Genome-wide characterization of *de novo* tandem repeat mutations in the human genome

T. Mokveld[1], E. Dolzhenko[1], H. Dashnow[2], B. van der Sanden[3], B. Pedersen[2], Z. Kronenberg[1], T. Nicholas[2], C. Fanslow[1], C. Lambert[1], N. Koundinya[4], W. Harvey[4], K. Hoekzema[4], J. Knuth[4], G. Garcia[4], K. M. Munson[4], B. Jadhav[5], A. J. Sharp[5], A.Tucci[6], S. Watkins[2], D. W. Neklason[2], A. R. Quinlan[2], C. Gilissen[3], A. Hoischen[3], E. E. Eichler[4], M. A. Eberle[1]; [1]PacBio, Menlo Park, CA, [2]Univ. of Utah, Salt Lake City, UT, [3]Radboudumc, Nijmegen, Netherlands, [4]Univ. of Washington, Seattle, WA, [5]Icahn School of Medicine at Mount Sinai, New York, NY, [6]Genomics England, London, UK

Introduction

A tandem repeat (TR) is a DNA sequence consisting of repetitions of some motif:

TRs are implicated in Mendelian disease, cancer, and complex traits, and are a major source of structural variants. Standard approaches have limitations in effectively analyzing these regions. Utilizing PacBio HiFi sequencing data, TRGT[1] and TRVZ[1] were developed to:

- Estimate repeat lengths and mosaicism
- Analyze sequence composition
- Measure CpG methylation in repeats
- Support repeats up to 10Kb
- Visualize tandem repeats

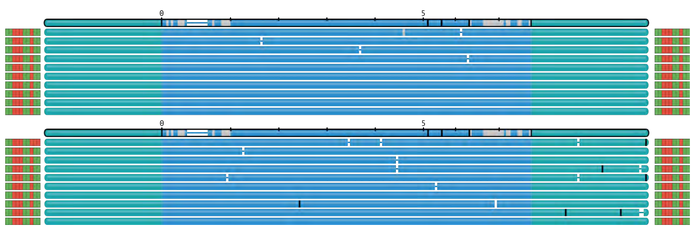

Figure 1. **Allele phasing.** TRGT can phase (near)-homozygous tandem repeat alleles using flanking information obtained from HiFi data.

Introducing TRGT-denovo

TRGT achieves over 99% Mendelian consistency when genotyping repeats in family trios, making it highly reliable for most applications. However, when analyzing millions of tandem repeats across the genome in family trios, there is still a potential for many false positive *de novo* calls. TRGT-denovo is designed to account for and eliminate these false positives. It detects *de novo* TR mutations—both expansions (Fig. 2) and contractions—that are frequently associated with rare diseases and genetic anticipation.

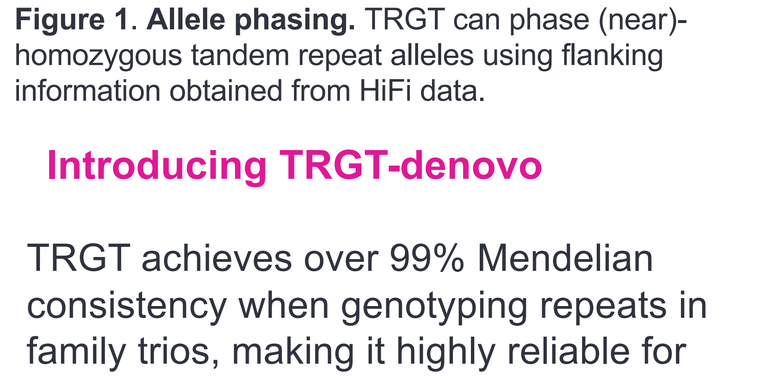

Figure 2. **TR *de novo* expansion.** TRVZ plot highlighting a *de novo* expansion in the short allele inherited from the father.

By working in tandem with TRGT, and by using familial data and pedigree-aware realignment, TRGT-denovo enhances the detection of *de novo* TR mutations, ranging from small (~2 bp) to large (>1000 bp) events.

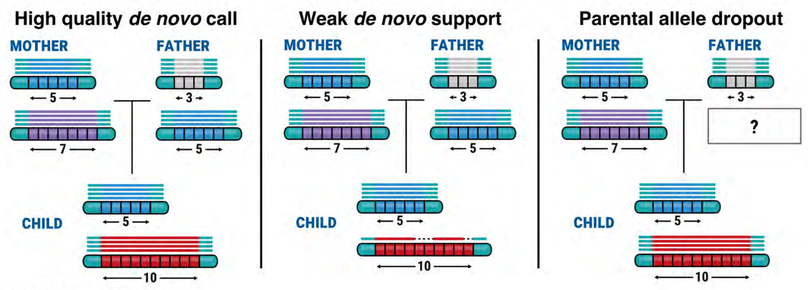

Figure 3. **Confidence of *de novo* TR calls.** High-confidence calls are supported by multiple reads in all family members; weaker calls have few reads in the child for the de novo allele; instances of parental allele dropout are a major source of false positives.

The TRGT, TRVZ, and TRGT-denovo source code and binaries are available at:

github.com/PacificBiosciences/trgt

github.com/PacificBiosciences/trgt-denovo

Validation of putative *de novo* calls

In a genome-wide analysis of approximately 1 million TR loci across eight rare-disease trios, TRGT-denovo:

- Reclassified all previously mislabeled false-positive putative *de novo* TR calls as true negatives[2].
- 19 out of 20 putative *de novo* TR calls were confirmed through targeted sequencing (Table 1).

Type	Range (bp)	Size μ bp (±SD)	Validated/Candidates
Large	[40, 1200]	301 (±372)	7/7
Small	[2, 5]	3.6 (±0.94)	8/9
NDD genes	[8, 24]	19.25 (±6.53)	4/4

Table 1. **Validation results.** Targeted sequencing results of 20 putative *de novo* TR calls detected by TRGT-denovo in eight trios. Calls are categorized by their size (large or small) or their genomic location, specifically in genes associated with neurodevelopmental disorders (NDD).

A pathogenic *de novo* TR in *AFF3*

Expansions of a GCC motif within *AFF3* were found to have a strong negative effect on educational attainment and cognitive function in large cohorts[3]. In two trios with HiFi data available, TRGT and TRGT-denovo were used to characterize, detect, and rank *de novo* TR mutations (Fig 4.).

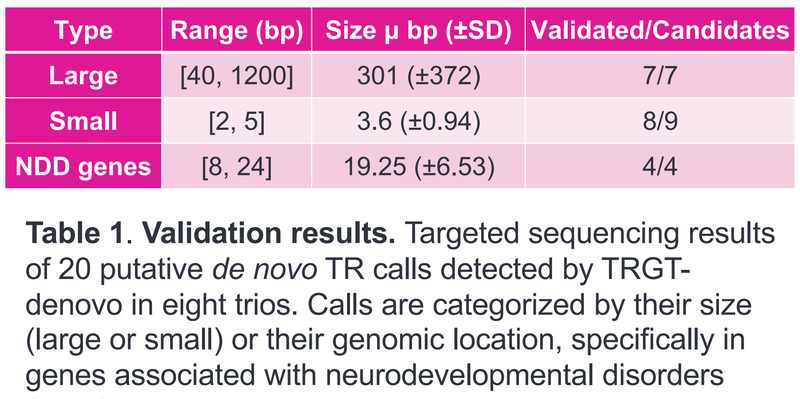

Figure 4. ***AFF3* genotyping with TRGT[3].** Two trios (**A,D**) with per-allele motif counts and TRVZ plots. The *de novo* TR mutations are highlighted in the probands (**B,E**), derived from the parental long alleles (**C,F**).

The largest *de novo* TR in both trios occurred in *AFF3*, far exceeding other events in size (Table 2).

A

Region	Gene	Motif	Father	Mother	Proband	Size Δ (bp)
chr2:100104799-100104824	**AFF3**	3	8,**424**	8,8	8,**157**	-906
chr6:141331488-141331726	NA	42	6,6	6,6	6,5	-22
chr5:33297973-33298142	NA	13	13,15	10,13	13,14	13
chr5:174558306-174559068	NA	4	189,203	193,203	180,203	-12
chr14:70254913-70255140	NA	32	7,8	6,8	6,8	-6
chr19:4805854-4805953	NA	5	20,22	22,22	21,22	-5
chr4:183841149-183841252	NA	4	26,27	26,26	26,28	4
chr18:71548770-71548912	NA	4	36,39	35,37	37,38	-4
chr5:156567250-156567461	NA	4	53,52	50,52	50,51	-4
chr1:16504803-16504906	NA	4	31,33	24,30	30,32	-4

B

Region	Gene	Motif	Father	Mother	Proband	Size Δ (bp)
chr2:100104799-100104824	**AFF3**	3	5,16	5,**133**	5,**822**	2,026
chr2:2462066-2462667	NA	37	16,16	16,16	16,15	-37
chr10:18041471-18042025	NA	32	17,16	15,18	15,15	-32
chr12:25077144-25077434	NA	11	36,57	53,56	53,55	-24
chr3:1245167-1245366	NA	2	89,101	136,138	101,152	20
chr3:74344372-74344468	NA	1	71,98	71,97	71,81	-17
chr14:27333818-27333984	NA	8	20,21	21,21	19,20	-16
chr2:162033832-162033924	NA	2	52,53	46,49	38,50	-16
chr17:8794948-8795035	NA	4	22,22	22,22	22,20	-8
chr16:1750172-1750259	NA	4	22,22	22,22	22,24	8

Table 2. ***De novo* calls *r*anked by size.** The top 10 putative *de novo* TR mutations in the two trios, sorted by the mean absolute magnitude of their size difference relative to the corresponding parental allele.

Multi-generational *de novo* TR calling

The four-generation CEPH K1463 pedigree[4] offers the unique opportunity to not only characterize genome-wide *de novo* TR mutation patterns (Fig. 5) but also validate detected *de novo* candidates through their observed transmission across generations (Fig 6.).

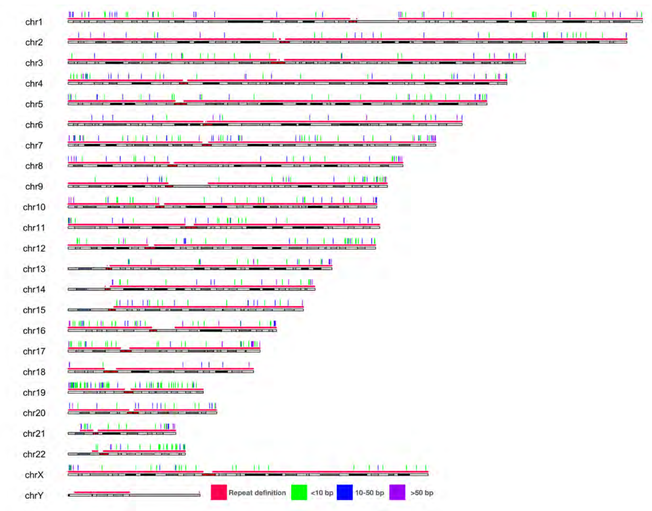

Figure 5. Distribution of *de novo* calls. The genome-wide distribution of putative de novo calls within the pedigree.

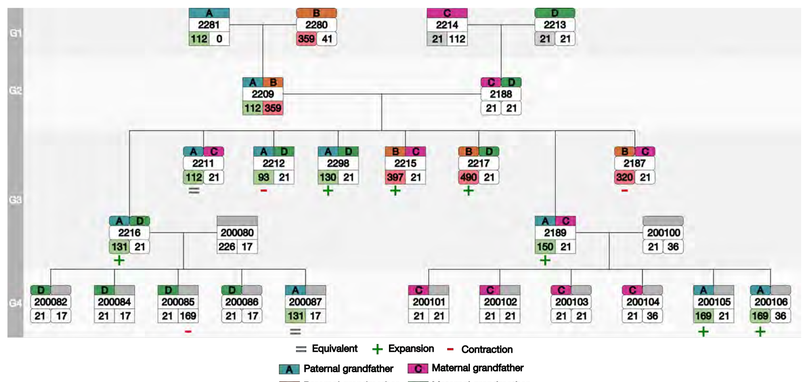

Figure 6. Highly mutable TR. Tracking the transmission of a TR site (chr8:2,623,322-2,623,462) through generations G1 to G4, showing its tendency to expand or contract. Colored blocks above IDs mark inheritance blocks[4].

Conclusion

We developed TRGT-denovo, a tool for genome-wide detection of *de novo* TR mutations.

References

1. E. Dolzhenko, et al. (2023) Resolving the unsolved: Comprehensive assessment of tandem repeats at scale.
2. E. Kucuk, et al. (2023) Comprehensive de novo mutation discovery with HiFi long-read sequencing.
3. A phenome-wide association study of methylated GC-rich repeats identifies a GCC repeat expansion in AFF3 as a significant cause of intellectual disability. [**Board No. PB4611**]
4. Building the spectrum of ground truth genetic variation in a four-generation 21-member CEPH family. [**Board No. PB3334**]