HHS Public Access

Author manuscript

Hum Genet. Author manuscript; available in PMC 2017 July 01.

Published in final edited form as:

Hum Genet. 2016 July ; 135(7): 727–740. doi:10.1007/s00439-016-1667-5.

Anchored pseudo-de novo assembly of human genomes identifies extensive sequence variation from unmapped sequence reads

Joshua J. Faber-Hammond[1] and **Kim H. Brown**[1]

Kim H. Brown: kibr2@pdx.edu

[1]Department of Biology, Portland State University, 1719 SW 10th Ave., SRTC 246, Portland 97207-0751, USA

Abstract

The human genome reference (HGR) completion marked the genomics era beginning, yet despite its utility universal application is limited by the small number of individuals used in its development. This is highlighted by the presence of high-quality sequence reads failing to map within the HGR. Sequences failing to map generally represent 2–5 % of total reads, which may harbor regions that would enhance our understanding of population variation, evolution, and disease. Alternatively, complete de novo assemblies can be created, but these effectively ignore the groundwork of the HGR. In an effort to find a middle ground, we developed a bioinformatic pipeline that maps paired-end reads to the HGR as separate single reads, exports unmappable reads, de novo assembles these reads per individual and then combines assemblies into a secondary reference assembly used for comparative analysis. Using 45 diverse 1000 Genomes Project individuals, we identified 351,361 contigs covering 195.5 Mb of sequence unincorporated in GRCh38. 30,879 contigs are represented in multiple individuals with ~40 % showing high sequence complexity. Genomic coordinates were generated for 99.9 %, with 52.5 % exhibiting high-quality mapping scores. Comparative genomic analyses with archaic humans and primates revealed significant sequence alignments and comparisons with model organism RefSeq gene datasets identified novel human genes. If incorporated, these sequences will expand the HGR, but more importantly our data highlight that with this method low coverage (~10–20×) next-generation sequencing can still be used to identify novel unmapped sequences to explore biological functions contributing to human phenotypic variation, disease and functionality for personal genomic medicine.

Introduction

Since its completion, the human genome reference (HGR) has been instrumental for evaluating disease variants, population differentiation, archaic hominid admixture, and more

Correspondence to: Kim H. Brown, kibr2@pdx.edu.

Electronic supplementary material The online version of this article (doi:10.1007/s00439-016-1667-5) contains supplementary material, which is available to authorized users.

Compliance with ethical standards

Competing interests The authors declare no competing interest, financial or otherwise, with the publication of this manuscript.

(Collins et al. 2004). Its completion and maintenance (i.e., error correction/gap filling), along with parallel technical advancements in sequencing methodology (i.e., next-generation sequencing) and data processing power, have provided an ability to perform genome-wide human studies. Much of this work is done using short-read technologies (i.e., Illumina; Life Sciences 454) with highly fragmented genomic DNA sequences mapped against the HGR (i.e., genome resequencing). Genome resequencing is preferred over de novo assembly for most applications, given it is less computationally expensive, requires lower sequence coverage, and does not require specific library preparations (Gnerre et al. 2010). Resequencing technologies have facilitated large-scale human genomic studies including the 1000 Genomes, 100,000 Genomes, African Genome Variation, deCODE, and Human Genome Diversity Projects (Altshuler et al. 2012; Cavalli-Sforza 2005; Colonna et al. 2014; Gudbjartsson et al. 2015; Udpa et al. 2014). These projects generated immense amounts of data to characterize SNPs, indels and copy number variants in diverse populations and assisted in identifying human disease variants (Altshuler et al. 2012; Colonna et al. 2014; Khurana et al. 2013; Mills et al. 2011a, b; Montgomery et al. 2013; Stewart et al. 2011).

While such projects have proved extremely valuable, the HGR has limitations that impact large-scale population and disease studies. Two primary limitations are the small number of individuals used to create the HGR and its reliance on a single individual genome (RPCI-11) for the majority (~75 %) of the references (Church et al. 2011; Green et al. 2010; Lander et al. 2001; Reich et al. 2009; Venter et al. 2001). These issues are highlighted, in part, by the percentage (~2–5 %) of high-quality sequence reads, with potentially valuable information, that fail to map in every individual re-sequenced to date with an estimated 3–10 Mb of sequence missing per individual and 40 Mb or more from the HGR (Dogan et al. 2014; Fujimoto et al. 2010; Li et al. 2010). While some of these sequences have scaffolds in the 1000 Genomes decoy reference (http://www.1000genomes.org), a large number remain unmapped and unanalyzed. Sequences not found in the HGR or decoy genome, One-End Anchor reads (i.e., paired-end sequences where only one mate properly maps; OEA) and non-mapping read pairs are flagged. Although there has been some efforts to use OEA sequences to identify structural variants (Kidd et al. 2010; Liu et al. 2014), these sequences are generally overlooked or underutilized in downstream analyses. Previous studies have been able to produce novel de novo assemblies in single, high coverage individuals (Dogan et al. 2014; Fujimoto et al. 2010) and pooled unmapped sequences (Korbel et al. 2007; Lander et al. 2001), but little work has been done to analyze de novo assemblies from unmapped reads. While new long single-molecule sequencing technologies (i.e., PacBio) are making de novo assembly easier (Pendleton et al. 2015), they remain cost prohibitive because of their high depth long read sequence requirements and large-scale studies still rely on genome resequencing. The broad presence of HGR unmapped sequences is evident based on 1000 Genomes copy number variant analysis, which indicate 21 % of 11,254 structural variants detected as deletions were present in the ancestral genome and subsequently lost in individuals used to generate the HGR (Mills et al. 2011b). Given that the current HGR lacks many ancestral human genome sequences (Sudmant et al. 2015), resequencing efforts miss these “lost” ancestral sequences by aligning to an incomplete HGR, and traditional de novo assembly is not yet a viable method to produce enough complete genomes to capture this unknown variation (Alkan et al. 2010, 2011). While a cost effective, high-quality de novo

assembler would be the best option for personal genomic medicine needs, such an assembler remains elusive due to library requirements, computing power (i.e., RAM) needs and overall genome complexity (Gnerre et al. 2010). Here, we describe and make available a hybrid bioinformatics pipeline that maps individual sequence reads, exports reads failing to map, and then uses these reads for de novo assembly. The resulting unmapped contigs represent large structural variants and novel genome sequences that can then be mapped to the HGR. Using this pipeline we identified previously unknown genomic sequences from 45 1000 Genomes individuals. This method serves as a cost- and time-efficient alternative to both pure genome resequencing and pure de novo assembly. Our results will enhance the HGR, highlight previously unidentified human genome diversity, and will ultimately assist with personal genomic medicine by ensuring individual-specific sequences not found in the HGR are considered when making important medical decisions.

Methods

Bioinformatics

Raw genome sequences for five individuals each from nine populations (45) covering three geographic ancestries were downloaded from the 1000 Genomes website (Supplemental Table 1). Using Bowtie2 v2.1.0 default alignment parameters (Langmead and Salzberg 2012), each individual's raw sequences were aligned to GRCh37 and common contaminant sequences, with unmappable sequences sent to individual-specific output files. Bowtie2 was selected given its base quality awareness, lower base tail quality tolerance and multi-thread capability. Although raw data was paired-end, we performed initial alignment without regard to paired relationships. Unmapped reads were de novo assembled per individual using MIRA v4.0 with parameters *denovo*, *genome*, *accurate*, and *solexa* as the sequencing technology (Chevreux et al. 1999). MIRA v4.0 was selected as it outperformed alternative de novo assembly programs Trinity and Velvet (Supplemental Figure 1) based on its production of larger average contig lengths and assembly sizes (total bp) (Gnerre et al. 2010; Grabherr et al. 2011; Zerbino and Birney 2008). The algorithm in MIRA automatically detected and labeled certain contigs as repetitive ("rep" in contig name) based on relative sequencing depth per contig. Following individual primary de novo assembly, a secondary de novo assembly was generated using all primary assembly contigs as input. This served as a reference for calculating insertion/deletion (in/del) frequencies among populations and ancestries. Analysis began prior to GRCh38 publication, and following publication all generated contigs were realigned to GRCh38, including associated patches and alternative assemblies, with successfully aligning contigs removed from downstream analysis. A command line executable version of the pipeline is available for public use at https://github.com/jfaberha/pdn_pipeline.

Sequence complexity was analyzed using DUST and Entropy algorithms in Prinseq software v0.20.4 (Schmieder and Edwards 2011). Both algorithms calculate complexity based on repeat frequencies, with values >4 or <75 considered low complexity by DUST or Entropy for the secondary assembly, respectively (Morgulis et al. 2006; Schmieder and Edwards 2011). We analyzed primary and secondary assemblies for possible redundancy and sample contamination given MIRA's conservative tendencies, which can result in more redundant

contigs than other assemblers (Mundry et al. 2012). We used Prinseq and Simplifier v0.4 (Ramos et al. 2012) to assess redundancy in the non-repetitive assembly checking for exact sequence matches as well as overlapping contigs with 5′ and/or 3′ overhangs. For Simplifier, we ignored 5 bp on both ends to account for potential low-quality base calls. BLASTn results for 1000 random primary and secondary assembly contigs were analyzed to assess sample contaminants. Random contigs were assigned using the Galaxy (Blankenberg et al. 2010; Giardine et al. 2005; Goecks et al. 2010) pipeline: 'fasta to tabluar', 'select random lines' and 'tabular to fasta'. Exported sequences were aligned to GenBank's non-redundant (nr) nucleotide database. Top species hits were examined to estimate assembly portions originating from non-primates or non-vertebrates. For presence/absence analysis, primary assemblies were aligned against the secondary assembly using Bowtie2 default parameters. Sorting and filtering alignment output in SAM format allowed us to identify all individuals having sequences matching all or part of each secondary assembly contig.

We searched the assembly for unknown genes, gene variants and genome scaffold variants containing known genes through comparisons to human (GENCODE; gencode.v22.pc_translations.fa), chimpanzee (*Pan troglodytes*) (Ensembl; Pan_troglodytes.CHIMP2.1.4.pep.abinitio.fa), and mouse (*Mus musculus*) (GENCODE; gencode.vM4. pc_translations.fa) RefGene files. Following construction of local BLAST databases using NCBI's *makeblastdb* tool (BLAST package v2.2.28+) non-repetitive contigs were queried against chimpanzee and mouse RefGene databases using BLASTx default parameters. The results were filtered to include only hits with a minimum 70 % sequence identity and BLAST alignment scores of at least 50. Contigs meeting these criteria were queried against the human RefGene database with results filtered as above.

Contigs showing stronger sequence identity to non-human RefGene sequences were examined further to determine whether they matched any submitted human genome scaffold sequence in the NCBI nr database. BLASTed contigs fit into one or more of the following categories: (1) placement in unmapped, non-HGR scaffolds, human BAC clones or other human sequences, (2) high-quality hits to primate sequences with no known human sequences, (3) low-quality hits to human/primate sequences (likely duplicated regions), (4) high-quality hits to highly conserved vertebrate sequences, not found in humans, (5) sequences with HGR terminal end homology lacking central region homology, and (6) contigs lacking any nr sequence homology.

To determine contig physical genomic locations, we used coordinates and alignment scores from SAM outputs generated by Bowtie2. Raw paired-end reads from eight individuals, having the highest detected proportion of contigs, were aligned to the secondary assembly and GRCh38 using default Bowtie2 parameters. Results from SAM outputs were exported when a single sequence mate aligned in either database. OEA sequences were cross-referenced to match genomic coordinates with specific contigs. Predicted loci were filtered to include only data with Bowtie2 alignment scores ≥30 to ensure mapping quality.

Mapping scores of OEA sequences were compared between the genome and assembly to derive qualitative insights toward the content of our unmapped assembly. Scores for 100 k random informative read pairs were visualized in a heat-map generated using the LSD

package in R (http://CRAN.R-project.org/package=LSD). To gain additional information, predicted loci ranges were mapped to the human genome using IGV v2.3, which allowed for the identification of hot spots of structural variation and concentrations of unmapped sequences (Robinson et al. 2011). For verification of our contig mapping method, we performed BLASTn analyses for a sample of 200 random contigs, which allowed us to assess whether possible BLAST results matched predicted loci coordinates. Although most secondary assembly sequences were previously undescribed in humans, a portion yielded BLASTn hits with human chromosomal coordinates. These hits were often at extreme terminal ends with the interior sequence remaining unknown.

To estimate whether the assembled contigs originated as novel in/dels in the human genome, we mapped secondary assembly sequences to four outgroup species in two comparative analyses. First, our secondary assembly was aligned to chimpanzee (GCA_000001515.4) and gorilla (*Gorilla gorilla)* (GCA_000151905.1) genomes using Bowtie2 default parameters. Second, raw sequences from one Denisovan (*Homo altai*) (ERP001519) (Meyer et al. 2012) and one Neanderthal (*Homo neanderthalensis*) (ERP000119:SAMEA845543) (Fujimoto et al. 2010) were aligned to our assembly using Bowtie2. Alignment scores ≥30 were considered high-quality alignments and reported. Each contig has a unique evolutionary history, and referencing comparative results on a fine scale provides more accurate insight into individual contig origins.

We compared sequences from our dataset to structural variant sequences identified in similar studies of Kidd et al. (2010) and Liu et al. (2014). We downloaded fasta files for 2363 contigs from the former study and 309 micSeqs from the latter. Both datasets were aligned to contigs from our primary and secondary assemblies and to the updated GRCh38 genome using Bowtie2. The number and percentage of contigs from other studies found in the current genome build and our assembly were calculated and reported.

Validation

To validate the predicted loci and population frequencies, we designed primers (Supplemental Table 2) to 20 unmapped contigs and performed PCR amplification using 45 human DNA samples. DNA samples were obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research for 45 individuals (Supplemental Table 1). Twenty-seven (27) original individuals (3/population) analyzed in our pipeline were used for confirmation of predicted presence/absence calls. The additional 18 samples (2/population, 6/ancestry) were previously unexamined and utilized to check predicted population frequencies. PCR amplification utilized 50 ng DNA, 1.25 μL (4 μM) forward and reverse primer, 6.25 μL LongAmp Hot Start 2× Master Mix (NEB) and H_20 to total 13 μL. Thermocycler conditions were: 95 °C for 30 s, 30 cycles of [95 °C for 30 s, 56–60 °C for 30 s, 65 °C for 1–10 m], then 65 °C for 10 m. Products were run on 1.2 % agarose gels for 1 h and visually scored for presence. Contigs were selected based on content and population frequencies. Primers were designed to amplify in/dels with flanking regions yielding variable product sizes, indicating presence/absence, while other internally designed primers provided presence/absence. The amplified products were Exo/Sap treated, labeled with BigDye Terminator Kit (ABI) and sent to Oregon Health Sciences University DNA Services

Core for Sanger sequencing to confirm proper amplification and genomic location. For the 27 DNA samples used in assembly construction, PCR-validated presence/absence was compared to the predicted presence/absence per individual for false-positive/negative rate calculations. For previously unexamined individuals, the observed presence/absence frequencies per geographic region were compared to predicted frequencies to test the accuracy of extrapolating predicted rates. Due to the small sample size, we performed binomial exact tests and calculated two-tailed *p* values for all amplified contigs in each ancestry (60 significance tests).

Results

Assembly

An average of 4.3 % of raw sequence reads from 45 1000 Genomes individuals failed to map to the HGR-build GRCh37 (Table 1). Unmapped sequence datasets generated were 4.5× larger than 1000 Genomes unmapped BAM files. These unmapped BAM files are not regularly updated with all new sequencing runs per individual, therefore our dataset encompassed a larger sequencing effort for certain individuals. 18.4 % of unmapped reads assembled into primary de novo contigs. The remaining reads likely failed to assemble due to low quality, high repetitiveness or failure to meet minimum coverage requirements. Assemblies averaged ~64,000 contigs at ~17× coverage per individual. Combining primary assemblies resulted in a secondary assembly with 37,144 contigs covering 72.5 Mb (Supplemental Table 3). The average GC content was 40.72 % ± 7.26 (Fig. 1a). Although only 34 % of primary assembly contigs were used to build the secondary assembly, remapping input sequences to the secondary assembly indicated that 88.4 % of primary contig sequences were represented in the final assembly.

Following GRCh38 publication, realignment of the secondary assembly to the new genome build revealed that 15.1 % (10.9 Mb) of assembled contigs not present in GRCh37 now aligned, leaving 84.9 % (30,879 contigs; 61.6 Mb) of our assembly unmapped (Supplemental Data 1). Transition from GRCh37 and GRCh38 “de facto” validated 6265 contigs. These contigs mapped primarily on chromosomes in either the main build or alternative assemblies (79.7 %) with the remainder found in unmapped contigs (Supplemental Figure 2). Combined with unassembled primary assembly contigs, our pipeline yielded 351,361 contigs covering 195.5 Mb not mapped to GRCh38 (Supplemental Data 2).

Secondary assembly sequence complexity exhibited a bimodal distribution, separating into repetitive and non-repetitive sequences (Fig. 1b). DUST and Entropy identified 40 and 35.8 % of the secondary assembly as having high sequence complexity, respectively. The unincorporated primary assembly sequences show comparable complexity rates with 41.3 % identified by Dust and 39.9 % by entropy. Secondary assembly redundancy was 0.08 % with no contigs identified by Prinseq and nine by Simplifier. Primary assembly redundancy was 4.2 % with 863 contigs removed by Prinseq and 4500 contigs by Simplifier. Filtering out repetitive and redundant sequences left 133,509 unique high-complexity contigs (64.9 Mb) of unmapped genomic sequence (EBI Biostudies:S-BSMS2).

Top BLASTn hits for 1000 random assembled contigs yielded 56.9 % with human BLAST hits (>85 % identity), suggesting contigs are real and likely duplicates of annotated human sequences. Among the remaining contigs, 33.6 % had no BLAST hits, 6.3 % hit to non-human primates and 3.1 % matched known human parasites including *Spirometra erinaceieuropaei*, *Onchocerca flexuosa* and *Dracunculus medinensis* (Fig. 2; Supplemental Table 4). Confirmed non-primate sequences were rare and likely represent contaminants introduced during DNA collection or pathogen DNA integrated into sequenced genomes. Sequences without BLAST alignments warrant further study, but likely do not represent contaminants given the number of pathogenic genomes annotated in the NCBI non-redundant database (Agarwala et al. 2015).

Population and frequency analyses

Presence/absence analysis identified 324 contigs appearing exclusive to/absent from entire ancestries. Among these, 30 non-repetitive contigs were exclusive to African ancestry, 6 to Asian and 5 to European. The remainder were absent from one ancestry, but present in the other two. This assessment excludes contigs where only a portion of the sequence exhibited population specificity. The average number of secondary assembly contigs found in European individuals was 17.7 k, Asians 18.8 k, and Africans 20.5 k. Among secondary contigs, 3267 were observed in at least 40 individuals (high frequency contigs) with 72 present in all 45 individuals, while only 88 contigs were found in 5 or fewer (low frequency contigs). Contigs were observed in a median frequency of 62.2 % of individuals (Supplemental Data 3).

Genome placement and frequencies

Raw sequence alignments provided map coordinates for 30,852 contigs (99.9 %) with 16,234 exhibiting high-quality map coordinates (52.5 %) based on strong Bowtie2 alignment scores (Supplemental Data Files 4, 5). Considering only high-quality map coordinates, 7412 produced single genomic loci (24 %), 3799 yielded two (12.3 %), and 5023 yielded more than two (16.2 %) (Supplemental Figure 3). For partial validation of predicted loci, 200 random mapping contigs were checked against the NCBI nr database for alignment. Seventy-two (72) shared high sequence identity (>99 %) with mapped human sequences in Genbank with 65 showing exact locus matches (Agarwala et al. 2015). Six (6) contigs had hits to sequences with multiple loci with at least one result matching predictions. Only one contig showed contradictory loci between methods. BLAST hits primarily appeared as HGR deletions with alignments to 5′ and 3′ termini (Fig. 3) or were ungapped alignments with non-HGR BAC clones.

Further examination of Bowtie2 mapping scores for informative read pairs provides insights into the overall content of the unassembled genome (Fig. 4). When we compare scores of split reads in the genome and assembly, we can visualize estimates of uniqueness of the aligned genomic content. A large proportion of read pairs mapped to repetitive regions in both the genome and assembly, and a second large proportion mapped to repetitive assembled contigs within unique flanking genomic sequence. These patterns are somewhat expected since repetitive elements are among the most likely sequences to vary between individuals and are masked, or otherwise difficult to assemble in the reference genome.

Beyond repetitive content in the unmapped genome, there is a high concentration of read pairs with mapping scores of 42 (maximum score) in both the genome and assembly. These read pairs aligning to single copy sequences in both databases likely contain a majority of the variable genic content between individuals and populations.

The distribution of predicted contig loci ranges throughout the reference genome shows a large number of unmapped sequences concentrating in the centromeric regions and occasionally at the telomeric regions on the short arm of acrocentric chromosomes (Supplemental Figure 4). These results are consistent with observations by other studies looking at locations of unmapped sequences (Eichler et al. 2004; Miga et al. 2015). When predicted loci ranges are filtered to include only high confidence loci with mapping scores ≥30 in both the genome and assembly, we instead see a relatively even distribution of structural variants throughout the genome with few obvious hot spots. Many of these variant regions with high confidence loci overlap with annotated human genes, underscoring the need to identify unknown gene variants within the human population that could directly alter gene function or expression.

Comparative analyses

Contig sequences were mapped to chimpanzee and gorilla genomes to explore sequence origin. The chimpanzee had 1482 secondary assembly contigs map within its genome, 754 of which had high Bowtie2 alignment scores (Supplemental Data 6A) and 717 of those being primarily non-repetitive. In the gorilla, 1079 contigs mapped with 481 having high alignment scores and 461 of those being non-repetitive. Nine-hundred and thirty-nine (939) contigs successfully mapped in both species with 396 having high-quality scores in each genome (Fig. 5a; Supplemental Data 6B, C). In ancestral human genomes, raw whole genome sequences mapping from single Neanderthal and Denisovan individuals were mapped against our secondary assembly (Supplemental Data 6A). Neanderthal sequences mapped to 18,157 contigs (58.8 %) and Denisovan sequences mapped to 21,377 contigs (69.2 %). Roughly half of the secondary assembly contigs, 14,867 (48.1 %), were present in both genomes (Fig. 5b). The low number of unmapped orthologous sequences shared with extant primate species compared to those shared with ancient hominids can be attributed to false-negative mapping hits due to higher sequence divergence from humans.

Structural variant comparison

The study by Kidd et al. (2010) generated 2363 contigs containing structural variants. Through Bowtie2 alignment, 670 of those contigs (28.4 %) mapped to contigs from our assemblies and an additional 1390 contigs (58.8 %) mapped to the current genome build, GRCh38. In total, 2060 Kidd et al. contigs (87.2 %) map to either our assembly or the current genome. A more recent study by Liu et al. (2014) found and reported another 309 structural variants. Two-hundred and nineteen (219; 70.8 %) mapped to contigs from our assembly. An additional 20 contigs (6.4 %) mapped to the current genome build. A total of 239 Liu et al. contigs (77.3 %) mapped to either our assembly or the human genome. Of all the secondary assembly contigs found in these other SV datasets, 75 were found to be in >90 % of individuals in our study compared to 5 contigs found in <10 % of individuals.

Overall, these strong concordances between studies further demonstrate that our approach generates true structural variants in the human genome.

Annotation

In total, 271 secondary assembly contigs had stronger sequence identity to non-human protein sequences than known human genes and were individually examined for potential novel functional element discovery. The results are summarized in Supplemental Table 3 and reported fully in Supplemental Data 7A–C. One potential previously unidentified human gene was identified, *Olfactory receptor 9K2-like* (*OR9K2-like*) (c14823). Two additional candidates were identified as missing from GRCh37; however, these contigs were included in GRCh38. Despite the HGR update, these genes (*FAM182B-like* and *TAS2R64-like*) remain unannotated despite sharing >98 % protein sequence identity to annotated primate genes. Three (3) additional contigs contain novel exons or transcripts, matching only to non-human sequences, and 16 contigs contain known human gene sequences in previously unplaced scaffolds or variable scaffolds from published genomic sequences. Seven (7) contigs represent unknown sequences with strong sequence identity (>90 %) to annotated pseudogenes or conserved, transcribed elements in non-humans. Another 243 contigs exhibited stronger sequence identity to non-human RefGene sequences than known human proteins and may contain additional pseudogenes or divergent functional elements.

The 271 contigs with strong non-human hits were queried against the non-human NCBI nr protein database, with the greatest number resembling chimpanzee, crab-eating macaque (*Macaca fascicularis*) and gorilla (Fig. 6). Among the previously identified 324 population-specific contigs, 12 were found in this dataset, including c14823 containing *OR9K2-like*, which shares high sequence identity with the predicted proteins (274 amino acids) from mRNAs in species ranging from bonobo (*P. paniscus*, 98 %) to American pika (*Ochotona princeps*, 71 %). The contig was also identified in the chimpanzee and gorilla genomes as well as the Denisovan raw sequence reads (Supplemental Data 6A) and appears to be maintained specifically in African ancestry (Supplementary Figure 3).

PCR validation

The ten primer sets based on variable PCR product sizes allowed for accurate genotyping while the remaining ten presence/absence primer sets could not distinguish between heterozygotes and homozygotes (Supplemental Figure 5). For consistency, the results are reported positive if individuals had at least one allele copy. Among the 27 individuals with the predicted presence/absence data, PCR results were concordant 93.3 % of the time (96.3 % median value). Five primer sets were 100 % concordant, while the lowest was 77.8 %. For the 540 alleles scored (27 individuals × 20 contigs), 3 false positives and 33 false negatives were observed (Supplemental Table 5). The predicted presence/absence frequencies for geographic ancestries (i.e., European, Asian, African) varied from observed frequencies in previously untested individuals by a mean of 16.8 % ± 15.3 (Supplemental Table 5). The largest variation was 50 %, while 26 of the 60 comparisons varied by 10 % or less. Binomial exact tests showed that 47 (78.3 %) empirical observations did not significantly differ from expectations ($p < 0.05$). Six tests produced significant results due to at least one positive PCR result when the expected value was zero (i.e., non-binomial).

Considering only tests fitting assumptions, 81.6 % (31/38) of the observed frequencies did not significantly differ from expectations. Twenty-two (22) tests had expected values of zero, and in 16 no individuals from the respective populations amplified, as predicted.

Sanger sequencing confirmed 18 anticipated target sequences with at least 97 % identity. The remaining contigs (c2621 and c14969) showed lower target sequence similarity; however, electropherograms showed overlapping peaks indicating multiple product amplification, indicating possible copy number variants. The presence/absence predictions for both contigs matched PCR results in 26 of 27 individuals. Additionally, c2621 mapped to a single predicted genome locus and c14969 mapped to two primary loci with high alignment scores (Supplemental Data 3).

Discussion

The human genome reference (HGR) completion marked the beginning of the genomics era. This valuable resource was the foundation for many fields of study and made personal genomic medicine a possibility. Personal genomic medical decisions require that the HGR be as complete and accurate as possible, but it is not yet representative of the broader human population due primarily to the small number of individuals used in its creation (Lander et al. 2001; Venter et al. 2001; Wheeler et al. 2008). Large numbers of high-quality sequence reads fail to align and are often ignored or underutilized. Here, we demonstrate that unmappable reads represent 2–5 % of sequencing efforts with ~40 % exhibiting high complexity. These regions contain important population variable markers which may be linked to population-specific disease phenotypes. Beyond its potential application for human disease research, the approaches described in this study, and similar studies, will serve as effective tools for mining unmapped sequences from previous next-generation sequencing projects across taxa (Faber-Hammond and Brown 2016).

Although short-read de novo sequence assembly can be problematic and prone to misassembly, MIRA is more conservative and less prone to such events than many other assemblers (Butler et al. 2008; Kidd et al. 2010; Mundry et al. 2012). While it is possible to have misassemblies in our secondary assembly, we expect multiple, exact repeated misassemblies in separate primary assemblies to be rare, or non-existent, prior to a second round of conservative assembly with specific coverage filters. Supporting this, only 88 secondary assembly contigs were represented in 5 or fewer primary assemblies. Assemblies were randomly sampled with little contamination observed beyond the occasional expected human pathogen sequences (i.e., viruses, parasites). Despite only 34 % of primary assembly contigs being utilized by MIRA for secondary assembly construction, 88.4 % of primary assembly sequences are represented in the secondary assembly when input contigs are aligned, suggesting the vast majority are not assembly artifacts. The unintegrated contigs contain high-quality, complex sequences and likely represent singletons excluded from secondary assembly. Sampling more individuals would likely lead to higher incorporation of rare sequences into the secondary assembly. Regardless, our assembly analysis revealed sequences resembling paralogous human and orthologous primate sequences with highly accurate predicted loci and presence/absence rates, all suggesting proper assembly. Moreover, transitioning between genome builds resulted in an additional “de facto”

validation of 6265 originally unmapped contigs (10.9 Mb) in GRCh37 that map in GRCh38, further confirming the pipeline's accurate identification and mapping of unknown sequences.

A recent publication described a slightly different assembly and characterization method for unmapped sequences (Liu et al. 2014). Although the pipelines are similar, a few key differences enabled us to identify thousands of contigs instead of hundreds. While we aligned split paired-end reads and exported all unmappable reads, Liu and colleagues used only OEA reads for assembly and analysis. This allowed us to capture unmappable read pairs entirely within unmapped regions and retroactively map a greater proportion back to the genome. The assembly methods also varied, with Liu et al. pooling reads from multiple individuals per chromosome and assembling using Velvet. For our dataset, we assembled all reads per individual and found that MIRA significantly outperformed Velvet in both the number of contigs and total bp coverage (Supplemental Figure 2). Finally, the post-assembly filtering criteria differed. Their confirmation criterion requires 50 % or more of mapped reads to cluster around a single genomic locus, which would have confirmed only ~11 k of our secondary assembly contigs. This threshold discounts dispersed (i.e., multiple mapping location) copy number variants which have been shown to be prevalent (Sudmant et al. 2010).

Comparing our contigs to those generated using slightly different methods, a high degree of overlap between datasets was observed, further confirming the validity of our assembly pipeline. Considering datasets from Liu et al. (2014) and Kidd et al. (2010), a majority of the their contigs not currently mapped in GRCh38 were found in our assembly. Differences between datasets from these separate studies and ours likely result from a combination of individuals chosen for sequencing and variations in the assembly pipeline. There is evidence for each of these contributing factors. First, contigs predicted to be of higher frequency in the total population in this study were more likely to be found in other datasets, as expected. Second, Mira performed better than two other assemblers tried for our pipeline. The vast majority of contigs from our assembly were not found in either of these databases, suggesting our pipeline captures additional structural variants that were either filtered out or not initially isolated through their approaches. By comparison, our pipeline seems to serve as a "catch all" approach that generates a large number of real and biologically interesting sequences in addition to the expected unmapped repetitive regions in the human genome. As a result, our dataset size made description and annotation more labor intensive, but we demonstrated several easily reproducible methods of fishing out sequences containing potentially biologically relevant content.

Within the secondary assembly, we observed population specificity for hundreds of contigs. The highest number observed was found in African ancestries in accordance with the "out of Africa" hypothesis (Cann et al. 1987; Stringer and McKie 1996; Templeton 2002), although this may also be related to the origin of individuals chosen for primary HGR construction. These regions included novel gene containing contigs and potential disease-related sequences. We also identified ubiquitously distributed contigs, common in all individuals examined. These regions clearly represent HGR deletions and highlight the continuing need to use additional individuals from diverse genetic backgrounds in the reference. The ancestry

of contigs was also examined in cross-species analyses by determining representation of sequences in two non-human primates and two ancient hominid genomes. This allowed us to identify significant overlap among groups. While secondary assembly contigs identified in multiple genomes are likely shared ancestral sequences, contigs shared with only one may be candidates for historical introgression. Furthermore, given that the Denisovan and Neandertal genomes were constructed using human scaffolds (Green et al. 2006, 2010; Meyer et al. 2012), most unmapped common orthologous regions remain unstudied in these species as well.

To assess possible functional genetic content, we annotated unmapped contgs based on well-established non-human datasets. Analysis identified protein-coding regions, ncRNAs, various regulatory elements and vast unmapped genomic regions containing duplicated genes. For example, *TAK1-binding protein 3* (*TAB 3*), *Double homeobox protein 4* (*DUX4*), *Cell division cycle protein 27* (*CDC27*), *T cell receptor beta* (*TCRB*) and many other genes were identified in multiple unique contigs. Although most showed weak alignments, it is possible that they are extremely copy number variable or are large duplicated and/or pseudogenized regions that remain incompletely assembled due to gaps between smaller adjacent contigs. The former hypothesis has support, given *DUX4* has previously been identified as one of the most highly copy number variable human genes (Sudmant et al. 2010). In most cases, multiple contigs with identical annotations show identical presence/absence data across individuals, suggesting they are physically linked or share evolutionary histories.

Important cancer- and disease-related loci with in/dels in known scaffolds were also identified due to in/dels preventing proper HGR alignment. One example, c3498, contains *breast carcinoma amplified sequence 4* (*BCAS4*) varying from known human scaffolds and exhibiting apparent population specificity (Supplemental Figure 5). The variant has two exon-adjacent inserts, 21 and 46 bp, present in 11/15 Africans, 3/15 Asians and no Europeans. While appearing to lie within introns, their location adjacent to exons provides the potential to alter transcription efficiency or post-transcriptional modification. Although further research is necessary, the high frequency of these insertions in Africans may contribute to known differences in breast cancer subtype rates between Caucasians and Africans (Carey et al. 2006; Stark et al. 2010). Additional population-specific sequences with potential disease-related paralogs include: *Ellis van Creveld syndrome 2* (*EVC2*), *Disrupted in schizophrenia 1* (*DISC1*), *Down syndrome encephalopathy related protein 1* (*DSERG1*), and an MHC variable region. These and other new unmapped regions have gone unexamined in previous sequencing projects, underscoring the utility for our pipeline in disease research.

Putative novel human genes were also identified using BLASTx protein homology searches. One, *OR9K2-like*, that was found only in African individuals mapped in chimpanzees, gorillas and Denisovans and is widespread among diverse mammals. Two new genes identified prior to transition from GRCh37 to GRCh38, *FAM182B-like* and *TAS2R64-like*, now map, although neither are yet recognized as human genes. Regardless, their inclusion in GRCh38 is a further validation of the pipeline's ability to identify novel sequences and

genes. Moreover, additional undiscovered genes, exons and transcripts may exist that do not resemble sequences found in chimpanzee or mouse and warrant more in-depth exploration.

Conclusions

Our pipeline identified 30,879 contigs covering 61.6 Mb of unmapped sequence within GRCh38. Contigs included repetitive and non-repetitive sequences containing a wide variety of predicted functional elements. Most contigs exhibited variable population frequencies and many represent potential disease-associated regions deserving further study. The pseudo-de novo assembly method used here can be applied to unmapped sequences from any species for single individuals or groups. Using raw sequence reads from individuals with and without phenotypes of interest, one could examine contig frequencies between groups and map candidate sequences back to the genome using paired-end relationships. Moreover, while it is generally accepted that the HGR needs additional individuals to incorporate the broader genomic diversity found in humans, most current studies incorporate extremely high coverage (>100×) and/or long-read sequence technologies (i.e., PacBio sequencing) to identify novel genomic regions. Our study highlights the fact that previously unknown genomic regions can be identified through mapping short-read sequences to the HGR at a much lower cost or by simply mining publicly available raw sequence data. This pipeline will allow future studies to explore the biological functions of previously unmapped human variation and gain new insights into human evolution, population structure, disease, and personal genomic medicine.

Availability of supporting data

A set of scripts that implement the pseudo-de novo assembly pipeline are available at GitHub: https://github.com/jfaberha/pdn_pipeline. The Supplemental Data sets supporting the results of this article are available in the Portland State University Library PDXScholar repository. Data can be found using the link: http://archives.pdx.edu/ds/psu/16928. In addition, the non-repetitive and non-redundant portions of these assemblies are available at EBI: https://wwwdev.ebi.ac.uk/biostudies/studies/S-BSMS2/.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by start-up funds from the Portland State University Department of Biology and NIEHS grant R00ES018892 to KHB.

References

Agarwala R, Barrett T, Beck J, Benson DA, Bollin C, Bolton E, Bourexis D, Brister JR, Bryant SH, Canese K, Clark K, DiCuccio M, Dondoshansky I, Federhen S, Feolo M, Funk K, Geer LY, Gorelenkov V, Hoeppner M, Holmes B, Johnson M, Khotomlianski VE, Kimchi A, Kimelman M, Kitts P, Klimke W, Krasnov S, Kuznetsov A, Landrum MJ, Landsman D, Lee JM, Lipman DJ, Lu ZY, Madden TL, Madej T, Marchler-Bauer A, Karsch-Mizrachi I, Murphy T, Orris R, Ostell J, O'Sullivan C, Panchenko A, Phan L, Preuss D, Pruitt KD, Rubinstein W, Sayers EW, Schneider V,

Schuler GD, Sherry ST, Sirotkin K, Siyan K, Slotta D, Soboleva A, Soussov V, Starchenko G, Tatusova TA, Trawick BW, Vakatov D, Wang YL, Ward M, Wilbur WJ, Yaschenko E, Zbicz K. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2015; 43:D6–D17. DOI: 10.1093/nar/gku1130 [PubMed: 25398906]

Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. Nat Methods. 2010; 8:61–65. DOI: 10.1038/nmeth.1527 [PubMed: 21102452]

Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet. 2011; 12:363–376. DOI: 10.1038/nrg2958 [PubMed: 21358748]

Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, Gabriel SB, Gibbs RA, Green ED, Hurles ME, Knoppers BM, Korbel JO, Lander ES, Lee C, Lehrach H, Mardis ER, Marth GT, McVean GA, Nickerson DA, Schmidt JP, Sherry ST, Wang J, Wilson RK, Dinh H, Kovar C, Lee S, Lewis L, Muzny D, Reid J, Wang M, Fang XD, Guo XS, Jian M, Jiang H, Jin X, Li GQ, Li JX, Li YR, Li Z, Liu X, Lu Y, Ma XD, Su Z, Tai SS, Tang MF, Wang B, Wang GB, Wu HL, Wu RH, Yin Y, Zhang WW, Zhao J, Zhao MR, Zheng XL, Zhou Y, Gupta N, Clarke L, Leinonen R, Smith RE, Zheng-Bradley X, Grocock R, Humphray S, James T, Kingsbury Z, Sudbrak R, Albrecht MW, Amstislavskiy VS, Borodina TA, Lienhard M, Mertes F, Sultan M, Timmermann B, Yaspo ML, Fulton L, Fulton R, Weinstock GM, Balasubramaniam S, Burton J, Danecek P, Keane TM, Kolb-Kokocinski A, McCarthy S, Stalker J, Quail M, Davies CJ, Gollub J, Webster T, Wong B, Zhan YP, Auton A, Yu F, Bainbridge M, Challis D, Evani US, Lu J, Nagaswamy U, Sabo A, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491:56–65. DOI: 10.1038/nature11632 [PubMed: 23128226]

Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J. Galaxy: a web-based genome analysis tool for experimentalists. Curr Protoc Mol Biol. 2010; Chapter 19(Unit 19.10):1–21.

Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, Jaffe DB. ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Res. 2008; 18:810–820. DOI: 10.1101/gr.7337908 [PubMed: 18340039]

Cann RL, Stoneking M, Wilson AC. Mitochondrial DNA and human-evolution. Nature. 1987; 325:31–36. DOI: 10.1038/325031a0 [PubMed: 3025745]

Carey LA, Perou CM, Livasy CA, Dressler LG, Cowan D, Conway K, Karaca G, Troester MA, Tse CK, Edmiston S, Deming SL, Geradts J, Cheang MCU, Nielsen TO, Moorman PG, Earp HS, Millikan RC. Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. JAMA J Am Med Assoc. 2006; 295:2492–2502. DOI: 10.1001/jama.295.21.2492

Cavalli-Sforza LL. Opinion—the human genome diversity project: past, present and future. Nat Rev Genet. 2005; 6:333–340. DOI: 10.1038/nrg1579 [PubMed: 15803201]

Chevreux B, Wetter T, Suhai S. Genome sequence assembly using trace signals and additional sequence information. Comput Sci Biol Proc German Conf Bioinf (GCB). 1999; 99:45–56.

Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, Chen HC, Agarwala R, McLaren WM, Ritchie GRS, Albracht D, Kremitzki M, Rock S, Kotkiewicz H, Kremitzki C, Wollam A, Trani L, Fulton L, Fulton R, Matthews L, Whitehead S, Chow W, Torrance J, Dunn M, Harden G, Threadgold G, Wood J, Collins J, Heath P, Griffiths G, Pelan S, Grafham D, Eichler EE, Weinstock G, Mardis ER, Wilson RK, Howe K, Flicek P, Hubbard T. Modernizing reference genome assemblies. PLoS Biol. 2011; doi: 10.1371/journal.pbio.1001091

Collins FS, Lander ES, Rogers J, Waterston RH. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004; 431:931–945. DOI: 10.1038/nature03001 [PubMed: 15496913]

Colonna V, Ayub Q, Chen Y, Pagani L, Luisi P, Pybus M, Garrison E, Xue Y, Tyler-Smith C. Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. Genome Biol. 2014; 15:R88. [PubMed: 24980144]

Dogan H, Can H, Otu HH. Whole genome sequence of a Turkis individual. PLoS One. 2014; 9:e85233.doi: 10.1371/journal.pone.0085233 [PubMed: 24416366]

Eichler EE, Clark RA, She XW. An assessment of the sequence gaps: unfinished business in a finished human genome. Nat Rev Genet. 2004; 5:345–354. DOI: 10.1038/nrg1322 [PubMed: 15143317]

Faber-Hammond JJ, Brown KH. Pseudo-de novo assembly and analysis of unmapped genome sequence reads in wild zebrafish reveals novel gene content. Zebrafish. 2016; 13:95–102. DOI: 10.1089/zeb.2015.1154 [PubMed: 26886859]

Fujimoto A, Nakagawa H, Hosono N, Nakano K, Abe T, Boroevich KA, Nagasaki M, Yamaguchi R, Shibuya T, Kubo M, Miyano S, Nakamura Y, Tsunoda T. Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massibely parallel sequencing. Nat Genet. 2010; 42:931–936. DOI: 10.1038/ng.691 [PubMed: 20972442]

Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A. Galaxy: a platform for interactive large-scale genome analysis. Genome Res. 2005; 15:1451–1455. DOI: 10.1101/gr.4086505 [PubMed: 16169926]

Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, Berlin AM, Aird D, Costello M, Daza R, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES, Jaffe DB. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci. 2010; 108:1513–1518. DOI: 10.1073/pnas.1017351108 [PubMed: 21187386]

Goecks J, Nekrutenko A, Taylor J, Galaxy T. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010; doi: 10.1186/gb-2010-11-8-r86

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng QD, Chen ZH, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011; 29:644–652. DOI: 10.1038/nbt.1883 [PubMed: 21572440]

Green RE, Krause J, Ptak SE, Briggs AW, Ronan MT, Simons JF, Du L, Egholm M, Rothberg JM, Paunovic M, Paabo S. Analysis of one million base pairs of Neanderthal DNA. Nature. 2006; 444:330–336. DOI: 10.1038/nature05336 [PubMed: 17108958]

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai WW, Fritz MHY, Hansen NF, Durand EY, Malaspinas AS, Jensen JD, Marques-Bonet T, Alkan C, Prufer K, Meyer M, Burbano HA, Good JM, Schultz R, Aximu-Petri A, Butthof A, Hober B, Hoffner B, Siegemund M, Weihmann A, Nusbaum C, Lander ES, Russ C, Novod N, Affourtit J, Egholm M, Verna C, Rudan P, Brajkovic D, Kucan Z, Gusic I, Doronichev VB, Golovanova LV, Lalueza-Fox C, de la Rasilla M, Fortea J, Rosas A, Schmitz RW, Johnson PLF, Eichler EE, Falush D, Birney E, Mullikin JC, Slatkin M, Nielsen R, Kelso J, Lachmann M, Reich D, Paabo S. A draft sequence of the neandertal genome. Science. 2010; 328:710–722. DOI: 10.1126/science.1188021 [PubMed: 20448178]

Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, Besenbacher S, Magnusson G, Halldorsson BV, Hjartarson E, Sigurdsson GT, Stacey SN, Frigge ML, Holm H, Saemundsdottir J, Helgadottir HT, Johannsdottir H, Sigfusson G, Thorgeirsson G, Sverrisson JT, Gretarsdottir S, Walters GB, Rafnar T, Thjodleifsson B, Bjornsson ES, Olafsson S, Thorarinsdottir H, Steingrimsdottir T, Gudmundsdottir TS, Theodors A, Jonasson JG, Sigurdsson A, Bjornsdottir G, Jonsson JJ, Thorarensen O, Ludvigsson P, Gudbjartsson H, Eyjolfsson GI, Sigurdardottir O, Olafsson I, Arnar DO, Magnusson OT, Kong A, Masson G, Thorsteinsdottir U, Helgason A, Sulem P, Stefansson K. Large-scale whole-genome sequencing of the Icelandic population. Nat Genet. 2015; doi: 10.1038/ng.3247

Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, Sboner A, Lochovsky L, Chen JM, Harmanci A, Das J, Abyzov A, Balasubramanian S, Beal K, Chakravarty D, Challis D, Chen Y, Clarke D, Clarke L, Cunningham F, Evani US, Flicek P, Fragoza R, Garrison E, Gibbs R, Guemues ZH, Herrero J, Kitabayashi N, Kong Y, Lage K, Liluashvili V, Lipkin SM, Mac-Arthur DG, Marth G, Muzny D, Pers TH, Ritchie GRS, Rosenfeld JA, Sisu C, Wei XM, Wilson M, Xue YL, Yu FL, Dermitzakis ET, Yu HY, Rubin MA, Tyler-Smith C, Gerstein M. Genomes Project Consortium (2013) Integrative annotation of variants from 1092 humans: application to cancer genomics. Science. 1000; 342:84.doi: 10.1126/science.1235587

Kidd JM, Sampas N, Antonacci F, Graves T, Fulton R, Hayden HS, Alkan C, Malig M, Ventura M, Giannuzzi G, Kallicki J, Anderson P, Tsalenko A, Yamada NA, Tsang P, Kaul R, Wilson RK, Bruhn L, Eichler EE. Characterization of missing human genome sequences and copy-number

polymorphic insertions. Nat Methods. 2010; 7:365–371. DOI: 10.1038/nmeth.1451 [PubMed: 20440878]

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders ACE, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, Snyder M. Paired-end mapping reveals extensive structural variation in the human genome. Science. 2007; 318:420–426. [PubMed: 17901297]

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, et al. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001; 409:860–921. DOI: 10.1038/35057062 [PubMed: 11237011]

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9:357–359. DOI: 10.1038/nmeth.1923 [PubMed: 22388286]

Li R, Li Y, Zheng H, Luo R, Zhu H, Li Q, Qian W, Ren Y, Tian G, Li J, Zhou G, Zhu X, Wu H, Qin J, Jin X, Li D, Cao H, Hu X, Blanche H, Cann H, Zhang X, Li S, Bolund L, Kristiansen K, Yang H, Wang J, Wang J. Building the sequence map of the human pangenome. Nat Biotechnol. 2010; 28:57–63. DOI: 10.1038/nbt.1596 [PubMed: 19997067]

Liu Y, Koyutürk M, Maxwell S, Xiang M, Veigl M, Cooper RS, Tayo BO, Li L, LaFramboise T, Wang Z, Zhu X, Chance MR. Discovery of common sequences absent in the human reference genome using pooled samples from next generation sequencing. BMC Genom. 2014; 15:685.doi: 10.1186/1471-2164-15-685

Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prufer K, de Filippo C, Sudmant PH, Alkan C, Fu QM, Do R, Rohland N, Tandon A, Siebauer M, Green RE, Bryc K, Briggs AW, Stenzel U, Dabney J, Shendure J, Kitzman J, Hammer MF, Shunkov MV, Derevianko AP, Patterson N, Andres AM, Eichler EE, Slatkin M, Reich D, Kelso J, Paabo S. A high-coverage genome sequence from an archaic Denisovan individual. Science. 2012; 338:222–226. DOI: 10.1126/science.1224344 [PubMed: 22936568]

Miga KH, Eisenhart C, Kent WJ. Utilizing mapping targets of sequences underrepresented in the reference assembly to reduce false positive alignments. Nucleic Acids Res. 2015; doi: 10.1093/nar/gkv671

Mills RE, Pittard WS, Mullaney JM, Farooq U, Creasy TH, Mahurkar AA, Kemeza DM, Strassler DS, Ponting CP, Webber C, Devine SE. Natural genetic variation caused by small insertions and deletions in the human genome. Genome Res. 2011a; 21:830–839. DOI: 10.1101/gr.115907.110 [PubMed: 21460062]

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, Chinwalla A, Conrad DF, Fu Y, Grubert F, Hajirasouliha I, Hormozdiari F, Iakoucheva LM, Iqbal Z, Kang S, Kidd JM, Konkel MK, Korn J, Khurana E, Kural D, Lam HYK, Leng J, Li R, Li Y, Lin CY, Luo R, Mu XJ, Nemesh J, Peckham HE, Rausch T, Scally A, Shi X, Stromberg MP, Stütz AM, Urban AE, Walker JA, Wu J, Zhang Y, Zhang ZD, Batzer MA, Ding L, Marth GT, McVean G, Sebat J, Snyder M, Wang J, Ye K, Eichler EE, Gerstein MB, Hurles ME, Lee C, McCarroll SA, Korbel JO. 1000 Genomes Project. Mapping copy number variation by population scale genome sequencing. Nature. 2011b; 470:59–65. [PubMed: 21293372]

Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZDD, Mu XJ, Ananda G, Howie B, Karczewski KJ, Smith KS, Anaya V, Richardson R, Davis J, MacArthur DG, Sidow A, Duret L, Gerstein M, Makova KD, Marchini J, McVean G, Lunter G. Genomes Project Consortium (2013) The origin, evolution, and functional impact of short insertion-deletion variants identified in 179

human genomes. Genome Res. 1000; 23:749–761. DOI: 10.1101/gr.148718.112 [PubMed: 23478400]

Morgulis A, Gertz EM, Schaffer AA, Agarwala R. Window-Masker: window-based masker for sequenced genomes. Bioinformatics. 2006; 22:134–141. DOI: 10.1093/bioinformatics/bit774 [PubMed: 16287941]

Mundry M, Bornberg-Bauer E, Sammeth M, Feulner PGD. Evaluating characteristics of de novo assembly software on 454 transcriptome data: a simulation approach. PLoS One. 2012; doi: 10.1371/journal.pone.0031410

Pendleton M, Sebra R, Pang AWC, Ummat A, Franzen O, Rausch T, Stutz AM, Stedman W, Anantharaman T, Hastie A, Dai H, Fritz MHY, Cao H, Cohainl A, Deikusl G, Durrett RE, Blanchard SC, Altman R, Chin CS, Guo Y, Paxinos EE, Korbe JO, Darne RB, McCombiemii WR, Kwok PY, Mason CE, Schadt EE, Bashirl A. Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat Methods. 2015; 12:780–786. DOI: 10.1038/nmeth.3454 [PubMed: 26121404]

Ramos RTJ, Carneiro A, Azevedo RV, Schneider MP, Barh D, Silva A. Simplifier: a web tool to eliminate redundant NGS contigs. Bioinformation. 2012; 8:996–999. [PubMed: 23275695]

Reich D, Nalls MA, Kao WH, Akylbekova EL, Tandon A, Patterson N, Mullikin J, Hsueh WC, Cheng CY, Coresh J, Boerwinkle E, Li M, Waliszewska A, Neubauer J, Li R, Leak TS, Ekunwe L, Files JC, Hardy CL, Zmuda JM, Taylor HA, Ziv E, Harris TB, Wilson JG. Reduced neutrophil count in people of African descent is due to a regulatory variant in the Duffy antigen receptor for chemokines gene. PLoS Genet. 2009; 5:e1000360.doi: 10.1371/journal.pgen.1000360 [PubMed: 19180233]

Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. Nat Biotechnol. 2011; 29:24–26. DOI: 10.1038/nbt.1754 [PubMed: 21221095]

Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011; 27:863–864. DOI: 10.1093/bioinformatics/btr026 [PubMed: 21278185]

Stark A, Kleer CG, Martin I, Awuah B, Nsiah-Asare A, Takyi V, Braman M, Quayson SE, Zarbo R, Wicha M, Newman L. African ancestry and higher prevalence of triple-negative breast cancer findings from an International Study. Cancer. 2010; 116:4926–4932. DOI: 10.1002/cncr.25276 [PubMed: 20629078]

Stewart C, Kural D, Stromberg MP, Walker JA, Konkel MK, Stutz AM, Urban AE, Grubert F, Lam HYK, Lee WP, Busby M, Indap AR, Garrison E, Huff C, Xing JC, Snyder MP, Jorde LB, Batzer MA, Korbel JO, Marth GT, Genomes P. A comprehensive map of mobile element insertion polymorphisms in humans. PLoS Genet. 2011; doi: 10.1371/journal.pgen.1002236

Stringer, C.; McKie, R. African exodus: the origins of modern humanity. Henery Holt and Company; New York: 1996.

Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J, Eichler EE. Diversity of human copy number variation and multicopy genes. Science. 2010; 330:641–646. DOI: 10.1126/science.1197005 [PubMed: 21030649]

Sudmant, PH.; Mallick, S.; Nelson, BJ.; Hormozdiari, F.; Krumm, N.; Huddleston, J.; Coe, BP.; Baker, C.; Nordenfelt, S.; Bamshad, M.; Jorde, LB.; Posukh, OL.; Sahakyan, H.; Watkins, WS.; Yepiskoposyan, L.; Abdullah, MS.; Bravi, CM.; Capelli, C.; Hervig, T.; Wee, JTS.; Tyler-Smith, C.; Driem, G.; Romero, IG.; Jha, AR.; Karachanak-Yankova, S.; Toncheva, D.; Comas, D.; Henn, B.; Kivisild, T.; Ruiz-Linares, A.; Sajantila, A.; Metspalu, E.; Parik, J.; Villems, R.; Starikovskaya, EB.; Ayodo, G.; Beall, CM.; Rienzo, AD.; Hammer, M.; Khusainova, R.; Khusnutdinova, E.; Klitz, W.; Winkler, C.; Labuda, D.; Metspalu, M.; Tishkoff, SA.; Dryomov, S.; Sukernik, R.; Patterson, N.; Reich, D.; Eichler, EE. Global diversity, population stratification, and selection of human copy number variation. Science. 2015. http://sciencemag.org/content/early/recent/6August2015/Page2/

Templeton AR. Out of Africa again and again. Nature. 2002; 416:45–51. DOI: 10.1038/416045a [PubMed: 11882887]

Udpa N, Ronen R, Zhou D, Liang J, Stobdan T, Appenzeller O, Yin Y, Du Y, Guo L, Cao R, Wang Y, Jin X, Huang C, Jia W, Cao D, Guo G, Claydon VE, Hainsworth R, Gamboa JL, Zibenigus M, Zenebe G, Xue J, Liu S, Frazer KA, Li Y, Bafna V, Haddad GG. Whole genome sequencing of

Ethiopian highlanders reveals conserved hypoxia tolerance genes. Genome Biol. 2014; 15:R36. [PubMed: 24555826]

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XQH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang JH, Miklos GLG, Nelson C, Broder S, Clark AG, Nadeau C, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng ZM, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge WM, Gong FC, Gu ZP, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke ZX, Ketchum KA, Lai ZW, Lei YD, Li ZY, Li JY, Liang Y, Lin XY, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue BX, Sun JT, Wang ZY, Wang AH, Wang X, Wang J, Wei MH, Wides R, Xiao CL, Yan CH, et al. The sequence of the human genome. Science. 2001; 291:1304.doi: 10.1126/science.1058040 [PubMed: 11181995]

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM. The complete genome of an individual by massively parallel DNA sequencing. Nature. 2008; 452:872–876. DOI: 10.1038/nature06884 [PubMed: 18421352]

Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008; 18:821–829. DOI: 10.1101/gr.074492.107 [PubMed: 18349386]

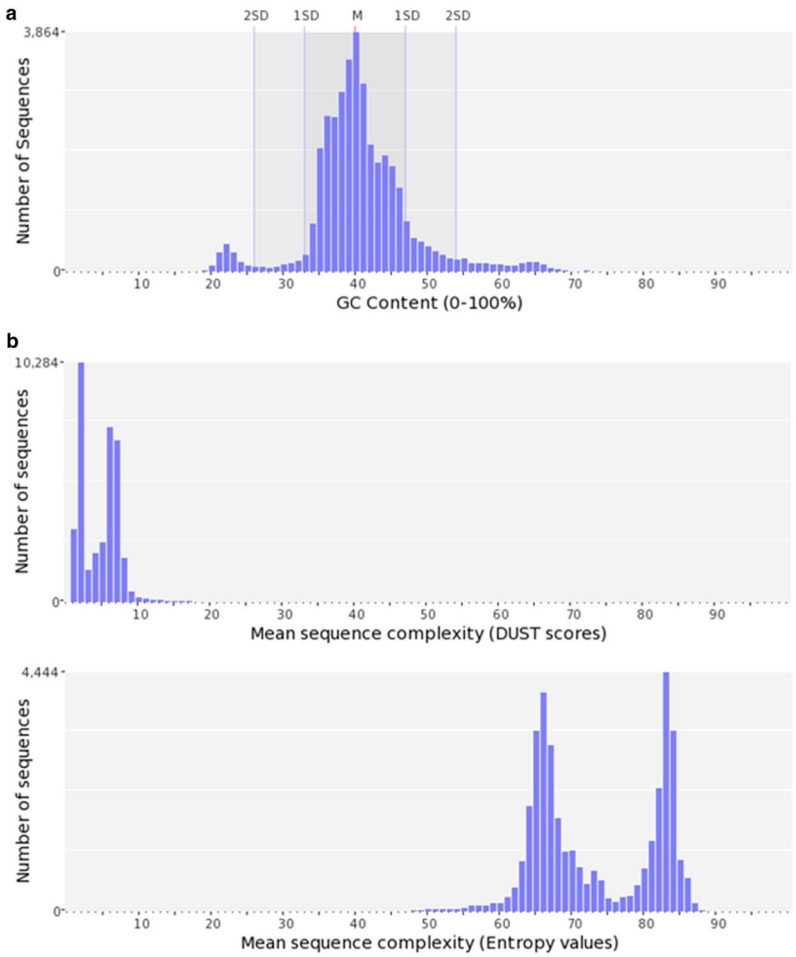

Fig. 1.

Assembly characteristics. **a** Graph showing GC content of the secondary assembly. **b** Plots of sequence complexity as calculated by DUST and Entropy algorithms. Sequences are considered to be of low complexity with values >3 using the DUST algorithm and <75 using the Entropy algorithm. Both plots show bimodal distributions based on repetitiveness

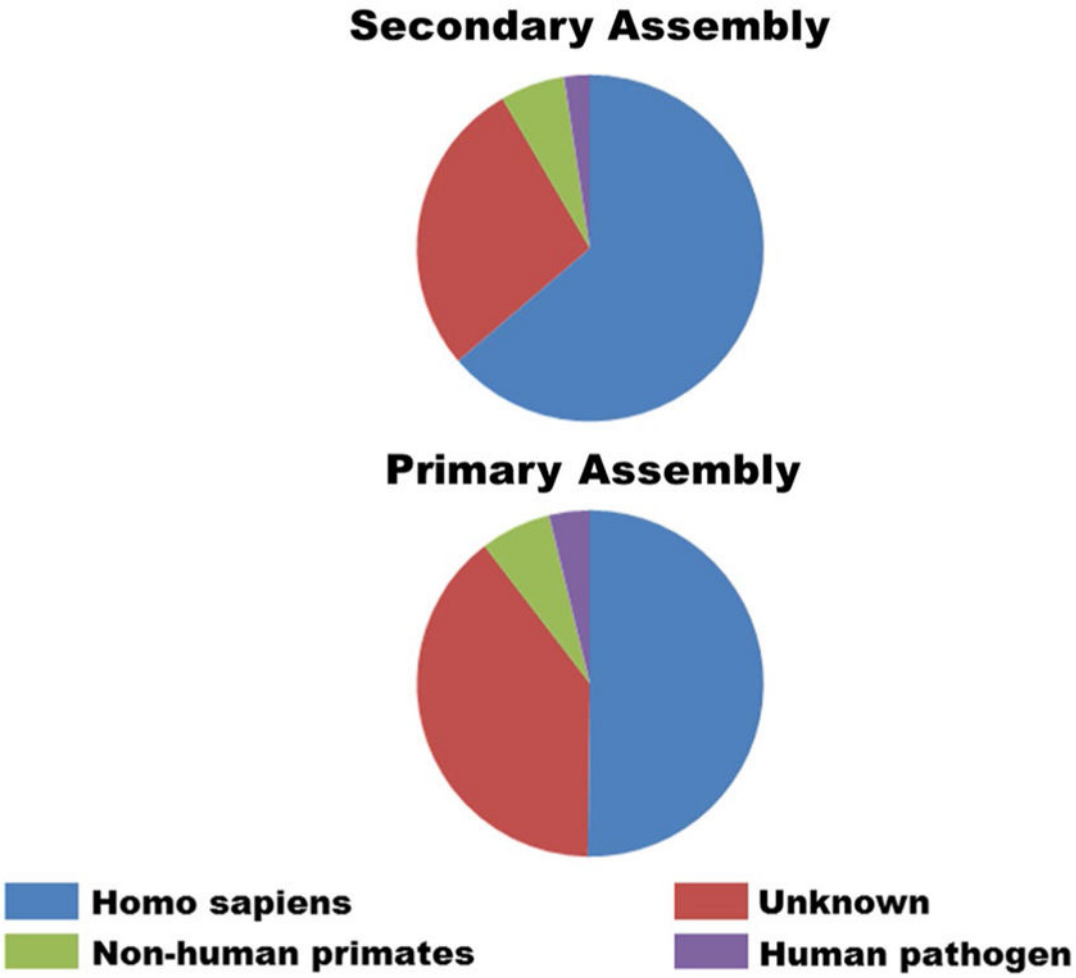

Fig. 2.

Top species BLASTn hits for 500 random contigs from the secondary reference assembly and the primary assembly contigs that remain unincorporated in the secondary assembly

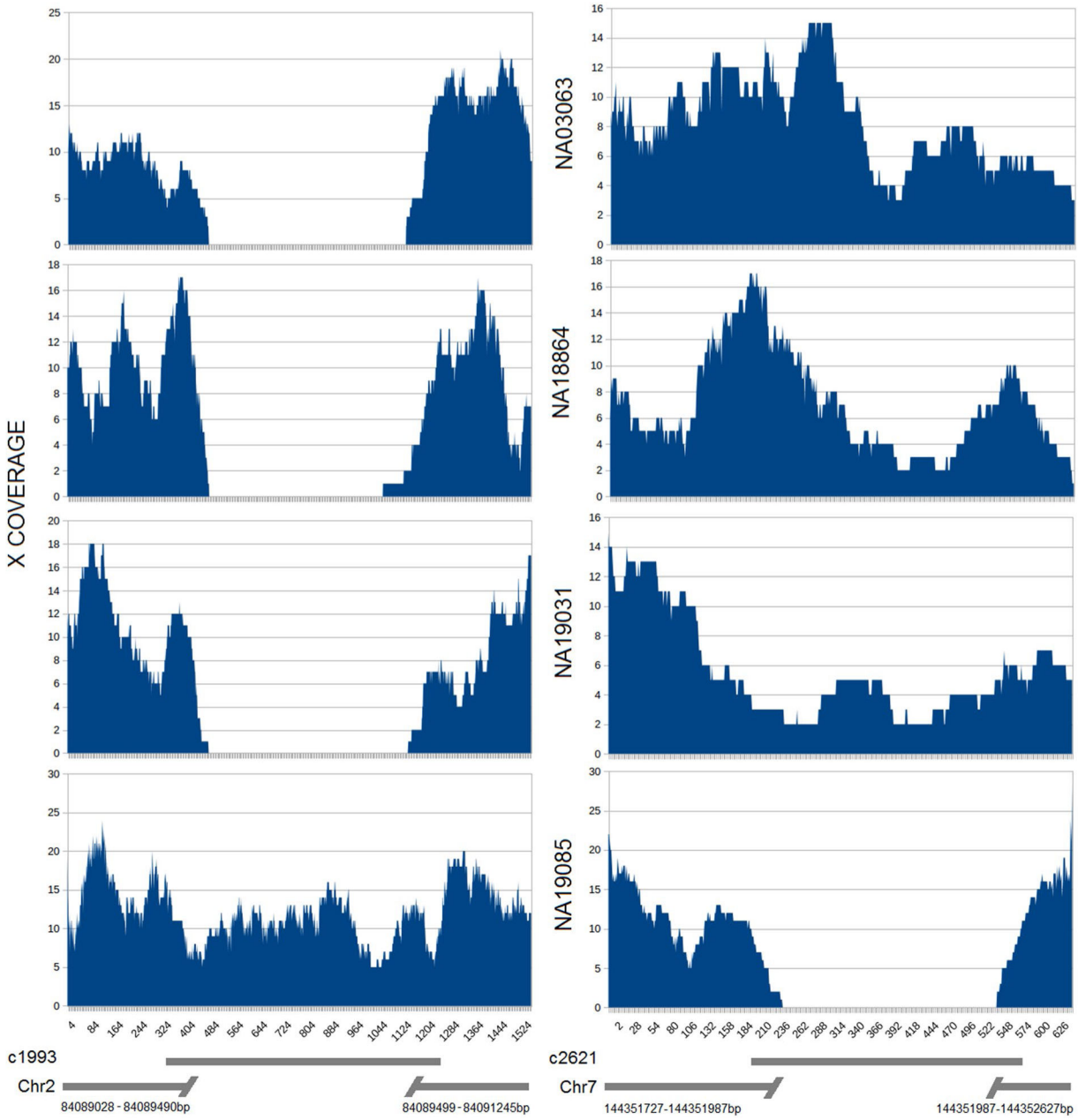

Fig. 3.

Plots showing coverage across in/dels in two non-repetitive contigs c1993 and c2621 using raw reads from four individuals (1000 Genomes Project). Contigs show sequence identity to Chr2 and Chr7 in the HGR on the 5′ and 3′ ends. Scaffolds are extended to include genomic sequence outside the respective contig boundaries for context

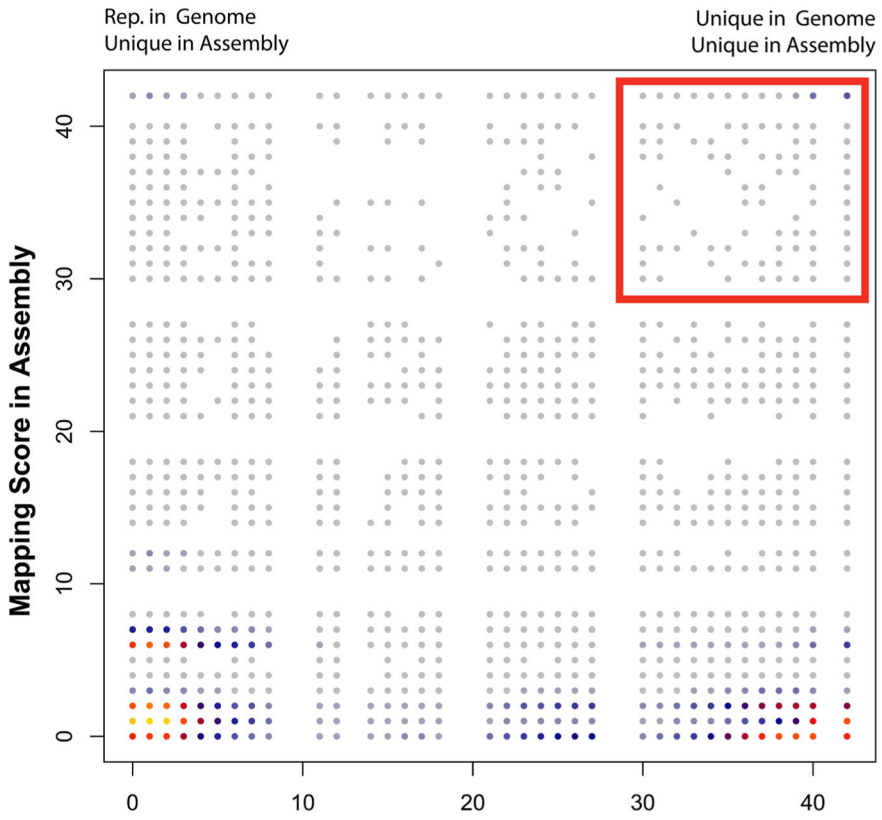

Fig. 4.

Heat map distribution of Bowtie2 mapping scores for 100 k randomly selected one-end anchored read pairs. The figure was created using LSD in R. *Cool colors* represent low relative density of points and *warm colors* represent high density of points. Scores range from highly repetitive (0) to completely unique (42). Bowtie2 alignment scores are reported as discrete values; therefore each plot point represents overlapping data points. The *red box* designates read pairs with mapping scores of at least 30 in both the genome and assembly (high confidence loci). A strong concentration of read pairs that are completely unique in the genome (score = 42 in both the assembly and reference genome). Unsurprisingly, we also see a large portion of read pairs that appear repetitive in both the assembly and genome (color figure online)

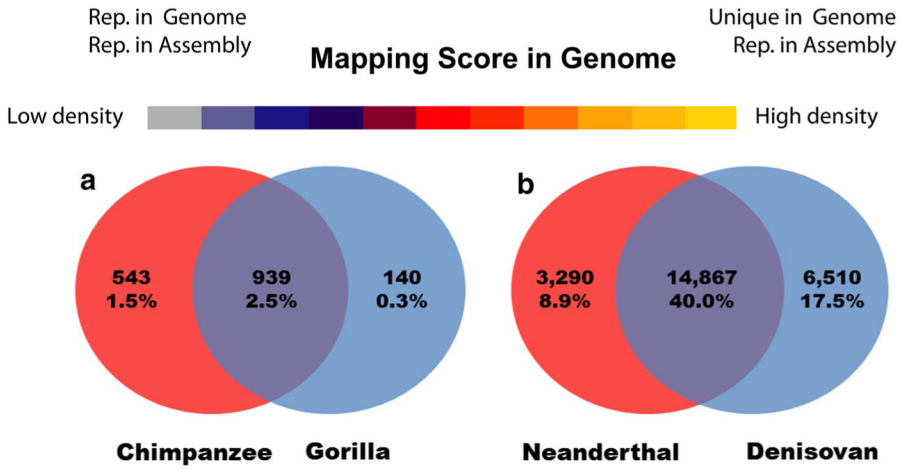

Fig. 5.

Venn diagrams show the number of contigs and portion of the ~31 k assembly found in **a** chimpanzee and gorilla genomes, and in **b** Neanderthal and Denisovan next-generation sequence read pools. Only high-quality alignments (Bowtie2 score ≥30) are reported

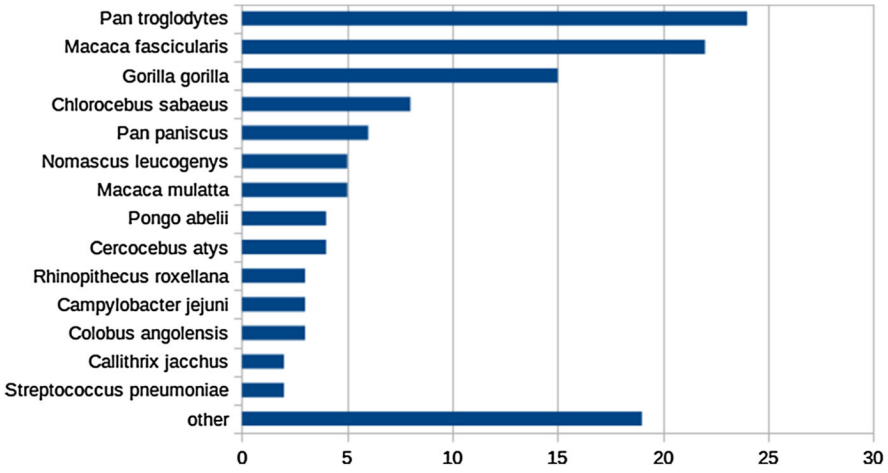

Fig. 6.

Top non-human species BLASTx hit distribution in the NCBI nr database for contigs identified as having stronger hits to chimpanzee and mouse RefGene protein sequences than to human sequences

Table 1

HRG mapping statistics for 45 sequence sets from the 1000 Genomes Project utilizing paired-end reads with each read mapped independently. Disparities in percentages are due to variation in sequencing depth between individuals

	Illumina raw reads	Unmapped reads	Percent
Average	287,005,118	11,499,233	4.00
Total	11,482,204,736	494,467,016	4.30