIP2C
1	11	47767112	-2	ACC	0.430267	FNBP4
1	12	118068523	1	AGG	0.024395	VSIG10
1	16	27219017	-1	AGG	0	KDM8
1	17	2057111	3	CCG	0	HIC1
1	17	67959646	1	AGCCCCTCC	0.174449	BPTF

Table 9: All *de novo* repeat expansions.

The table lists 78 *de novo* TR expansion mutations. Columns include: Phenotype (1=control, 2=ASD proband), TR chromosome, TR start position (hg 38 reference), mutation unit size (number of repeats), repeat motif, gene name, and known associated phenotypes.

Phenotype (2=proband, 1=control)	Chromosome	Position (hg38)	Mutation size (# units)	Repeat motif	Gene	Gene Association
2	1	108180617	13	AG	SLC25A24 (intron)	Fontaine progeroid syndrome (Ehmke et al. 2017)
2	1	156817048	12	AG	SH2D2A (promoter)	No known phenotype associations.
2	1	161851142	5	AC	ATF6 (intron)	Type 2 Diabetes (Meex et al. 2007). Achromatopsia (Kohl et al. 2015)
2	1	164716686	5	AC	PBX1 (intron)	ASD (De Rubeis S , et al. 2014). Developmental disorders (DDD Study. 2017). Developmental delay and ASD (Coe et al. 2018)
2	2	30304937	12	AG	LBH (intron)	
2	2	34119626	5	AC	intergenic	
2	2	45432004	6	AAAT	intergenic	
2	3	31216580	8	AG	intergenic	
2	3	133370872	10	AC	TMEM108 (intron)	No known phenotype associations.
2	4	78629320	5	AG	intergenic	
2	5	59131598	6	AC	PDE4D (intron)	Developmental delay and ASD (Coe et al. 2018). Stroke (Gretarsdottir et al. 2002). Acrodysostosis 2 (Michot et al. 2012)
2	5	123962238	5	AC	intergenic	
2	5	152863986	18	AG	intergenic	
2	6	16687658	7	AC	ATXN1 (intron)	Repeat expansions in ATXN1 are associated with Spinocerebellar Ataxia 1 and Hereditary Ataxia (Orr et al. 1993. Banfi et al. 1994. Zuhlke et al. 2002).
2	6	120460511	5	AAAAT	intergenic	
2	6	143427097	10	AC	ADAT2 (3' UTR)	No known phenotype associations.
2	7	39005485	18	AC	POU6F2 (intron)	Wilms tumor (Perotti et al. 2004)
2	7	54111194	13	AG	intergenic	
2	7	92688582	6	AC	CDK6 (intron)	Primary microcephaly (Hussain et al. 2013).

						cancer (Shennan et al. 2000)
2	7	100634178	7	AC	TFR2 (intron)	Hereditary hemochromatosis (Feder et al. 1996)
2	7	131868327	10	AC	intergenic	
2	7	150755610	13	AG	intergenic	
2	8	8429599	12	AC	intergenic	
2	9	27573529	5	CCCCGG	C9orf72 (intron)	Repeat expansions in C9orf72 are associated with amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD) (Balendra et al. 2018).
2	11	36270138	7	AG	COMMD9 (downstream)	No known phenotype associations.
2	11	126091248	6	AAAAG	intergenic	
2	12	23726119	6	AG	SOX5 (intron)	Intragenic SOX5 deletions associated with ASD and intellectual disability (Rosenfeld et al. 2010. Lamb et al. 2012). Developmental delay and ASD (Coe et al. 2018).
2	12	50505002	12	CCG	DIP2B (5' UTR)	5' UTR CGG repeat expansions are associated with Mental retardation (Winnepenninckx et al. 2007). Developmental delay and ASD (Coe et al. 2018).
2	14	97155210	6	AC	intergenic	
2	16	20691752	6	AC	ACSM1 (intron)	No known phenotype associations.
2	16	50672497	5	AC	SNX20 (3' UTR)	No known phenotype associations.
2	16	86777643	15	AC	intergenic	
2	17	7885308	8	CCG	CHD3 (intron)	ASD (Iossifov et al. 2014). Developmental delay and ASD (Coe et al. 2018).
2	17	51831668	11	AGC	CA10 (intron)	No known phenotype associations.
2	18	591973	15	AG	intergenic	
2	18	591973	9	AG	intergenic	
2	19	8694854	13	AG	intergenic	
2	20	63015968	6	AGAGGCAGGG	intergenic	
2	X	96775354	10	TG	DIAPH2 (intron)	Mutations in DIAPH2 are associated with premature ovarian failure 2 (Philippe et

						al. 1995. Sala et al. 1997, Bione et al. 1998).
1	1	28247135	6	AAAG	intergenic	
1	2	30304937	11	AG	intergenic	
1	2	44497710	20	AG	intergenic	
1	2	57120223	6	AC	intergenic	
1	2	70561093	5	AAGG	intergenic	
1	2	233086120	5	AAAG	intergenic	
1	3	5788918	5	AAATGCACAGGAAT	intergenic	
1	3	183712192	5	AAAAT	intergenic	
1	4	152264129	11	AG	intergenic	
1	5	23931476	13	AG	intergenic	
1	5	77551981	8	AAAAT	intergenic	
1	5	178644795	14	AC	intergenic	
1	6	41155376	12	AG	intergenic	
1	7	1023625	9	AC	C7orf50 (intron)	No known phenotype associations.
1	8	104730454	13	AG	intergenic	
1	9	103006018	10	AC	CYLC2 (3' UTR)	Loss-of-function CYLC2 deletions associated with ASD (Gonzalez-Mantilla et al., 2016, Levy et al., 2011, Stobbe et al., 2013).
1	9	123740900	15	AG	intergenic	
1	10	107285591	6	AATGG	intergenic	
1	10	128010546	6	AAAAG	intergenic	
1	11	110884225	8	AG	intergenic	
1	12	2420103	13	AC	intergenic	
1	12	31108534	5	AAAT	intergenic	
1	12	89164804	15	AG	intergenic	
1	12	115909728	5	AAAAC	intergenic	
1	13	112811162	5	AC	ATP11A (intron)	Developmental delay and ASD (Coe et al. 2018).
1	14	83715680	8	AG	intergenic	
1	15	22963495	7	AACAT	intergenic	
1	17	14300434	7	AC	intergenic	
1	17	39183338	7	AAG	intergenic	
1	17	48677964	9	AG	intergenic	
1	18	30625693	7	AG	intergenic	
1	19	8156492	6	AAAAT	intergenic	
1	19	32721680	8	AAAAT	intergenic	
1	21	37563882	6	ACGCGG	intergenic	

1	21	40453041	9	AC	intergenic	
1	X	22249209	7	GT	PHEX (exon)	Mutations in PHEX associated with X-linked hypophosphatemic rickets (Filisetti et al. 1999. Gaucher et al. 2009, etc.)
1	X	32199781	8	TG	DMD (intron)	Developmental delay and ASD (Coe et al. 2018). Duchenne muscular dystrophy (DMD), Becker muscular dystrophy (BMD), and cardiomyopathy (Kunkel 1986. Yoshida et al. 1998. Daoud et al. 2009, etc.).
1	X	32948114	9	AC	DMD (intron)	Developmental delay and ASD (Coe et al. 2018). Duchenne muscular dystrophy (DMD), Becker muscular dystrophy (BMD), and cardiomyopathy (Kunkel 1986. Yoshida et al. 1998. Daoud et al. 2009, etc.).
1	X	135306289	5	AC	ZNF75D (intron)	No known phenotype associations.

Table 10: Top candidate pathogenic *de novo* TR mutations.

The table lists *de novo* TR mutations resulting in previously unobserved alleles with most severe pathogenicity score (top 1% of pathogenicity scores). Columns include: Phenotype (1=unaffected sibling, 2=ASD proband), TR chromosome, TR start position (hg 38 reference), mutation size (number of repeats), repeat motif, pathogenicity score, gene, known associated phenotypes, and ASD SFARI score.

Phenotype (2=proband, 1=control)	Chromosome	Position (hg38)	Mutation size (# units)	Repeat motif	Pathogenicity Score	Gene	Gene Association	SFARI Gene (https://gene.sfari.org/database/human-gene)
2	14	31071890	3	AAAT	0.14475	AP4S1 (intron)	Spastic paraplegia 52 (Abou Jamra et al. 2011)	
2	2	241850724	-3	AGC	0.08979	PDCD1 (3' UTR)	Developmental delay and ASD (Coe et al. 2018). Multiple sclerosis (Kroner et al. 2005). Systemic lupus erythematosus (Prokunina et al. 2002)	Score=3
2	5	123374214	-4	AAC	0.07336	CEP120 (intron)	Developmental delay and ASD (Coe et al. 2018). Joubert syndrome (Roosing et al. 2016). Short-rib thoracic dysplasia (Shaheen et al. 2015)	
2	12	116290124	-1	AAAT	0.07209	MED13L (upstream)	ASD (Satterstrom et al. 2020). Developmental delay and ASD (Coe et al. 2018). Mental retardation	Score=1

							(Asadollahi et al. 2013)	
2	5	50954696	5	AATT	0.06805	intergenic		
2	20	49348524	-4	AC	0.03784	KCNB1 (downstream)	Developmental delay and ASD (Coe et al. 2018). Epileptic encephalopathy (Torkamani et al. 2014)	Score=1
2	3	25897920	1	AC	0.03528	intergenic		
2	1	35866936	-7	AC	0.02716	AGO1 (upstream)	Developmental delay and ASD (Coe et al. 2018).	Score=2
2	10	124884378	-4	AAAT	0.02492	intergenic		
2	3	54249663	-11	AC	0.02266	CACNA2D3 (intron)	ASD (Satterstrom et al. 2020). Developmental delay and ASD (Coe et al. 2018).	Score=2
2	7	68421891	4	AAAT	0.02208	intergenic		
2	9	3441309	-3	AG	0.02175	RFX3 (intron)	ASD (Satterstrom et al. 2020). Developmental delay and ASD (Coe et al. 2018).	Score=1
2	13	45831180	-5	AAAT	0.02085	SIAH3 (intron)	No known phenotype associations.	
2	8	90210309	-5	AAAT	0.0195	LINC00534 (upstream)	No known phenotype associations.	
2	9	4121621	-10	AC	0.01932	GLIS3 (intron)	Developmental delay and ASD (Coe et al. 2018). Neonatal diabetes (Taha et al. 2003)	
2	3	70813032	3	AATG	0.019	FOXP1 (downstream)	ASD (Satterstrom et al. 2020).	Score=1

							Developmental delay and ASD (Coe et al. 2018). Mental retardation with language impairment and with or without autistic features (Hamdan et al. 2010).
2	8	94226294	2	AC	0.01862	CDH17 (upstream)	No known phenotype associations.
2	2	55870644	-3	AAC	0.0177	EFEMP1 (intron near splice site)	Doyne honeycomb degeneration of retina (Stone et al. 1999)
2	12	67113369	-10	AC	0.0174	intergenic	
2	5	55067784	-4	ATCC	0.01715	intergenic	
2	2	151878551	-9	AC	0.0162	CACNB4 (intron)	Epilepsy (Escayg et al. 2000)
2	2	112876339	-1	AAAT	0.01592	IL37 (upstream)	No known phenotype associations.
2	7	141807438	-5	AAT	0.01578	intergenic	
2	3	48888571	2	AAAT	0.01448	SLC25A20 (intron)	Carnitine-acylcarnitine translocase deficiency (Huizing et al. 1997)
2	13	97232852	2	AC	0.01422	MBNL2 (intron)	No known phenotype associations.
1	4	30716894	1	AGC	0.05552	PCDH7 (upstream)	Developmental delay and ASD (Coe et al. 2018).
1	11	48001350	5	AC	0.02979	PTPRJ (intron)	Colon cancer (Ruivenkamp et al. 2002)
1	13	43770687	-1	AC	0.0261	ENOX1 (intron)	Developmental delay

							and ASD (Coe et al. 2018).
1	13	89466656	-8	AGAT	0.02504	LINC01040 (intron)	
1	9	21636638	-2	AC	0.01917	intergenic	
1	13	106650603	-3	AGAT	0.01876	LINC00443 (intron)	
1	7	91380294	-8	AC	0.01876	intergenic	
1	3	139114125	-4	AC	0.01815	BPESC1 (intron)	
1	17	75482525	-2	ACAT	0.01516	TMEM94 (intron)	Intellectual developmental disorder with cardiac defects and dysmorphic facies (Stephen et al. 2018)
1	3	48888571	2	AAAT	0.01448	SLC25A20 (intron)	Carnitine-acylcarnitine translocase deficiency (Huizing et al. 1997)