A reference-quality, fully annotated genome from a Puerto Rican individual

Aleksey V. Zimin[1, 2]*, Alaina Shumate[1, 2]*, Ida Shinder[2,3], Jakob Heinz[1,2], Daniela Puiu[1, 2], Mihaela Pertea[1,2], and Steven L. Salzberg[1,2,4,5]

[1]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD
[2]Center for Computational Biology, Johns Hopkins University, Baltimore, MD
[3]Cross Disciplinary Graduate Program in Biomedical Sciences, Johns Hopkins University School of Medicine, Baltimore, MD
[4]Department of Computer Science, Johns Hopkins University, Baltimore, MD
[5]Department of Biostatistics, Johns Hopkins University, Baltimore, MD
*These authors contributed equally to this work.

Abstract
Until 2019, the human genome was available in only one fully-annotated version, which was the result of 18 years of continuous improvement and revision. Despite dramatic improvements in sequencing technology, no other individual human genome was available as an annotated reference until 2019, when the genome of an Ashkenazi individual was released. In this study, we describe the assembly and annotation of a second individual genome, from a Puerto Rican individual whose DNA was collected as part of the Human Pangenome project. The new genome, called PR1, is the first true reference genome created from an individual of African descent. Due to recent improvements in both sequencing and assembly technology, PR1 is more complete and more contiguous than either the human reference genome (GRCh38) or the Ashkenazi genome. Annotation revealed 42,217 genes (of which 20,168 are protein-coding), including 107 additional gene copies that are present in PR1 and missing from GRCh38. 180 genes have fewer copies in PR1 than in GRCh38, 13 map only partially, and 3 genes (1 protein-coding) from GRCh38 are entirely missing from PR1.

Introduction

Since 2001, most work in human genomics and genetics has relied upon a single genome, which was originally sequenced from a small number of anonymous individuals of varying, unknown backgrounds [1]. The human reference genome, currently in release GRCh38, is a mosaic of those individuals, originally created by stitching together individual assemblies of thousands of short bacterial artificial chromosomes (BACs), which averaged ~150 kilobases (Kb) in length.

The human reference genome is the basis for an enormous range of scientific and medical studies, and for large databases of genome variation such as gnomAD [2]and dbSNP [3] that aim to catalog the genetic variation inherent in diverse populations. A reference genome that is more representative of the genetic diversity of African genomes may help drive discovery of clinically relevant variation in that population. Considering that the African pan-genome contains 10% more DNA than the current human reference genome [4] and that the clinical relevance of this genetic variation remains largely unknown [5], analyses of admixed or non-European populations that rely on GRCh38 may overlook medically relevant genetic variation and skew the discovery of diverse haplotypes toward individuals who more closely match the reference genome.

Although thousands of human genomes have been sequenced in the past decade, until recently none of them were available to function as reference genomes, for at least two reasons. First, none were as contiguous as GRCh38, which has undergone many years of careful refinement to fill gaps and fix errors. This is now changing as a result of extremely long reads that can be generated by the latest sequencing technology, allowing us to generate assemblies that are more complete than GRCh38. Second, an essential requirement for a reference genome is that it is annotated with all known human genes, including both protein-coding and noncoding genes. Until the publication of the Ash1 genome in 2019 [6], no human genome other than GRCh38 was both highly contiguous and fully annotated.

A distinctive feature of Ash1 is that it is based on a single individual of Ashkenazi descent, unlike GRCh38 which was created from a mixture of multiple individuals with different genetic backgrounds. For some genetic studies, particularly those focused on a specific population, an ideal genetic reference would be a genome assembled from a healthy individual from a similar genetic background. One reason for this, as discussed in [6], is that comparisons between new study subjects and an appropriately matched reference will turn up many fewer variants, allowing investigators to focus on those variants most likely to be relevant. The Ash1 genome demonstrated that it was technically possible to create such reference genomes, and that more are needed.

The current study describes the assembly and annotation of PR1, the first genome of a Puerto Rican individual, and the first reference-quality genome of any human of African descent. As we show in this study, the contiguity and completeness of PR1 are considerably better than GRCh38, with more total DNA, fewer gaps, and more genes. For these and other reasons, PR1 should be ideal to use as a reference genome in studies of individuals from a similar genetic background.

Results and discussion

Data and genetic background

The data for the HG01243 genome was generated and made publicly available by the Human Pangenome project (https://github.com/human-pangenomics/hpgp-data), which in turn used data from the 1000 Genomes Project [7] and the Genome In A Bottle (GIAB) [7] sample collections. Several earlier assemblies of HG01243 appeared as part of an assembly methods comparison study [8]; we refer to our assembly as PR1. HG01243 was chosen for the current study because the individual is distinct from the only previous individual human reference-quality assembly, that of an Ashkenazi individual [6], and who represents a population, Puerto Ricans, that is not well-represented by the GRCh38 reference genome. Another benefit of choosing HG01243 is that sequence data is available from both parents, HG01241 and HG01242, which offers future opportunities to study recombination and genetic differences between parents and offspring.

HG01243 is a male individual from Puerto Rico of mixed, primarily African, ancestry. His maternal haplogroup based on the mitochondrial genome is L1b2a, which is most common in West Africa [9]. The paternal haplogroup based on the Y chromosome is R1b1a1b (more specifically, R1b1a1b1a1a), which is the most common haplogroup in western Europe [10]. We compared HG01243's Y chromosome to males of diverse backgrounds in the 1000 Genomes (1KG) and Human Genome Diversity Project (HGDP) callset available from gnomAD v3.1 [2] and determined that his Y chromosome has the highest similarity to males of Basque descent, which is consistent with the R1b1a1b haplogroup.

To infer PR1's ancestry from the whole-genome data, we followed an approach developed by Stevens et al. [11] that uses identity-by-descent (IBD) and identity-by-state (IBS) to determine the extent of relatedness between a pair of individuals. Briefly, in a pairwise comparison of two individuals at a single locus or allele, there are three possible IBS outcomes: IBS0, IBS1, and IBS2. Loci where the individuals do not have any alleles in common (for example, a pair of individuals with genotypes AA and BB) are labeled IBS0. Loci where the individuals share a single allele (e.g., AA and AB) are IBS1, and loci where individuals share both alleles (e.g., AA and AA) are IBS2. A subtype of IBS2 called IBS2* [11] occurs when two individuals are heterozygous and share both alleles (i.e., AB and AB). Pairwise relationships for all individuals in a given cohort can be plotted by integrating IBS values over a set of variants. These relationships for PR1 are shown in Figure 1, where the x-axis ratio uses the IBS2* ratio, IBS2*/(IBS2* + IBS0) [11]. For unrelated individuals from the same population, where all loci are assumed to be biallelic and follow Hardy-Weinberg

equilibrium, the IBS2* ratio is expected to average ⅔, whereas the IBS1 heterozygosity ratio (the y-axis of Figure 1) is expected to average 1 [12]. As shown in Figure 1, the IBS2* ratio for pairwise comparisons of PR1 to African-descent individuals averages just below ⅔, indicating that PR1 is of African ancestry.

Figure 1. IBS relationship between PR1 and 2,371 individuals of African/African-American (afr), Non-Finish European (nfe), and East Asian (eas) descent in the 1KG and HGDP callset. Population labels were obtained from gnomAD v3.1 [2]. The x-axis ratio has two components, IBS0 and IBS2*. IBS0 is an aggregate count of times that a given individual does not share any alleles with PR1 (i.e., PR1 is AA and the target individual is BB). IBS2* is the aggregate count of times that the target individual PR1 is heterozygous and shares two alleles with PR1 (genotypes AB and AB). The y-axis ratio ($IBS1_{het2}/IBS1_{het1}$), is a proxy for heterozygosity and level of genetic variation. $IBS1_{het1}$ is the aggregate count of times that PR1 is heterozygous and the target individual is homozygous and they share one allele (i.e., AB and AA, respectively). $IBS1_{het2}$ is the aggregate count of times that PR1 is homozygous and the target individual is heterozygous and they share one allele (i.e., AA and AB, respectively). As expected for parent-child relationships, the IBS2* ratio for PR1's parents, who appear as two dots on the far right of the plot, is 1.

Assembly

To create the PR1 assembly, we used both long and short reads from multiple technologies, including PacBio continuous long reads (CLR), PacBio high-fidelity reads (HiFi), Oxford

Table 1. Data used for assembly of the PR1 genome. N50 read length is defined as the length such that 50% of the total sequence is contained in reads of length N50 or longer. Note that the effective length of the Illumina reads was shorter after removing barcodes and adapter sequences.

Sequence type	Number of reads	N50 read length (bp)	Genome coverage
Illumina paired-end	1,307,137,030	150	59x
PacBio CLR	23,672,554	28,837	113x
PacBio HIFI	5,485,642	19,528	34x
ONT Ultralong	16,293,849	44.052	58x

Nanopore (ONT) ultra-long reads, and 150bp Illumina paired-end reads (**Table 1**). Our

goal was to produce a high-quality, traditional haplotype-merged assembly with a single sequence representing both chromosomes. For long-range scaffolding, we used as a reference the assembly of CHM13 released by the Telomere-to-Telomere (T2T) consortium, a very high-quality human assembly that is far more complete and contiguous than the GRCh38 reference genome, with only 5 gaps in version 1.0 [13] [14].

We initially produced three different assemblies using three different assemblers, and evaluated them for contiguity and consensus accuracy. To create these assemblies, we used Hifiasm [15] version 0.13-r308 with PacBio HiFi data, Flye [16] version 2.5 with Oxford Nanopore data, and MaSuRCA [17] version 4.0.1 with Illumina, Oxford Nanopore and PacBio CLR data. (Note that for the MaSuRCA assembly, we only used PacBio CLR reads that were longer than 20Kb. A MaSuRCA assembly that included the HiFi reads contained slightly more errors and was not used.) The exact commands used to create each assembly are listed in the Supplementary text. **Table 2** contains summary statistics for these three assemblies. We measured consensus accuracy by aligning the Illumina data to the assemblies and calling variants using Freebayes [18]. A variant was considered an error if no Illumina reads supported the assembly, and if at least 2 Illumina reads supported an alternative base. The MaSuRCA assembly had the best consensus accuracy, while the Flye assembly had the best contiguity, with an N50 contig size of 27.4 Mbp.

Table 2. Statistics for the preliminary PR1 assemblies. N50 values were calculated using an estimated genome size of 3.2Gbp; thus an N50 size of 6.7 Mbp means that 1.6 Gbp of the total assembly was contained in contigs at least that large. ONT: Oxford Nanopore Technology.

Assembler	**Data used**	**Assembled sequence**	**Contig N50**	**Number of contigs**
Hifiasm	PacBio HiFi	3,096,276,394	6,760,611	1,473
MaSuRCA	Illumina+ONT +PacBio CLR >20Kb	2,848,911,829	7,622,810	1,496
Flye	ONT	2,906,307,381	27,370,234	1,693

From these initial assemblies, we chose the Flye assembly for subsequent refinements because it had the best contiguity. To improve base-level accuracy, we merged the MaSuRCA assembly with the Flye assembly to use the MaSuRCA consensus where possible (see Methods). The resulting MaSuRCA+Flye merged assembly contained 2,892,693,836 bp in 1,185 contigs. Note that this assembly merging process has advantages over simply "polishing" one assembly using Illumina data, because the large contigs in the assemblies can be mapped unambiguously in many places where short Illumina reads would not map reliably enough for accurate polishing.

We then used the MaSuRCA chromosome scaffolder [6] to order and orient the merged contigs onto the human reference using chromosomes 1-22 and X from the CHM13 1.0

reference genome (https://github.com/nanopore-wgs-consortium/CHM13) produced by the T2T consortium [13] [19], and using the Y chromosome from GRCh38.p12. This is similar to the procedure used to build the Ashkenazi reference genome, with the major difference being that Ash1 used GRCh38 as the primary backbone for scaffolding. (See Methods for more details of the scaffolding process.) The final assembly has each chromosome in a single scaffold, and contains only 67 unplaced contigs containing 1.17 Mbp of sequence.

We closed gaps in the chromosome scaffolds by aligning the MaSuRCA contigs to the scaffolds with nucmer, retaining the unique 1-to-1 alignments from delta-filter [20], and closing each gap where the MaSuRCA contig was anchored by unique alignments of 5 Kbp or longer on both sides of the gap. This step closed 272 gaps. We filled in the remaining gaps in the chromosome scaffolds using the CHM13 genome sequence. The final assembly contains these CHM13 sequences in lowercase to indicate that they did not derive directly from HG01243. Essentially all of the gaps that were filled in this way represented hard-to-assemble repetitive sequences. The biggest gap filled using the CHM13 assembly was the centromere of chromosome 9, where we used ~35 Mbp of CHM13 sequence. We included a total of 218 Mbp of sequence from CHM13 and GRCh38 (for chromosome Y only) in the final 3.1 Gbp assembly (**Table 3**).

Table 3. Chromosome sizes and the amount of non-HG01243 sequence per chromosome in the final PR1 assembly.

Chromosome	**Total sequence**	**Non-HG01243 sequence**	**Number of contigs**
1	248,243,692	23,873,221	1
2	242,441,519	6,317,197	1
3	200,623,860	5,696,644	1
4	193,343,089	7,406,780	1
5	181,923,926	5,513,656	1
6	171,764,581	3,762,549	1
7	160,350,280	6,017,454	1
8	146,002,196	4,418,110	1
9	150,081,504	36,131,972	1
10	134,238,072	3,896,169	1
11	134,644,853	4,129,054	1
12	133,390,335	2,461,456	1
13	110,077,781	11,656,137	2
14	100,344,869	10,519,739	2
15	97,474,498	15,930,758	2
16	96,221,913	18,324,976	1

17	83,931,792	4,573,052	1
18	80,716,684	5,494,999	1
19	61,708,235	4,649,059	1
20	65,906,100	3,980,784	1
21	42,594,448	8,609,782	3
22	50,188,155	12,199,724	2
X	154,359,490	4,964,647	1
Y	24,141,209	7,270,630	79
M	16,568	0	1
unplaced	1,165,067	0	67
Total	3,065,894,716	217,943,071	176

After a final polishing step, the assembly had 3,064,729,649 bp of sequence in 109 contigs on the 24 chromosomes and the mitochondrial genome, plus 1,165,067 bp in 67 unplaced contigs, for a total of 3,065,894,716 bp. In the final PR1 assembly, 18 chromosomes are complete, assembled end-to-end with no gaps. Four chromosomes have a single gap, and chr21 has 2 gaps. The Y chromosome is far more fragmented, in 79 contigs, because we had to rely on the less-contiguous GRCh38 genome as a backbone for that sequence. All chromosome sizes and the amount of non-PR1 sequence are listed in **Table 3**.

Chromosomes 9 and 22 stand out in terms of their proportion of non-PR1 sequence, which is about 20% of the chromosome in each case. This is due to very large repeat regions in those chromosomes, primarily centromeres, that are correctly resolved in the CHM13 assembly, but which could not be assembled *de novo* from the data for HG01243. **Supplementary Figure S1** shows the mapping of assembled contigs to CHM13 chromosome 9, illustrating the numerous fragmented contigs from the PR1 assembly that align to the chr9 centromere region.

Novel sequence in PR1

In addition to the sequences placed on chromosomes, the PR1 assembly has 67 small contigs containing 1,165,067 bp that could not be placed and that are non-repetitive, representing sequences that are unique to PR1. 42 of these contigs totalling 297 Kbp do not align elsewhere in the genome (defined as not having any alignment that covers more than 5% of the contig), while the remaining 20 contigs align partially, ranging from 5–75% of their length.

To identify additional novel sequence in PR1, we aligned all chromosomes in the assembly to the corresponding CHM13 chromosomes. After removing simple sequence repeats and other recognizably repetitive sequences, we were left with 10.6 Mbp in 1,103 insertions in PR1, with lengths ranging from 200–395,532 bp (of which 20 insertions were longer than

100 Kb). Because ChrY was derived from the relatively incomplete GRCh38 genome, it was over-represented in the set of novel sequences, accounting for 1,814,738 bp of sequence unique to PR1. When we add in the unplaced contigs, the PR1 genome contains ~11.1 Mb of novel sequence, likely representing regions that are distinctive in the Puerto Rican population.

Comparison to other reference genomes

PR1 is the third fully-annotated reference genome to be created, and we compared it to the longstanding GRCh38 reference as well as to the Ashkenazi reference genome, Ash1. Because the Ash1 assembly used GRCh38 as a template, both it and GRCh38 are less complete than PR1, which used the more-complete CHM13 assembly as a guide. For our comparisons here, we only considered regions that are present in all three assemblies, and we ignored any sequence in PR1 that was filled in using CHM13, and in Ash1 that was filled in using GRCh38.

Table 4. PR1 and Ash1 variants compared to GRCh38. Variants include both substitutions and indels (up to 200bp) in Ash1 or PR1 compared to GRCh38. We used the original Illumina reads to determine whether variants in PR1 and Ash1 were homozygous or heterozygous.

Variant type	Both variants disagree with GRCh38	One variant agrees with GRCh38
PR1 homozygous	3,663,158	n/a
PR1 heterozygous	22,924	4,107,663
Ash1 homozygous	3,249,658	n/a
Ash1 heterozygous	15,564	2,719,200

Tables 4 and 5 show a summary of these findings. In Table 4, which compares PR1 and Ash1 to GRCh38, we see that PR1 has a total of 7,793,745 variants as compared to GRCh38, out of which 4,107,663 are heterozygous variants where one of the haplotypes agrees with GRCh38. In contrast, the Ashkenazi genome has fewer differences, 5,984,422 in total, of which 2,719,200 are heterozygous sites where one variant agrees with GRCh38.

Table 5. Comparison between Ash1 and PR1 at sites where both differ from GRCh38 and where they differ from each other. Most of the locations correspond to variants that are homozygous in both genomes. For cases where both genomes are homozygous or both are heterozygous, numbers in parentheses indicate the number of positions where PR1 and Ash1 match each other.

	Ash1 homozygous	Ash1 heterozygous
PR1 homozygous	1,242,413 (1,193,366)	388,455
PR1 heterozygous	559,115	606,985 (593,803)

In Table 5, we compare PR1 and Ash1 directly at 2,124,816 locations where they both differ from GRCh38. In 1,242,413 of these locations, both PR1 and Ash1 are homozygous, and at 1,193,366 (96%) of these sites, PR1 and Ash1 agree with each other, suggesting that GRCh38 has a less-common variant at those sites. Similarly, the two genomes agree on both variants at 593,803 of the 606,985 sites (98%) where both PR1 and Ash1 are heterozygous and one of the variants disagrees with GRCh38. Data from the heterozygous sites also suggest that PR1 is more heterozygous than Ash1, because it has a greater number of heterozygous sites than Ash1 when they both have a variant compared to GRCh38 at the same location.

Gene Annotation

We used Liftoff [21] to map all of the genes from GRCh38 onto PR1, including protein-coding and non-coding RNA genes. We used two different human annotation databases as sources for the mapping, CHESS [22] and GENCODE [23], creating two different annotation sets. CHESS includes all protein-coding genes from both GENCODE and RefSeq [24], but the noncoding genes are substantially different, as are the precise splice variants for many protein-coding genes. Both databases use GRCh38 as a reference; i.e., all genes are defined by their coordinates on GRCh38.

Liftoff successfully mapped 42,110 out of 42,306 genes (99.54%) from the CHESS 2.2 annotation onto the PR1 assembly, leaving 196 genes unmapped, of which 107 were protein-coding (Table S1). Liftoff considers a gene to be mapped if the alignment coverage and sequence identity of the exons are greater than or equal to 50%; however, the vast majority of genes in PR1 greatly exceed this threshold, with 95.8% of genes mapping with >= 99% coverage and sequence identity.

Out of the 196 genes that failed to map, 91 aligned end-to-end with another copy of the gene present elsewhere in the assembly (i.e., a paralog), suggesting that PR1 simply has fewer copies (Table S1). Another 89 genes had partial copies present in the assembly (see Methods). Of the 16 remaining unmapped genes, all but 3 genes mapped partially did not meet the 50% minimum coverage and sequence identity threshold. Out of these 3 genes completely missing from PR1, only one was protein-coding, and that gene (CHS.2756, a

single-exon gene encoding a 74-amino-acid protein of unknown function) is novel to the CHESS2.2 annotation, suggesting that it might not be a valid protein-coding gene.

Extra gene copies in PR1

Given the greater completeness of PR1 as compared to GRCh38, we expected that it might include additional copies of known genes. We used an optional feature of Liftoff to identify 107 extra gene copies (i.e., genes that have extra copies in PR1 as compared to GRCh38) giving a total gene count of 42,217 (Figure 2 and Table S2). 18 of these are extra copies of protein-coding genes and 89 are copies of non-coding genes. 56 of the extra copies occur in tandem, including 38 tandem copies of CHS.53849, a miscellaneous RNA gene on chromosome 8 (Figure 2). (A gene copy was considered tandem if it occurred within 100 Kbp of another copy.) All isoforms of the additional copies are 100% identical at the mRNA level to the original copy in GRCh38. Some of these extra genes have previously been reported in the CHM13 genome on chromosomes 8 [14], X [13], and other chromosomes [19].

Of the 107 extra gene copies, only three occurred in genomic regions that are present in PR1 and missing from both CHM13 and GRCh38. These three genes are copies of: (1) CHS.19312, a lncRNA gene on chr16 (GENCODE gene ENSG00000253978.2); (2) CHS.59337, a protein-coding gene on chrY that is unique to CHESS; and (3) CHS.59187, a lncRNA gene (GENCODE gene ENSG00000229308.1) on chrY .

We also mapped the GENCODE human annotation onto PR1. Using Gencode v35, which contains 19,954 protein coding and 16,899 lncRNAs, Liftoff mapped all but 117 protein-coding genes and 53 lncRNA genes (0.46% of the total) onto PR1. In addition, we identified 18 additional copies of five distinct protein-coding genes, and 48 extra copies of 34 lncRNA genes in the PR1 assembly. Including all gene types (adding microRNAs, pseudogenes, and other types), we successfully mapped 60,171 out of 60,656 genes from GENCODE v35, plus 396 extra gene copies, bringing the total number of mapped GENCODE genes to 60,567. The GENCODE genes that failed to map, including all gene types, are shown in Supplementary Table S3. All extra gene copies from both CHESS and GENCODE, along with the gene names and chromosomal locations on PR1, are shown in Supplemental Tables S2 and S4.

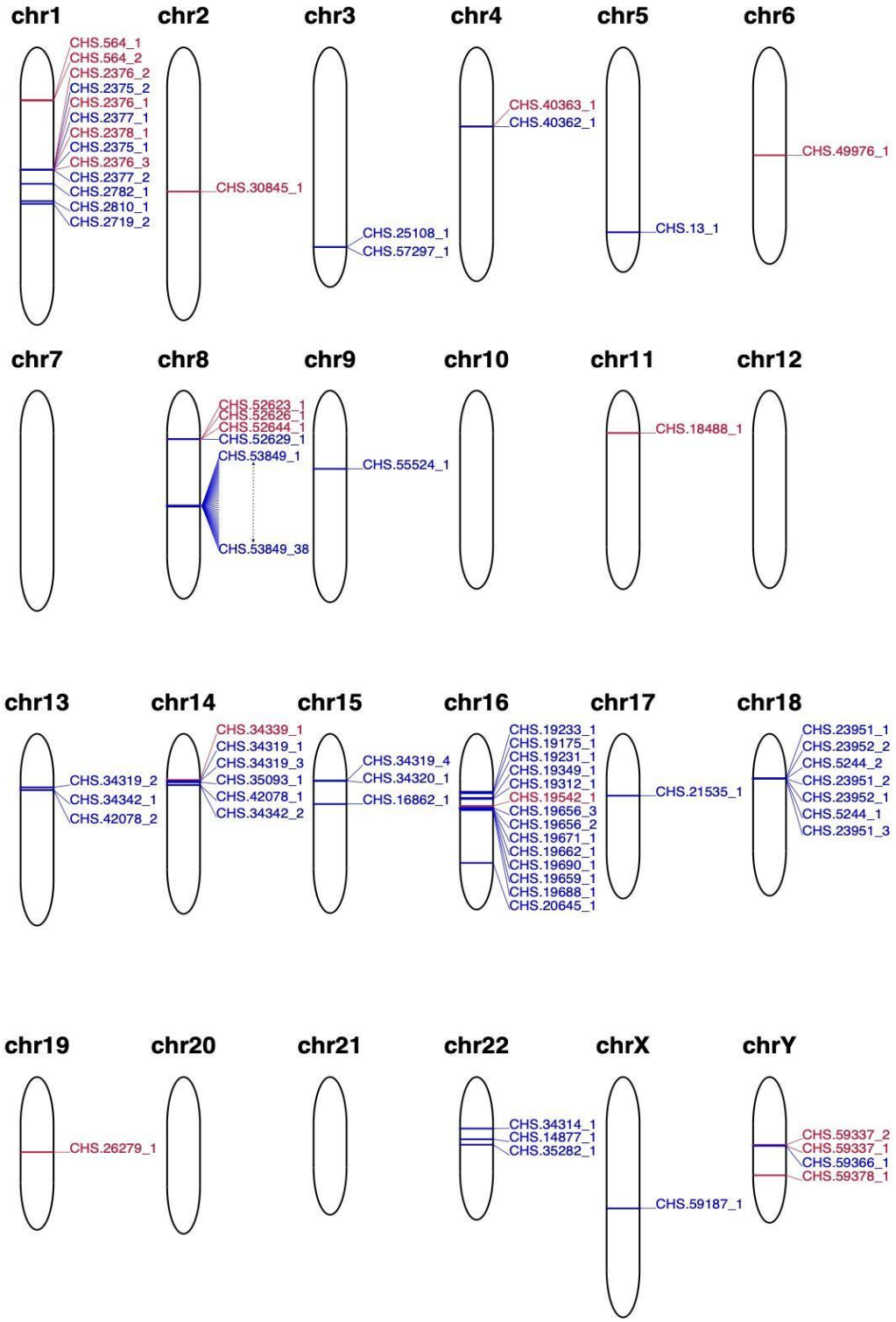

Figure 2. Ideogram of the PR1 assembly showing the locations and IDs of novel gene copies. Protein-coding genes are labeled in red and non-coding genes in blue. A complete list of these genes is provided in Supplementary Table S2.

Methods

Assembly polishing and scaffolding. To improve base-level accuracy, we aligned the MaSuRCA assembly to the Flye assembly using Nucmer [25], and identified the alignments with 1-to-1 best matches using delta-filter. We merged alignments when a MaSuRCA contig matched a Flye contig in several chunks that had the same order and orientation, producing spanning alignments that started at the beginning of the first chunk and ended at the end of the last chunk. Next, we replaced the Flye assembly's consensus sequence with the higher-quality MaSuRCA consensus in places where MaSuRCA and Flye contigs aligned. This process replaced 99.4% of the Flye sequence with the MaSuRCA consensus.

Next, we added all MaSuRCA contigs that aligned to the Flye assembly over less than 25% of their length, on the assumption that these contigs contained sequence missing from the Flye assembly. This process added 90 MaSuRCA contigs containing 2,263,294 bp to the assembly. We aligned the assembly to itself and removed any contigs that mapped over 90% of their length to the interior of larger contigs. These contigs were likely to represent redundant haplotype copies; i.e., places where the two haplotypes diverged sufficiently that the assembly algorithm created two separate contigs.

To scaffold the contigs, the MaSuRCA chromosome scaffolder first aligned the assembly to the CHM13 reference genome and automatically identified breaks in the alignments where the mapping indicated a possible misassembly. Each such location was then checked by mapping the underlying reads to the assembly and examining read coverage. If read coverage dropped below 10% of the average coverage C within 50Kb of the breakpoint, the assembled contig was assumed to be erroneous and was broken at the lowest coverage point. If read coverage was higher than $C/ln(2)$, indicating likely presence of a collapsed repeat (according to a 2-copy Poisson test), we looked for two breakpoints flanking the high-coverage region, and split the assembly at both breakpoints.

This process introduced 703 gaps (breakpoints) into the MaSuRCA+Flye merged assembly, resulting in 1888 contigs. We were able to close 134 of these gaps after the contigs were placed on the chromosomes, which confirmed that they were correctly assembled. Then we used the CHM13 1.0 assembly (plus chromosome Y from GRCh38) as a reference to order and orient the contigs, placing each contig where it aligned best. This procedure ordered and oriented 1359 contigs, containing 98.8% of the sequence, onto chromosome-scale scaffolds. 34.2 Mbp of sequence remained in unplaced contigs. When we aligned these contigs to the CHM13 assembly, we found that 33.0 Mbp of these sequences were collapsed repeat copies that aligned with 90-95% identity to centromeric regions. Because of the high error rates in the Nanopore data, the assembler likely collapsed these repeats producing spurious "consensus" of multiple slightly divergent repeat copies. (Note that

these sequences contained no genes.) Thus we decided to exclude these sequences from the assembly.

The last step in assembly was polishing the assembly with POLCA [26] using the Illumina reads. The polishing was applied to the entire genome, including the sequences that were filled in from CHM13. Any small variants that were not supported by the HG01243 Illumina data were corrected to properly represent the PR1 genome sequence. All corrected bases are indicated in uppercase in the final assembly, even those within CHM13-derived regions.

Genotyping. To identify the paternal haplogroup of PR1, we used Yleaf (18) applied to the Y chromosome. To identify the maternal haplogroup, we used Haplogrep 2 (19) on the mitochondrial genome. To compare the PR1, GRCh38, and Ash1 assemblies, we aligned all pairs of genomes using nucmer [25] and then computed SNPs and indels using nucmer's delta2vcf utility. We then compared these variants to the original Illumina reads for both PR1 and Ash1 in order to identify which variants were homozygous and heterozygous.

For IBS analysis, we used Hail v0.2.67 [27] to parse the 1KG and HGDP variant files from gnomAD v3.1 (https://cloud.google.com/life-sciences/docs/resources/public-datasets/gnomad). To select high quality variants, we downsampled the 1KG Phase I high-confidence SNPs to randomly keep 1% of the variants, resulting in 281,308 biallelic variants. For each variant, we performed a pairwise comparison of HG01243 alleles to all 3,941 individuals in the 1KG and HGDP call set, with an average of 280,000 non-missing variants per pairwise comparison. Counts of IBS2*, IBS1 and IBS0 for each pairwise comparison were aggregated to calculate the IBS2* ratio (IBS2*/(IBS2* + IBS0) and the IBS1 heterozygosity ratio ($IBS1_{het2}/IBS1_{het1}$). To determine the population closest to PR1's Y chromosome lineage, we filtered the 1KG and HGDP callset for SNPs in the male-specific region of the Y chromosome with a quality value ≥ 20 and calculated the proportion of shared variants to total variants.

Gene annotation. To annotate the PR1 genome, we mapped both CHESS v2.2 and GENCODE v35 annotations using Liftoff version 1.6.1 with the following parameters: -copies -polish -exclude_partial -chroms <chroms.txt>. After the initial mapping, we aligned every unmapped transcript to every mapped transcript using Blastn [28] in order to determine if the unmapped genes were copies of mapped genes (where we define a copy as an end-to-end alignment with a mapped transcript). We also re-ran Liftoff allowing for overlapping genes (-overlap 1.0). By comparing the results to the initial Liftoff output, we were able to identify genes that only mapped when allowed to overlap another gene. These overlapping genes are either complete copies of one another if they map to exactly the same locus, or partial copies if they map to different but overlapping loci. The overlapping genes were identified by intersecting the Liftoff-generated annotation file with itself using Bedtools

intersect [29]. The output file was then filtered to remove self-overlaps, and genes identified by this process were classified as partial copies. We further attempted to identify partial copies at the protein level by using gffread [30] to extract the protein sequences and blastp [28] to align them to mapped proteins, using an e-value threshold of 10^{-6} to filter the results.

Data availability

Sequence data including raw signal files and basecalls for HG01243 were previously released by Shafin et al. (20) and are available as an AWS Open Data set for download from https://github.com/human-pangenomics/hpgp-data. Nanopore sequence data are additionally archived at the European Nucleotide Archive under accession code PRJEB37264. The PR1 genome assembly and annotation are available from NCBI and GenBank under accession JAHLSK000000000, and on GitHub at https://github.com/PuertoRicanGenome.

Acknowledgements

Thanks to Adam Phillippy, Sergey Koren, and Karen Miga for help with sample selection and data access. This work was supported in part by NIH grants R01 HG006677 and R01 MH123567, by NSF grants ISO-1744309 and DBI-1759518, and by USDA National Institute of Food and Agriculture grant 2018-67015-28199.

References

1. International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome. *Nature*, 2001. **409**(6822): p. 860-921.
2. Karczewski, K.J., et al., The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 2020. **581**(7809): p. 434-443.
3. Sherry, S.T., et al., dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, 2001. **29**(1): p. 308-11.
4. Sherman, R.M., et al., Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat Genet*, 2019. **51**(1): p. 30-35.
5. Choudhury, A., et al., High-depth African genomes inform human migration and health. *Nature*, 2020. **586**(7831): p. 741-748.
6. Shumate, A., et al., Assembly and annotation of an Ashkenazi human reference genome. *Genome Biol*, 2020. **21**(1): p. 129.
7. 1000 Genomes Project Consortium, et al., A global reference for human genetic variation. *Nature*, 2015. **526**(7571): p. 68-74.
8. Shafin, K., et al., Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat Biotechnol*, 2020. **38**(9): p. 1044-1053.
9. Cerezo, M., et al., Reconstructing ancient mitochondrial DNA links between Africa and Europe. *Genome Res*, 2012. **22**(5): p. 821-6.

10. Busby, G.B., et al., The peopling of Europe and the cautionary tale of Y chromosome lineage R-M269. *Proc Biol Sci*, 2012. **279**(1730): p. 884-92.
11. Stevens, E.L., et al., Inference of relationships in population data using identity-by-descent and identity-by-state. *PLoS Genet*, 2011. **7**(9): p. e1002287.
12. Lee, W.C., Testing the genetic relation between two individuals using a panel of frequency-unknown single nucleotide polymorphisms. *Ann Hum Genet*, 2003. **67**(Pt 6): p. 618-9.
13. Miga, K.H., et al., Telomere-to-telomere assembly of a complete human X chromosome. *Nature*, 2020. **585**(7823): p. 79-84.
14. Logsdon, G.A., et al., The structure, function and evolution of a complete human chromosome 8. *Nature*, 2021. **593**(7857): p. 101-107.
15. Cheng, H., et al., Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*, 2021. **18**(2): p. 170-175.
16. Kolmogorov, M., et al., Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol*, 2019. **37**(5): p. 540-546.
17. Zimin, A.V., et al., Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res*, 2017. **27**(5): p. 787-792.
18. Garrison, E. and G. Marth, Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*, 2012.
19. Nurk, S., et al., The complete sequence of a human genome. *bioRxiv*, 2021.
20. Kurtz, S., et al., Versatile and open software for comparing large genomes. *Genome Biol*, 2004. **5**(2): p. R12.
21. Shumate, A. and S.L. Salzberg, Liftoff: accurate mapping of gene annotations. *Bioinformatics*, 2020.
22. Pertea, M., et al., CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol*, 2018. **19**(1): p. 208.
23. Frankish, A., et al., GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res*, 2019. **47**(D1): p. D766-D773.
24. O'Leary, N.A., et al., Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*, 2016. **44**(D1): p. D733-45.
25. Marcais, G., et al., MUMmer4: A fast and versatile genome alignment system. *PLoS Comput Biol*, 2018. **14**(1): p. e1005944.
26. Zimin, A.V. and S.L. Salzberg, The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. *PLoS Comput Biol*, 2020. **16**(6): p. e1007981.
27. Hail Team, *Hail version 0.2.67*. 2021.
28. Altschul, S.F., et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 1997. **25**(17): p. 3389-402.
29. Quinlan, A.R. and I.M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 2010. **26**(6): p. 841-2.
30. Pertea, G. and M. Pertea, GFF Utilities: GffRead and GffCompare. *F1000Research*, 2020. **9**(304).