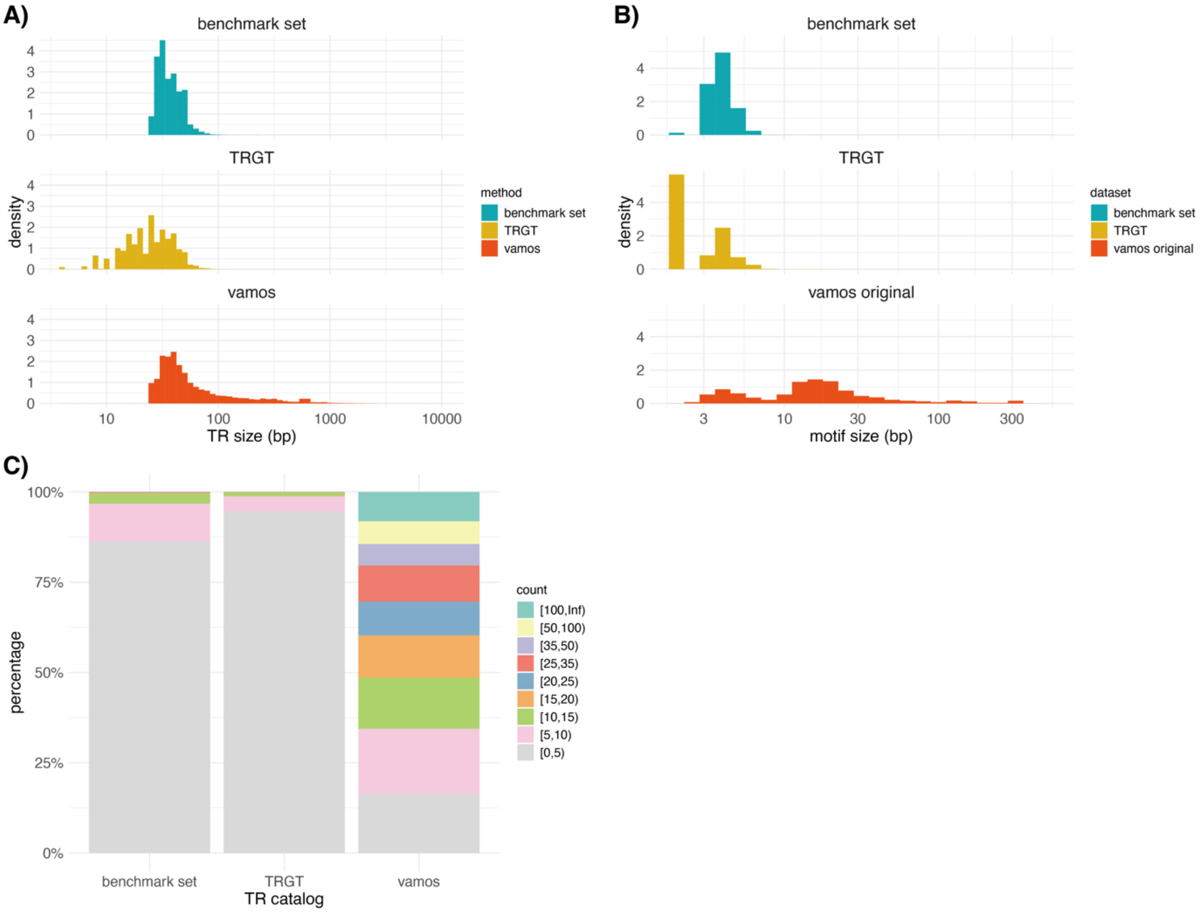
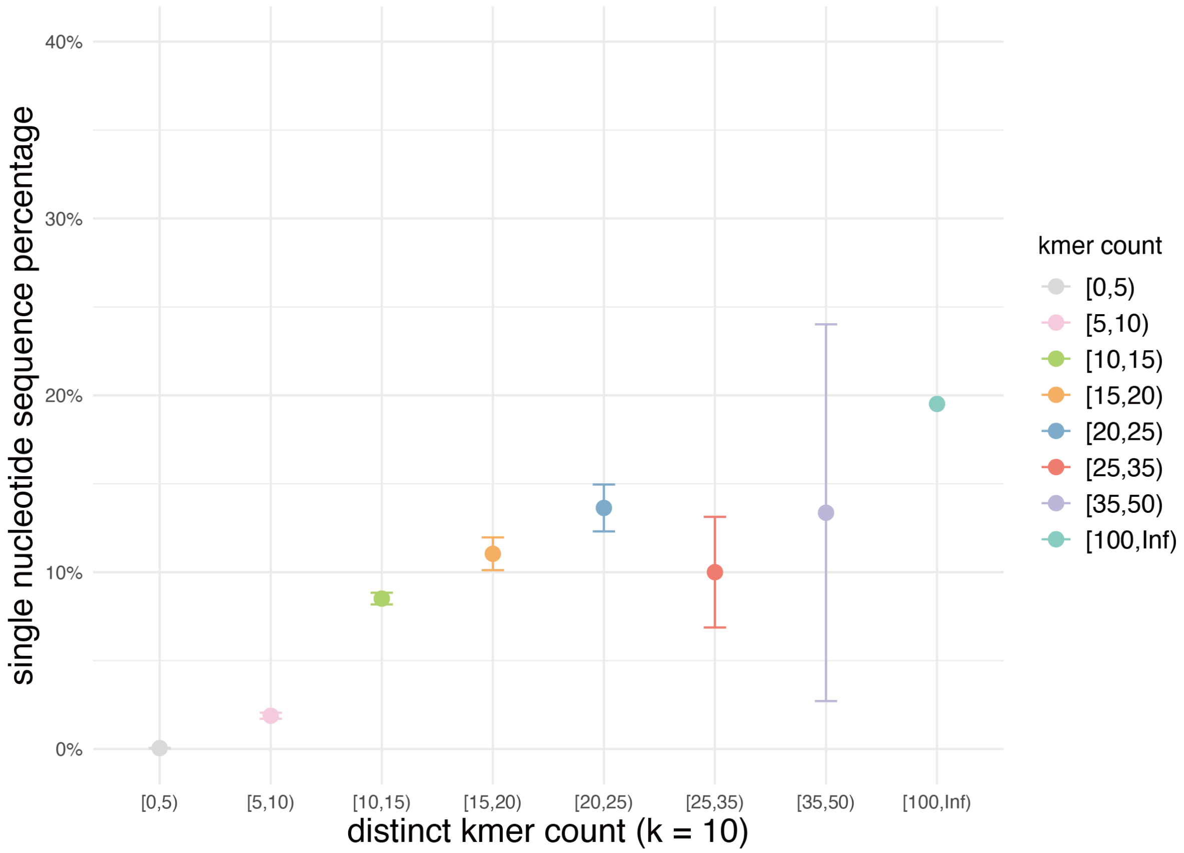
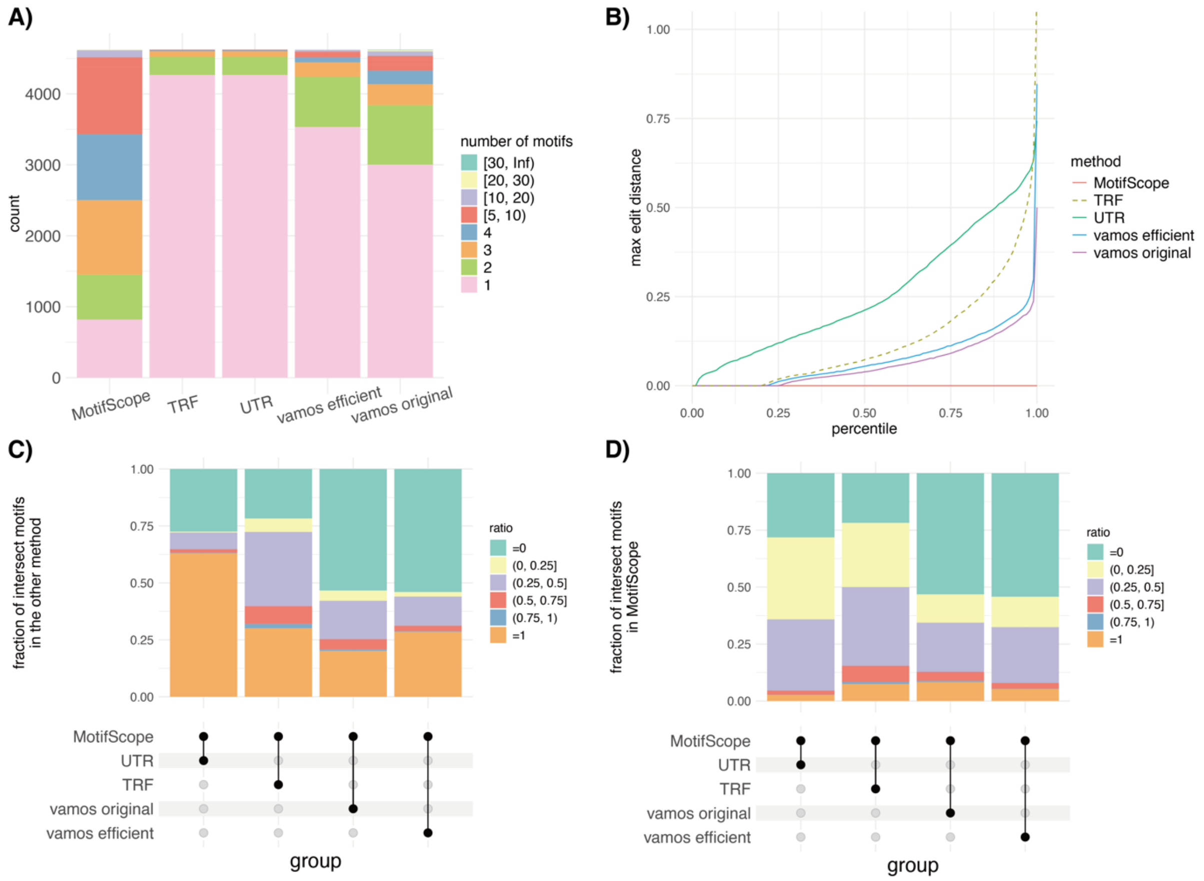


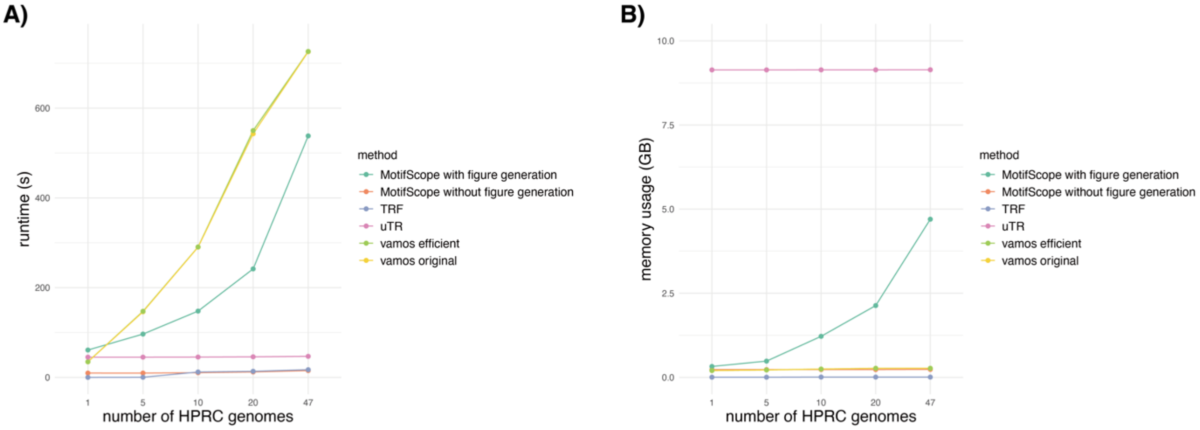
Supplementary Figure 1. Summary of the PacBio TRGT repeat catalog, the vamos VNTR catalog and the set of TRs used for benchmarking. This figure provides an overview of three catalogs: the PacBio TRGT repeat catalog, comprising 171,146 repeats; the vamos VNTR catalog, comprising 467,104 repeats; and the overlap between these two catalogs that are used for benchmarking, comprising 5,486 loci. (A) compares the size distribution of these three datasets. (B) compares the motif size distribution of these three TR sets. For the PacBio set and the benchmarking set, the length of the motif indicated for each locus in the PacBio catalog is used. For the vamos set, the median length of the motifs of each locus provided in the original vamos motif set is used. (C) presents a stacked bar plot showing the number of distinct k-mers (k = 10) for each locus in all datasets in GRCh38. These counts are corrected for cyclic shifts.



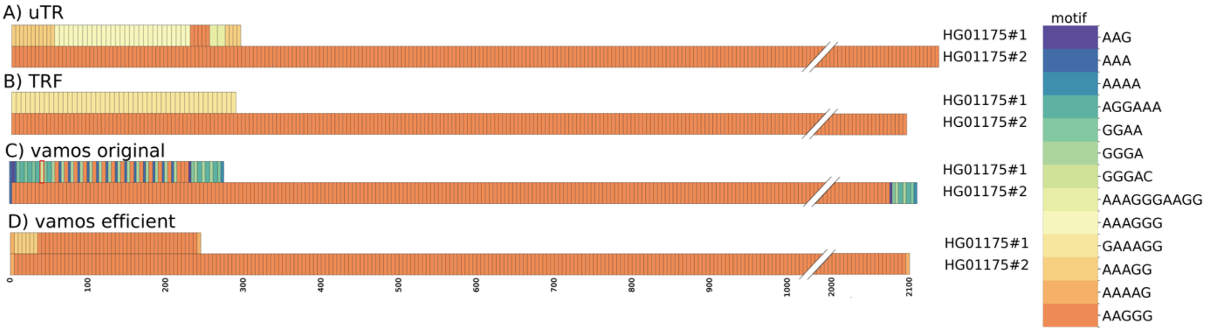
Supplementary Figure 2. The percentage of sequence annotated as single-nucleotide motifs by MotifScope. This figure shows the percentage of single-nucleotide sequence per allele sequence in the benchmarking set of TRs on genome HG002. The percentages are stratified by the number of distinct k-mers (k = 10) of the corresponding sequence, corrected for cyclic shifts.



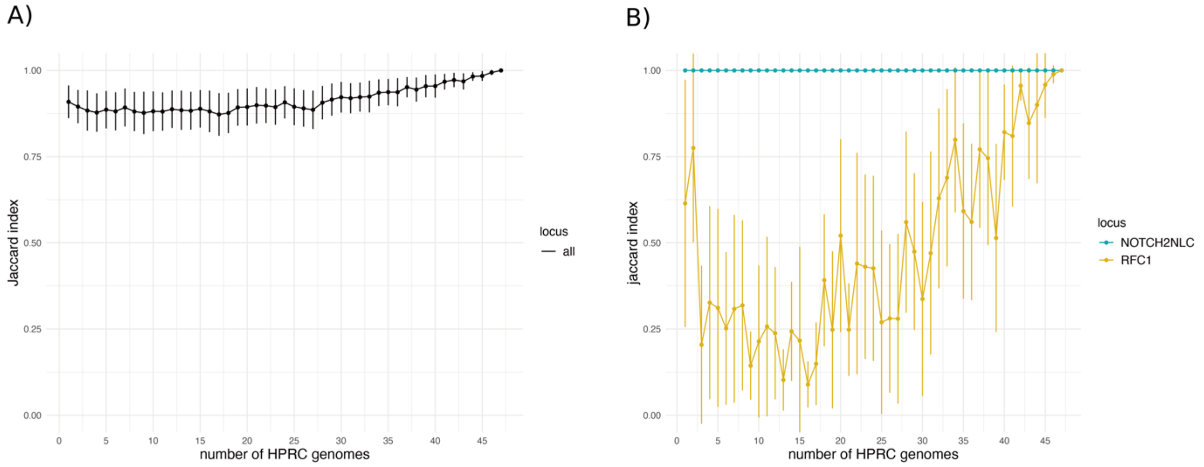
Supplementary Figure 3. Comparative analysis of tandem repeat characterization in vamos repeat catalog in 5000 randomly sampled regions. Four tools were tested in the analysis, MotifScope, TRF, uTR, and vamos. For vamos, here we used both the original motif set (vamos original) and the efficient motif set (vamos efficient). (A) shows the number of motifs discovered by each of the four tools. (B) shows the normalized edit distance between the actual sequence and the results obtained from the four tools, normalized with respect to repeat length. (C) and (D) shows the intersection of motifs between MotifScope and other tools (dots connected by lines below the X axis): for (C), the stacked bar plot at the top shows the fraction of intersected motifs over the total number of motifs found by the other tool; for (D) the stacked bar plot at the top shows the fraction of intersected motifs over the total number of motifs found by MotifScope.



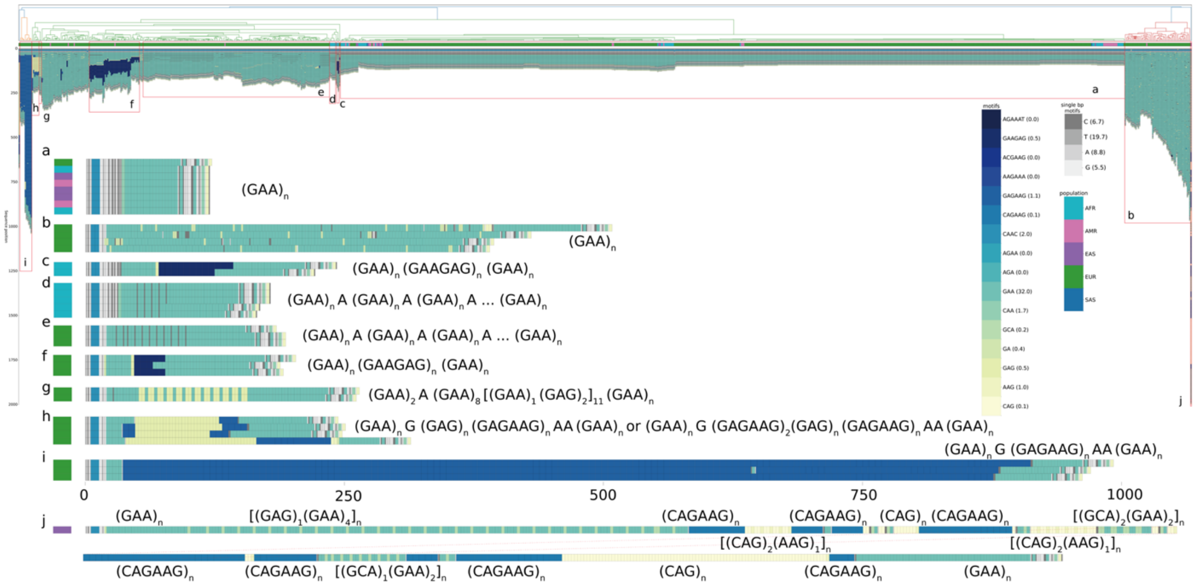
Supplementary Figure 4. Computational performance of MotifScope on pathogenic repeats on HPRC genomes. Total runtime (A) and maximum memory usage (B) analyses are performed on each of the known 48 pathogenic loci provided by PacBio (available here: https://github.com/PacificBiosciences/trgt/blob/main/repeats/pathogenic_repeats.hg38.bed) using varying numbers of genomes: 1 (HG002), 5 (HG002, HG00438, HG005, HG00621, HG00673), 10 (HG002, HG00438, HG005, HG00621, HG00673, HG00733, HG00735, HG00741, HG01071, HG01106), 20 (HG002, HG00438, HG005, HG00621, HG00673, HG00733, HG00735, HG00741, HG01071, HG01106, HG01109, HG01123, HG01175, HG01243, HG01258, HG01358, HG01361, HG01891, HG01928, HG01952), and 47 (all the HPRC genomes). For both runtime and memory usage analyses, MotifScope was run with (parameter “-figure” was set to “True”) and without (parameter “-figure” was set to “False”) generating the figure using the “reads” mode (parameter “--sequence-type” was set to “reads”).



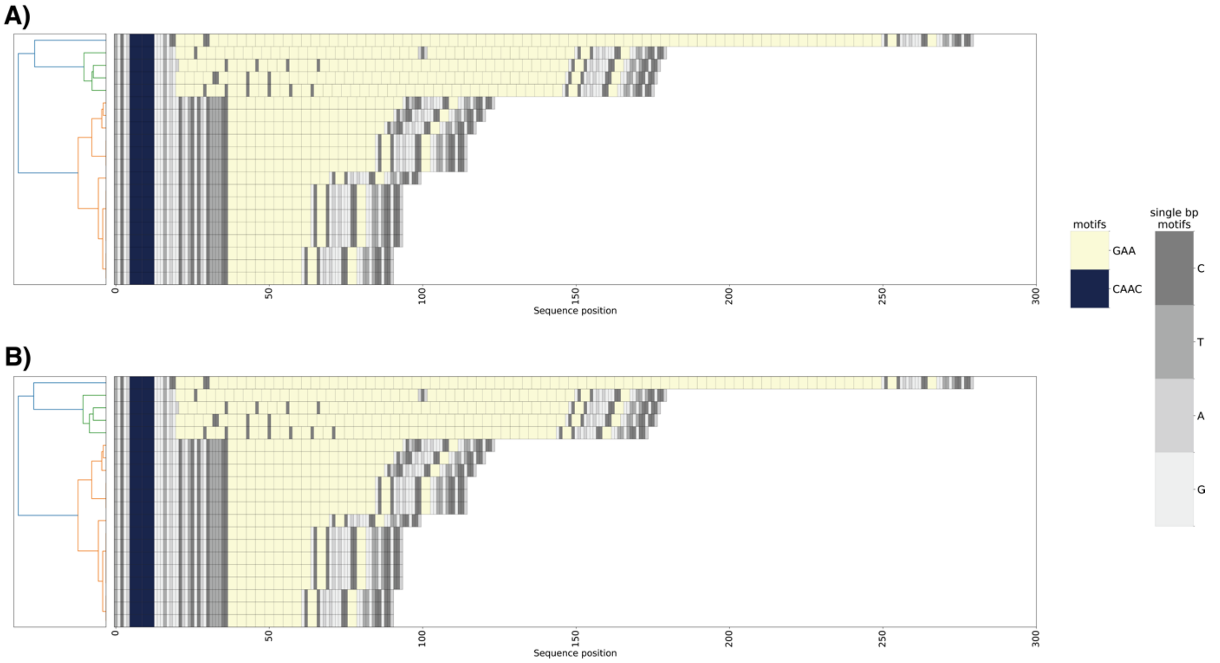
Supplementary Figure 5. Motif characterization of the RFC1 repeat in HG01175 with different tools. The results from (A) uTR, (B) TRF, (C) vamos original and (D) vamos efficient of the RFC1 repeat in the HG01175 assembly. The “GACGG” motif is highlighted in a red box.



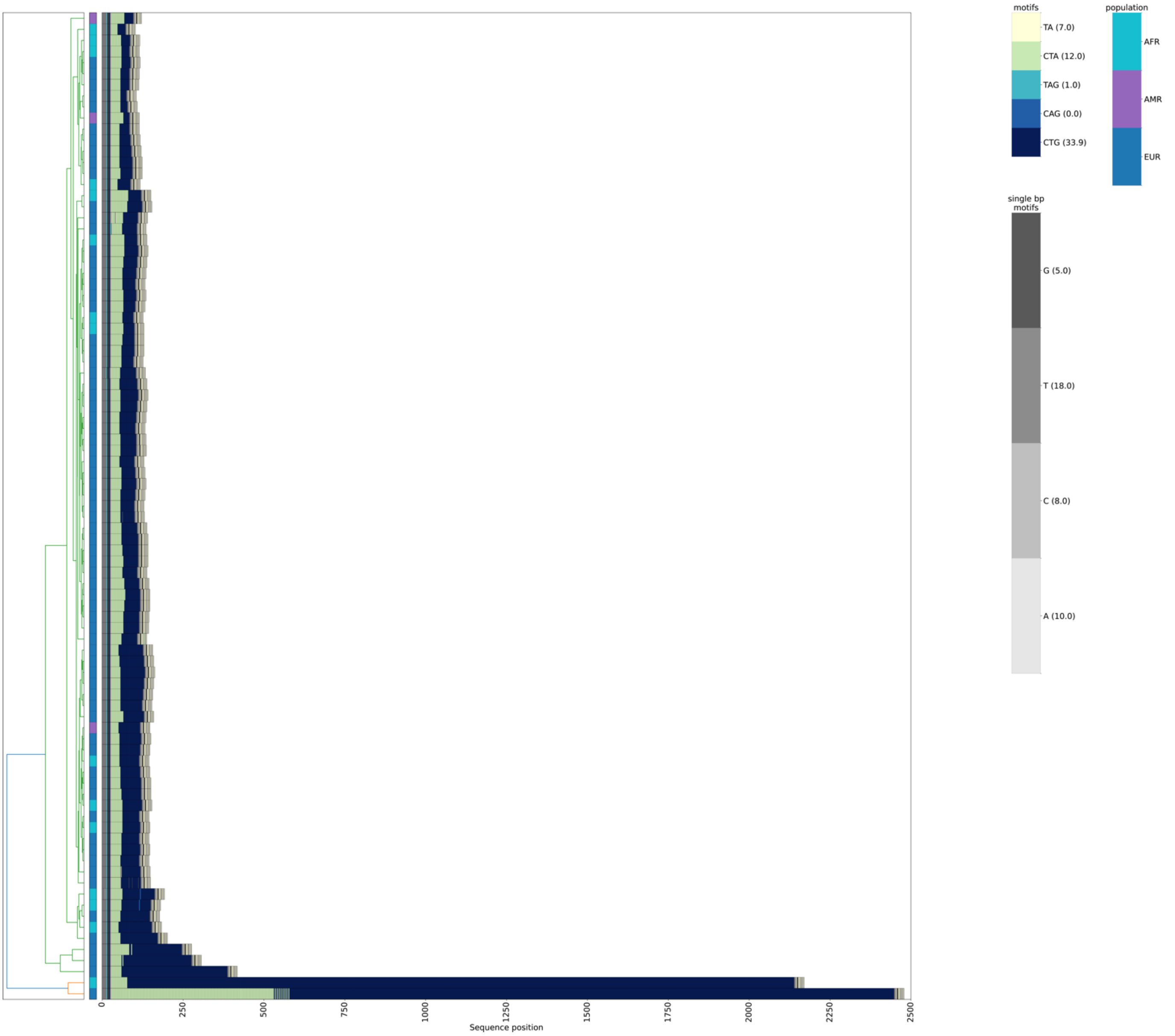
Supplementary Figure 6. Stability of motifs identified by MotifScope on pathogenic repeats on HPRC genomes. In (A), the 56 known pathogenic loci from PacBio data were used. For each locus, two equal-sized subsets of genomes (ranging from 1 to 47 samples each) were randomly sampled from the HPRC dataset. The Jaccard index for motifs identified in these subsets (excluding single nucleotide motifs) was calculated directly from MotifScope's output. This analysis was repeated ten times. (B) shows the distribution of the mean Jaccard index for each locus is shown with solid lines, while the blue and yellow lines represent the mean Jaccard index for the NOTCH2NLC TR and RFC1 TR loci, respectively.



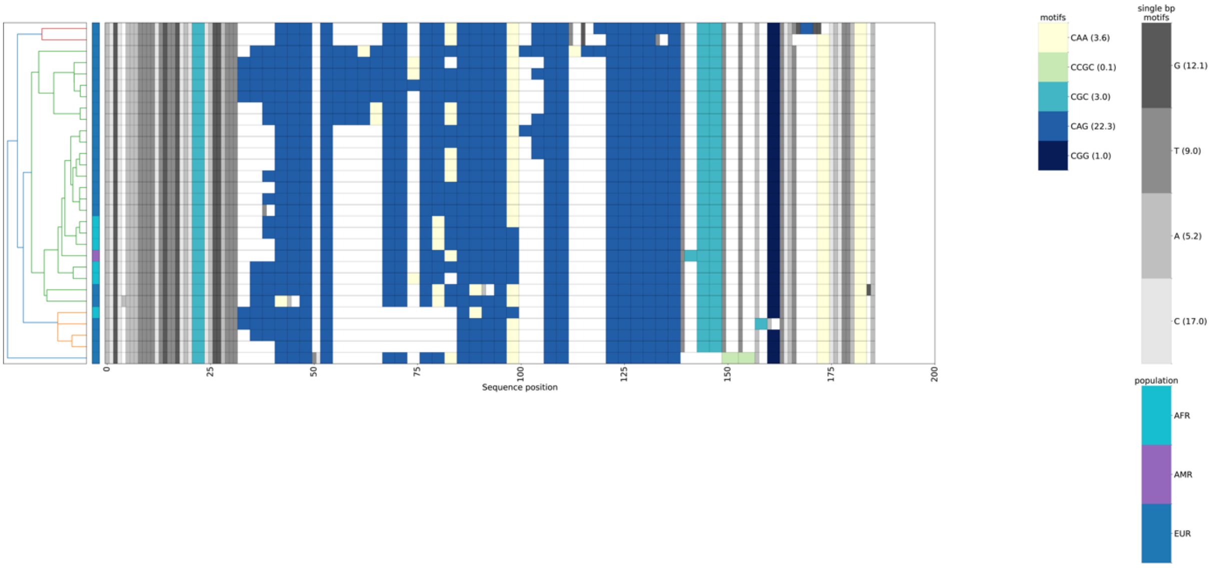
Supplementary Figure 7. FGF14 repeat in HPRC, Dutch AD patients and centenarians. The assembly sequences on the FGF14 repeat with 31 bp flanking the repeat, of 47 HPRC samples, 250 Dutch AD patients and 251 Dutch centenarians. The top figure showed the result from MotifScope. The alleles below showed the zoomed in of different alleles that are present in these assemblies with the left column indicating the corresponding population (second-to-right color bar, SAS: South Asian; EUR: European; EAS: East Asian; AMR: Admixed American; AFR: African) of the assembly and the right figure showing the motif structure of these different alleles. The corresponding motif color bar is on the right side of the figure. The number following the motif on the color bar indicates the number of average occurrences of the motif per allele sequence.



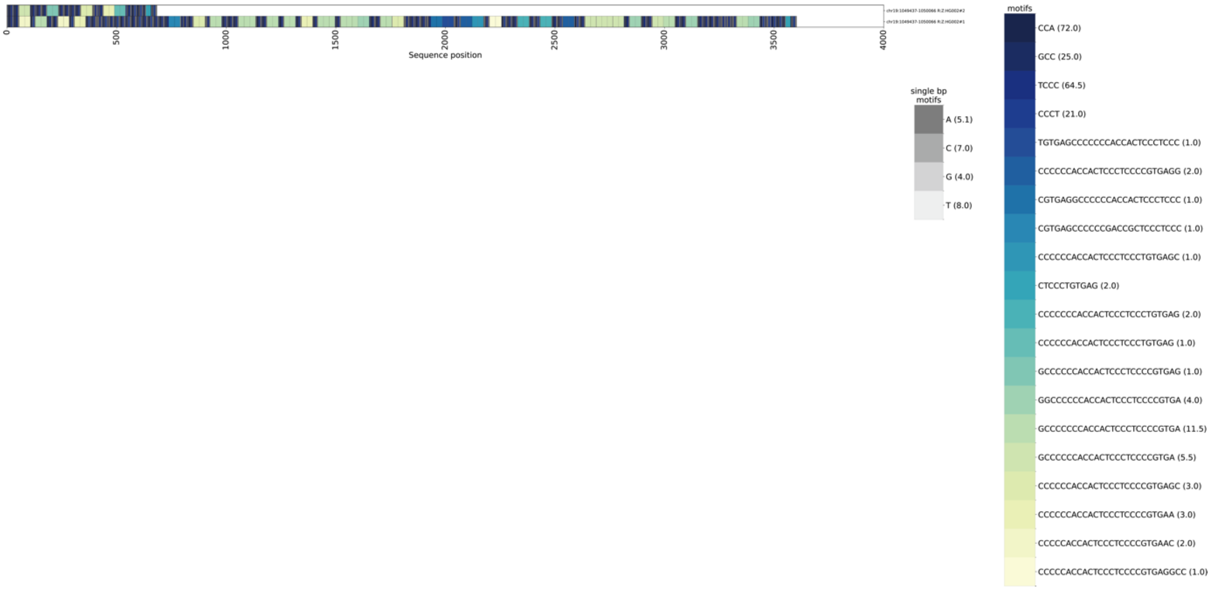
Supplementary Figure 8. Motif characterization with assemblies generated by different assemblers. Assemblies of the FGF14 TR were produced using (A) hifiasm and (B) otter in 10 Dutch centenarians. The TR region is flanked by 30 bp on both sides. In each figure, the left panel shows the clustering of the sequences, while the right panel displays the motif composition of the repeat, with distinct motifs represented by different colors, as indicated by the color bar on the right side of the figure. Only minor differences were observed between the assemblies, confirming that otter (version 1.0) is consistent with a standard tool like hifiasm.



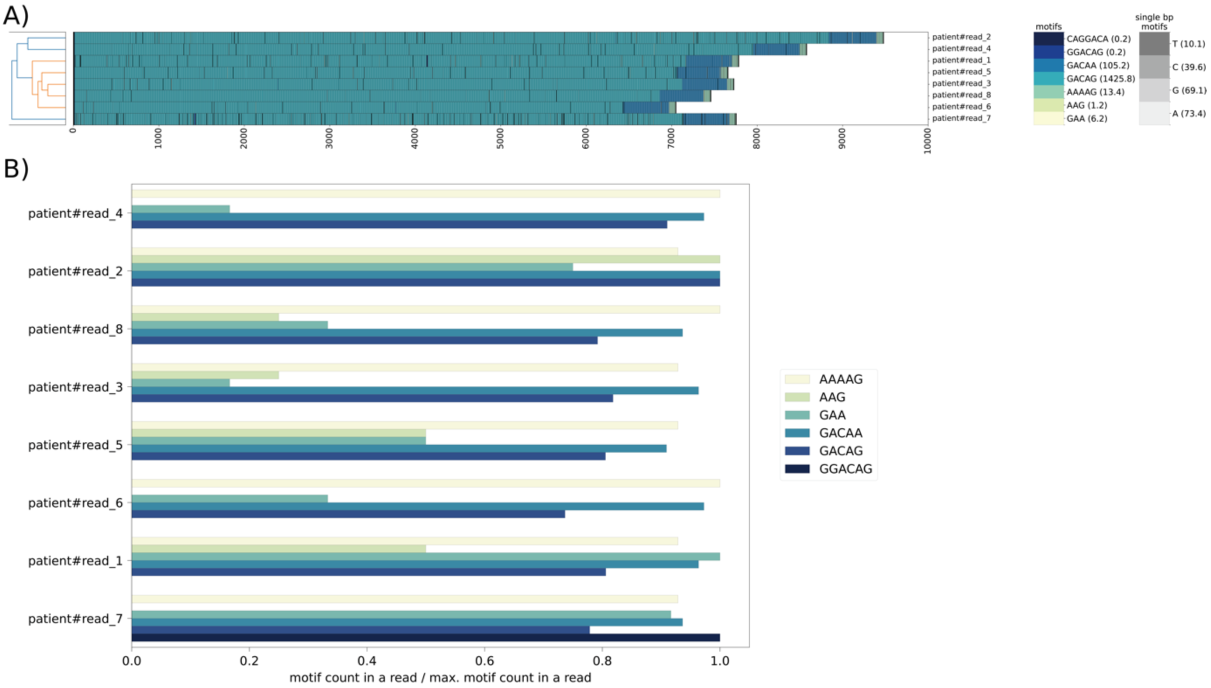
Supplementary Figure 9. ATXN8 repeat in HPRC, Dutch AD patients and centenarians. The unique allele sequences of the assemblies on the ATXN8 repeat with 31 bp flanking the repeat, of 47 HPRC samples, 250 Dutch AD patients and 251 Dutch centenarians. The leftmost panel presents the clustering of assembly sequences, along with population origin of the alleles in the adjacent column. The color code, denoting population origin, is in the second-to-right color bar (SAS: South Asian; EUR: European; EAS: East Asian; AMR: Admixed American; AFR: African). The middle panel visually represents repeat composition, with distinct colors signifying different motifs.

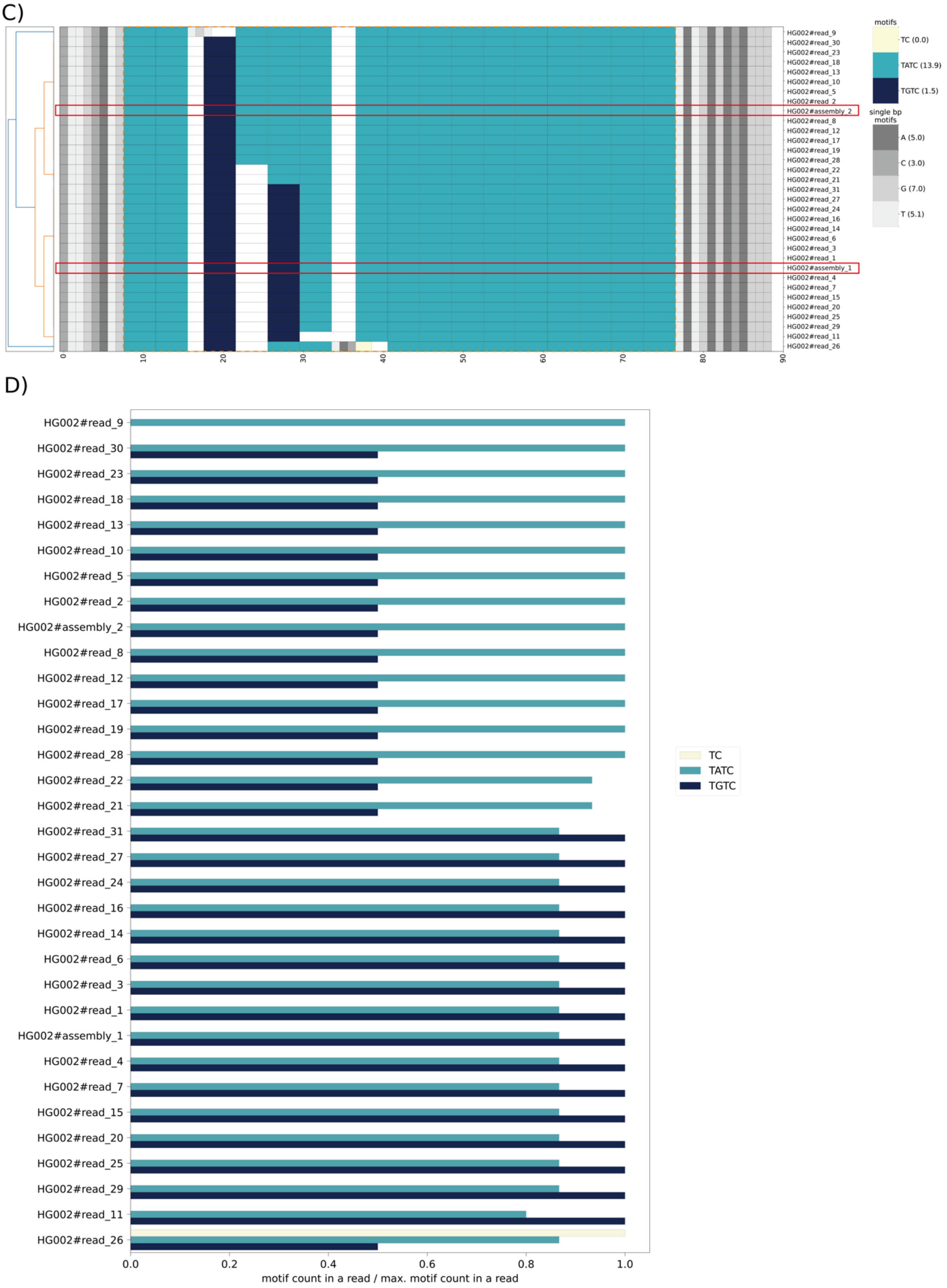


Supplementary Figure 10. ATXN2 repeat in HPRC, Dutch AD patients and centenarians. The unique allele sequences of the assemblies on the ATXN2 repeat with 31 bp flanking the repeat, of 47 HPRC samples, 250 Dutch AD patients and 251 Dutch centenarians. The leftmost panel presents the clustering of assembly sequences, along with population origin of the alleles in the adjacent column. The color code, denoting population origin, is in the second-to-right color bar (SAS: South Asian; EUR: European; EAS: East Asian; AMR: Admixed American; AFR: African). The middle panel visually represents repeat composition, with distinct colors signifying different motifs. The sequences are aligned based on their motif compositions.

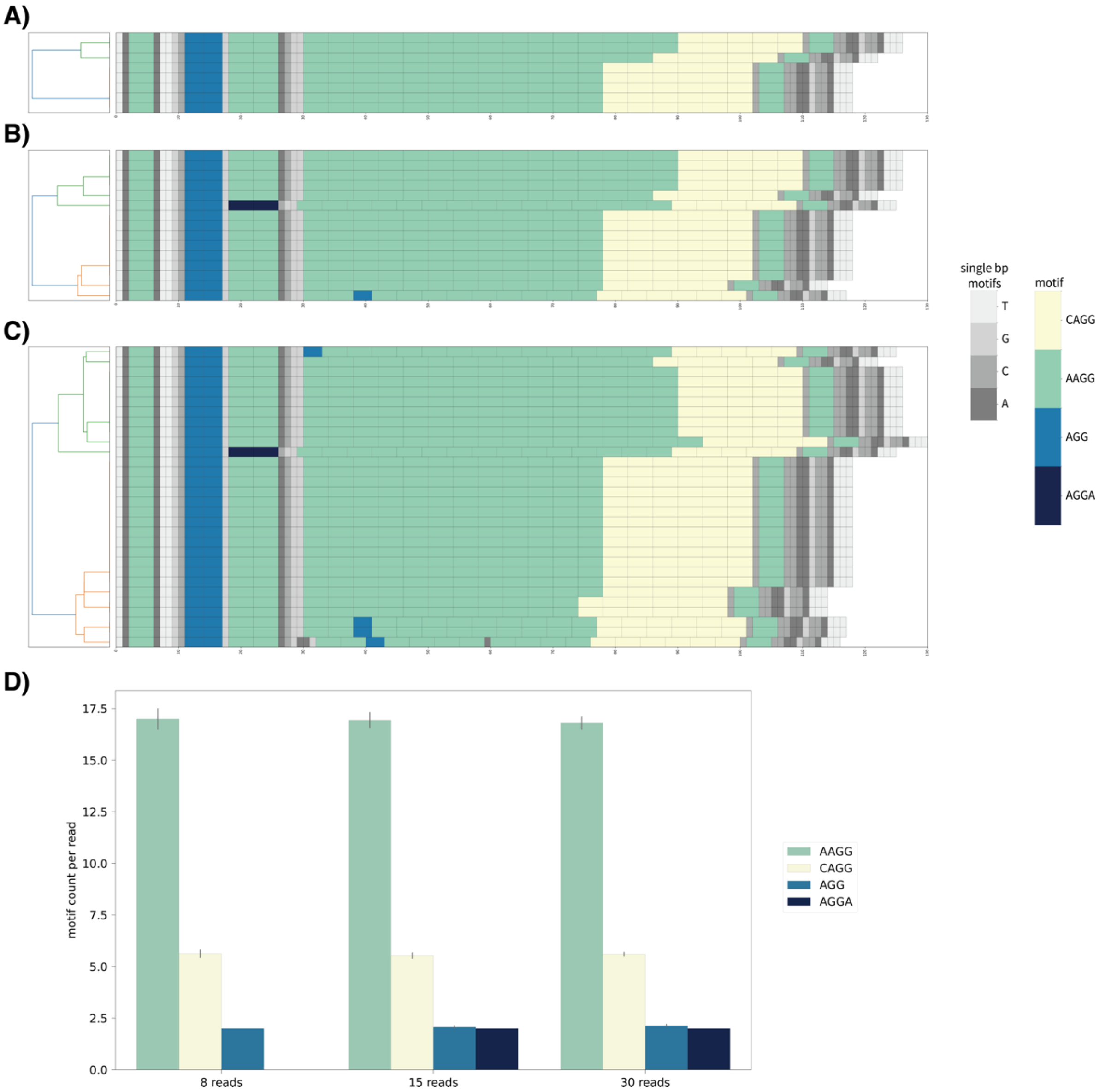


Supplementary Figure 11. Motif characterization of the ABCA7 repeat in HG002 with MoitfScope. This shows the visualization of the decomposing result from MotifScope of the local assembly of the ABCA7 repeat in the HG002 genome assembly, with 10 bp flanking on both sides. Distinct motifs are represented in different colors, as indicated by the color bar on the right side of the figure, followed by the average occurrence of motifs per read in the figure.

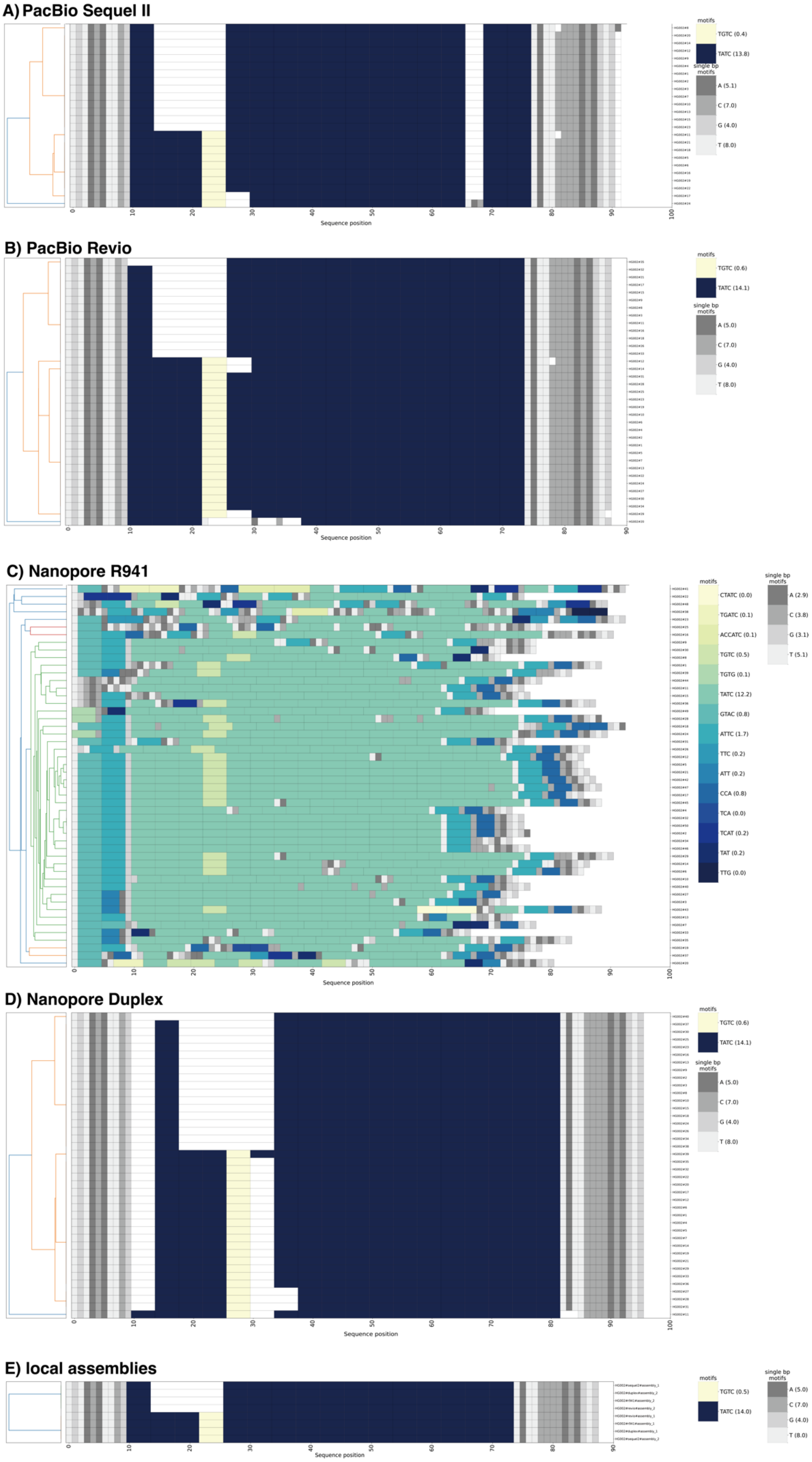

A)
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
patient#read_2
patient#read_4
patient#read_1
patient#read_5
patient#read_3
patient#read_8
patient#read_6
patient#read_7
motifs
CAGGACA (0.2)
GGACAG (0.2)
GACAA (105.2)
GACAG (1425.8)
AAAAG (13.4)
AAG (1.2)
GAA (6.2)
single bp motifs
T (10.1)
C (39.6)
G (69.1)
A (73.4)
B)
patient#read_4
patient#read_2
patient#read_8
patient#read_3
patient#read_5
patient#read_6
patient#read_1
patient#read_7
0.0
0.2
0.4
0.6
0.8
1.0
motif count in a read / max. motif count in a read
AAAAG
AAG
GAA
GACAA
GACAG
GGACAG



Supplementary Figure 12. Motif counts in reads. (A) shows motif characterization of the spanning reads of the RFC1 repeat in the blood of a CANVAS patient. The sequences contain the RFC1 repeat and 10 bp sequences flanking both sides of this region. (B) shows the counts of different motifs in each read. These counts were normalized by dividing by the maximum count of that motif across all reads. (C) shows motif characterization of the spanning reads of forensic loci D3S135. The two assembly alleles of the HG002 genome were highlighted in red boxes and the repeat is highlighted in the orange box with 10 bp flanking this region. The normalized count of different motifs for each spanning read is presented in (D).



Supplementary Figure 13. Motif characterization with varying coverage in D2S1338. A total of 8 (A), 15 (B), and 30 (C) spanning reads from HG002 were randomly selected and characterized by MotifScope using the reference-guided motifs mode. The corresponding motif counts per read are visualized in (D), with error bars representing the standard error.


A) PacBio Sequel II
motifs
TGTC (0.4)
TATC (13.8)
single bp motifs
A (5.1)
C (7.0)
G (4.0)
T (8.0)
Sequence position
B) PacBio Revio
motifs
TGTC (0.6)
TATC (14.1)
single bp motifs
A (5.0)
C (7.0)
G (4.0)
T (8.0)
Sequence position
C) Nanopore R941
motifs
CTATC (0.0)
TGATC (0.1)
ACCATC (0.1)
TGTC (0.5)
TGTG (0.1)
TATC (12.2)
GTAC (0.8)
ATTC (1.7)
TTC (0.2)
ATT (0.2)
CCA (0.8)
TCA (0.0)
TCAT (0.2)
TAT (0.2)
TTG (0.0)
single bp motifs
A (2.9)
C (3.8)
G (3.1)
T (5.1)
Sequence position
D) Nanopore Duplex
motifs
TGTC (0.6)
TATC (14.1)
single bp motifs
A (5.0)
C (7.0)
G (4.0)
T (8.0)
Sequence position
E) local assemblies
motifs
TGTC (0.5)
TATC (14.0)
single bp motifs
A (5.0)
C (7.0)
G (4.0)
T (8.0)
Sequence position

Supplementary Figure 14. Motif characterization of the forensic locus D8S1179 sequenced with different technologies. The spanning reads of this locus in HG002 sequenced with (A) PacBio Sequel II, (B) PacBio Revio, (C) Nanopore R941 chemistry, (D) Nanopore Duplex were extracted with 10 bp flanking this region. (E) shows the local assemblies based on the above sequencing technologies.

Supplementary Algorithm 1:

Input: a set of TR sequences of one region, S, and parameters kmin and kmax, defining the range for screening kmers.

Output: the best kmer to annotate the sequences k_{best}

```
Function SelectBestKmer(S, kmin, kmax)
    Kmer_Counts <- an empty set
    for k <- kmin to kmax do
        kmers <- a set contains all kmers present in S
        foreach kmer in kmers
            Kmer_Counts <- Kmer_Counts ∪ {(kmer, count, k*count)}
    Sorted_Kmers <- Kmer_counts sorted based on k*count in descending order
    max_mask <- 0
    $l_{mcs}$ <- an empty set
    foreach (kmer, count, k*count) in Sorted_Kmers do
            if k*count >  max_mask then
                $l_{mcs}$(kmer)<- the longest length of uninterrupted sequence with kmer (≥ 2 copies of kmer) among S
                max_mask = max($l_{mcs}$(kmer), max($l_{mcs}$))
            else
                break
    $k_{best}$ <- kmer with max_mask
    return $k_{best}$
```

Supplementary Algorithm 2:

Input: a single TR sequence S_j, motif m to annotate the sequences, a set contains all start positions of m in s to be masked, Start_Positions

Output: a set of regions that are annotated with m and the remaining sequences

```
Function MaskMotif($S_j$, m, P)
    Masked_Regions <- an empty set
    foreach position in Start_Positions do
    Masked_Regions <- Masked_Regions $\cup$ {(j, m, position, position + |m|)}
    foreach (j, m, position, position + |m|) in Mask_Regions
        S <- S - $S_j$
        $S_j$ <- the remaining sequence of $S_j$ after masking [position, position + |m|] in $S_j$
        S <- S $\cup$ {$S_j$}
    return Masked_Regions, S
```

Supplementary Algorithm 3:

Input: the iteration number i, the set of previously identified motifs M, the best kmer, kbest, for annotation in this iteration

Output: canonicalized kmer as the candidate motif of this iteration mi

```
Function CanonicalizeKmer(i, M, $k_{best}$)
    if i == 1 then #if it is the first iteration
        $m_i$ <- $k_{best}$
    else
        $K_{best}$ <- {$k_{best}$[i:] + $k_{best}$[:i] for i <- 0 to k}
        Kmer_Alignment_Score <- an empty set of (kmer, Sum_Alignment_Score)
        foreach kmer in $K_{best}$ do
            Sum_Alignment_Score <- 0
            foreach m in M do
                Alignment_Score <- pairwise alignment score between kmer and m
                Sum_Alignment_Score <- Sum_Alignment_Score + Alignment_Score
            Kmer_Alignment_Score <- Kmer_Alignment_Score ∪ {(kmer, Sum_Alignment_Score)}
        Sorted_Alignment <- Kmer_Alignment_Score sorted based on Kmer_Alignment_Score in descending order
        $m_i$ <- Sorted_Alignment[0]
    return $m_i$
```