nature genetics

Supplementary information

https://doi.org/10.1038/s41588-024-01808-5

A common flanking variant is associated with enhanced stability of the *FGF14*-SCA27B repeat locus

In the format provided by the authors and unedited

Supplementary Note

Comparison of *FGF14* reference sequences from the T2T-CHM13v2.0 (chr13:101,377,550-101,377,792) and GRCh38.p14 (chr13:102,161,575-102,161,726) assemblies revealed differences in both 5' and 3' flanking regions (Supplementary Fig. 1). Of the 4,382 alleles analyzed by PacBio HiFi sequencing in this study, we found that 768 (17.53%) matched the T2T-CHM13 5'-flanking sequence, while only 6 (0.14%) matched the GRCh38 5'-flanking sequence. In addition, 95 alleles (2.17%) carried a single terminal T>A substitution of the T2T-CHM13 5'-flanking reference sequence. These alleles did not contain the common 5'-flanking variant (5'-CFV). Furthermore, we found that only 10 of the 4,382 alleles analyzed by PacBio matched the T2T-CHM13 reference 3'-flanking sequence, which lacked a (TAG) sequence immediately following the GAA repeat. In comparison, 3,756 alleles (85.71%) matched the GRCh38 reference 3'-flanking sequence. The sequence of the major *FGF14* allele, identified through the analysis of our PacBio dataset, is shown in Supplementary Figure 1. In the PacBio dataset, the sizes of the 17 alleles with 30 or more GAA triplets bearing the 5'-CFV were nine alleles of 30 repeats, five alleles of 31 repeats, one allele of 32 repeats, and two alleles of 34 repeats, while the sizes of the two alleles with less than 30 GAA triplets bearing the 5'-RFS were both of 25 repeats.

Since the 5'-flanking region of the *FGF14* repeat locus harbored variants related to the repeat stability, we studied the 3'-flanking region for similar patterns. We analyzed the 15 nucleotides immediately following the termination of the GAA repeat sequence in 2,191 individuals (4,382 alleles) by whole-genome long-read PacBio HiFi sequencing. The only common variant observed was an A>C polymorphism (rs61965263) four nucleotides after the termination of the GAA repeat sequence: $(GAA)_n$TAGCAA. This variant was observed in 329 alleles (7.51%). No other variant was observed in more than 1% of alleles. All alleles carrying the polymorphism rs61965263 were below 30 repeat units (Extended Data Fig. 2). Furthermore, the majority of the alleles with the common 3'-flanking sequence (5'-TAGAAATGTGTTTAAGAA) were also below 30 repeat units. None of the 3'-flanking variants distinguished unique populations of alleles as observed with 5'-flanking variants (Extended Data Fig. 2).

A total of 411 intergenerational events were measured by PacBio sequencing (from the Genomic Answers for Kids cohort) and 67 intergenerational events were measured by Sanger sequencing. Of these, 297 were from parents with the 5'-CFV and 169 were from parents with the 5'-RFS. The 12 remaining intergenerational events were from cases where the parental alleles contained the less common 5'-flanking sequences: three harbored a C1 flanking variant, one harbored a C2 flanking variant, two harbored a C5 flanking variant, and six harbored other flanking variants. The 478 intergenerational events together show a clear picture of the length-dependent instability of the GAA repeat tract (Extended Data Fig. 3a). We observed stable intergenerational transmission in three of four events with the C1 or C2 flanking variants,

two of two events with the C5 flanking variant, and one of six events with other 5'-flanking variants (Extended Data Fig. 3b). Analyzing intergenerational stability according to motif purity further showed the stabilizing effect of sequence interruption and impurity on transmission of the repeat, particularly when the locus contains over 75 repeat units (Fig. 2 and Supplementary Fig. 3). Specifically, alleles with 30 to 75 triplets and containing the 5'-RFS were stably transmitted in 72 of 97 alleles (74.23%). In this size range, 69.23% (45/65) of GAA-pure repeats were stably transmitted compared to 84.37% (27/32) of non-GAA-pure alleles. Larger alleles with 75 to 250 triplets and containing the 5'-RFS were stably transmitted in only 2 of 33 alleles (6.06%), both of which were non-GAA-pure. Finally, only one of the 39 alleles larger than 250 repeats with the 5'-RFS were passed stably (2.56%), which was non-GAA-pure (Fig. 2). The greatest degree of instability was observed in GAA-pure alleles, with expansions observed upon transmission from female parents and contractions upon transmission from male parents, similar to what has previously been observed in Friedreich ataxia.[1]

In order to disambiguate the effect of the 5'-CFV on repeat stability from that of another sequence element with which it is in strong linkage disequilibrium, we analyzed the haplotypes of the long-read PacBio samples. First, we collected stable population variants from a recent analysis of the 1000 Genomes and Human Genome Diversity Project (HGDP) datasets,[2] which had been statistically phased and include persons from a wide range of self-reported geographic origins. From it, we collected 15 biallelic variants with a minor allele frequency >0.2 which were within 10 kb of the *FGF14* repeat locus. We then fitted a *k*-means clustering model so that it could assign a haplotype group label to new alleles, given phased calls of these 15 biallelic variants. We chose ten clusters for this analysis to define ten 'major haplotype groups' from this diverse set of short-read data. PacBio genomes had single nucleotide variants (SNVs) and insertion/deletion polymorphisms (indels) called with PEPPER-Margin-DeepVariant,[3] structural variants called with PBSV, and their *FGF14* repeat locus and flanking variant evaluated with TRGT.[4] The VCF files from these three methods were then input, along with the BAM file, to HiPhase,[5] which returned physical phase blocks. Samples where physical phasing could not be established for at least 10 kb on either side of the *FGF14* repeat were excluded, leaving 1,674 samples for the haplotype analysis (Extended Data Fig. 4a). We then used the fitted *k*-means clustering model to assign haplotype group IDs to each of the 3,348 haplotypes so that the labels on the long-read samples matched those on the short-read samples. This analysis revealed that the 5'-CFV is observed on all ten haplotype groups defined by the 1000 Genomes and HGDP dataset, while the 5'-RFS is observed on only six of those haplotype groups (Extended Data Fig. 4b). This finding demonstrates a difference in the distribution of haplotypes carrying the 5'-RFS compared to the 5'-CFV: all haplotype groups carrying the 5'-RFS are also observed to carry the 5'-CFV, though the reverse is not true. This analysis supports the hypothesis that the 5'-CFV is the more ancestral allele, while also suggesting that the 5'-RFS may represent a recurrent variant. The short flank variant has a similar distribution of haplotypes to that seen in the 5'-RFS, being observed on five of the same six

haplotype groups, plus two others. The degenerate forms of the 5'-CFV with two to four nucleotide variations from the 5'-CFV had a different distribution of haplotype groups than the 5'-CFV but were still detected on nine of the ten haplotype groups. This evidence supports the hypothesis that the flanking variant is the primary determinant of the observed patterns in repeat stability, rather than any other genomic elements in linkage disequilibrium. We observed a similar distribution in the frequency of haplotypes from our cohort as from the 1000 Genomes and HGDP dataset, supporting that at this locus our cohort is reasonably representative of that diverse, short-read dataset (Extended Data Fig. 4b). A visualization of all tandem repeat alleles seen more than three times in the PacBio dataset, along with their interrupting motifs, flanking sequence type, and distribution of haplotype groups, is shown in Supplementary Figure 4.

To assess whether the pathogenic *FGF14* repeat expansion is present on a single 'disease haplotype' or if it can arise recurrently, we generated long-read sequencing data on seven unrelated patients with SCA27B (two of self-reported Han Chinese descent and five of self-reported European descent) and found that the pathogenic alleles were present on three distinct haplotype groups. As expected, all seven of these pathogenic repeats were flanked by the 5'-RFS, and the three haplotypes were a subset of the six haplotypes previously observed to harbor 5'-RFS sequences (Supplementary Fig. 7).

The allele frequency and haplotype distribution in humans presented thus far support the hypothesis that the 5'-CFV is the ancestral allele and the 5'-RFS represents a recurrent variant. To study this further, we attempted to inspect the *FGF14* repeat locus in Neanderthal and Denisova reference assemblies, which had, however, no coverage, likely due to the repetitive nature of this locus. We next checked the reference assemblies of other hominoids and found that Chimpanzees (*Pan troglodytes*; panTro6 assembly), Bonobos (*Pan paniscus*; panPan3 assembly), Gorillas (*Gorilla gorilla gorilla*; gorGor6 assembly), and Gibbons (*Nomascus leucogenys*; nomLeu3 assembly) carry a 5'-flanking sequence identical to the human 5'-CFV sequence (Supplementary Fig. 5a,b). The Sumatran orangutan (*Pongo abelii*) reference genome ponAbe3 contains a C3A3 variation of the 5'-CFV (Supplementary Fig. 5b). Each of these hominoid species carried 4 to 7 GAA triplets in their reference genome (Supplementary Fig. 5b). Beyond hominoids, the 5'-flanking sequences of each of the four Old World monkeys that we inspected (Green monkey / *Chlorocebus sabaeus*, chlSab2 assembly; Crab-eating macaque / *Macaca fasicularis*, macFas5 assembly; Rhesus macaque / *Macaca mulatta*, rheMac10 assembly; Baboon / *Papio anubis*, papAnu4 assembly) differed by a single nucleotide variation from the 5'-CFV (5'-TAGTCATAGTATCCCAA) (Supplementary Fig. 5c), while the 5'-flanking sequences of the two New World monkeys we analyzed (Marmoset / *Callithrix jacchus*, calJac4 assembly; Squirrel monkey / *Saimiri boliviensis*, saiBol1 assembly) differed by several nucleotide variations from the 5'-CFV (Supplementary Fig. 5c). This analysis strongly suggests that the 5'-CFV is indeed more ancestral than the 5'-RFS.

Given the high degree of flanking sequence, repeat length, and repeat configuration polymorphism observed in the human population, we examined population variation data of other great ape species for similar polymorphisms. We re-analyzed the 79 short-read great ape genome sequences from Prado-

Martinez et al.[6] and found minimal intra-species variation in GAA repeat length. All species carried 5 to 7 GAA triplets (Supplementary Fig. 6). All 50 alleles of the 25 Chimpanzees (*Pan troglodytes*; ten *ellioti*, six *schweinfurthii*, five *verus*, and four *troglodytes* subspecies), all 26 alleles of the 13 Bonobos (*Pan paniscus*), all 62 alleles of the 31 Gorillas (*Gorilla*; 27 *gorilla gorilla*, three *beringei graueri*, and one *gorilla diehli* subspecies), all ten alleles of the five Bornean orangutans (*Pongo pygmaeus*), and six of ten alleles of the five Sumatran orangutans (*Pongo abelii*) carried the 5'-CFV. We observed two C3A3 variations and two 5'-TAGTCACAGTACCCCAA variations in the Sumatran orangutans (*Pongo abelii*). This analysis further supports that the 5'-CFV is the ancestral allele by showing no evidence of the existence of the 5'-RFS within any of the great ape species examined. It also reinforces the notion that the 5'-CFV may aid the stability of *the FGF14* repeat region, as the repeat is shorter in these great apes than in most modern humans and exhibits very little intra-species polymorphism.

We hypothesized that the 5'-CFV could be aiding repeat stability relative to the reference sequence through an effect on the epigenetic architectures of the locus and its neighboring regions. To test if there is any relationship between the 5'-flanking sequence and the epigenetic state of the repeat locus, we performed Fiber-seq[7] on post-mortem human cortex. Nuclei were FACS sorted into $NeuN^+$ and $NeuN^-$ groups, comprising neurons and glia, respectively. Each fiber spanning the repeat locus was assessed for the number of GAA repeats, the flanking sequence composition, 5-methyl cytosine indications, and open chromatin indications (detected as 6-methyl adenines). From this data, open chromatin regions and nucleosome placements were inferred. In $NeuN^+$ nuclei, we observed increased chromatin accessibility in the ~120 bp on each side of the GAA repeat on fibers bearing the 5'-CFV as compared to those bearing the 5'-RFS (Supplementary Fig. 8). The repetitive sequence itself did not bear many marks of accessible chromatin. The 5'-CFV fibers carried 9 GAA repeats each, while the 5'-RFS fibers carried 53 repeats each. Fibers with the C2A1 variation of the 5'-CFV (58 repeats), which are associated with moderate repeat lengths (Extended Data Fig. 1b), showed a level of chromatin accessibility intermediate to that of the 5'-CFV and 5'-RFS. Finally, fibers with the 5'-CFV in $NeuN^-$ nuclei (9 GAA repeats) showed a similar pattern of chromatin accessibility to what was seen for the 5'-CFV in $NeuN^+$ nuclei, but with lower signal intensity (Supplementary Fig. 8). Flanking sequences other than the 5'-CFV were not observed in the $NeuN^-$ nuclei sequenced. This shows first functional evidence for a sequence-dependent role for the 5'-CFV in influencing repeat stability through an epigenetic mechanism that is likely dependent in part on its four terminal cytosines. Furthermore, we found by querying the Homo Sapiens Comprehensive Model Collection (HOCOMOCO v12)[8] of transcription factor binding sites (at a *P* value threshold of < 0.0001) that the 5'-CFV sequence likely introduces binding sites for Jun-related (JunB and JunD) and zinc finger (zinc finger protein 573, 677, and 749) transcription factors, all expressed in the central nervous system.[9] Specifically, Jun-related transcription factors were predicted to bind the proximal and mid portion of the 5'-CFV while the zinc finger transcription factors were predicted to bind to the terminal portion of the 5'-CFV encompassing the four terminal cytosines and two terminal adenines. This provides a potential explanation for the greater chromatin accessibility observed with the 5'-CFV and the intermediate accessibility observed with the

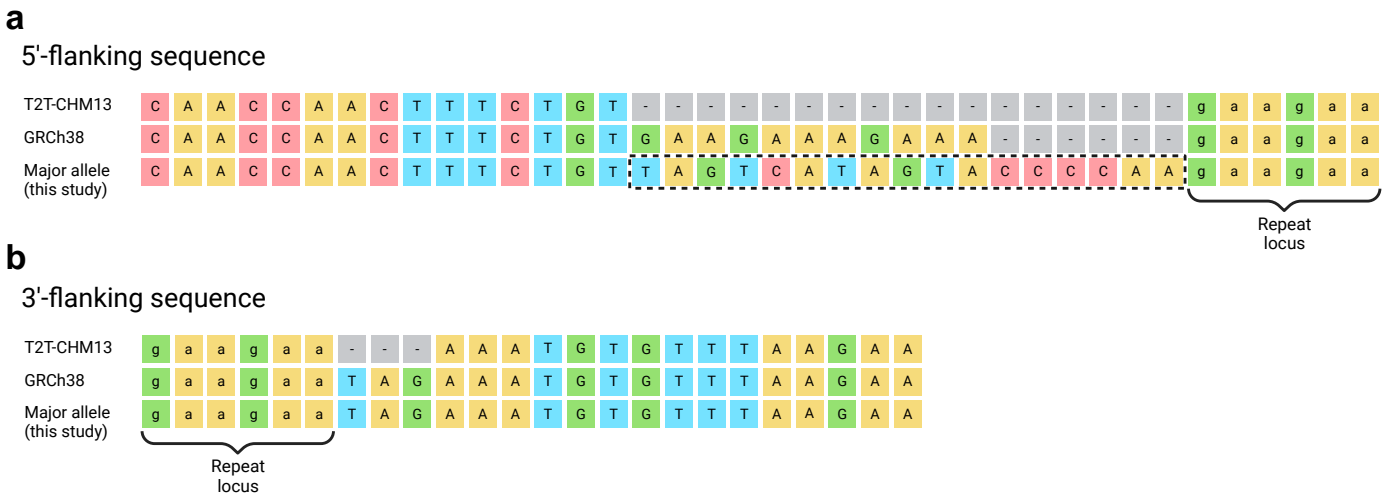
degenerate C2A1 variation, which is predicted to abolish the binding site for the zinc finger proteins 677 and 749 due to the loss of two terminal cytosines and one terminal adenine.[8]

We further studied whether the 5'-CFV, and in particular the four terminal cytosines, could be stabilizing the repeat locus by preventing the formation of a DNA secondary structure that is more prone to DNA replication and transcription errors. We analyzed the predicted DNA/RNA secondary structure of the 60 nucleotides upstream of the *FGF14* repeat locus in the presence of the 5'-RFS sequence, the 5'-CFV sequence, and the degenerate CFV sequence (C2A1 variant) using the RNAStructure web server (http://rna.urmc.rochester.edu/RNAstructure.html; Supplementary Fig. 9). We observed that the 5'-RFS sequence creates an inverted repeat sequence with the 50 bases upstream, which is predicted to form a multi-branched loop (ΔG = -5.5 kcal mol^{-1}). The 5'-CFV sequence appears less likely to form such a structure as the CCCC tract in this sequence creates a point of resistance to this folding since there is no poly-G sequence with which it can pair. However, the C2A1 variant seen in the Fiber-seq experiment, which is also associated with longer repeat lengths in the population data, creates a structure with similar stability to that associated with the 5'-CFV (ΔG = -6.2 kcal mol^{-1} vs ΔG = -5.5 kcal mol^{-1}). The absence of significant difference in stability between the predicted DNA secondary structures is unlikely to account for the difference in repeat stability observed with each flanking sequence.

Alternatively, the four final cytosines of the 5'-CFV may hypothetically decrease the rate of triplex formation at the repeat locus, thus preventing repeat instability.[10] GC-rich flanking sequences, even when unilateral, have been shown to act as a clamp in reducing the kinetics of triplex formation.[11] Degenerate sequences of the 5'-CFV, which are less GC-rich due to a shorter final stretch of cytosines, and 5'-RFS may less effectively decrease the rate of triplex formation, which in turn can induce instability of the repeat locus. However, further studies will be required to full unravel the mechanisms by which the 5'-CFV leads to repeat stability.

Supplementary Figures

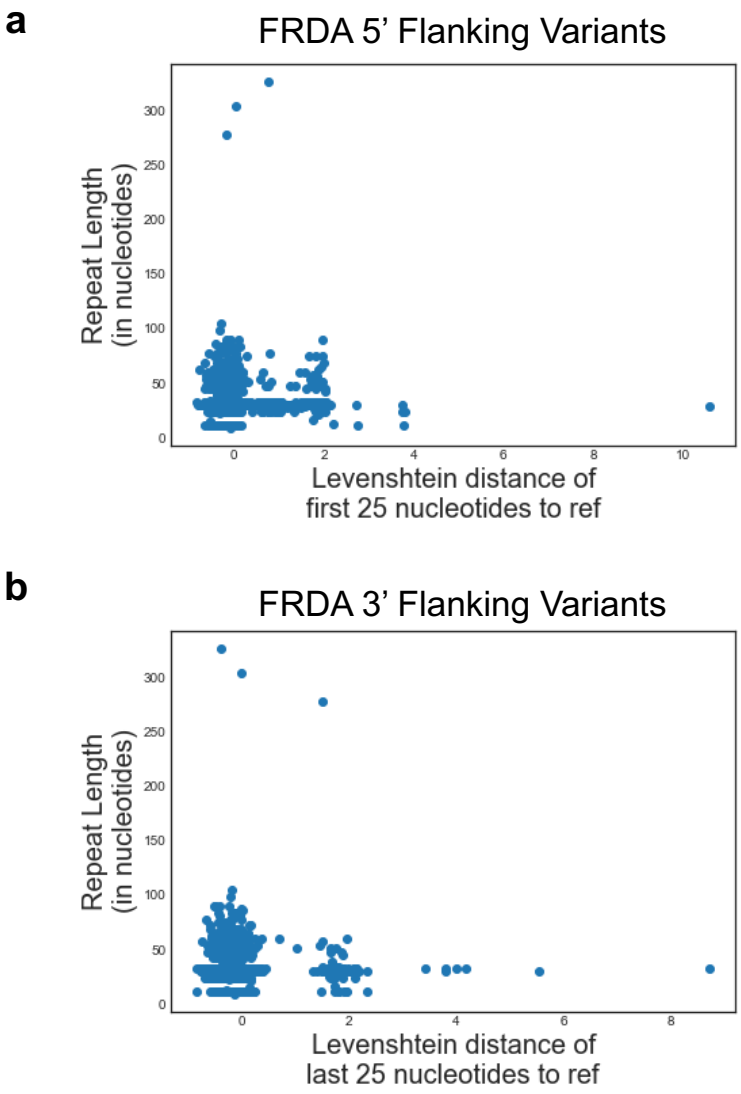
Supplementary Figure 1



Supplementary Figure 1 | Reference flanking sequences from T2T-CHM13 and GRCh38 assemblies.

Reference (**a**) 5'-flanking and (**b**) 3'-flanking sequences from T2T-CHM13 and GRCh38 assemblies. The sequence of the major *FGF14* allele, identified through analysis of our PacBio dataset, is shown for comparison. The dashed line shows the sequence of the common 5'-flanking variant present in the majority of alleles. Figure created with BioRender.com.

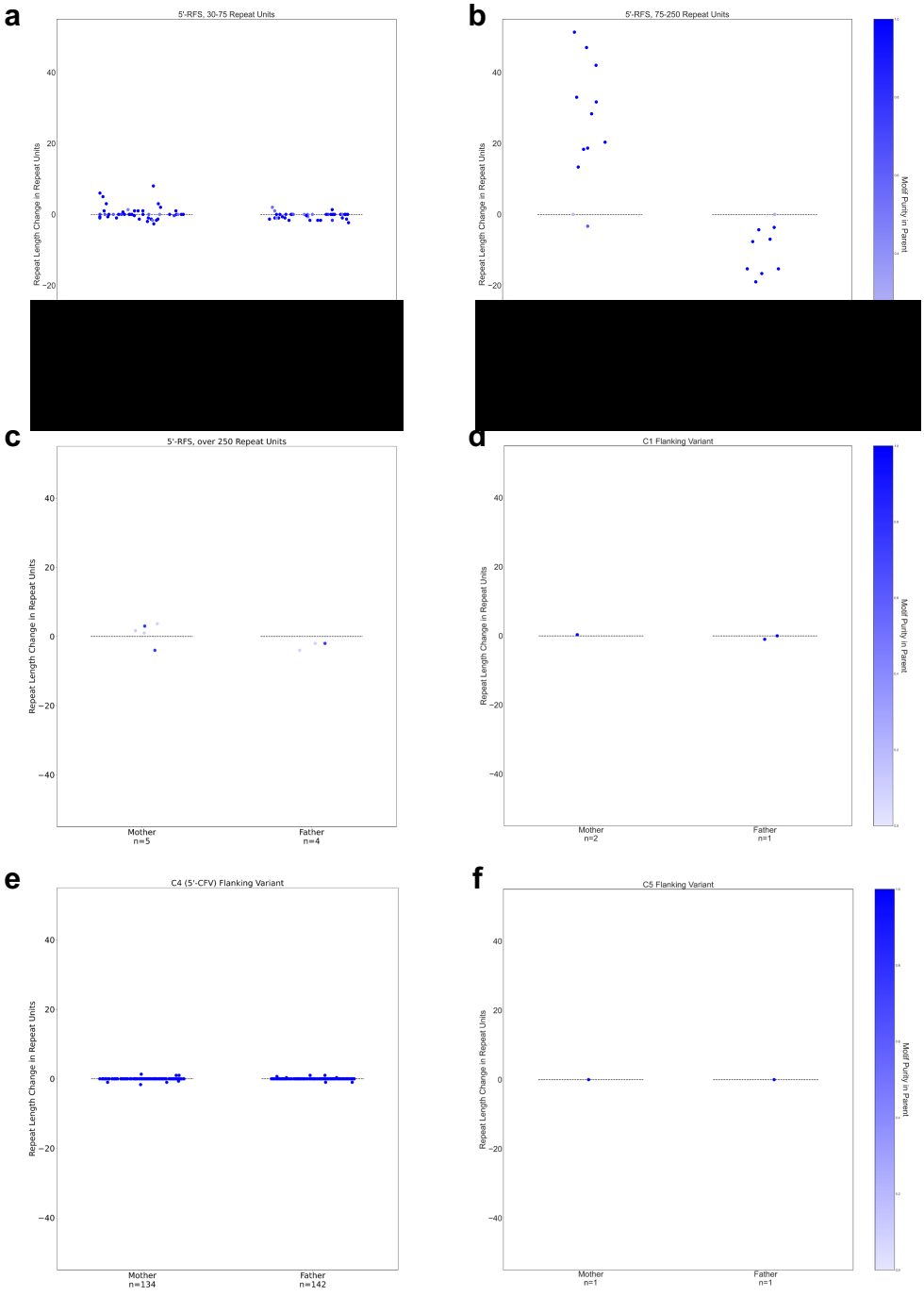
Supplementary Figure 2



Supplementary Figure 2 | Effect of flanking sequence variation on repeat size at the Friedreich ataxia locus.

Scatterplots of 2,054 alleles (from the All of Us cohort) at the Friedreich ataxia locus (*FRDA*) where the *y*-axis plots the GAA repeat length in nucleotides and the *x*-axis plots the Levenshtein distance between the reference and the observed sequences for the 25 nucleotides immediately (**a**) 5' to the GAA repeat locus, and (**b**) 3' to the GAA repeat locus. Gaussian noise was added to the *x*-axis values to mitigate overlap of data points. No clear segregation of allele sizes by flanking variants was observed.

Supplementary Figure 3

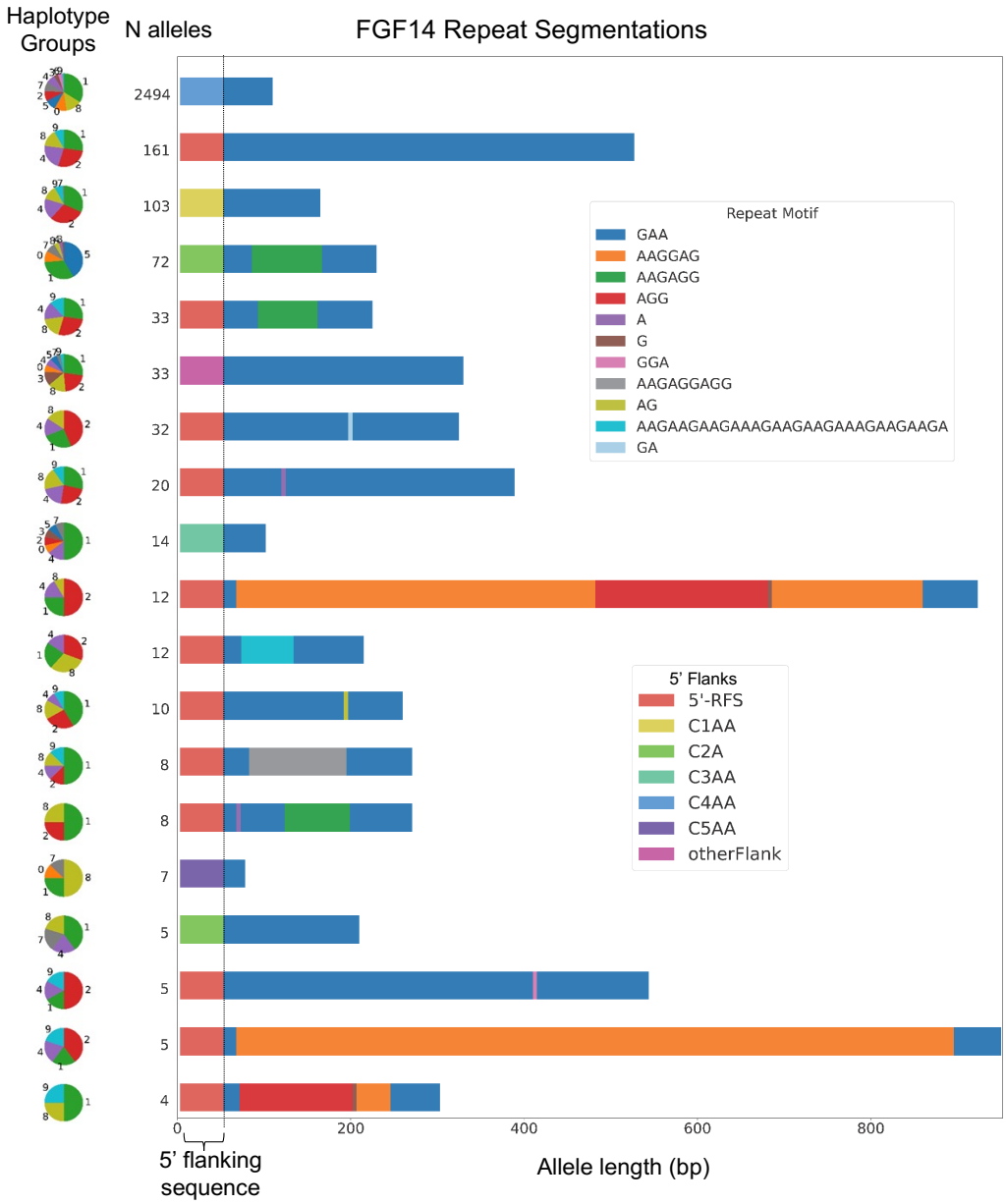


Supplementary Figure 3 | Analysis of parent-offspring transmission of the *FGF14* repeat according to the 5'-flanking sequence variant and GAA repeat purity.

Change in GAA repeat length across intergenerational transmission events (from the Genomic Answers for Kids cohort) as measured by PacBio HiFi sequencing. The color of the data points is a function of the GAA repeat motif purity, with dark blue indicating pure and lighter blue impure motif (a hue scale is shown on the

right *y*-axis). The *y*-axis shows the change in repeat length from parent to child. Contractions are plotted below the dashed lines while expansions are plotted above them. Random noise was applied across the *x*-axis within each category to maximize data visualization. (**a**) 5'-RFS group of alleles, ranging from 30 to 75 repeat units long, (**b**) 5'-RFS group of alleles, ranging from 75 to 250 repeat units long, (**c**) 5'-RFS group of alleles, over 250 repeat units long (one transmission is not shown here for visual clarity in which a GAA-pure, paternally inherited allele contracted by 132 repeat units), (**d**) C1 group of alleles, (**e**) C4 group of alleles, and (**f**) C5 group of alleles. The C2 group is not shown here as it included a single intergenerational event. The number of intergenerational transmission events in each group is indicated below the *x*-axis.

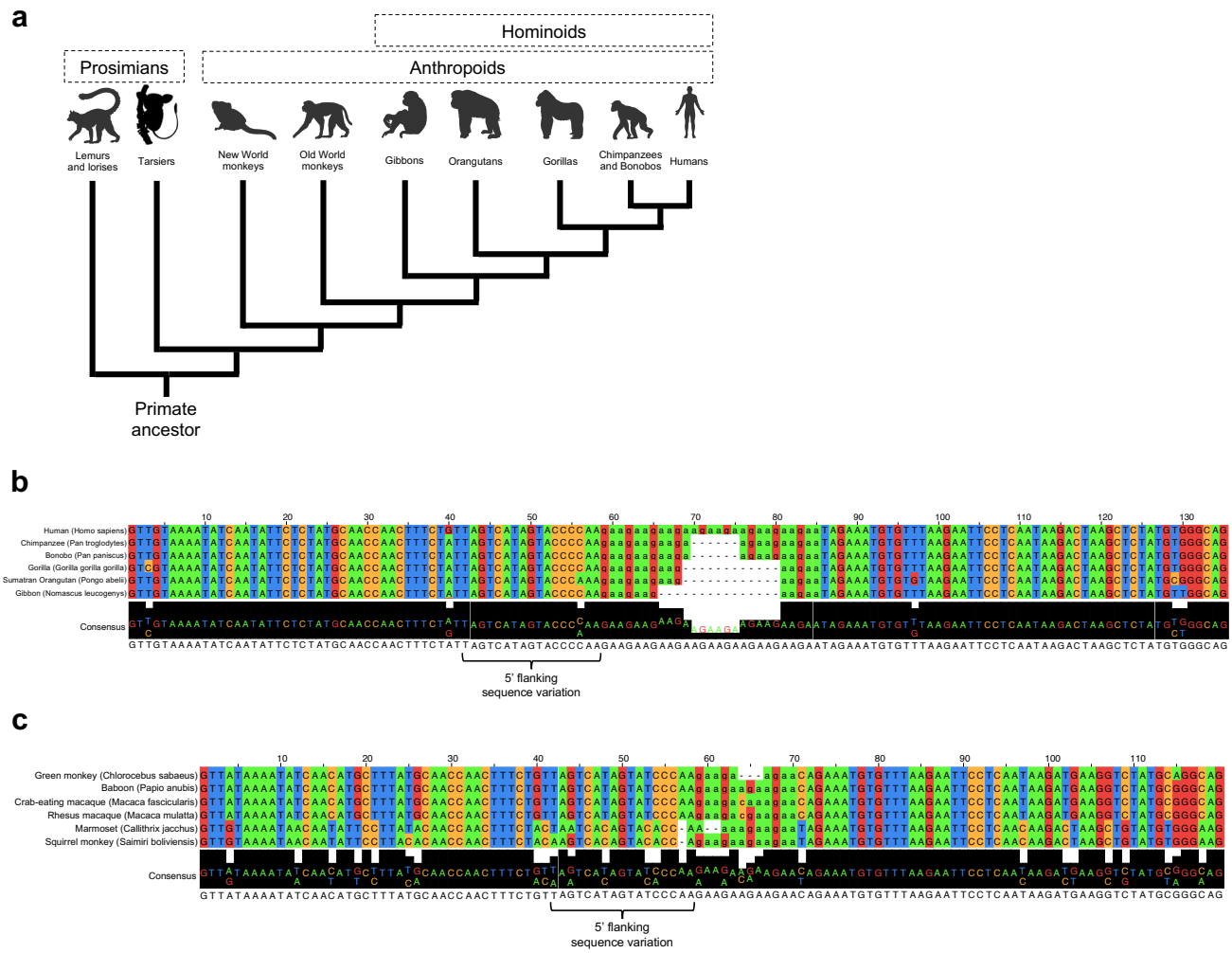
Supplementary Figure 4



Supplementary Figure 4 | Segmentation analysis of the *FGF14* repeat locus and flanking sequence variant.

Average patterns of repeat lengths, motif interruptions and variations, and flanking sequence type in 3,038 alleles (from the Genomic Answers for Kids and All of Us cohorts). The length of each segment (in nucleotides) is represented by the median count of triplets in that repeat segment across all alleles sharing the same pattern. Only allele patterns observed at least three times (allele frequency >0.1%) are depicted. For each allele pattern, the distribution of haplotype groups on which it was observed is depicted on the left as a pie chart.

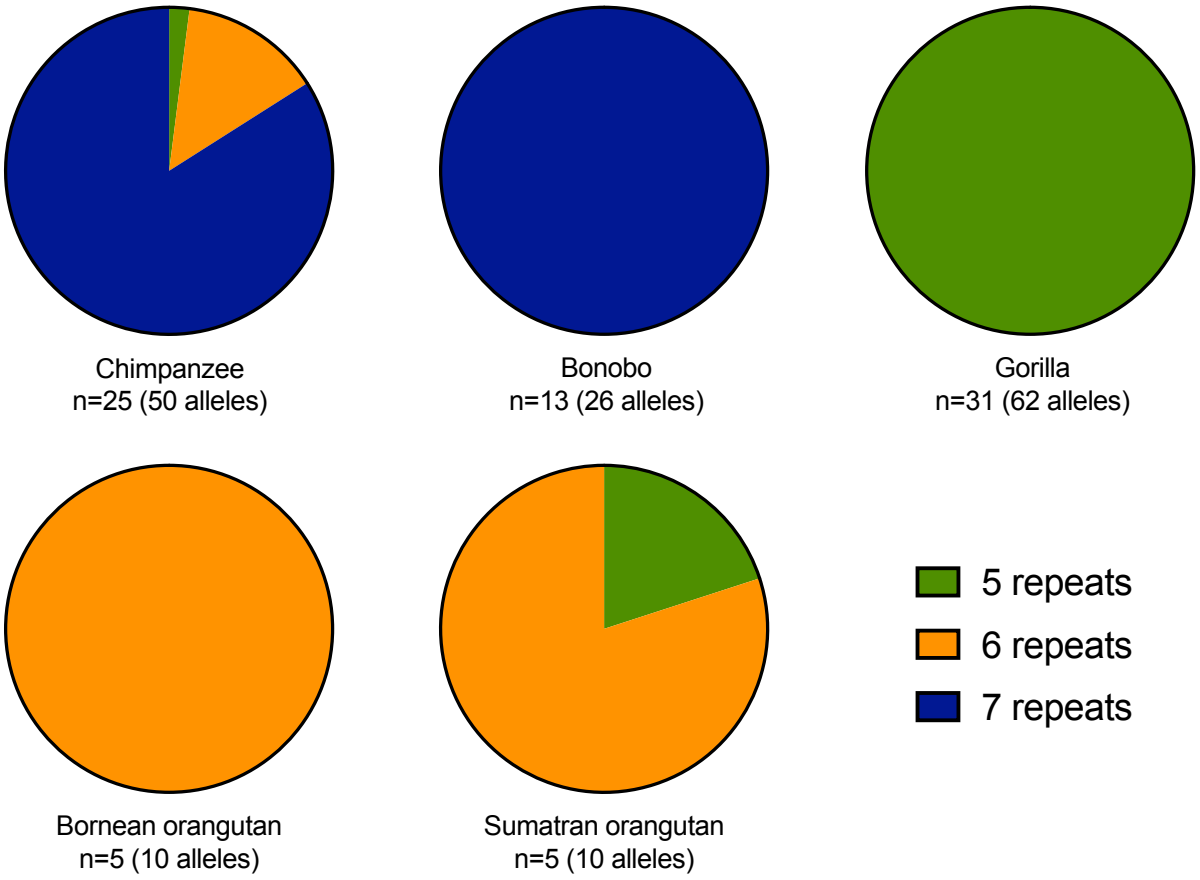
Supplementary Figure 5



Supplementary Figure 5 | *FGF14* repeat locus and 5'-flanking sequence variation in primates.

(**a**) Phylogenetic tree of the prosimian, anthropoid, and hominoid primates. Panel created with BioRender.com. (**b**) Multiple sequence alignment of the reference genomes of non-human hominoid species and human (using the 5'-CFV with the modal repeat length of 9 GAA triplets as the reference sequence) shows high conservation of the 5'-flanking sequence in all species. (**c**) Multiple sequence alignment of the reference genomes of Old World monkeys (Green monkey, Baboon, Crab-eating macaque, Rhesus macaque) and New World monkeys (Marmoset, Squirrel monkey) shows the presence of a 5'-flanking sequence reminiscent of the 5'-CFV in all examined species.

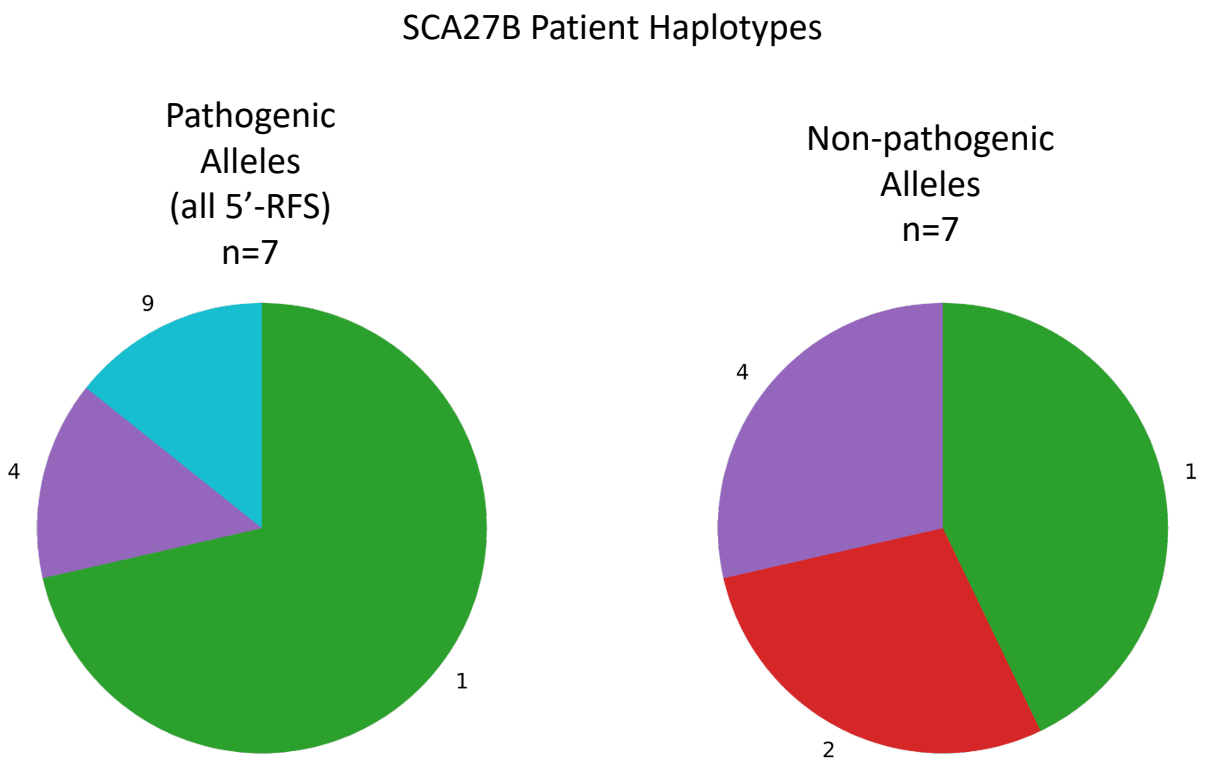
Supplementary Figure 6



Supplementary Figure 6 | Repeat length distribution of the *FGF14* allele in great apes.

Analysis of 79 great ape genome sequences reveals low polymorphism of the GAA repeat tract length in Chimpanzees (*n* = 25), Bonobos (*n* = 13), Gorillas (*n* = 31), Bornean orangutans (*n* = 5), and Sumatran orangutans (*n* = 5), which is always associated with the common 5'-flanking variant or similar sequence.

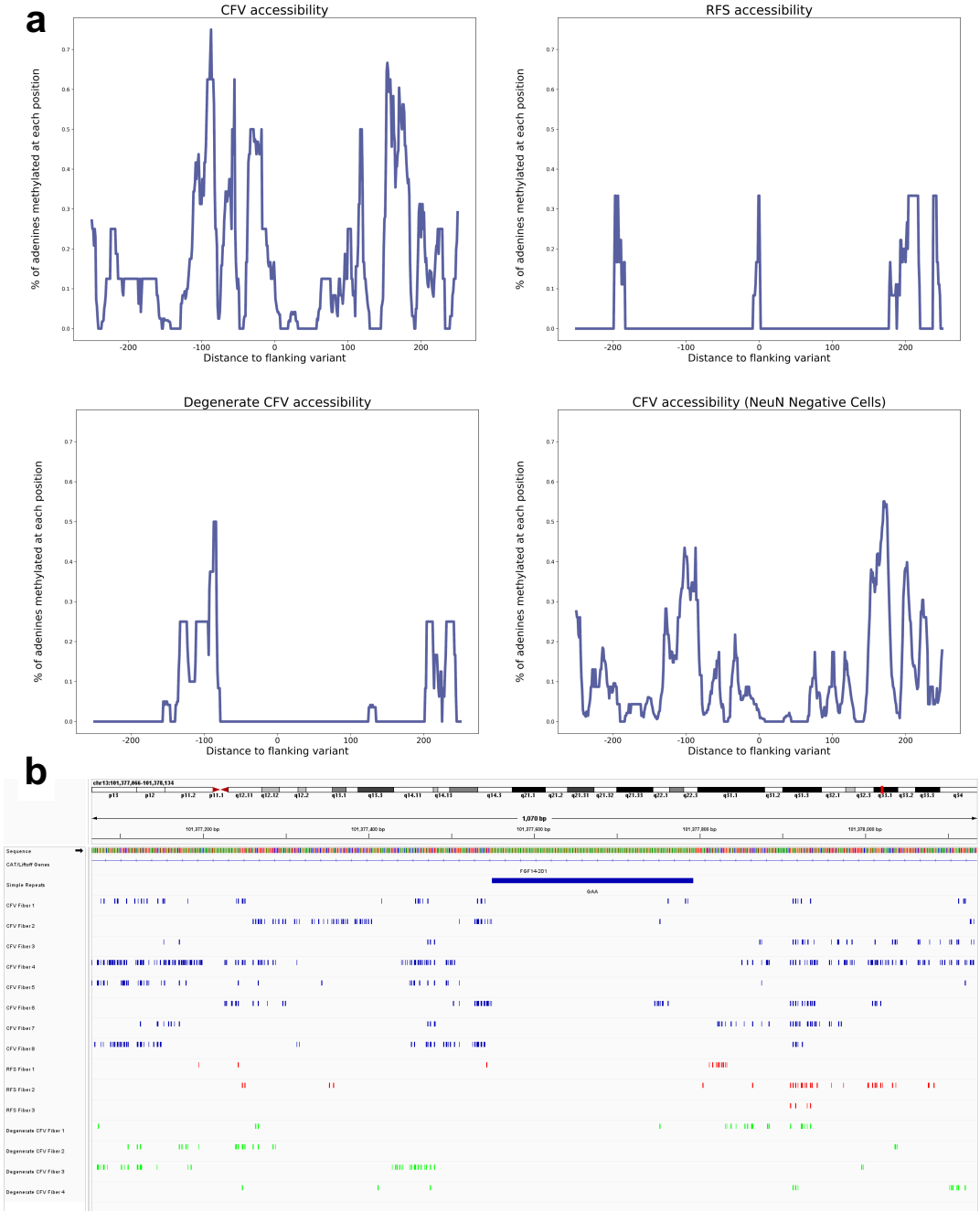
Supplementary Figure 7



Supplementary Figure 7 | Repeat expansion haplotypes in seven patients with SCA27B.

Haplotype analysis of seven unrelated patients with SCA27B (two patients of self-reported Han Chinese ancestry and five patients of self-reported European ancestry, none of which were included in the Genomic Answers for Kids, Care4Rare-SOLVE, or All of Us cohorts). Pie charts of the haplotypes for the pathogenic, expanded (left) and non-pathogenic (right) alleles of seven patients with SCA27B. The haplotype IDs displayed in this figure are directly comparable to those used in Extended Data Figure 4.

Supplementary Figure 8

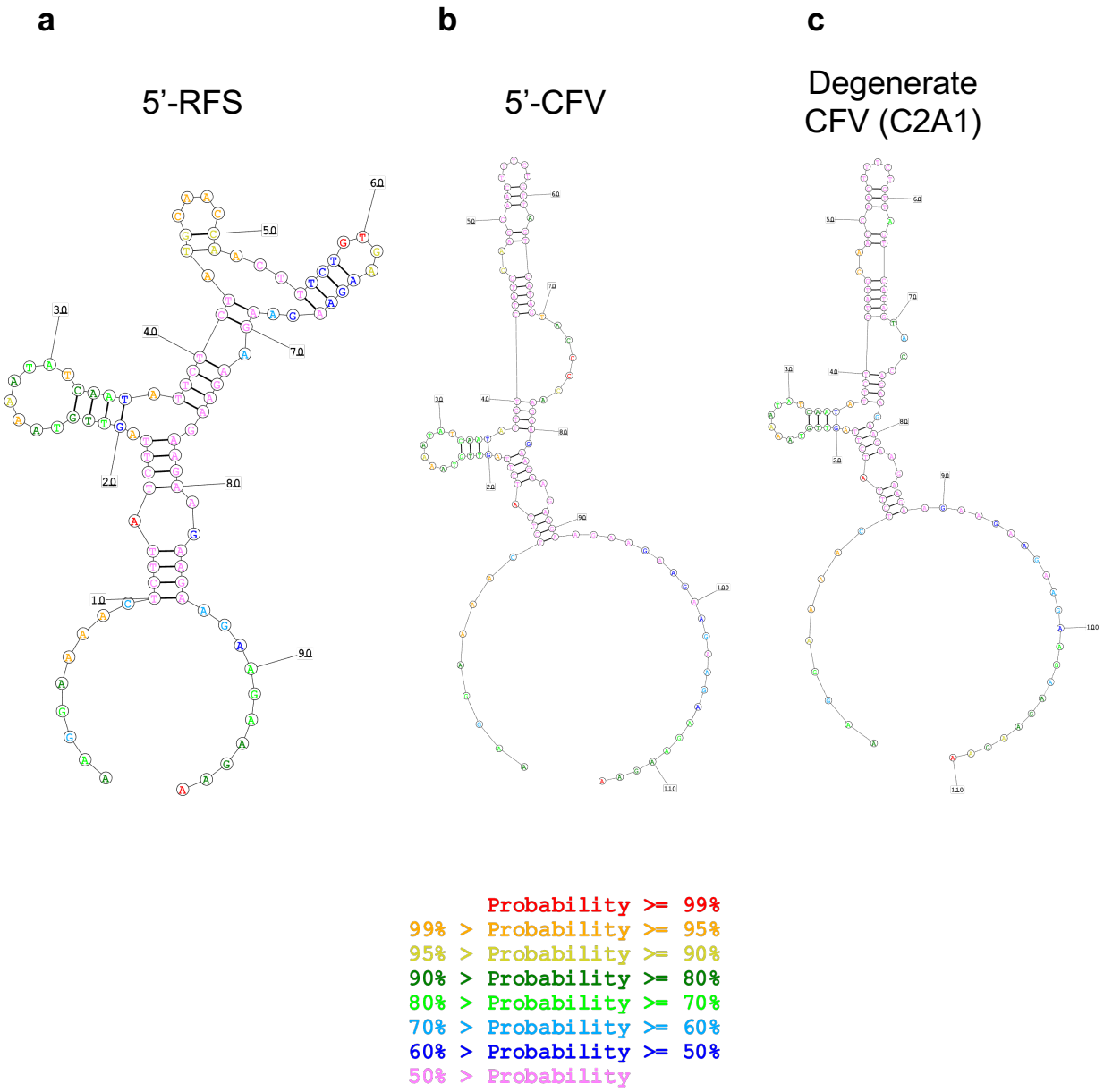


Supplementary Figure 8 | Fiber-seq analysis of single-molecule chromatin architectures surrounding the *FGF14* repeat locus.

Fiber-seq analysis reveals increased chromatin accessibility surrounding the GAA tract on DNA strands carrying the 5'-CFV relative to the 5'-RFS. (**a**) Density plots of the proportion of adenines methylated in a 10 bp sliding window surrounding the GAA tract ranging from 250 bp upstream to 250 bp downstream of

the start of that tract. In this assay, methylated adenines indicate accessible chromatin. The upper-left, upper-right, and lower-right plots are from NeuN$^+$ cells, which have high levels of *FGF14* expression, while the lower-right plot is from NeuN$^-$ cells, which have low levels of *FGF14* expression. (**b**) Visualization on IGV of the adenine methylation events recorded for each fiber spanning the GAA tract in NeuN$^+$ nuclei, colored and grouped according to the flanking sequence of each allele: blue for 5'-CFV, red for 5'-RFS, and green for degenerate CFV (C2A1). There are two alleles containing the 5'-CFV, and one allele each carrying the 5'-RFS and degenerate CFV sequences.

Supplementary Figure 9



Supplementary Figure 9 | Predicted effect of flanking sequences on DNA secondary structure.

Predicted DNA/RNA secondary structure using the RNAStructure web server of the 60 nucleotides upstream of the *FGF14* (GAA)•(TTC) repeat locus along with the first 12 GAA triplet repeats. Panel **a** shows the reference 5'-flanking sequence (5'-RFS), panel **b** shows the C4 (5'-CFV) variant, and panel **c** shows the C2A1 degenerate 5'-CFV variant. The probabilities indicate model confidence in correctness of prediction. The stability of each of the structures is: (**a**) ΔG = -5.5 kcal mol^{-1}, (**b**) ΔG = -5.5 kcal mol^{-1}, and (**c**) ΔG = -6.2 kcal mol^{-1}.

Supplementary References

1. De Michele, G. et al. Parental gender, age at birth and expansion length influence GAA repeat intergenerational instability in the X25 gene: pedigree studies and analysis of sperm from patients with Friedreich's ataxia. *Hum. Mol. Genet.* **7**, 1901-1906 (1998).
2. Koenig, Z. et al. A harmonized public resource of deeply sequenced diverse human genomes. *bioRxiv* doi: 10.1101/2023.01.23.525248 (2023).
3. Shafin, K. et al. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nat. Methods* **18**,1322-1332 (2021).
4. Dolzhenko, E. et al. Characterization and visualization of tandem repeats at genome scale. *Nat. Biotechnol.* doi: 10.1038/s41587-023-02057-3 (2024).
5. Holt, J. M. et al. HiPhase: Jointly phasing small and structural variants from HiFi sequencing. *Bioinformatics* **40**, btae042 (2024).
6. Prado-Martinez, J. et al. Great ape genetic diversity and population history. *Nature* **499**, 471-475 (2013).
7. Stergachis, A. B., Debo, B. M., Haugen, E., Churchman. L. S. & Stamatoyannopoulos, J. A. Single-molecule regulatory architectures captured by chromatin fiber sequencing. *Science* **368**, 1449-1454 (2020).
8. Vorontsov, I. E. et al. HOCOMOCO in 2024: a rebuild of the curated collection of binding models for human and mouse transcription factors. *Nucleic Acids Res.* **52**, D154-D163 (2024).
9. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580-585 (2013).
10. Khristich, A. N. & Mirkin, S. M. On the wrong DNA track: Molecular mechanisms of repeat-mediated genome instability. *J. Biol. Chem.* **295**, 4134-4170 (2020).
11. Kang, S., Wohlrab, F. & Wells, R. D. GC-rich flanking tracts decrease the kinetics of intramolecular DNA triplex formation. *J. Biol. Chem.* **267**, 19435-19442 (1992).