Uncovering Functional Sequence Gaps in Human Reference Genomes using African Pan Genome Contig Sequences

Rachel Martini[1,2*], Abdulfatai Tijjani[3,4*], Kyriaki Founta[3,4,5], Daniel Cha[6], Sebastian Maurice[7], Jason A. White[1,2], Onyinye Balogun[8], Melissa B. Davis[1,2,9,10+], Nyasha Chambwe[3,4,5+]
[1]Institute of Translational Genomic Medicine, Morehouse School of Medicine, Atlanta, GA
[2]Department of Microbiology, Biochemistry & Immunology, Morehouse School of Medicine, Atlanta, GA
[3]Northwell, New Hyde Park, NY
[4]Feinstein Institutes for Medical Research, Northwell Health, Manhasset, NY
[5]Donald and Barbara Zucker School of Medicine at Hofstra/Northwell, Hempstead, NY
[6]Carnegie Mellon University, Pittsburgh, PA
[7]City College of New York, New York, NY
[8]Weill Cornell Medicine, New York, NY
[9]Georgia Research Alliance, Atlanta, GA
[10]New York Genome Center, New York, NY
[*]*denotes equal contribution*
[+]*corresponding authors*

CRediT author contributions

Conceptualization: RM, AT, KF, MBD, NC
Methodology: RM, DC, AT, MBD, NC
Software, Formal analysis: RM, AT, KF, DC, SM
Data Curation: RM, AT, KF, DC
Visualization: RM, AT, KF, DC
Writing - Original Draft: RM, AT, KF, NC
Writing - Review & Editing: RM, AT, KF, DC, SM, JW, OB, MBD, NC
Supervision: RM, NC, MBD
Project administration: RM, NC
Funding acquisition: RM, NC, OB, MBD

Abstract

Predominantly used human reference genomes, including GRCh38 and the gapless T2T-CHM13 references, remain limited in their representation of African genomic diversity. We analyzed African Pan-Genome (APG) contig sequences representing 296.5 Mb of African-ancestry-specific sequence not represented in GRCh38 to assess their representation and functional potential in newer long-read assemblies. Alignments to T2T-CHM13 and the 47 Human Pangenome Reference Consortium (HPRC) linear assemblies positioned 40% and 83% of APG contigs, respectively, with high identity and coverage. Most T2T-CHM13 placements corresponded to sequences absent from GRCh38 (94.5%) and were enriched in centromeric and satellite repeats (94.2%). Functional overlap included annotated genes (2.6%) and CpG islands (3.6%), with enrichment in immune, synaptic and intracellular signaling pathways. HPRC alignments revealed ancestry-associated patterns, with African and Admixed American genomes showing the highest numbers of unique contig placements and shared alignments. A subset of 742 APG contigs showed weak or no mapping to any of the reference assemblies. These contigs were not enriched for repeat elements, and ~58% showed predicted gene content. These findings highlight persistent gaps in even the most complete reference genomes and underscore the importance of incorporating ancestry-enriched sequences into future genome frameworks to reduce reference bias and advance equitable discovery in genomics.

Keyword/Indexing Terms: pangenome, alignment, ancestry, reference genome, repeat elements

Introduction

Despite the recent release of gapless and more diverse reference genome assemblies, GRCh37 and GRCh38 remain the predominant backbones for sequence alignment and downstream analytical pipelines in human genomics and clinical sequencing pipelines[1,2]. GRCh38 contains unresolved gaps comprising approximately 7% of the genome[3,4], and critically, both GRCh37 and GRCh38 reflect a limited view of human diversity, composed overwhelmingly of European ancestry[5]. This Eurocentric reference bias has broad implications for genomic medicine, as sequencing read alignment, variant interpretation, gene annotation and downstream analyses are all influenced by the structure and content of the reference genome[1,2,6]. One of the clearest examples of reference bias in genomic research is genome-wide association studies (GWAS), the majority of which have been conducted in populations of European ancestry[7,8]. GWAS findings often fail to replicate in populations that differ ancestrally from the original discovery cohorts[9,10]. This lack of transferability is due in large part to reference bias, where ancestry-associated risk variants may be missed entirely or misinterpreted when the reference genomes do not adequately represent haplotypes present in diverse populations. Additionally, GWAS of individuals with African ancestry have largely focused on African American populations[7,11]; however, even studies of this admixed group are limited in their ability to capture the full breadth of genetic diversity, as African genomes are the most genetically diverse among all human populations[11,12]. As a result, risk loci, regulatory regions and even gene structures may be incompletely mapped in underrepresented populations, with far-reaching implications from disease etiology studies to routine implementation of genomic medicine.

Recent long-read sequencing initiatives have begun to address these reference genome limitations. The complete Telomere-to-Telomere (T2T-CHM13) assembly provides a gapless linear reference genome[3]. At the same time, the Human Pangenome Reference Consortium (HPRC) has constructed individual genome assemblies and a graph-based reference using multiple, high-coverage long-read sequencing from 47 individuals from diverse ancestral backgrounds[4]. These newer assemblies improve representation of previously unresolved regions, such as centromeres[3], short tandem repeats[13], copy number variations[14], and structural variants[4], and offer better sensitivity in variant calling across diverse populations[15]. Despite these advances, GRCh38 remains the commonly used reference in research and clinical workflows, in part due to a lack of standardized tools and practices for analyzing data with alternative references.

GRCh38, and to a greater extent T2T-CHM13, show limited ancestral diversity[5,15], reinforcing concerns that current genomic interpretations may overlook ancestry-associated variants or haplotypes. Critically, the question remains: how well do these newer references capture the breadth of genomic variation from populations underrepresented in genomics research?

To address this gap, prior to the release of the T2T-CHM13 and HPRC reference genomes, a landmark study conducted whole genome sequencing on 910 individuals of African descent, identifying 296.5 million base pairs absent from GRCh38[16]. These sequences were assembled into 125,715 contigs, collectively referred to as the African Pan Genome (APG). While these contigs revealed previously uncharacterized genomic content specific to individuals of African ancestry, their biological relevance and overlap with newer reference builds remained unclear. In this study, we sought to determine whether

APG sequences, absent on GRCh38, are represented in the updated T2T-CHM13 and HPRC linear reference genomes, and whether they correspond to novel, functionally relevant regions of the genomes. By evaluating the representation and potential functional relevance of these sequences, our work underscores that continued reliance on incomplete and Eurocentric reference genomes risks missing ancestry-associated genomic content with potential biological and clinical relevance.

Results

Study overview

The African Pan Genome contig sequences were published prior to the recent release of the long-read sequencing-based T2T-CHM13 reference genome and HPRC linear reference assemblies. We therefore sought to determine if these APG contig sequences assembled by Sherman et. al[16] were present in gaps filled by newer reference genome releases and or assemblies built primarily on the backbone of long-read sequencing data **(Figure 1, SF1)**. Across all reference genomes and assemblies, alignment results were classified as "Reasonably Good" (RG, ≥ 50% coverage and ≥ 80% identity) or "Nearly Perfect" (NP, ≥ 80% coverage and ≥ 90% identity) as described previously[16] (see **Methods**).

APG contigs map primarily to T2T-CHM13 regions unique from GRCh38

Of the 124,240 APG contig sequences we analyzed (**ST1**, see **Methods**), 87.74% (108,423) had RG alignments, and 39.71% (49,070) had NP alignments to T2T-CHM13 **(Table 1, Figure 2A, ST2)**. For both RG and NP alignments, more than 90% of contigs mapped to T2T-CHM13 regions unique from GRCh38 (92.6% and 94.5%, respectively) **(Figure 2B)**. We report the highest number of alignments to T2T-CHM13 Chromosomes 9, 21 and 14, and the least number of alignments to Chromosomes 18, 16 and Y **(Figure 2C)**. This is overwhelmingly driven by mapping of previously unplaced APG contigs, where neither end of the contig sequence could be resolved to a location on GRCh38[16] **(Figure 2D).**

We assessed alignment metrics to T2T-CHM13 in the context of whether the APG contigs were Two End Placed (TE, n = 272), One End Placed (OE, n = 1,246), or unplaced (n = 122,054) on GRCh38, where OE or TE placed contigs had one or both ends of the contig placement determined on the GRCh38 reference genome, respectively[16] **(ST3)**. Unplaced contigs showed the highest level of RG alignments (87.39%), followed by OE (81.06%) and TE (63.58%). In contrast, unplaced contigs had the lowest proportion meeting the NP threshold (39.31%), where TE and OE had similar NP alignment rates (51.32% and 55.06%, respectively) **(ST3)**.

Given that the OE and TE placed contigs had reference locators on the GRCh38 genome, we sought to determine if alignments to T2T-CHM13 mapped to the same chromosome. For TE placed contigs with NP alignments, we report only one contig where the chromosome placement on GRCh38 did not match the T2T-CHM13 alignment output **(Figure 2E)**. Three TE placed contigs that had been placed on GRCh38 unknown chromosomes mapped to T2T-CHM13 Chromosomes 15 and 22 (n = 2, n = 1,

respectively). However, for OE placed contigs, we report that 53.21% (n = 365) of OE contigs with NP alignments to T2T-CHM13 were discordant with the GRCh38 OE chromosomal placement **(Figure 2F)**. The majority of discordant calls are largely driven by groupings of OE contig discordant calls between the acrocentric chromosomes (64.89%, n = 231), where the short arms of these chromosomes were largely unsequenced before the release of the T2T-CHM13 reference genome[3]. T2T-CHM13 chromosomes 22, 15 and 14 had the most discordant OE contig alignments (n = 80, n = 57 and n= 44, respectively), driven by mapping to acrocentric chromosomes where 16 contigs OE placed on GRCh38 chromosome 14 map to T2T-CHM13 chromosome 22, 13 contigs OE placed on GRCh38 chromosome 21 map to T2T-CHM13 chromosome 15, 10 contigs OE placed on GRCh38 chromosome 22 map to T2T-CHM13 chromosome 15 **(Figure 2F)**. In line with our expectations, we find that the vast majority of APG contigs map to the T2T-CHM13 genome at locations that represent novel sequences not represented in earlier versions of the reference genome.

APG contigs predominantly overlap functional regions in T2T-CHM13 resolved gaps

Based on the observation that the overwhelming majority of the contigs (94.5%) mapped to T2T-CMH13 unique regions **(Figure 2B)**, we sought to determine the overlap of APG contig sequence alignments to T2T-CHM13 functional annotations. 94.2% of NP alignment contigs to T2T-CMH13 mapped to centromeres and satellite repeats. A small proportion of the contigs were annotated to genes (2.6%), CpG islands (3.6%) and composite elements (4.4%) **(Figure 3A) (Table 2).** For the RG-aligned contigs, we report similar enrichment patterns across T2T-CHM13 elements **(Table 2)**.

For contigs that mapped to T2T-CMH13 coding regions (373 genes), we found an enrichment for genes involved in biological processes related to antigen presentation via the MHC class II protein complex across all Gene Ontology (GO) categories **(Figure 3B)**. Enrichment of genes involved in neuron recognition and regulation/modulation of synaptic signaling was also observed across cellular component and molecular function processes **(Figure 3B)**. Interestingly, the enrichment across MHC class II processes was driven by a single region on chromosome 6, specifically the HLA locus that displayed a high number of annotated contigs. 84 contigs mapped to this 170Kb region on chromosome 6, while 44 of these mapped to 4 HLA genes (*HLA-DRB3*, *HLA-DRB1*, *HLA-DQA1*, *HLA-DQRB1*) **(Figure 3A)**. The main driver of contig overrepresentation across genes involved in neuronal signaling processes was *PPFIA1*, with 33 contigs mapping to it. Interestingly, 33 of the contigs that mapped to genes overlap with previously reported GWAS hits, while 60 of the mapped genes have been linked to at least one phenotype or disease in the Online Mendelian Inheritance in Man[17] (**ST4-5**).

APG contig alignments to HPRC assemblies reveal population-associated enrichment patterns

Next, we aligned the APG contigs to the 47 haploid paternal assemblies from the Human Pangenome Reference Consortium (HPRC)[4] based on the rationale that these assemblies represent a more diverse range of population origins, including complete genomes from African populations. Compared to our mapping result to T2T-CHM13, we report a significant improvement in the mapping of APG contigs, where 99.40% APG contigs mapped RG, and 82.91% mapped NP to at least one HPRC assembly (n = 123,498 and n = 103,003, respectively) **(Table 3)**. Twenty-four of the 47 HPRC linear assemblies are from individuals of African ancestry (AFR, **ST6**). We found that 80.44% of APG contigs have at least 1 NP mapping to at least one of the AFR ancestry HPRC assemblies (n = 99,944) **(Table 3)**. Across all HPRC

linear assemblies, we also report increased RG and NP mappings of TE, OE and unplaced APG contigs, where for NP, we see 95.03% of TE, 94.06% of OE, and 82.76% of unplaced APG contigs mapping to at least one HPRC assembly (n = 287, n = 1172, n = 101,544, respectively) **(ST7-9)**.

Comparing our APG contig NP alignment results between reference assemblies, 49,020 contigs mapped to both T2T-CHM13 and the HPRC linear assemblies, whereas 53,983 APG contigs had an NP mapping to the HPRC assemblies only **(Figure 4A)**. Of these 58,983 NP HPRC-mapped only, there is a notable enrichment of APG contig alignments within the AFR superpopulation (n = 7,969). In contrast, alignments were less frequent in the other reference superpopulations **(Figures 4B, C)**. The most significant overlap between any two superpopulations consisted of 8,006 contigs that aligned to both the AFR and Admixed American (AMR) populations. Furthermore, unsupervised clustering of alignment patterns emphasized contigs that were exclusively or predominantly mapped to the AFR and AMR populations **(Figure 4B)**. AFR and AMR assemblies are enriched in the HPRC 47 linear assembly set (n = 24 and n = 16, respectively), highlighting the population-associated nature of these sequences and pointing to the presence of shared ancestry signals or structural variations between these groups. In contrast, shared mappings to the EAS, EUR, and SAS populations were considerably smaller, possibly indicating a limited representation of APG contigs within these genomes **(Figures 4B, C)**.

To validate the observed enrichment of APG contig alignments in African HPRC assemblies, we compared the likelihood of APG contigs aligning in AFR assemblies relative to other superpopulations (**Figure 4D**). We report a significant overrepresentation of APG contig alignments in AFR genomes ($p <$ 2.2 e-16), where the OR of contig alignment in AFR compared to AMR was 1.53, indicating a 53% higher likelihood of contigs aligning with AFR genomes. Comparisons with other populations revealed even greater enrichment (ORs: AFR vs. EAS = 2.74, AFR vs. EUR = 7.23, AFR vs. SAS = 8.55), highlighting a significantly higher likelihood of APG contig alignment in AFR genomes compared to those of European and South Asian populations. Hierarchical clustering of the alignment patterns revealed ancestry-associated groupings among individuals, with AFR samples clustering together and AMR samples forming a separate cluster **(Figure 4E)**.

Sequence Characteristics and Functional Landscape of Below-threshold APG Contigs

21,187 APG contigs failed to map to the T2T-CHM13 and 47 HPRC linear reference assemblies using a strict "nearly perfect" alignment threshold, defined as coverage ≥ 80% and sequence identity ≥ 90% (**Figure 4A**). We further evaluated the mappability of these contigs under a more relaxed “reasonably good” alignment threshold (coverage ≥ 50%, identity ≥ 80%). We found that 14,796 contigs (~25 Mbp) aligned to both the T2T and HPRC assemblies, while 5,649 contigs (10.4 Mbp) aligned only to HPRC assemblies. The remaining 742 contigs, totaling ~1.5 Mbp, either failed to align or were aligned at below the relaxed alignment threshold. Thus, we designated these contigs as “Below-threshold contigs” (**Figure 5A, ST10**).

To determine whether repeat content influenced this mappability profile, we annotated repeat elements across all three contig subsets using RepeatMasker (**Table 4**). Contigs mapped to T2T and HPRC were highly enriched for repetitive sequences, with over 93% of their total length composed of repeats (**Figure 5B**). These mapped contigs primarily comprised satellite repeats (~74%) and simple repeats (~17–19%), typical of centromeric and subtelomeric regions. Conversely, the 742 unmapped contigs had an average

repeat content of around 19%, with more than 80% of them lacking detectable repeats (**Figure 5B, Table 4**). LINEs (3.2%), SINEs (1.9%), and LTR elements (0.89%) were the most prevalent repeats detected. To better understand the sequence features of the 742 below-threshold contigs (BTC), we examined the relationship between GC content and contig length, annotating each contig by its repeat proportion and the number of distinct repeat classes (**Figure 5C**). We found that the overall GC content varied widely from approximately 20% to 70%, indicating a heterogeneous base composition with no strong correlation to contig length or repeat content. To further determine the functional potential of these sequences, *ab initio* gene and CpG island (CGI) prediction revealed that 58.5% of these contigs contained predicted gene sequences, 0.3% contained predicted CGIs, 3.6% exhibited both predicted genes and CGIs, and the remaining 37.6% did not show any predicted features **(Figure 5D)**. Taken together, this suggests that these unmapped contigs may represent gene-rich regions and/or regulatory features absent from current reference datasets, while some of the sequence remains yet unknown or could represent sequencing errors.

The spatial distribution of predicted genes and CGIs along the normalized contig length revealed distinct yet overlapping patterns (**Figure 5E**). Both features were most enriched near the start of the contigs, consistent with the well-established association of high-CGI promoters and transcription start sites in gene-rich regions[18,19]. The density of both features also declines towards the 3' ends. Predicted genes exhibited an additional density peak at the mid-contig region, whereas CGIs showed a steady decline after the initial peak, with minimal enrichment towards distal contig ends. Additionally, we assessed the co-occurrence of predicted genes and CGIs on a per-contig basis. Most contigs exhibited modest numbers of either feature; however, the longest contig (CAAPA_25902) showed the highest enrichment of both features. This contig spans ~50 kb and contains multiple predicted genes (n = 8) and CGIs (n = 24), indicating the presence of regulatory-rich genomic segments **(Figure 5F)**. Along this long contig, CGIs are distributed throughout the contig, often upstream of genes but also within intergenic regions, suggesting the presence of both proximal and distal regulatory elements (**Figure 5G**).

Discussion

In this study, we examined the genetic sequences absent from GRCh38 that were identified within an African ancestry cohort, a highly admixed population reflecting African, European, and Native American ancestry[16,20]. Our findings highlight the persistent and significant gaps in current human reference genomes, particularly for genomic sequences enriched in African and admixed populations, which are partially driven by repetitive sequences.

The failure of some APG contigs to align completely to T2T-CHM13 highlights that African populations may harbor ancestry-specific sequences that remain missing from the current reference assemblies. While T2T-CHM13 has enhanced read mappability, particularly in previously inaccessible regions, these improvements do not fully resolve biases affecting non-European genomes. Thus, it remains evident that no single linear reference can adequately represent the full spectrum of human genomic variation[11,12,21]. While structural gaps in reference genomes are narrowing, functional gaps persist. Specifically, genes we found within these regions are involved in critical biological processes, such as immune function, synaptic signaling, and intracellular transport, and are not only evolutionarily conserved but also functionally diverse. The identification of genes and pathways previously implicated in GWAS studies, or

associated with known disease phenotypes, suggests that the historical underrepresentation of African genomes in reference construction may have led to the systematic omission of biologically relevant regions[8,16]. It is also essential to note that genome assembly is distinct from functional annotation[22]. The recovery of such genes within the context of APG contigs underscores both the limitations of earlier assemblies and the potential for discovering functional elements pertinent to disease, population-specific traits, and regulatory biology. These findings highlight that even a small subset of previously unmapped contigs may contain genes involved in essential biological pathways, thereby supporting ongoing efforts to enhance reference resources that account for both structural and functional diversity.

When using the more ancestrally diverse set of 47 HPRC linear assemblies, we observed an approximately twofold increase in APG contig mappings compared to T2T-CHM13, with notable enrichment in AFR and AMR assemblies, thereby emphasizing the need for population-specific reference resources. Moreover, the approximately 21,000 contigs that did not align nearly perfectly to any of these references still possess functional signatures. We report increased APG contig mapping at the RG level to the HPRC linear assemblies, demonstrating the different sensitivities of the two reference assemblies in capturing the full spectrum of African genomic diversity, where the higher alignment rate to HPRC may reflect better representation of global haplotypes and reduced reference bias in the pangenome structure[4,22].

APG Contigs that mapped NP or RG to either T2T-CHM13 or HPRC linear assemblies were highly enriched in repeat elements (> 90% sequence content), whereas the smaller subset of 742 below-threshold contigs showed limited detected repeat content (19%). This indicates that the reference assemblies effectively encompass sequences with satellite and tandem repeats, likely due to improved telomere-to-telomere assembly[23]. However, the non-satellite repeats and repeat-poor regions, as seen in the below-threshold contigs, may reflect population-specific insertions, structural variants, or gaps in current references, especially for African genomes[24,25]. Functional annotation of these below-threshold contigs also showed predicted gene content among 58% of contigs, with 3.6% of contigs having both predicted genes and CGIs. The co-occurrence and spatial proximity of these features indicate the presence of intact promoter architectures, many of which align with established gene regulatory models, such as CpG island promoters at transcription start sites (TSSs)[19,26]. These features were more common near the ends of contigs, with density plots showing a distinct double-peaked distribution of predicted genes. This arrangement is consistent with complex, bidirectional regulatory domains. The observed architecture also coincides with multi-gene loci and open chromatin regions marked by CGIs, which are often linked to housekeeping genes and broadly expressed loci[18,27]. Specifically, the co-occurrence of gene-dense regions and CpG-rich regulatory features in CAAPA_25902 suggests that this contig likely represents a functionally important genomic segment that is absent from current reference assemblies. The structural novelty and feature density underscore the need to integrate such sequences into future reference models to improve the representation of diverse African haplotypes. These results demonstrate that these unmapped sequences encompass both repetitive and non-repetitive elements and underscore the limitations of even the most comprehensive linear references in capturing the full spectrum of genomic diversity, particularly among underrepresented populations[28,29].

It is likely that a graph-based pangenome model that enables the representation of multiple haplotypes and alleles in a unified framework, offers a more promising alternative, as it reduces reference bias and increases the potential to uncover novel, clinically relevant variation across diverse ancestries[22].

However, bridging the gap between ancestrally diverse pangenomes and practical implementation will be essential to ensure that advances in genomic representation translate into equitable insights across diverse populations[30]. This study provides foundational evidence for incorporating ancestry-enriched, functionally significant sequences into future reference frameworks. The inclusion of African-specific sequences is not merely a technical enhancement; it is foundational to enabling accurate and globally relevant genomic medicine[31,32]. However, a major limitation lies in the current infrastructure of genomics research, as the overwhelming majority of analytic pipelines, variant annotation tools, and reference-based workflows remain tightly coupled to the GRCh38 reference genome. This dependence has created inertia within the field, as researchers may be reluctant to transition to alternative references due to a lack of compatibility, comprehensive annotations, and stable tool sets. Although pangenome models offer substantial advantages, particularly for detecting complex structural variation and interrogating highly polymorphic regions, the tools required to implement them are still in development. Many remain inaccessible to the broader research community due to computational complexity, a lack of standardization, and uncertainty about how to reconcile graph-based results with the broader existing knowledge built on linear reference frameworks. Nonetheless, our findings support recent initiatives that advocate for graph-based reference genomes and ancestry-aware genomic models[33,34].

Several limitations should be considered when interpreting our findings. Although APG contigs provide a valuable resource for capturing ancestry-associated sequence content, the high levels of admixture within the cohort may affect the generalizability of alignment results to specific African ancestry populations. Our analyses did not include native mapping of short-read whole-genome sequencing data to the T2T-CHM13 or HPRC linear references, which would more directly assess improvements in alignment. Moreover, our approach relies on existing reference assemblies that remain incomplete for certain ancestries and complex genomic regions, potentially leaving highly divergent APG contigs unplaced or ambiguously mapped. Additionally, mapping specificity was evaluated at broad population levels without incorporating local ancestry inference, limiting the resolution of fine-scale signals. The HPRC panel used is unevenly sampled, with an overrepresentation of African and admixed American genomes and an underrepresentation of European and Asian lineages, which may bias estimates of contig population enrichment. While we underscore the potential of graph-based references, our analyses were conducted within a linear framework, and functional enrichment assessments depended on existing gene annotations that may be incomplete or biased toward well-characterized pathways and ancestries. Additionally, predictions for unmapped contigs were based solely on computational models, lacking transcriptomic or epigenomic validation, which may lead to overestimation of their functional relevance. Future studies integrating local ancestry inference, experimental validation, and graph-based methodologies will be essential to confirm the biological significance of these sequences and to advance the development of more inclusive human references.

Continued use of GRCh38 potentially neglects biologically active regulatory domains that likely influence gene expression and disease susceptibility/plasticity and its continued use in studies of diverse populations is no longer defensible.

Methods

African Pan Genome contigs data access

We accessed the published African Pan Genome contigs sequences at GenBank with accession code PDBU01000000. All APG contig annotations were available via supplementary data from Sherman et al (2019)[16]. 125,715 APG contig sequences were published, however, 1475 were annotated as "dead" and were therefore excluded in all subsequent analyses (**ST1**). In all analyses described below, we proceeded with the remaining 124,240 APG contig sequences.

Mapping APG contigs to GRCh38, T2T-CHM13 and the HPRC linear references

To determine if any of the APG contig sequences are represented in the T2T-CHM13 and 47 HPRC linear reference genomes, we utilized a similar mapping and assessment approach as outlined by Sherman et al 2019[16], to assess the representation of APG contigs among unique reference genomes representative of African, South American, European, East and South Asian ancestries **(SF1)**. Specifically, we aligned the APG contigs to GRCh38.p10 as a benchmark to the Sherman et al analysis,[16] the T2T-CHM13 v2.0 reference genome,[3] and the HPRC Year 1 haploid paternal linear assemblies v2.[4] Alignment to the various reference genomes and assemblies was performed using bwa-mem v0.7.17 with default parameters.[35] Alignments to individual locations within at least 300 bp were "chained" and analyzed as a single mapping output.

Assessment of APG contig mapping to GRCh38, T2T-CHM13 and the HPRC linear references

To assess the quality of APG contig alignment to the various reference genomes and assemblies, we determined the coverage and identity of each alignment. Coverage is defined as the fraction of aligned length over the entire contig length. Identity is the fraction of correctly aligned bases over the aligned length. For individual alignments within 300 bp, we took the aligned length to be the sum of each alignment length, and measured identity by taking the weighted average of identities across each alignment, using the following formulae:

$$Coverage = Aligned\ length \div Query\ Sequence\ Length$$

$$Identity = (Aligned\ length - Distance) \div Aligned\ Length$$

After calculating the coverage and identity, we applied two thresholds to determine if the mapping result was a "Reasonably Good" or "Nearly Perfect" alignment.[16] A Reasonably Good alignment was defined as having greater than or equal to 50% coverage, and greater than or equal to 80% identity. $Reasonably\ Good\ Alignment = \geq 50\%\ Coverage + \geq 80\%\ Identity$

A Nearly Perfect alignment was defined as having greater than or equal to 80% coverage, and greater than or equal to 90% identity.

$$Nearly\ Perfect\ Alignment = \geq 80\%\ Coverage + \geq 90\%\ Identity$$

Using these metrics and thresholds, we determined the number of contigs passing each threshold for each reference assembly. A binary matrix was created to indicate the alignment status of each APG contig relative to T2T-CHM13 and each HPRC assembly, with values of 1 indicating alignment and 0 indicating no alignment/threshold for alignment not met.

To further assess the representativeness of APG contigs across HPRC superpopulations, the HPRC assemblies were categorized into superpopulation groups based on the designations from the 1000 Genomes Project: African (AFR), Admixed American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS).[36] For HPRC samples that were not a part of the 1000 Genomes project (HG002 and HG005), these were assigned to a corresponding superpopulation group (EUR and EAS, respectively) **(ST2).** A corresponding superpopulation-level and individual-level binary matrix was then generated, summarizing whether each contig aligned with at least one assembly within each superpopulation **(ST7-8)**. These matrices facilitated the evaluation of contig distribution and potential population-specific enrichment across HPRC assemblies. Additionally, to statistically assess the population-specific enrichment of APG contig alignments, a 2 × 5 contingency table was constructed, summarizing the counts of aligned and unaligned contigs across the superpopulations. The alignment counts were normalized according to the number of assemblies representing each superpopulation (AFR: 24, AMR: 16, EAS: 5, EUR: 1, SAS: 1) to account for the differences in sample sizes. A chi-square test of independence was conducted to determine the association between alignment status and population group. Odds ratios (ORs) were calculated for each pairwise comparison between AFR and other populations (AMR, EAS, EUR, SAS) to quantify the relative likelihood of contig alignment in AFR genomes.

Functional enrichment of APG contigs across T2T-CHM13 genomic elements

To explore the potential localization of contigs mapped to T2T-CHM13 regions within functional genomic elements, we first obtained genomic coordinates for T2T-CHM13 unmasked CpG islands, NCBI RefSeq genes, problematic regions, as well as T2T-CMH13 unique regions in comparison to GRCh38 from the UCSC Genome Browser[37] Table Browser (https://genome.ucsc.edu/cgi-bin/hgTables) in *.bed* file format. Additionally, we retrieved T2T-CHM13 RepeatMasker annotations generated by the T2T-CHM13 consortium (https://github.com/marbl/CHM13?tab=readme-ov-file). We then used BEDTools[38] v2.30.0 to identify the presence of contigs across these various T2T-CHM13 genomic elements.

For colocalization of contigs with previously reported GWAS hits, we obtained the GWAS associations catalog from the EMBL-EBI GWAS Catalog (https://www.ebi.ac.uk/gwas/docs/file-downloads). GRCh38 variant positions were converted to the T2T-CHM13 coordinate system using the annotated GWAS variant positions provided by the T2T-CHM13 consortium (https://github.com/marbl/CHM13?tab=readme-ov-file). To identify genes associated with phenotypes or diseases among those mapped by contigs, we downloaded OMIM's *Synopsis of the Human Gene Map* file (https://www.omim.org/downloads/) following licensing and registration, and extracted genes with reported phenotype or disease associations[17].

Gene ontology over-representation analysis

We extracted gene annotations using Ensembl gene models mapped to the T2T-CHM13 assembly, identifying a total of 373 genes that overlap with the regions where the APG contigs are placed. For the enrichment analysis, we utilized the clusterProfiler package[39] (version 4.14.6) in R to compare our test gene set against a background gene set composed of all annotated genes in the T2T-CHM13 assembly. We evaluated the overrepresentation of Gene Ontology (GO) terms across functional categories, including Biological Process (BP), Cellular Components (CC), and Molecular Functions (MF). Statistical significance was assessed using Benjamini-Hochberg-adjusted p-values with a false discovery rate (FDR) of less than 0.05. We generated bar plots to summarize the enriched GO terms related to the genes within the APG-contig mapped regions.

Assessment of Repetitive Elements in APG Contigs

Repetitive DNA elements in the African pan-genome (APG) contigs were identified using RepeatMasker v4.1.6, a widely-used tool for detecting and classifying interspersed repeats and low-complexity sequences[40]. The analysis was performed on three subsets of contig categories based on our Reasonably Good alignment threshold. These are: (i) those aligning to both T2T and HPRC assemblies, (ii) those aligning only to HPRC, and (iii) Belo threshold contigs (those that failed to align or are aligned to either reference at a below reasonably good threshold. RepeatMasker was executed using the Dfam 3.7 and RepBase repeat libraries[41,42], utilizing the human-specific repeat database for precise classification. The analysis was performed with the parameters -species set to human to confine the search to repeats curated for the human genome.

Gene and CpG Island Predictions and Annotation

We employed AUGUSTUS v3.5.0[43], an *ab initio* gene prediction software trained on human gene models to annotate potential protein-coding genes within the 742 below-threshold contig sequences that did not align nearly perfectly to HPRC assemblies. Contigs were individually processed, and gene structures including start and stop codons, exons, and coding sequences (CDS) were predicted. The resulting gene predictions were filtered to retain only transcripts with complete CDS features, representing putative protein-coding genes for downstream annotation (**ST11**).

Additionally, to assess the epigenetic potential of APG contig sequences, we utilized CpGplot v6.6.0.0, a component of the European Molecular Biology Open Software Suite (EMBOSS, https://www.ebi.ac.uk/Tools/seqstats/emboss_cpgplot/). This tool detects CpG-rich sequences based on windowed GC content and the observed-to-expected CpG dinucleotide ratio. We used default thresholds commonly applied in genome-wide CpG prediction: a minimum island length of 200 bp, GC content ≥ 50%, an observed/expected CpG ratio of ≥ 0.6, and a 100 bp sliding window. The identified CpG islands were filtered to show only contigs with islands, where we quantified the number of islands predicted. The output was parsed and converted to BED format for comparative positional analysis and overlay with predicted genes (**ST11**).

We performed both positional normalization and overlap analysis to assess the relative positions and relationships between predicted CpG islands and gene features. CpG and gene start coordinates were normalized by contig length to generate a relative position profile ranging from 0 (contig start) to 1 (contig end). The identified CpG islands were classified based on their genomic context relative to predicted genes.

Acknowledgements

This work was made possible by the MacMillan Family Foundation as part of the MacMillan Center for the Study of the Non-Coding Cancer Genome at the New York Genome Center. This work was performed using the CPU and GPU compute nodes of the high-performance computing (HPC) cluster at Cold Spring Harbor Laboratory (CSHL). This cluster is supported by US National Institutes of Health grant S10OD028632-01. We would like to acknowledge the Human Pangenome Reference Consortium (BioProject ID: PRJNA730823) and its funder, the National Human Genome Research Institute (NHGRI).

Figures

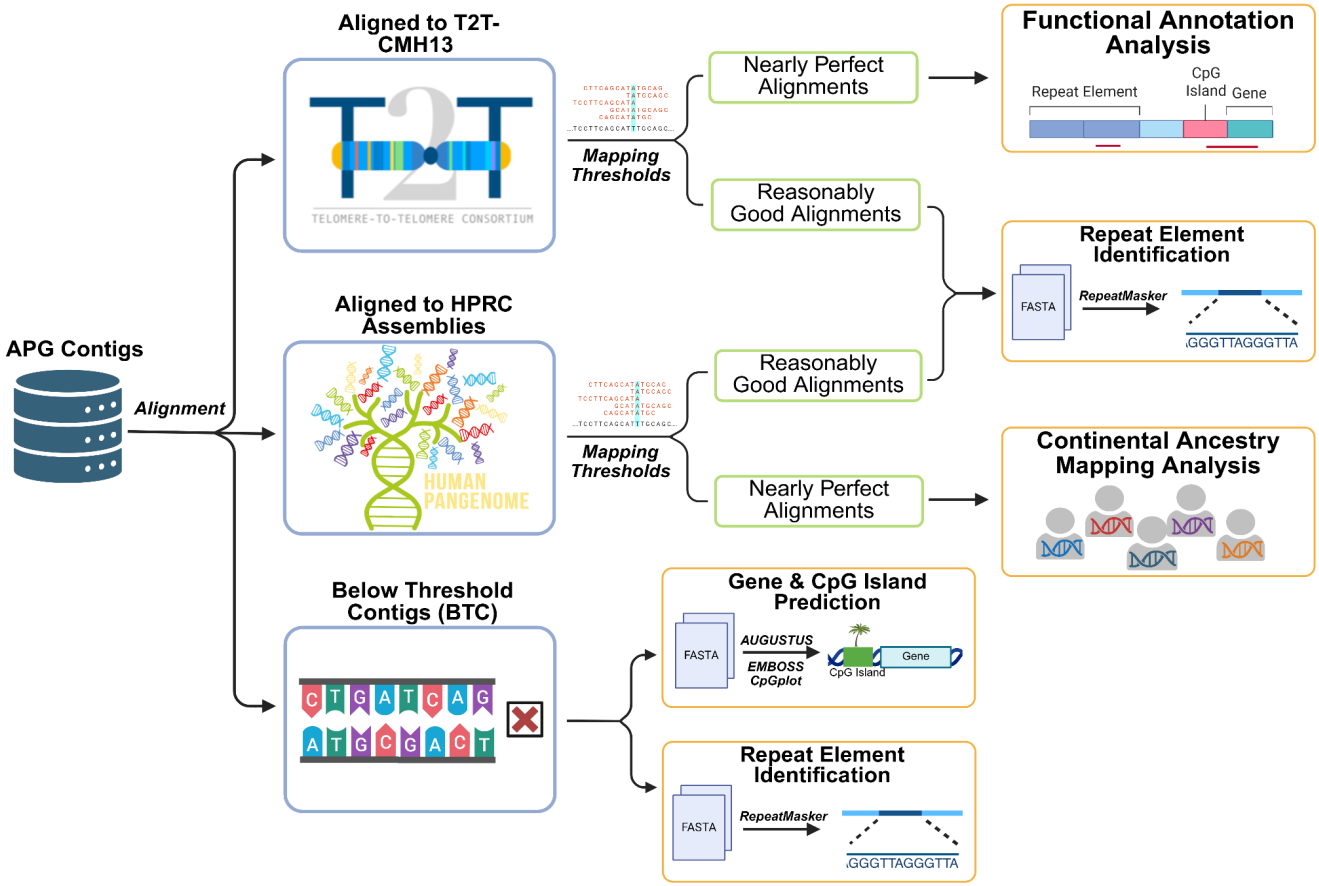


Figure 1. Study Design. APG contigs were aligned to the T2T-CHM13 reference genome and to 47 haploid paternal linear assemblies from the Human Pangenome Reference Consortium (HPRC; Year 1, Release 1). Based on predefined mapping quality thresholds, alignments were classified as “Reasonably Good” or “Nearly Perfect”. Contigs that failed to align, or that produced only weak alignments to both T2T-CHM13 and HPRC references, were designated as “Below Threshold Contigs (BTC)”. Contigs with “Nearly Perfect” alignments to T2T-CHM13 were further examined based on their functional annotations, whereas those aligning “Nearly Perfectly” to HPRC assemblies were used for continental ancestry analysis. For “Below Threshold Contigs (BTC)”, we performed de novo gene prediction and CpG island identification. Repeat element analysis was conducted on both the “Below Threshold Contigs (BTC)” and the “Reasonably Good” alignments to any of the reference genomes.

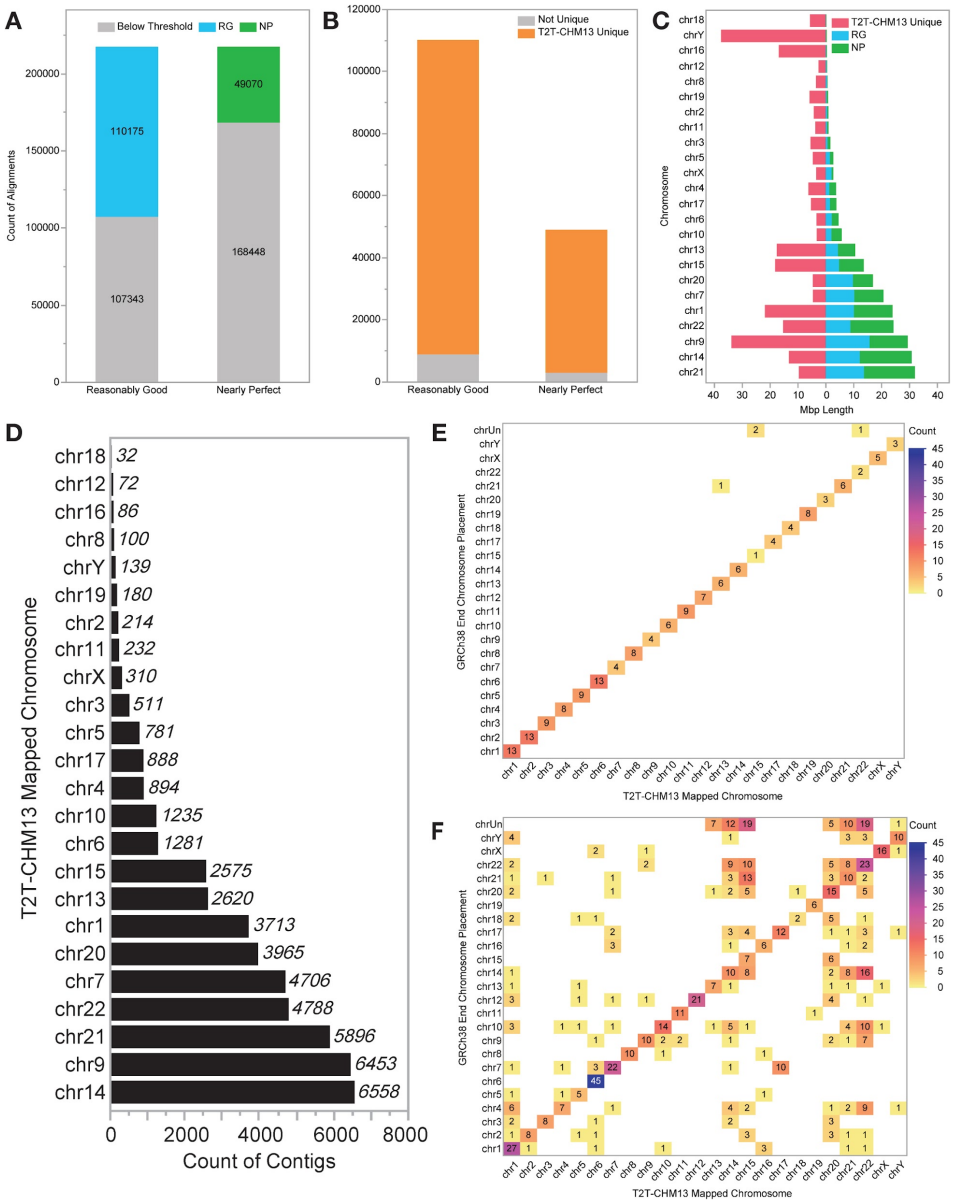


Figure 2. APG Contigs align to T2T-CHM13. (A) Stacked bar charts showing distribution of contigs with Reasonably Good (RG, blue) or Nearly Perfect (NP, green) alignments to T2T-CHM13. Alignments that did not pass RG or NP thresholds are designated as 'below threshold contigs' (BTC) and are shown in gray. (B) Stacked bar chart depicting T2T-CHM13 RG or NP alignments that map to T2T-CHM13 unique regions compared to the GRCh38 reference (orange). Alignments to non-unique regions are shown in gray. (C) Butterfly plot of T2T-CHM13 alignment counts per chromosome (right) and Mb T2T-CHM13 unique content from GRCh38 (left, pink). All alignments shown passed the RG threshold (blue), where those RG alignments that passed the NP threshold are shown in green. (D) Bar chart of counts of unplaced APG contigs mapping to T2T-CHM13 chromosomes. Heatmaps of GRCh38 chromosome placements and T2T-CHM13 chromosome NP mapping results for (E) Two-End placed and (F) One-End placed APG contigs.

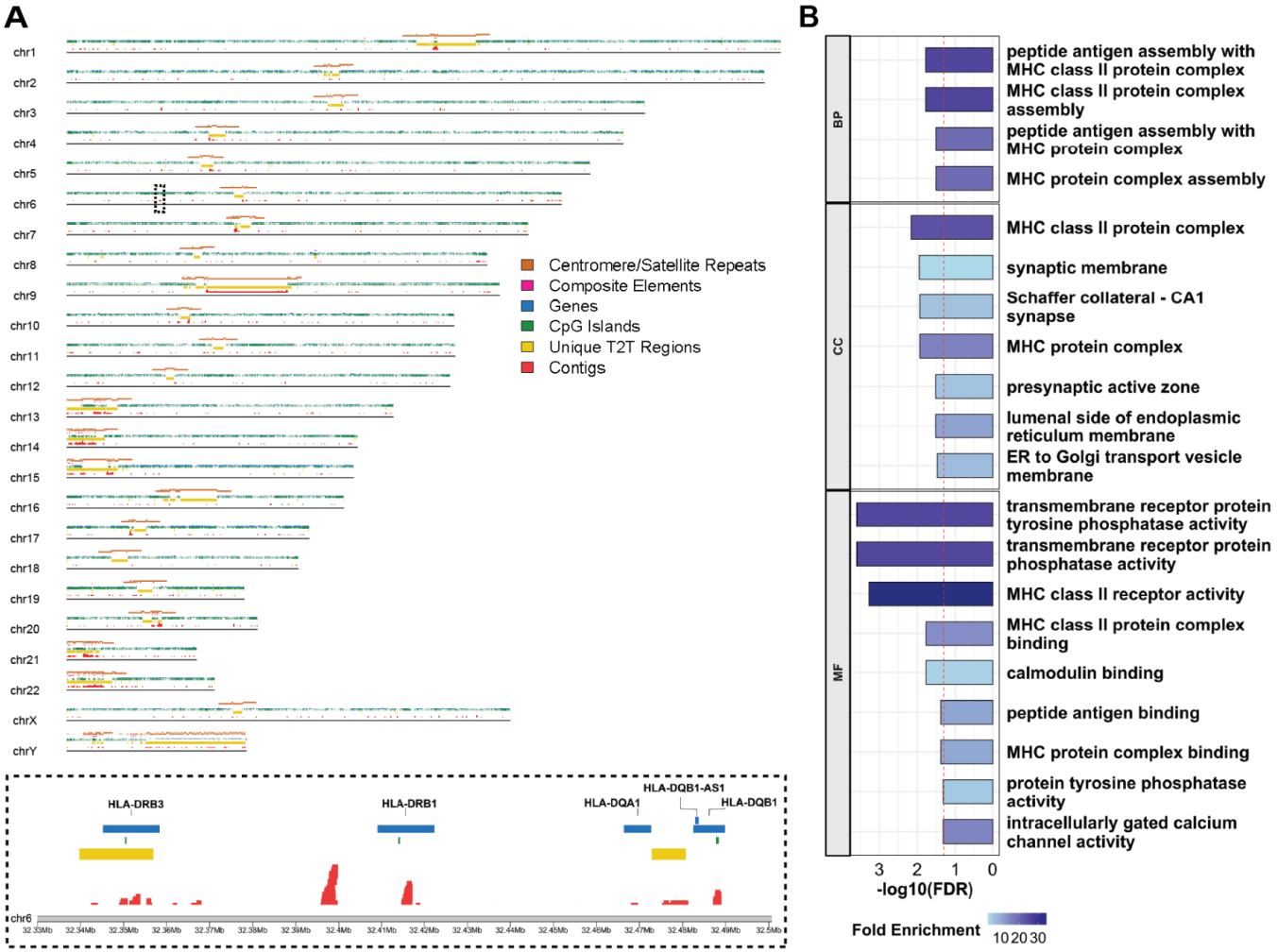


Figure 3. Distribution of Functional Features in T2T-Mapped Contigs. (A) Placement of nearly perfectly aligned contigs across T2T-CMH13 chromosomes and various genomic elements. Contigs are highlighted in red. Enrichment was examined across T2T-CMH13 unique regions in comparison to GRCh38 (yellow), CpG Islands (green), RefSeq annotated genes (blue), composite elements (pink), and centromeres/satellite repeats (brown). A zoomed-in 170 Kb region on chromosome 6 HLA locus with the highest number of mapped contigs across genes is shown at the bottom of the panel. This region is also indicated on the chromosomal plot inside a black dashed-line box. (B) Significantly overrepresented Gene Ontologies (BP: Biological Processes, CC: Cellular Component, MF: Molecular Function) across the 373 T2T-CMH13 genes with at least one APG contig nearly perfectly aligned to them. The X-axis represents the enrichment significance (-log10(FDR)). The color-coded fold enrichment score is also visualized.

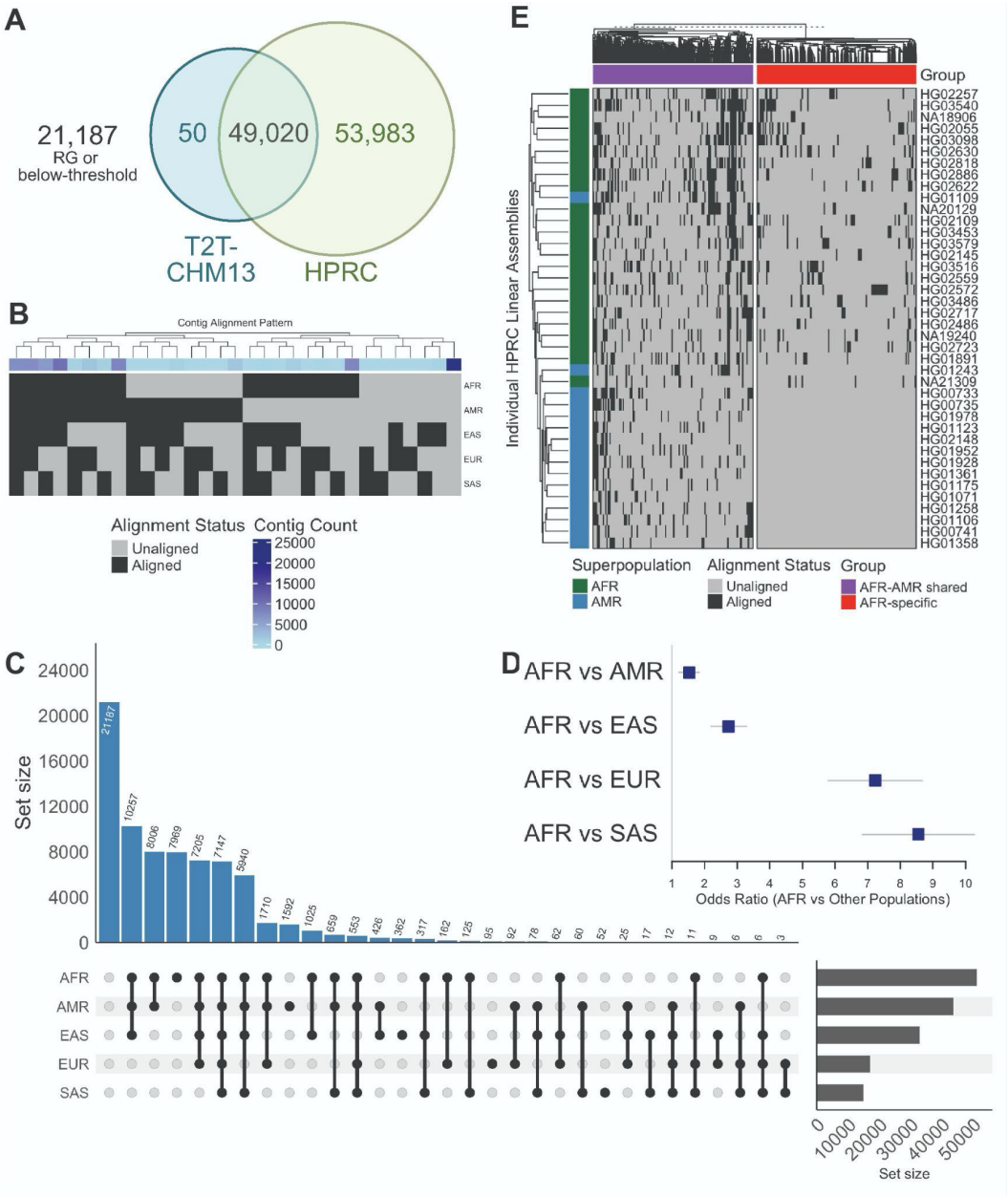


Figure 4. Population and Individual Sharing Summary of APG Contig Alignment across HPRC Assemblies. (A) Venn diagram depicting the overlap of NP APG contig mapping to T2T-CHM13, HPRC linear assemblies. (B) Continental-level binary alignment patterns across HPRC linear assemblies, where a contig is aligned to at least one member of each of the five human superpopulations. (C) Summary of the intersections of contig alignments across human superpopulations. (D) Odds ratios (OR) comparing the likelihood of APG contigs alignment in African (AFR) genomes versus other superpopulations. (E) Heatmap showing alignment across individual linear assemblies from African (AFR) and Admixed American (AMR) HPRC samples.

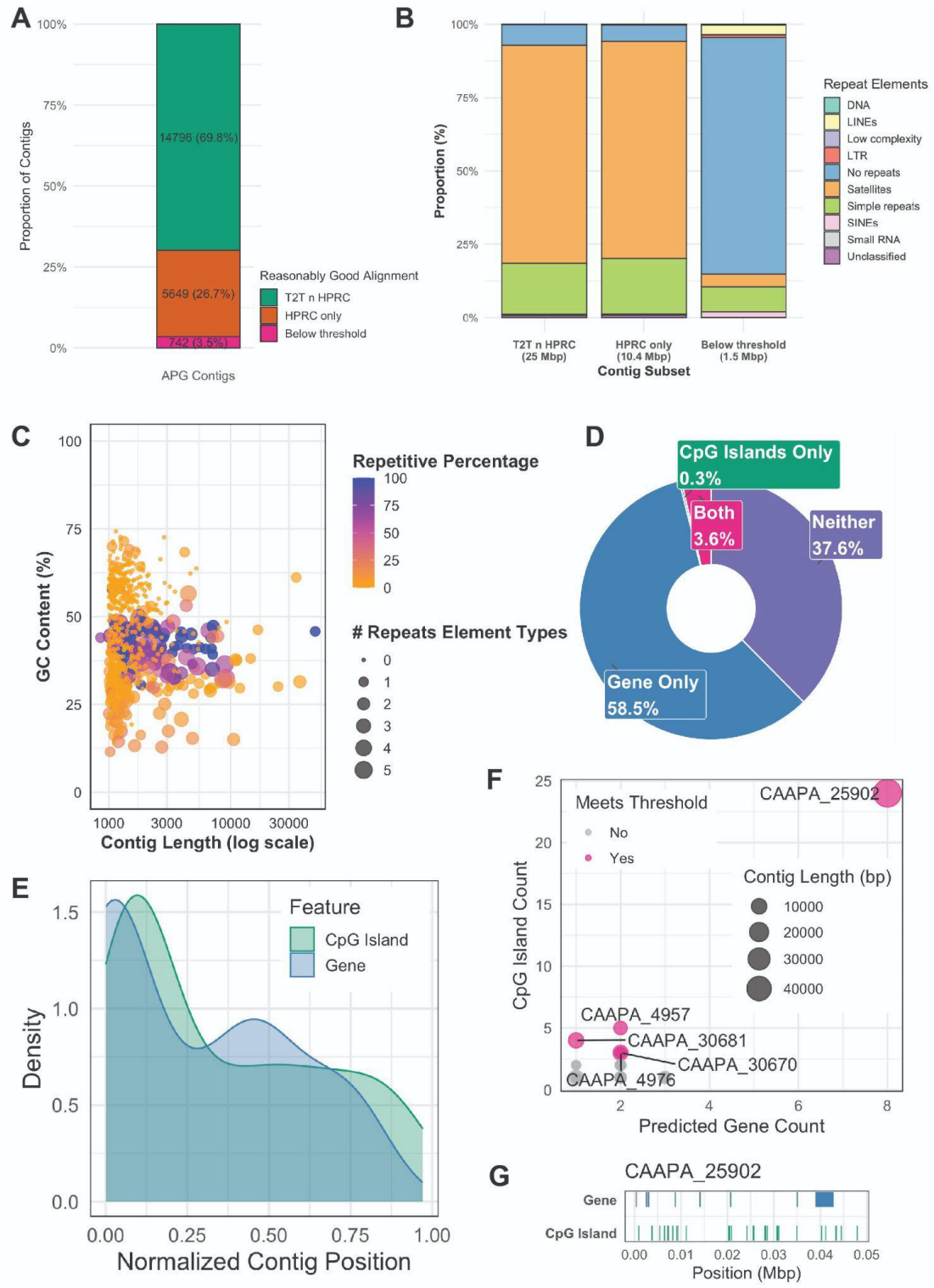


Figure 5. Sequence Characteristics and Functional Landscape of Below-threshold APG Contigs. (A) Mapping categories of APG contigs to T2T and HPRC at Reasonably Good Threshold. (B) Repeat profiles of APG contigs across mapping categories. (C) GC content and length distribution of below-

threshold APG contigs with repeat annotation. Contigs are shown by GC content (%) versus length (log scale), with color indicating repeat proportion and point size representing the number of repeat classes. (D) Proportion of ab-initio predicted gene and CpG islands in below-threshold APG contigs. (E) Density of predicted gene and CpG island start positions normalized to contig length. (F) Scatter plot comparing the number of predicted CpG islands and genes per APG contig. (G) Positional landscape of CpG islands by their genomic context relative to predicted genes.

Tables

Table 1. APG Contigs to T2T-CHM13 Alignment Metrics

Alignment Threshold	Alignment Counts N = 217,518	Contig Counts N = 124,240
Reasonably Good	**N (%)**	**N (%)**
No	107,343 (49.35%)	15,817 (12.73%)
Yes	110,175 (50.65%)	108,423 (87.26%)
Nearly Perfect	**N (%)**	**N (%)**
No	168,448 (77.44%)	75,170 (60.50%)
Yes	49,070 (22.56%)	49,070 (39.50%)

Table 2. APG Contig to T2T-CHM13 Functional Annotation Enrichment

	RG		NP	
	All Mapped Contigs N = 108,423	**Uniquely Mapped Elements N = 1,639**	**All Mapped Contigs N = 49,070**	**Uniquely Mapped Elements N = 1,978**
T2T-CHM13 Element	**N (%)**	**N**	**N (%)**	**N**
Unique T2T-CHM13 Region	100,408 (92.6%)	158	46,347 (94.5%)	137
CpG Islands	2,227 (2.0%)	452	1,772 (3.6%)	431
Genes	3,061 (2.8%)	504	1,297 (2.6%)	373
Composite Elements	2,523 (2.3%)	4	2,153 (4.4%)	3
Centromere/ Satellite Repeats	99,864 (92.1%)	521	46,204 (94.2%)	434
Not Annotated	2 (~0.0%)	-	2 (~0.0%)	-

Table 3. APG Contigs with at least 1 RG or NP Mapping to HPRC linear assemblies

Ancestry Group		**Reasonably Good** *n Contigs (%)*	**Nearly Perfect** *n Contigs (%)*
African	*n* = 24	123,434 (99.35%)	99,944 (80.44%)
European	*n* = 1	111,915 (90.08%)	59,630 (48.00%)
Ad Mixed American	*n* = 16	123,033 (99.03%)	92,577 (74.51%)
East Asian	*n* = 5	121,457 (97.76%)	79,892 (64.30%)
South Asian	*n* = 1	109,491 (88.13%)	54,998 (44.27%)
All HPRC Assemblies	***n* = 47**	**123,498 (99.40%)**	**103,003 (82.91%)**

Table 4. Repeat Content of APG Contigs by Mapping Category

Contig Subset	All		T2T-CHM13 & HPRC RG		HPRC only RG		Below threshold	
# of Contigs	124240		14796		5649		742	
Total Length (bp)	293,418,311		24,924,937		10,354,272		1,532,611	
GC level	40.99%		38.88%		39.00%		43.13%	
bases masked (bp)	274,133,549 (93.43%)		23,187,533 (93.03%)		9,784,387 (94.50%)		296,699 (19.36%)	
	# Elements	**bp (%)**	**# Elements**	**bp (%)**	**# Elements**	**bp (%)**	**# Elements**	**bp (%)**
DNA elements	1144	232436 (0.08%)	76	8322 (0.03%)	29	4405 (0.04%)	22	4978 (0.32%)
LINEs	4026	1671234 (0.57%)	55	9195 (0.04%)	37	14014 (0.14%)	92	48462 (3.16%)
Low complexity	763	239340 (0.08%)	45	15470 (0.06%)	17	5523 (0.05%)	57	3217 (0.21%)
LTR elements	2432	863264 (0.29%)	37	4872 (0.02%)	33	4865 (0.05%)	43	13593 (0.89%)
Satellites	142519	165015904 (56.24%)	14271	18543553 (74.4%)	5535	7677137 (74.14%)	57	65402 (4.27%)
Simple Repeats	99086	100762542 (34.34%)	5620	4312141 (17.3%)	2017	1949807 (18.83%)	279	131287 (8.57%)
SINEs	46438	3918252 (1.34%)	1671	108380 (0.43%)	591	43208 (0.42%)	168	28956 (1.89%)
Small RNA	136	80123 (0.03%)	2	960 (0%)	0	0 (0%)	3	209 (0.01%)
Unclassified	12694	1349759 (0.46%)	1038	184640 (0.74%)	449	85428 (0.83%)	6	595 (0.04%)

References

1. Li, H. *et al.* Exome variant discrepancies due to reference-genome differences. *Am. J. Hum. Genet.* **108**, 1239–1250 (2021).
2. Nie, J., Tellier, J., Tarasova, I., Nutt, S. L. & Smyth, G. K. T2T-CHM13 versus hg38: accurate identification of immunoglobulin isotypes from scRNA-seq requires a genome reference matched for ancestry. *NAR Genom. Bioinform.* **7**, lqaf074 (2025).
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. *Nature* **617**, 312–324 (2023).
5. Miga, K. H. & Wang, T. The need for a human pangenome reference sequence. *Annu. Rev. Genomics Hum. Genet.* **22**, 81–102 (2021).
6. Clavell-Revelles, P. *et al.* Long-read transcriptomics of a diverse human cohort reveals widespread ancestry bias in gene annotation. *Genomics* (2025).
7. Fatumo, S. *et al.* A roadmap to increase diversity in genomic studies. *Nat. Med.* **28**, 243–250 (2022).
8. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–164 (2016).
9. Martin, A. R. *et al.* Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).
10. Carlson, C. S. *et al.* Generalization and dilution of association results from European GWAS in populations of non-European ancestry: the PAGE study. *PLoS Biol.* **11**, e1001661 (2013).
11. Ojewunmi, O. O. & Fatumo, S. Driving Global Health equity and precision medicine through African genomic data. *Hum. Mol. Genet.* ddaf025 (2025).
12. Bentley, A. R., Callier, S. L. & Rotimi, C. N. Evaluating the promise of inclusion of African ancestry populations in genomics. *NPJ Genom. Med.* **5**, 5 (2020).
13. Hadar, N. *et al.* STRavinsky STR database and PGTailor PGT tool demonstrate superiority of CHM13-T2T over hg38 and hg19 for STR-based applications. *Eur J Hum Genet* **31**, 738–743

(2023).

14. Ma, W. & Chaisson, M. Genotyping sequence-resolved copy number variation using pangenomes reveals paralog-specific global diversity and expression divergence of duplicated genes. *bioRxiv* (2025) doi:10.1101/2024.08.11.607269.
15. Aganezov, S. *et al.* A complete reference genome improves analysis of human genetic variation. *Science* **376**, eabl3533 (2022).
16. Sherman, R. M. *et al.* Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat Genet* **51**, 30–35 (2019).
17. McKusick-Nathans Institute of Genetic Medicine, John Hopkins University (Baltimore, MD). Online Mendelian Inheritance in Man, OMIM. https://omim.org/.
18. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev* **25**, 1010–1022 (2011).
19. Saxonov, S., Berg, P. & Brutlag, D. L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A* **103**, 1412–1417 (2006).
20. Mathias, R. A. *et al.* A continuum of admixture in the Western Hemisphere revealed by the African Diaspora genome. *Nat. Commun.* **7**, 12522 (2016).
21. Sibomana, O. Genetic diversity landscape in African population: A review of implications for personalized and precision medicine. *Pharmgenomics. Pers. Med.* **17**, 487–496 (2024).
22. Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).
23. Miga, K. H. *et al.* Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020).
24. Kazazian, H. H., Jr. Mobile elements: drivers of genome evolution. *Science* **303**, 1626–1632 (2004).
25. Almarri, M. A. *et al.* Population structure, stratification, and introgression of human structural

variation. *Cell* **182**, 189–199.e15 (2020).

26. Takai, D. & Jones, P. A. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci U S A* **99**, 3740–3745 (2002).
27. Illingworth, R. S. *et al.* Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS Genet* **6**, e1001134 (2010).
28. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
29. Hsieh, P. *et al.* Adaptive archaic introgression of copy number variants and the discovery of previously unknown human genes. *Science* **366**, eaax2083 (2019).
30. Wang, T. *et al.* The Human Pangenome Project: a global resource to map genomic diversity. *Nature* **604**, 437–446 (2022).
31. Martini, R., Gebregzabher, E., Newman, L. & Davis, M. B. Enhancing the trajectories of cancer health disparities research: Improving clinical applications of diversity, equity, inclusion, and accessibility. *Cancer Discov.* **12**, 1428–1434 (2022).
32. Davis, M. B. & Martini, R. Precision oncology and genetic ancestry: The science behind population-based cancer disparities. *Cancer Cell* **43**, 619–622 (2025).
33. Garrison, E. *et al.* Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat Biotechnol* **36**, 875–879 (2018).
34. Paten, B., Novak, A. M., Eizenga, J. M. & Garrison, E. Genome graphs and the evolution of genome inference. *Genome Res* **27**, 665–676 (2017).
35. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).
36. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
37. Perez, G. *et al.* The UCSC Genome Browser database: 2025 update. *Nucleic Acids Res* **53**, D1243–D1249 (2025).

38. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

39. Xu, S. *et al.* Using clusterProfiler to characterize multiomics data. *Nat. Protoc.* **19**, 3292–3320 (2024).

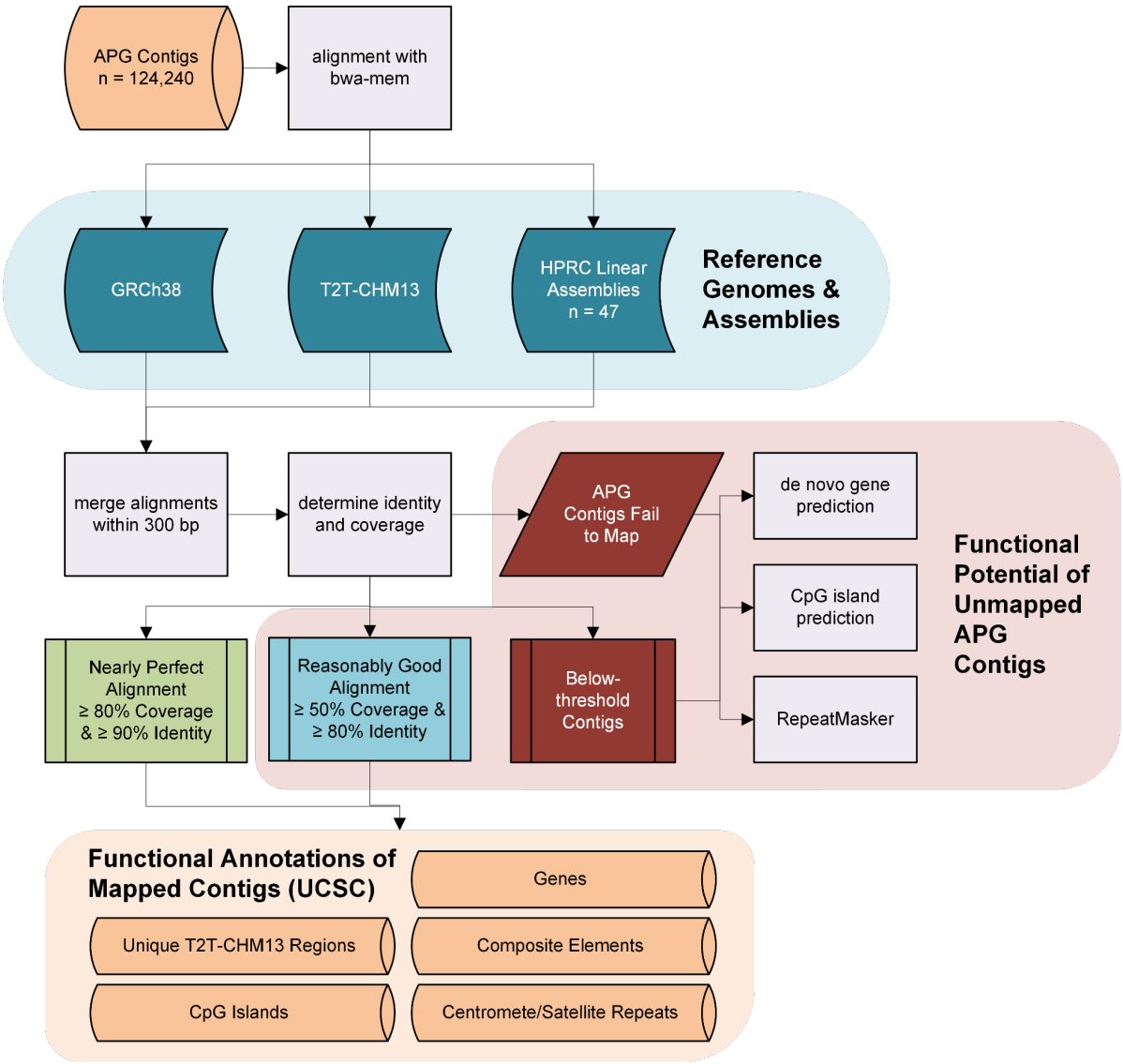
40. Smit, A. F. A., Hubley, R. & Green, P. *RepeatMasker Open-4.0*. (2013-2015).

41. Hubley, R. *et al.* The Dfam database of repetitive DNA families. *Nucleic Acids Res.* **44**, D81–9 (2016).

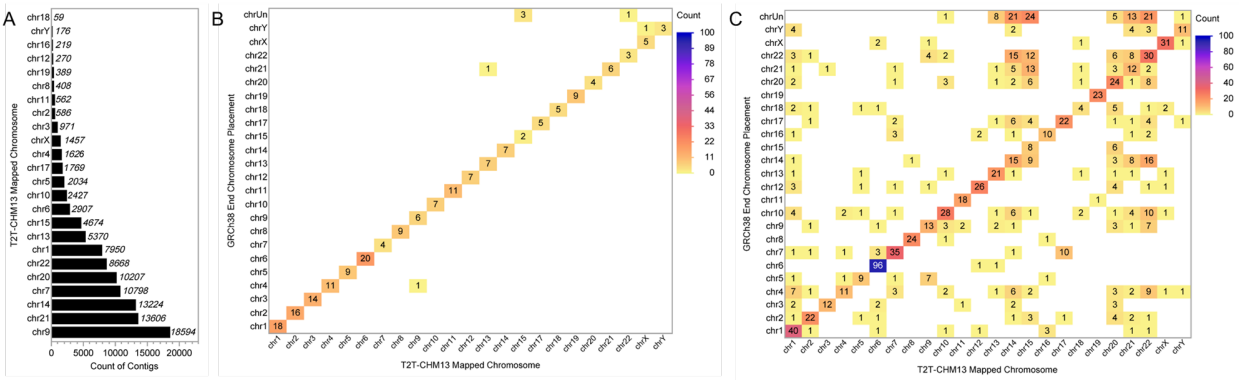
42. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).

43. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res* **32**, W309–12 (2004).

Supplemental Figures



Supplemental Figure 1. African Pan Genome Contig Alignment Computational Workflow. Overview of the computational steps and resources utilized in the characterization of the 124,240 APG Contigs.



Supplemental Figure 2. GRCh38 chromosome placements and T2T-CHM13 chromosome mapping results for Reasonably Good mapping APG Contigs. (A) Bar chart of RG mapping to T2T-CHM13 chromosomes. Heatmaps of GRCh38 chromosome placements and T2T-CHM13 chromosome mapping results for (B) Two-End placed and (C) One-End placed APG contigs for Reasonably Good alignments.

Supplemental Tables

Supplemental Tables Excel File contains the following:

- **Supplemental Table 1.** African Pan Genome Contig Status
- **Supplemental Table 2.** APG Contig alignments to T2T with GRCh38 Two and One End Placements
- **Supplemental Table 3.** APG Contigs to T2T-CHM13 Alignment Metrics for Two End, One End, Unplaced
- **Supplemental Table 4.** T2T-CHM13 NP APG Contigs overlap with GWAS hits
- **Supplemental Table 5.** T2T-CHM13 NP APG Contigs overlap with OMIM genes
- **Supplemental Table 6.** Human PanGenome Reference Consortium Ancestry Annotations
- **Supplemental Table 7.** APG Contigs to HPRC Alignment Metrics for Two End, One End and Unplaced Contigs
- **Supplemental Table 8.** APG Contigs to HPRC Supergroup Level Metrics
- **Supplemental Table 9.** APG Contigs to HPRC Individual Assemblies
- **Supplemental Table 10.** Annotation of the 21,187 RG and Below-threshold Mapping Status
- **Supplemental Table 11.** Gene and CpG Island Prediction of 742 Below-threshold APG contigs