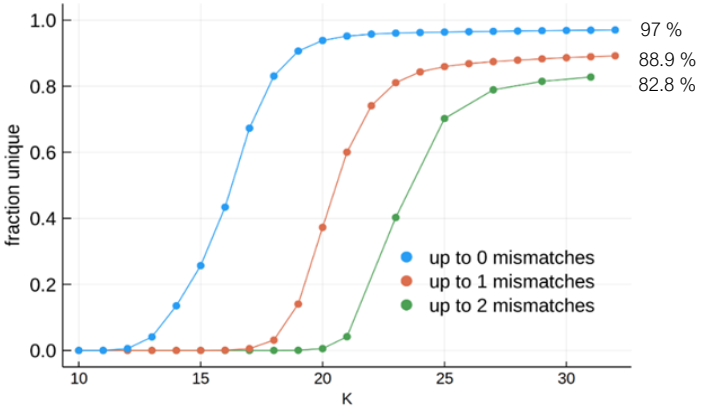
Cell Reports Methods, Volume *3*

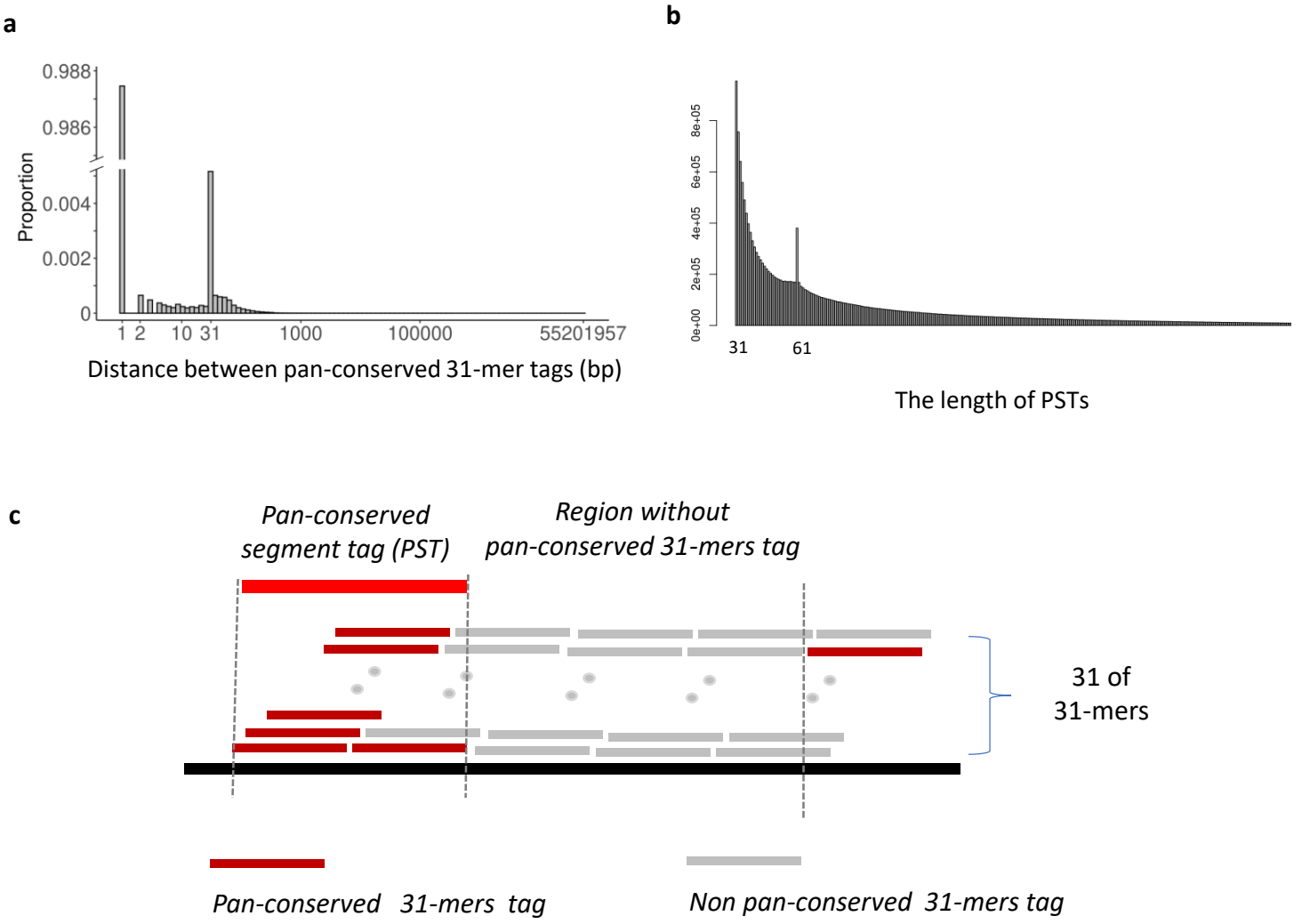
Supplemental information

Pan-conserved segment tags identify ultra-conserved sequences across assemblies in the human pangenome

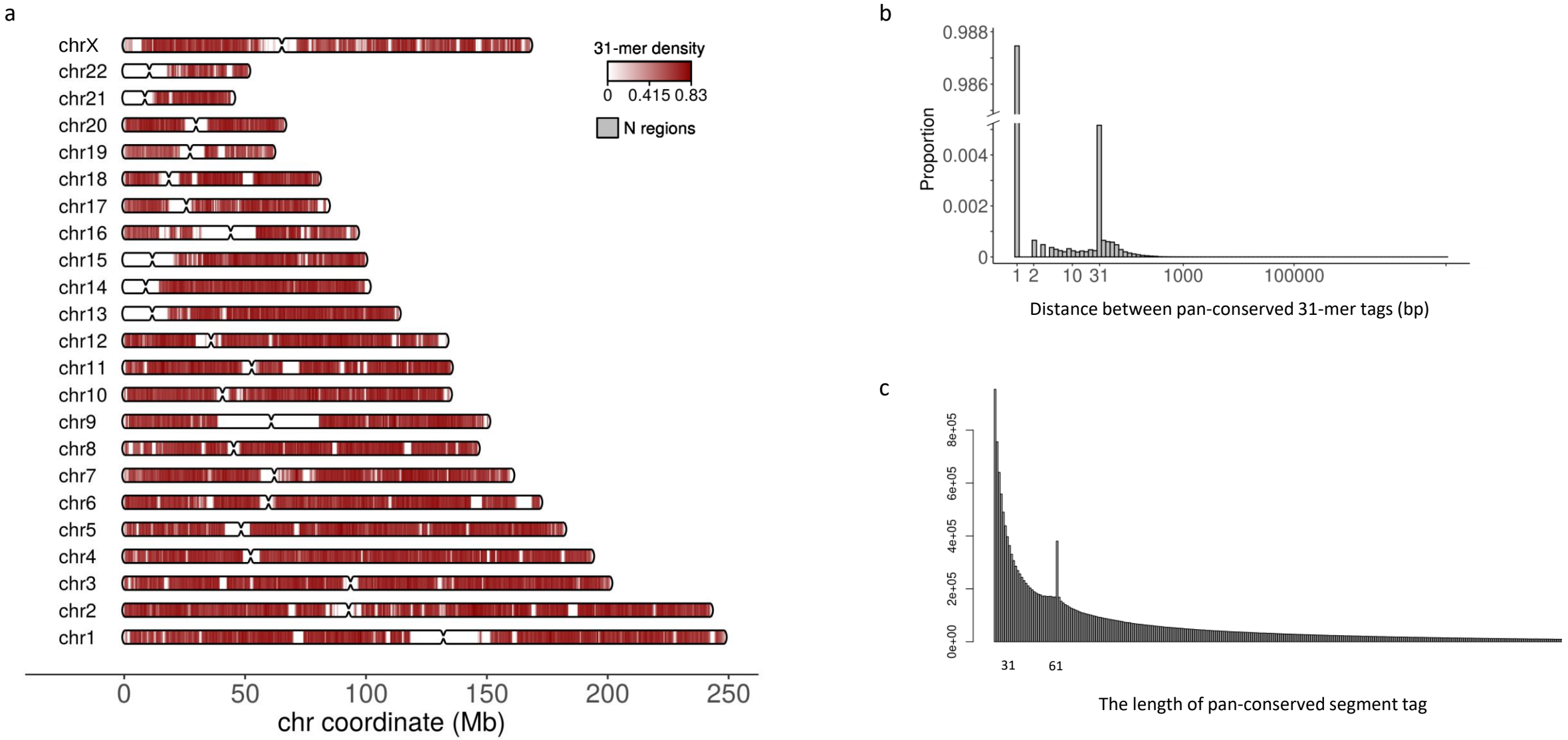
HoJoon Lee, Stephanie U. Greer, Dmitri S. Pavlichin, Bo Zhou, Alexander E. Urban, Tsachy Weissman, Human Pangenome Reference Consortium, and Hanlee P. Ji



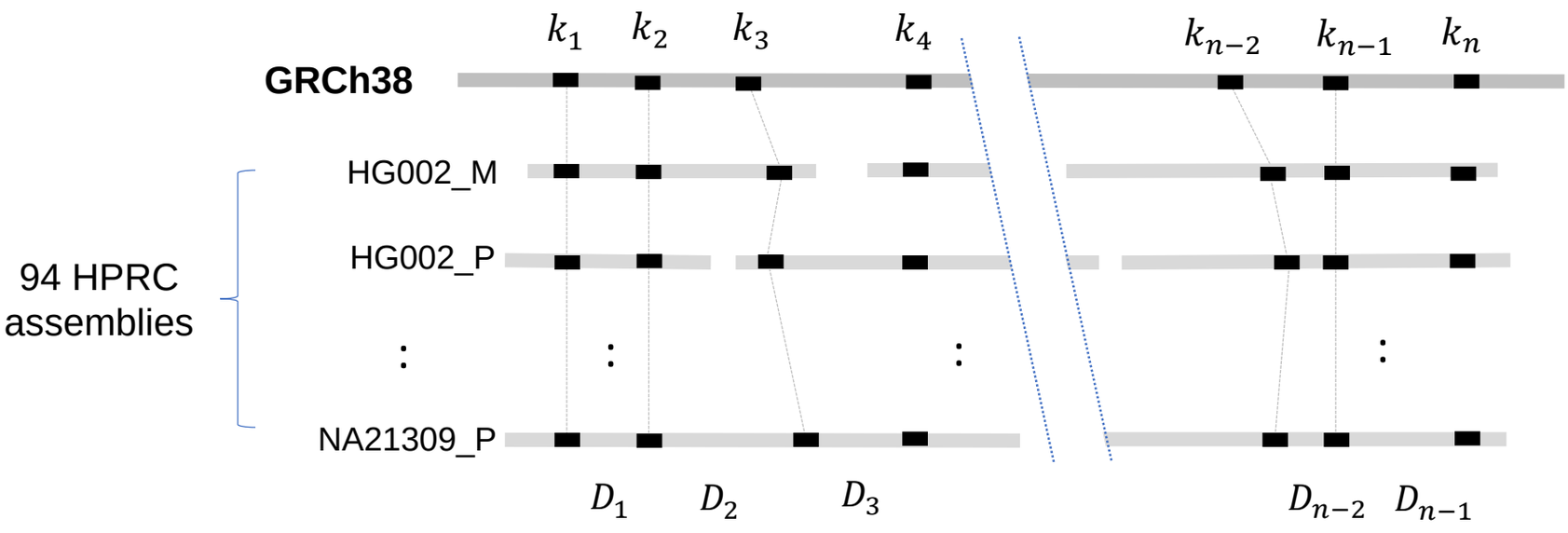
Supplementary Figure 1. The extent of uniqueness by k-mer length within 2 mismatches, related to Figure 1A and STAR Methods



Supplementary Figure 2. The spatial distribution of pan-conserved 31-mers based on GRCH38 coordinates, related to Figure 1A and B. **a**, The proportion of distance between adjacent 31-mer pairs. Most of pan-conserved 31-mer are consecutive and another peak at the distance of 31, which could be explained by the presence of single nucleotide polymorphism (**SNP**) across assemblies. **b**, The distribution of the length of pan-conserved segments. **c**, "**pan-conserved segment tag**", stretch of consecutive pan-conserved 31-mer tags, and regions with depletion of pan-conserved 31-mer tags



Supplementary Figure 3. The spatial distribution of pan-conserved 31-mers based on CHM13 coordinates, related to Figure 1A and B. a, The distribution of pan-conserved segment tags (PSTs) on CHM13. b, The proportion of distance between adjacent 31-mer pairs. c, The distribution of the length of pan-conserved segments.



	D_1	D_2	D_3	- -	D_{n-2}	D_{n-1}
GRCh38	***70***	***332***	***530***	***- -***	***400***	***450***
HG002 maternal	70	460	NA	- -	310	450
HG002 paternal	70	445	500	- -	200	450
:	:	:	:	:	:	:
NA21309 paternal	70	600	460	- -	260	450
Num. same distances	94	82	30	- -	0	94
Num. different distance	0	11	62	- -	94	0
Median differences	0	152	-80	- -	-200	0

Distance matrix

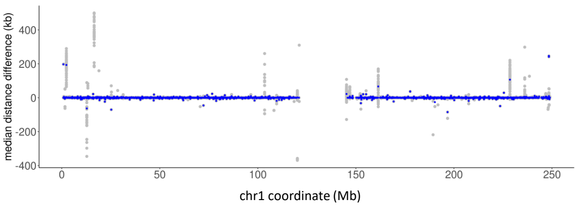
Summary

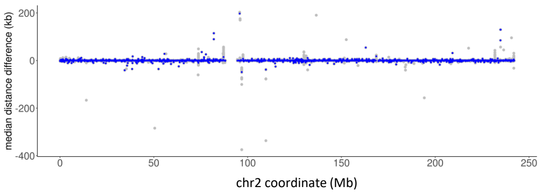
Constant Intervals

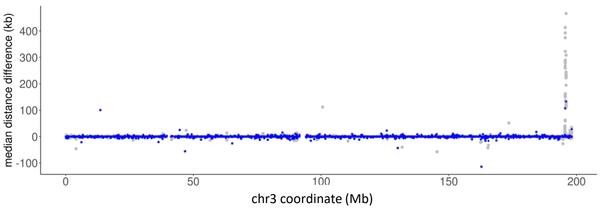
Polymorphic Intervals

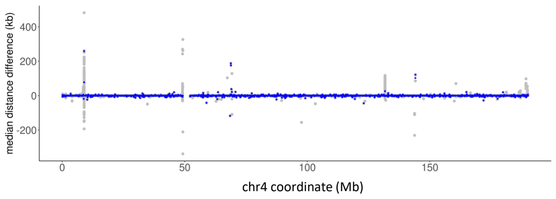
Supplementary Figure 4. The matrix for interval lengths between pan-conserved sequences across all 94 HPRC assemblies, related to Figure 2 and supplementary file 2. The distance between tandem pan-conserved sequences cannot be measured when they are on different contigs, indicated as 3000001 (meaning NA). The transpose of this matrix is provided as supplementary file 2.

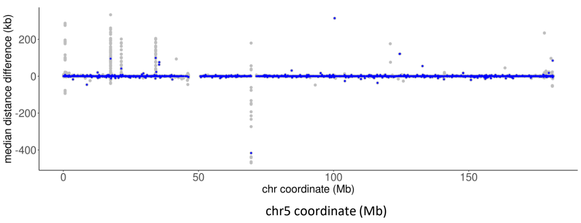
Supplementary Figure 5. Polymorphic intervals on chromosomes, related to Figure 3A

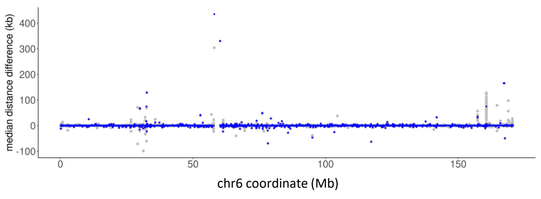


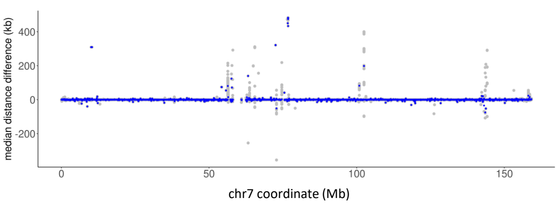


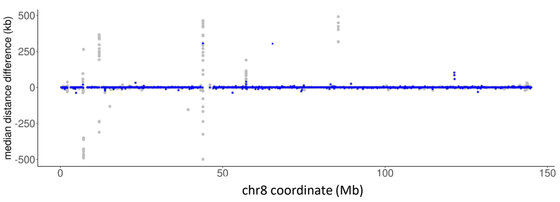


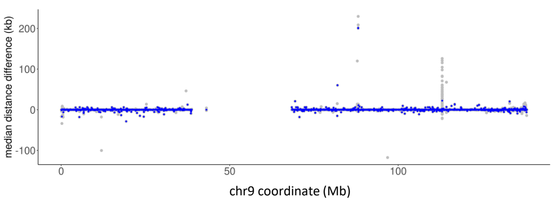


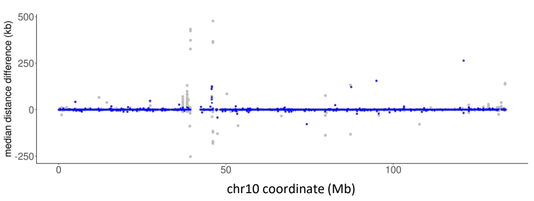


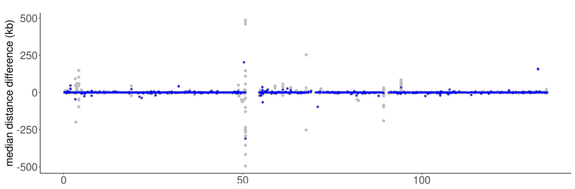


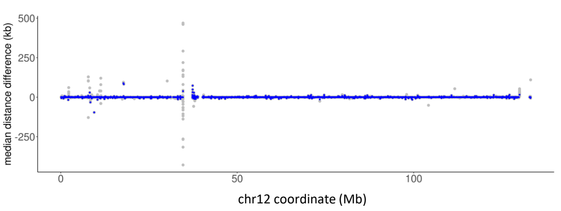


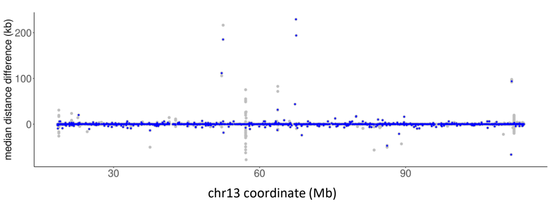


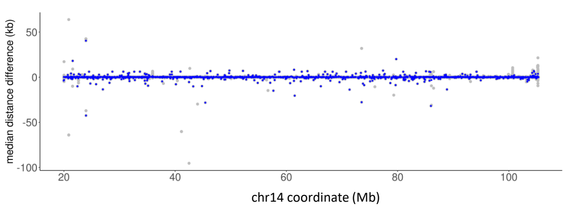


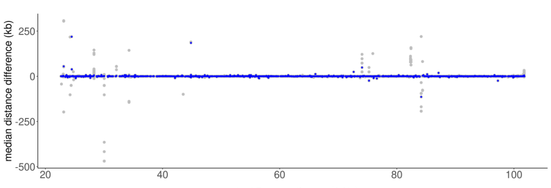


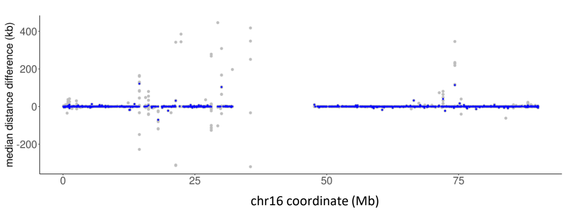


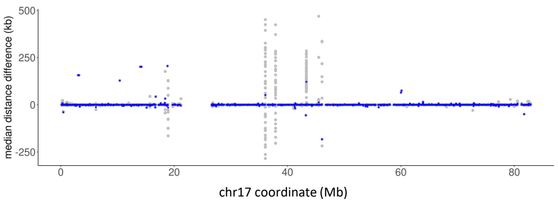


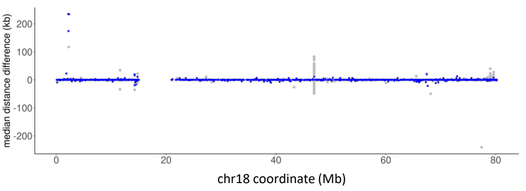


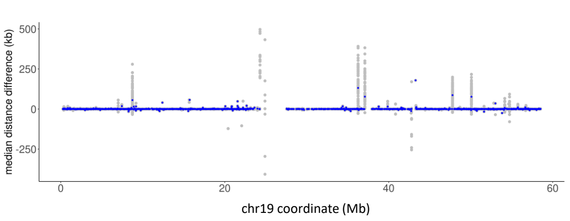


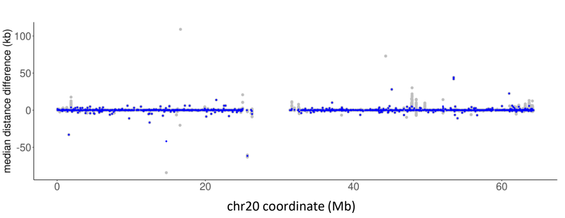


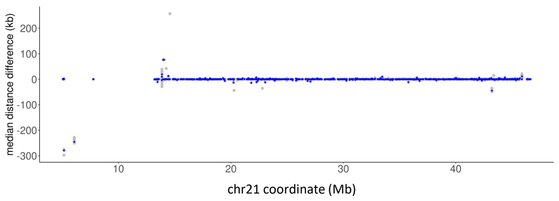


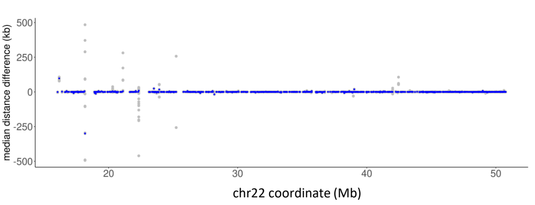


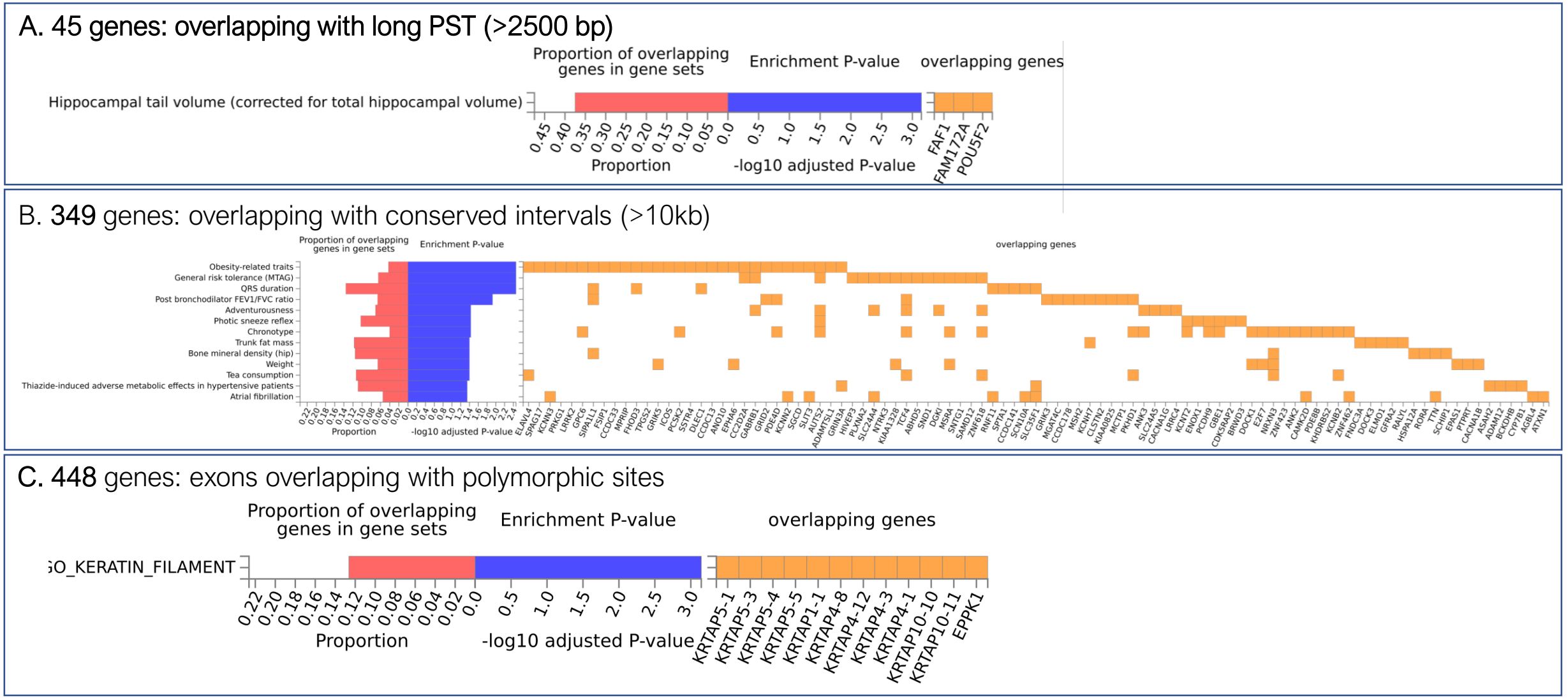






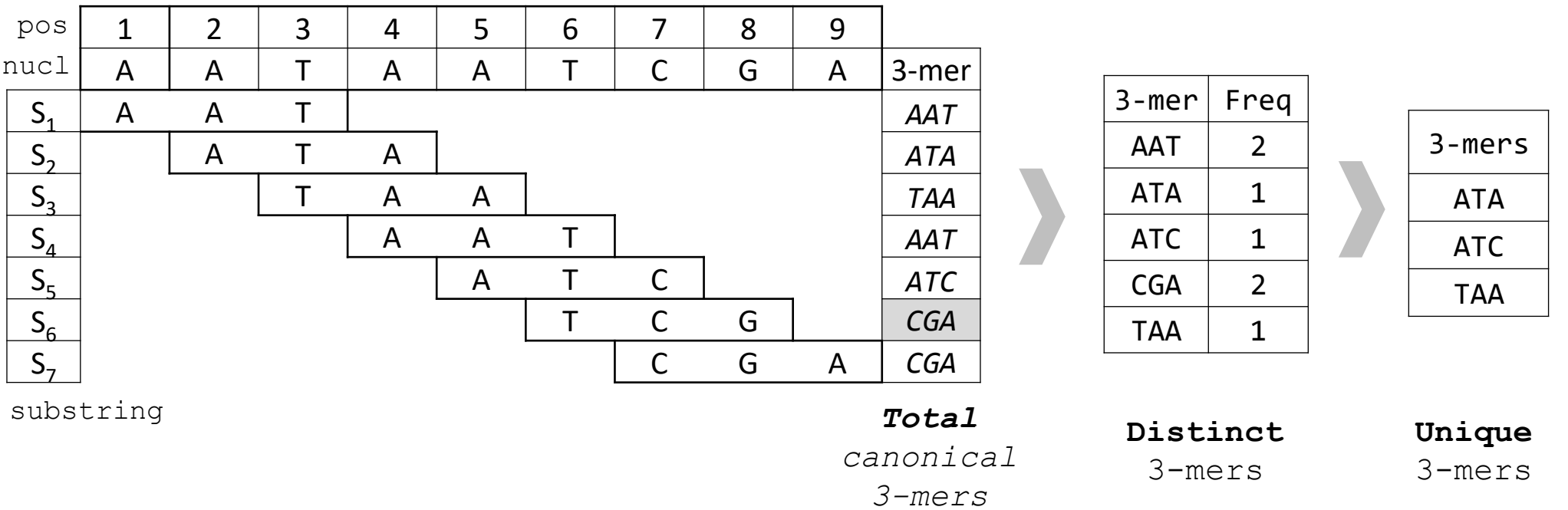




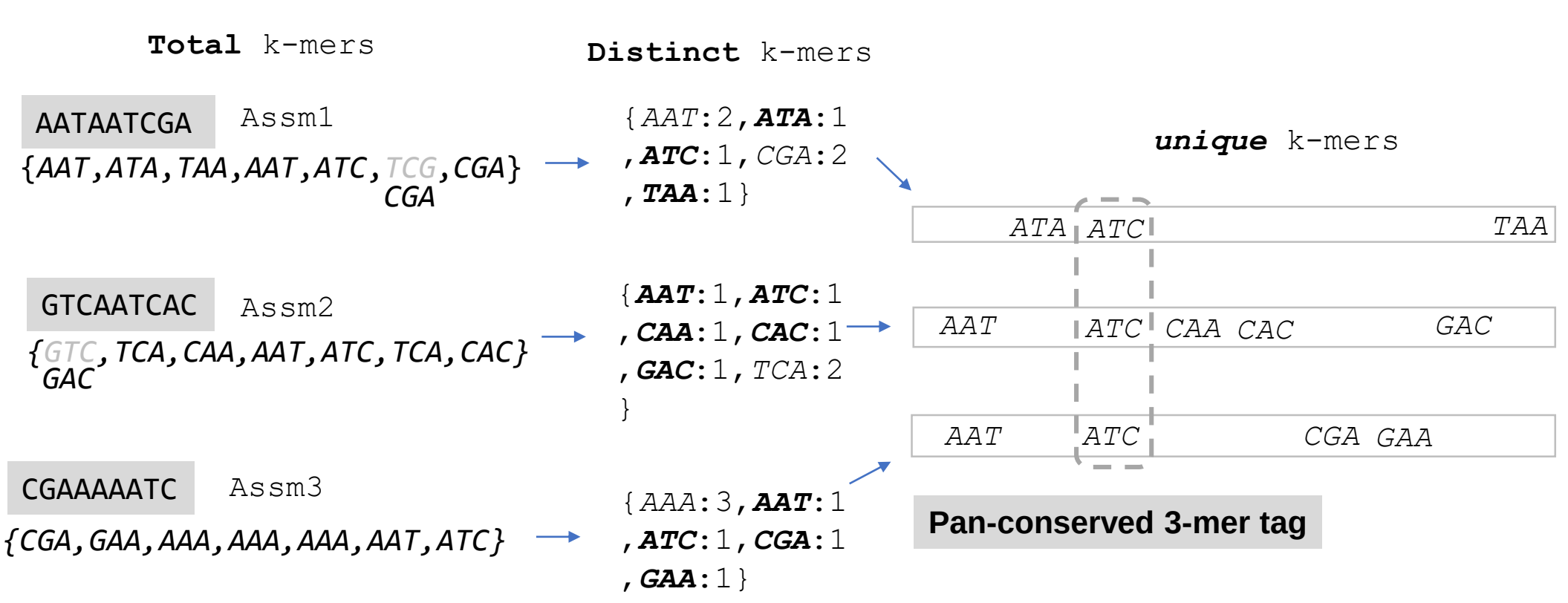


Supplementary Figure 6. Enrichment analysis, related to Figure 1B and Figure 2C, for genes in (A) long PSTs, (B) conserved, and (C) polymorphic intervals.

a



b



Supplementary Figure 7. The indexing of assembly using k-mers, related to STAR Methods. a, Different type of k-mers, b, The identification of pan-conserved segment tag (PST).