Detection of repeat expansions with PureTarget

M. Eberle[1], G. De Sena Brandine[2], V. Gaysinskaya[2], J. Aiyedun[2], J. Rocha[3], D. Kilburn[2], S. Kingan[4], E. Dolzhenko[2], Z. Kontogeorgiou[5], A. Szabo[6], C. Zarouchlioti[6], R. Thaenert[7], P. Alvarez Jerez[8], K. Billingsley[8], S. Lameiras[9], S. Baulande[9], A. Davidson[10], G. Koutsis[5], G. Karadima[5], S. Tome[11]; [1]PacBio, Oceanside, CA, [2]PacBio, Menlo Park, CA, [3]PacBio, Bel Air, MD, [4]PacBio, San Mateo, CA, [5]Natl. and Kapodistrian Univ. of Athens, Athens, Greece, [6]Univ. Coll. London, London, United Kingdom, [7]Quest Diagnostics, Marlborough, MA, [8]NIH, Bethesda, MD, [9]Inst. Curie, Paris, France, [10]UCL, London, United Kingdom, [11]INSERM, Paris, France

Abstract: Short tandem repeats (STRs) are DNA sequences composed of repetitions of 1-6bp motifs. Expansions of STRs are the cause of over 60 monogenic diseases, including Huntington's disease, Fragile X syndrome, and amyotrophic lateral sclerosis. In addition to their length, the pathogenicity of these STRs is impacted by sequence composition, methylation status and mosaicism. One such example is a repeat in an intron of the *RFC1* gene whose reference sequence consists of a short stretch of AAAAGs while expansions that span hundreds of AAGGGs cause cerebellar ataxia with neuropathy and vestibular areflexia syndrome. Another example is the *FMR1* repeat whose expansions are typically hypermethylated. Detecting all the characteristics associated with pathogenic repeat expansions traditionally required multiple assays, however long-read sequencing of unamplified DNA holds the promise to resolve all of the required features in a single assay.

We describe a robust amplification-free protocol to generate long-read HiFi sequencing libraries containing a panel of loci associated with 20 pathogenic STR expansions. The protocol can be multiplexed to sequence 48 samples at up to 1000x coverage per locus in one sequencing run. To assess the accuracy of this protocol, we sequenced 129 samples with validated pathogenic expansions at 20 loci including *CNBP*, *DMPK*, *RFC1* and *C9orf72*.

Combined, we tested 2580 sample-expansion combinations, including technical replicates, for expansions between 66 bp and >10kb. Our assay correctly categorized all (129/129) expansions, including the detection of hypermethylation in the *FMR1* expansion and differentiating the pathogenic AAGGG motif in *RFC1*. We identified additional expansions in *FXN*, *RFC1* and *TCF4*, consistent with these loci having carrier frequencies between 1:50 and 1:20. Excluding these three genes, we found no unexpected expansions (0/2064) in any sample-loci combination.

We will also present a detailed characterization of lengths, sequence composition, mosaicism, and methylation of normal and expanded alleles in 150 genomes. Most repeats we profiled exhibit high genetic or epigenetic polymorphism and also mosaicism at the expanded size ranges. Motivated by these results, we describe a novel computational approach that will capture all these modalities to robustly differentiate between normal and abnormal variation at known pathogenic or any other repeats in the human genome. In summary, we will present a protocol and a set of computational methods for accurately assessing tissue-level molecular landscapes of various pathogenic STRs, which can be further adapted to other loci in the human genome.

Repeat expansion panel of 20 targets

Gene(s)	Associated disease
ATN1, ATXN1, ATXN2, ATXN3, ATXN7, ATXN8, ATXN10, CACNA1A, PPP2R2B, TBP	Spinocerebellar ataxia
FMR1	Fragile X-associated disorders
C9orf72	Amyotrophic lateral sclerosis and Frontotemporal dementia
DMPK, CNBP	Myotonic dystrophy (DM1, DM2)
FXN	Friedreich's ataxia
RFC1	CANVAS
HTT	Huntington's disease
AR	Spinal-bulbar muscular atrophy
PABPN1	Oculopharyngeal muscular dystrophy
TCF4	Fuchs endothelial corneal dystrophy

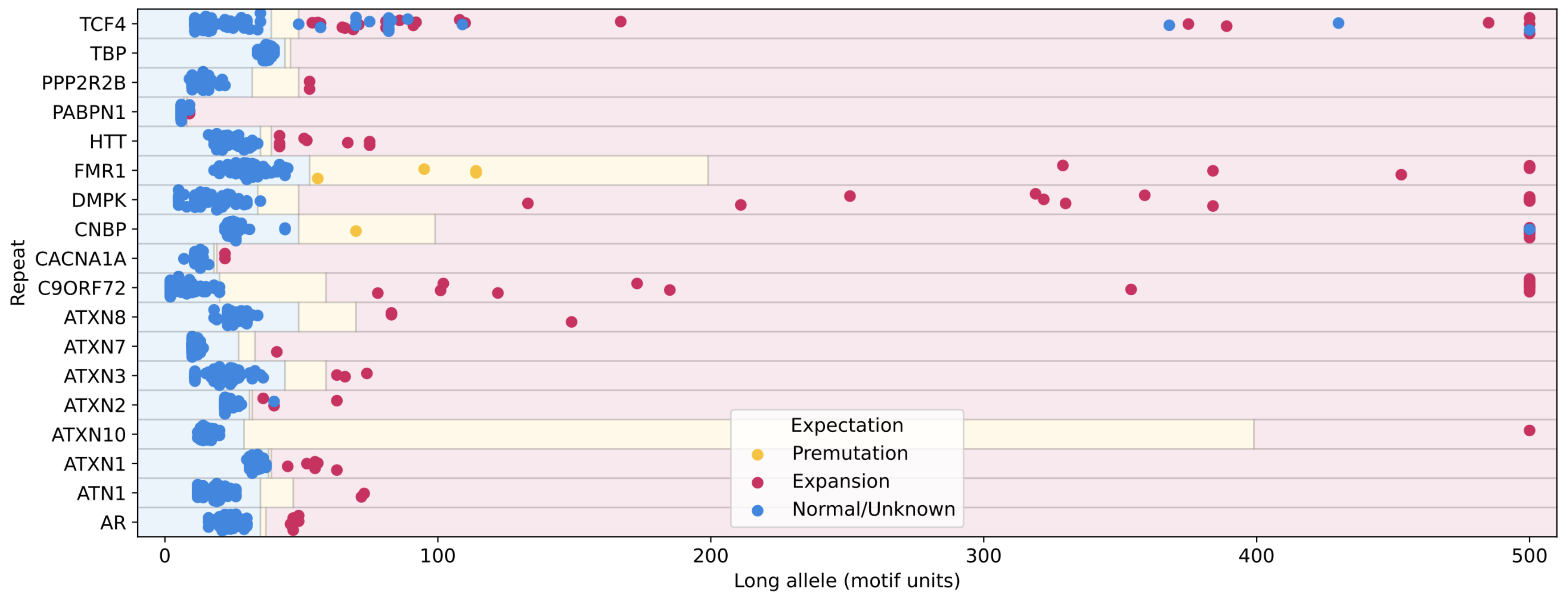

Figure 1. "Swim lane plot" showing the long allele length of 150 samples for 18 autosomal and X-linked dominant loci. Dots are colored by expected genotype.

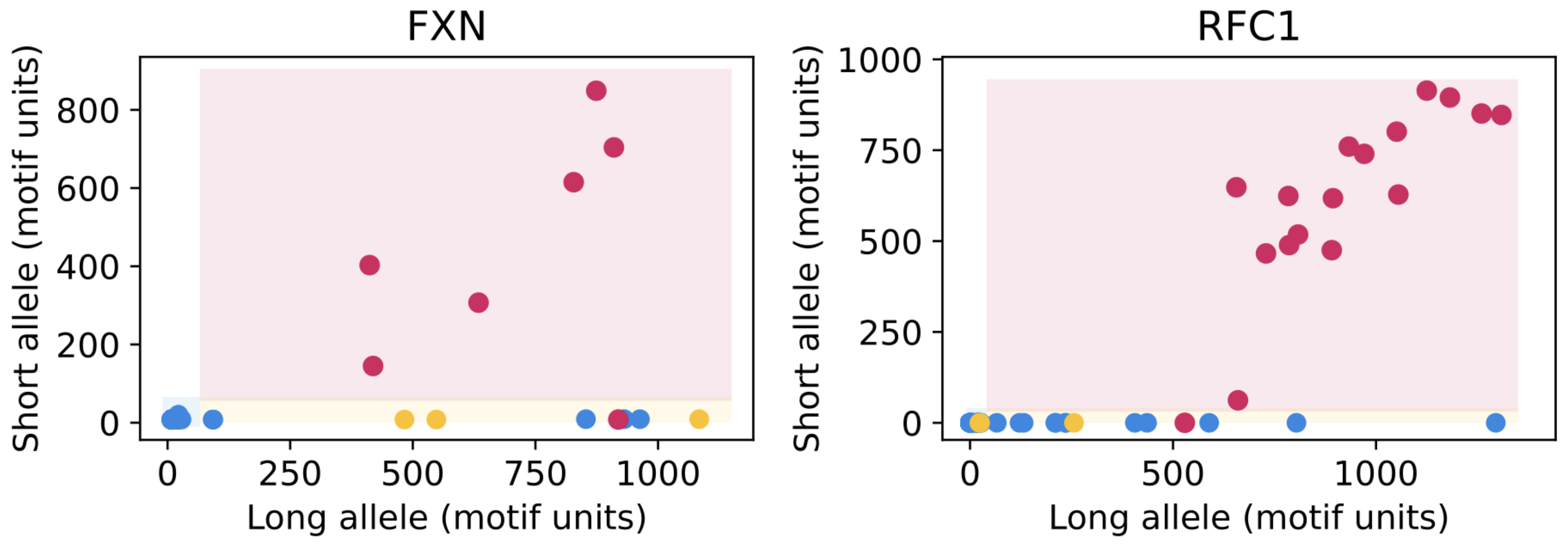

Figure 2. Scatterplot showing short and long allele lengths of 150 samples for 2 autosomal recessive loci FXN and RFC1. All the known pathogenic samples have been identified carrying two expansions (of the pathogenic allele for *RFC1*). In addition, we identified several previously unknown carriers consistent with the high carrier frequencies of these repeat expansions.

A paper describing tandem repeat genotyping tool (TRGT)

A talk on resolving complex tandem repeat regions with TRGT and other methods

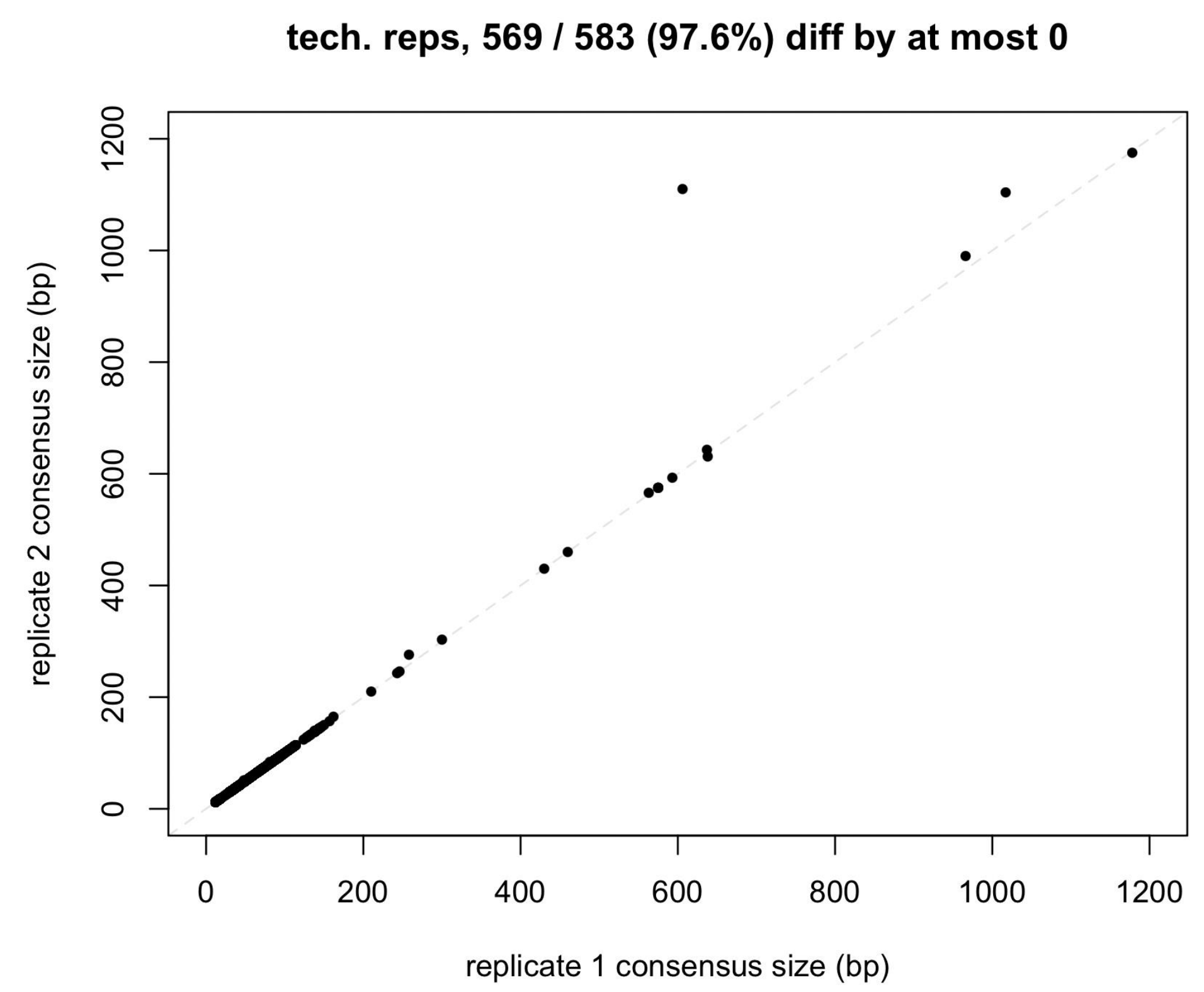

Figure 3. Consensus comparison in 15 pairs of technical replicates, 8 males and 7 females. 570/584 have identical consensus sequences, 577 are at most off by 1, and all (584/584) have concordant ranges, meaning the range of allele sizes overlap between replicates.

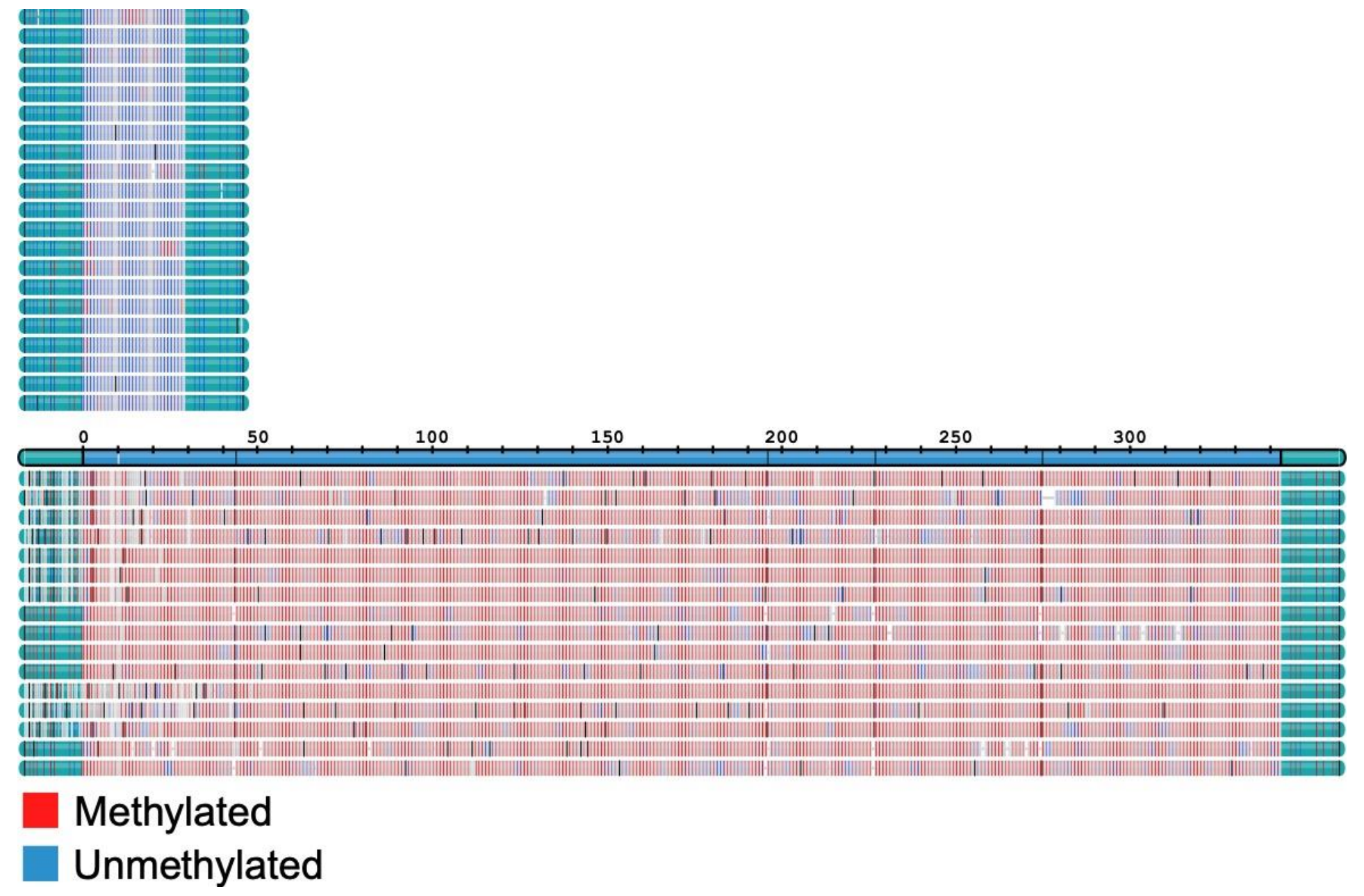

Figure 4. TRVZ plots of the two alleles for an *FMR1* repeat showing the methylation on the longer allele. Top image shows the shorter allele with very little methylation while the expanded allele is fully methylated.

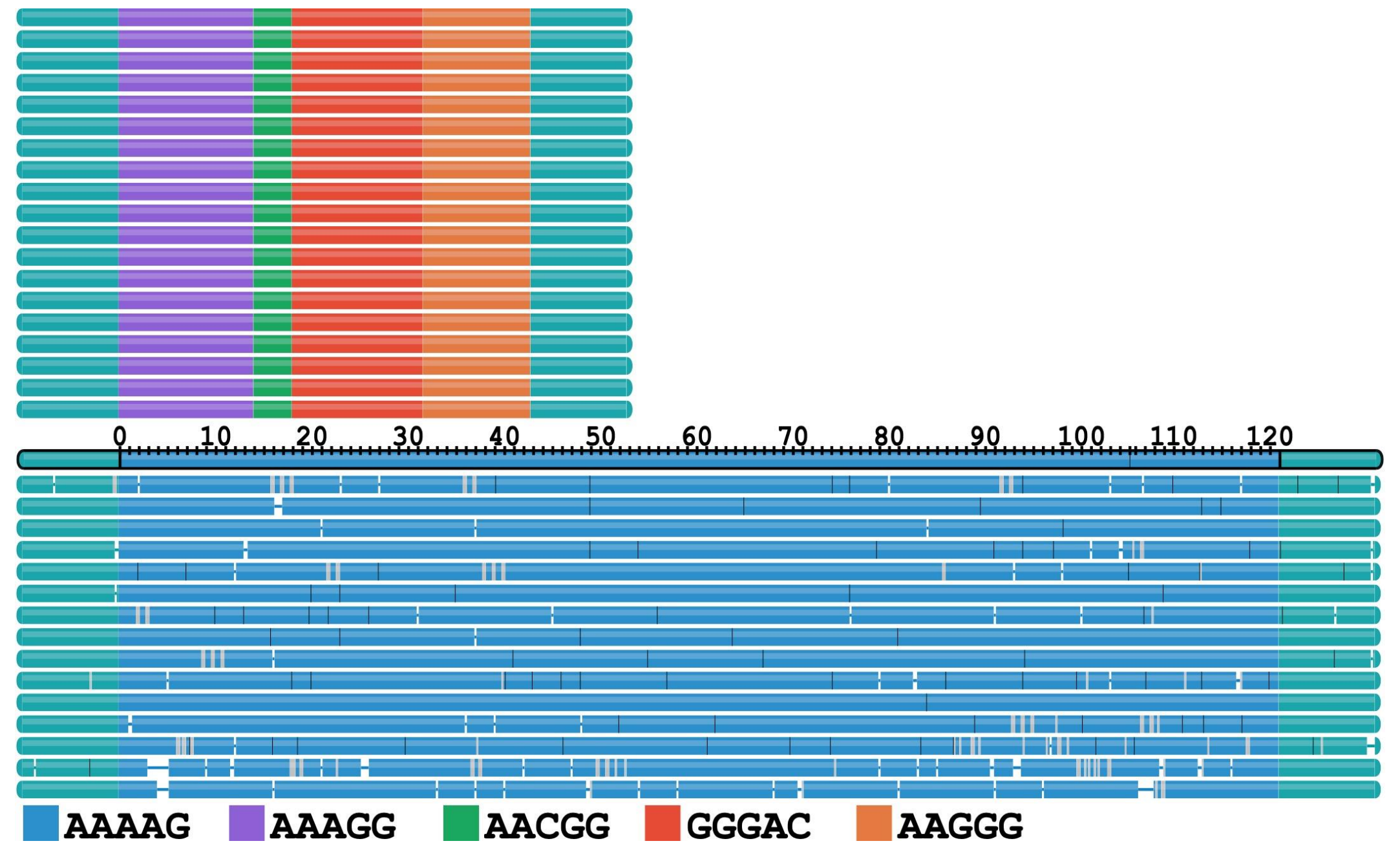

Figure 5. TRVZ plots of the two alleles for the *RFC1* repeat known to cause CANVAS. Top panel shows the short allele comprised of four different repeat motifs. Bottom panel shows the expanded allele but because the expansion is the non-pathogenic AAAAG repeat, this individual is not a carrier.

Conclusions:

- The PureTarget assay combined with the software tools TRGT and TRVZ allows accurate measurments of 20 known pathogenic repeats.
- Applying PureTarget to 150 samples with known repeat expansions correctly identified all known repeat expansions. While the unexpected expansions all occur in genes known to have high carrier frequencies.
- The visualization package, TRVZ, enables visual inspection of the genotype calls made by TRGT providing additional confidence in the expansion calls.
- Technical replication experiments show that >97% of genotype calls are identical with the differences often occurring due to mosaicism.