Recombination between heterologous human acrocentric chromosomes

Andrea Guarracino[1], Silvia Buonaiuto[2], Tamara Potapova[3], Arang Rhie[4], Sergey Koren[4], Boris Rubinstein[3], Christian Fischer[5], Human Pangenome Reference Consortium[#], Jennifer L. Gerton[3], Adam M. Phillippy[4], Vincenza Colonna[2,5], and Erik Garrison[5,+]

A complete list of authors appears at the end of the manuscript
1 Genomics Research Centre, Human Technopole, Viale Rita Levi-Montalcini 1, Milan, 20157, Italy
2 Institute of Genetics and Biophysics, National Research Council, Naples 80111, Italy
3 Stowers Institute for Medical Research, Kansas City, MO 64110, USA
4 Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA
5 Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN 38163, USA
+ Corresponding author. Email: egarris5@uthsc.edu

Abstract

The short arms of the human acrocentric chromosomes 13, 14, 15, 21, and 22 share large homologous regions, including the ribosomal DNA repeats and extended segmental duplications (1,2). While the complete assembly of these regions in the Telomere-to-Telomere consortium's (T2T) CHM13 provided a model of their homology (3), it remained unclear if these patterns were ancestral or maintained by ongoing recombination exchange. Here, we use pangenomic resources and methods to show that specific pseudo-homologous regions of the acrocentrics recombine. Considering an all-to-all comparison of the high-quality human pangenome from the Human Pangenome Reference Consortium (HPRC) (4), we find that contigs from all of the acrocentric short arms form a community similar to those formed by single chromosomes or the sex chromosome pair. A variation graph (5) constructed from centromere-spanning acrocentric contigs indicates the presence of pseudo-homologous regions where most contigs appear as recombinants of heterologous CHM13 acrocentrics. A diploid T2T assembly of a target sample cross-validates these patterns (6). On chromosomes 13, 14, 21, and 22, we observe lower levels of linkage disequilibrium in pseudo-homologous regions than in their short and long arms, indicating higher rates of recombination and/or a larger effective population size (7). The ubiquity of signals of heterologous recombination seen in the HPRC draft pangenome's acrocentric assemblies suggests that this phenomenon is a basic feature of human cellular biology, providing sequence and population-based confirmation of hypotheses first developed cytogenetically fifty years ago (8).

Introduction

Although the human reference genome is now 22 years old (9), fundamental limitations of the bacterial artificial chromosome libraries on which it was built prevented its completion. Incomplete regions amount to 8% of the Genome Reference Consortium's Human Build 38, and include heterochromatic regions in the centromeres and the short arms of the acrocentrics. Advances in long-read DNA sequencing recently enabled the creation of a complete reference assembly—the Telomere-to-Telomere consortium's CHM13 (T2T-CHM13) (3)—from a homozygous human cell line, providing a reference system for these regions for the first time and exposing them to potential study. In parallel, our ongoing work in the Human Pangenome Reference Consortium (HPRC) has yielded 94 haplotype-resolved assemblies for human cell lines (HPRCy1), based on the same Pacific Biosciences circular consensus (HiFi) sequencing that forms the foundation of T2T-CHM13 (4). These resources enable us to characterize patterns of variation in these previously-invisible regions for the first time. In this work, we apply them to study variation in the largest non-centromeric regions made visible in T2T-CHM13: the short arms of the acrocentrics (SAACs).

Most human chromosomes (N=18) are metacentric, with the centromere found in a median position between short ("p") and long ("q") arms, while 5 are acrocentric, and feature one arm that is significantly shorter than the other. The short arms of human acrocentric chromosomes (13p, 14p, 15p, 21p, and 22p) host the nucleolus organizing regions (NOR), the genomic segments that harbor ribosomal DNA (rDNA) genes and that give rise to the interphase nucleoli (10,11). Due to their repetitive nature, rDNA repeat arrays facilitate intramolecular recombination (12). rDNA repeats incur double-strand breaks at a high rate due to transcription-replication conflicts (11). Moreover, rDNA from multiple acrocentric chromosomes can be co-located in nucleoli during interphase, and multiple acrocentric chromosomes often colocalize to a single nucleolus during pachytene (13), when chromosomes synapse and recombine. Because it causes them to occupy the same constrained physical space, the positioning of rDNA-adjacent sequences at the nucleolar periphery could be a driver of genetic exchange between heterologous chromosomes (**Supplementary Note 1**). In line with this, distal and proximal sequences to rDNA repeat arrays are conserved among the acrocentric chromosomes, suggesting that recombination homogenizes them (1,2). Double-stranded breaks within the rDNA recruit homology directed repair systems throughout the cell cycle, even leading to unscheduled DNA synthesis in G1 phase cells where no sister chromatid is available as a repair template (14). Furthermore, experimental evidence indicates the presence of a common subfamily of alpha satellite DNA on the acrocentric chromosomes that provides evidence for an evolutionary process consistent with heterologous recombination (15). Analysis of somatic cell hybrids showed a higher-order repeating structure common to chromosomes 13, 14, and 21, but not to 15 and 22, providing direct evidence for sequence homogenization in this domain among chromosomes 13, 14 and 21 (16). Related alpha-repeat subfamilies are also shared exclusively by chromosomes 13/21 (17–19) and 14/22 (20) pairs.

The T2T-CHM13 reference fully resolves the genomic structure of the short arms of acrocentric chromosomes for the first time, confirming their strong similarity and providing a complete view

of the homologies in this single genome (3). However, T2T-CHM13 does not provide information on how SAACs vary in the human population—we require further studies on multiple genomes to understand if T2T-CHM13's representation of the SAACs is typical. An alignment-based assembly quality evaluation of the HPRCy1 revealed a pattern of putative mis-joins in all assemblies wherein single HPRCy1 contigs would map preferentially to distinct T2T-CHM13 acrocentric chromosomes (4). While suggestive, this analysis is necessarily relative to only a single frame of reference used as target in alignment, leaving open questions as to the relationships between pairs of HPRCy1 haplotypes. A complete study of this region thus requires improvements in both sequence assembly and pangenome analysis to enable an unbiased assessment of its structure and variation in the population. Here, we combine T2T-CHM13 and HPRCy1 assemblies in a reference-free pangenome variation graph (PVG) model of the SAACs. Using this model and other symmetric analyses of T2T-CHM13 and the HPRCy1 assemblies, we establish the first coherent model of population-scale variation in the short arms of the human acrocentrics.

Results

Chromosome community detection

We sought to study the chromosome groupings implied by the homologies found in all 94 of the HPRCy1 pangenome's assemblies. We show that an all-to-all model of homology in HPRCy1's 94 genomes implies "communities", or modular components, that contain single chromosomes or chromosome arms—with the exception of the sex and acrocentric chromosomes. We posit that both the sex and acrocentric chromosome communities are maintained by pseudo-homologous regions shared among heterologs.

Homology mapping

We first used all-to-all mapping to build a reference-free model of homology relationships in the HPRCy1 pangenome. This models the full HPRCy1 as a mapping graph in which nodes are contigs and edges represent mappings between them. To build the HPRCy1 mapping graph, we generated homology mappings based on chains of 50 kbp seeds of 95% average nucleotide identity—which we expect to support homologous recombination (21)—allowing up to (N-1)=93 alternative mappings over any part of each contig. We first applied this process to map all 38,325 HPRCy1 contigs against all others, obtaining mappings for 38,036 of them covering the 99.9% of the total assembly sequence. This indicated that 38,036 / 38,325 or 99.2% of the HPRCy1 assembly contigs are homologous to at least one other contig. Complex tangles in the assembly graphs used to build the HPRCy1 generate short contigs and tend to result in higher rates of error (4). Thus, to simplify later analysis and focus on well-resolved regions of the assemblies, we narrowed our focus to consider only mappings involving the 16,118 contigs at least 1 Mbps long, covering the 98.72% of the total assembly sequence.

We then built a graph where nodes are contigs and edges represent the mappings between them. Edges in this mapping graph have weights equal to the inferred identity multiplied by the length of the mapping. To infer the chromosome represented by each contig, in parallel, we mapped all contigs against both T2T-CHM13 and GRCh38 references and assigned them a chromosome identity based on this mapping. To ease visualization, we examined a subset of the full graph containing only the best 3 mappings per contig segment, labeling each contig with its reference-relative assignment (**Figure 1A**). The shape of the acrocentric cluster (**Figure 1B**) matches our prior expectations of higher similarity between chromosomes 13/21 and 14/22 (17–20). A region of the 2D graph layout contains a mixture of contigs which individually best-match different T2T-CHM13 or GRCh38 reference chromosomes.

Community detection on the homology mapping graph

To quantify the significance of these patterns, we then applied a community detection algorithm to the full mapping graph. The algorithm assigns each contig to a community such that the total assignment maximizes modularity, which can be understood as the density of (weighted) links inside communities compared to links between communities (22). This process yielded 31 communities (**Supplementary Table 1**). We hypothesized that each cluster represented one chromosome or chromosome arm. Around half of the chromosomes (N=11) were each represented by a single community. Chromosomes 1, 2, 3, 6, and 18 were each represented in two communities corresponding to their short and long arms, likely due to frequent assembly breaks across their centromeres in the HPRCy1 (**Figure 1C-1D**). Contigs from chromosome X and Y fell in the same community, although the short arm of chromosome X was represented in two communities (**Figure 1D**). The SAACs formed the community with the most distinct chromosomes and most contigs (1706 contigs containing 3.91% of HPRCy1 sequence), composed of contigs belonging to the short arms of all the five acrocentrics plus 21q and 22q (**Figure 1C-1D**). 13p, 14p, and 15p each had their own community. We obtained similar results with homology mappings based on chains of 10 kbp seeds of 95% average nucleotide identity, suggesting that the recovered communities are stable relative to basic changes in homology mapping parameters (**Supplementary Figure 1**).

In the homology mapping graph of the HPRCy1, only the acrocentric and sex chromosomes form combined communities containing multiple chromosomes. The sex chromosome community reflects the pseudoautosomal regions (PARs) on X and Y (23), which are telomeric regions where these otherwise non-homologous chromosomes recombine as if they were homologs. We hypothesized that the acrocentric community might also reflect ongoing pseudo-homologous recombination.

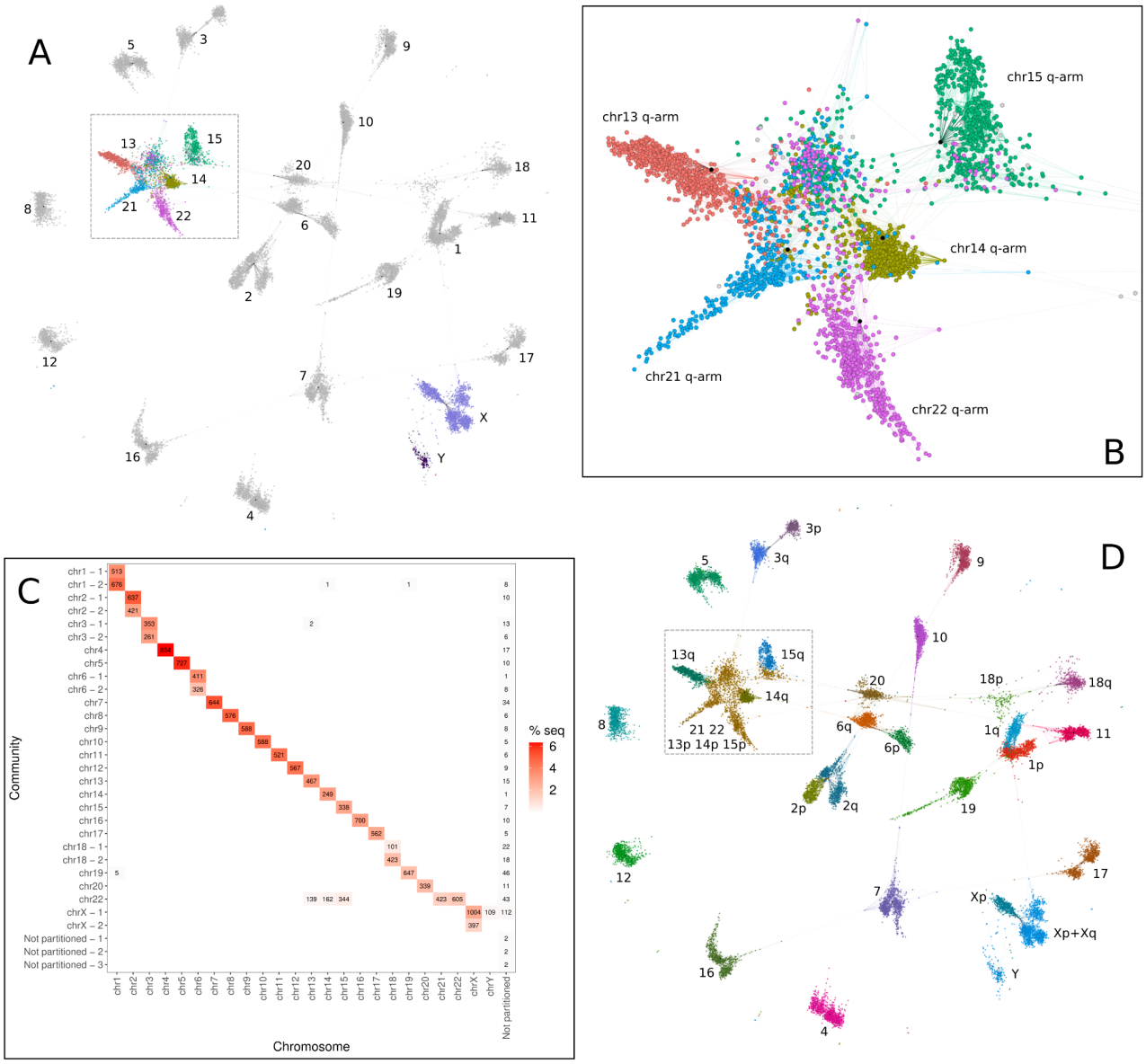

Figure 1. Community detection in the HPRCy1 pangenome. (**A**) The reduced all-to-all mapping graph of HPRCy1 versus itself, with contigs represented as nodes and mappings as edges, rendered in Gephi (24). Colors indicate the acrocentric or sex chromosome to which each contig was assigned by competitive mapping against T2T-CHM13 and GRCh38, with text labels indicating the chromosome for each visual cluster. (**B**) A close-up of the region indicated in A and D containing virtually all contigs matching acrocentric chromosomes. (**C**) Results of community assignment on the mapping graph. The x-axis gives the assigned partition of a contig based on competitive mapping to the chromosomes of T2T-CHM13 and GRCh38, while the y-axis indicates the community, which is named by the chromosome partition that contributes the largest number of contigs to it. The sex and acrocentric chromosomes repsectively participate in the only clusters which mix many (N>100) contigs of different reference-relative partitions. (**D**) Coloring the reduced homology mapping graph in A with community assignments (colors do not correlate with panel A or B). The short arms of 13, 14, 15, and all of 21 and 22 are involved in one community, as are Y and most of X.

An all-acrocentric pangenome graph

To evaluate the hypotheses of heterologous recombination between the SAACs, we built a pangenome graph from acrocentric contigs in the HPRCy1 draft pangenome. These assemblies have high quality, but are nevertheless incomplete and frequently break at centromeres. Thus, to allow their assignment to specific acrocentric chromosomes, we collected long (multi-Mbp) HPRCy1 contigs that span the centromere of an acrocentric chromosome. We then applied the PanGenome Graph Builder (PGGB) (25) to construct a single pangenome variation graph (PVG) from these contigs. Annotation and visualization of the structure of this PVG demonstrates the structure of relationships in the SAACs across the HPRCy1.

Centromere-spanning contigs in the acrocentrics

Due to our interest in relating patterns of variation between acrocentric chromosomes, we required a mechanism to assign contigs to particular T2T-CHM13 chromosomes. We thus focused our analysis on contigs mapping across the acrocentric centromeres (**Figure 2**). To obtain this subset, we mapped HPRCy1 contigs against T2T-CHM13 and considered only those with mappings on both sides of the centromere (±1 Mbp) of the acrocentric chromosomes. We find 132 contigs that meet this requirement. For HG002, we considered contigs from two different versions of its assembly: HG002-HPRCy1 (4), obtained from HiFi reads, and HG002-Verkko, a T2T diploid assembly constructed from both HiFi and ONT reads as described in (**Materials and Methods**) and (6). We later used these two versions of HG002 to provide cross-validation of the patterns observed in HiFi-only assemblies. To support this application, we retain all 50 acrocentric contigs from HG002-HPRCy1, although only 1 meets our centromere-spanning requirements, and all 10 rDNA-proximal assemblies from HG002-Verkko. We additionally include both T2T-CHM13 and GRCh38. We call this sequence set HPRCy1-acro.

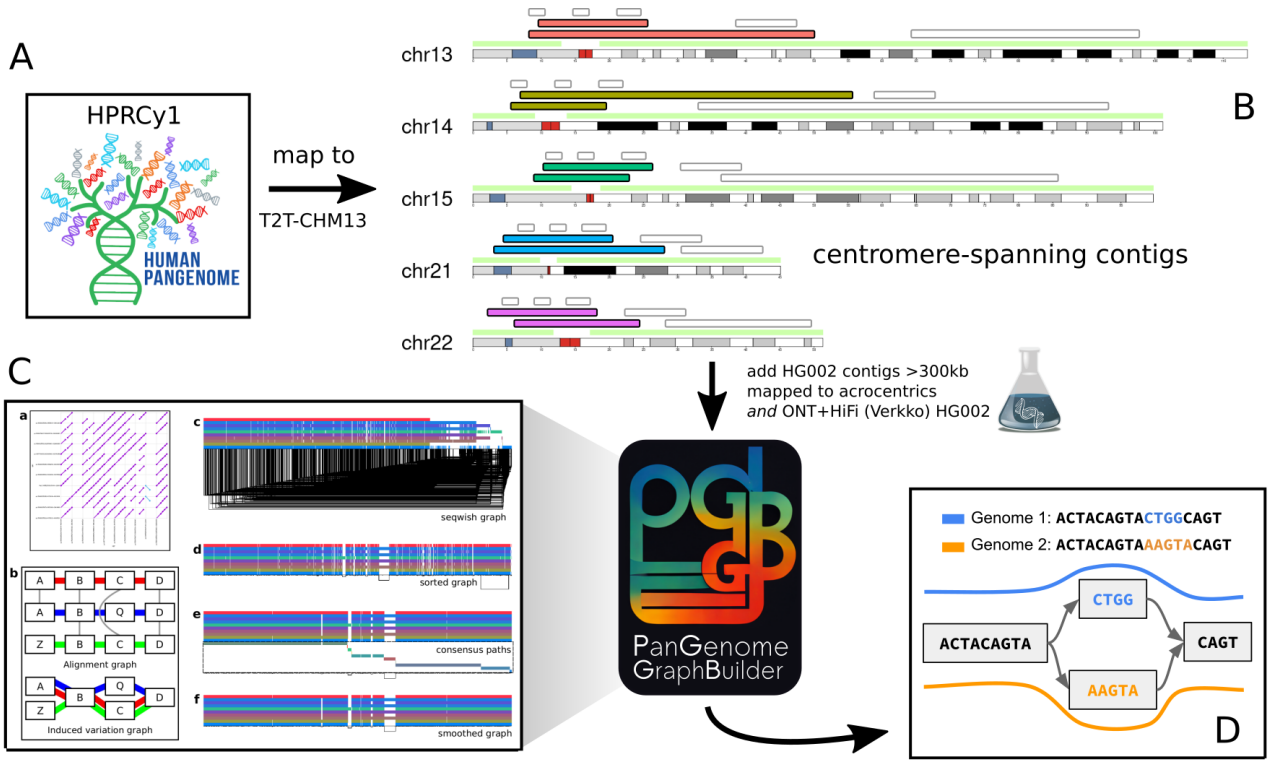

Figure 2. An overview of our approach to build a PVG for HPRCy1 contigs that can be anchored to a specific acrocentric q-arm. (**A**) As input, we take the entire HPRCy1 and map it to T2T-CHM13. (**B**) This yields mappings to acrocentric chromosomes, which we filter to select contigs that map across the centromeres (red cytobands) between non-centromeric regions (over-labeled green). We include HG002 assemblies based on both standard HiFi (from HPRCy1) and on HiFi with Oxford Nanopore scaffolding and tangle resolution (Verkko). (**C**) We then apply PGGB to build a PVG from the HPRCy1-acro collection. PGGB first obtains an all-to-all alignment of the input (C.a.), which is converted to a variation graph with SEQWISH (26) (C.b.), then normalized with sorting and multiple sequence alignment steps in SMOOTHXG (C.c-f). (**D**) The resulting PVG expresses genomes as paths, or walks, through a common sequence graph. This model thus contains all input sequences and their relative alignments to all others—in the example we see a CTGG/AAGTA block substitution between genomes 1 and 2.

Base-pair resolution of acrocentric homologies in a pangenome graph

Pangenomes unite sequences, their homology relationships, and variation in a single system (4,5,27). Pangenome variation graphs (PVGs) represent identical regions between genomes as nodes, or alleles, while individual DNA sequences are represented as paths through nodes (5). PVGs provide a unified multi-scale perspective on complex patterns of variation, including both single nucleotide polymorphisms (SNPs) and structural variants in the same model. Assembly of the SAACs remains challenging, but we hypothesized that a PVG of a subset of assemblies that propose contiguous solutions to the SAACs would provide insight into the aggregate structure of inter-acrocentric sequence variation and homology among HPRCy1 samples.

PVGs can be constructed by performing pairwise alignments between a collection of genomes (26). However, due to the high cost of the all-to-all alignment problem, current approaches for building pangenome graphs are progressive, and construct the graph by serial or phylogeny-guided incremental inclusion of genomes in the PVG (28,29). This introduces bias in the resulting alignment models, in that the choice of reference or genome inclusion order will change the topology of the resulting PVG, a significant problem when approaching complex loci like the SAACs. PGGB specifically addresses the issue of reference bias by building the graph out of symmetric all-to-all alignment between pangenome sequences (**Figure 2C**), a process enabled by the low cost of alignment with the Wavefront algorithm (30,31).

We applied PGGB to construct a PVG representing all pairwise alignments between contigs in HPRCy1-acro. The resulting acrocentric PVG (acro-PVG) presents structures that echo those observed in T2T-CHM13 and the community structure of the homology mapping graph (**Figure 3A**). In more detail, the main connected component including all chromosomes presented a tangled region, anchored at the ribosomal DNA repeats, but also extending towards the centromere-proximal end of the short arms. The alpha satellite higher-order repeats (HORs) in the centromeres of chromosomes 13/21 and 14/22 pairs shared high similarity within each pair (19,20), leading to collapsed motifs in the graph (**Figure 3B**). 13/21 and 14/22 diverge in centromere-proximal regions of the q-arms. Furthermore, a single region in the pangenome graph centered on the GC-rich SST1 array was present in a single copy in chromosomes 13, 14, and 21, indicating a high degree of similarity of the samples in those regions (**Figure 3C** and **Supplementary Figure 2**). This is compatible with the frequent involvement of these chromosomes in Robertsonian translocations (32,33). These graph motifs were also confirmed by building a pangenome graph without embedding the T2T-CHM13 and GRCh38 references (**Supplementary Figures 3 and 4**), indicating that the observed structure is independent of the reference assemblies.

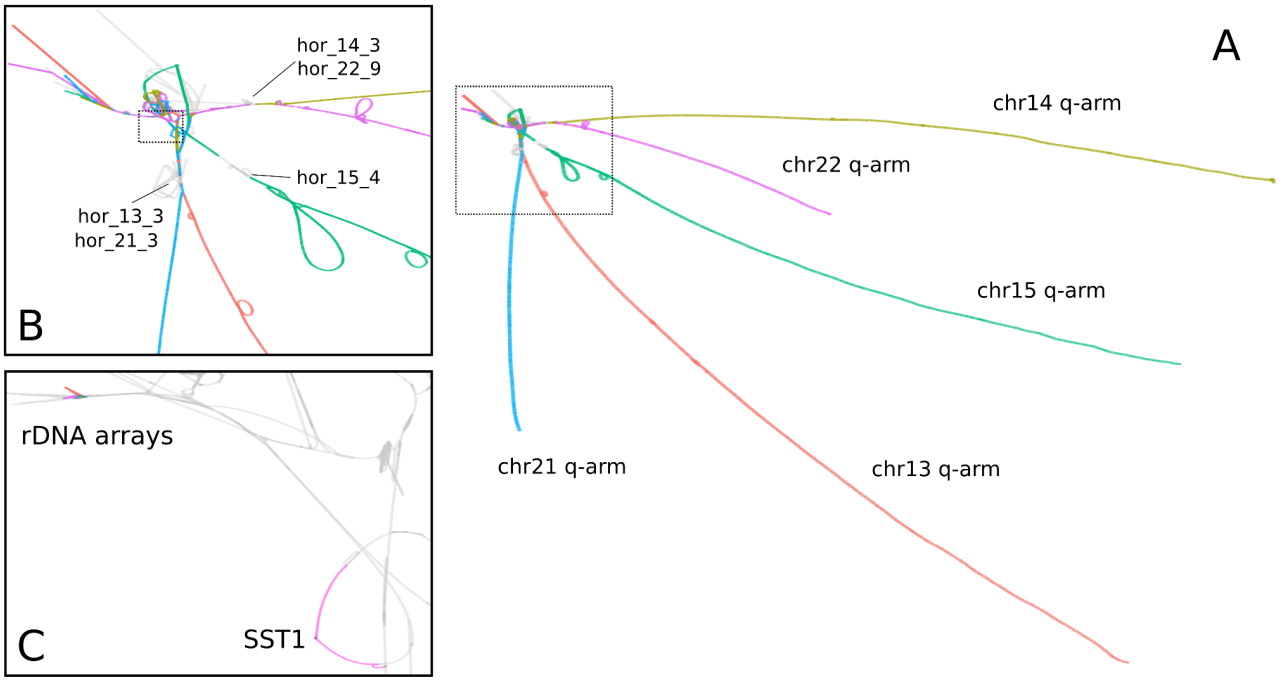

Figure 3. The HPRCy1 acrocentric PVG. We apply GFAESTUS (34) to visualize the 2D layout generated by ODGI (35) in the PGGB pipeline (25). This renders sequences and chains of small variants as linear structures, while repeats caused by segmental duplications, inversions, and other structural variants tend to form loops. (**A**) The full graph contains a dominant connected component, with nodes in T2T-CHM13 chromosomes labeled using colors also used in other figures. The acrocentric q-arms are almost completely separated, while the p-arms unite in a tangle adjacent to the rDNA array. (**B**) A close-up of the SAACs' junction, showing the separation of chromosome 15's centromeric high-order repeats (*hor_15_4*) from the other chromosomes, while 13/21, and 14/22 share substantial homology in their arrays that causes them to collapse in the PVG. A few assemblies span the rDNA array into its distal junction, which presents as a single homologous region across all chromosomes (2), and then fray into diverse sequences visible as tips in the top left. (**C**) Zooming further, we focus on the segmentally duplicated core centered in the SST1 array and the rDNA arrays of the various chromosomes. In this panel, we label the rDNA arrays and SST1 subsequences in T2T-CHM13. The highlighted region around the SST1 array is in direct orientation on T2T-CHM13's 13p11.2, 14p11.2, and 21p11.2 at pairwise identity above 99% (3).

Genomic information flow between heterologous SAACs

The acro-PVG provides a representation of the multiple alignment of SAACs found in the human population. In the acro-PVG, we observe many regions in the graph where multiple T2T-CHM13 chromosomes are aligned together. We expect these regions to potentially support homologous recombination (HR), which largely depends on sequence homology and physical proximity (36) both of which are common among heterologous SAACs (8,13). We hypothesized that sequence exchange patterns would be visible in the PVG as regions where contigs switch the T2T-CHM13 chromosome to which they are most similar.

Untangling the acrocentric PVG

We sought to test the hypothesis that homologous regions of the SAACs feature ongoing sequence exchange by looking for regions in the acrocentric PVG where individual contigs are best-described as recombinants of diverse T2T-CHM13 acrocentric chromosomes. We derived pairwise alignments—a graph *untangling*—from the acro-PVG which map each part (segment) of each HPRCy1-acro contig against T2T-CHM13. The untangling essentially allows us to project the graph into an alignment between a set of query (HPRCy1-acro) and reference (T2T-CHM13) sequences while simultaneously considering all possible alignments represented by the pangenome graph.

In the PVG untangling, each HPRCy1-acro contig is thus represented as a patchwork, or painting, of T2T-CHM13 chromosomes to which it best matches within the pangenome graph. In regions of ongoing recombination between non-homologs, each HPRCy1-acro contig should be painted as a patchwork of diverse T2T-CHM13 chromosomes (7). However, this will not be true if ancient, recombination-inactive homologies link contigs together in the graph—in this case contigs should all match their assigned chromosome.

We queried the PVG (35) to obtain a mapping from segments of all PVG paths onto T2T-CHM13. This segments the graph, and for each HPRCy1 contig (query) subpath through each graph segment, we find the most-similar reference segment (**Supplementary Figure 5**). To reduce the possibility of error, we focused the alignment projection only on the confidently-assembled regions of the HPRCy1-acro contigs (4) and we filtered the mappings to retain only those at greater than 90% estimated identity, removing a total of 1.17 Gbps, or 2.52% of the total SAAC contig segments (**Supplementary Table 2**).

To support a simple reference-relative interpretation of the results, we “grounded” these segments in the reference frame of the T2T-CHM13 chromosome (**Materials and Methods**). This grounding provides a reference-relative aggregation of homology information in the PVG, letting us look inside its structure from the various frames of reference offered by T2T-CHM13’s acrocentric chromosomes. In result, we produced a painting that indicated which T2T-CHM13 chromosome is most similar to each contig in each reference-relative region, which we provide in full for each chromosome (**Supplementary Figures 6-10**).

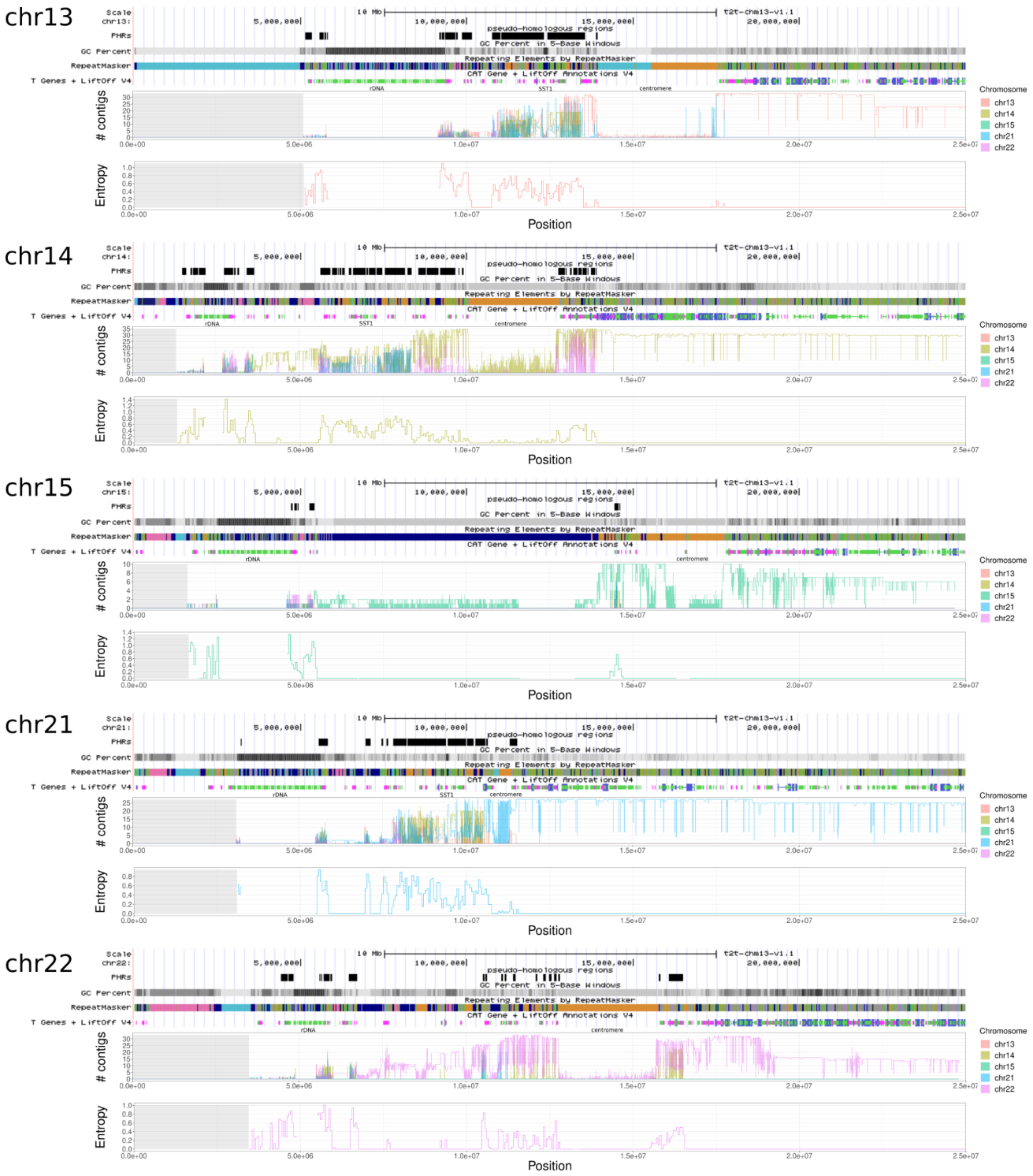

Figure 4. Aggregated untangle results in the SAACs. For each acrocentric chromosome, we show the count of its HPRCy1 q-arm-anchored contigs mapping itself and all other acrocentrics (**# contigs**), as well as the untangle entropy metric (**Entropy**) computed over the contigs' T2T-CHM13-relative untanglings. Gray background indicates regions with missing data due to the lack of non-T2T-CHM13 contigs. We provide the RepeatMasker track legend in **Supplementary Figure 11**.

Aggregating information flow relative to T2T-CHM13

For each chromosome grouping (contigs anchored to a T2T-CHM13 q-arm), over regions in T2T-CHM13, we counted the number of contigs that best-match each one of the acrocentrics inside the PVG (**Figure 4**). We observe a clear pattern in which, on the q-arm, all contigs in each group uniformly best-match their homologous T2T-CHM13 chromosome (**Figure 4 right**). This pattern provides a positive control—we expect it based on our selection of the contigs and the observed structure of the graph (**Figure 3A**). However, as we approach the centromere from the q-arm, we observe regions of homology between 13 and 21—which share approximately 250 kbp in 13q12.11 and 21q11.2, as well as between 14 and 22—which share around 1 Mbp of 14q11.2 and 22q11.21. In contrast, homology with other acrocentrics begins closer to the rDNA in chromosome 15, corroborating the patterns observed in the PVG topology (**Figure 3B**).

Although the 13+21 and 14+22 HORs are both collapsed in the PVG (**Figure 3B**), we observe sparse >90% identity mappings within the centromeres. This is consistent with other reports of high divergence within centromeric satellites (37). HPRCy1 contigs anchored on the q-arms of chromosomes 13, 14, and 21 share a segmental duplication, or homologous region, centered on the SST1 array (**Figure 4**), in line with what is seen in the pangenome graph topology (**Figure 3C and Supplementary Figure 2**). All chromosomes provide similarly good matches for contigs in the regions immediately proximal and distal to the rDNA, however the phenomenon is supported by the relatively few q-arm-anchored contigs (N=9) claim to cross an rDNA array.

To simplify our assessment of homology between the acrocentrics, we developed a single aggregated *diversity entropy* metric which describes the degree of disorder in the untangling of HPRCy1-acro contigs over 50 kbp regions of T2T-CHM13 (**Materials and Methods**). Conceptually, this measures the amount of noise or inconsistency in the untangling over each contig, which we then average across regions of T2T-CHM13. This metric captures the patterns seen in the count-based untangle metric (**Figure 4**), and further demonstrates a clear increasing trend as we progress over each SAAC. The entropy metric begins at 0 in the q-arm (**Figure 4 right**) where by design all contigs in each acrocentric group exclusively match a single T2T-CHM13 chromosome, and becomes higher as we approach the rDNA arrays, reaching a maximum immediately on their proximal flanks.

Diversity entropy greater than 0 indicates regions where contigs match multiple T2T-CHM13 chromosomes. This in turn provides indirect evidence for homology-driven recombination between heterologous chromosomes. To provide a form of positive control, we additionally apply an identical approach to look at contigs in the PARs on the sex chromosomes (**Supplementary Figures 12 and 13**). We confirm increased untangle diversity entropy in the PARs relative to the non-recombining regions of the chromosomes.

Diverse patterns of homology indicative of ongoing sequence exchange

A *multiple* untangling, which shows multiple best T2T-CHM13 matches for every contig segment, reveals patterns consistent with recombination in several regions with high diversity

entropy (**Supplementary Figures 14-18**). For exposition, we focus on a subset of representative contigs for 13, 14, and 21 (**Figure 5**). We hypothesize that regions where multiple T2T-CHM13 chromosomes provide near-identical mappings will be capable of supporting large-scale homologous recombination. These regions form a type of “checkerboard” pattern on the multiple untangling plots (**Figure 5**).

Some homologies, such as those beyond the centromeres between 13q12.11 and 21q11.2, or 14q11.2 and 22q11.21, do not appear to recombine frequently, and we typically see the same order of untangling hits consistent with ancient, or ancestral, homology. For example, in the 13q12.11/21q11.2 region we almost always observe “[13, 21]” from the perspective of chromosome 13 contigs (**Figure 5**). However, in the region near the SST1 array, on 13p11.2, 14p11.2, and 21p11.2, and among all chromosomes in proximity to the rDNA array, we observe a complex mix of different orders in contigs partitioned to these acrocentrics. This mix implies diversity in the local phylogeny of haplotypes across these loci relative to T2T-CHM13 chromosomes, a phenomenon typically associated with recombination (38). To quantify this pattern, we generalized the diversity entropy metric to work over orders of the top 5 untangling hits and using all contigs jointly. We compute this multi-hit, multi-contig diversity entropy for each position of T2T-CHM13 acrocentrics (**Supplementary Figures 19-23**).

Aggregating these results into T2T-CHM13 relative regions and removing short regions less than 30 kbp in length yields regions totalling 15.09 Mbp (**Supplementary Table 4 and Supplementary File 1**). The length of these pseudo-homologous regions (PHRs) differs by chromosome: chr13 = 3.92 Mbp, chr14 = 5.344 Mbp, chr15 = 452.67 kbp, chr21 = 3.435 Mbp, and chr22 = 1.943 Mbp. To provide a sense of scale, we identify a similar amount of PHRs on each sex chromosome (chrX = 5.36 Mbp, chrY 5.59 Mbp; **Supplementary Table 5**). In these chromosomes, the PHRs precisely match the PARs and contain sparse hits in the X-transposed regions (XTRs) (**Supplementary Figures 24 and 25**), which would be compatible with reports of X-Y interchange in the XTR (39). Biallelic SNP calls from a full HPRCy1 graph released in our companion work (4) show that variant density in the PARs and the acrocentric p-arms is markedly higher than elsewhere in these chromosomes (**Supplementary Figure 26**), which is compatible with increased rates of recombination in these regions (40).

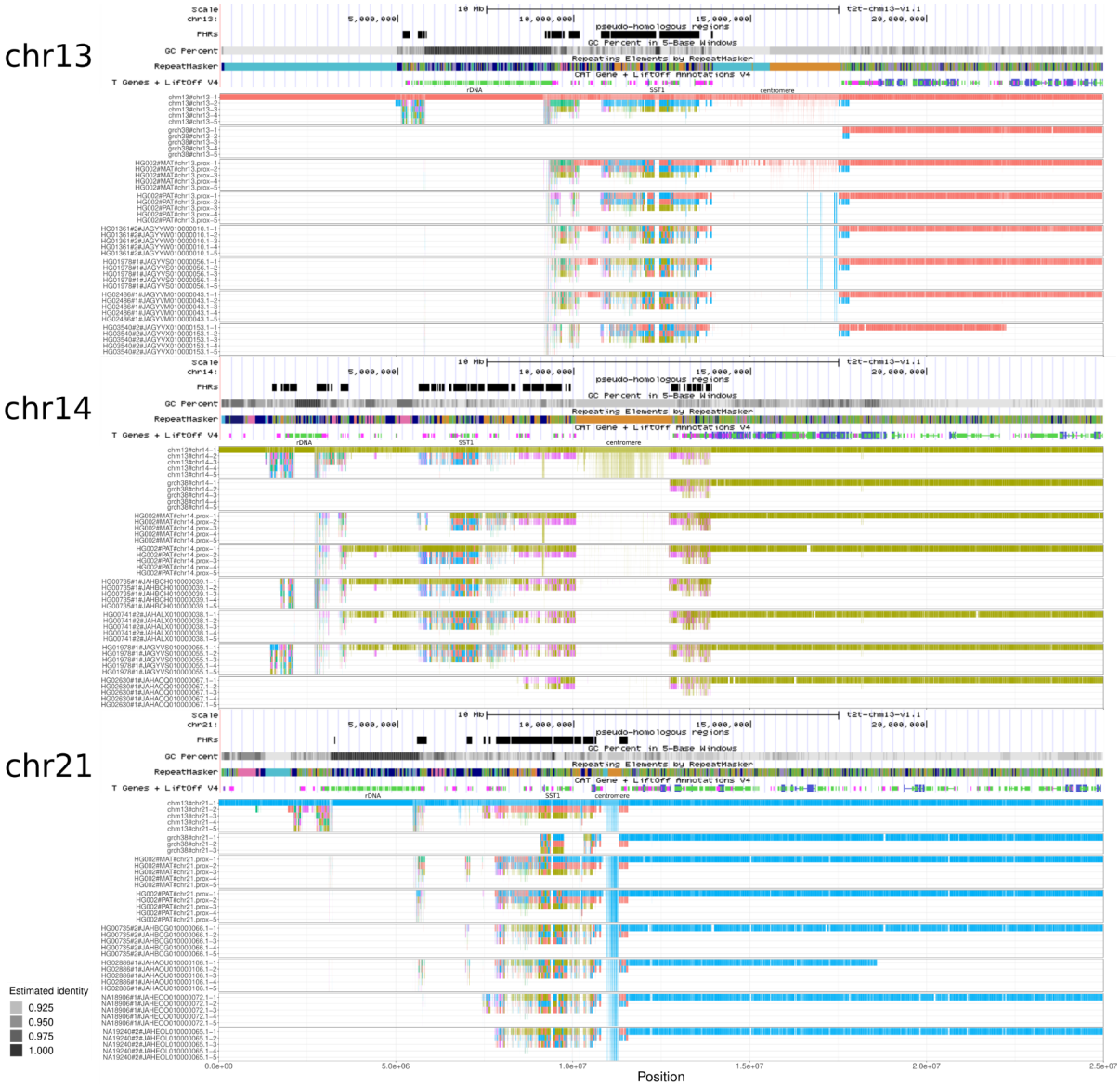

Figure 5. Multiple untangling of T2T-CHM13, GRCh38, HG002-Verkko haplotypes, and HPRCy1-acro contigs versus T2T-CHM13. Three of the five acrocentric chromosomes are represented (from top: 13, 14, and 21), with contigs and chromosome assemblies below. Transparency shows the estimated identity of the mappings. We display all mappings above 90% estimated pairwise identity. To allow the display of simultaneous hits to all acrocentrics, each grouping shows the first 5 best alternative mappings. Checkerboard patterns observed in several regions of the SAACs correspond to contexts that may permit ongoing recombination. These patterns are less common on chromosomes 15 and 22 (**Supplementary Figures 16 and 18**).

Linkage disequilibrium decay in pseudo-homologous regions

If SAACs are homologous but do not recombine, we would expect the T2T-CHM13-relative untangling of the graph paths to have very low entropy, i.e., each contig would tend to map optimally to a single T2T-CHM13 chromosome. What we observe is inconsistent with this

pattern. In large regions of the SAACs, most HPRCy1-acro contigs match patchworks of diverse T2T-CHM13 chromosomes (**Figure 4**), even in diverse orders relative to T2T-CHM13 (**Figure 5**). This is compatible with ongoing information flow between heterologous acrocentrics.

To quantify the magnitude of putative recombination in the PHRs, observed the rate linkage disequilibrium (LD) decay between markers of all HPRCy1-acro contigs within the variant space defined by the pangenome graph (41). Specifically, we obtained a set of SNPs relative to T2T-CHM13 from the acro-PVG and computed the LD for all pairs of variants separated up to 70 kbp (the median size of PHR intervals, **Supplementary Figure 27**) on the T2T-CHM13 acrocentrics. We then plotted the r^2 allele correlation versus distance for three sets of variants. We considered variants on the q-arm, the p-arm, and within the PHRs. We see that for chromosomes 13, 14, and 22, the LD decay is faster in the PHRs compared to p- and q-arms, while for 21 the decay curve is lower than for the whole p-arm and overlaps that of the q-arm (**Figure 6**). We do not observe a significant pattern on 15, likely due to the small size of its pseudo-homologous regions.

The fast decay in LD in the PHRs supports the hypothesis of ongoing recombination exchange. In the HPRCy1-acro PVG, we find many cases where homologous chromosomes map together. If these mappings were based on ancestral homology, we would not expect to see information flow between heterologous pairs. In our variant set, we would in turn expect to find higher relative levels of structuring in the PHRs than in the acrocentric q-arms. In general on the SAACs we find a higher level of LD than on the q-arms, perhaps due to lower recombination in heterochromatic regions (42). However, in the PHRs, we observe the opposite—even when considering haplotypes from different chromosomes, our overall level of LD is lower than in other regions. This effect is strongest within the chromosomes that other analyses suggest to share the most homologous sequence: 13, 14, and 22, weaker on 21, and weakest on 15, which appears to be an outlier (**Figures 1 and 3**). This pattern is consistent with increased recombination rate and or effective population size in these regions (7).

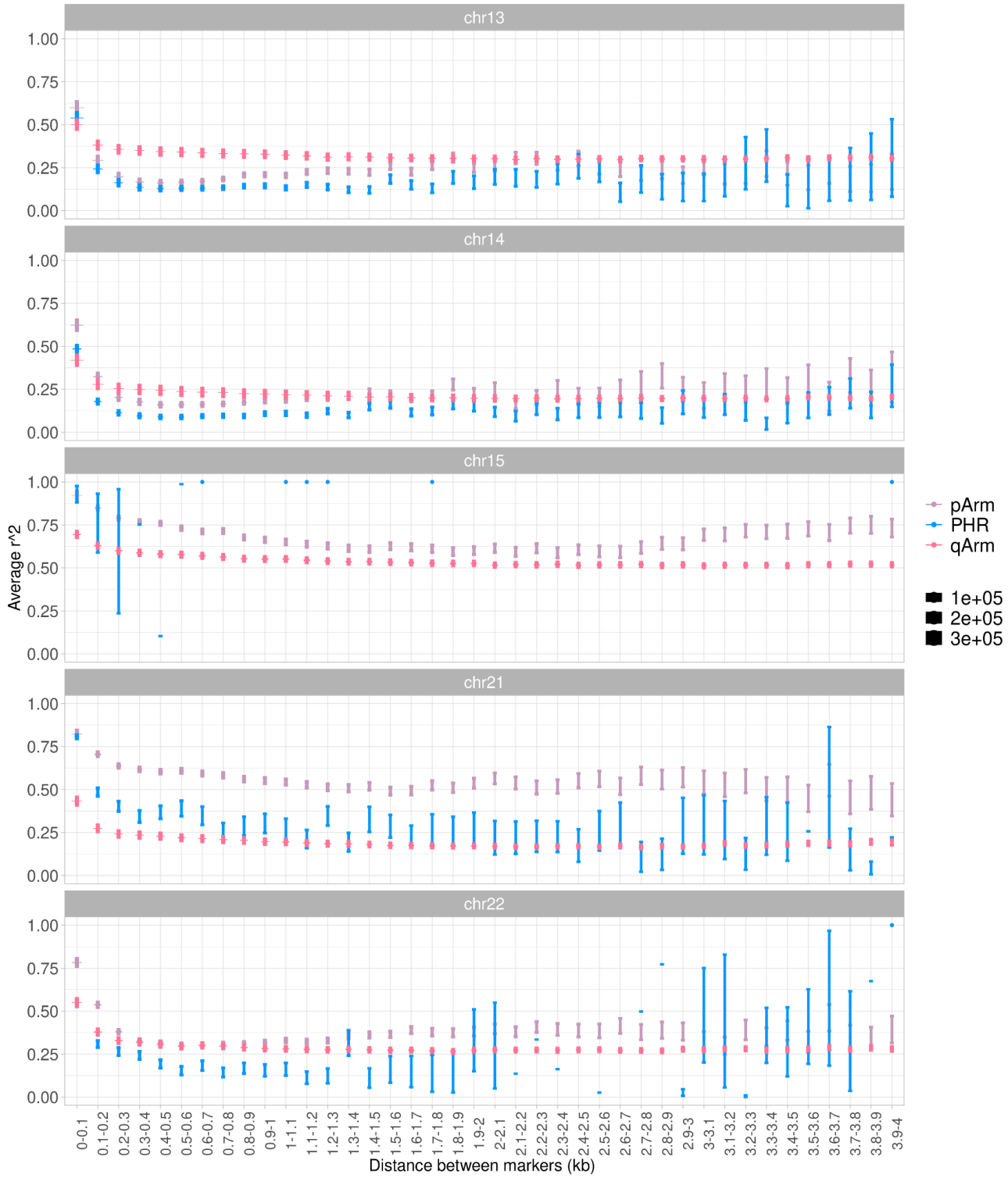

Figure 6. Linkage disequilibrium decay with distance between markers per acrocentric chromosome. Each LD decay plot shows the p-arm (purple), q-arm (pink), and PHR (blue) mean r^2 and 95% confidence intervals in bins of the given inter-marker distance. Dot size is proportional to the number of pairwise comparisons within a bin. LD decay is faster in PHRs for chromosomes 13, 14, and 22.

No significant LD decay is observed in PHRs for chromosome 15, likely due to small numbers of markers.

Validating observed patterns of sequence exchange

To provide cross-validation of the HiFi contig assemblies and our analysis of them, we compared the untangling of two assemblies of the same sample (HG002). One was made with the HPRCy1 pipeline, while the other used the Verkko diploid T2T assembler (6). Verkko employs ONT-UL to untangle ambiguous regions in a HiFi-based assembly graph, automating techniques first developed in production of T2T-CHM13. Verkko's assembly aggregates information from ONT-UL, thus providing a single integrated target for cross-validation of our analysis using an alternative sequencing and assembly approach.

We observe a high degree of concordance between the two methods at a level of the chromosome painting. The best-hit untangling shows similar patterns in the HG002-Verkko assembly as those seen in HG002-HPRCy1 (**Supplementary Figures 28-32**). However, some SAAC haplotypes appear to be poorly-assembled in HG002-HPRCy1. On the q-arms we measured 99.93% concordance between HG002-HPRCy1 and HG002-Verkko untangling results, but only 87.45% concordance on the p-arms (**Supplementary Table 3**). This lower level is consistent with greater difficulty in assembling the SAACs due to their multiple duplicated sequences (including the PHRs), satellite arrays, and the rDNA. We found the discordance was driven by a single chromosome haplotype: while most p-arms achieve around 90% concordance, HG002-HPRCy1 14p-maternal exhibits a high degree of discordance in the assemblies (66.19% concordance), and the nature of the disagreement is largely visible when comparing the two assemblies across 14p (**Supplementary Figure 29**). Although this N=1 validation focuses on only 10 haplotypes, it incorporates many independent reads provided by deep HiFi (105x) and ONT-UL (85x) data used in HG002-Verkko (6). We thus have compared the concordance between structures observed in single molecule reads across the SAACs and the HG002-HPRCy1 assembly that represents the HiFi-only assembly process that produced our pangenome.

However, this analysis should be seen as presenting a lower bound on our process accuracy. We are considering *all* HG002-HPRCy1 contigs that map to the acrocentrics—not only those that would meet our centromere-crossing requirement, and HG002-HPRCy1 is itself more fragmented than the other assemblies which we have selected for the acro-PVG (4). The fragmented nature of its contigs (only one from chromosome 22 meets our p-q mapping requirement) may introduce additional disagreements with HG002-Verkko. The overall result indicates that most patterns observed in HiFi-only assemblies are likely to be supported by an automated near-T2T assembly of the same sample.

Discussion

We develop multiple lines of evidence indicating active recombination between heterologous human acrocentric chromosomes. First, we find that a symmetric comparison of the sequences of a draft human pangenome contains multi-chromosome communities corresponding to both the sex chromosomes and acrocentric chromosomes (**Figure 1**). An acrocentric pangenome graph reveals base-level homologies that outline patterns of exchange between the heterologous chromosomes (**Figures 2 and 3**). We then find that chromosome paintings extracted from the pangenome graph contain regions featuring interspersed best-match patterns involving non-homologous T2T-CHM13 chromosomes (**Figure 4**), and we cross-validate these findings with a T2T diploid assembly of a target sample. We develop an entropy metric suitable to highlight this phenomenon and apply it to derive a set of putatively recombining pseudo-homologous regions (PHRs) across the SAACs (**Figure 5 and Supplementary File 1**). Finally, we show that on chromosomes 13, 14, 21, and 22, the resulting 15 Mbp of sequence in their PHRs presents a higher rate of linkage disequilibrium decay than sequences from non-PHR regions of the same chromosomes, even when comparing sets of haplotypes from diverse acrocentric chromosomes, indicating higher rates of recombination or a larger effective population size in these regions (**Figure 6**). These lines of evidence all suggest that heterologous SAACs recombine.

Our findings consistently describe a pattern that implicates the ribosomal DNA repeats—GC-rich variable number tandem repeat arrays found on the SAACs—as nucleation points for recombination between heterologous chromosomes. This would confirm a large body of evidence accumulated over the last 50 years of research (1,2,8,15,16,43). The HPRC pangenome provides base-level resolution of recombination patterns across many SAAC haplotypes, letting us add further detail to this picture. We observe that the GC-rich SST1 array lies at the center of a segmentally duplicated region on 13p, 14p, and 21p which shows a clear pattern of haplotype mixing between these chromosomes. This may also implicate this array as a nucleation point for recombination, as suggested by the observation (N=1/220) of heterologous 14p-21p synapse formation in pachytene oocytes (33). We also find that regions on the centromere-adjacent q-arms of certain pairs (13+21 and 14+22) show evidence of heterologous interchange. From this border, our analyses suggest an increasing rate of recombination exchange between heterologous chromosomes that rises until a maximum in the regions flanking the rDNA.

Although we find that assembly methods have difficulty in the SAACs (4), our validation based on ONT-UL and HiFi data integration shows that the patterns that we observe in HiFi-only assemblies are consistent with ONT-UL reads from the same sample. HG002-Verkko recapitulates key T2T-CHM13-relative untangling patterns also seen in HPRCy1 HiFi assemblies, such as the SST1-linked PHR at 13p11.2 and 21p11.2, rDNA-proximal mixing of all SAACs, and mixing of 22q11.21 and 14q11.2. Our analyses rest on the extensive assembly validation carried out by the HPRC, and observations used to establish signals for recombination are based only on assembly regions deemed to be reliable by mapping analyses

(4). Our study confirms prior hypotheses based on decades of diverse inquiry, which provides additional assurance that patterns observed bioinformatically are biologically grounded.

Our study is fundamentally population-based. We cannot directly observe recombination of SAACs in this context, leaving open questions about recombination mechanisms that may be difficult to resolve with sequence information alone (44). However, recombination, like mutation, is a rare event, making it easier to measure its distribution over chromosomes in a population genetic context as we have done here (7). This addresses key issues with many previous studies of recombination in the SAACs, which often feature small numbers of individuals (33,45) presenting medically-relevant genomic states like trisomy and Robertsonian translocation (46). Our resolution of the acrocentric PHRs confirms reported homologies between the SAACs (2,3) providing a reference for their structure that will be useful to future genomic studies.

The rDNA arrays are of essential importance to cell function. Positioning them on homologous short chromosome arms is possibly a solution to the problem of obtaining sufficient redundancy for the rDNA such that protein translation is not interrupted by germline or somatic mutation. The high copy number of the PHRs of the SAACs, as well as their frequent physical co-localization to the nucleolar membranes (13), helps to ensure that there is always a suitable partner for homology-directed repair at hand in the event of double-stranded breaks in the rDNA (47). Our analyses suggest that the rate of recombination between heterologous pairs of acrocentrics varies, which leads to characteristic patterns in the homology spaces that we have explored. Insofar as recombination represents physical proximity, an implication is that groups of acrocentrics between which we observe stronger homology and recombination—such as 13, 14, and 21 —may be more likely to co-localize to the same nucleolus, as observed in pachytene spermatocytes (13). If driven by sequences in the nucleolar organizing regions, then these patterns may vary and be heritable (8). The HPRC draft human pangenome has allowed us to approach genome evolution from a chromosome scale. By stepping away from a reference-centric model and directly comparing whole chromosome assemblies of the acrocentrics, we obtain sequence-resolved responses to long-standing questions first posed in early cytogenetic studies of human genomes.

Materials and Methods

Genome assemblies

We analyzed the 47 T2T phased diploid de novo assemblies (94 haplotypes in total) produced by the HPRC and described in (4). We included both T2T-CHM13 version 2 (3) and GRCh38.

Community detection

We performed pairwise mapping for all contigs from the 47 T2T phased diploid de novo assemblies with the WFMASH sequence aligner (48) (commit ad8aeba). We set the following parameters:

```
wfmash HPRCy1.fa -s 50k -l 250k -p 95 -n 93 -Y '#' -H 0.001 -m
```

We used segment seed length of 50kbps (`-s`), requiring homologous regions at least ~250 kbp long (`-l`) and estimated nucleotide identity of at least ~95% (`-p`). Having 94 haplotypes in total, we kept up to 93 mappings for each contig (`-n`). Moreover, we skipped mappings when the query and target had the same prefix before the '#' character (`-Y`), that is when involving the same haplotype. To properly map through repetitive regions, only the 0.001% of the most frequent kmers were ignored (`-H`). We skipped the base-level alignment (`-m`). We also generated pairwise mapping with the same parameters, but using a segment seed length of 10kbps and requiring homologous regions at least ~50kbp long.

From the resulting mappings, we excluded those involving contigs shorter than 1 Mbps to reduce the possibility of spurious matches. We then used the `paf2net.py` Python script (delivered in the PGGB repository) to build a graph representation of the result (a mapping graph), with nodes and edges representing contigs and mappings between them, respectively.

```
python3 ~/pggb/scripts/paf2net.py -p HPRCy1.1Mbps.paf
```

The script produces a file representing the edges, a file representing the edge weights, and a file to map graph nodes to sequence names. The weight of an edge is given by the product of the length and the nucleotide identity of the corresponding mapping (higher weights were associated with longer mappings at higher identities). Finally, we used the `net2communityes.py` Python script (delivered in the PGGB repository) to apply the Leiden algorithm (22), implemented in the igraph tools (49), to detect the underlying communities in the mapping graph.

```
python3 ~/pggb/scripts/net2communities.py \
     -e HPRCy1.1Mbps.edges.list.txt \
     -w HPRCy1.1Mbps.edges.weights.txt \
     -n HPRCy1.1Mbps.vertices.id2name.chr.txt --accurate-detection
```

To identify which chromosomes were represented in each community, we partitioned all contigs by mapping them against both T2T-CHM13v1.1 and GRCh38 human reference genomes with WFMASH, this time requiring homologous regions at least 150 kbp long and nucleotide identity of at least 90%.

```
wfmash chm13+grch38.fa HPRCy1.fa -s 50k -l 150k -p 90 -n 1 -H 0.001
-m -N
```

We disabled the contig splitting (`-N`) during mapping to obtain homologous regions covering the whole contigs. For the unmapped contigs, we repeated the mapping with the same parameters, but allowing the contig splitting (without specifying `-N`). We labeled contigs with 'p' or 'q' depending on whether they cover the short arm or the long arm of the chromosome they

belonged to. Contigs fully spanning the centromeres were labeled with ‘pq'. We used such labels to identify the chromosome composition of the communities detected in the mapping graph obtained without reference sequences, and to annotate the nodes in the mapping graph.

To obtain a clean visualization of the homology relationships between the HPRC assemblies, we generated a simpler mapping graph by using the same parameters used for the main graph, but keeping up to 3 mappings for each contig and adding the T2T-CHM13 reference genome version 2, which includes also the complete HG002 chromosome Y (https://www.ncbi.nlm.nih.gov/assembly/GCF_009914755.1):

```
wfmash HPRCy1+chm13v2.fa -s 50k -l 250k -p 95 -n 3 -Y '#' -H 0.001 -m
-w 5000
```

We set window size for sketching equal to 5000 (`-w`) to reduce the runtime by sampling fewer kmers. We used the `paf2net.py` Python script to build the mapping graph and then used Gephi (50) (version 0.9.4) to visualize it. We computed the mapping graph layout by running ‘Random Layout’ and then the ‘Yifan Hu’ algorithm.

Pangenome graph building

For each of the 47 T2T phased diploid de novo assemblies, we mapped all contigs against the T2T-CHM13 human reference genome with the WFMASH sequence aligner (commit ad8aeba). For the HG002 sample, we included two assemblies: the HG002-HPRCy1 phased diploid de novo assembly (built with HiFi reads) and a phased diploid de novo assembly based on both HiFi and ONT-UL reads, built with the Verkko assembler (6). We set the following parameters:

```
wfmash chm13.fa assembly.fa -s 50k -l 150k -p 90 -n 1 -H 0.001 -m
```

We used segment seed length of 50kbps (`-s`), requiring homologous regions at least ~150 kbp long (`-l`) and estimated nucleotide identity of at least ~90% (`-p`). We kept only 1 mapping (the best one) for each contig (`-n`). To properly map through repetitive regions, only the 0.001% of the most frequent kmers were ignored (`-H`). We skipped the base-level alignment (`-m`). For the HG002-HPRCy1 contigs, we disabled the contig splitting (`-N`).

Then, we identified contigs originating from acrocentric chromosomes and covering both the short and long arms of the chromosome they belonged to. We considered only contigs with mappings at least 1 kbp long on both arms and at least 1 Mbps away from the centromere. We call such contigs “p-q acrocentric contigs”. For HG002-HPRCy1, only contigs longer than or equal to 300 kbps were considered, regardless of covering both arms of the belonging chromosomes.

Finally, we built a pangenome graph with all the p-q acrocentric contigs and both T2T-CHM13 and GRCh38 human reference genomes by applying PGGB (25) (commit a4a6668). We set the following parameters:

```
pggb -i contigs.fa.gz -s 50k -l 250k -p 98 -n 162 -F 0.001 -k 311 -G
13117,13219 -O 0.03
```

We used segment seed length of 50kbps (`-s`), requiring homologous regions at least ~250 kbp long (`-l`) and estimated nucleotide identity of at least ~98% (`-p`). Having 142 p-q acrocentric contigs in input (132 from HG002-HPRCy1 and 10 from HG002-Verkko) plus 10 acrocentric chromosomes from the T2T-CHM13 and GRCh38 reference genomes plus 49 HG002-HPRCy1 contigs representing other 10 acrocentric haplotypes (5 maternal and 5 paternal), we kept up to 162 mappings (142 + 10 + 10) for each contig (`-n`). To properly map through repetitive regions, only the 0.001% of the most frequent kmers were ignored (`-F`). We filtered out alignment matches shorter than 311 bp to remove possible spurious relationships caused by short repeated homologies (`-k`). We set big target sequence lengths and a small sequence padding for two rounds of graph normalization (`-G` and `-O`). To visualize the acrocentric pangenome graph, we built the graph layout with ODGI LAYOUT (35) (commit e2de6cb) and visualized with GFAESTUS (34) (commit 50fe37a).

Pangenome graph untangling

We untangled the pangenome graph by applying ODGI UNTANGLE (commit e2de6cb). Practically, we projected the graph into an alignment between a set of query (HPRCy1 contigs) and reference (T2T-CHM13) sequences. We set the following parameter:

```
odgi untangle -i graph.og -e 50000 -m 1000 -n 100 -j 0 -R targets.txt
-d cuts.txt
```

We segmented the graph into regular-sized regions of ~50kbps (`-e`), merging regions shorter than 1kbps (`-m`). We reported up to the 100th best target mapping for each query segment (`-n`), not applying any threshold for the jaccard similarity (`-j`). We used all paths in the graphs as queries and projected them against the 5 acrocentric chromosomes of the T2T-CHM13 genome (`-R`). Moreover, we emit the cut points used to segment the graph (`-d`).

For each query segment, if there were multiple best hits against different targets (i.e., hits with the same, highest jaccard similarity), we put as the first one the hit having as target the chromosome of origin of the query (obtained from the chromosome partitioning of the contigs).

We repeated the graph untangling another 5 times, but constrained the algorithm to use only one T2T-CHM13's acrocentric chromosome as a target at a time (`-r`) and return the best-matching hit (`-n`).

```
odgi untangle -i graph.og -e 50000 -m 1000 -n 1 -j 0 -r chr13 -c
cuts.txt
```

```
odgi untangle -i graph.og -e 50000 -m 1000 -n 1 -j 0 -r chr14 -c
cuts.txt
odgi untangle -i graph.og -e 50000 -m 1000 -n 1 -j 0 -r chr15 -c
cuts.txt
odgi untangle -i graph.og -e 50000 -m 1000 -n 1 -j 0 -r chr21 -c
cuts.txt
odgi untangle -i graph.og -e 50000 -m 1000 -n 1 -j 0 -r chr22 -c
cuts.txt
```

We used the cut points generated when using all T2T-CHM13's acrocentric chromosomes as targets (`-c`). In this way, all untangling runs (six in total) used the same cut points for the segment boundaries.

Finally, we "grounded" the untangled output generated with all acrocentric chromosomes as targets: in more detail, each untangled query segment was placed against a particular acrocentric chromosome (not only the best-matching one) by using the untangled outputs constrained to a single target. We split the result by acrocentric chromosome and kept only queries untangling both p- and q-arms of the targets. Furthermore, we removed query segments overlapping regions flagged in the assemblies as unreliable (i.e., having coverage issues) by FLAGGER (4). To focus on the more similar query-target hits, we used the jaccard metric to estimate the sequence identity by applying the corrected formula reported by (51), and retained only results at greater than 90% estimated identity. We displayed the untangling results for each acrocentric chromosome with the R development environment (version 3.6.3), equipped with the following packages: tidyverse (version 1.3.0), RColorBrewer (version 1.1.2), ggplot2 (version 3.3.3) and ggrepel (version 0.9.1).

Information flow aggregation and diversity entropy

Aggregating best-hit untangle results

For each group of HPRCy1-acro contigs anchored to a T2T-CHM13 acrocentric q-arm, we counted the number of contigs having as best-hit each one of the acrocentrics. In particular, for each base position of the T2T-CHM13 acrocentric chromosome of each group, we quantified how many times each of the acrocentrics appeared as best-hit in the pangenome graph untangling. We considered only best-hits with an estimated identity of at least 90%.

Best-hit untangling diversity index

To quantify the degree of disorder in the untangling result, we calculated the diversity entropy across the different acrocentrics that were present as best-hit. In more detail, we projected each HPRCy1 acrocentric p-q contig against the T2T-CHM13 acrocentric to which it is anchored via the q-arm and associated each reference base position to the corresponding acrocentric best-hit. We considered only best-hits with an estimated identity of at least 90%. Then, we computed the Shannon Diversity Index (SDI) in windows 50 kbp long over the contigs. We used

-1 as missing SDI value in the regions where the contigs do not match any targets. For each group of contigs, we aggregated the SDI results by computing their average (ignoring the missing SDI values) for each reference base position.

We call this metric “Best-SDI”, and it serves to show regions where contigs can be described as patchworks of different reference chromosomes. However, it cannot distinguish regions where there are different orders of reference chromosome similarity—that might be indicative of recombination exchange—from those where there is regional diversity in each contig’s relationship to T2T-CHM13. The latter case could occur if T2T-CHM13 itself contains rare recombinations between acrocentrics, or where ancient homology might result in “noise” in untangling alignments as contigs pick from two equally-good alternative mappings. To avoid these pitfalls and establish a more stringent graph-space recombination metric, we then extended the untangling diversity metric to operate on multiple mappings.

Multi-hit untangling diversity index and PHR derivation

To take into account the other hits in addition to the first one, including their order, we generalized the diversity entropy metric to work over orders of the top 5 untantling hits and consider all contigs jointly. For each reference segment, we collected the corresponding best 5 untantling hits for each of the HPRCy1 acrocentric p-q contig; this is possible because the reference segments are stable across all contigs. We considered only best-hits with an estimated identity of at least 90%. To avoid driving the untangle entropy by intra-chromosomal similarity caused by segmental duplications modeled in the structure of the PVG (as seen in loops on 13q, 15q, and 22q, **Figure 2**), we ignored consecutive duplicate target hits—in other words, we took the ordered set of unique reference targets. When multiple contig segments were grounded against the same reference segment, we considered the first contig segment having the best grounding, that is having the highest estimated identity when placed against the current reference segment. Then, we ranked the 5 best-hits by estimated similarity. Finally, for each reference segment, we computed the SDI across all available 5 best-hits orders. We used -1 as missing SDI value in the reference regions without any contig matches. We kept in the output also the information about how many HPRCy1 acrocentric p-q contigs contributed to the entropy computation in each reference segment. This yielded the “Multi-SDI” metric. To obtain the pseudo-homologous regions (PHRs), we aggregated the final results by considering regions with Multi-SDI greater than 0 and supported by at least 1 contig, merging with BEDtools (52) those that were less than 10 kbps away, and removing merged regions shorter than 30 kbps.

Display and annotation of untangle paintings

We displayed the aggregated results for each acrocentric chromosome. We used genome annotations for the first 25 Mbp of each acrocentric chromosome, using T2T-CHM13v1.1 UCSC trackhub (https://genome.ucsc.edu/cgi-bin/hgTracks?db=hub_2397929_t2t-chm13-v1.1). We made the figures with the scripts available at https://github.com/pangenome/chromosome_communities/tree/main/scripts. To plot the figures, we used R (version 3.6.3), equipped with the following packages: tidyverse (version 1.3.0), RColorBrewer (version 1.1.2), ggplot2 (version 3.3.3) and ggrepel (version 0.9.1). Finally, we

used Inkscape (https://inkscape.org/) to compose main-text figures based on the results, and provide supplementary figures directly produced by these methods.

HPRCy1 SNP density plots

We displayed biallelic SNP density in the full HPRCy1 draft pangenome built with PGGB (4) versus both GRCh38 and T2T-CHM13. To do so, we extracted biallelic SNPs from the released VCF files versus each chromosome, for both references (get_bisnp.sh). Because T2T-CHM13 version 1.1, which was used in HPRCy1, does not have a Y chromosome, we used GRCh38's, which includes masked PAR1 and PAR2 regions. We displayed biallelic SNP density in bins of 100kbp, using R (version 4.1.1) and tidyverse (version 1.3.1) package (plot_bisnp_dens.R).

Linkage disequilibrium analysis

We identified variants embedded in the pangenome graph by using VG DECONSTRUCT (5):

```
vg deconstruct -P chm13 -H '?' --ploidy 1 -e -a graph.gfa > variants.vcf
```

We called variants with respect to the T2T-CHM13 reference genome (`-P`), reporting variants for each HPRCy1 acrocentric p-q contig (`-H` and `--ploidy`). We considered only traversals that correspond to paths (i.e., contigs) in the graph (`-e`) and also reported nested variation (`-a`). From the variant set, we considered only single nucleotide variants. We estimated linkage disequilibrium between pairs of markers within 70kb by using PLINK v1.9 (53) upon specification of haploid sets and retaining all values of r2 >0 (plot_ld_1.R). Finally, we generated binned LD decay plots with confidence intervals using R (version 3.6.3).

Validating observed patterns of sequence exchange

HG002-Verkko assembler

We applied an earlier version of Verkko (beta 1, commit vd3f0b941b5facf5807c303b0c0171202d83b7c74) to build a diploid assembly graph for the HG002 cell line using the HiFi (105x) and ONT (85x) reads as described in (6). The resulting assembly graph resolves the proximal junction in single contigs for each haplotype up to multi-mega bases, while the distal junctions (DJ) remain to be resolved. We used homopolymer compressed markers from the parental Illumina reads to assign unitigs to maternal, paternal haplotype or ambiguous when not enough markers supported either haplotype. For estimating the number of times a unitig has to be visited, we aligned HiFi and ONT reads to the assembly graph using GraphAligner with the following parameters: --seeds-mxm-length 30 --seeds-mem-count 10000 -b 15 --multimap-score-fraction 0.99 --precise-clipping 0.85 --min-alignment-score 5000 --clip-ambiguous-ends 100 --overlap-incompatible-cutoff 0.15 --max-trace-count 5 --hpc-collapse-reads --discard-cigar (54). Four DJ junctions were connected to the rDNA arrays with ambiguous nodes connecting the maternal and paternal nodes,

supporting they belong to the same chromosome. Two DJ unitigs, one maternal and paternal, were disconnected from each other but connected to the rDNA arrays, which were assigned to the same chromosome. Using the marker and ONT alignments, we identified paths in the graph and assigned them according to the most supported haplotype. If only ambiguous nodes were present between the haplotype assigned unitigs, with no ONT reads to resolve the path, nodes were randomly assigned to one haplotype to build the contig. After all paths were identified, we produced the consensus using `verkko --assembly <path-to-original-assembly> --paths <path-to-paths>`. The entire procedure to produce parental markers, tagging unitigs according to its haplotype on the assembly graph and finding the path using ONT reads is now available in the latest Verkko (v1) in a more automated way.

Validating untangling patterns

We validated the results of the pangenome untangling by comparing the best-hits of the two HG002 assemblies, the one built with HiFi reads and the other based on both HiFi and ONT reads, built with the Verkko assembler (6). For each base position of each T2T-CHM13 acrocentric chromosome, we compared the untangling best-hit of the HG002-HPRCy1 contigs with the best-hit supported by HG002-Verkko contigs. We considered only best-hits with an estimated identity of at least 90%. We defined reference regions as concordant when both HG002 assemblies supported the same T2T-CHM13 acrocentric as best-hit. We treated the two haplotypes (maternal and parental) separately.

Acknowledgements

Our work depends on the HPRC draft human pangenome resource established in our companion paper (4), and we thank production and assembly groups for their efforts in establishing this resource. This work used the computational resources of the UTHSC Octopus cluster and NIH HPC Biowulf cluster. We acknowledge support in maintaining these systems that was critical to our analyses.

Funding

This work was supported, in part, by National Institutes of Health/NIDA U01DA047638 (E.G.), National Institutes of Health/NIGMS R01GM123489 (E.G.), NSF PPoSS Award #2118709 (E.G., C.F.), the Tennessee Governor's Chairs program (C.F., E.G.), and the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health (A.R., S.K., A.M.P.). We acknowledge support from Human Technopole (A.G.), Consiglio Nazionale delle Ricerche, Italy (S.B., V.C.), and Stowers Institute for Medical Research (T.P., B.R., J.L.G).

Supplementary Materials

Online Materials

Code and links to materials and methods used to perform all the analyses and produce all the figures can be found at the following repository: https://github.com/pangenome/chromosome_communities.

Combined figures and supplementary material
Supplementary materials may be accessed at https://doi.org/10.5281/zenodo.6993789.

We provide alternative links to materials including all source files for figures and tables in the manuscript and supplement. These include graph builds and layouts not included in the Zenodo upload due to technical constraints.

Main Figures
https://drive.google.com/drive/folders/1QdVCrnYnPgxWf2YD8V2M26O-Dng0CwCL

Supplementary Figures
https://drive.google.com/drive/folders/102iFYLEnqD5FraJN3h771qFw8HSu1vyU

Tables https://drive.google.com/drive/folders/1wOTyxKdv3PjwarFvHLtq-1YJMnXVqGZG

Notes https://drive.google.com/drive/folders/1qpmn2L5lAJilMRP4sWAmQDOX5yVHp3_P

Files https://drive.google.com/drive/folders/1FFb-uF60qXg5kBeambBYvQVDUYc3AshW

Figure captions

Supplementary Figure 1. Results of community assignment on the mapping graph when using chains of 10 kbp seeds of 95% average nucleotide identity. The x-axis gives the assigned partition of a contig based on competitive mapping to the chromosomes of T2T-CHM13 and GRCh38, while the y-axis indicates the community, which is named by the chromosome partition that contributes the largest number of contigs to it.
Supplementary Figure 2. Regions of a PVG built from centromere-spanning HPRCy1 contigs plus T2T-CHM13 and GRCh38 references. We apply GFAESTUS to visualize the 2D layout generated by ODGI in the PGGB pipeline. We focus on the segmentally duplicated core centered in the SST1 array (labeled).
Supplementary Figure 3. A PVG built from centromere-spanning HPRCy1 contigs, without embedding the T2T-CHM13 and GRCh38 references. We apply ODGI to visualize the 2D layout generated in the PGGB pipeline. This renders sequences and chains of small variants as linear structures, while repeats caused by segmental duplications, inversions, and other structural

variants tend to form loops and tangles. The acrocentric q-arms are almost completely separated, while the p-arms unite in a structure adjacent to the rDNA array.

Supplementary Figure 4. Regions of a PVG built from centromere-spanning HPRCy1 contigs without the T2T-CHM13 and GRCh38 references. We apply GFAESTUS to visualize the 2D layout generated by ODGI in the PGGB pipeline. We focus on the segmentally duplicated core centered in the SST1 array (labeled).

Supplementary Figure 5. Scheme of the graph untangling. We applied ODGI UNTANGLE to obtain a mapping from segments of all PVG paths onto T2T-CHM13. The segmentation cuts the graph into regular-sized regions whose boundaries occur at structural variant breakpoints. For each query subpath through a graph segment, we use a Jaccard metric over the sequence space of the subpaths to find the best-matching reference segment.

Supplementary Figures 6-10. Untangling of HPRCy1-acro's sequences belonging to chromosome 13, 14, 15, 21, 22 versus T2T-CHM13. Transparency shows the estimated identity of the mappings. We display all mappings above 90% estimated pairwise identity. Checkerboard patterns observed in several regions of the SAACs correspond to contexts that may permit ongoing recombination.

Supplementary Figure 11. RepeatMasker track legend.

Supplementary Figures 12 and 13. For HPRCy1 contigs from chromosome X and Y, we show the tracks for the pseudoautosomal regions (PARs) and the X-transposed region (XTRs) with respect to T2T-CHM13 (on the top) as well as the untangle entropy metric (**Entropy**, on the bottom) computed over the contigs' T2T-CHM13-relative untanglings.

Supplementary Figures 14-18. Multiple untangling of HPRCy1-acro's sequences belonging to chromosome 13, 14, 15, 21, 22 versus T2T-CHM13. Transparency shows the estimated identity of the mappings. We display all mappings above 90% estimated pairwise identity. To allow the display of simultaneous hits to all acrocentrics, each grouping shows the first 5 best alternative mappings. Checkerboard patterns observed in several regions of the SAACs correspond to contexts that may permit ongoing recombination.

Supplementary Figures 19-23. Multi-hit untangling diversity for HPRCy1-acro's sequences belonging to chromosome 13, 14, 15, 21, 22 versus T2T-CHM13. We considered all mappings above 90% estimated pairwise identity.

Supplementary Figures 24 and 25. For HPRCy1 contigs from chromosome X and Y, we show the tracks for the pseudoautosomal regions (PARs) and the X-transposed region (XTRs) with respect to T2T-CHM13 as well as the pseudo-homologous regions (last track on the bottom).

Supplementary Figure 26. Biallelic SNP density in the full HPRCy1 draft pangenome. Colors indicate the count of SNPs in each bin at given variant nesting levels (41), with gray showing the base level and a red to violet spectrum for increasing level.

Supplementary Figure 27. *Left*) Histograms showing missing genotypes per site for variants called with respect to the T2T-CHM13 acrocentrics. *Middle*) Histograms showing the allele frequency spectrum for variants called with respect to the T2T-CHM13 acrocentrics. MAF = Minor Allele Frequency. *Right-Top*) Segment size distributions of the pseudo-homologous regions with respect to the T2T-CHM13 acrocentrics. *Right-Bottom*) Distributions of the distance between consecutive markers with respect to the T2T-CHM13 acrocentrics.

Supplementary Figures 28-32. For each base position of T2T-CHM13 chromosome 13, 14, 15, 21, 22, we compared the untangling best-hit of the HG002-HPRCy1 contigs with the best-hit

supported by HG002-Verkko's contigs. We considered only best-hits with an estimated identity of at least 90%. We reported the number of different targets for each reference position. On the bottom, the untangling of HG002's contigs from HG002-Verkko assembly as those seen in HG002-HPRCy1 assemblies, for chromosome 13, 14, 15, 21, 22 versus T2T-CHM13. Transparency shows the estimated identity of the mappings. We display all mappings above 90% estimated pairwise identity. Checkerboard patterns observed in several regions of the SAACs correspond to contexts that may permit ongoing recombination.

Table captions

Supplementary Table 1. Results of community assignment on the mapping graph.
Supplementary Table 2. For each HPRCy1-acro contig, we report the amount of projected alignments, the amount of reliable projected alignments, and the fraction of unreliable alignments that we ignored in our analysis.
Supplementary Table 3. For each T2T-CHM13 chromosome, we report the number of bases where HG002-HiFi and HG00-Verkko assemblies are concordant and discordant, that these the bases where the untangling best-hits from both assemblies is on the same T2T-CHM13 chromosome or not.
Supplementary Table 4. Pseudo-homologous regions identified in the acrocentric chromosomes of T2T-CHM13.
Supplementary Table 5. Pseudo-homologous regions identified in the sex chromosomes of T2T-CHM13.

Note captions

Supplementary Note 1. Physical proximity modeling of acrocentric short arms.

References

1. Floutsakou, I. *et al.* The shared genomic architecture of human nucleolar organizer regions. *Genome Res.* **23,** 2003–2012 (2013).
2. van Sluis, M. *et al.* Human NORs, comprising rDNA arrays and functionally conserved distal elements, are located within dynamic chromosomal regions. *Genes Dev.* **33,** 1688–1701 (2019).
3. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376,** 44–53 (2022).
4. Liao, W.-W. *et al.* A Draft Human Pangenome Reference. *bioRxiv* 2022.07.09.499321 (2022). doi:10.1101/2022.07.09.499321

5. Garrison, E. *et al.* Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36,** 875–879 (2018).
6. Rautiainen, M. *et al.* Verkko: telomere-to-telomere assembly of diploid chromosomes. *bioRxiv* 2022.06.24.497523 (2022). doi:10.1101/2022.06.24.497523
7. Pritchard, J. K. & Przeworski, M. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69,** 1–14 (2001).
8. Henderson, A. S., Warburton, D. & Atwood, K. C. Letter: Ribosomal DNA connectives between human acrocentric chromosomes. *Nature* **245,** 95–97 (1973).
9. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
10. Spinner, N. B. in *Brenner's Encyclopedia of Genetics (Second Edition)* (eds. Maloy, S. & Hughes, K.) 546–548 (Academic Press, 2013). doi:10.1016/B978-0-12-374984-0.00238-2
11. Lindström, M. S. *et al.* Nucleolus as an emerging hub in maintenance of genome stability and cancer pathogenesis. *Oncogene* **37,** 2351–2366 (2018).
12. Kobayashi, T. Regulation of ribosomal RNA gene copy number and its role in modulating genome integrity and evolutionary adaptability in yeast. *Cell. Mol. Life Sci.* **68,** 1395–1403 (2011).
13. Holm, P. B. & Rasmussen, S. W. Human meiosis I. The human pachytene karyotype analyzed by three dimensional reconstruction of the synaptonemal complex. *Carlsberg Res. Commun.* **42,** 283 (1977).
14. van Sluis, M. & McStay, B. A localized nucleolar DNA damage response facilitates recruitment of the homology-directed repair machinery independent of cell cycle stage. *Genes Dev.* **29,** 1151–1163 (2015).
15. Choo, K. H., Vissel, B., Brown, R., Filby, R. G. & Earle, E. Homologous alpha satellite sequences on human acrocentric chromosomes with selectivity for chromosomes 13, 14 and 21: implications for recombination between nonhomologues and Robertsonian

translocations. *Nucleic Acids Res.* **16,** 1273–1284 (1988).

16. Choo, K. H., Vissel, B. & Earle, E. Evolution of alpha-satellite DNA on human acrocentric chromosomes. *Genomics* **5,** 332–344 (1989).
17. Devilee, P. *et al.* Two subsets of human alphoid repetitive DNA show distinct preferential localization in the pericentric regions of chromosomes 13, 18, and 21. *Cytogenet. Cell Genet.* **41,** 193–201 (1986).
18. Jørgensen, A. L., Bostock, C. J. & Bak, A. L. Homologous subfamilies of human alphoid repetitive DNA on different nucleolus organizing chromosomes. *Proc. Natl. Acad. Sci. U. S. A.* **84,** 1075–1079 (1987).
19. Greig, G. M., Warburton, P. E. & Willard, H. F. Organization and evolution of an alpha satellite DNA subset shared by human chromosomes 13 and 21. *J. Mol. Evol.* **37,** 464–475 (1993).
20. Jørgensen, A. L., Kølvraa, S., Jones, C. & Bak, A. L. A subfamily of alphoid repetitive DNA shared by the NOR-bearing human chromosomes 14 and 22. *Genomics* **3,** 100–109 (1988).
21. Peng, Z. *et al.* Correlation between frequency of non-allelic homologous recombination and homology properties: evidence from homology-mediated CNV mutations in the human genome. *Hum. Mol. Genet.* **24,** 1225–1233 (2015).
22. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9,** 5233 (2019).
23. Helena Mangs, A. & Morris, B. J. The Human Pseudoautosomal Region (PAR): Origin, Function and Future. *Curr. Genomics* **8,** 129–136 (2007).
24. Cherven, K. *Network Graph Analysis and Visualization with Gephi*. (Packt Publishing Ltd, 2013). at <https://play.google.com/store/books/details?id=-bP4AAAAQBAJ>
25. Garrison, E., Guarracino, A., Heumos, S., Hagmann, J. & petersudmant. *pggb: the PanGenome Graph Builder*. (2022). doi:10.5281/zenodo.6949381

26. Garrison, E. & Guarracino, A. Unbiased pangenome graphs. *bioRxiv* 2022.02.14.480413 (2022). doi:10.1101/2022.02.14.480413
27. Eizenga, J. M. *et al.* Pangenome Graphs. *Annu. Rev. Genomics Hum. Genet.* **21,** 139–162 (2020).
28. Li, H., Feng, X. & Chu, C. The design and construction of reference pangenome graphs with minigraph. *Genome Biol.* **21,** 265 (2020).
29. Armstrong, J. *et al.* Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587,** 246–251 (2020).
30. Marco-Sola, S., Moure, J. C., Moreto, M. & Espinosa, A. Fast gap-affine pairwise alignment using the wavefront algorithm. *Bioinformatics* **37,** 456–463 (2021).
31. Marco-Sola, S. *et al.* Optimal gap-affine alignment in O(s) space. *bioRxiv* 2022.04.14.488380 (2022). doi:10.1101/2022.04.14.488380
32. Sullivan, B. A., Jenkins, L. S., Karson, E. M., Leana-Cox, J. & Schwartz, S. Evidence for structural heterogeneity from molecular cytogenetic analysis of dicentric Robertsonian translocations. *Am. J. Hum. Genet.* **59,** 167–175 (1996).
33. Cheng, E. Y. & Naluai-Cecchini, T. FISHing for acrocentric associations between chromosomes 14 and 21 in human oogenesis. *Am. J. Obstet. Gynecol.* **190,** 1781–5; discussion 1785–7 (2004).
34. Fischer, C. & Garrison, E. *chfi/gfaestus: a pangenome graph browser*. (2022). doi:10.5281/zenodo.6954036
35. Guarracino, A., Heumos, S., Nahnsen, S., Prins, P. & Garrison, E. ODGI: understanding pangenome graphs. *Bioinformatics* (2022). doi:10.1093/bioinformatics/btac308
36. Kaniecki, K., De Tullio, L. & Greene, E. C. A change of view: homologous recombination at single-molecule resolution. *Nat. Rev. Genet.* **19,** 191–207 (2018).
37. Logsdon, G. A. *et al.* The structure, function and evolution of a complete human chromosome 8. *Nature* **593,** 101–107 (2021).

38. Arenas, M. The importance and application of the ancestral recombination graph. *Front. Genet.* **4,** 206 (2013).
39. Veerappa, A. M., Padakannaya, P. & Ramachandra, N. B. Copy number variation-based polymorphism in a new pseudoautosomal region 3 (PAR3) of a human X-chromosome-transposed region (XTR) in the Y chromosome. *Funct. Integr. Genomics* **13,** 285–293 (2013).
40. Cotter, D. J., Brotman, S. M. & Wilson Sayres, M. A. Genetic Diversity on the Human X Chromosome Does Not Support a Strict Pseudoautosomal Boundary. *Genetics* **203,** 485–492 (2016).
41. Paten, B. *et al.* Superbubbles, Ultrabubbles, and Cacti. *J. Comput. Biol.* **25,** 649–663 (2018).
42. Roberts, P. A. Difference in the Behaviour of Eu- and Hetero-chromatin: Crossing-over. *Nature* **205,** 725–726 (1965).
43. Gonzalez, I. L. & Sylvester, J. E. Human rDNA: evolutionary patterns within the genes and tandem arrays derived from multiple chromosomes. *Genomics* **73,** 255–263 (2001).
44. Ahuja, J. S., Harvey, C. S., Wheeler, D. L. & Lichten, M. Repeated strand invasion and extensive branch migration are hallmarks of meiotic recombination. *Mol. Cell* **81,** 4258–4270.e4 (2021).
45. Guissani, U., Facchinetti, B., Cassina, G. & Zuffardi, O. Mitotic recombination among acrocentric chromosomes' short arms. *Ann. Hum. Genet.* **60,** 91–97 (1996).
46. Bandyopadhyay, R. *et al.* Mosaicism in a patient with Down syndrome reveals post-fertilization formation of a Robertsonian translocation and isochromosome. *Am. J. Med. Genet. A* **116A,** 159–163 (2003).
47. van Sluis, M. & McStay, B. Nucleolar DNA Double-Strand Break Responses Underpinning rDNA Genomic Stability. *Trends Genet.* **35,** 743–753 (2019).
48. Guarracino, A., Mwaniki, N., Marco-Sola, S. & Garrison, E. *wfmash: a pangenome-scale*

pairwise aligner. (2021). doi:10.5281/zenodo.6949373

49. igraph citation info. at <https://cran.r-project.org/web/packages/igraph/citation.html>
50. Bastian, M., Heymann, S. & Jacomy, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. *ICWSM* **3,** 361–362 (2009).
51. Belbasi, M., Blanca, A., Harris, R. S., Koslicki, D. & Medvedev, P. The minimizer Jaccard estimator is biased and inconsistent. *Bioinformatics* **38,** i169–i176 (2022).
52. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
53. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
54. Rautiainen, M. & Marschall, T. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biol.* **21,** 253 (2020).

Human Pangenome Reference Consortium

Haley J. Abel[1], Lucinda L Antonacci-Fulton[2], Mobin Asri[3], Gunjan Baid[4], Carl A. Baker[5], Anastasiya Belyaeva[4], Konstantinos Billis[6], Guillaume Bourque[7,8,9], Silvia Buonaiuto[10], Andrew Carroll[4], Mark JP Chaisson[11], Pi-Chuan Chang[4], Xian H. Chang[3], Haoyu Cheng[12,13], Justin Chu[12], Sarah Cody[2], Vincenza Colonna[10,14], Daniel E. Cook[4], Omar E. Cornejo[15], Mark Diekhans[3], Daniel Doerr[16], Peter Ebert[16], Jana Ebler[16], Evan E. Eichler[5,17], Jordan M. Eizenga[3], Susan Fairley[6], Olivier Fedrigo[18], Adam L. Felsenfeld[19], Xiaowen Feng[12,13], Christian Fischer[14], Paul Flicek[6], Giulio Formenti[18], Adam Frankish[6], Robert S. Fulton[2], Yan Gao[20], Shilpa Garg[21], Erik Garrison[14], Carlos Garcia Giron[6], Richard E. Green[22,23], Cristian Groza[24], Andrea Guarracino[25], Leanne Haggerty[6], Ira Hall[26,27], William T Harvey[5], Marina Haukness[3], David Haussler[3,17], Simon Heumos[28,29], Glenn Hickey[3], Kendra Hoekzema[5], Thibaut Hourlier[6], Kerstin Howe[30], Miten Jain[31], Erich D. Jarvis[32,17], Hanlee P. Ji[33], Alexey Kolesnikov[4], Jan O. Korbel[34], Sergey Koren[41], Jennifer Kordosky[5], HoJoon Lee[33], Alexandra P. Lewis[5], Heng Li[12,13], Wen-Wei Liao[2,35,26], Shuangjia Lu[26], Tsung-Yu Lu[11], Julian K. Lucas[3], Hugo Magalhães[16], Santiago Marco-Sola[36,37], Pierre Marijon[16], Charles Markello[3], Tobias Marschall[16], Fergal J. Martin[6], Ann McCartney[41], Jennifer McDaniel[38], Karen H. Miga[3], Matthew W. Mitchell[39], Jean Monlong[3], Jacquelyn Mountcastle[18], Katherine M. Munson[5], Moses Njagi Mwaniki[40], Maria Nattestad[4], Adam M. Novak[3], Sergey Nurk[41], Hugh E. Olsen[3], Nathan D. Olson[38], Benedict Paten[3], Trevor Pesout[3], Adam M. Phillippy[41], David Porubsky[5], Alice B. Popejoy[52], Pjotr Prins[14], Daniela Puiu[42], Mikko Rautiainen[41], Allison A Regier[2], Arang Rhie[41], Samuel Sacco[43], Ashley D. Sanders[44], Valerie A. Schneider[45], Baergen I. Schultz[19], Kishwar Shafin[4], Jonas A. Sibbesen[46], Jouni Sirén[3], Michael W. Smith[19], Heidi J. Sofia[19], Ahmad N. Abou Tayoun[47,48], Françoise Thibaud-Nissen[45], Chad Tomlinson[2], Francesca Floriana Tricomi[6], Flavia Villani[14], Mitchell R. Vollger[5,49], Justin Wagner[38], Brian Walenz[41], Ting Wang[50], Jonathan M. D. Wood[30], Aleksey V. Zimin[42,51], Justin M. Zook[38]

1 Division of Oncology, Department of Internal Medicine, Washington University School of Medicine, St. Louis, MO 63110, USA
2 McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO 63108, USA
3 UC Santa Cruz Genomics Institute, University of California, Santa Cruz, 1156 High St, Santa Cruz, CA, USA
4 Google LLC, 1600 Amphitheater Pkwy, Mountain View, CA 94043, USA
5 Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA 98195, USA
6 European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Cambridge, CB10 1SD, UK
7 Department of Human Genetics, McGill University, Montreal, Québec H3A 0C7, Canada
8 Canadian Center for Computational Genomics, McGill University, Montreal, Québec H3A 0G1, Canada
9 Institute for the Advanced Study of Human Biology (WPI-ASHBi), Kyoto University, Kyoto 606-8501, Japan
10 Institute of Genetics and Biophysics, National Research Council, Naples 80111, Italy
11 University of Southern California, Quantitative and Computational Biology, 3551 Trousdale, Pkwy, Los Angeles, CA, USA
12 Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA 02215, USA
13 Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02215, USA
14 Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN 38163, USA
15 School of Biological Sciences, Washington State University, Pullman WA 99163, USA
16 Institute for Medical Biometry and Bioinformatics, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
17 Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA

18 The Vertebrate Genome Laboratory, The Rockefeller University, New York, NY 10065, USA
19 National Institutes of Health (NIH)–National Human Genome Research Institute, Bethesda, MD, USA
20 Center for Computational and Genomic Medicine, The Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA.
21 Department of Biology, University of Copenhagen, Denmark
22 Department of Biomolecular Engineering, University of California, Santa Cruz, 1156 High St., Santa Cruz, CA 95064, USA
23 Dovetail Genomics, Scotts Valley, CA 95066, USA
24 Quantitative Life Sciences, McGill University, Montreal, Québec H3A 0C7, Canada
25 Genomics Research Centre, Human Technopole, Milan 20157, Italy
26 Department of Genetics, Yale University School of Medicine, New Haven, CT 06510, USA
27 Center for Genomic Health, Yale University School of Medicine, New Haven, CT 06510, USA
28 Quantitative Biology Center (QBiC), University of Tübingen, Tübingen 72076, Germany
29 Biomedical Data Science, Department of Computer Science, University of Tübingen, Tübingen 72076, Germany
30 Tree of Life, Wellcome Sanger Institute, Hinxton, Cambridge, CB10 1SA, UK
31 Northeastern University, Boston, MA 02115, USA
32 The Rockefeller University, New York, NY 10065, USA
33 Division of Oncology, Department of Medicine, Stanford University School of Medicine, Stanford, CA, 94305, USA
34 European Molecular Biology Laboratory, Genome Biology Unit, Meyerhofstr. 1, 69117 Heidelberg, Germany
35 Department of Medicine, Washington University School of Medicine, St. Louis, MO 63110, USA
36 Computer Sciences Department, Barcelona Supercomputing Center, Barcelona, Spain
37 Departament d'Arquitectura de Computadors i Sistemes Operatius, Universitat Autònoma de Barcelona, Barcelona, Spain
38 Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20877, USA
39 Coriell Institute for Medical Research, Camden, NJ 08103, USA
40 Department of Computer Science, University of Pisa, Pisa 56127, Italy
41 Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA
42 Department of Biomedical Engineering, Johns Hopkins University, Baltimore 21218, MD, USA
43 Department of Ecology & Evolutionary Biology, University of California, Santa Cruz, 1156 High St, Santa Cruz, CA, USA
44 Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Berlin, Germany
45 National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA
46 Center for Health Data Science, University of Copenhagen, Denmark
47 Al Jalila Genomics Center of Excellence, Al Jalila Children's Specialty Hospital, Dubai, UAE
48 Center for Genomic Discovery, Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai, UAE
49 Division of Medical Genetics, University of Washington School of Medicine, Seattle, WA 98195, USA
50 Department of Genetics, Washington University School of Medicine, St. Louis, MO 63110, USA
51 Center for Computational Biology, Johns Hopkins University, Baltimore, MD 21218, USA
52 Department of Public Health Sciences, University of California, Davis, One Shields Avenue, Medical Sciences 1C, Davis, CA 95616