Doctoral Thesis

The Korean Reference Genome

Yun Sung Cho

Department of Biomedical Engineering

Graduate School of UNIST

2017

The Korean Reference Genome

Yun Sung Cho

Department of Biomedical Engineering

Graduate School of UNIST

The Korean Reference Genome

A thesis/dissertation
submitted to the Graduate School of UNIST
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Yun Sung Cho

5. 30. 2017
Approved by

____

Advisor
Jong Bhak

The Korean Reference Genome

Yun Sung Cho

This certifies that the thesis/dissertation of Yun Sung Cho is approved.

5. 30. 2017

signature

____

Advisor: Jong Bhak

signature

____

Cheol-Min Ghim: Thesis Committee Member #1

signature

____

Dougu Nam: Thesis Committee Member #2

signature

____

Taejoon Kwon: Thesis Committee Member #3

signature

____

Semin Lee: Thesis Committee Member #4;

Abstract

Human genomes are routinely compared against a universal human reference. However, this strategy could miss population-specific and personal genomic variations, which may be detected more efficiently using an ethnically-relevant or personal reference. Here I describe principles and methods in constructing a hybrid assembly of the first Korean reference genome (KOREF) by compiling all the major contemporary sequencing and mapping technologies: short and long paired-end sequences, synthetic and single molecule long reads, and optical and nanochannel genome maps. This low-cost hybrid approach shows the feasibility of routine reference-quality *de novo* assembled genomes to precisely analyze many personal and ethnic genomes in the future. I also introduce the concept of the consensus variome reference, providing information on millions of variants incorporated directly from 40 additional ethnically homogeneous genomes from the Korean Personal Genome Project. KOREF is the first *de novo* assembled consensus variome reference. KOREF has been constructed according to standardized production and evaluation procedures, and registered as a standard reference data for ethnic Korean genomes by evaluating its traceability, uncertainty, and consistency. By comparing KOREF against other ethnic references, I find that the ethnically-relevant consensus reference can be beneficial for efficient variants detection and possibly other purposes in the future. Therefore, I propose that, despite the limited level of divergence within our species, the level of genomic scale variation is sufficiently high to warrant the use of ethnically-relevant references for large-scale personal and disease genome projects. Systematic comparison of human assemblies also shows the importance of assembly quality, suggesting the necessity of new technologies to comprehensively map ethnic and personal genomic structure variations. In the era of large-scale population genome projects, the leveraging of ethnicity-specific genome assemblies as well as the human reference genome will accelerate mapping all human genome diversity on Earth.

Contents

List of Figures

List of Tables

Nomenclature

1KGP, 1000 genomes project

BAC, bacterial artificial chromosome

CHM, complete hydatidiform mole

dbRIP, database of retrotransposon insertion polymorphisms

DGV, database of genomic variants

EBV, Epstein-Barr virus

FBS, fetal bovine serum

GATK, genome analysis toolkit

GRCh38, genome reference consortium human build 38 patch

GRCh38_C, consensus Asian GRCh38

HGDP, human genome diversity project

indel, insertion or deletion

KCLB, Korean cell line bank

KOREF, Korean reference genome

KOREF_C, single Korean reference + consensus variome

KOREF_S, single Korean reference assembly

KPGP, Korean personal genome project

LD, linkage disequilibrium

MAF, minor allele frequency

MDS, multidimensional scaling

mtDNA, mitochondrial DNA

NCSRD, national center for standard reference data

NGS, next-generation sequencing

nsSNV, non-synonymous single nucleotide variation

PAPGI, Pan-Asian population genomics initiative

PBMC, peripheral blood mononuclear cell

PCR, polymerase chain reaction

PGP, personal genome project

SMRM, single molecule restriction map

SMRT, single-molecule real-time sequencing technology

SNV, single nucleotide variation

SV, structural variation

TSLR, TruSeq synthetic long read

I. Introduction

The standard human reference (currently GRCh38), which is mostly based on Caucasian and African ancestry[1,2], is accurate, precise, and extensive. Because of the relatively small long term effective population size of anatomically modern humans (estimated to be as small as ~10,000)[3,4], such a reference is adequate for most purposes and routinely used in research and biomedical applications. However, certain population specific variants could be missed with such a universal reference, and the current research efforts to map human diversity, including low frequency and structural variants, would benefit from ethnically relevant references[5,6]. Since the publication of the first draft of the human reference genome in 2001[7], sequencing technologies have advanced rapidly. In 2007, the diploid genome of a Caucasian male was sequenced and assembled using Sanger sequencing technology (HuRef)[8]. Later, the genomes of a Chinese (YH), an African (2009), a Caucasian (HsapALLPATHS1, here called NA12878_Allpaths, 2011), and a Mongolian (2014) were built using Illumina short-read sequencing data[9-11]. In 2014, a complete hydatidiform mole genome (CHM1_1.1) was assembled, albeit reference-guided, using Illumina short-reads and indexed bacterial artificial chromosome (BAC) clones[12]. In 2015, a haplotype-resolved diploid YH genome was assembled using fosmid pooling together with short-read sequence data[13]. These assemblies, although useful and important for genomics researches, are not of sufficient accuracy or overall quality to be considered a general purpose standard reference genome[14].

The recent increased availability of long-range sequencing and mapping methods has important implications for the generation of references for ethnic groups and even personal genomes, especially for disease associated structural variations (SVs). Long range data can improve draft genome assemblies by increasing the scaffold size, efficiently closing gaps, resolving complex regions, and identifying SVs[15-22] at relatively low costs. Notable approaches are single-molecule real-time sequencing technology (SMRT) and highly-parallel library preparation and local assembly of short reads (synthetic long reads) for resolving complex DNA regions and filling genomic gaps[15-17]. For instance, single haplotype human genomes were constructed using single-molecule long read sequencing (CHM1_PacBio_r2 and CHM13). Long-read methods can be complemented and validated by two high-throughput mapping methods: optical mapping and nanochannel-based genome mapping. The most representative cases are the NA12878 (ASM101398v1; here called NA12878_single) and HX1 (a Chinese individual) genomes, which were hybrid assembled by combining single-molecule long reads with single-molecule genome maps[21,22]. Assemblies incorporating high-throughput short reads and long range mapping or sequencing data, or hybrid assemblies, can enhance the quality, providing much longer scaffolds with validation and adjustment of complex genomic regions[19-22].

Complementary to reference genome projects, which provide accurate templates, population genome projects, such as Personal Genome Project (PGP)[23] and the 1,000 Genomes Project (1KGP)[24,25], provide valuable variome information that is fundamental to many biomedical research projects. The PGP was initiated in 2005 to publicly share personal genome, health, and trait data, crucial in understanding the diverse functional consequences associated with genetic variation. Recently, large scale population genome projects in Britain and the Netherlands have been launched to identify population-specific rare genetic variations and disease-causing variants[26,27]. The single reference and population derived genomic variation types and frequencies (variome) are the pillars of genomics.

Here, I report two versions of the Korean reference (KOREF) genome (KOREF_S: a single reference assembly and KOREF_C: single reference + consensus variome), produced as part of PGP, by utilizing hybrid sequencing and mapping data. KOREF provides another high quality East-Asian reference to complement GRCh38. KOREF was initiated by the Korean Ministry of Science and Technology in 2006 to generate a national genome and variome references. To deal with the issues inherent to short reads, I use data from a number of different technologies (short and long paired-end sequences, synthetic and single molecule long reads, and optical and nanochannel genome maps) to build a high quality hybrid assembly of a male donor, KOREF_S (Fig. 1). Furthermore, I integrate information from 40 high-coverage whole genomes (based on short reads) from the Korean PGP (KPGP)[28] to generate a population-wide consensus Korean reference, KOREF_C. I compared the genomic structure of KOREF_C with other human genome assemblies, uncovering many structural differences, including ethnic-specific highly frequent structural variants. Importantly, the identification of SVs is largely affected by the sequencing platform used and assembly quality, suggesting the need for long-read sequences and a higher quality assembly to comprehensively map the ethnic and personal genomic structures. Accompanied by multi-ethnic PGP data, in the future, many low-cost personal, national, and ethnic genome references will accelerate the completion of mapping all human genome diversity in both single nucleotide variations (SNVs) and SVs. My endeavor to construct KOREF is not limited to one ethnic group, but it is towards the era of personalized complete reference genome where everyone has his or her own reference of genome, transcriptome, proteome, and other omics data at a fraction of the cost spent for the human genome project decades ago. This has a far reaching implications in the society as well as in science as it will revolutionize how humans are born, live, and die in the future with an extensive amount of omics data of their own life. In a way, this is an important part of the ultimate democratization of genomics data and associated technologies for humanity.

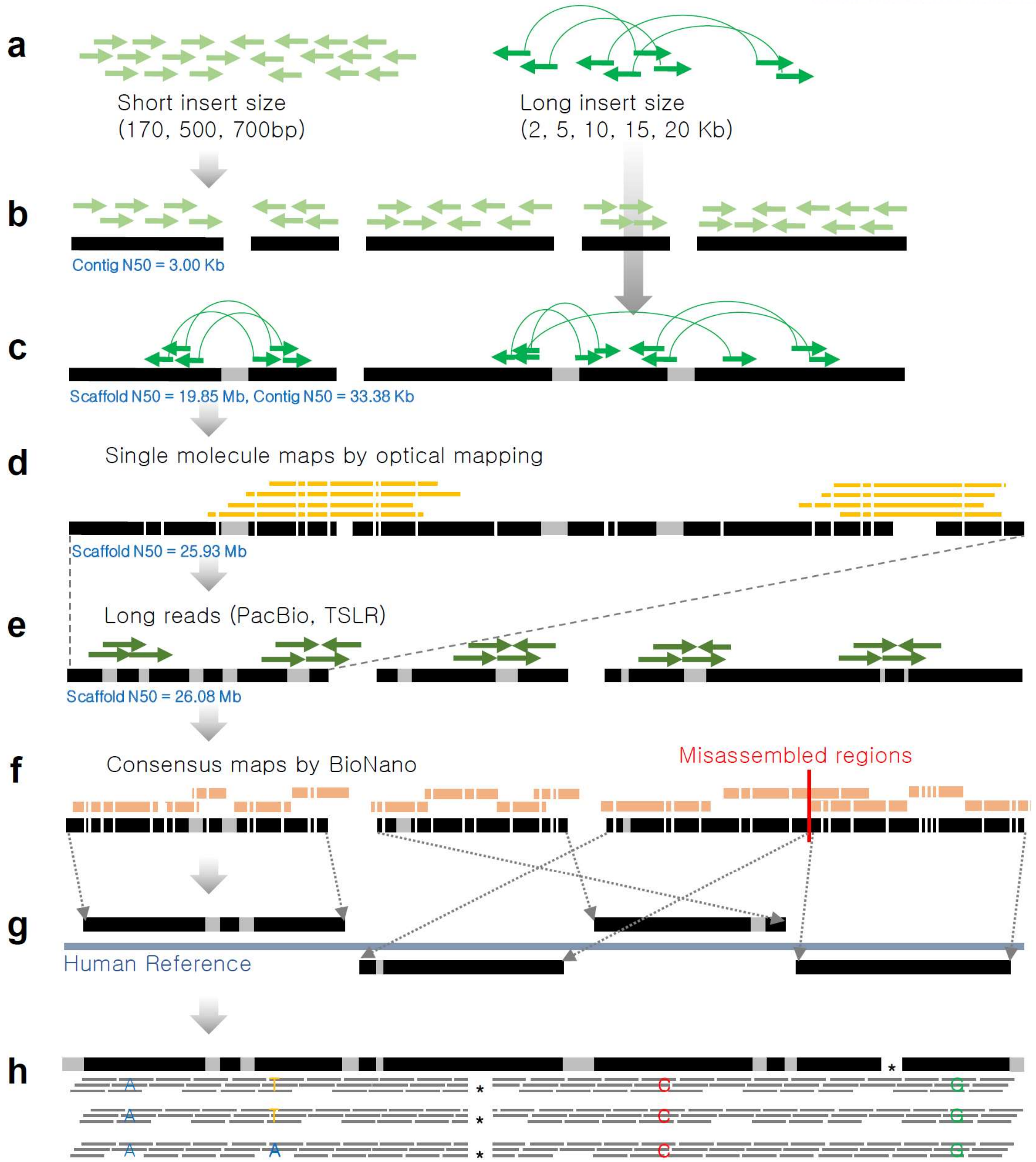


Figure 1. Schematic overview of KOREF assembly procedure. (**a**) Short and long insert size libraries by Illumina whole genome sequencing strategy. (**b**) Contig assembly using *K*-mers from short insert size libraries. (**c**) Scaffold assembly using long insert size libraries. (**d**) Super-scaffold assembly using OpGen whole genome mapping approach. (**e**) Gap closing using PacBio long reads and Illumina TruSeq synthetic long reads (TSLR). (**f**) Assembly assessment using BioNano consensus maps. (**g**) Chromosome sequence building using whole genome alignment information into the human reference (GRCh38). (**h**) Common variants substitution using 40 Korean whole genome sequences.

II. Methods

2.1 Sample preparation

All sample donors in this study signed written informed consent to participate, and the Institutional Review Board on Genome Research Foundation (IRB-201307-1 and IRB-201501-1 for KOREF, and 20101202-001 for KPGP) provided approval for this study. Genomic DNA and RNA used for genotyping, sequencing, and mapping data were extracted from the peripheral blood of sample donors. My colleagues and I conducted genotyping experiments with 16 Korean male participants using Infinium omni1 quad chip to check if the 16 donors had certain genetic biases. A total of 45 Korean whole genomes (40 for variant substitution and five for variant comparison) were used in this study (from the KPGP), sequenced using Illumina HiSeq2000/2500. For the comparison with the 16 donors, 34 Korean whole genome sequences from the KPGP and 86 Japanese, 84 Chinese, 112 Caucasians, and 113 Africans genotyping data from HAPMAP phase 3 were used. After filtering for MAF (< 5 %), genotyping rate (< 1 %), and LD ($R^2 \leq 0.2$) using PLINK[29], 90,462 and 72,578 shared nucleotide positions were used to calculate genetic distances for three ethnic groups (East-Asians, Caucasians, and Africans) and three East-Asian groups (Koreans, Chinese, and Japanese), respectively.

Epstein-Barr virus (EBV)-transformed B-cell line was constructed from the KOREF_S donor's blood[30], with minor modification. Briefly, peripheral blood mononuclear cells (PBMCs) were purified by Ficoll-Paque™ Plus (GE Healthcare, UK) density gradient centrifugation. For EBV infection, the cells were pre-incubated for 1 h with spent supernatant from the EBV producer cell line B95-8, and then cultured in RPMI-1640 containing 10-20% fetal bovine serum (FBS), 2 mM L-glutamine, 100 U per ml penicillin, 0.1 mg/ml streptomycin, 0.25 µg per ml amphotericin B (all from Gibco, Grand Island, NY, USA). The EBV-transformed B-cells were maintained at a concentration between $4\times10^5 - 1\times10^6$ cells per ml and expanded as needed.

2.2 Genome sequencing and scaffold assembly

For the *de novo* assembly of KOREF_S, 24 DNA libraries (three libraries for each insert size) with multiple insert sizes (170bp, 500bp, 700bp, 2 Kb, 5 Kb, 10 Kb, 15 Kb, and 20 Kb) were constructed according to the protocol of Illumina sample preparation. The libraries were sequenced using HiSeq2500 (three 20 Kb libraries) and HiSeq2000 (others) with a read length of 100bp. PCR duplicated, sequencing and junction adaptor contaminated, and low quality (<Q20) reads were filtered out, leaving only highly accurate reads to assemble the Korean genome. Additionally, short insert size and long insert size reads were trimmed into 90bp and 49bp, respectively, to remove poly-A tails and

low quality sequences in both ends. Error corrected read pairs by *K*-mer analysis from the short insert size libraries (<1 Kb) were assembled into distinct contigs based on the *K*-mer information using SOAPdenovo2[31]. Then, read pairs from all the libraries were used to concatenate the contigs into scaffolds step by step from short insert size to long insert size libraries using the *scaff* command of SOAPdenovo2 with default options except the –F option (filling gaps in scaffolds). To obtain scaffolds with the longest N50 length, I assembled the Korean genome (KOREF_S) with various *K*-mer values (29, 39, 49, 55, 59, 63, 69, 75, and 79) and finally selected an assembly derived from *K*=55, which has the longest contig N50 length. To reduce gaps in the scaffolds, I closed the gaps twice using the short insert size reads iteratively.

2.3 Super-scaffold assembly

I used whole-genome optical mapping data to generate a restriction map of the KOREF_S and assemble scaffolds into super-scaffolds[18]. First, 13 restriction enzymes were evaluated for compatibility with the Korean genome draft assembly, and *SpeI* enzyme was deemed suitable for the Korean genome analysis. High molecular weight DNA was extracted, and 4,217,937 single molecule restriction maps (62,954 molecules on each map card on overage) were generated from 67 high density MapCards. Among them, 2,071,951 molecules exceeding 250 Kb with ~360 Kb of average size were collected for the genome assembly. The Genome Builder bioinformatics tool of OpGen[18] was used to compare the optical mapping data to the scaffolds. The distance between restriction enzyme sites in the scaffolds were matched to the lengths of the optical fragments in the optical maps, and matched regions were linked into super-scaffolds. Only scaffolds exceeding 200 Kb were used in this step.

Additionally, I generated two types of long reads for KOREF_S building: PacBio long reads and TSLRs. The PacBio long reads were generated using a Pacific Biosciences RSII instrument (P4C2 chemistry, 78 SMRT cells; P5C3 chemistry, 51 SMRT cells), and the TSLRs were sequenced by Illumina HiSeq2500. Both long reads were simultaneously used in additional scaffolding and gap closing processes using PBJelly2 program[32] with default options.

2.4 Assembly assessment and chromosome building

For a large-scale assessment of the scaffolds, I generated nanochannel-based genome mapping data (~145 Gb of single-molecule maps exceeding 150 Kb) on five irysChips and assembled the mapping data into 2.8 Gb of consensus genome maps using BioNano Genomics Irys genome mapping system.

The consensus genome maps were compared to KOREF_S scaffolds and GRCh38 using irysView software package[21] (version 2.2.1.8025). To identify misassembles in KOREF_S scaffolds in detail, I manually checked alignment results of the consensus genome map into KOREF_S scaffolds and human reference. For a smaller resolution assessment, I aligned all the filtered short and long reads into the scaffolds using BWA-MEM[33] (version 0.7.8) with default options. My colleagues and I conducted a whole genome alignment between KOREF_S scaffolds (≥ 10 Kb) and human reference (soft repeat masked) using SyMap[34] with default comparison parameters (mapped anchor number ≥ 7) to detect possible inter- or intra-chromosomal rearrangements. My colleagues and I manually checked all the whole genome alignment results.

To build the chromosome sequence of KOREF_S, first I used the whole genome alignment information (chromosomal location and ordering information) of the final scaffolds (≥ 10 Kb) onto GRCh38 chromosomes. Then, unmapped scaffolds were re-aligned to GRCh38 chromosome with a mapped anchor number ≥ 4 option. Small length scaffolds (from 200bp to 10 Kb) were aligned to GRCh38 chromosomes using BLASR[35], and only alignments with mapping quality = 254 were used. Unused scaffolds (a total 88.3 Mb sequences) for this chromosome building process were located in an unplaced chromosome (chrUn). Gaps between the aligned scaffolds were estimated based on the length information of the human reference sequences. If some scaffold locations overlapped, 10 Kb was used as the size of gap between the scaffolds. I added 10 Kb gaps on both sides of KOREF_S chromosome sequences as telomeric regions just as done for GRCh38. The mitochondrial sequences of KOREF_S were independently sequenced using Nextera XT sample prep kit and then assembled using ABySS[36] (version 1.5.1) with *K*=64. Haplogroup of the mitochondrial DNA was assigned using MitoTool[37].

The 40 Korean whole genome sequences from KPGP database were aligned onto KOREF_S chromosomes using BWA-MEM with default options, in order to remove individual specific sequence biases of KOREF_S and generate KOREF_C. SNVs and small indels in the 40 Koreans were called using the Genome Analysis Toolkit (GATK, version 2.3.9)[38]. IndelRealigner was conducted to enhance mapping quality, and base quality scores were recalibrated using the TableRecalibration algorithm of GATK. Commonly found variants in the 40 Korean genomes were used to substitute KOREF_S sequences. For the SNV substitution, I calculated allele ratio of each position, and then I substituted any KOREF_S sequence with the most frequent allele only if the KOREF_S sequence and most frequent allele were different. For the indel substitution, I used only indels that were found in over 40 haploids out of the 40 Korean whole genomes (80 haploids). In cases of sex chromosomes, I used 25 male (25 haploids) whole genomes for Y chromosome and 15 female whole genomes (30 haploids) for X chromosome comparison.

2.5 Genome annotation

KOREF_C was annotated for repetitive elements and protein coding genes. For the repetitive elements annotation, my colleagues and I searched KOREF_C for tandem repeats and transposable elements using Tandem Repeats Finder (version 4.07)[39], Repbase (version 19.02)[40], RepeatMasker (version 4.0.5)[41], and RepeatModeler (version 1.0.7)[42]. For the protein coding gene prediction, homology-based gene prediction was first conducted by searching nucleotides of protein coding genes in Ensembl database 79 against KOREF_C using Megablast[43] with identity 95 criterion. The matched sequences were clustered based on their positions in KOREF_C, and a gene model was predicted using Exonerate software[44] (version 2.2.0). I also conducted *de novo* gene prediction. To certify expression of a predicted gene, I sequenced three different timeline whole transcriptome data of the KOREF_S sample using a TruSeq RNA sample preparation kit (v2) and HiSeq2500. I predicted protein coding genes with the integrated transcriptome data using AUGUSTUS[45] (version 3.0.3). I filtered out genes shorter than 50 amino acids and possible pseudogenes having stop-codons. I searched *de novo* predicted genes against primate (human, bonobo, chimpanzee, gorilla, and orangutan) protein sequences from NCBI, and filtered out *de novo* predicted genes if identity and coverage were below 50 %. For the assembly quality comparison purpose, I only used homology-based search for RefSeq[46] human protein-coding genes and repetitive elements. The homology-based segmental duplicated region search was conducted using DupMasker program[47]. To calculate GRCh38 genome recovery rates of human assemblies, my colleagues and I conducted whole genome alignments between each assembly (KOREF_S final contigs, KOREF_S final scaffolds, and other assemblies) and GRCh38 using LASTZ[48] (version 1.03.54) and Kent utilities (written by Jim Kent at UCSC)[49] with GRCh38 self-alignment options (--step 19 --hspthresh 3000 --gappedthresh 3000 --seed=12of19 --minScore 3000 --linearGap medium). After generating a MAF file, my colleagues and I calculated genome recovery rates using mafPairCoverage in mafTools[50].

To estimate the amount of novel KOREF_C sequences, I aligned the short insert size and long mate pair library sequences into GRCh38 using BWA-MEM with default options and then extracted unmapped reads using SAMtools[51] (version 0.1.19) and Picard (version 1.114, http://picard.sourceforge.net) programs. I filtered out possible microbial contamination by searching against Ensembl databases of bacterial genomes and fungal genomes using BLAST with default options. The remaining reads were sequentially aligned into other human genome assemblies (CHM1_1.1, HuRef, African, Mongolian, and YH sequentially) using BWA-MEM with default options, and then removed duplicated reads using MarkDuplicate program in Picard. The alignment results were extracted to an unmapped BAM file using SAMtools view command with -u -f 4 options. I extracted final unmapped reads from the unmapped BAM file using SamToFastq program in Picard. Finally, unmapped reads to the other human genome assemblies were aligned to KOREF_C. The regions with length ≥100bp and covered by at least three unmapped reads were considered as novel in KOREF_C.

2.6 Variant and genome comparison

A total of 15 whole genome re-sequencing data results (five Caucasians, five Africans, and five East-Asians) were downloaded from the 1KGP, HGDP, and PAPGI projects. The re-sequencing data (five Caucasians, five Africans, five East-Asians, and five Koreans from KPGP) was filtered (low quality with a Q20 criterion and PCR duplicated reads) and then mapped to KOREFs (KOREF_S and KOREF_C) with unplaced scaffolds, GRCh38, and GRCh38_C chromosomes using BWA-MEM with default options. To generate GRCh38_C, common variants (2,043,259 SNVs and 197,885 small indels) of East-Asians were collected from the 1KGP database and used to substitute GRCh38 sequences. The variants (SNVs and small indels) were called for only chromosome sequences using GATK, in order to exclude variants in unmatched and partially assembled repetitive regions[14]. Variants were annotated using SnpEff[52], and biological function altering was predicted using PROVEAN[53]. I considered all of the nsSNVs causing stop codon changes and frame shift indels as function altered. Enrichment tests and annotation of variants were conducted using WebGestalt[54] and ClinVar[55]. The variants were compared with dbSNP[56] (version 144) to annotate known variants information.

For linking variants found compared to KOREFs, GRCh38, and GRCh38_C, the genome to genome alignment was conducted between GRCh38 and KOREF_C reference genomes using LASTZ[48]. The LASTZ scoring matrix used was with M=254 (--masking=254), K=4500 (--hspthresh=4500), L=3000 (--gappedthresh=3000), Y=15000 (--ydrop=15000), H=0 (--inner=9), E=150 / O=600 (--gap=<600,150>), and T=2 options. The LASTZ output was translated to the chain format with axtChain, then merged and sorted by the chainMerge and chainSort programs, respectively. The alignable regions were identified with chainNet, and then selected by netChainSubSet programs for creating a lift-over file. All programs run after LASTZ were written by Jim Kent at UCSC[49].

To detect SVs among the human genome assemblies, my colleagues and I conducted whole genome alignments between each assembly and GRCh38 using LASTZ. Then, the whole genome alignment results were corrected and re-aligned based on a dynamic-programming algorithm using SOAPsv package. SVs that could be derived from possible misassembles were filtered out by comparing the S/P ratio for each structural variation region in the assembly and GRCh38; authentic SVs would be covered by sufficient paired-end reads, whereas spurious SVs would be covered by wrongly mapped single-end reads. My colleagues and I implemented the S/P ratio filtering system according to the previous published algorithm[57], because the S/P ratio filtering step in the SOAPsv package is designed for only assembled sequences by SOAPdenovo. *P*-value was calculated by performing Fisher's exact test to test whether the S/P ratio of each SV and the S/P ratio of the whole genome are significantly different (*P*-value < 0.001). I confirmed that commonly shared SVs were not caused by the mis-assembly by checking the mapping status of KOREF_S short and long reads into both GRCh38 and KOREF_C. SVs by mapping CHM1's PacBio SMRT reads to the human reference genome were derived by lift-over SV results found against GRCh37 in the published paper[15]. When I compared SVs in the different genome assemblies and available database, I considered SVs to be the same if SVs were reciprocally 50 % covered and had the same SV type. Novel SVs were determined as not found in dbVar, Database of Genomic Variants (DGV)[58], Database of Retrotransposon Insertion Polymorphisms (dbRIP)[59], dbSNP146, Mills[60], and 1000 Genome phase 3 database.

III. Results & Discussion

3.1 Choosing a representative genome donor

My colleagues and I recruited 16 Korean volunteers, who signed an informed consent (based on the PGP protocol, with minor country-specific adaptations) for use of their genomic data and agreed to their public release (Table 1). After extracting DNA from peripheral blood (Table 2), we genotyped each volunteer using an Infinium omni1 quad chip.

Table 1. 16 Korean male volunteers in KOREF construction

ID	Age	Sex
KR01	**47**	**male**
KR02	27	male
KR03	30	male
KR04	31	male
KR05	30	male
KR06	50	male
KR07	48	male
KR08	56	male
KR09	34	male
KR10	31	male
KR11	29	male
KR12	29	male
KR13	27	male
KR14	39	male
KR15	31	male
KR16	35	male

Table 2. Quality control results of the 16 blood sample donors in KOREF construction

ID	Conc_ Quant-iT (ng/ul)	Vol. (ul)	Fluorescence amount (ug)	Conc_UV (ng/ul)	260/ 280	260/ 230	UV amount
KR-01	127	45	5.72	195.9	1.78	2.07	8.82
KR-02	137	48	6.58	208.1	1.79	2.19	9.99
KR-03	159	49	7.79	234.0	1.78	2.02	11.47
KR-04	376	43	16.17	467.1	1.81	2.14	20.09
KR-05	200	49	9.80	286.3	1.81	2.18	14.03
KR-06	270	41	11.07	525.7	1.82	2.05	21.55
KR-07	328	40	13.12	579.9	1.82	2.00	23.20
KR-08	131	41	5.37	183.5	1.81	2.17	7.52
KR-09	101	42	4.24	172.5	1.80	2.13	7.25
KR-10	125	43	5.38	192.8	1.80	2.18	8.29
KR-11	103	43	4.43	156.8	1.81	2.12	6.74
KR-12	129	43	5.55	177.8	1.81	2.12	7.65
KR-13	98.9	52	5.14	152.4	1.81	1.82	7.92
KR-14	164	43	7.05	238.7	1.82	2.17	10.26
KR-15	186	43	8.00	275.1	1.80	2.14	11.83
KR-16	147	980	144.06	228.8	1.79	2.10	224.22

Multidimensional scaling (MDS) plots of pairwise genetic distances were constructed, using an additional 34 Korean whole genome sequences from the KPGP database, as well as 86 Japanese, 84 Chinese, 112 Caucasians, and 113 Africans genotype data from HAPMAP phase 3[61] (Fig. 2). All the 16 Korean samples fell into a tight population cluster, indicating they represent one ethnic group. A healthy male donor was chosen as KOREF_S by considering a list of parameters such as centrality of the genetic distance, the participant's age, parental sample availability, the availability for continuous blood sample donation, and normality of the G-banded karyotype (Fig. 3). To supply reference material, an immortalized cell line was constructed from the KOREF_S donor's blood and deposited in the Korean Cell Line Bank (KCLB, #60211).

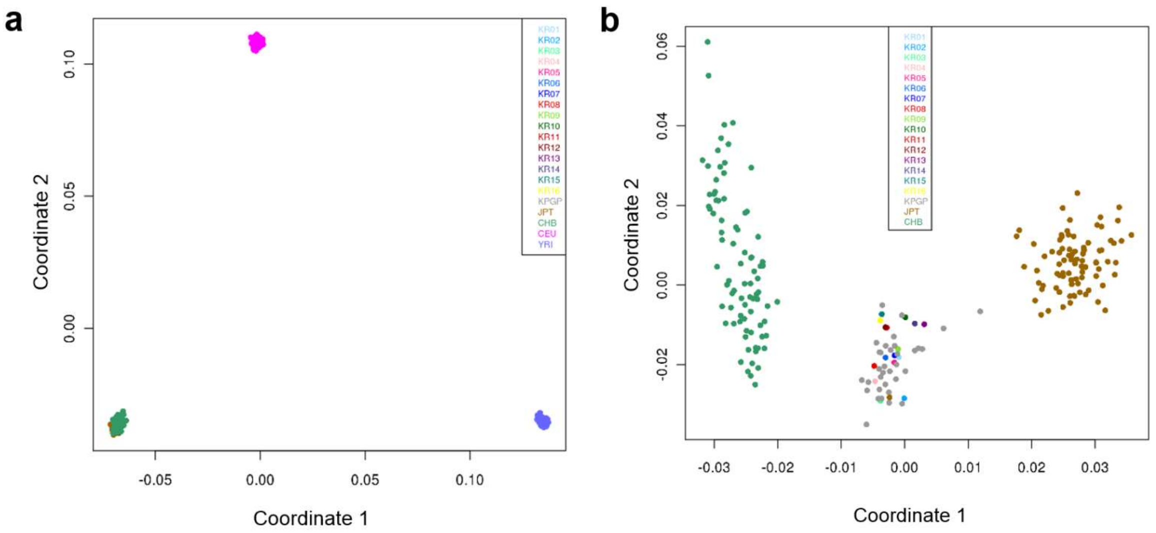


Figure 2. MDS plot of 445 human samples. (**a**) MDS plots of the 16 donors (KR) were drawn by comparing to other 34 Koreans (KPGP), 86 Japanese (JPT), 84 Chinese (CHB), 112 Caucasians (CEU), and 113 Africans (YRI) using 90,462 SNV markers. (**b**) MDS plot among Koreans, Chinese, and Japanese using 72,578 SNV markers. The span of the genetic distance of the 16 did not fall outside the common Korean population range.

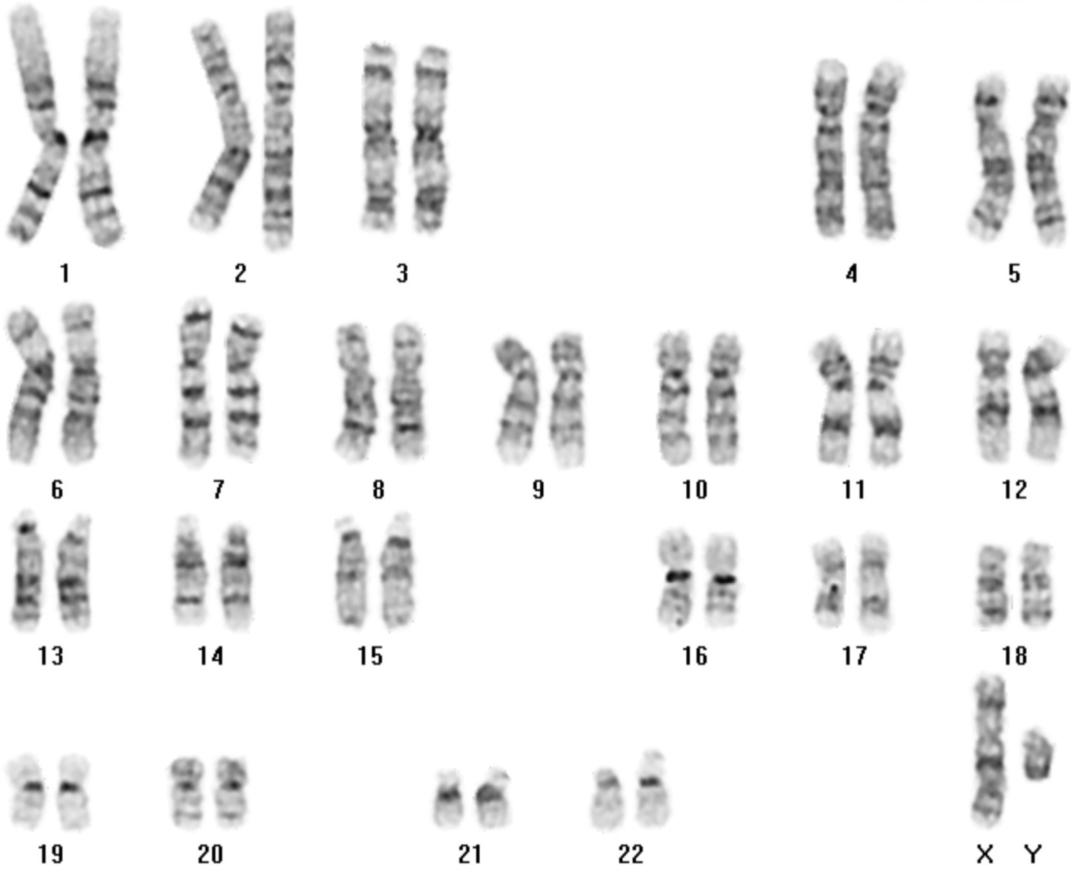


Figure 3. G-banded karyotype of the Korean genome. There were no abnormalities in the chromosomes (2n=46).

3.2 KOREF_S assembly

I obtained short-read sequencing data from the Illumina HiSeq2000 and HiSeq2500 platforms, using the same approach adopted by other draft reference genome projects[9-11,13,31]. A total 964 Gb of paired-end DNA reads were generated from 24 libraries with different fragment sizes (170bp, 500bp, and 700bp of short insert size, and 2 Kb, 5 Kb, 10 Kb, 15 Kb, and 20 Kb of long insert size), giving a total sequencing depth coverage of ~311 fold (Tables 3 and 4). From a *K*-mer analysis, the size of KOREF_S was estimated to be ~3.03 Gb (Table 5 and Fig. 4). Error corrected reads by *K*-mer analysis from the short insert size libraries (<1 Kb) were assembled into distinct contigs based on the *K*-mer information (Table 6). As the target fragment sizes can be biased by library construction process, I estimate the real fragment sizes of all the libraries by mapping the DNA reads onto the contigs (Fig. 5). Then, read pairs from all the libraries were used to concatenate the contigs into scaffolds step by step from short insert size to long insert size libraries. A total of 68,170 scaffolds (≥ 200bp) were generated, totaling 2.92 Gb in length reaching an N50 length of almost 20 Mb (19.85 Mb) and containing only 1.65 % gaps (Table 7 and Fig. 6). Approximately, 90 % of the genome draft (N90) was covered by 178 scaffolds, each larger than 3.09 Mb, with the largest spanning over 80 Mb (81.9) on Chromosome 6.

Table 3. Statistics regarding Illumina whole-genome shotgun sequence

Type	Insert size	Read length (bp)	Number of read pairs	Total data (Gb)	Sequence depth (×)	
Short-insert size libraries	170bp	101	254,562,947	51.42	16.59	48.69
			246,624,330	49.82	16.07	
			246,007,078	49.70	16.03	
	500bp	101	246,418,836	49.78	16.06	46.71
			230,109,465	46.48	14.99	
			240,361,539	48.55	15.66	
	700bp	101	207,193,678	41.85	13.50	39.17
			188,159,956	38.01	12.26	
			205,873,335	41.59	13.41	
Long-mate pair libraries	2Kb	101	196,290,337	39.65	12.79	38.22
			232,858,099	47.04	15.17	
			157,507,662	31.82	10.26	
	5Kb	101	152,201,289	30.74	9.92	32.81
			177,874,430	35.93	11.59	
			173,383,733	35.02	11.30	
	10Kb	101	205,215,277	41.45	13.37	40.05
			209,859,354	42.39	13.67	
			199,617,521	40.32	13.01	
	15Kb	101	156,336,183	31.58	10.19	30.65
			166,036,249	33.54	10.82	
			147,927,209	29.88	9.64	
	20Kb	101	181,506,276	36.66	11.83	34.72
			177,434,679	35.84	11.56	
			173,929,946	35.13	11.33	
Total			4,773,289,408	964.19	311.02	311.02

Table 4. Statistics regarding filtered and trimmed whole-genome shotgun sequence

Type	Insert size	Read length (bp)	Number of read pairs	Total data (Gb)	Sequence Depth (×)	
Short-insert size libraries	170bp	90	238,901,578	43.00	13.87	40
			225,934,916	40.67	13.12	
			224,145,725	40.35	13.01	
	500bp	90	220,100,704	39.62	12.78	37.57
			207,716,033	37.39	12.06	
			219,165,329	39.45	12.73	
	700bp	90	189,043,000	34.03	10.98	32.24
			173,545,699	31.24	10.08	
			192,535,557	34.66	11.18	
Long-mate pair libraries	2Kb	49	102,368,796	10.03	3.24	9.64
			118,485,351	11.61	3.75	
			83,704,400	8.20	2.65	
	5Kb	49	74,199,538	7.27	2.35	8.08
			93,060,115	9.12	2.94	
			88,156,446	8.64	2.79	
	10Kb	49	52,521,514	5.15	1.66	5.03
			54,759,429	5.37	1.73	
			51,874,811	5.08	1.64	
	15Kb	49	60,904,413	5.97	1.93	5.3
			55,631,632	5.45	1.76	
			51,042,581	5.00	1.61	
	20Kb	49	20,374,949	2.00	0.64	2.08
			26,561,512	2.60	0.84	
			19,032,195	1.87	0.60	
Total			2,843,766,223	433.77	139.94	139.94

Table 5. Statistics regarding 23-mer analysis results

K-mer size	*K*-mer total number	Peak depth	Genome size (bp)	Used base (bp)	Used reads number	Depth coverage (×)	Average read length (bp)	*K*-mer species number
23	87,989,560,976	29	3,034,122,792	116,456,771,880	1,293,964,132	38.3824	90	5,689,732,938

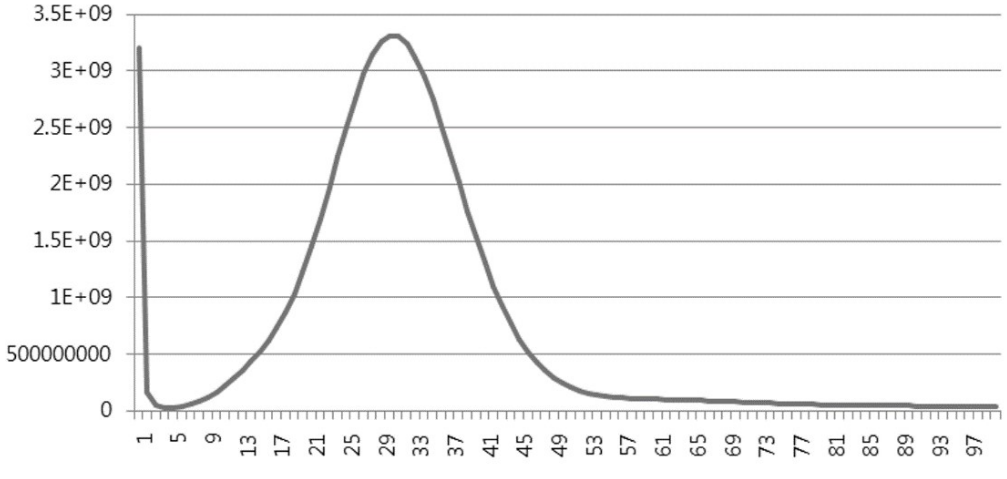


Figure 4. *K*-mer (*K*=23) analysis. The x-axis represents the depth coverage of each unique 23-mer in the Korean genome, and the y-axis represents the proportion of frequency at that depth divided by the total frequency at all depths.

Table 6. Contig assembly results based on various *K*-mer information

K-mer size	All sequences			Longer than 100bp		
	Total size	Longest	N50	Total size	Longest	N50
29	5,187,304,717	16,946	90	2,275,359,750	16,946	1,099
39	4,459,796,947	35,726	300	2,529,816,579	35,726	1,939
49	4,066,593,737	51,838	980	2,740,134,913	51,838	2,375
55	**3,860,731,497**	**44,789**	**1,447**	**2,915,054,629**	**44,789**	**2,559**
59	3,744,446,380	48,982	1,773	2,990,197,206	48,982	2,735
63	3,641,677,654	54,683	2,113	3,029,961,853	54,683	2,964
69	3,524,281,519	54,689	2,589	3,072,247,309	54,689	3,295
75	3,429,622,648	62,488	2,918	3,097,380,667	62,488	3,466
79	3,343,414,611	80,399	2,789	3,086,359,621	80,399	3,187

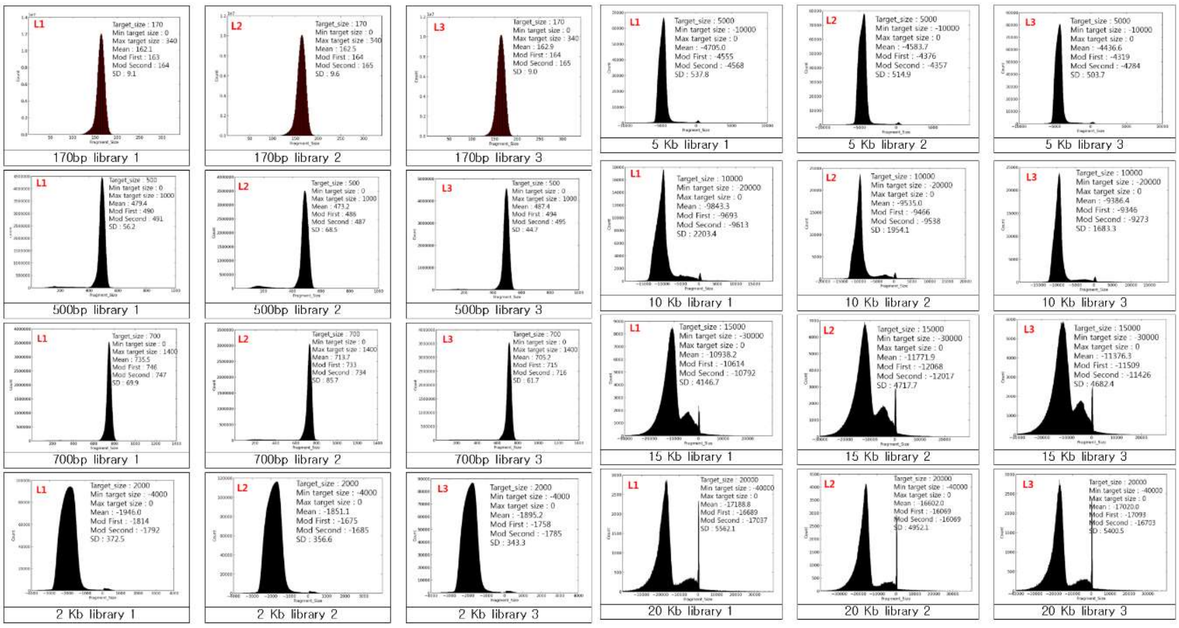


Figure 5. The real fragment size estimation for all the short and long insert size libraries

Table 7. KOREF build statistics along the assembly steps

	Contig		Scaffold		Whole-genome optical mapping		Long reads (PacBio and TSLR)		Chromosomes (Assessment using BioNano maps) *Unplaced scaffolds were excluded.	
	Size (Kb)	No.	Size (Mb)	No.	Size (Mb)	No.	Size (Mb)	No.	Size (Mb)	No.
N90	8.59	89,240	3.09	178	3.86	140	3.53	143	81.54	19
N80	14.62	63,987	6.45	116	9.45	92	9.26	93	103.05	16
N70	20.42	47,417	10.45	81	14.47	67	14.53	67	136.43	13
N60	26.58	35,099	16.16	59	19.56	49	19.36	50	137.59	11
N50	33.38	25,446	19.85	42	25.93	36	26.08	36	155.88	8
Longest	334.16	-	81.91	-	101.22	-	101.48	-	251.92	-
Gaps	0 %	-	1.65 %	-	1.75 %	-	1.06 %	-	9.44 %	-
Total (≥ 200bp)	2.87 Gb	230,514	2.92 Gb	68,170	2.92 Gb	68,103	2.94 Gb	68,451	3.12 Gb	24
Total (≥10 Kb)	2.52 Gb	82,254	2.88 Gb	1,243	2.88 Gb	1,176	2.90 Gb	1,369	3.12 Gb	24

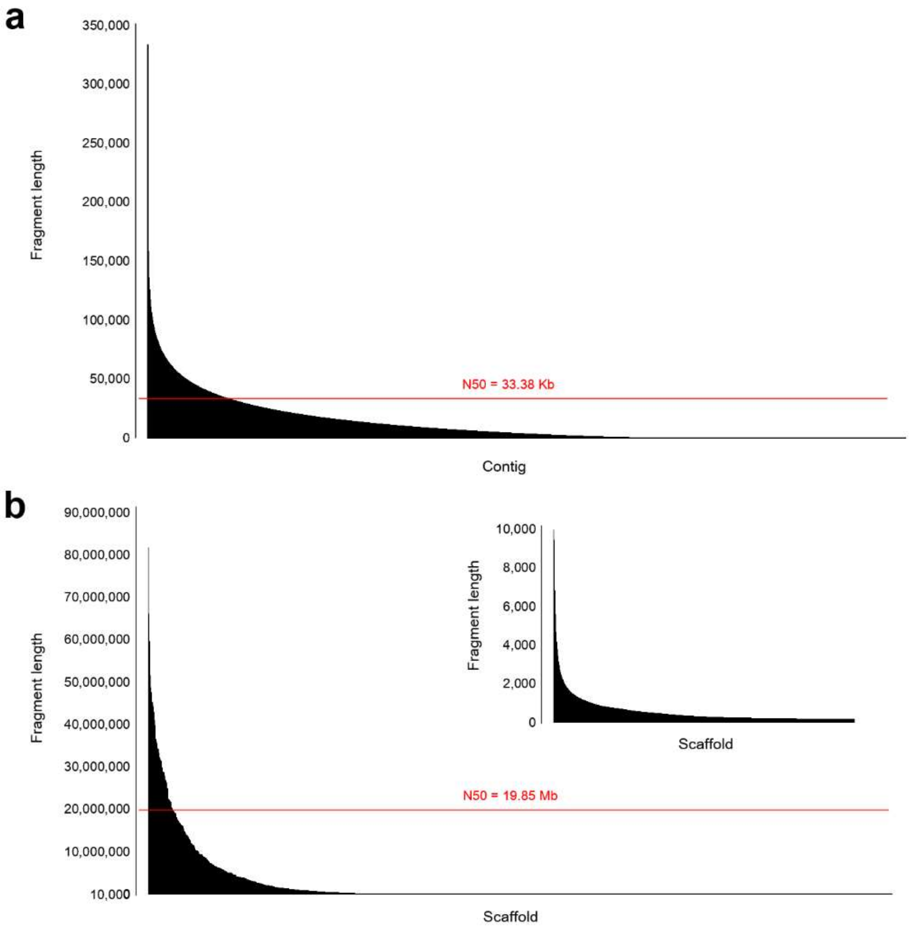


Figure 6. Length distribution of KOREF_S assembled fragments. (**a**) KOREF_S contig using only NGS short read data. (**b**) KOREF_S scaffold using only NGS short read data. Fragments (contigs/scaffolds) were sorted by their lengths.

In order to further extend the scaffolds, I used a high-throughput whole-genome optical mapping instrument, as previously suggested[18]. I extracted high molecular weight DNA and generated 745.5 Gb of single-molecule restriction maps (about two million molecules with 360 Kb of average size) from 67 high density MapCards, resulting in 240-fold optical map coverage (Tables 8 and 9).

Table 8. *In silico* restriction enzyme selection on the KOREF_S scaffolds. *SpeI* enzyme was used for the KOREF_S whole genome optical map building.

Enzyme	Usable% 5Kb-20Kb	Usable% 6Kb-15Kb	Usable% 6Kb-12Kb	Ave. Frags size (kb)	# of Frags > 100kb	Max Frag size (Kb)
AflII	25.12	10.31	10.07	4.58	4	117.49
BamHI	94.94	82.36	72.76	8.08	19	159.82
KpnI	98.76	91.89	69.64	10.35	50	154.09
NcoI	17.1	3.37	3.35	3.85	0	84.46
NheI	98.08	89.26	65.1	10.67	62	149.61
SpeI	94.8	73.17	67.9	7.44	63	196.12
BglII	7.01	2.12	2.07	3.79	1	104.69
EcoRI	7.86	2.87	2.85	3.65	0	71.37
MluI	0.76	0.23	0.09	130.62	9422	1529.97
NdeI	12.35	6.4	6.21	3.25	3	105.73
PvuII	2.2	0.4	0.4	2.7	3	149.7
XbaI	9.27	3.33	3.26	3.64	3	147.38
XhoI	26.46	11.1	4.88	23.64	2612	372.38

Table 9. OpGen single molecule restriction map (SMRM) statistics

Summary of SMRM data	Maps used in analysis
Total Size (Gb)	745.51
Number of Molecules	2,071,951
Average Size of Molecules (Kb)	359.81
Minimum molecule size (Kb)	250
Average Size of Fragments (Kb)	13.24

To join the scaffolds, the single-molecule optical maps were compared to the assembled scaffolds that were converted into restriction maps by *in silico* restriction enzyme digestion. As a result, a total of 67 scaffolds (>200 Kb) were joined (Table 10). This resulted in the increase of scaffold N50 length from 19.85 Mb to 25.93 Mb (Table 7).

Table 10. Scaffold joining results using optical mapping data

Scaffold1	size1(kb)	strand1	Scaffold2	size2 (kb)	strand2	Gap (kb)	Score
SCAFFOLD317	1022.416	1	SCAFFOLD743	842.84	-1	14.466	99.4236
SCAFFOLD210	11746.639	1	SCAFFOLD940	551.059	-1	12.506	97.8962
SCAFFOLD244	882.071	1	SCAFFOLD150	8643.747	1	-4.539	92.9372
SCAFFOLD532	495.294	1	SCAFFOLD280	1697.743	-1	16.755	87.2892
SCAFFOLD103	8759.181	1	SCAFFOLD431	2527.758	1	4.325	80.7857
SCAFFOLD8	18209.097	1	SCAFFOLD122	5972.151	1	17.543	69.7056
SCAFFOLD79	778.308	1	SCAFFOLD292	913.969	1	1.636	66.4837
SCAFFOLD77	4752.716	1	SCAFFOLD89	4287.167	1	0.067	64.7672
SCAFFOLD89	4287.167	1	SCAFFOLD140	10524.263	1	-5.698	62.3363
SCAFFOLD63	8355.854	-1	SCAFFOLD163	6250.598	1	14.348	55.3254
SCAFFOLD356	1363.545	-1	SCAFFOLD743	842.84	1	71.197	55.2638
SCAFFOLD70	19845.87	1	SCAFFOLD42	6341.468	1	202.32	54.2056
SCAFFOLD110	6289.28	1	SCAFFOLD170	3210.067	1	2.994	53.1726
SCAFFOLD19	29018.184	1	SCAFFOLD364	2266.538	1	39.026	47.5055
SCAFFOLD485	689.059	1	SCAFFOLD343	2303.617	1	57.217	43.8511
SCAFFOLD428	511.544	1	SCAFFOLD31	2851.399	-1	116.431	43.2197
SCAFFOLD126	5708.801	1	SCAFFOLD219	1429.49	-1	85.562	43.2175
SCAFFOLD353	2639.995	1	SCAFFOLD15	2258.516	1	10.722	38.5231
SCAFFOLD91	5409.31	1	SCAFFOLD63	8355.854	-1	190.878	38.2565
SCAFFOLD169	5101.962	1	SCAFFOLD653	227.433	1	12.551	32.557
SCAFFOLD87	12817.817	-1	SCAFFOLD212	3045.171	1	16.396	29.8232
SCAFFOLD264	14081.586	1	SCAFFOLD575	626.29	1	25.872	28.7976
SCAFFOLD24	15566.053	1	SCAFFOLD3	13712.728	1	-0.342	28.4213
SCAFFOLD502	381.379	-1	SCAFFOLD533	1080.224	1	0.859	27.1306
SCAFFOLD1072	619.532	1	SCAFFOLD189	12056.91	-1	51.438	26.8774
SCAFFOLD246	13977.981	-1	SCAFFOLD206	20601.118	1	5.588	24.7277
SCAFFOLD322	4940.238	1	SCAFFOLD201	6752.265	1	2.859	23.4562
SCAFFOLD337	286.159	1	SCAFFOLD787	520.497	1	25.873	22.9017
SCAFFOLD103	8759.181	-1	SCAFFOLD11	5130.215	1	0.002	22.6392
SCAFFOLD85	5575.593	1	SCAFFOLD302	1599.441	-1	-5.59	21.6902
SCAFFOLD82	5897.044	1	SCAFFOLD43	28037.362	1	-0.311	21.4608
SCAFFOLD533	1080.224	1	SCAFFOLD27	4154.534	-1	5.276	21.2813
SCAFFOLD246	13977.981	1	SCAFFOLD112	34485.537	-1	-3.432	19.0796
SCAFFOLD392	875.318	1	SCAFFOLD289	1425.336	-1	6.962	18.2247
SCAFFOLD142	7148.482	1	SCAFFOLD59	5549.968	1	-0.24	18.0723
SCAFFOLD7	40570.24	-1	SCAFFOLD199	16436.955	1	10.033	17.6323
SCAFFOLD233	3346.963	1	SCAFFOLD147	30048.452	-1	3.123	17.3518
SCAFFOLD377	1560.501	1	SCAFFOLD233	3346.963	1	7.023	16.3624
SCAFFOLD455	3872.703	1	SCAFFOLD85	5575.593	1	-3.332	16.165
SCAFFOLD872	333.932	1	SCAFFOLD243	2305.143	1	82.932	16.098
SCAFFOLD350	999.02	-1	SCAFFOLD142	7148.482	1	236.727	16.0549
SCAFFOLD197	9499.216	1	SCAFFOLD12	2823.635	1	-6.936	15.8702
SCAFFOLD569	387.885	-1	SCAFFOLD119	1305.15	1	5.536	15.3893
SCAFFOLD434	1008.885	1	SCAFFOLD423	472.166	-1	16.713	15.3473
SCAFFOLD153	18967.221	1	SCAFFOLD353	2639.995	1	29.897	14.2316
SCAFFOLD161	943.876	1	SCAFFOLD87	12817.817	-1	147.087	14.2259
SCAFFOLD98	48842.997	1	SCAFFOLD235	10164.153	1	6.502	13.9087
SCAFFOLD232	242.678	1	SCAFFOLD218	444.904	1	0.834	13.8088
SCAFFOLD296	792.382	1	SCAFFOLD35	1500.96	1	37.211	13.7568
SCAFFOLD54	14806.717	1	SCAFFOLD214	5135	1	4.606	13.3133
SCAFFOLD502	381.379	1	SCAFFOLD222	4068.33	-1	11.1	12.7174
SCAFFOLD100	6592.548	1	SCAFFOLD359	2048.679	-1	27.867	12.3654
SCAFFOLD49	36078.134	-1	SCAFFOLD100	6592.548	1	0.002	12.3407
SCAFFOLD243	2305.143	1	SCAFFOLD940	551.059	1	8.69	12.3289
SCAFFOLD146	6416.747	1	SCAFFOLD40	20409.372	1	4.306	11.054
SCAFFOLD350	999.02	1	SCAFFOLD570	524.193	-1	51.022	10.734
SCAFFOLD39	8825.901	-1	SCAFFOLD104	7398.895	1	-11.052	10.6812
SCAFFOLD306	1232.982	1	SCAFFOLD99	3038.256	-1	-6.299	10.29
SCAFFOLD42	6341.468	1	SCAFFOLD263	2671.43	1	52.694	10.0097
SCAFFOLD638	678.726	1	SCAFFOLD79	778.308	1	116.653	9.9301
SCAFFOLD86	16308.764	1	SCAFFOLD16	19543.299	-1	-1.287	9.8459
SCAFFOLD170	3210.067	1	SCAFFOLD306	1232.982	1	254.75	9.683
SCAFFOLD120	19315.79	1	SCAFFOLD38	81906.269	1	5.027	9.6118
SCAFFOLD649	661.586	1	SCAFFOLD570	524.193	1	576.918	9.3685
SCAFFOLD392	875.318	-1	SCAFFOLD169	5101.962	1	19.408	9.1531
SCAFFOLD178	423.463	1	SCAFFOLD28	12121.666	-1	67.343	9.12
SCAFFOLD364	2266.538	1	SCAFFOLD74	3948.894	1	4.863	9.0136

Additionally, I generated two types of long reads for KOREF_S: PacBio SMRT (~31.1 Gb, ~10-fold coverage; Fig. 7 and Table 11) and Illumina TruSeq Synthetic Long Reads (TSLR, ~16.3 Gb, ~5.3-fold coverage; Fig. 8 and Table 12).

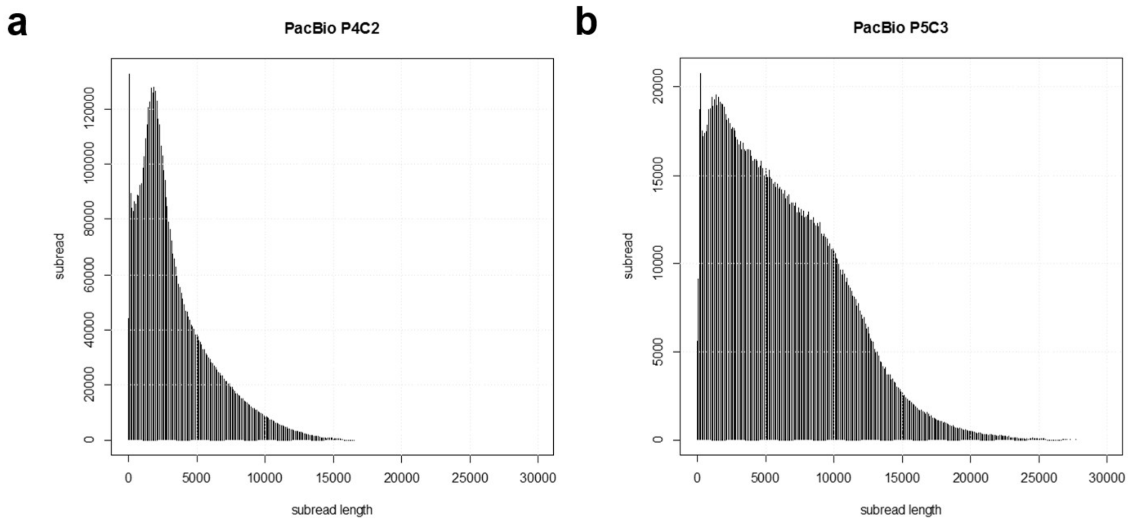


Figure 7. Length distribution of PacBio RSII DNA sequence reads. (**a**) PacBio RSII P4C2. (**b**) PacBio RSII P5C3.

Table 11. PacBio RSII long reads statistics

a. PacBio P4C2

Size	Number of bases (bp)	Number of reads	Mean length (bp)
~2kb	2,200,375,125	2,023,326	1,088
~3kb	2,598,138,881	1,054,927	2,463
~4kb	2,253,729,183	650,819	3,463
~5kb	1,993,913,569	445,503	4,476
~6kb	1,868,335,867	341,037	5,478
~7kb	1,692,679,373	261,244	6,479
~8kb	1,490,151,540	199,293	7,477
~9kb	1,264,147,938	149,166	8,475
~10kb	1,025,254,470	108,261	9,470
10kb~	2,404,653,532	202,921	11,850
Total	18,791,379,478	5,436,497	3,457

b. PacBio P5C3

Size	Number of bases (bp)	Number of reads	Mean length (bp)
~2kb	376,691,922	352,650	1,068
~3kb	448,189,058	179,744	2,493
~4kb	581,090,138	166,158	3,497
~5kb	707,030,086	157,272	4,496
~6kb	815,006,427	148,315	5,495
~7kb	905,881,157	139,481	6,495
~8kb	978,965,060	130,607	7,496
~9kb	1,063,290,046	125,158	8,496
~10kb	1,084,089,752	114,232	9,490
10kb~	5,347,185,274	406,019	13,170
Total	12,307,418,920	1,919,636	6,411

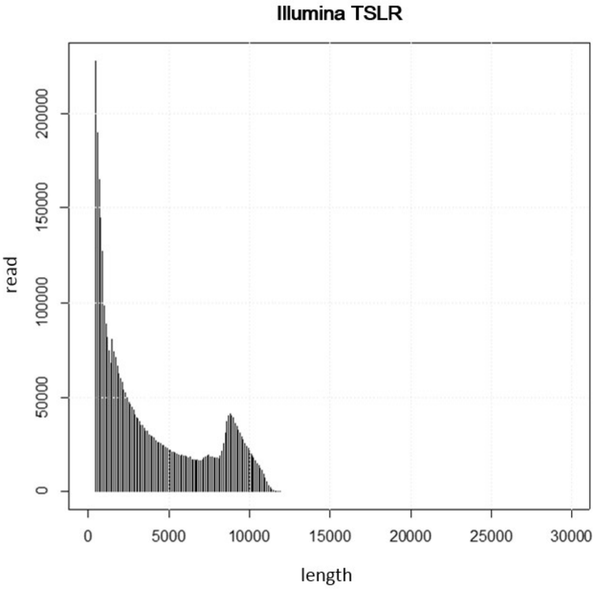


Figure 8. Length distribution of Illumina TruSeq synthetic long reads

Table 12. Illumina TruSeq synthetic long reads statistics

Size	Number of bases (bp)	Number of reads	Mean length (bp)
~2kb	1,745,885,089	1,627,362	1,073
~3kb	1,227,839,348	498,112	2,465
~4kb	1,200,052,670	345,449	3,474
~5kb	1,170,624,980	261,313	4,480
~6kb	1,141,935,546	208,259	5,483
~7kb	1,132,652,780	174,578	6,488
~8kb	1,358,992,691	181,044	7,506
~9kb	2,532,232,743	294,819	8,589
~10kb	2,879,791,577	304,656	9,453
10kb~	1,910,098,184	181,128	10,546
Total	16,300,105,608	4,076,720	3,998

Both types were used simultaneously, resulting in a decrease number of gaps from 1.75 % to 1.06 % of the expected genome size and a small increase in the final scaffold N50 length from 25.93 Mb to 26.08 Mb (Table 7). I suspect that the low quantity of long reads (only 1.2 % of read numbers compared to mate-pairs) is one reason for the small increase in the scaffold length (Table 13). Also, it was possible that the continuity information of the long reads were overlapping with those of next-generation sequencing (NGS) mate-pair sequences (various insert sizes to ~20 Kb).

Table 13. The number of sequence reads for scaffolding

	Mate-pairs (read depth: ~20×)	PacBio reads (read depth: ~10×)	TSLRs (read depth: ~5.3×)
The number of read information that can be used for scaffolding	952,677,682	7,356,133	4,076,720
The ratio to mate-pair number	100 %	0.77 %	0.43 %

Scaffolds usually contain misassembles[14,16]. I carefully and systematically assessed the quality of KOREF_S by generating nanochannel-based genome mapping data (~145 Gb of single-molecule maps; Fig. 9). I assembled the mapping data into 2.8 Gb of genome maps having an N50 length of 1.12 Mb (Table 14).

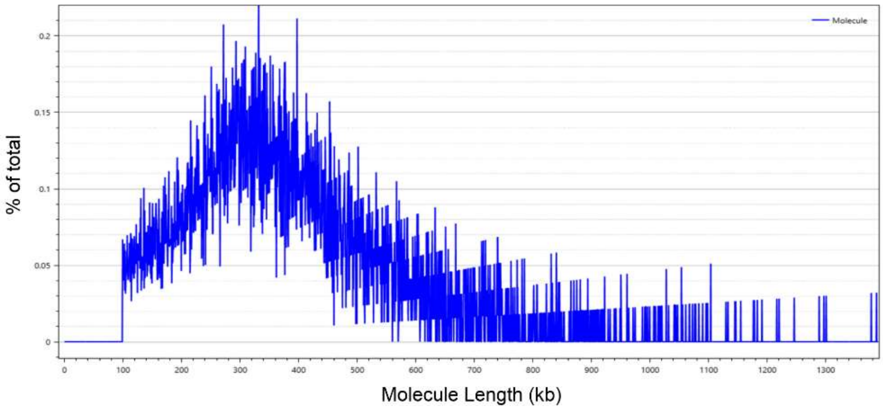


Figure 9. Length distribution of BioNano single molecule maps

Table 14. BioNano genome mapping data statistics

	BioNano single molecules	BioNano consensus maps
Total data	210 Gb	-
Single molecule N50	273 Kb	-
Molecules above 150Kb	145 Gb	-
Coverage depth	45 ×	-
Assembly size	-	2.78 Gb
Consensus map N50	-	1.12 Mb

A total of 93.1 % of KOREF_S scaffold regions (≥ 10 Kb) were covered by these genome maps, confirming their continuity (Fig. 10). To pinpoint misassembles of KOREF_S scaffolds, I manually checked all the alignment results of the genome maps (3,216 cases with align confidence ≥ 20) onto KOREF_S and GRCh38. Seven misassembled regions were detected in KOREF_S and were split for correction (Fig. 10). Next, my colleagues and I conducted a whole genome alignment of KOREF_S and GRCh38 to detect possible inter- or intra-chromosomal translocations (indicative of misassembled sequences; Fig. 11a). A total of 280 of the KOREF_S scaffolds (≥ 10 Kb) covered 93.5 % of GRCh38's chromosomal sequences (non-gaps). I found no large scale inter- or intra-chromosomal translocations. Additionally, as a fine-scale assessment, I aligned the short and long read sequence data to the KOREF_S scaffolds (self-to-self alignment). A total of 98.69 % of the scaffold sequences (≥ 2 Kb) were covered by equal or more than 20-fold (Table 15). My colleagues and I assigned KOREF_S's scaffolds to chromosomes using whole genome alignment information (chromosomal location and ordering information of scaffolds on GRCh38 chromosomes), to obtain KOREF_S chromosome sequences (~3.12 Gb of total length; Table 7 and Fig. 11b).

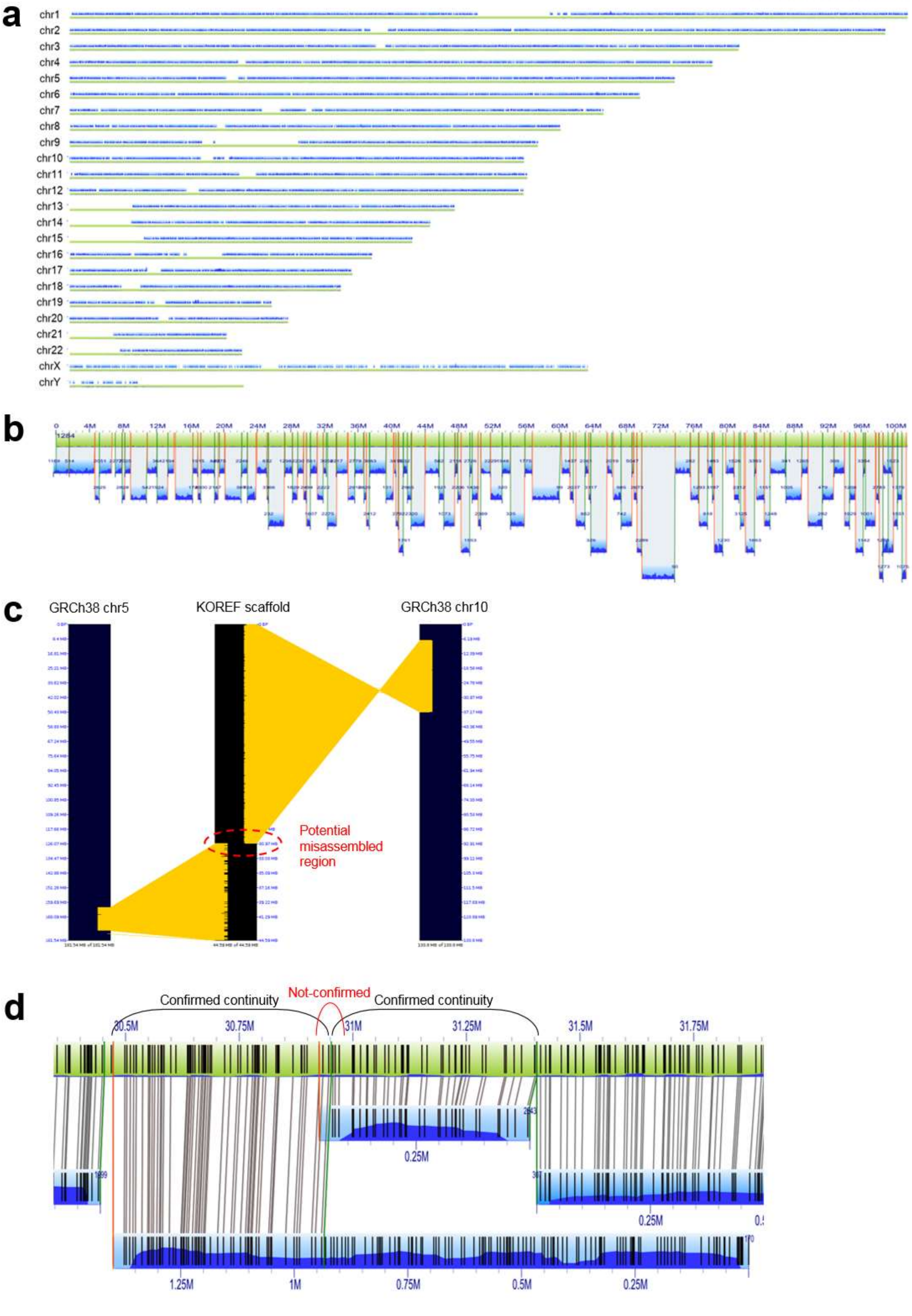


Figure 10. Assessment of scaffold assembly using BioNano genome mapping data. (**a**) Overall view of BioNano consensus maps compared to KOREF_S assembly. Green bars indicate KOREF_S scaffolds, and blue ones are assembled BioNano genome maps. (**b**) The longest KOREF_S scaffold (~101 Mb) confirmed by BioNano consensus maps. (**c**) An example of potentially misassembled region. (**d**) The confirmation of the potentially misassembled region in the panel (c), using the consensus maps.

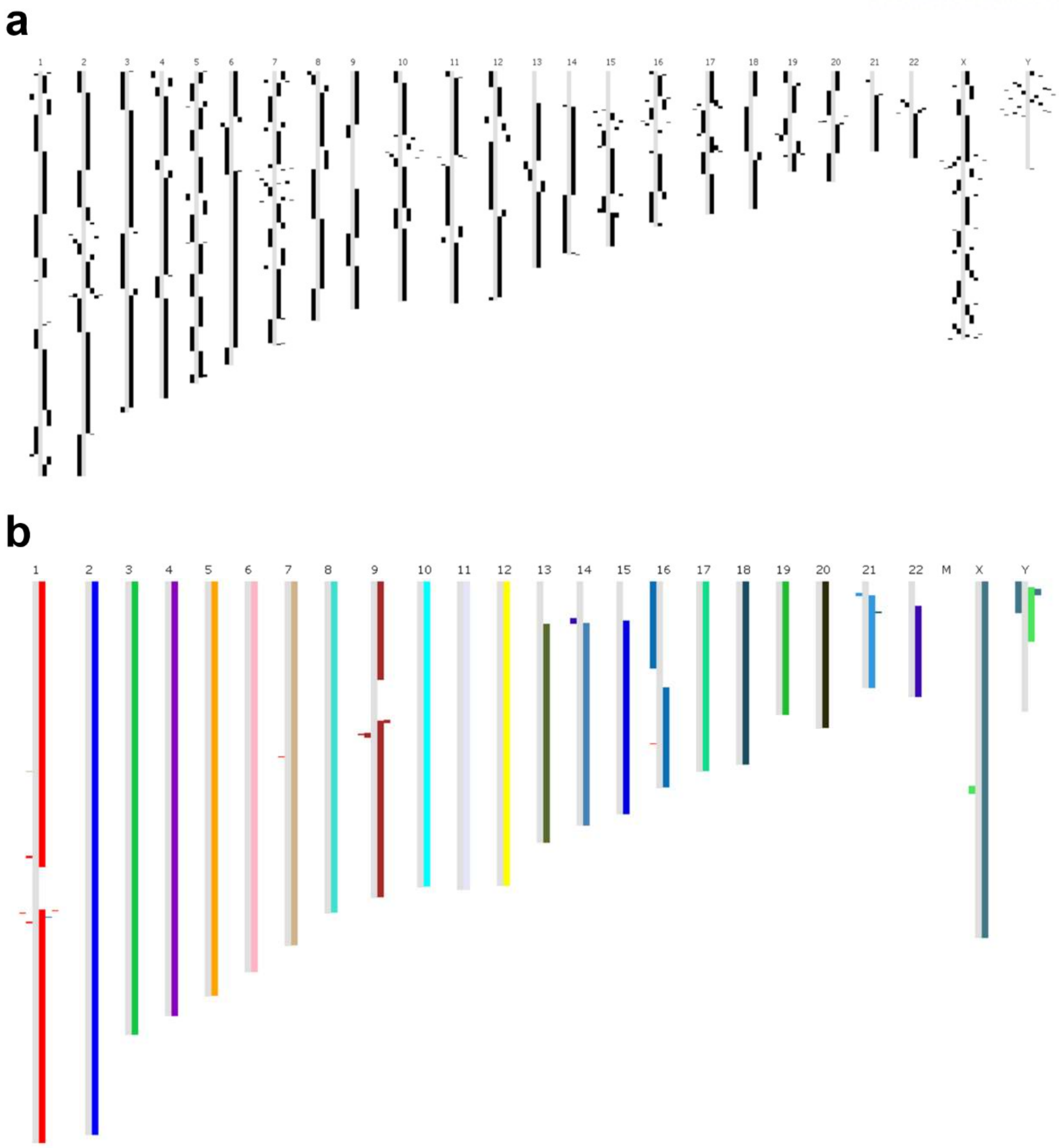


Figure 11. Whole genome alignment results between the human reference and KOREFs. (**a**) Whole genome alignments between GRCh38 and KOREF_S scaffolds. Gray bars are GRCh38 chromosomes, and black bars are KOREF_S scaffolds. (**b**) Whole genome alignments between GRCh38 and chromosome version of KOREF. Gray bars are GRCh38 chromosomes, and other color bars are KOREF chromosomes.

Table 15. Assessment of genome coverage based on the alignment of sequence reads

	≥ 10-depth	≥ 20-depth	≥ 30-depth
Percentage of covered regions (≥ 2Kb, without gaps)	98.94%	98.69%	98.46%

3.3 KOREF_C construction and genome annotation

Recently, Dewey *et al*. demonstrated much improved genotype accuracy for disease-associated variant loci using major allele reference sequences[5], which were built by substituting the ethnicity specific major allele (single base substitutions from the 1KGP) in the low-coverage European, African, and East-Asian reference genomes. I followed the same approach for KOREF_S by substituting sequences with both SNVs and small insertions or deletions (indels) that were commonly found in the 40 Korean PGP high-depth (average 31-fold mapped reads) whole genomes. This removed individual specific biases, and thus better represents common variants in the Korean population as a consensus reference (KOREF_C; Table 16).

Table 16. Mapping and variants statistics of 40 Korean whole genomes aligned to KOREF_S

Sample ID	Total number of raw reads	Mapped read depth (except 'N')	Read mapping rate (%)	Homozygous SNVs	Homozygous INDELs	Heterozygous SNVs	Heterozygous INDELs	All variants
KPGP-00002	98,317,515,960	27.64	99.29	962,066	146,462	2,958,707	292,082	4,359,317
KPGP-00006	93,448,081,980	24.73	99.28	1,431,527	204,234	2,915,971	276,219	4,827,951
KPGP-00032	112,190,946,660	30.36	99.29	1,444,163	215,475	2,955,815	296,145	4,911,598
KPGP-00033	108,196,466,760	29.95	99.30	1,406,058	211,651	2,961,708	297,035	4,876,452
KPGP-00039	101,141,448,400	30.19	99.16	1,391,102	212,028	2,991,047	315,678	4,909,855
KPGP-00056	111,361,334,200	32.24	99.34	1,419,373	230,317	3,100,438	340,429	5,090,557
KPGP-00086	102,626,322,600	29.88	99.34	1,423,097	228,216	3,074,640	335,156	5,061,109
KPGP-00125	118,670,365,980	33.12	99.31	1,438,747	211,687	2,932,733	291,074	4,874,241
KPGP-00127	118,883,354,760	32.81	99.33	1,416,527	206,959	2,948,523	288,104	4,860,113
KPGP-00128	117,849,278,700	32.76	99.29	1,407,530	208,532	2,941,634	292,805	4,850,501
KPGP-00129	107,124,150,780	29.96	99.28	1,440,746	203,979	2,908,731	271,108	4,824,564
KPGP-00131	120,142,829,340	33.36	99.29	1,432,319	211,261	2,970,372	289,604	4,903,556
KPGP-00132	122,237,363,160	33.93	99.30	1,411,276	210,946	2,946,694	297,988	4,866,904
KPGP-00134	119,540,641,320	32.54	99.28	1,416,157	207,904	2,931,855	288,305	4,844,221
KPGP-00136	114,984,689,940	30.71	99.30	1,429,777	204,804	2,940,492	274,170	4,849,243
KPGP-00137	118,027,255,140	32.97	99.28	1,403,331	207,581	2,940,643	289,256	4,840,811
KPGP-00138	123,868,546,380	33.39	99.32	1,398,902	207,327	2,938,964	289,045	4,834,238
KPGP-00139	105,730,760,700	29.32	99.28	1,397,287	207,216	2,918,240	291,707	4,814,450
KPGP-00141	111,508,577,820	31.41	99.24	1,405,400	207,892	2,926,108	288,957	4,828,357
KPGP-00142	125,024,326,200	32.62	99.29	1,443,241	211,075	2,943,175	292,818	4,890,309
KPGP-00144	127,001,127,600	33.96	99.30	1,422,369	211,512	2,973,541	296,396	4,903,818
KPGP-00145	111,861,808,380	31.18	99.29	1,438,003	210,730	2,953,375	293,052	4,895,160
KPGP-00205-B01-G	123,835,438,866	37.24	98.41	1,422,423	221,835	3,072,207	332,313	5,048,778
KPGP-00220	106,317,727,560	28.21	99.28	1,411,132	201,485	2,931,702	284,397	4,828,716
KPGP-00227	115,164,844,920	34.39	99.30	1,419,518	217,159	3,039,274	308,248	4,984,199
KPGP-00228	112,898,405,520	33.34	99.30	1,455,818	221,343	3,052,488	303,008	5,032,657
KPGP-00230	110,458,697,940	32.86	99.31	1,414,415	214,448	3,031,789	301,182	4,961,834
KPGP-00232	109,620,112,860	32.01	99.29	1,442,223	214,897	3,020,544	292,548	4,970,212
KPGP-00233	107,091,428,940	32.08	99.27	1,421,451	216,917	3,014,334	302,473	4,955,175
KPGP-00235	114,400,539,900	34.74	99.31	1,414,391	218,911	3,047,216	309,518	4,990,036
KPGP-00245-B01-G-PE500	102,078,086,860	31.40	99.11	1,465,527	223,235	3,031,190	322,301	5,042,253
KPGP-00254	122,277,928,000	34.56	99.24	1,427,301	221,720	3,080,569	313,709	5,043,299
KPGP-00255	102,221,657,600	29.67	99.34	1,414,140	227,857	3,083,228	336,527	5,061,752
KPGP-00256	127,033,362,000	36.61	99.35	1,422,753	235,874	3,174,628	355,538	5,188,793
KPGP-00265-B01-G-P500	90,922,729,400	27.53	99.29	1,414,977	216,811	2,964,359	306,126	4,902,273
KPGP-00266-B01-G-P500	91,666,078,800	27.38	99.32	1,374,215	212,665	2,962,424	307,516	4,856,820
KPGP-00269-B01-G-PE500	100,240,975,874	30.81	99.32	1,449,250	219,822	3,052,622	324,886	5,046,580
KPGP-00317-B01-G-PE500	103,075,371,660	26.76	87.15	1,400,454	208,300	3,002,602	306,055	4,917,411
KPGP-00318-B01-G-PE500	101,805,865,370	28.22	95.42	1,440,304	218,383	2,971,844	319,451	4,949,982
KPGP-00319-B01-G-PE500	100,957,938,100	27.77	97.17	1,403,626	213,564	3,063,114	315,785	4,996,089

Roughly two million variants (1,951,986 SNVs and 219,728 indels), commonly found in the 40 high quality short read Korean genome data, were integrated. Additionally, KOREF_S's mitochondrial DNA (mtDNA) was independently sequenced and assembled, resulting in a 16,570bp mitogenome that was similar, in structure, to that of GRCh38. A total of 34 positions of KOREF_S mtDNA were different from that of GRCh38 (Table 17). KOREF_S's mtDNA could be assigned to the D4e haplogroup that is common in East-Asians, whereas GRCh38 mtDNA belongs to European haplogroup H.

Table 17. Variations found in KOREF_S mtDNA compared to GRCh38 mtDNA

Position	Ref	Alt	Gene	Variant type	Amino acid Change	dbSNP143
73	A	G	*TRNF*	Upstream variant	-	rs3087742
263	A	G	*TRNF*	Upstream variant	-	rs2853515
310	T	CTC	*TRNF*	Upstream variant	-	rs66492218
489	T	C	*TRNF*	Upstream variant	-	rs28625645
750	A	G	*RNR1*	Noncoding variant	-	rs2853518
1438	A	G	*RNR1*	Noncoding variant	-	rs2001030
2706	A	G	*RNR2*	Noncoding variant	-	rs2854128
3010	G	A	*RNR2*	Noncoding variant	-	rs3928306
3107	N	-	*RNR2*	Noncoding variant	-	-
4769	A	G	*ND2*	Synonymous variant	-	rs3021086
4883	C	T	*ND2*	Synonymous variant	-	rs200763872
5178	C	A	*ND2*	Missense variant	Met237Leu	rs28357984
7028	C	T	*COX1*	Synonymous variant	-	rs2015062
8414	C	T	*ATP8*	Missense variant	Leu17Phe	rs28358884
8701	A	G	*ATP6*	Missense variant	Thr58Ala	rs2000975
8860	A	G	*ATP6*	Missense variant	Thr112Ala	rs2001031
9010	G	A	*ATP6*	Missense variant	Ala162Thr	-
9540	T	C	*COX3*	Synonymous variant	-	rs2248727
10398	A	G	*ND3*	Missense variant	Thr114Ala	rs2853826
10400	C	T	*ND3*	Synonymous variant	-	rs28358278
10873	T	C	*ND4*	Synonymous variant	-	rs2857284
11215	C	T	*ND4*	Synonymous variant	-	rs386419997
11719	G	A	*ND4*	Synonymous variant	-	-
12705	C	T	*ND5*	Synonymous variant	-	-
14668	C	T	*ND6*	Synonymous variant	-	rs28357678
14766	C	T	*CYTB*	Missense variant	Thr7Ile	rs527236041
14783	T	C	*CYTB*	Synonymous variant	-	rs527236042
15043	G	A	*CYTB*	Synonymous variant	-	rs527236043
15148	G	A	*CYTB*	Synonymous variant	-	rs527236206
15184	T	C	*CYTB*	Synonymous variant	-	-
15301	G	A	*CYTB*	Synonymous variant	-	rs527236045
15326	A	G	*CYTB*	Missense variant	Thr194Ala	rs2853508
16223	C	T	*CYTB*	Downstream variant	-	rs2853513
16362	T	C	*CYTB*	Downstream variant	-	rs62581341

KOREF_C GC content and distribution were similar to other human assemblies except the African assembly, which has the lowest quality among them (Fig. 12). My colleagues and I annotated KOREF_C for repetitive elements by integrating *de novo* prediction and homology-based alignments. Repetitive elements occupied 1.51 Gb (47.13 %) of KOREF_C (Table 18), which is slightly less than found in GRCh38 (1.59 Gb). On the other hand, KOREF_C contained more repeats than the Mongolian genome (1.36 Gb), which was assembled by NGS short reads only. I predicted 20,400 protein coding genes for KOREF_C (Table 19). By comparing KOREF_C with other human assemblies (GRCh38, CHM1_1.1, HuRef, African, Mongolian, and YH), a total of 875.8 Kb KOREF_C sequences (≥100bp of fragments) were defined as novel (Table 20).

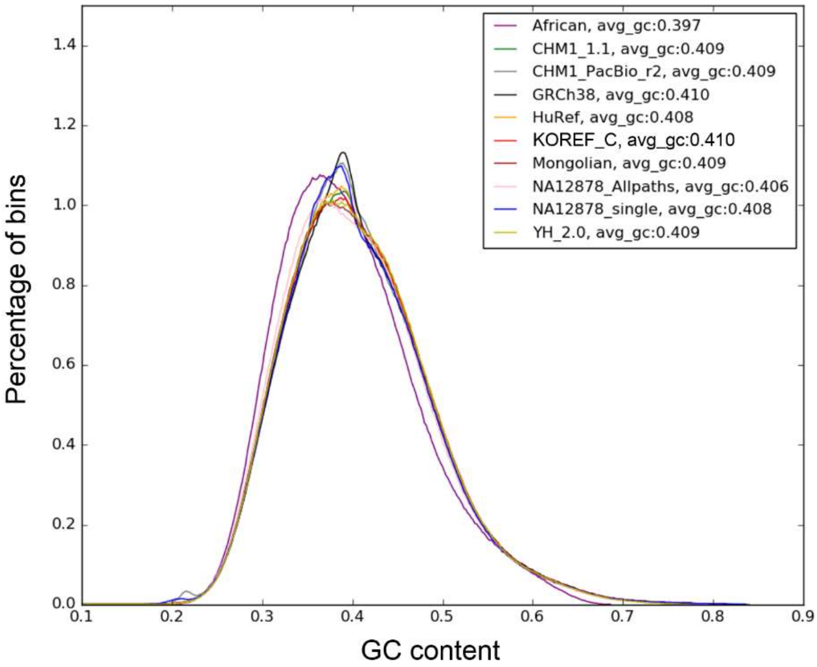


Figure 12. GC content distributions in the human genome assemblies. The *x*-axis is GC content, and the *y*-axis is the proportion of the bin count with the specified GC content.

Table 18. KOREF_C repeat annotation

	Repbase TEs		*De novo*		Combined	
	Length (bp)	% in Genome	Length (bp)	% in Genome	Length (bp)	% in Genome
DNA	106,469,686	3 %	24,415,664	1 %	108,618,651	3 %
LINE	610,159,517	19 %	536,712,478	17 %	745,903,228	23 %
SINE	390,299,729	12 %	254,443,404	8 %	425,991,881	13 %
LTR	267,766,723	8 %	112,840,399	4 %	270,236,817	8 %
Unknown	837,329	0 %	17,216,396	1 %	18,050,168	1 %
Total	1,450,469,642	45 %	994,936,953	31 %	1,513,511,651	47 %

Table 19. KOREF_C protein-coding gene prediction

Gene set	Gene number	Average transcript length (bp)	Average CDS length (bp)	Average exon per gene	Average exon length (bp)	Average intron length (bp)
Homology (Human)	18,564	51,797.23	1,701.15	9.80	173.59	5,847.84
de novo	18,988	51,291.92	1,485.38	9.16	162.07	6,099.12
mtDNA	13	876.54	876.54	1.00	876.54	-
Combined	20,400	49,584.30	1,635.35	9.41	173.76	5,847.28

Table 20. KOREF_C-specific novel sequence identification

a. KOREF_S short reads mapped to each human genome assembly

	Mapped KOREF_S short reads	Unmapped KOREF_S short reads	Mapped reads (out of unmapped reads to other human assemblies) to KOREF_C assembly	Total length of novel sequences (bp; the regions with length ≥100bp and covered by at least three reads)
GRCh38	-	4,087,416	1,340,733	4,676,384
Microbial sequences	2,485,935	1,601,481		
CHM1_1.1	261,251	1,340,230	1,080,569	4,012,692
HuRef	900,070	440,160	182,060	1,305,352
African	74,658	365,502	108,958	1,024,687
Mongolian	40,008	325,494	69,725	890,472
YH	3,705	321,789	67,385	875,820

b. Length distribution of KOREF_C novel sequences

Length	Number of fragments (≥100bp)
100 – 500bp	2,531
501 – 1,000bp	240
1,001 – 5,000bp	89
5,001bp – 10 Kb	1
Above 10 Kb	1
Total length of novel sequences (bp)	875,820

3.4 KOREF_C compared with other human genomes

I assessed the quality of ten human genome assemblies (CHM1_PacBio_r2, CHM1_1.1, NA12878_single, NA12878_Allpaths, HuRef, Mongolian, YH_2.0, African, KOREF_C, and another Korean single individual assembly AK1[62]) by comparing assembly statistics, and the recovery rates for GRCh38 genome, segmentally-duplicated regions, and repetitive sequences (Tables 21–24).

Table 21. Systematic comparison of assembly quality. Major sequencing and mapping data used in the assembly are marked by superscript letters: NGS short reads, S; long reads, L; genome maps, M; indexed BAC end sequences, B; chain-terminating Sanger sequences; C.

Assembly (level)	Total sequence length (bp)	Scaffold or Contig N50 (Mb) / L50	GRCh38 recovery rate (%)	Segmental duplication length (bp)	Repeat length (bp)	Detected RefSeq genes (intact only)
GRCh38^{C} (chromosome)	3,209,286,105	67.79 / 16	-	212,777,868 (6.63 %)	1,564,209,365 (48.74 %)	20,135
KOREF_C^{S,L,M} (chromosome)	**3,211,075,818**	**26.46 / 35**	**88.47 (scaffolds)**	**149,353,191 (4.65 %)**	**1,452,404,484 (45.23 %)**	**17,758**
AK1^{L,M} (scaffold)	2,904,207,228	44.85 / 21	87.90	144,868,735 (4.99 %)	1,454,888,506 (50.10 %)	17,759
CHM1_PacBio_r2^{L} (contig)	2,996,426,293	26.90 / 30	88.02	205,559,250 (6.86 %)	1,541,211,387 (51.43 %)	17,657
CHM1_1.1^{S,B} (reference-guided)	3,037,866,619	50.36 / 20	-	157,426,845 (5.18 %)	1,417,977,130 (46.68 %)	18,040
NA12878_single^{L,M} (scaffold)	3,176,574,379	26.83 / 37	88.26	168,652,649 (5.31 %)	1,545,168,387 (48.64 %)	6,610
NA12878_Allpaths^{S} (scaffold)	2,786,258,565	12.08 / 67	82.89	90,343,965 (3.24 %)	1,250,655,296 (44.89 %)	16,995
HuRef^{C} (chromosome)	2,844,000,504	17.66 / 48	85.85	134,317,812 (4.72 %)	1,411,487,301 (49.63 %)	16,968
Mongolian^{S} (scaffold)	2,881,945,563	7.63 / 111	86.54	121,384,034 (4.21 %)	1,399,420,366 (48.56 %)	17,189
YH_2.0^{S} (scaffold)	2,911,235,363	20.52 / 39	86.31	127,254,909 (4.37 %)	1,397,013,571 (47.99 %)	17,125
African^{S} (scaffold)	2,676,008,911	0.062 / 11,689	69.47	55,830,170 (2.09 %)	968,988,149 (36.21 %)	9,167

Table 22. Global assembly statistics of human assemblies. Major sequencing and mapping data used in the assembly are marked by superscript letters: NGS short reads, S; long reads, L; genome maps, M; indexed BAC end sequences, B; chain-terminating Sanger sequences; C.

Statistics	GRCh38^{C}	KOREF_C^{S,L,M}	AK1^{L,M}	CHM1_PacBio_r2^{L}	CHM1_1.1^{S,B}	NA12878_single^{L,M}
Assembly level	Chromosome	Chromosome	Scaffold	Contig	Chromosome	Scaffold
Total sequence length	3,209,286,105	3,211,075,818	2,904,207,228	2,996,426,293	3,037,866,619	3,176,574,379
Total assembly gap length	159,970,007	297,934,127	37,339,479	0	210,229,812	146,352,286
Gaps between scaffolds	349	4,495	0	-	225	0
Number of scaffolds	735	4,481	2,832	-	163	18,903
Scaffold N50	67,794,873	26,457,717	44,846,623	-	50,362,920	26,834,081
Scaffold L50	16	35	21	-	20	37
Number of contigs	1,385	198,871	3,096	3,641	40,828	21,235
Contig N50	56,413,054	47,858	18,080,262	26,899,841	143,936	1,557,716
Contig L50	19	17,749	46	30	5,635	532
Total number of chromosomes and plasmids	25	25	0	0	23	0

Statistics	NA12878_Allpaths^{S}	HuRef^{C}	Mongolian^{S}	YH_2.0^{S}	African^{S}
Assembly level	Scaffold	Chromosome	Scaffold	Scaffold	Scaffold
Total sequence length	2,786,258,565	2,844,000,504	2,881,945,563	2,911,235,363	2,676,008,911
Total assembly gap length	171,353,127	34,429,377	58,452,127	105,204,230	592,227,090
Gaps between scaffolds	0	1,396	0	0	0
Number of scaffolds	11,393	4,530	221,013	125,643	314,786
Scaffold N50	12,084,118	17,664,250	7,632,466	20,520,932	62,478
Scaffold L50	67	48	111	39	11,689
Number of contigs	231,194	71,333	321,009	361,157	5,313,377
Contig N50	23,924	108,431	56,244	20,516	887
Contig L50	30,971	7,164	14,915	40,005	642,142
Total number of chromosomes and plasmids	0	24	0	0	0

Table 23. GRCh38 genome recovery rates of human assemblies. Whole genome alignment approach was used to calculate GRCh38 genome recovery rates of human assemblies. Major sequencing and mapping data used in the assembly are marked by superscript letters: NGS short reads, S; long reads, L; genome maps, M; chain-terminating Sanger sequences; C.

Assembly	GRCh38 length (bp)	Assembly length (bp)	Total alignment results (including duplicated alignments)		Non-redundant alignment results (excluding duplicated alignments)	
			Length of aligned regions (bp)	GRCh38 coverage (%)	Length of aligned regions (bp)	GRCh38 coverage (%)
KOREF_S_scaffold^{S,L,M}	3,209,286,105	2,944,499,428	2,956,077,148	92.11	2,839,274,905	88.47
KOREF_S_contig^{S,L,M}	3,209,286,105	2,913,213,215	2,944,669,829	91.75	2,755,264,778	85.85
AK1^{L,M}	3,209,286,105	2,904,207,228	2,960,869,067	92.26	2,821,038,382	87.90
CHM1_PacBio_r2^{L}	3,209,286,105	2,996,426,293	2,968,736,981	92.50	2,824,727,975	88.02
NA12878_single^{L,M}	3,209,286,105	3,176,574,379	2,948,546,881	91.88	2,832,488,088	88.26
NA12878_Allpaths^{S}	3,209,286,105	2,786,258,565	2,753,492,425	85.80	2,660,094,223	82.89
HuRef_contig^{C}	3,209,286,105	2,809,571,127	2,942,411,659	91.68	2,755,302,479	85.85
Mongolian^{S}	3,209,286,105	2,881,945,563	2,916,062,756	90.86	2,777,307,567	86.54
YH_2.0^{S}	3,209,286,105	2,911,235,363	2,885,254,871	89.90	2,769,798,873	86.31
African^{S}	3,209,286,105	2,676,008,911	2,354,016,286	73.35	2,229,410,403	69.47

Table 24. Predicted segmentally-duplicated and repetitive sequence regions in human assemblies. Homology search was used to identify segmentally-duplicated and repetitive regions. Major sequencing and mapping data used in the assembly are marked by superscript letters: NGS short reads, S; long reads, L; genome maps, M; indexed BAC end sequences, B; chain-terminating Sanger sequences; C.

Assembly	Assembly length	SD length	SD %	Repeat length	Repeat %
GRCh38^{C}	3,209,286,105	212,777,868	6.63	1,564,209,365	48.74
CHM1_PacBio_r2^{L}	2,996,426,293	205,559,250	6.86	1,541,211,387	51.43
NA12878_single^{L,M}	3,176,574,379	168,652,649	5.31	1,545,168,387	48.64
CHM1_1.1^{S,B}	3,037,866,619	157,426,845	5.18	1,417,977,130	46.68
KOREF_C^{S,L,M}	3,211,075,818	149,353,191	4.65	1,452,404,484	45.23
KOREF_S_scaffold^{S}	2,921,901,481	139,246,009	4.77	1,438,015,194	49.22
AK1^{L,M}	2,904,207,228	144,868,735	4.99	1,454,888,506	50.10
HuRef^{C}	2,844,000,504	134,317,812	4.72	1,411,487,301	49.63
YH_2.0^{S}	2,911,235,363	127,254,909	4.37	1,397,013,571	47.99
Mongolian^{S}	2,881,945,563	121,384,034	4.21	1,399,420,366	48.56
NA12878_Allpaths^{S}	2,786,258,565	90,343,965	3.24	1,250,655,296	44.89
African^{S}	2,676,008,911	55,830,170	2.09	968,988,149	36.21

The results showed that KOREF_C was more contiguous (26.46 Mb of N50) than any of the short-read based *de novo* assemblies, but comparable to two long-read based assemblies (26.83 Mb of N50 for NA12878_single; 26.90 Mb of N50 for CHM1_PacBio_r2); KOREF_C was hybrid-assembled by compiling heterogeneous sequencing and mapping technologies, however, a majority of KOREF_C sequences was derived from NGS short reads. However, KOREF_C's contig size is small (47.86 Kb of N50 and 17,749 of L50; Table 22) compared to long-read based assemblies due to the low level of continuity information of short reads. KOREF_C showed a comparable GRCh38 recovery rate with other long-read assemblies (Tables 21 and 23). KOREF (KOREF_S scaffolds) recovered duplicated and repetitive regions more efficiently than other short-read based *de novo* assemblies. However, KOREF recovered duplicated and repetitive regions less than the two (CHM1_PacBio_r2 and NA12878_single) PacBio long-read assemblies (Table 24); importantly, KOREF recovered those regions more efficiently than the other Korean PacBio long-read based assembly, AK1. Notably, a higher sequencing depth long-read assembly, CHM1_PacBio_r2, recovered the most segmentally-duplicated regions, almost as well as GRCh38, indicating that long read information is important to recover such challenging genomic regions. Also, structural polymorphisms between the two haplotypes in a donor is one of the most significant factors affecting the assembly quality[15,63]. Therefore, it was expected that CHM1_PacBio_r2, a haploid assembly, showed a superior genome recovery for segmentally-duplicated regions than other assemblies using a diploid source. Additionally, I compared the assembly quality by mapping the re-sequencing data of a single haplotype genome (CHM1) to the human assemblies (Fig. 13). Ideally, CHM1 should lack heterozygous variants, if the human assembly recovered the entire genome efficiently. CHM1_PacBio_r2 was the most accurate (having the lowest number of heterozygous variants) in resolving the entire human genome, and KOREF_C was the most accurate among the short-read based assemblies. These results confirm that short-reads based *de novo* assemblies have a reduced power to fully resolving the entire genome sequences accurately[14].

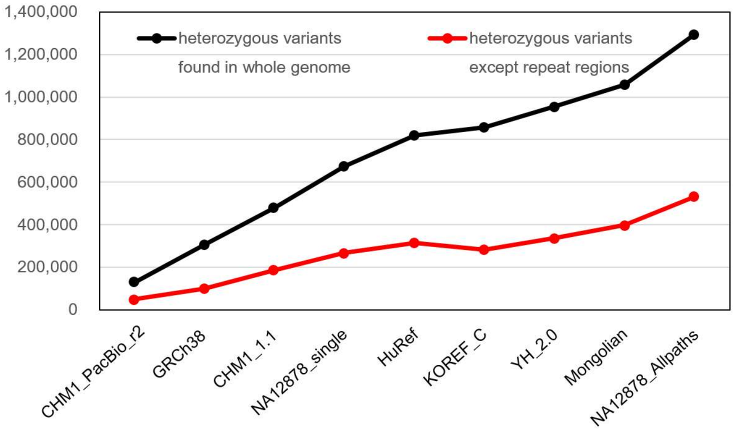


Figure 13. Numbers of heterozygous variants found in re-sequencing data from a single haplotype (CHM1) genome

I also conducted gene content assessments by comparing the number of detected RefSeq[46] protein-coding genes in each human assembly (Tables 21 and 25). The RefSeq genes were the best recovered in CHM1_1.1 (18,040), which was assembled using that reference as a guide. Among the *de novo* assembled genomes, KOREF_C showed the highest level of intact RefSeq gene recovery (17,758), even more than the two Caucasian long-read based assemblies (~17,657). Notably, the NA12878_single genome, which was hybrid assembled by combining single-molecule long reads with genome maps, had the lowest number (6,610) of intact protein-coding genes, even lower than the low quality African genome (9,167). I confirmed that NA12878_single had many frame-shifts in its coding regions. This can be explained by the higher error rates of PacBio single-molecule long reads, which could not be corrected by an error correction step due to its low sequencing depth (46× coverage)[21,64].

Table 25. Predicted protein-coding genes in human assemblies. Homology search was used to identify RefSeq protein-coding genes. Major sequencing and mapping data used in the assembly are marked by superscript letters: NGS short reads, S; long reads, L; genome maps, M; indexed BAC end sequences, B; chain-terminating Sanger sequences; C.

# of genes in RefSeq	# of intact genes in RefSeq (without genes having premature stop codons)	Assembly	# of searched genes by TblastN (E-value > 1E-05, Best hit only)	# of gene models by Exonerate prediction (at least 50% of the maximal score obtainable for query)	# of detected RefSeq genes (by removing genes having premature stop codons)
20,196	20,135	African^{S}	19,924	12,282	9,167
		CHM1_1.1^{S,B}	20,167	19,848	18,040
		CHM1_PacBio_r2^{L}	20,176	19,888	17,657
		HuRef^{C}	20,165	19,578	16,968
		KOREF_C^{S,L,M}	20,181	19,748	17,758
		KOREF_S_scaffold^{S}	20,179	19,719	17,750
		Mongolian^{S}	20,174	19,458	17,189
		NA12878_Allpaths^{S}	20,117	18,978	16,995
		NA12878_single^{L,M}	20,119	19,482	6,610
		YH_2.0^{S}	20,161	19,241	17,125
		AK1^{L,M}	-	-	17,759

3.5 Structural variation comparison

My colleagues and I investigated SVs, such as large insertions, deletions, and inversions, in the eight human assemblies by comparing to GRCh38 (since there were no paired-end read data, HuRef was not used in this analysis; AK1 was also not used, as it was not published at that time when the analysis was performed). The analysis showed that the assembly quality is determined primarily by sequencing platform (i.e., sequence read lengths), and therefore, I had to consider that mis-assemblies could generate erroneous SVs. Two Caucasian samples (CHM1 and NA12878) were assembled using short-read sequences as well as long reads, and therefore, allow an examination of the association between the assembly quality and SV identification. The CHM1 sample's ethnicity was confirmed to be Caucasian using ancestry-sensitive DNA markers in autosomes[65] and mitochondrial DNA sequences (Fig. 14). SVs that could have been derived from possible misassembles were filtered out by comparing the ratio of aligned single-end reads to paired-end reads (S/P ratio) as previously suggested[57] (see Methods).

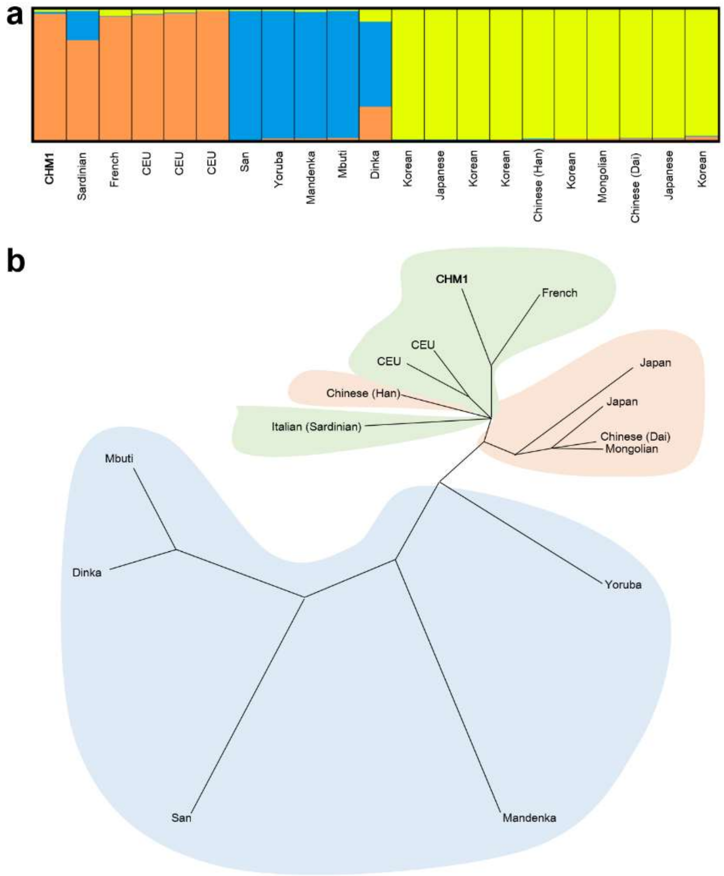


Figure 14. CHM1 ethnicity confirmation. (**a**) STRUCTURE analysis using 47 ancestry-sensitive DNA markers in autosomes. For *K*=3, CHM1 is grouped together with Europeans. (**b**) Mitochondrial DNA (mtDNA) sequence comparison. The mtDNA sequences were generated by mapping CHM1's Illumina short reads into GRCh38 mtDNA sequences and building consensus sequences.

A total of 6,397 insertions (> 50bp), 3,399 deletions (> 50bp), and 42 inversions were found in KOREF_C compared to GRCh38, making up 9,838 total SVs. This is slightly fewer than those found in the Mongolian (12,830 SVs) and African (10,772 SVs) assemblies, but greater than those found in CHM1 and NA12878 assemblies (~5,179 SVs; Tables 26–28).

Table 26. Summary of structural variations in eight human assemblies compared to GRCh38. Major sequencing and mapping data used in the assembly are marked by superscript letters: NGS short reads, S; long reads, L; genome maps, M; indexed BAC end sequences, B.

Assembly	Total SVs	Novel SVs (insertions and deletions only)	SVs in repetitive regions	SVs in segmentally-duplicated regions	Assembly specific SVs (insertions and deletions only)	SVs shared with the CHM1 PacBio read mapping results (insertions and deletions only)
KOREF_C^{S,L,M}	9,838	8,392 (85.7 %)	6,992 (71.1 %)	912 (9.3 %)	6,691 (68.3 %)	955 (9.7 %)
Mongolian^{S}	12,830	10,775 (87.7 %)	8,929 (69.6 %)	1,242 (9.7 %)	9,101 (74.1 %)	834 (6.8 %)
YH_2.0^{S}	5,027	4,664 (93.8 %)	4,119 (81.9 %)	633 (12.6 %)	3,063 (61.6 %)	148 (3.0 %)
CHM1_PacBio_r2^{L}	3,454	3,130 (92.0 %)	2,340 (67.7 %)	1,002 (29.0 %)	2,448 (72.0 %)	301 (8.8 %)
CHM1_1.1^{S,B}	3,926	3,258 (83.7 %)	2,848 (72.5 %)	394 (10.0 %)	2,800 (71.9 %)	487 (12.5 %)
NA12878_single^{L,M}	4,859	4,171 (86.7 %)	3,339 (68.7 %)	1,041 (21.4 %)	3,492 (72.6 %)	400 (8.3 %)
NA12878_Allpaths^{S}	5,179	4,649 (91.0 %)	4,014 (77.5 %)	378 (7.3 %)	3,787 (74.1 %)	269 (5.3 %)
African^{S}	10,772	10,026 (94.0 %)	8,362 (77.6 %)	425 (3.9 %)	8,935 (83.8 %)	212 (2.0 %)

Table 27. Structural variations found in human assemblies compared to GRCh38. Major sequencing and mapping data used in the assembly are marked by superscript letters: NGS short reads, S; long reads, L; genome maps, M; indexed BAC end sequences, B.

Assembly	Types	Insertion	Deletion	Inversion	Total
KOREF_C^{S,L,M}	No. of Confident SVs	6,397	3,399	42	9,838
	Minimum (bp)	51	51	45	-
	Maximum (bp)	37,813	36,793	44,546	-
Mongolian^{S}	No. of Confident SVs	6,904	5,386	540	12,830
	Minimum (bp)	51	51	90	-
	Maximum (bp)	44,580	44,577	22,225	-
YH_2.0^{S}	No. of Confident SVs	3,896	1,077	54	5,027
	Minimum (bp)	51	51	53	-
	Maximum (bp)	37,683	43,540	39,965	-
CHM1_PacBio_r2^{L}	No. of Confident SVs	2,969	433	52	3,454
	Minimum (bp)	51	51	14	-
	Maximum (bp)	37,524	24,278	50,943	-
CHM1_1.1^{S,B}	No. of Confident SVs	2,415	1,477	34	3,926
	Minimum (bp)	51	51	44	-
	Maximum (bp)	35,612	18,511	16,592	-
NA12878_single^{L,M}	No. of Confident SVs	3,896	914	49	4,859
	Minimum (bp)	51	51	23	-
	Maximum (bp)	43,701	16,093	20,342	-
NA12878_Allpaths^{S}	No. of Confident SVs	4,012	1,097	70	5,179
	Minimum (bp)	51	51	53	-
	Maximum (bp)	40,018	7,860	46,762	-
African^{S}	No. of Confident SVs	7,991	2,673	108	10,772
	Minimum (bp)	51	51	12	-
	Maximum (bp)	24,657	23,065	39,807	-

Table 28. Structural variations found in genic regions. Major sequencing and mapping data used in the assembly are marked by superscript letters: NGS short reads, S; long reads, L; genome maps, M; indexed BAC end sequences, B.

Region	KOREF_C^{S,L,M}	Mongolian^{S}	YH_2.0^{S}	CHM1_PacBio_r2^{L}	CHM1_1.1^{S,B}	NA12878_single^{L,M}	NA12878_Allpaths^{S}	African^{S}
CDS	403	559	122	134	173	149	192	288
UTR	193	277	60	48	92	70	105	115
Intron	2,958	3,388	783	884	1,444	1,184	1,261	1,629
Gene (Total)	2,985	3,427	792	899	1,466	1,205	1,281	1,650

Notably, YH_2.0 (5,027 SVs) had a similar number of SVs compared to those found in the Caucasian assemblies, than other Asian assemblies. The length distribution of SVs found in these assemblies showed a similar pattern (Figs. 15 and 16), with a peak at the 200-400bp size range, due to *Alu* element insertions and deletions[15,57].

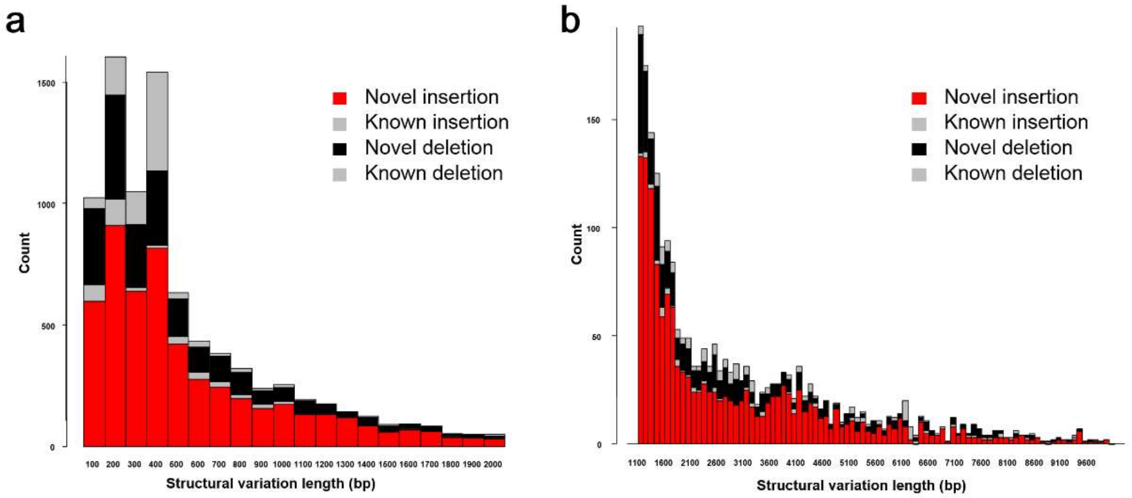


Figure 15. Length distributions of KOREF_C structural variations compared to GRCh38. (**a**) Structural variation lengths range from 50bp to 2 Kb. (**b**) Structural variation lengths range from 1 Kb to 10 Kb.

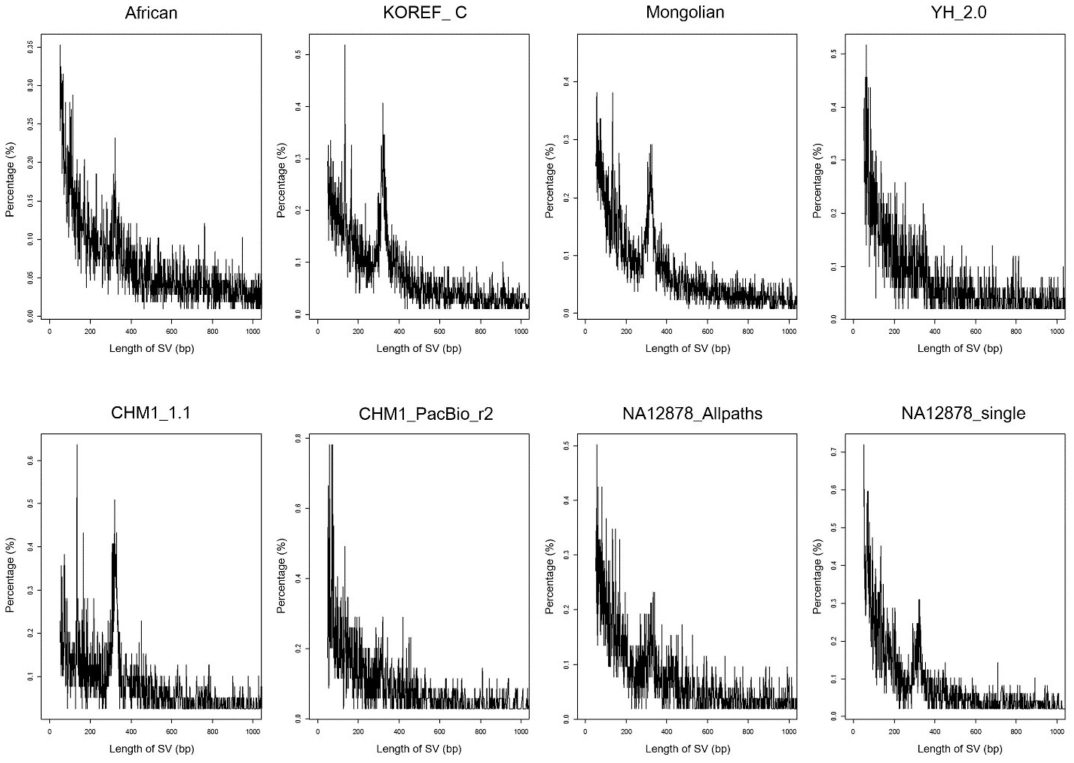


Figure 16. Length distributions of structural variations found in human assemblies compared to GRCh38

The fractions of SVs in repeat regions were higher in the short-read based assemblies (69.6~81.9 %) than long-read assemblies (67.7~68.7 %; Tables 26 and 29). On the other hand, the fractions of SVs in the segmentally-duplicated regions were much higher in the long-read assemblies (21.4~29.0 %) than short-read assemblies (3.9~12.6 %; Tables 26 and 30).

Table 29. Structural variations in repetitive regions. Major sequencing and mapping data used in the assembly are marked by superscript letters: NGS short reads, S; long reads, L; genome maps, M; indexed BAC end sequences, B.

Assembly	Total SVs	SVs in repeats	SVs in non-repeats	The percentage of SVs in repeats
KOREF_C^{S,L,M}	9,838	6,992	2,846	71.1
Mongolian^{S}	12,830	8,929	3,901	69.6
YH_2.0^{S}	5,027	4,119	908	81.9
CHM1_PacBio_r2^{L}	3,454	2,340	1,114	67.7
CHM1_1.1^{S,B}	3,926	2,848	1,078	72.5
NA12878_single^{L,M}	4,859	3,339	1,520	68.7
NA12878_Allpaths^{S}	5,179	4,014	1,165	77.5
African^{S}	10,772	8,362	2,410	77.6

Table 30. Structural variations in segmentally-duplicated regions. Major sequencing and mapping data used in the assembly are marked by superscript letters: NGS short reads, S; long reads, L; genome maps, M; indexed BAC end sequences, B.

Assembly	Total SVs	SVs in segmental duplicated regions	SVs not in segmental duplicated regions	The percentage of SVs in segmental duplicated regions
KOREF_C^{S,L,M}	9,838	912	8,926	9.3
Mongolian^{S}	12,830	1,242	11,588	9.7
YH_2.0^{S}	5,027	633	4,394	12.6
CHM1_PacBio_r2^{L}	3,454	1,002	2,452	29.0
CHM1_1.1^{S,B}	3,926	394	3,532	10.0
NA12878_single^{L}	4,859	1,041	3,818	21.4
NA12878_Allpaths^{S}	5,179	378	4,801	7.3
African^{S}	10,772	425	10,347	3.9

Of the KOREF_C SVs, 93.8 % of insertions and 70.4 % of deletions were not found in public SV databases and hence defined as novel (Tables 26 and 31, Fig. 15, and Methods). The fraction of novel SVs in KOREF_C was similar to those found in other human assemblies but smaller than other short-read only *de novo* assemblies. Regardless of sequencing platform, all assemblies showed a greater fractions of novel SVs than those found by mapping CHM1's PacBio SMRT reads to the human reference genome (here termed CHM1_mapping)[15]. Notably, CHM1_PacBio_r2, which was assembled using the same sample's PacBio long reads, also showed a much higher fraction of novel SVs.

Table 31. Novel structural variations found in the human assemblies. Major sequencing and mapping data used in the assembly are marked by superscript letters: NGS short reads, S; long reads, L; genome maps, M; indexed BAC end sequences, B.

Assembly	Insertion				Deletion			
	# of insertions	Novel insertions	Known insertions	% of novel insertions	# of deletions	Novel deletions	Known deletions	% of novel deletions
CHM1 PacBio read mapping approach	10,978	10,029	949	91.4	7,071	3,164	3,907	44.7
KOREF_C^{S,L,M}	6,397	5,999	398	93.8	3,399	2,393	1,006	70.4
Mongolian^{S}	6,904	6,500	404	94.1	5,386	4,275	1,111	79.4
YH_2.0^{S}	3,896	3,806	90	97.7	1,077	858	219	79.7
CHM1_PacBio_r2^{L}	2,969	2,802	167	94.4	433	328	105	75.8
CHM1_1.1^{S,B}	2,415	2,374	41	98.3	1,477	884	593	59.8
NA12878_single^{L,M}	3,896	3,633	263	93.2	914	538	376	58.9
NA12878_Allpaths^{S}	4,012	3,897	115	97.1	1,097	752	345	68.6
African^{S}	7,991	7,893	98	98.8	2,673	2,133	540	79.8

I found a correlation between N50 length of fragments and the fraction of novel SVs (R^2 = 0.44; Fig. 17). When I compared SVs of the human assemblies with the SVs by the CHM1_mapping, only small portions of SVs (~12.51 %) were shared (Tables 26 and 32). The shared portion of SVs (8.85 %) between the CHM1_PacBio_r2 and CHM1_mapping was small, and the shared portions of NA12878 assemblies were quite different (NA12878_single: 8.32 %, NA12878_Allpaths: 5.27 %). There was a correlation between the assembly quality (N50 length) and shared portion (R^2 = 0.71; Fig. 18). These results suggest that even for the same sample there was a large difference between the long-read sequence mapping and *de novo* assembly-based whole genome alignment methods.

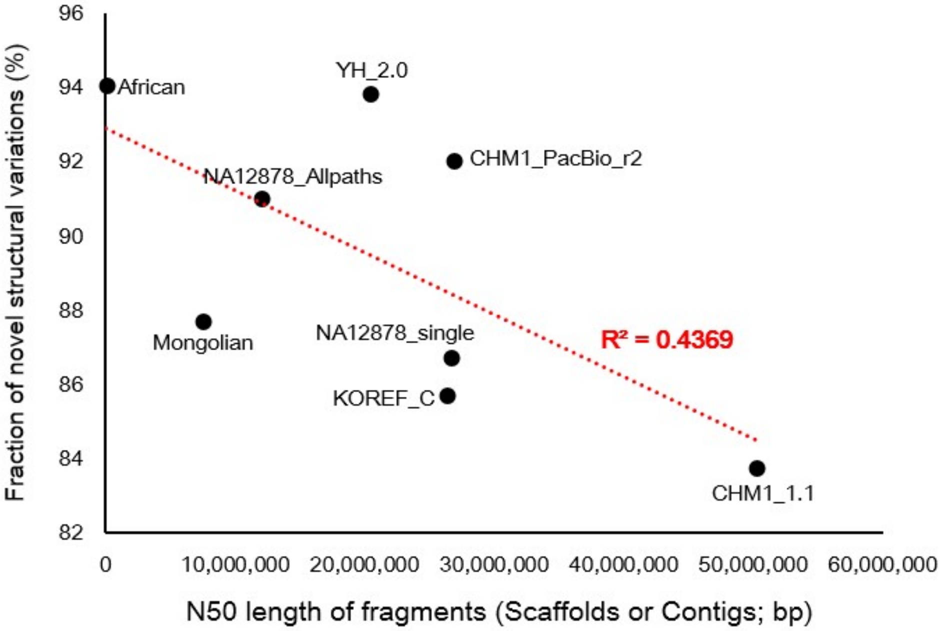


Figure 17. The correlation between N50 length of fragments (scaffolds or contigs) and fraction of novel structural variations

Table 32. Structural variations shared with CHM1 PacBio read mapping results. Major sequencing and mapping data used in the assembly are marked by superscript letters: NGS short reads, S; long reads, L; genome maps, M; indexed BAC end sequences, B.

Assembly	Total SVs (only insertions or deletions)	The number of shared SVs with the CHM1 PacBio read mapping results			
		Shared SVs	Shared Insertions	Shared Deletions	% of shared SVs
KOREF_C^{S,L,M}	9,796	955	477	478	9.75
Mongolian^{S}	12,290	834	362	472	6.79
YH_2.0^{S}	4,973	148	113	35	2.98
CHM1_PacBio_r2^{L}	3,402	301	258	43	8.85
CHM1_1.1^{S,B}	3,892	487	87	400	12.51
NA12878_single^{L,M}	4,810	400	224	176	8.32
NA12878_Allpaths^{S}	5,109	269	137	132	5.27
African^{S}	10,664	212	50	162	1.99

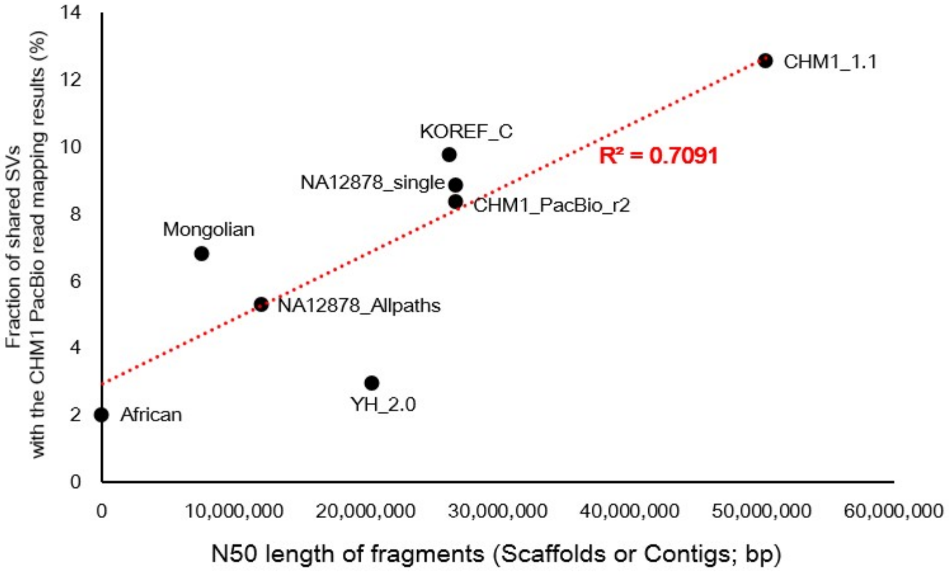


Figure 18. The correlation between N50 length of fragments and fraction of structural variations shared with the CHM1 PacBio read mapping method

Human genomes contain population-specific sequences and population stratified copy number variable regions[6,66]. Therefore, I assumed that ethnically-relevant human assemblies should share similar genome structures. To investigate the genomic structure among human assemblies, I grouped SVs that were shared by the human assemblies (Fig. 19).

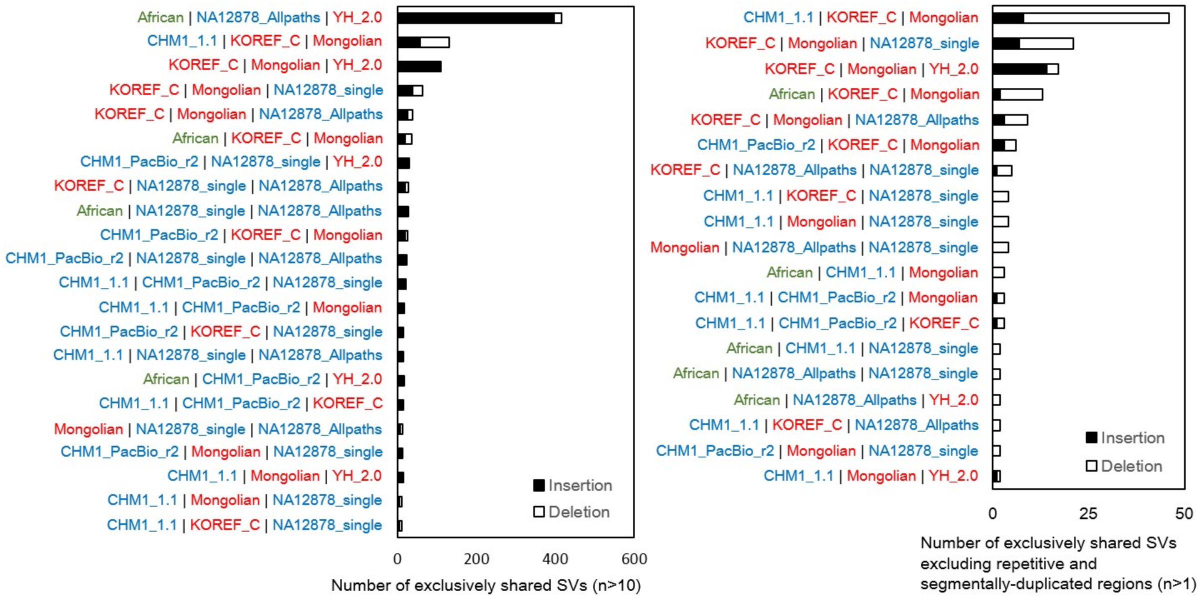


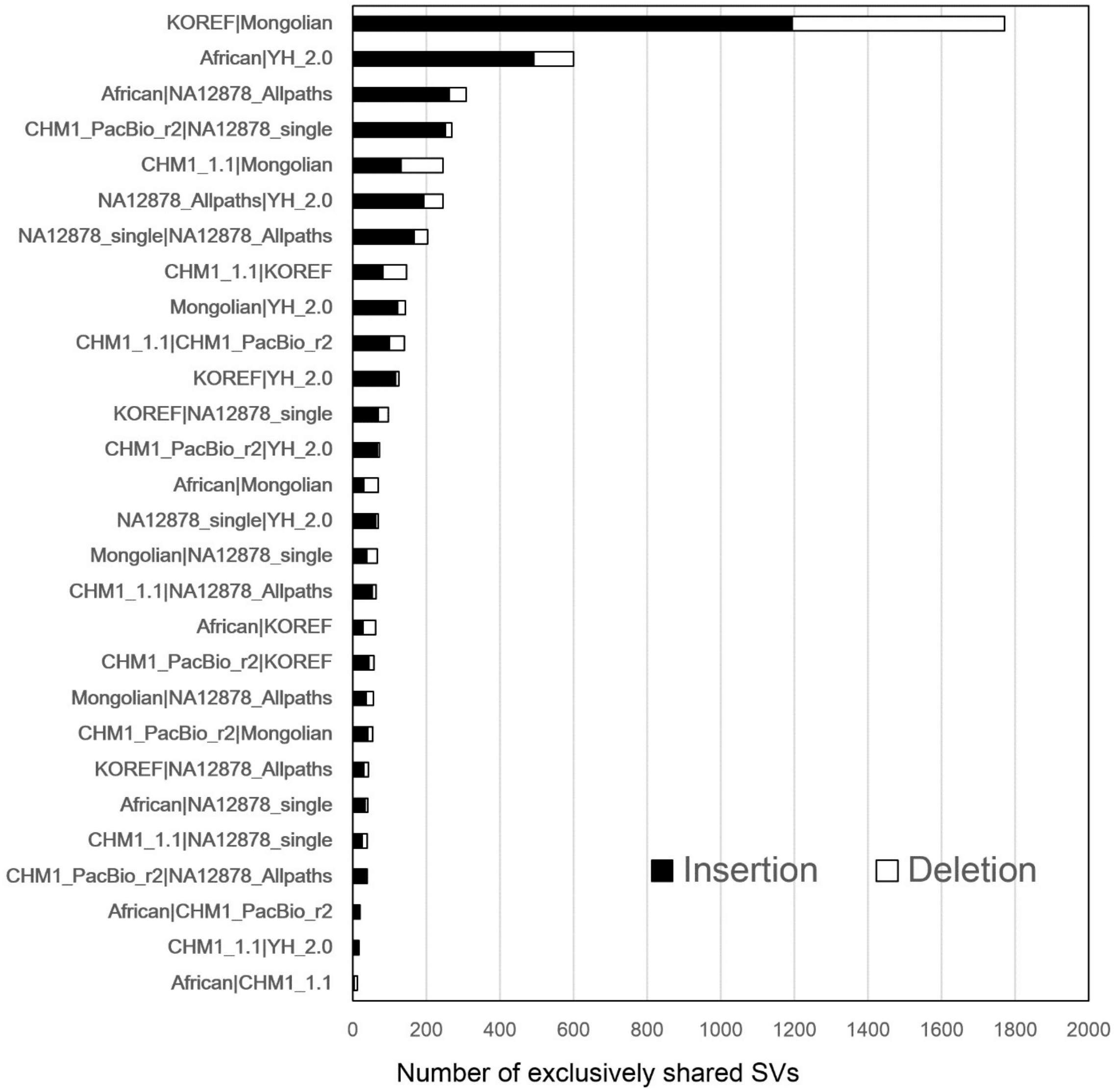
Figure 19. Exclusively shared structural variations. Structural variations shared (reciprocally 50 % covered) by only denoted assemblies were considered in this figure.

Most SVs (above 61.6 %) were assembly specific (Table 33). When I consider SVs that were shared by only two assemblies, two Asian genomes (KOREF_C and Mongolian) shared the highest number of SVs (Fig. 20). However, YH_2.0 shared only small numbers of SVs with KOREF_C and Mongolian assemblies. Notably, YH_2.0 and African genomes shared SVs abundantly, which cannot be explained by my assumption that similar ethnic genomes should have a higher genome structure similarity. CHM1_PacBio_r2 and NA12878_single, which are Caucasian assemblies using PacBio long read sequences, shared more SVs than those between the same sample's assemblies (NA12878 assemblies and CHM1 assemblies). In cases of SVs shared by only three assemblies, African, NA12878_Allpaths, and YH_2.0 had the largest number of shared SVs, whereas the three Asian genomes had smaller numbers of shared SVs (Figs. 19 and 20). However, when SVs detected in the repetitive and segmentally-duplicated regions were excluded, the three Asian assemblies had the largest number of shared insertions, whereas African, NA12878_Allpaths, and YH_2.0 shared no insertions at all (Fig. 21). These results indicate that SV identification was critically affected by the sequencing platform and assembly quality. I therefore suggest that long-read sequencing methods are necessary to improve the assembly quality and SV identification for the better characterization of genome structural differences.

Table 33. Assembly-specific structural variations. Major sequencing and mapping data used in the assembly are marked by superscript letters: NGS short reads, S; long reads, L; genome maps, M; indexed BAC end sequences, B.

Assembly	Total SVs (only insertions or deletions)	The number of assembly specific SVs	The number of shared SVs with other assemblies	The percentage of the specific SVs
KOREF_C^{S,L,M}	9,796	6,691	3,105	68.3
Mongolian^{S}	12,290	9,101	3,189	74.1
YH_2.0^{S}	4,973	3,063	1,910	61.6
CHM1_PacBio_r2^{L}	3,402	2,448	954	72.0
CHM1_1.1^{S,B}	3,892	2,800	1,092	71.9
NA12878_single^{L,M}	4,810	3,492	1,318	72.6
NA12878_Allpaths^{S}	5,109	3,787	1,322	74.1
African^{S}	10,664	8,935	1,729	83.8

a. Structural variations shared by only two assemblies



b. Structural variations shared by only three assemblies

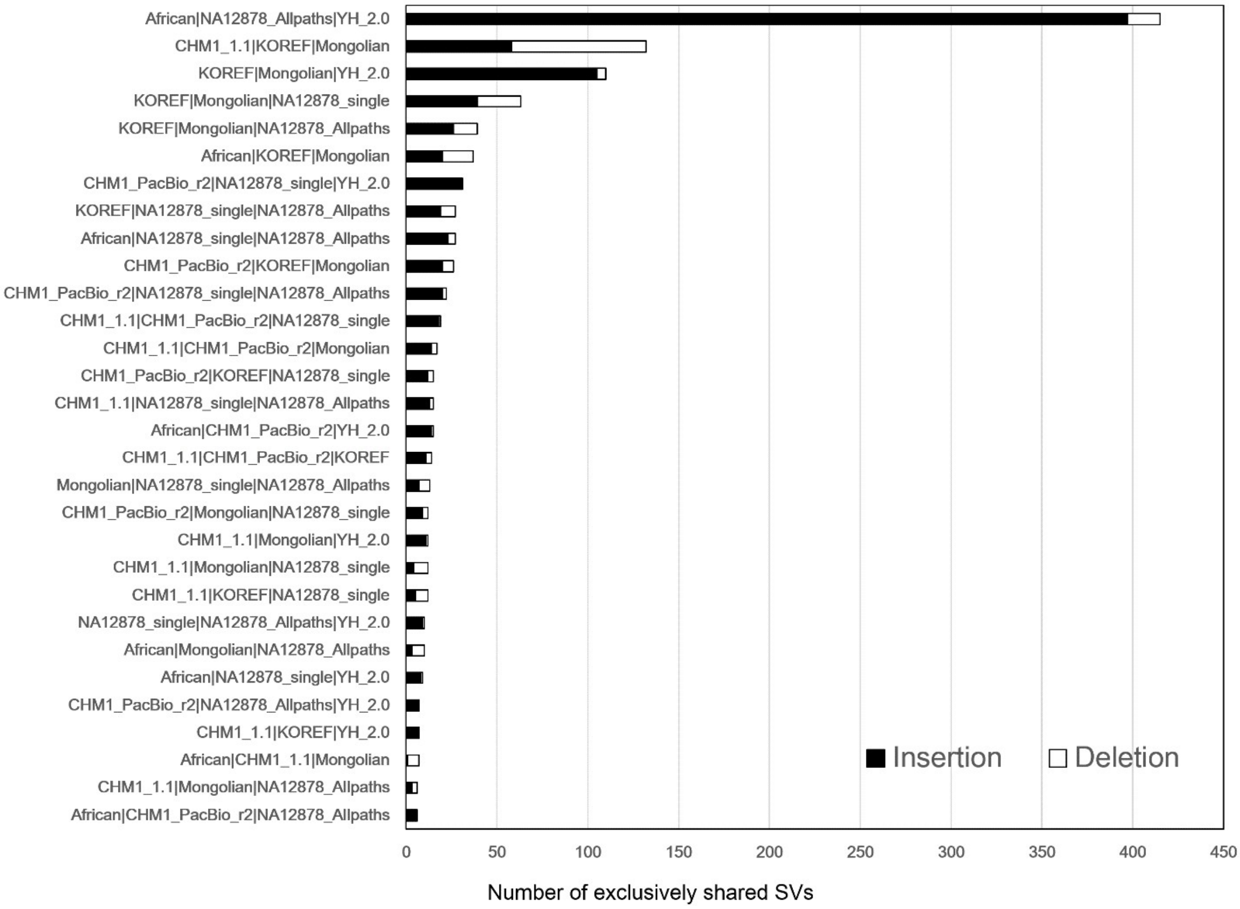
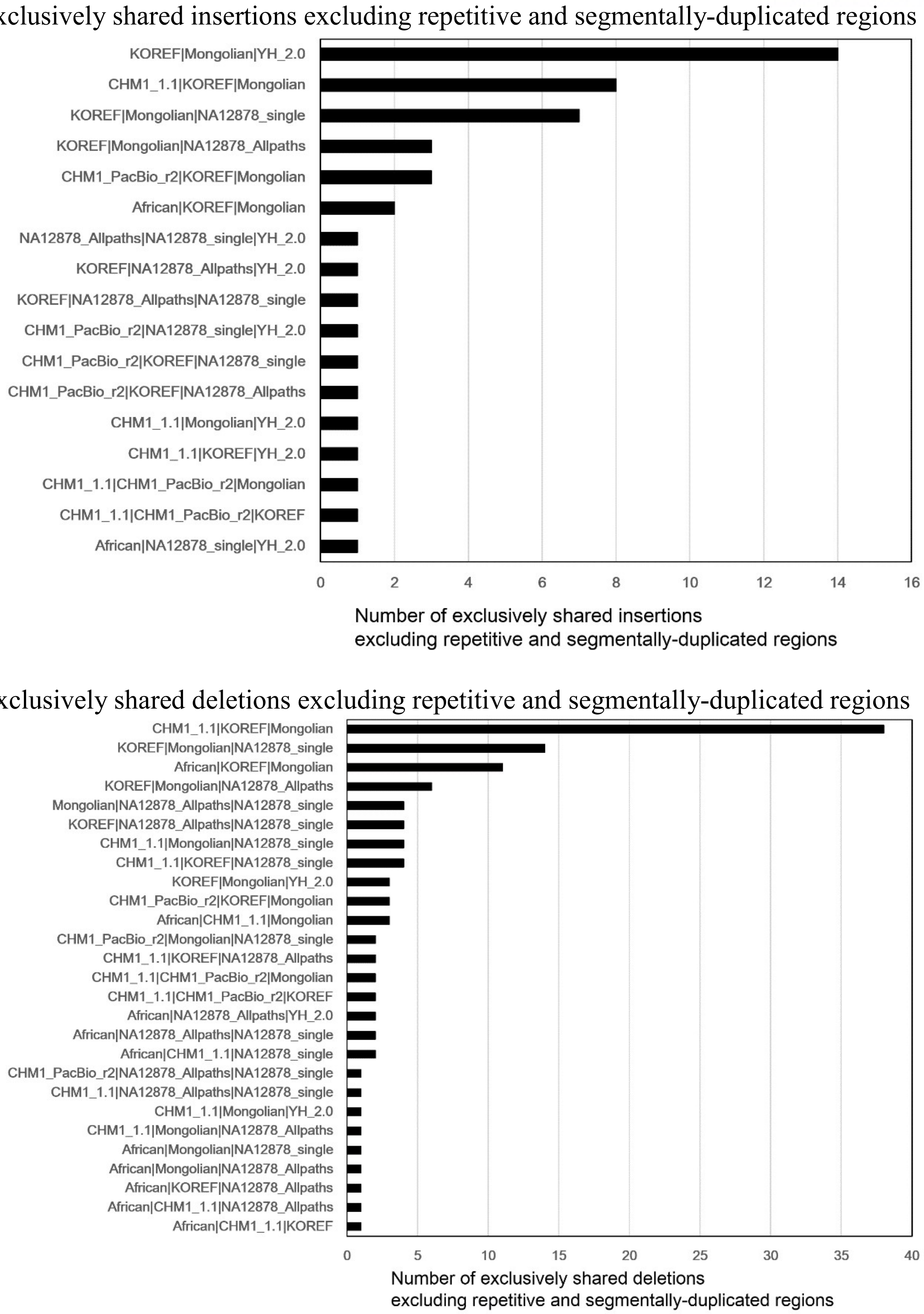


Figure 20. Exclusively shared structural variations among human assembly sets. Structural variations shared (reciprocally 50 % covered) by only denoted assemblies (y-axis: assembly sets) were considered in this figure. KOREF indicates KOREF_C. (**a**) Structural variations shared by only two assemblies. (**b**) Structural variations shared by only three assemblies. Only cases with five or more shared structural variations are shown.



a. Exclusively shared insertions excluding repetitive and segmentally-duplicated regions

b. Exclusively shared deletions excluding repetitive and segmentally-duplicated regions

Figure 21. Exclusively shared structural variations excluding repetitive and segmentally-duplicated regions. Structural variations shared by only three assemblies were considered in this figure (reciprocally 50 % covered). KOREF indicates KOREF_C. (**a**) Exclusively shared insertions. (**b**) Exclusively shared deletions.

Given these limitations, I continued to identify commonly-shared SVs by ethnic group. To do this, my colleagues and I checked S/P ratios for the SVs using the whole genome re-sequencing data from five Koreans, four East-Asians, four Caucasians, and one African, from the KPGP, 1KGP, Human Genome Diversity Project (HGDP)[67], and the Pan-Asian Population Genomics Initiative (PAPGI). First, I found one SV that was shared by all human assemblies (Fig. 22). This SV was also commonly found in re-sequencing data (13 out of the 14 re-sequencing data).

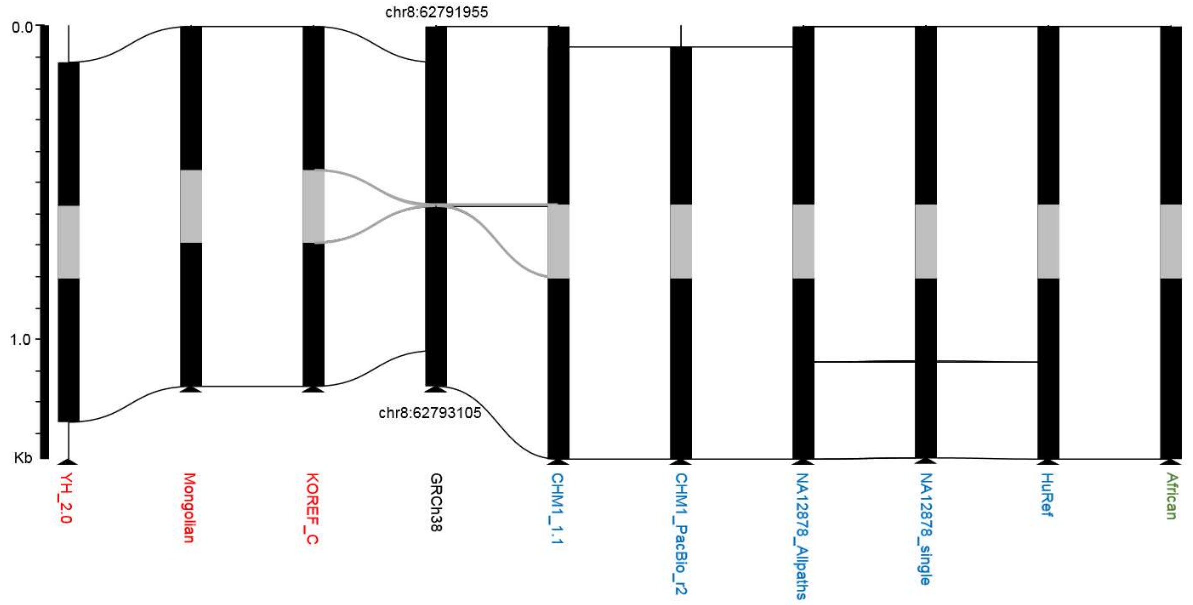


Figure 22. An example of structural variation that was shared by nine human assemblies. Gray regions denote structural differences shared among all the assemblies, and horizontal lines indicate homologous sequence regions.

Out of the 110 SVs that were shared by the three Asian assemblies, 18 were frequently found in eleven Asian genomes (one Mongolian assembly, one Chinese assembly, and nine Asian re-sequencing data) compared to ten non-Asian genomes (five non-Asian assemblies and five re-sequencing data, *P*-value <0.05, *Fisher*'s exact test; Table 34). Although the SV analysis had limitations due to the heterogeneity of sequencing platform and assembly quality, these results may indicate that the genomic structure is more similar within the same ethnic group[6,66], suggesting that ethnically-relevant reference genomes are necessary for efficiently performing large-scale comparative genomics.

Table 34. Structural variations that were frequently found only in Asian genomes

chr	SV type	GRCh38 start	GRCh38 end	Asian_ support	Asia_ not_support	Non-Asian_ support	Non-Asian_ not_support	*P*-value	Ensembl Gene	Confirmed by short/ long read alignments
chr20	Insertion	60764249	60764435	11	0	3	7	0.0010	-	Confirm
chr1	Insertion	75619372	75619500	5	6	0	10	0.023	-	Undefinable
chr1	Insertion	1565637	1565733	7	4	1	9	0.017	*SSU72*	Confirm
chr5	Insertion	96535023	96535129	5	6	0	10	0.023	*CAST*	Confirm
chr9	Insertion	86053597	86053801	9	2	3	7	0.024	*GOLM1*	Confirm
chr9	Insertion	4345943	4346583	10	1	3	7	0.0067	-	Confirm
chr6	Insertion	161000000	161000000	6	5	1	9	0.043	-	Confirm
chr11	Insertion	134000000	134000000	10	1	4	6	0.021	-	Confirm
chr4	Insertion	86343248	86343734	7	4	1	9	0.017	*MAPK10*	Confirm
chr12	Insertion	10915394	10916410	7	4	1	9	0.017	*PRH1 , PRH1-PRR4, PRR4*	Confirm
chr6	Insertion	169000000	169000000	7	4	0	10	0.0028	-	Confirm
chr6	Insertion	157000000	157000000	10	1	4	6	0.021	-	Confirm
chr11	Insertion	70951060	70951170	9	2	1	9	0.0016	*SHANK2*	Confirm
chr20	Insertion	35564449	35564597	11	0	3	7	0.0010	-	Confirm
chr6	Insertion	40655117	40655181	11	0	5	5	0.012	-	Confirm
chr7	Deletion	117000000	117000000	11	0	2	8	0.00022	-	Confirm
chr6	Deletion	161000000	161000000	8	3	1	9	0.0058	-	Confirm
chr5	Deletion	9411654	9411968	11	0	4	6	0.0039	*SEMA5A*	Confirm

3.6 Variant comparison mapped to KOREFs

Ethnicity-specific genomic sequences that are absent from the reference genome may be important for precise detection of genomic variations[22]. It is also known that the current human reference sequence contains both common and rare disease risk variants[68], and the use of the current human reference for variant identification may complicate the detection of rare disease risk alleles[5]. Using re-sequencing data on five whole genomes from each population (Caucasian, African, East-Asian, and Korean), I compared the number of variants (SNVs and small indels) detected using KOREF_S, KOREF_C, GRCh38, and consensus Asian GRCh38 (GRCh38_C, the implementation of Dewey *et al.*'s Asian major allele reference[5] but including small indels for this study; Tables 35 and 36).

Table 35. Mapping statistics of 20 individuals from different populations

Sample ID	Nation /tribe	Ethnicity	GRCh38		GRCh38_C		KOREF_S		KOREF_C	
			Mapped read depth (except 'N')	Read mapping rate (%)	Mapped read depth (except 'N')	Read mapping rate (%)	Mapped read depth (except 'N')	Read mapping rate (%)	Mapped read depth (except 'N')	Read mapping rate (%)
HGDP01286	Mandenka	African	35.39	98.64	35.39	98.55	36.78	98.78	36.70	98.78
HGDP00936	Yoruba	African	37.93	98.71	37.93	98.60	39.49	98.86	39.40	98.86
HGDP01036	San	African	37.10	98.82	37.11	98.74	38.55	98.93	38.46	98.93
HGDP00982	Mbuti	African	35.63	98.45	35.65	98.36	35.71	98.56	37.00	98.56
DNK07	Dinka	African	33.66	85.50	33.66	84.25	35.01	85.86	34.94	85.86
HGDP01076	Sardinia	Caucasian	36.51	98.56	36.51	98.45	37.85	98.72	37.78	98.72
HGDP00533	France	Caucasian	40.05	98.46	40.04	98.35	41.44	98.64	41.35	98.64
SRR622457	CEU	Caucasian	65.36	99.82	65.37	99.78	67.19	99.84	67.12	99.84
SRR622458	CEU	Caucasian	58.88	99.36	58.88	99.32	60.94	99.40	60.83	99.40
SRR622459	CEU	Caucasian	58.02	99.45	58.04	99.40	60.06	99.46	59.98	99.46
PAP-MGL0002 -U01-G	Mongolia	Asian	27.81	99.85	27.81	99.82	28.62	99.98	28.57	99.99
HGDP00775	China (Han)	Asian	32.79	98.81	32.78	98.74	33.98	98.94	33.90	98.94
HGDP01308	China (Dai)	Asian	34.26	98.86	34.26	98.78	35.41	99.01	35.34	99.01
PUB-JPN0003 -U01-G	Japan	Asian	60.79	99.97	60.80	99.96	62.96	99.98	62.87	99.99
PUB-JPN0005 -U01-G	Japan	Asian	47.25	99.96	47.25	99.94	48.97	99.98	48.87	99.98
KPGP-00120	Korea	Asian	32.50	99.97	32.51	99.94	33.49	99.99	33.42	99.99
KPGP-00121	Korea	Asian	32.19	99.97	32.19	99.95	32.94	99.99	32.89	99.99
KPGP-00122	Korea	Asian	26.33	99.97	26.33	99.95	27.32	99.99	27.28	99.99
KPGP-00124	Korea	Asian	31.17	99.97	31.17	99.94	31.98	99.99	31.91	99.99
KPGP-00117	Korea	Asian	36.61	99.91	36.62	99.53	37.57	99.94	37.49	99.94

Table 36. All variants compared to GRCh38, GRCh38_C, and KOREFs

a. Variants compared to GRCh38

Nation/tribe	Ethnicity	homozygous SNV	homozygous INDEL	heterozygous SNV	heterozygous INDEL	all variants
Mandenka	African	1,614,344	250,110	3,252,486	423,957	5,540,897
Yoruba	African	1,623,397	259,325	3,287,388	453,110	5,623,220
San	African	1,929,708	299,317	3,330,631	443,792	6,003,448
Mbuti	African	1,834,909	284,177	3,282,740	429,029	5,830,855
Dinka	African	1,640,520	254,011	3,153,108	410,603	5,458,242
Sardinia	Caucasian	1,560,599	253,087	2,507,882	337,472	4,659,040
France	Caucasian	1,512,052	244,518	2,550,429	344,208	4,651,207
CEU	Caucasian	1,495,963	243,410	2,643,275	437,506	4,820,154
CEU	Caucasian	1,517,099	245,221	2,586,786	385,858	4,734,964
CEU	Caucasian	1,483,765	237,393	2,630,607	394,254	4,746,019
Mongolia	Asian	1,602,333	232,843	2,479,567	344,493	4,659,236
China (Han)	Asian	1,650,342	254,437	2,401,103	293,025	4,598,907
China (Dai)	Asian	1,643,907	256,270	2,406,494	300,865	4,607,536
Japan	Asian	1,639,601	267,938	2,516,845	362,831	4,787,215
Japan	Asian	1,668,037	269,589	2,450,423	342,790	4,730,839
Korea	Asian	1,631,396	239,837	2,305,755	292,243	4,469,231
Korea	Asian	1,597,954	230,450	2,367,444	288,357	4,484,205
Korea	Asian	1,601,168	228,671	2,231,534	274,009	4,335,382
Korea	Asian	1,657,144	237,764	2,283,548	276,815	4,455,271
Korea	Asian	1,640,010	248,200	2,335,993	325,122	4,549,325

b. Variants compared to GRCh38_C

Nation/tribe	Ethnicity	homozygous SNV	homozygous INDEL	heterozygous SNV	heterozygous INDEL	all variants
Mandenka	African	1,211,982	243,431	3,305,587	414,358	5,175,358
Yoruba	African	1,231,018	252,663	3,345,092	443,006	5,271,779
San	African	1,516,945	292,609	3,389,637	435,794	5,634,985
Mbuti	African	1,423,658	277,114	3,336,610	420,853	5,458,235
Dinka	African	1,213,904	244,908	3,206,560	401,425	5,066,797
Sardinia	Caucasian	984,396	227,947	2,558,644	327,025	4,098,012
France	Caucasian	914,364	218,931	2,599,928	333,270	4,066,493
CEU	Caucasian	916,802	220,826	2,703,296	422,778	4,263,702
CEU	Caucasian	944,366	222,936	2,643,462	372,847	4,183,611
CEU	Caucasian	907,366	215,316	2,688,540	381,097	4,192,319
Mongolia	Asian	658,202	189,942	2,536,644	329,663	3,714,451
China (Han)	Asian	622,947	201,688	2,449,243	283,102	3,556,980
China (Dai)	Asian	622,883	203,148	2,454,664	290,356	3,571,051
Japan	Asian	624,433	214,155	2,571,845	349,498	3,759,931
Japan	Asian	651,368	215,298	2,503,848	330,550	3,701,064
Korea	Asian	621,435	189,908	2,353,181	280,468	3,444,992
Korea	Asian	581,684	181,304	2,415,107	276,944	3,455,039
Korea	Asian	585,745	178,753	2,280,669	262,940	3,308,107
Korea	Asian	630,821	187,158	2,330,619	265,688	3,414,286
Korea	Asian	625,752	197,653	2,388,942	310,568	3,522,915

c. Variants compared to KOREF_S

Nation/tribe	Ethnicity	homozygous SNV	homozygous INDEL	heterozygous SNV	heterozygous INDEL	all variants
Mandenka	African	1,899,606	271,185	3,301,289	420,873	5,892,953
Yoruba	African	1,919,941	284,415	3,334,640	449,129	5,988,125
San	African	2,188,629	317,078	3,364,703	439,159	6,309,569
Mbuti	African	2,100,096	301,653	3,325,523	425,646	6,152,918
Dinka	African	1,887,255	270,418	3,168,465	406,489	5,732,627
Sardinia	Caucasian	1,728,462	257,330	2,560,749	334,792	4,881,333
France	Caucasian	1,664,474	247,083	2,628,682	343,003	4,883,242
CEU	Caucasian	1,679,236	263,547	2,719,341	431,376	5,093,500
CEU	Caucasian	1,650,211	258,262	2,678,162	384,000	4,970,635
CEU	Caucasian	1,629,051	252,415	2,708,918	389,169	4,979,553
Mongolia	Asian	1,433,902	187,832	2,499,056	331,562	4,452,352
China (Han)	Asian	1,408,738	196,399	2,451,061	288,071	4,344,269
China (Dai)	Asian	1,431,892	203,261	2,458,007	295,599	4,388,759
Japan	Asian	1,399,464	219,953	2,575,275	351,982	4,546,674
Japan	Asian	1,407,595	216,866	2,514,712	334,576	4,473,749
Korea	Asian	1,411,971	188,996	2,377,237	285,905	4,264,109
Korea	Asian	1,383,188	180,090	2,413,482	279,755	4,256,515
Korea	Asian	1,388,544	177,490	2,282,391	265,137	4,113,562
Korea	Asian	1,419,583	184,724	2,350,290	270,957	4,225,554
Korea	Asian	1,415,274	201,750	2,413,606	316,357	4,346,987

d. Variants compared to KOREF_C

Nation/tribe	Ethnicity	homozygous SNV	homozygous INDEL	heterozygous SNV	heterozygous INDEL	all variants
Mandenka	African	1,212,596	206,550	3,292,369	421,829	5,133,344
Yoruba	African	1,237,976	219,861	3,323,619	450,244	5,231,700
San	African	1,505,723	254,670	3,356,014	440,571	5,556,978
Mbuti	African	1,420,095	238,949	3,316,613	427,357	5,403,014
Dinka	African	1,209,682	206,620	3,160,340	407,555	4,984,197
Sardinia	Caucasian	993,587	183,953	2,552,486	335,486	4,065,512
France	Caucasian	922,712	172,431	2,616,202	343,503	4,054,848
CEU	Caucasian	926,900	185,792	2,701,042	431,538	4,245,272
CEU	Caucasian	927,687	182,975	2,649,135	383,506	4,143,303
CEU	Caucasian	903,211	176,639	2,678,229	388,519	4,146,598
Mongolia	Asian	652,322	114,456	2,499,555	328,313	3,594,646
China (Han)	Asian	616,323	115,941	2,441,811	287,953	3,462,028
China (Dai)	Asian	635,841	121,720	2,449,488	295,368	3,502,417
Japan	Asian	576,063	127,970	2,466,876	339,653	3,510,562
Japan	Asian	573,960	123,705	2,450,926	330,414	3,479,005
Korea	Asian	583,492	105,543	2,377,620	284,977	3,351,632
Korea	Asian	554,680	98,501	2,414,149	278,757	3,346,087
Korea	Asian	557,668	96,310	2,283,849	264,161	3,201,988
Korea	Asian	593,228	102,429	2,349,328	269,759	3,314,744
Korea	Asian	590,524	116,896	2,408,198	316,139	3,431,757

I found that the number of variants was considerably different, depending on the reference used. Variant numbers of all individuals (Caucasian, African, and East-Asian) decreased when KOREF_C was used as a reference. However, because the lower number of actual bases (non-gapped) in KOREFs (KOREF_S and KOREF_C) could affect the accuracy of genotype reconstruction, I compared variant numbers only within the regions shared by KOREFs, GRCh38, and GRCh38_C (Table 37).

Table 37. Variants within the regions shared by GRCh38, GRCh38_C, and KOREFs

a. Variants compared to GRCh38

Nation/tribe	Ethnicity	homozygous SNV	homozygous INDEL	heterozygous SNV	heterozygous INDEL	all variants
Mandenka	African	1,537,873	243,192	2,984,279	410,079	5,175,423
Yoruba	African	1,546,651	252,162	3,008,067	437,903	5,244,783
San	African	1,841,485	291,111	3,045,419	428,629	5,606,644
Mbuti	African	1,753,016	276,634	3,007,386	414,524	5,451,560
Dinka	African	1,567,818	247,302	2,879,582	396,375	5,091,077
Sardinia	Caucasian	1,480,532	245,823	2,247,876	323,930	4,298,161
France	Caucasian	1,437,117	237,741	2,295,762	331,586	4,302,206
CEU	Caucasian	1,413,798	235,749	2,366,980	421,059	4,437,586
CEU	Caucasian	1,435,451	237,311	2,304,493	370,296	4,347,551
CEU	Caucasian	1,406,198	230,185	2,357,944	379,214	4,373,541
Mongolia	Asian	1,523,758	225,799	2,231,015	330,974	4,311,546
China (Han)	Asian	1,575,375	247,925	2,150,845	281,443	4,255,588
China (Dai)	Asian	1,567,327	249,339	2,158,728	289,153	4,264,547
Japan	Asian	1,555,213	259,770	2,233,166	347,895	4,396,044
Japan	Asian	1,585,887	261,995	2,171,749	328,495	4,348,126
Korea	Asian	1,555,627	233,655	2,080,105	281,432	4,150,819
Korea	Asian	1,525,401	224,823	2,137,357	277,678	4,165,259
Korea	Asian	1,532,866	223,272	2,027,996	263,914	4,048,048
Korea	Asian	1,579,741	231,933	2,053,255	266,294	4,131,223
Korea	Asian	1,564,621	241,984	2,102,604	313,656	4,222,865

b. Variants compared to GRCh38_C

Nation/tribe	Ethnicity	homozygous SNV	homozygous INDEL	heterozygous SNV	heterozygous INDEL	all variants
Mandenka	African	1,140,674	234,795	3,029,895	398,762	4,804,126
Yoruba	African	1,158,743	243,701	3,057,738	426,039	4,886,221
San	African	1,435,562	282,503	3,094,711	419,031	5,231,807
Mbuti	African	1,347,609	267,723	3,053,888	404,815	5,074,035
Dinka	African	1,145,701	236,530	2,924,800	385,595	4,692,626
Sardinia	Caucasian	912,707	219,775	2,290,437	311,798	3,734,717
France	Caucasian	846,238	211,278	2,337,491	318,846	3,713,853
CEU	Caucasian	842,235	212,174	2,417,190	404,301	3,875,900
CEU	Caucasian	869,407	214,106	2,352,705	355,499	3,791,717
CEU	Caucasian	836,982	207,081	2,407,010	364,228	3,815,301
Mongolia	Asian	595,351	182,774	2,279,226	314,497	3,371,848
China (Han)	Asian	565,008	195,176	2,191,491	270,050	3,221,725
China (Dai)	Asian	562,927	196,225	2,199,261	277,139	3,235,552
Japan	Asian	559,044	206,148	2,278,808	332,765	3,376,765
Japan	Asian	586,127	207,706	2,215,849	314,543	3,324,225
Korea	Asian	562,965	183,668	2,120,913	268,212	3,135,758
Korea	Asian	526,445	175,743	2,177,998	264,832	3,145,018
Korea	Asian	535,049	173,397	2,070,424	251,501	3,030,371
Korea	Asian	571,978	181,302	2,093,657	253,797	3,100,734
Korea	Asian	566,370	191,396	2,148,469	297,654	3,203,889

c. Variants compared to KOREF_S

Nation/tribe	Ethnicity	homozygous SNV	homozygous INDEL	heterozygous SNV	heterozygous INDEL	all variants
Mandenka	African	1,838,584	261,031	3,193,260	411,299	5,704,174
Yoruba	African	1,855,419	273,126	3,226,650	438,830	5,794,025
San	African	2,121,591	305,602	3,254,073	429,336	6,110,602
Mbuti	African	2,035,747	290,769	3,217,689	416,122	5,960,327
Dinka	African	1,827,074	260,185	3,067,251	397,427	5,551,937
Sardinia	Caucasian	1,663,730	246,752	2,465,532	326,422	4,702,436
France	Caucasian	1,603,905	237,107	2,531,956	334,458	4,707,426
CEU	Caucasian	1,614,403	250,863	2,616,108	420,811	4,902,185
CEU	Caucasian	1,592,455	245,827	2,572,676	373,934	4,784,892
CEU	Caucasian	1,571,221	239,991	2,608,477	379,413	4,799,102
Mongolia	Asian	1,378,650	179,021	2,413,749	323,974	4,295,394
China (Han)	Asian	1,356,821	187,835	2,359,226	280,404	4,184,286
China (Dai)	Asian	1,377,144	194,272	2,367,079	288,021	4,226,516
Japan	Asian	1,344,144	209,132	2,470,649	342,741	4,366,666
Japan	Asian	1,356,151	207,138	2,413,387	325,581	4,302,257
Korea	Asian	1,359,667	181,002	2,292,465	278,932	4,112,066
Korea	Asian	1,331,466	172,387	2,330,028	273,137	4,107,018
Korea	Asian	1,338,598	170,149	2,208,679	259,363	3,976,789
Korea	Asian	1,366,749	176,853	2,265,556	264,238	4,073,396
Korea	Asian	1,361,208	192,930	2,322,623	308,795	4,185,556

d. Variants compared to KOREF_C

Nation/tribe	Ethnicity	homozygous SNV	homozygous INDEL	heterozygous SNV	heterozygous INDEL	all variants
Mandenka	African	1,169,556	199,035	3,177,568	412,169	4,958,328
Yoruba	African	1,192,463	211,369	3,208,543	439,883	5,052,258
San	African	1,457,870	245,947	3,238,385	430,637	5,372,839
Mbuti	African	1,373,409	230,635	3,201,885	417,685	5,223,614
Dinka	African	1,167,547	199,131	3,052,556	398,438	4,817,672
Sardinia	Caucasian	949,611	176,166	2,449,771	327,061	3,902,609
France	Caucasian	883,063	165,329	2,512,690	334,951	3,896,033
CEU	Caucasian	883,635	176,306	2,590,930	420,995	4,071,866
CEU	Caucasian	888,351	173,695	2,537,367	373,554	3,972,967
CEU	Caucasian	864,194	167,370	2,572,881	378,897	3,983,342
Mongolia	Asian	617,152	108,283	2,406,654	320,873	3,452,962
China (Han)	Asian	584,392	110,287	2,342,911	280,331	3,317,921
China (Dai)	Asian	600,804	115,643	2,350,939	287,795	3,355,181
Japan	Asian	542,593	120,860	2,360,342	330,763	3,354,558
Japan	Asian	543,428	117,352	2,347,537	321,787	3,330,104
Korea	Asian	555,357	100,726	2,285,893	278,057	3,220,033
Korea	Asian	527,941	94,040	2,323,498	272,173	3,217,652
Korea	Asian	530,235	92,058	2,203,812	258,373	3,084,478
Korea	Asian	563,836	97,792	2,257,652	262,967	3,182,247
Korea	Asian	560,149	111,283	2,310,227	308,538	3,290,197

As expected, the numbers of homozygous variants from all the Asian genomes (two Chinese, two Japanese, one Mongolian, and five Korean) decreased largely (35.5 % of SNVs and 43.9 % of indels remained) when KOREF_C was used as a reference compared to GRCh38 (Fig. 23a and 23b); on the contrary, the numbers of homozygous variants from Caucasian and African genomes decreased little.

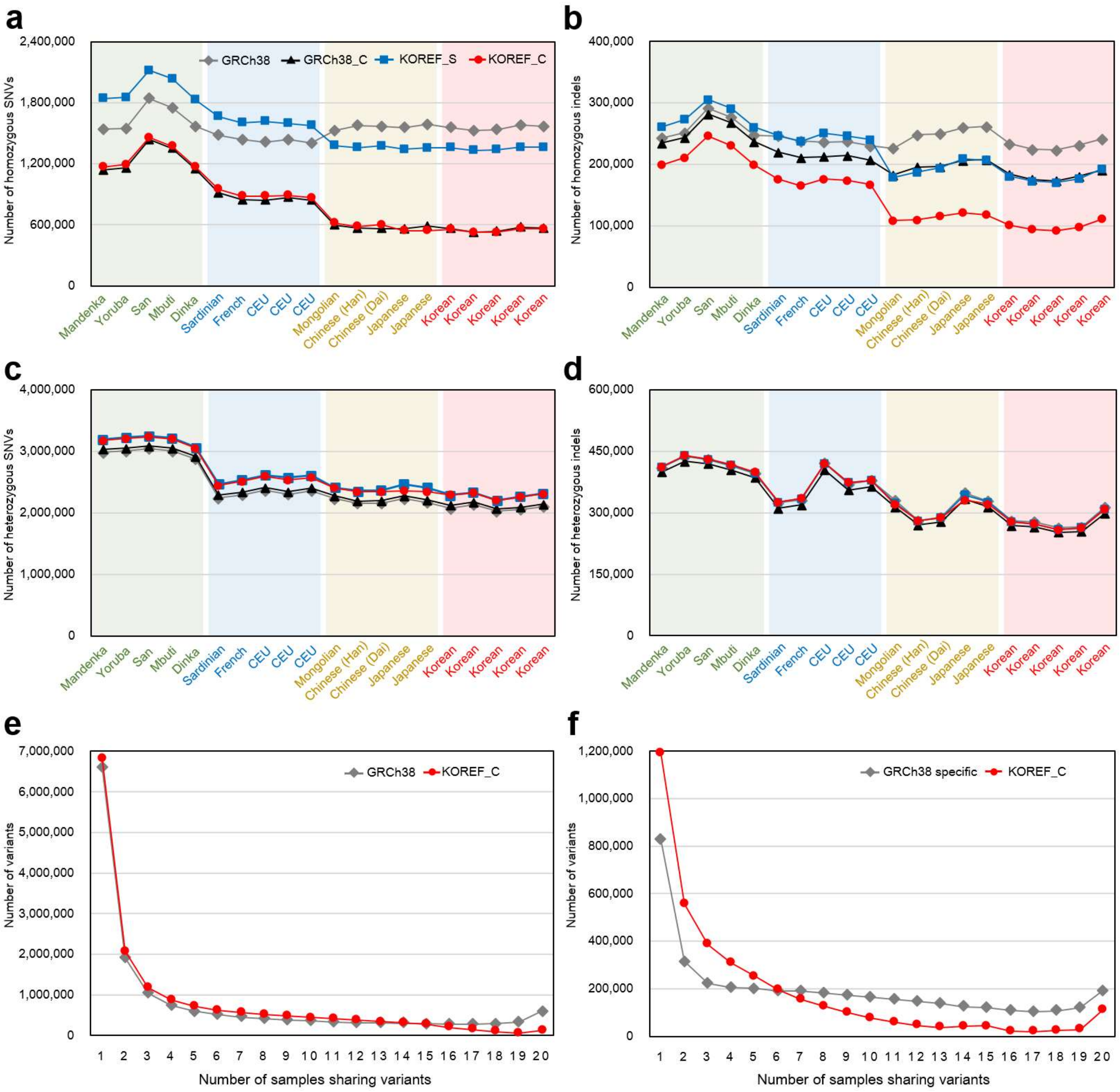


Figure 23. Variants difference depending on the reference genome. Variants (SNVs and small indels) numbers within the regions shared by KOREFs, GRCh38, and GRCh38_C were compared using whole genome re-sequencing data from three different ethnic groups (Africans: Mandenka, Yoruba, San, Mbuti, and Dinka; Caucasians: Sardinian, French, and three CEPH/Utah (CEU); East-Asians: Mongolian, two Chinese, two Japanese, and five Koreans). (**a**) Number of homozygous SNVs. (**b**) Number of homozygous small indels. (**c**) Number of heterozygous SNVs. (**d**) Number of heterozygous small indels. (**e**) The number of variants (referenced by GRCh38 and KOREF_C) at different levels of sharedness. (**f**) The number of reference-specific variants at different levels of sharedness.

In cases of homozygous SNVs, a similar pattern was observed between GRCh38_C and GRCh38. However, the numbers of homozygous indels when using GRCh38_C as a reference were higher than when using KOREF_C as a reference. I speculate that this is because fewer common indels were substituted for GRCh38_C when compared to KOREF_C due to low sequencing depths of 1KGP data. The numbers of homozygous variants found in non-Korean Asians were similar to those found among Koreans, suggesting that KOREFs can be used for other East-Asian genomes. On the other hand, the numbers of heterozygous SNVs were slightly higher in KOREFs, which is consistent with the mapping result of the CHM1 re-sequencing data as described above (Fig. 13). However, I confirmed that the numbers of heterozygous SNVs were similar when restricted the analysis to non-repetitive regions. The numbers of heterozygous indels were also largely constant regardless of reference used (Fig. 23c and 23d).

Focusing on differently called variants (variants found in GRCh38 but not found in KOREF_C, and vice versa), I found that there were differences in the number of variants among populations (i.e., population stratification in terms of variant number). The differences of variants among populations were more prominent when using KOREF_C specifically called variants (Table 38).

Table 38. Differently called variants between KOREF_C and GRCh38

Re-sequenced genome	KOREF_C					GRCh38				
	Total variants	Linkable variants by lift-over	Commonly called variants	Specifically called variants	% of known (dbSNP 144)	Total variants	Lift-overed variants	Commonly called variants	Specifically called variants	% of known (dbSNP 144)
HGDP01286	5,133,344	4,817,523	3,724,661	1,092,862	62.84	5,540,897	5,092,299	3,724,661	1,367,638	90.89
HGDP00936	5,231,700	4,906,223	3,777,871	1,128,352	62.44	5,623,220	5,157,411	3,777,871	1,379,540	90.57
HGDP01036	5,556,978	5,225,218	4,059,097	1,166,121	63.98	6,003,448	5,515,678	4,059,097	1,456,581	89.91
HGDP00982	5,403,014	5,080,484	3,937,532	1,142,952	63.73	5,830,855	5,364,279	3,937,532	1,426,747	90.23
DNK07	4,984,197	4,682,755	3,620,328	1,062,427	63.67	5,458,242	5,023,822	3,620,328	1,403,494	90.59
HGDP01076	4,065,512	3,764,205	2,749,152	1,015,053	60.13	4,659,040	4,216,533	2,749,152	1,467,381	91.85
HGDP00533	4,054,848	3,760,102	2,775,122	984,980	58.46	4,651,207	4,226,021	2,775,122	1,450,899	92.23
SRR622457	4,245,272	3,932,305	2,892,034	1,040,271	56.67	4,820,154	4,353,229	2,892,034	1,461,195	91.78
SRR622458	4,143,303	3,835,419	2,806,641	1,028,778	57.83	4,734,964	4,268,877	2,806,641	1,462,236	91.98
SRR622459	4,146,598	3,850,185	2,857,748	992,437	58.25	4,746,019	4,295,574	2,857,748	1,437,826	92.25
PAP-MGL0002-U01-G	3,594,646	3,321,154	2,581,981	739,173	50.98	4,659,236	4,234,280	2,581,981	1,652,299	92.94
HGDP00775	3,462,028	3,191,805	2,472,583	719,222	48.40	4,598,907	4,182,000	2,472,583	1,709,417	93.33
HGDP01308	3,502,417	3,228,298	2,487,426	740,872	48.95	4,607,536	4,190,235	2,487,426	1,702,809	92.78
PUB-JPN0003-U01-G	3,510,562	3,228,695	2,492,415	736,280	46.55	4,787,215	4,304,660	2,492,415	1,812,245	92.52
PUB-JPN0005-U01-G	3,479,005	3,204,624	2,483,055	721,569	47.18	4,730,839	4,261,777	2,483,055	1,778,722	92.66
KPGP-00120	3,351,632	3,104,933	2,406,100	698,833	49.18	4,469,231	4,094,983	2,406,100	1,688,883	93.17
KPGP-00121	3,346,087	3,106,893	2,446,729	660,164	49.03	4,484,205	4,101,486	2,446,729	1,654,757	93.55
KPGP-00122	3,201,988	2,982,262	2,338,187	644,075	51.35	4,335,382	3,990,411	2,338,187	1,652,224	94.02
KPGP-00124	3,314,744	3,068,830	2,379,991	688,839	49.61	4,455,271	4,068,391	2,379,991	1,688,400	93.37
KPGP-00117	3,431,757	3,169,212	2,439,378	729,834	49.14	4,549,325	4,154,657	2,439,378	1,715,279	92.88

The number of commonly shared KOREF_C called variants (> 6 individuals) in the 20 whole genomes was much smaller, whereas the number of less common KOREF_C called variants, including individual-specific ones, was higher (Fig. 23e and 23f). Also, the number of KOREF_C specifically called variants was considerably lower in the ten Asians than those in the ten non-Asians. These results reflect the consensus variants components of KOREF_C and also confirm that GRCh38 lacks Asian specific sequences[5]. The majority (92.3 %) of the GRCh38 specifically called variants were found in dbSNP[56] (Table 38), whereas a smaller fraction (56.17 %) of the KOREF_C specifically called variants were defined as known. When variants in repetitive and segmentally-duplicated regions were excluded, a much larger fraction (86.21 %) of the KOREF_C specifically called variants were known (Table 39), indicating that the majority of novel variants found in KOREF_C was caused by the incompleteness of repetitive and segmentally-duplicated regions. Therefore, I conclude that although KOREFs have an advantage for efficient variant detection for the same ethnic genomes, KOREFs need to be improved using longer sequence reads to reconstruct genotypes properly.

Table 39. Differently called variants excluding repetitive and segmentally-duplicated regions

	KOREF_C				GRCh38			
Re-sequenced genome	Specifically called variants	Variants excluding repetitive and segmentally -duplicated regions	Variants found in dbSNP 144	% of known (dbSNP 144)	Specifically called variants	Variants excluding repetitive and segmentally -duplicated regions	Variants found in dbSNP 144	% of known (dbSNP 144)
HGDP01286	1,092,862	299,091	265,979	88.93	1,367,638	539,509	513,196	95.12
HGDP00936	1,128,352	306,275	270,824	88.43	1,379,540	540,882	513,927	95.02
HGDP01036	1,166,121	329,655	293,693	89.09	1,456,581	574,635	543,375	94.56
HGDP00982	1,142,952	317,551	283,225	89.19	1,426,747	562,818	532,827	94.67
DNK07	1,062,427	293,307	261,979	89.32	1,403,494	549,907	522,865	95.08
HGDP01076	1,015,053	263,759	231,572	87.80	1,467,381	581,224	557,298	95.88
HGDP00533	984,980	244,711	213,247	87.14	1,450,899	580,942	557,868	96.03
SRR622457	1,040,271	254,313	219,226	86.20	1,461,195	577,047	552,595	95.76
SRR622458	1,028,778	250,068	218,577	87.41	1,462,236	574,845	552,937	96.19
SRR622459	992,437	246,130	215,322	87.48	1,437,826	570,444	548,299	96.12
PAP-MGL0002-U01-G	739,173	150,497	125,793	83.59	1,652,299	671,208	646,308	96.29
HGDP00775	719,222	137,325	112,841	82.17	1,709,417	699,766	675,676	96.56
HGDP01308	740,872	144,349	119,116	82.52	1,702,809	692,124	666,145	96.25
PUB-JPN0003-U01-G	736,280	140,793	111,647	79.30	1,812,245	730,151	700,294	95.91
PUB-JPN0005-U01-G	721,569	139,637	111,743	80.02	1,778,722	717,353	689,257	96.08
KPGP-00120	698,833	137,357	113,773	82.83	1,688,883	693,106	667,117	96.25
KPGP-00121	660,164	132,394	109,914	83.02	1,654,757	686,470	661,936	96.43
KPGP-00122	644,075	133,855	112,752	84.23	1,652,224	686,656	662,991	96.55
KPGP-00124	688,839	136,820	114,290	83.53	1,688,400	695,874	670,298	96.32
KPGP-00117	729,834	145,430	118,223	81.29	1,715,279	701,389	674,777	96.21

Additionally, I found that the number of variants identified following substitution in the reference with the dominant variant (KOREF_S vs. KOREF_C) is much higher than the change caused by the ethnicity difference (KOREF_S vs. GRCh38; Fig. 23a and 23b). Also, the East-Asians' homozygous variant number decreased only slightly when the KOREF_S was used, compared to GRCh38 (87.0 % of homozygous SNVs and 77.9 % of homozygous indels remained), while it was greatly decreased when KOREF_C was used (36.1 % of homozygous SNVs and 44.5 % of homozygous indels remained). On the other hand, the number of non-East Asians' homozygous variants increased when the KOREF_S was used, compared to when GRCh38 was used. These results indicate that, at the whole genome variation level, intra-population variation is higher than the inter-population variation in terms of number of variants, supporting the notion that *Homo sapiens* is one population within one species with no genomically significant subspecies.

3.7 Ethnicity-specific reference and functional markers

I also found that depending on the reference used, different numbers of non-synonymous SNVs (nsSNVs) and small indels were found in genic regions (Tables 40 and 41). With the aforementioned ten East-Asian whole genomes, the number of homozygous nsSNVs (from 3,644 to 1,280 on average) and indels (from 95 to 40 on average) decreased most when using KOREF_C as a reference instead of GRCh38; whereas a smaller decrease was observed in the five Caucasians (nsSNVs from 3,467 to 2,098; indels from 89 to 65) and five Africans (nsSNVs from 4,216 to 3,007; indels from 134 to 109).

Table 40. Variant in genic regions compared to GRCh38 and KOREF_C

a. The number of variants found in genic regions compared to GRCh38

Ethnicity / Sample ID		nsSNV		small indels			
				Frame shift		Indels in codon (x3)	
		Homozygous	Heterozygous	Homozygous	Heterozygous	Homozygous	Heterozygous
African	HGDP01286	3,772	8,356	35	75	92	128
	HGDP00936	3,840	8,350	30	96	96	140
	HGDP01036	4,387	8,580	33	94	108	127
	HGDP00982	4,439	8,518	34	81	96	130
	DNK07	3,885	8,059	37	98	77	123
Caucasian	HGDP01076	3,584	6,607	29	66	65	102
	HGDP00533	3,466	6,717	23	73	73	120
	SRR622457	3,498	6,804	38	43	58	106
	SRR622458	3,374	6,567	29	64	48	72
	SRR622459	3,412	6,505	38	51	46	76
Asian	PAP-MGL0002-U01-G	3,651	6,893	25	64	75	122
	HGDP00775	3,769	6,207	33	55	83	108
	HGDP01308	3,683	6,342	28	67	82	94
	PUB-JPN0003-U01-G	3,705	6,710	31	68	85	113
	PUB-JPN0005-U01-G	3,823	6,648	32	75	88	114
	KPGP-00120	3,542	5,755	32	50	56	68
	KPGP-00121	3,525	5,595	28	46	44	61
	KPGP-00122	3,517	5,398	26	41	40	50
	KPGP-00124	3,550	5,616	27	52	54	68
	KPGP-00117	3,679	5,807	26	51	54	71

b. The number of variants found in genic regions compared to KOREF_C

Ethnicity / Sample ID		nsSNV		small indels			
				Frame shift		Indels in codon (x3)	
		Homozygous	Heterozygous	Homozygous	Heterozygous	Homozygous	Heterozygous
African	HGDP01286	2,731	7,999	35	93	71	130
	HGDP00936	2,863	8,039	28	111	85	144
	HGDP01036	3,339	8,141	37	102	88	130
	HGDP00982	3,352	8,071	33	99	83	128
	DNK07	2,751	7,549	26	107	57	132
Caucasian	HGDP01076	2,237	6,203	21	79	49	107
	HGDP00533	2,060	6,402	16	95	45	129
	SRR622457	2,070	6,436	33	58	38	95
	SRR622458	2,032	6,177	23	84	37	80
	SRR622459	2,091	6,129	25	69	37	78
Asian	PAP-MGL0002-U01-G	1,459	6,542	15	80	37	114
	HGDP00775	1,352	5,850	11	74	40	109
	HGDP01308	1,429	5,966	9	83	31	98
	PUB-JPN0003-U01-G	1,363	6,081	14	80	37	112
	PUB-JPN0005-U01-G	1,337	6,362	24	92	28	119
	KPGP-00120	1,188	5,301	10	57	24	68
	KPGP-00121	1,100	5,286	11	56	18	62
	KPGP-00122	1,151	4,983	9	49	14	52
	KPGP-00124	1,285	5,177	8	56	24	67
	KPGP-00117	1,131	5,373	9	51	24	68

Table 41. The number of genes with homozygous variants

Ethnicity / Sample ID		GRCh38			KOREF_C		
		nsSNV	small indel	total	nsSNV	small indel	total
African	HGDP01286	2,669	123	2,742	1,961	100	2,016
	HGDP00936	2,688	117	2,756	2,055	106	2,116
	HGDP01036	3,045	138	3,128	2,319	124	2,393
	HGDP00982	3,012	128	3,083	2,278	115	2,339
	DNK07	2,687	110	2,756	1,946	77	1,998
Caucasian	HGDP01076	2,428	92	2,481	1,503	68	1,546
	HGDP00533	2,374	92	2,435	1,454	54	1,487
	SRR622457	2,388	93	2,440	1,416	64	1,449
	SRR622458	2,376	76	2,418	1,424	58	1,464
	SRR622459	2,335	84	2,382	1,469	59	1,508
Asian	PAP-MGL0002	2,508	100	2,568	1,016	50	1,052
	HGDP00775	2,569	115	2,631	915	48	946
	HGDP01308	2,515	103	2,579	987	38	1,009
	PUB-JPN0003	2,552	112	2,622	933	50	965
	PUB-JPN0005	2,599	115	2,671	913	41	939
	KPGP-00120	2,446	88	2,492	847	33	864
	KPGP-00121	2,440	71	2,477	791	29	808
	KPGP-00122	2,435	66	2,470	837	22	849
	KPGP-00124	2,470	81	2,515	870	32	888
	KPGP-00117	2,521	80	2,563	817	33	838

When KOREF_C was used as the reference, predicted functionally altered (or damaged) genes by the homozygous variants also decreased the most among the East-Asians (East Asians, from 490 to 246 on average; Caucasians, from 448 to 362; Africans, from 448 to 415; Table 42).

Table 42. Predicted functionally altered genes by homozygous variants

Ethnicity / Sample ID		GRCh38			KOREF_C		
		nsSNV	small indel	total	nsSNV	small indel	total
African	HGDP01286	368	50	412	336	41	374
	HGDP00936	380	44	413	354	40	384
	HGDP01036	438	48	482	431	47	469
	HGDP00982	442	48	479	426	43	461
	DNK07	416	49	452	359	33	385
Caucasian	HGDP01076	432	45	468	362	29	387
	HGDP00533	404	36	434	327	20	344
	SRR622457	418	50	455	321	36	347
	SRR622458	412	39	442	317	29	342
	SRR622459	397	50	441	362	33	392
Asian	PAP-MGL0002	434	41	473	236	18	254
	HGDP00775	478	47	516	241	15	254
	HGDP01308	454	48	497	244	16	260
	PUB-JPN0003	433	42	469	222	16	236
	PUB-JPN0005	449	47	493	203	21	223
	KPGP-00120	458	50	500	244	18	259
	KPGP-00121	445	39	476	215	16	227
	KPGP-00122	468	40	501	247	12	256
	KPGP-00124	456	45	498	245	17	262
	KPGP-00117	436	42	475	215	18	232

Notably, in the ten East-Asians, the functionally altered genes, which were found only against GRCh38 but not KOREF_C, were enriched in several disease terms (myocardial infarction, hypertension, and genetic predisposition to disease), and olfactory and taste transduction pathways (Tables 43 and 44). Additionally, 13 nsSNVs, which are known as disease- and phenotype-associated variants, were called against GRCh38 but not KOREF_C (Table 45); I verified these loci by manually checking short reads alignment to both GRCh38 and KOREF_C (Fig. 24).

Table 43. Disease term enrichment test for genes predicted to be functionally altered when using GRCh38 but not KOREF_C

Group	Disease term	#Gene	*P*-value	Bonferroni *P*-value
Korean	Adhesion	26	3.20E-08	2.40E-05
	Hypertension	14	3.58E-07	3.00E-04
	Musculoskeletal Diseases	20	3.93E-07	3.00E-04
	Genetic Predisposition to Disease	27	6.77E-07	5.00E-04
	Gestational hypertension	10	6.52E-07	5.00E-04
	Bacterial Infections	12	8.01E-07	6.00E-04
	Myocardial Infarction	14	7.72E-07	6.00E-04
	metabolic syndrome	11	1.86E-06	1.40E-03
	Eclampsia	9	3.09E-06	2.30E-03
	Disease Susceptibility	26	3.15E-06	2.40E-03
	Infarction	13	3.27E-06	2.50E-03
	Bone Diseases	13	4.92E-06	3.70E-03
	Osteonecrosis	5	5.02E-06	3.80E-03
	Coronary Disease	13	6.97E-06	5.20E-03
	Coronary Artery Disease	13	7.27E-06	5.50E-03
	Myocardial Ischemia	13	9.73E-06	7.30E-03
	Collagen Diseases	8	1.35E-05	1.01E-02
	Pre-Eclampsia	8	2.42E-05	1.81E-02
	Dwarfism	7	2.66E-05	1.99E-02
	Gastroschisis	3	2.76E-05	2.07E-02
	Brain Ischemia	8	3.40E-05	2.55E-02
	Mycobacterium Infections	7	3.41E-05	2.56E-02
	Arteriosclerosis	11	3.50E-05	2.62E-02
	Arterial Occlusive Diseases	11	4.32E-05	3.24E-02
	Aggressive Periodontitis	5	4.33E-05	3.25E-02
	Mycobacterial infection	7	4.33E-05	3.25E-02
	Coxa plana	3	4.78E-05	3.58E-02
	Congenital dislocation of hip NOS	4	5.36E-05	4.02E-02
Asian including Korean	Musculoskeletal Diseases	23	1.83E-08	1.28E-05
	Adhesion	26	1.57E-07	1.00E-04
	Bone Diseases	15	4.22E-07	3.00E-04
	Bacterial Infections	12	1.86E-06	1.30E-03
	Myocardial Infarction	14	1.99E-06	1.40E-03
	Collagen Diseases	9	2.83E-06	2.00E-03
	metabolic syndrome	11	4.04E-06	2.80E-03
	Hypertension	13	5.14E-06	3.60E-03
	Genetic Predisposition to Disease	26	9.21E-06	6.40E-03
	Gestational hypertension	9	1.08E-05	7.50E-03
	Disease Susceptibility	25	3.65E-05	2.55E-02
	Infarction	12	3.82E-05	2.67E-02
	Dwarfism	7	4.44E-05	3.10E-02
	Eclampsia	8	4.60E-05	3.21E-02
Caucasian	Musculoskeletal Diseases	17	6.01E-07	3.00E-04
	Common Cold	11	9.83E-06	4.90E-03
	Dystonia Musculorum Deformans	4	1.12E-05	5.60E-03
	Adhesion	17	4.87E-05	2.43E-02
	Respiratory Tract Infections	10	5.19E-05	2.59E-02
	Bone Diseases	10	5.37E-05	2.68E-02
	metabolic syndrome	8	6.48E-05	3.23E-02
	Bacterial Infections	8	1.00E-04	4.99E-02
African	Bone Diseases	11	9.02E-07	4.00E-04
	Musculoskeletal Diseases	13	1.60E-05	6.30E-03
	Collagen Diseases	6	4.63E-05	1.83E-02
	Aggressive Periodontitis	4	8.55E-05	3.38E-02
	metabolic syndrome	7	9.35E-05	3.69E-02
	Adhesion	14	1.00E-04	3.95E-02

Table 44. Pathway enrichment test for genes predicted to be functionally altered when using GRCh38 but not KOREF_C

Group	Pathway	#Gene	*P*-value	Bonferroni *P*-value
Korean	Olfactory transduction	36	4.00E-22	2.16E-20
	ECM-receptor interaction	9	5.16E-07	2.79E-05
	Taste transduction	5	3.00E-04	1.62E-02
	Focal adhesion	9	5.00E-04	2.70E-02
Asian including Korean	Olfactory transduction	36	5.93E-21	3.50E-19
	ECM-receptor interaction	9	1.01E-06	5.96E-05
	Taste transduction	5	4.00E-04	2.36E-02
	Protein digestion and absorption	6	5.00E-04	2.95E-02
	Focal adhesion	9	8.00E-04	4.72E-02
Caucasian	Olfactory transduction	31	7.79E-21	2.57E-19
	ECM-receptor interaction	7	8.52E-06	3.00E-04
	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	5	4.00E-04	1.32E-02
	Hypertrophic cardiomyopathy (HCM)	5	7.00E-04	2.31E-02
	Dilated cardiomyopathy	5	1.10E-03	3.63E-02
African	Olfactory transduction	20	2.50E-12	6.50E-11
	ECM-receptor interaction	6	2.32E-05	6.00E-04

Table 45. Disease associated nsSNVs found against GRCh38 but not KOREF_C

Chr	Pos	Ref	Alt	Gene	A.A. Change	sig	name	acc	Freq. in Korean	Freq. in Asian	Freq. in Caucasian	Freq. in African
1	100206504	T	C	*DBT*	G323S	pathogenic	Intermediate maple syrup urine disease type 2	RCV000012727.21	5/5	4/5	5/5	4/5
1	196690107	C	T	*CFH*	Y402H	pathogenic	Basal laminar drusen	RCV000018016.27	4/5	4/5	4/5	4/5
4	186236880	G	A	*KLKB1*	N124S	pathogenic	Prekallikrein deficiency	RCV000012817.23	2/5	3/5	1/5	4/5
5	35871088	G	A	*IL7R*	I138V	pathogenic	Severe combined immunodeficiency, autosomal recessive, T cell-negative, B cell-positive, NK cell-positive	RCV000015965.24	2/5	0/5	4/5	4/5
5	74685445	T	C	*HEXB*	S62L	pathogenic	Sandhoff disease, infantile type	RCV000004086.1	3/5	5/5	5/5	5/5
7	150999023	T	G	*NOS3*	E298D	pathogenic	Hypertension resistant to conventional therapy	RCV000015056.2	5/5	4/5	0/5	4/5
8	18400806	G	A	*NAT2*	K268R	drug-response	NAT2:N-acetyltransferase 2 (arylamine N-acetyltransferase)	RCV000000760.2	5/5	5/5	3/5	1/5
11	17388025	T	C	*KCNJ11*	E23K	drug response	Exercise stress response, impaired, association with	RCV000009215.1	1/5	3/5	1/5	5/5
12	120999579	A	G	*HNF1A*	G574S	pathogenic	Maturity-onset diabetes of the young, type 3 (MODY3)	RCV000016077.24	5/5	5/5	5/5	4/5
12	121857429	T	C	*HPD*	A33T	pathogenic	4-Alpha-hydroxyphenylpyruvate hydroxylase deficiency	RCV000001643.1	5/5	4/5	3/5	5/5
15	48134287	A	G	*SLC24A5*	A111T	pathogenic	Skin/hair/eye pigmentation, variation in, 4 (SHEP4)	RCV000001552.2	5/5	5/5	0/5	5/5
16	56514589	C	T	*BBS2*	N70S	pathogenic	BARDET-BIEDL SYNDROME 2/6, DIGENIC	RCV000004838.2	5/5	5/5	5/5	5/5
22	18913491	C	T	*PRODH*	Q521R	pathogenic	Proline dehydrogenase deficiency (HYRPRO1)	RCV000004222.4	4/5	5/5	4/5	2/5

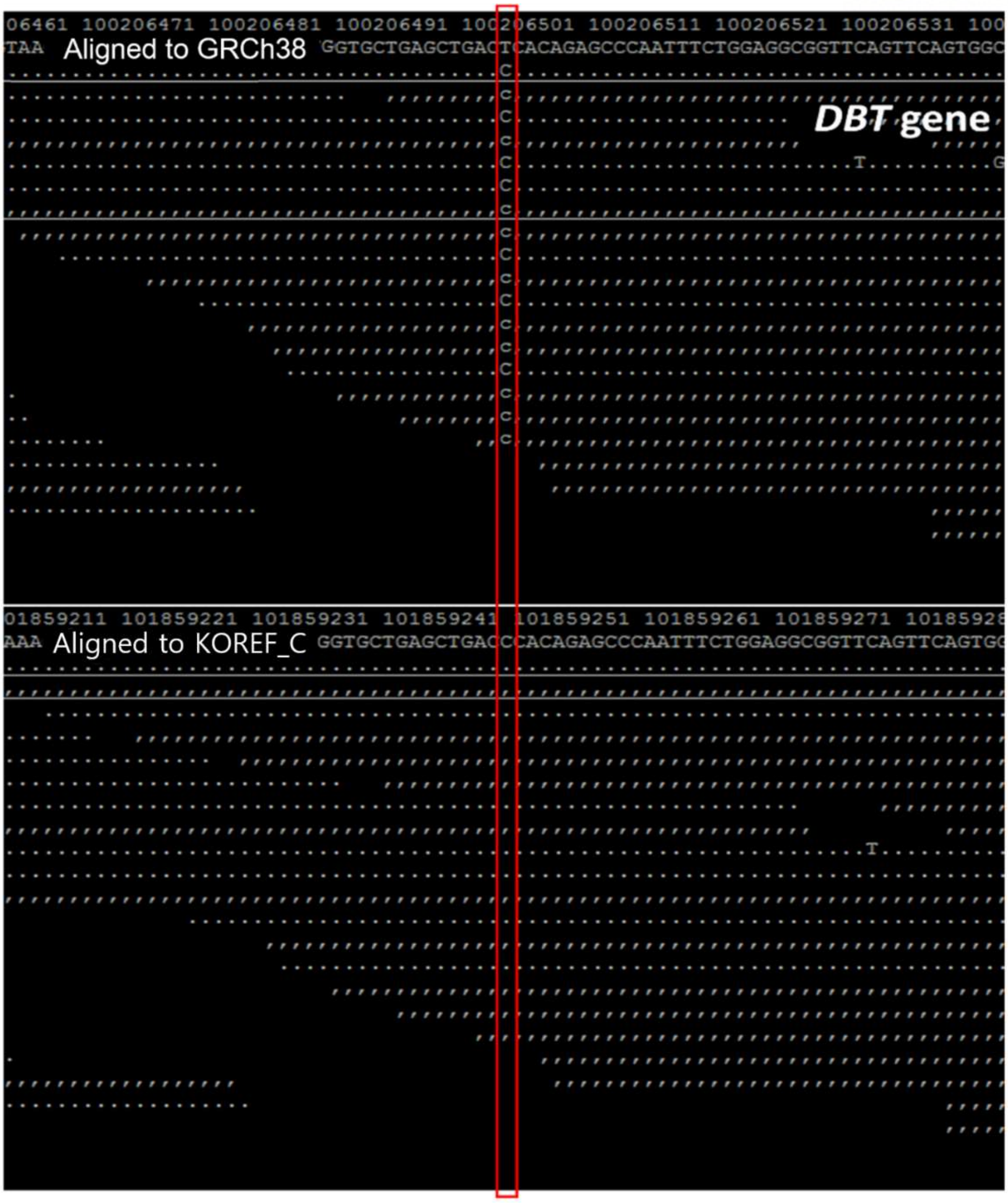


Figure 24. An example of variants that were called against GRCh38, but not KOREF_C. The 13 nsSNVs that are known as disease- and phenotype-associated were verified by visual inspection of short reads alignments.

IV. Conclusions

Each ethnic group has a specific variation repertoire, including single nucleotide polymorphisms and larger structural deviations[6,69]. Therefore, for large-scale population genome projects, leveraging ethnicity-specific reference genomes alongside GRCh38 can bring additional benefits in detecting variants more efficiently. The genotype reconstruction should bring similar results (without the assembly-specific sequence regions) regardless which reference is used, if the assembly quality is similarly high and all-sites of whole genome are called. Instead, the ethnic-relevant assembly has an additional utility in terms of fast and efficient variant-calling (lower number of variants) for the same ethnic genomes especially with the consensus variants components. Also, the population stratification (systematic difference in allele frequencies) can be a problem for association studies, where the association could be found due to the underlying structure of the population and not a disease associated locus[70]. In cancer genome analyses, it is a common practice to compare cancer sample sequencing data against public variants databases such as dbSNP[56] to remove previously described normal variants as a key filtering step in detecting somatic point mutations[71]. As a consensus reference, KOREF contains the Korean population variome from 40 additional Korean personal genomes, and can help researchers to efficiently process cancer-specific variants. Ethnicity-specific genomic regions such as novel sequences and copy number variable regions can affect precise genotype reconstruction. I demonstrate an example of a better genotype reconstruction in the copy number variable regions using KOREF (Fig. 25). Hence, the ethnicity-specific reference genome, KOREF, can also be useful for detecting disease-relevant variants in East-Asians.

As a national standard reference genome, KOREF has been constructed according to standardized production and evaluation procedures (document # GDC-KMP-004 and GDC-KEP-004) that were registered in National Center for Standard Reference Data (NCSRD) of Korea. In 2017, KOREF is officially registered as a standard reference for Korean genome by evaluating its traceability, uncertainty, and consistency and by expert committee's reviews.

De novo assembly based on Sanger sequencing is still too expensive to be used routinely. I have demonstrated that it is possible to produce a *de novo* assembly of relatively high quality at a fraction of the cost by combining the latest sequencing and bioinformatics methods. Additionally, I have shown that optical and nano technologies can extend the size of the large scaffolds while validating the initial assembly. I found that the identification of structural differences based on the genome assembly is largely affected by assembly quality, suggesting a need for new technologies and higher quality of assembly from additional individuals in various populations to better understand comprehensive maps of genomic structure. Also, it is important that the same coordinate system on the GRCh38 allows comparison of different individuals, to leverage the vast amount of previously established knowledge and annotations. Therefore, it is also crucial to investigate how to transfer those annotations to personal/ethnic reference genomes by preferentially supplementing additional references into GRCh38 to gain additional biological insights.

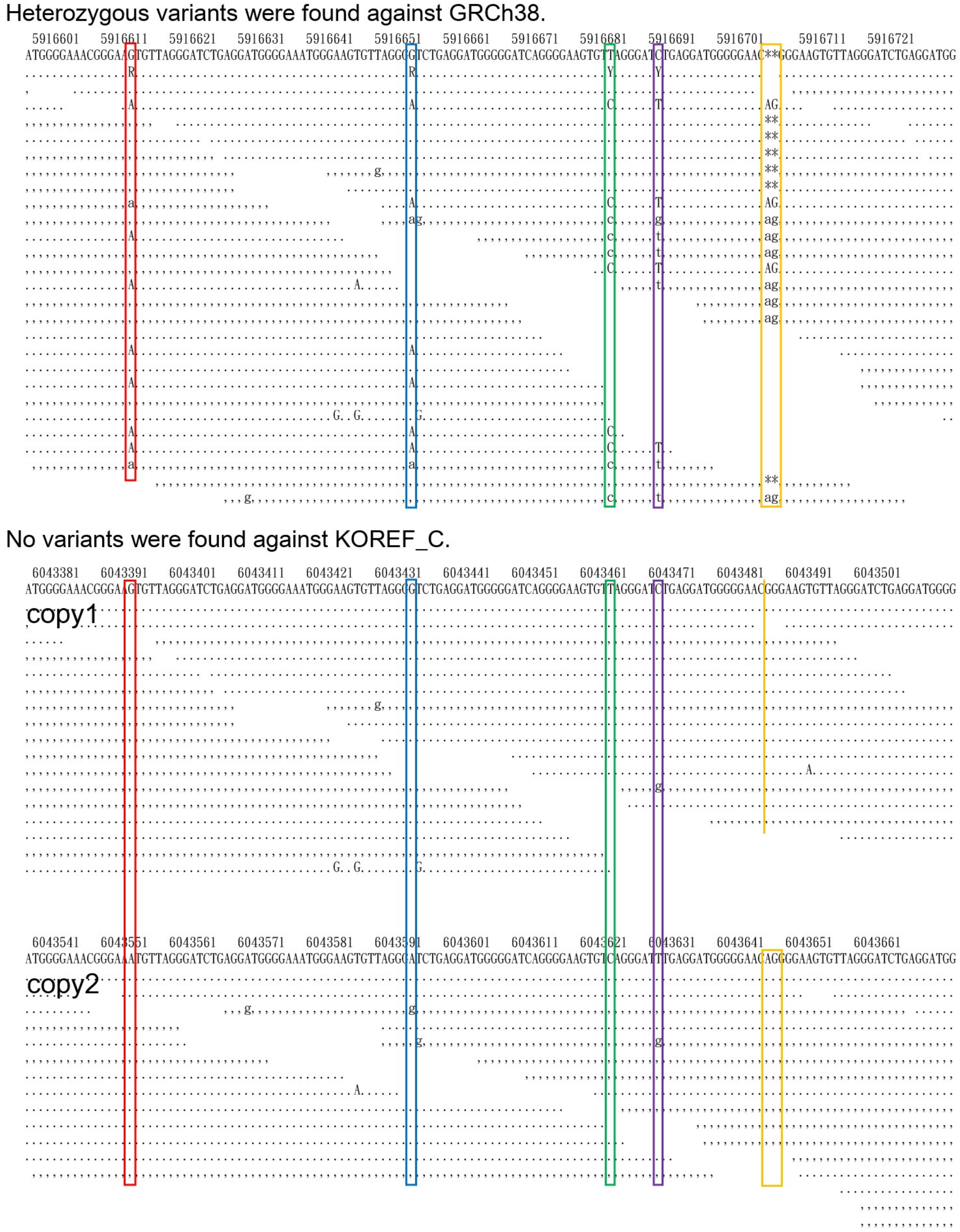


Figure 25. An example of genotype reconstruction difference in GRCh38 and KOREF_C. GRCh38 has one copy region, but KOREF_C has two copies for the same region. Heterozygous variants that may be caused by the copy number difference were not detected when using KOREF_C.

KOREFs cannot, and are not meant to, replace the human reference in general, and some of its genomic regions, such as centromeric and telomeric regions, and many gaps, are largely incomplete. However, KOREFs still can be useful in improving the alignment of East-Asian personal genomes, in terms of fast and efficient variant-calling and detecting individual- and ethnic-specific variations for large-scale genome projects. I think it is possible in the near future to use KOREF as a platform for constructing a complete reference genome that includes all the missing gaps and repeat regions using currently available long distance genome interaction information such as Hi-C[72] and other nanochannel and nanopore based sequencing technologies[73]. I also think that every individual should have his or her own reference genome for a high quality genotype reconstruction and genomic structure identification in the personalized medicine era. Therefore, a new era of the *de novo* assembly based personal reference will arrive together with and through improving genome technologies.

References

1. Reich, D.; Nalls, M. A.; Kao, W. H.; Akylbekova, E. L.; Tandon, A.; Patterson, N.; Mullikin, J.; Hsueh, W. C.; Cheng, C. Y.; Coresh, J.; Boerwinkle, E.; Li, M.; Waliszewska, A.; Neubauer, J.; Li, R.; Leak, T. S.; Ekunwe, L.; Files, J. C.; Hardy, C. L.; Zmuda, J. M.; Taylor, H. A.; Ziv, E.; Harris, T. B.; Wilson, J. G. 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.

2. Green, R. E.; Krause, J.; Briggs, A. W.; Maricic, T.; Stenzel, U.; Kircher, M.; Patterson, N.; Li, H.; Zhai, W.; Fritz, M. H.; Hansen, N. F.; Durand, E. Y.; Malaspinas, A. S.; Jensen, J. D.; Marques-Bonet, T.; Alkan, C.; Prüfer, K.; Meyer, M.; Burbano, H. A.; Good, J. M.; Schultz, R.; Aximu-Petri, A.; Butthof, A.; Höber, B.; Höffner, B.; Siegemund, M.; Weihmann, A.; Nusbaum, C.; Lander, E. S.; Russ, C.; Novod, N.; Affourtit, J.; Egholm, M.; Verna, C.; Rudan, P.; Brajkovic, D.; Kucan, Z.; Gusic, I.; Doronichev, V. B.; Golovanova, L. V.; Lalueza-Fox, C.; de la Rasilla, M.; Fortea, J.; Rosas, A.; Schmitz, R. W.; Johnson, P. L.; Eichler, E. E.; Falush, D.; Birney, E.; Mullikin, J. C.; Slatkin, M.; Nielsen, R.; Kelso, J.; Lachmann, M.; Reich, D.; Pääbo, S. A Draft Sequence of the Neandertal Genome. *Science* **2010,** *328*, 710–722.

3. Sheehan, S.; Harris, K.; Song, Y. S. Estimating Variable Effective Population Sizes from Multiple Genomes: a Sequentially Markov Conditional Sampling Distribution Approach. *Genetics* **2013,** *194*, 647–662.

4. Schiffels, S.; Durbin, R. Inferring Human Population Size and Separation History from Multiple Genome Sequences. *Nat. Genet.* **2014,** *46*, 919–925.

5. Dewey, F. E.; Chen, R.; Cordero, S. P.; Ormond, K. E.; Caleshu, C.; Karczewski, K. J.; Whirl-Carrillo, M.; Wheeler, M. T.; Dudley, J. T.; Byrnes, J. K.; Cornejo, O. E.; Knowles, J. W.; Woon, M.; Sangkuhl, K.; Gong, L.; Thorn, C. F.; Hebert, J. M.; Capriotti, E.; David, S. P.; Pavlovic, A.; West, A.; Thakuria, J. V.; Ball, M. P.; Zaranek, A. W.; Rehm, H. L.; Church, G. M.; West, J. S.; Bustamante, C. D.; Snyder, M.; Altman, R. B.; Klein, T. E.; Butte, A. J.; Ashley, E. A. Phased Whole-Genome Genetic Risk in a Family Quartet using a Major Allele Reference Sequence. *PLoS Genet.* **2011,** *7*, e1002280.

6. Sudmant, P. H.; Mallick, S.; Nelson, B. J.; Hormozdiari, F.; Krumm, N.; Huddleston, J.; Coe, B. P.; Baker, C.; Nordenfelt, S.; Bamshad, M.; Jorde, L. B.; Posukh, O. L.; Sahakyan, H.; Watkins, W. S.; Yepiskoposyan, L.; Abdullah, M. S.; Bravi, C. M.; Capelli, C.; Hervig, T.; Wee, J. T.; Tyler-Smith, C.; van Driem, G.; Romero, I. G.; Jha, A. R.; Karachanak-Yankova, S.; Toncheva, D.; Comas, D.; Henn, B.; Kivisild, T.; Ruiz-Linares, A.; Sajantila, A.; Metspalu, E.; Parik, J.; Villems, R.;

Starikovskaya, E. B.; Ayodo, G.; Beall, C. M.; Di Rienzo, A.; Hammer, M. F.; Khusainova, R.; Khusnutdinova, E.; Klitz, W.; Winkler, C.; Labuda, D.; Metspalu, M.; Tishkoff, S. A.; Dryomov, S.; Sukernik, R.; Patterson, N.; Reich, D.; Eichler, E. E. Global Diversity, Population Stratification, and Selection of Human Copy-Number Variation. *Science* **2015,** *349*, aab3761.

7. Lander, E. S.; Linton, L. M.; Birren, B.; Nusbaum, C.; Zody, M. C.; 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, J. P.; Miranda, C.; Morris, W.; Naylor, J.; Raymond, C.; Rosetti, M.; Santos, R.; Sheridan, A.; Sougnez, C.; Stange-Thomann, Y.; 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, J. C.; Mungall, A.; Plumb, R.; Ross, M.; Shownkeen, R.; Sims, S.; Waterston, R. H.; Wilson, R. K.; Hillier, L. W.; McPherson, J. D.; Marra, M. A.; Mardis, . ER.; Fulton, L. A.; Chinwalla, A. T.; Pepin, K. H.; Gish, W. R.; Chissoe, S. L.; Wendl, M. C.; Delehaunty, K. D.; Miner, T. L.; Delehaunty, A.; Kramer, J. B.; Cook, L. L.; Fulton, R. S.; Johnson, D. L.; Minx, P. J.; Clifton, S. W.; Hawkins, T.; Branscomb, E.; Predki, P.; Richardson, P.; Wenning, S.; Slezak, T.; Doggett, N.; Cheng, J. F.; Olsen, A.; Lucas, S.; Elkin, C.; Uberbacher, E.; Frazier, M.; Gibbs, R. A.; Muzny, D. M.; Scherer, S. E.; Bouck, J. B.; Sodergren, E. J.; Worley, K. C.; Rives, C. M.; Gorrell, J. H.; Metzker, M. L.; Naylor, S. L.; Kucherlapati, R. S.; Nelson, D. L.; Weinstock, G. M.; Sakaki, Y.; Fujiyama, A.; Hattori, M.; Yada, T.; Toyoda, A.; Itoh, T.; Kawagoe, C.; Watanabe, H.; Totoki, Y.; Taylor, T.; Weissenbach, J.; Heilig, R.; Saurin, W.; Artiguenave, F.; Brottier, P.; Bruls, T.; Pelletier, E.; Robert, C.; Wincker, P.; Smith, D. R.; Doucette-Stamm, L.; Rubenfield, M.; Weinstock, K.; Lee, H. M.; Dubois, J.; Rosenthal, A.; Platzer, M.; Nyakatura, G.; Taudien, S.; Rump, A.; Yang, H.; Yu, J.; Wang, J.; Huang, G.; Gu, J.; Hood, L.; Rowen, L.; Madan, A.; Qin, S.; Davis, R. W.; Federspiel, N. A.; Abola, A. P.; Proctor, M. J.; Myers, R. M.; Schmutz, J.; Dickson, M.; Grimwood, J.; Cox, D. R.; Olson, M. V.; Kaul, R.; Raymond, C.; Shimizu, N.; Kawasaki, K.; Minoshima, S.; Evans, G. A.; Athanasiou, M.; Schultz, R.; Roe, B. A.; Chen, F.; Pan, H.; Ramser, J.; Lehrach, H.; Reinhardt, R.; McCombie, W. R.; de la Bastide, M.; Dedhia, N.; Blöcker, H.; Hornischer, K.; Nordsiek, G.; Agarwala, R.; Aravind, L.; Bailey, J. A.; Bateman, A.; Batzoglou, S.; Birney, E.; Bork, P.; Brown, D. G.; Burge, C. B.; Cerutti, L.; Chen, H. C.; Church, D.; Clamp, M.; Copley, R. R.; Doerks, T.; Eddy, S. R.; Eichler, E. E.; Furey, T. S.; Galagan, J.; Gilbert, J. G.; Harmon, C.; Hayashizaki, Y.; Haussler, D.; Hermjakob, H.; Hokamp, K.; Jang, W.; Johnson, L. S.; Jones, T. A.; Kasif, S.; Kaspryzk, A.; Kennedy, S.; Kent, W. J.; Kitts, P.; Koonin, E. V.; Korf, I.; Kulp, D.; Lancet, D.; Lowe, T. M.; McLysaght, A.; Mikkelsen, T.; Moran, J. V.; Mulder, N.;

Pollara, V. J.; Ponting, C. P.; Schuler, G.; Schultz, J.; Slater, G.; Smit, A. F.; Stupka, E.; Szustakowki, J.; Thierry-Mieg, D.; Thierry-Mieg, J.; Wagner, L.; Wallis, J.; Wheeler, R.; Williams, A.; Wolf, Y. I.; Wolfe, K. H.; Yang, S. P.; Yeh, R. F.; Collins, F.; Guyer, M. S.; Peterson, J.; Felsenfeld, A.; Wetterstrand, K. A.; Patrinos, A.; Morgan, M. J.; de Jong, P.; Catanese, J. J.; Osoegawa, K.; Shizuya, H.; Choi, S.; Chen, Y. J.; Szustakowki, J.; International Human Genome Sequencing Consortium. Initial Sequencing and Analysis of the Human Genome. *Nature* **2001,** *409*, 860–921.

8. Levy, S.; Sutton, G.; Ng, P. C.; Feuk, L.; Halpern, A. L.; Walenz, B. P.; Axelrod, N.; Huang, J.; Kirkness, E. F.; Denisov, G.; Lin, Y.; MacDonald, J. R.; Pang, A. W.; Shago, M.; Stockwell, T. B.; Tsiamouri, A.; Bafna, V.; Bansal, V.; Kravitz, S. A.; Busam, D. A.; Beeson, K. Y.; McIntosh, T. C.; Remington, K. A.; Abril, J. F.; Gill, J.; Borman, J.; Rogers, Y. H.; Frazier, M. E.; Scherer, S. W.; Strausberg, R. L.; Venter, J. C. The Diploid Genome Sequence of an Individual Human. *PLoS Biol,* **2007,** *5*, e254.

9. Li, R.; Zhu, H.; Ruan, J.; Qian, W.; Fang, X.; Shi, Z.; Li, Y.; Li, S.; Shan, G.; Kristiansen, K.; Li, S.; Yang, H.; Wang, J.; Wang, J. De novo Assembly of Human Genomes with Massively Parallel Short Read Sequencing. *Genome Res.* **2010,** *20*, 265–272.

10. Bai, H.; Guo, X.; Zhang, D.; Narisu, N.; Bu, J.; Jirimutu, J.; Liang, F.; Zhao, X.; Xing, Y.; Wang, D.; Li, T.; Zhang, Y.; Guan, B.; Yang, X.; Yang, Z.; Shuangshan, S.; Su, Z.; Wu, H.; Li, W.; Chen, M.; Zhu, S.; Bayinnamula, B.; Chang, Y.; Gao, Y.; Lan, T.; Suyalatu, S.; Huang, H.; Su, Y.; Chen, Y.; Li, W.; Yang, X.; Feng, Q.; Wang, J.; Yang, H.; Wang, J.; Wu, Q.; Yin, Y.; Zhou, H. The Genome of a Mongolian Individual Reveals the Genetic Imprints of Mongolians on Modern Human Populations. *Genome Biol. Evol.* **2014,** *6*, 3122–3136.

11. Gnerre, S.; Maccallum, I.; Przybylski, D.; Ribeiro, F. J.; Burton, J. N.; Walker, B. J.; Sharpe, T.; Hall, G.; Shea, T. P.; Sykes, S.; Berlin, A. M.; Aird, D.; Costello, M.; Daza, R.; Williams, L.; Nicol, R.; Gnirke, A.; Nusbaum, C.; Lander, E. S.; Jaffe, D. B. High-Quality Draft Assemblies of Mammalian Genomes from Massively Parallel Sequence Data. *Proc. Natl. Acad. Sci. U. S. A.* **2011,** *108*, 1513–1518.

12. Steinberg, K. M.; Schneider, V. A.; Graves-Lindsay, T. A.; Fulton, R. S.; Agarwala, R.; Huddleston, J.; Shiryev, S. A.; Morgulis, A.; Surti, U.; Warren, W. C.; Church, D. M.; Eichler, E. E.; Wilson, R. K. Single Haplotype Assembly of the Human Genome from a Hydatidiform Mole. *Genome Res.* **2014,** *24*, 2066–2076.

13. Cao, H.; Wu, H.; Luo, R.; Huang, S.; Sun, Y.; Tong, X.; Xie, Y.; Liu, B.; Yang, H.; Zheng, H.; Li,

J.; Li, B.; Wang, Y.; Yang, F.; Sun, P.; Liu, S.; Gao, P.; Huang, H.; Sun, J.; Chen, D.; He, G.; Huang, W.; Huang, Z.; Li, Y.; Tellier, L. C.; Liu, X.; Feng, Q.; Xu, X.; Zhang, X.; Bolund, L.; Krogh, A.; Kristiansen, K.; Drmanac, R.; Drmanac, S.; Nielsen, R.; Li, S.; Wang, J.; Yang, H.; Li, Y.; Wong, G. K.; Wang, J. De novo Assembly of a Haplotype-Resolved Human Genome. *Nat. Biotechnol.* **2015,** *33*, 617–622.

14. Alkan, C.; Sajjadian, S.; Eichler, E. E. Limitations of Next-Generation Genome Sequence Assembly. *Nat. Methods* **2011,** *8*, 61–65.

15. Chaisson, M. J.; Huddleston, J.; Dennis, M. Y.; Sudmant, P. H.; Malig, M.; Hormozdiari, F.; Antonacci, F.; Surti, U.; Sandstrom, R.; Boitano, M.; Landolin, J. M.; Stamatoyannopoulos, J. A.; Hunkapiller, M. W.; Korlach, J.; Eichler, E. E. Resolving the Complexity of the Human Genome using Single-Molecule Sequencing. *Nature* **2015,** *517*, 608–611.

16. Huddleston, J.; Ranade, S.; Malig, M.; Antonacci, F.; Chaisson, M.; Hon, L.; Sudmant, P. H.; Graves, T. A.; Alkan, C.; Dennis, M. Y.; Wilson, R. K.; Turner, S. W.; Korlach, J.; Eichler, E. E. Reconstructing Complex Regions of Genomes using Long-Read Sequencing Technology. *Genome Res.* **2014,** *24*, 688–696.

17. McCoy, R. C.; Taylor, R. W.; Blauwkamp, T. A.; Kelley, J. L.; Kertesz, M.; Pushkarev, D.; Petrov, D. A.; Fiston-Lavier, A. S. Illumina TruSeq Synthetic Long-Reads Empower De novo Assembly and Resolve Complex, Highly-Repetitive Transposable Elements. *PLoS One* **2014,** *9*, e106689.

18. Dong, Y.; Xie, M.; Jiang, Y.; Xiao, N.; Du, X.; Zhang, W.; Tosser-Klopp, G.; Wang, J.; Yang, S.; Liang, J.; Chen, W.; Chen, J.; Zeng, P.; Hou, Y.; Bian, C.; Pan, S.; Li, Y.; Liu, X.; Wang, W.; Servin, B.; Sayre, B.; Zhu, B.; Sweeney, D.; Moore, R.; Nie, W.; Shen, Y.; Zhao, R.; Zhang, G.; Li, J.; Faraut, T.; Womack, J.; Zhang, Y.; Kijas, J.; Cockett, N.; Xu, X.; Zhao, S.; Wang, J.; Wang, W. Sequencing and Automated Whole-Genome Optical Mapping of the Genome of a Domestic Goat (Capra hircus). *Nat. Biotechnol.* **2013,** *31*, 135–141.

19. Cao, H.; Hastie, A. R.; Cao, D.; Lam, E. T.; Sun, Y.; Huang, H.; Liu, X.; Lin, L.; Andrews, W.; Chan, S.; Huang, S.; Tong, X.; Requa, M.; Anantharaman, T.; Krogh, A.; Yang, H.; Cao, H.; Xu, X. Rapid Detection of Structural Variation in a Human Genome using Nanochannel-based Genome Mapping Technology. *GigaScience* **2014,** *3*, 34.

20. Howe, K.; Wood, J. M. Using Optical Mapping Data for the Improvement of Vertebrate Genome Assemblies. *GigaScience* **2015,** *4*, 10.

21. Pendleton, M.; Sebra, R.; Pang, A. W.; Ummat, A.; Franzen, O.; Rausch, T.; Stütz, A. M.; Stedman, W.; Anantharaman, T.; Hastie, A.; Dai, H.; Fritz, M. H.; Cao, H.; Cohain, A.; Deikus, G.; Durrett, R. E.; Blanchard, S. C.; Altman, R.; Chin, C. S.; Guo, Y.; Paxinos, E. E.; Korbel, J. O.; Darnell, R. B.; McCombie, W. R.; Kwok, P. Y.; Mason, C. E.; Schadt, E. E.; Bashir, A. Assembly and Diploid Architecture of an Individual Human Genome via Single-Molecule Technologies. *Nat. Methods* **2015,** *12*, 780–786.

22. Shi, L.; Guo, Y.; Dong, C.; Huddleston, J.; Yang, H.; Han, X.; Fu, A.; Li, Q.; Li, N.; Gong, S.; Lintner, K. E.; Ding, Q.; Wang, Z.; Hu, J.; Wang, D.; Wang, F.; Wang, L.; Lyon, G. J.; Guan, Y.; Shen, Y.; Evgrafov, O. V.; Knowles, J. A.; Thibaud-Nissen, F.; Schneider, V.; Yu, C. Y.; Zhou, L.; Eichler, E. E.; So, K. F.; Wang, K. Long-Read Sequencing and De novo Assembly of a Chinese Genome. *Nat. Commun.* **2016,** *7*, 12065.

23. Church, G. M. The Personal Genome Project. *Mol. Syst. Biol.* **2005,** *1*, 2005.0030.

24. The 1000 Genomes Project Consortium; Abecasis, G. R.; Altshuler, D.; Auton, A.; Brooks, L. D.; Durbin, R. M.; Gibbs, R. A.; Hurles, M. E.; McVean, G. A. A Map of Human Genome Variation from Population-scale Sequencing. *Nature* **2010,** *467*, 1061–1073.

25. The 1000 Genomes Project Consortium; Abecasis, G. R.; Auton, A.; Brooks, L. D.; DePristo, M. A.; Durbin, R. M.; Handsaker, R. E.; Kang, H. M.; Marth, G. T.; McVean, G. A. An Integrated Map of Genetic Variation from 1,092 Human Genomes. *Nature* **2012,** *491*, 56–65.

26. Muddyman, D.; Smee, C.; Griffin, H.; Kaye, J. Implementing a Successful Data-Management Framework: the UK10K Managed Access Model. *Genome Med.* **2013,** *5*, 100.

27. Genome of the Netherlands Consortium. Whole-Genome Sequence Variation, Population Structure and Demographic History of the Dutch Population. *Nat. Genet.* **2014,** *46*, 818–825.

28. Zhang, W.; Meehan, J.; Su, Z.; Ng, H. W.; Shu, M.; Luo, H.; Ge, W.; Perkins, R.; Tong, W.; Hong, H. Whole Genome Sequencing of 35 Individuals Provides Insights into the Genetic Architecture of Korean Population. *BMC Bioinf.* **2014,** *15 Suppl 11*, S6.

29. Purcell, S.; Neale, B.; Todd-Brown, K.; Thomas, L.; Ferreira, M. A.; Bender, D.; Maller, J.; Sklar, P.; de Bakker, P. I.; Daly, M. J.; Sham, P. C. PLINK: a Tool Set for Whole-Genome Association and Population-based Linkage Analyses. *Am. J. Hum. Genet.* **2007,** *81*, 559–575.

30. Tosato, G.; Cohen, J. I. Generation of Epstein-Barr Virus (EBV)-Immortalized B Cell Lines. *Curr. Protoc. Immunol.* **2007,** *Chapter 7*, Unit 7.22.

31. Luo, R.; Liu, B.; Xie, Y.; Li, Z.; Huang, W.; Yuan, J.; He, G.; Chen, Y.; Pan, Q.; Liu, Y.; Tang, J.; Wu, G.; Zhang, H.; Shi, Y.; Liu, Y.; Yu, C.; Wang, B.; Lu, Y.; Han, C.; Cheung, D. W.; Yiu, S. M.; Peng, S.; Xiaoqian, Z.; Liu, G.; Liao, X.; Li, Y.; Yang, H.; Wang, J.; Lam, T. W.; Wang, J. SOAPdenovo2: an Empirically Improved Memory-Efficient Short-Read De novo Assembler. *GigaScience* **2012,** *1*, 18.

32. English, A. C.; Richards, S.; Han, Y.; Wang, M.; Vee, V.; Qu, J.; Qin, X.; Muzny, D. M.; Reid, J. G.; Worley, K. C.; Gibbs, R. A. Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. *PLoS One* **2012,** *7*, e47768.

33. Li, H. Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM. **2013,** Preprint at arXiv:1303.3997v2 [q-bio.GN].

34. Soderlund, C.; Bomhoff, M.; Nelson, W. M. SyMAP v3.4: a Turnkey Synteny System with Application to Plant Genomes. *Nucleic Acids Res.* **2011,** *39*, e68.

35. Chaisson, M. J.; Tesler, G. Mapping Single Molecule Sequencing Reads using Basic Local Alignment with Successive Refinement (BLASR): Application and Theory. *BMC Bioinf.* **2012,** *13*, 238.

36. Simpson, J. T.; Wong, K.; Jackman, S. D.; Schein, J. E.; Jones, S. J.; Birol, I. ABySS: a Parallel Assembler for Short Read Sequence Data. *Genome Res.* **2009,** *19*, 1117–1123.

37. Fan, L.; Yao, Y. G. MitoTool: a Web Server for the Analysis and Retrieval of Human Mitochondrial DNA Sequence Variations. *Mitochondrion* **2011,** *11*, 351–356.

38. McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernytsky, A.; Garimella, K.; Altshuler, D.; Gabriel, S.; Daly, M.; DePristo, M. A. The Genome Analysis Toolkit: a MapReduce Framework for Analyzing Next-Generation DNA Sequencing Data. *Genome Res.* **2010,** *20*, 1297–1303.

39. Benson, G. Tandem Repeats Finder: a Program to Analyze DNA Sequences. *Nucleic Acids Res.* **1999,** *27*, 573–580.

40. Jurka, J.; Kapitonov, V. V.; Pavlicek, A.; Klonowski, P.; Kohany, O.; Walichiewicz, J. Repbase Update, a Database of Eukaryotic Repetitive Elements. *Cytogenet. Genome Res.* **2005,** *110*, 462–467.

41. Bedell, J. A.; Korf, I.; Gish, W. MaskerAid: a Performance Enhancement to RepeatMasker. *Bioinformatics* **2000,** *16*, 1040–1041.

42. Abrusán, G.; Grundmann, N.; DeMester, L.; Makalowski, W. TEclass--a Tool for Automated Classification of Unknown Eukaryotic Transposable Elements. *Bioinformatics* **2009,** *25*, 1329–1330.

43. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T. L. BLAST+: Architecture and Applications. *BMC Bioinf.* **2009,** *10*, 421.

44. Slater, G. S.; Birney, E. Automated Generation of Heuristics for Biological Sequence Comparison. *BMC Bioinf.* **2005,** *6*, 31.

45. Stanke, M.; Keller, O.; Gunduz, I.; Hayes, A.; Waack, S.; Morgenstern, B. AUGUSTUS: Ab Initio Prediction of Alternative Transcripts. *Nucleic Acids Res.* **2006,** *34* (Web Server issue), W435–W439.

46. Pruitt, K. D.; Brown, G. R.; Hiatt, S. M.; Thibaud-Nissen, F.; Astashyn, A.; Ermolaeva, O.; Farrell, C. M.; Hart, J.; Landrum, M. J.; McGarvey, K. M.; Murphy, M. R.; O'Leary, N. A.; Pujar, S.; Rajput, B.; Rangwala, S. H.; Riddick, L. D.; Shkeda, A.; Sun, H.; Tamez, P.; Tully, R. E.; Wallin, C.; Webb, D.; Weber, J.; Wu, W.; DiCuccio, M.; Kitts, P.; Maglott, D. R.; Murphy, T. D.; Ostell, J. M. RefSeq: an Update on Mammalian Reference Sequences. *Nucleic Acids Res.* **2014,** *42* (Database issue), D756–D763.

47. Jiang, Z.; Hubley, R.; Smit, A.; Eichler, E. E. DupMasker: a Tool for Annotating Primate Segmental Duplications. *Genome Res.* **2008,** *18*, 1362–1368.

48. Harris, R. S. Improved Pairwise Alignment of Genomic DNA. Ph.D. Thesis, Pennsylvania State University 2007.

49. Kent, W. J.; Sugnet, C. W.; Furey, T. S.; Roskin, K. M.; Pringle, T. H.; Zahler, A. M.; Haussler, D. The Human Genome Browser at UCSC. *Genome Res.* **2002,** *12*, 996–1006.

50. Earl, D.; Nguyen, N.; Hickey, G.; Harris, R. S.; Fitzgerald, S.; Beal, K.; Seledtsov, I.; Molodtsov, V.; Raney, B. J.; Clawson, H.; Kim, J.; Kemena, C.; Chang, J. M.; Erb, I.; Poliakov, A.; Hou, M.; Herrero, J.; Kent, W. J.; Solovyev, V.; Darling, A. E.; Ma, J.; Notredame, C.; Brudno, M.; Dubchak, I.; Haussler, D.; Paten, B. Alignathon: a Competitive Assessment of Whole-Genome Alignment Methods. *Genome Res.* **2014,** *24*, 2077–2089.

51. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* **2009,** *25*, 2078–2079.

52. Cingolani, P.; Platts, A.; Wang le, L.; Coon, M.; Nguyen, T.; Wang, L.; Land, S. J.; Lu, X.; Ruden, D. M. A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff: SNPs in the Genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **2012,** *6*, 80–92.

53. Choi, Y.; Sims, G. E.; Murphy, S.; Miller, J. R.; Chan, A. P. Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS One* **2012,** *7*, e46688.

54. Zhang, B.; Kirov, S.; Snoddy, J. WebGestalt: an Integrated System for Exploring Gene Sets in Various Biological Contexts. *Nucleic Acids Res.* **2005,** *33* (Web Server issue), W741–W748.

55. Landrum, M. J.; Lee, J. M.; Riley, G. R.; Jang, W.; Rubinstein, W. S.; Church, D. M.; Maglott, D. R. ClinVar: Public Archive of Relationships among Sequence Variation and Human Phenotype. *Nucleic Acids Res.* **2014,** *42* (Database issue), D980–D985.

56. Sherry, S. T.; Ward, M. H.; Kholodov, M.; Baker, J.; Phan, L.; Smigielski, E. M.; Sirotkin, K. dbSNP: the NCBI Database of Genetic Variation. *Nucleic Acids Res.* **2001,** *29*, 308–311.

57. Li, Y.; Zheng, H.; Luo, R.; Wu, H.; Zhu, H.; Li, R.; Cao, H.; Wu, B.; Huang, S.; Shao, H.; Ma, H.; Zhang, F.; Feng, S.; Zhang, W.; Du, H.; Tian, G.; Li, J.; Zhang, X.; Li, S.; Bolund, L.; Kristiansen, K.; de Smith, A. J.; Blakemore, A. I.; Coin, L. J.; Yang, H.; Wang, J.; Wang, J. Structural Variation in Two Human Genomes Mapped at Single-Nucleotide Resolution by Whole Genome De novo Assembly. *Nat. Biotechnol.* **2011,** *29*, 723–730.

58. MacDonald, J. R.; Ziman, R.; Yuen, R. K.; Feuk, L.; Scherer, S. W. The Database of Genomic Variants: a Curated Collection of Structural Variation in the Human Genome. *Nucleic Acids Res.* **2014,** *42* (Database issue), D986–D992.

59. Wang, J.; Song, L.; Grover, D.; Azrak, S.; Batzer, M. A.; Liang, P. dbRIP: a Highly Integrated Database of Retrotransposon Insertion Polymorphisms in Humans. *Hum. Mutat.* **2006,** *27*, 323–329.

60. Mills, R. E.; Walter, K.; Stewart, C.; Handsaker, R. E.; Chen, K.; Alkan, C.; Abyzov, A.; Yoon, S. C.; Ye, K.; Cheetham, R. K.; Chinwalla, A.; Conrad, D. F.; Fu, Y.; Grubert, F.; Hajirasouliha, I.; Hormozdiari, F.; Iakoucheva, L. M.; Iqbal, Z.; Kang, S.; Kidd, J. M.; Konkel, M. K.; Korn, J.; Khurana, E.; Kural, D.; Lam, H. Y.; Leng, J.; Li, R.; Li, Y.; Lin, C. Y.; Luo, R.; Mu, X. J.; Nemesh, J.; Peckham, H. E.; Rausch, T.; Scally, A.; Shi, X.; Stromberg, M. P.; Stütz, A. M.; Urban, A. E.; Walker, J. A.; Wu, J.; Zhang, Y.; Zhang, Z. D.; Batzer, M. A.; Ding, L.; Marth, G. T.; McVean, G.; Sebat, J.; Snyder, M.; Wang, J.; Ye, K.; Eichler, E. E.; Gerstein, M. B.; Hurles, M. E.; Lee, C.;

McCarroll, S. A.; Korbel, J. O.; 1000 Genomes Project. Mapping Copy Number Variation by Population-scale Genome Sequencing. *Nature* **2011,** *470*, 59–65.

61. International HapMap 3 Consortium; Altshuler, D. M.; Gibbs, R. A.; Peltonen, L.; Altshuler, D. M.; Gibbs, R. A.; Peltonen, L.; Dermitzakis, E.; Schaffner, S. F.; Yu, F.; Peltonen, L.; Dermitzakis, E.; Bonnen, P. E.; Altshuler, D. M.; Gibbs, R. A.; de Bakker, P. I.; Deloukas, P.; Gabriel, S. B.; Gwilliam, R.; Hunt, S.; Inouye, M.; Jia, X.; Palotie, A.; Parkin, M.; Whittaker, P.; Yu, F.; Chang, K.; Hawes, A.; Lewis, L. R.; Ren, Y.; Wheeler, D.; Gibbs, R. A.; Muzny, D. M.; Barnes, C.; Darvishi, K.; Hurles, M.; Korn, J. M.; Kristiansson, K.; Lee, C.; McCarrol, S. A.; Nemesh, J.; Dermitzakis, E.; Keinan, A.; Montgomery, S. B.; Pollack, S.; Price, A. L.; Soranzo, N.; Bonnen, P. E.; Gibbs, R. A.; Gonzaga-Jauregui, C.; Keinan, A.; Price, A. L.; Yu, F.; Anttila, V.; Brodeur, W.; Daly, M. J.; Leslie, S.; McVean, G.; Moutsianas, L.; Nguyen, H.; Schaffner, S. F.; Zhang, Q.; Ghori, M. J.; McGinnis, R.; McLaren, W.; Pollack, S.; Price, A. L.; Schaffner, S. F.; Takeuchi, F.; Grossman, S. R.; Shlyakhter, I.; Hostetter, E. B.; Sabeti, P. C.; Adebamowo, C. A.; Foster, M. W.; Gordon, D. R.; Licinio, J.; Manca, M. C.; Marshall, P. A.; Matsuda, I.; Ngare, D.; Wang, V. O.; Reddy, D.; Rotimi, C. N.; Royal, C. D.; Sharp, R. R.; Zeng, C.; Brooks, L. D.; McEwen, J. E. Integrating Common and Rare Genetic Variation in Diverse Human Populations. *Nature* **2010,** *467*, 52–58.

62. Seo, J. S.; Rhie, A.; Kim, J.; Lee, S.; Sohn, M. H.; Kim, C. U.; Hastie, A.; Cao, H.; Yun, J. Y.; Kim, J.; Kuk, J.; Park, G. H.; Ryu, H.; Kim, J.; Roh, M.; Baek, J.; Hunkapiller, M. W.; Korlach, J.; Shin, J. Y.; Kim, C. De novo Assembly and Phasing of a Korean Human Genome. *Nature* **2016,** *538*, 243–247.

63. Church, D. M.; Schneider, V. A.; Graves, T.; Auger, K.; Cunningham, F.; Bouk, N.; Chen, H. C.; Agarwala, R.; McLaren, W. M.; Ritchie, G. R.; 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, E. E.; Weinstock, G.; Mardis, E. R.; Wilson, R. K.; Howe, K.; Flicek, P.; Hubbard, T. Modernizing Reference Genome Assemblies. *PLoS Biol.* **2011,** *9*, e1001091.

64. Koren, S.; Schatz, M. C.; Walenz, B. P.; Martin, J.; Howard, J. T.; Ganapathy, G.; Wang, Z.; Rasko, D. A.; McCombie, W. R.; Jarvis, E. D.; Adam, M. P. Hybrid Error Correction and De novo Assembly of Single-Molecule Sequencing Reads. *Nat. Biotechnol.* **2012,** *30*, 693–700.

65. Kersbergen, P.; van Duijn, K.; Kloosterman, A. D.; den Dunnen, J. T.; Kayser, M.; de Knijff, P. Developing a Set of Ancestry-Sensitive DNA Markers Reflecting Continental Origins of Humans. *BMC Genet.* **2009,** *10*, 69.

66. 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 Pan-Genome. *Nat. Biotechnol.* **2010,** *28*, 57–63.

67. Prüfer, K.; Racimo, F.; Patterson, N.; Jay, F.; Sankararaman, S.; Sawyer, S.; Heinze, A.; Renaud, G.; Sudmant, P. H.; de Filippo, C.; Li, H.; Mallick, S.; Dannemann, M.; Fu, Q.; Kircher, M.; Kuhlwilm, M.; Lachmann, M.; Meyer, M.; Ongyerth, M.; Siebauer, M.; Theunert, C.; Tandon, A.; Moorjani, P.; Pickrell, J.; Mullikin, J. C.; Vohr, S. H.; Green, R. E.; Hellmann, I.; Johnson, P. L.; Blanche, H.; Cann, H.; Kitzman, J. O.; Shendure, J.; Eichler, E. E.; Lein, E. S.; Bakken, T. E.; Golovanova, L. V.; Doronichev, V. B.; Shunkov, M. V.; Derevianko, A. P.; Viola, B.; Slatkin, M.; Reich, D.; Kelso, J.; Pääbo, S. The Complete Genome Sequence of a Neanderthal from the Altai Mountains. *Nature* **2014,** *505*, 43–49.

68. Chen, R.; Butte, A. J. The Reference Human Genome Demonstrates High Risk of Type 1 Diabetes and Other Disorders. *Pac. Symp. Biocomput.* **2011**, 231–242.

69. Rosenfeld, J. A.; Mason, C. E.; Smith, T. M. Limitations of the Human Reference Genome for Personalized Genomics. *PLoS One* **2012,** *7*, e40294.

70. Price, A. L.; Patterson, N. J.; Plenge, R. M.; Weinblatt, M. E.; Shadick, N. A.; Reich, D. Principal Components Analysis Corrects for Stratification in Genome-wide Association Studies. *Nat. Genet.* **2006,** *38*, 904–909.

71. Jung, H.; Bleazard, T.; Lee, J.; Hong, D. Systematic Investigation of Cancer-Associated Somatic Point Mutations in SNP Databases. *Nat. Biotechnol.* **2013,** *31*, 787–789.

72. Burton, J. N.; Adey, A.; Patwardhan, R. P.; Qiu, R.; Kitzman, J. O.; Shendure, J. Chromosome-scale Scaffolding of De novo Genome Assemblies based on Chromatin Interactions. *Nat. Biotechnol.* **2013,** *31*, 1119–1125.

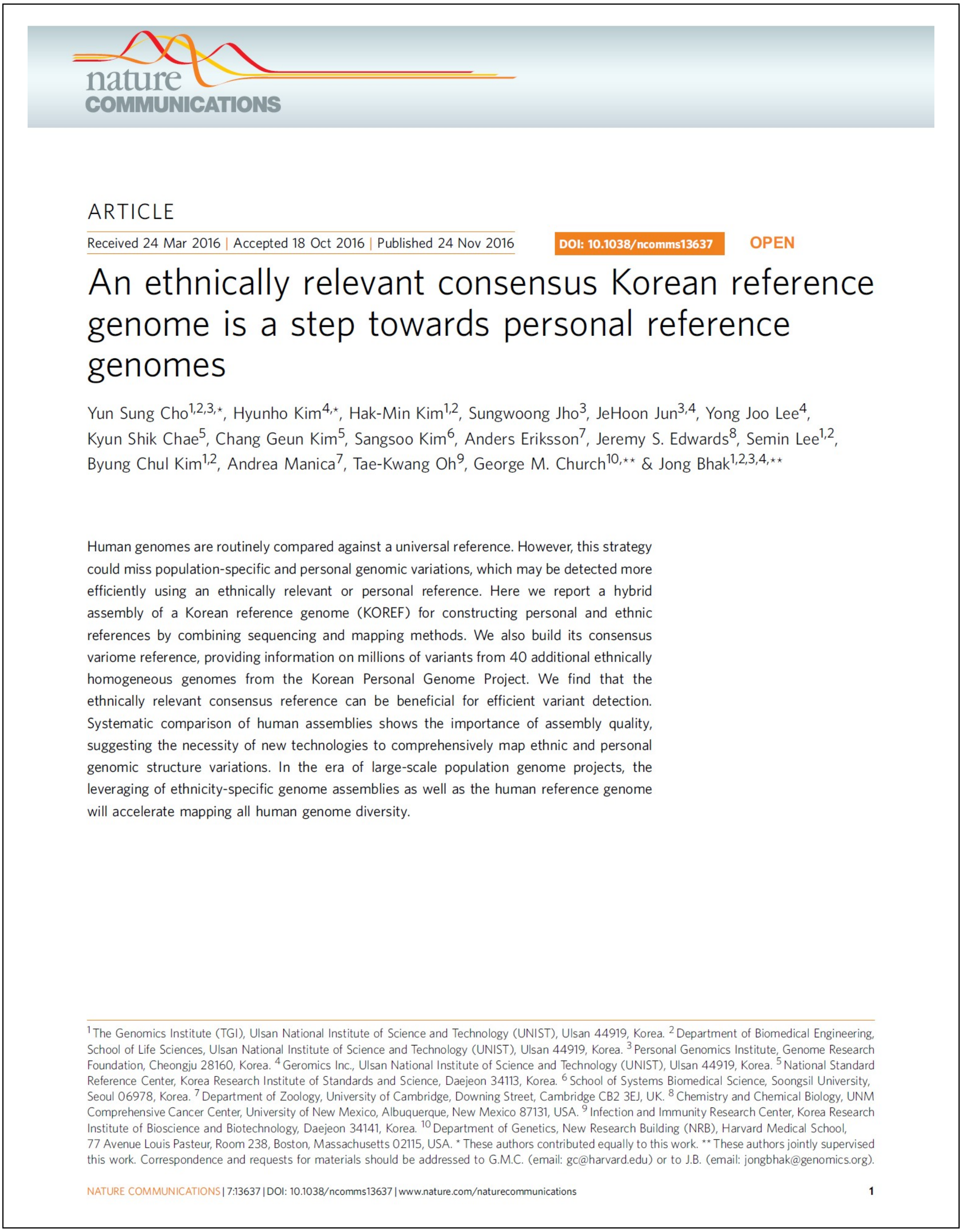
73. Howorka, S.; Siwy, Z. Nanopores and Nanochannels: From Gene Sequencing to Genome Mapping. *ACS Nano* **2016,** *10*, 9768–9771.

Acknowledgements

This work was supported by the Ministry of Trade, Industry & Energy (MOTIE, Korea) under Industrial Technology Innovation Programs (‘Pilot study of building of Korean Reference Standard Genome map’, No.10046043; ‘Developing Korean Reference Genome’, No.10050164; and ‘National Center for Standard Reference Data’, No.10063239) and Industrial Strategic Technology Development Program (‘Bioinformatics platform development for next generation bioinformation analysis’, No.10040231). This work was also supported by the Korea Research Institute of Bioscience and Biotechnology (KRIBB) under ‘Bioinformatics pipeline construction for de novo assembly’ program. This work was also supported by ‘Software Convergence Technology Development Program’ through the Ministry of Science, ICT and Future Planning (S0177-16-1046). This work was also supported by the 2015 Research fund (1.150014.01) of Ulsan National Institute of Science & Technology (UNIST). This work was also supported by the Ulsan city’s Genome Korea Project. This work was also supported by the Research Fund (14-BR-SS-03) of Civil-Military Technology Cooperation Program. Part of KPGP was supported by KT (Korea Telecom) Personal Genome Project grant and Genome Research Foundation. Korea Institute of Science and Technology Information (KISTI) provided us with Korea Research Environment Open NETwork (KREONET), which is the internet connection service for efficient information and data transfer.

Appendix

The Korean reference genome



nature COMMUNICATIONS

ARTICLE

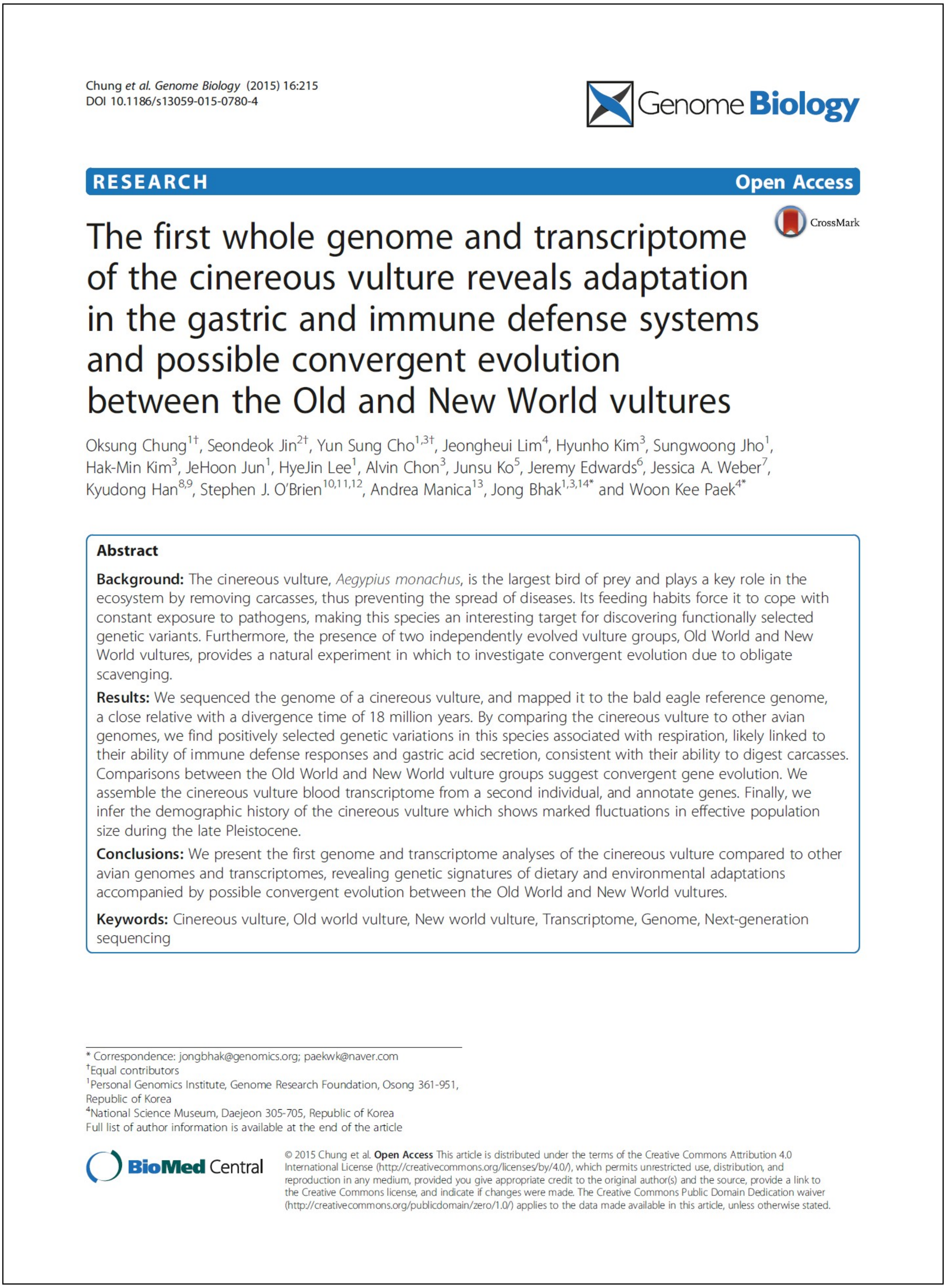
Received 24 Mar 2016 | Accepted 18 Oct 2016 | Published 24 Nov 2016 DOI: 10.1038/ncomms13637 OPEN

An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes

Yun Sung Cho[1,2,3,*], Hyunho Kim[4,*], Hak-Min Kim[1,2], Sungwoong Jho[3], JeHoon Jun[3,4], Yong Joo Lee[4], Kyun Shik Chae[5], Chang Geun Kim[5], Sangsoo Kim[6], Anders Eriksson[7], Jeremy S. Edwards[8], Semin Lee[1,2], Byung Chul Kim[1,2], Andrea Manica[7], Tae-Kwang Oh[9], George M. Church[10,**] & Jong Bhak[1,2,3,4,**]

Human genomes are routinely compared against a universal reference. However, this strategy could miss population-specific and personal genomic variations, which may be detected more efficiently using an ethnically relevant or personal reference. Here we report a hybrid assembly of a Korean reference genome (KOREF) for constructing personal and ethnic references by combining sequencing and mapping methods. We also build its consensus variome reference, providing information on millions of variants from 40 additional ethnically homogeneous genomes from the Korean Personal Genome Project. We find that the ethnically relevant consensus reference can be beneficial for efficient variant detection. Systematic comparison of human assemblies shows the importance of assembly quality, suggesting the necessity of new technologies to comprehensively map ethnic and personal genomic structure variations. In the era of large-scale population genome projects, the leveraging of ethnicity-specific genome assemblies as well as the human reference genome will accelerate mapping all human genome diversity.

[1] The Genomics Institute (TGI), Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Korea. [2] Department of Biomedical Engineering, School of Life Sciences, Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Korea. [3] Personal Genomics Institute, Genome Research Foundation, Cheongju 28160, Korea. [4] Geromics Inc., Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Korea. [5] National Standard Reference Center, Korea Research Institute of Standards and Science, Daejeon 34113, Korea. [6] School of Systems Biomedical Science, Soongsil University, Seoul 06978, Korea. [7] Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK. [8] Chemistry and Chemical Biology, UNM Comprehensive Cancer Center, University of New Mexico, Albuquerque, New Mexico 87131, USA. [9] Infection and Immunity Research Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon 34141, Korea. [10] Department of Genetics, New Research Building (NRB), Harvard Medical School, 77 Avenue Louis Pasteur, Room 238, Boston, Massachusetts 02115, USA. * These authors contributed equally to this work. ** These authors jointly supervised this work. Correspondence and requests for materials should be addressed to G.M.C. (email: gc@harvard.edu) or to J.B. (email: jongbhak@genomics.org).



Chung *et al. Genome Biology* (2015) 16:215
DOI 10.1186/s13059-015-0780-4

Genome Biology

RESEARCH Open Access

The first whole genome and transcriptome of the cinereous vulture reveals adaptation in the gastric and immune defense systems and possible convergent evolution between the Old and New World vultures

Oksung Chung[1†], Seondeok Jin[2†], Yun Sung Cho[1,3†], Jeongheui Lim[4], Hyunho Kim[3], Sungwoong Jho[1], Hak-Min Kim[3], JeHoon Jun[1], HyeJin Lee[1], Alvin Chon[3], Junsu Ko[5], Jeremy Edwards[6], Jessica A. Weber[7], Kyudong Han[8,9], Stephen J. O'Brien[10,11,12], Andrea Manica[13], Jong Bhak[1,3,14*] and Woon Kee Paek[4*]

Abstract

Background: The cinereous vulture, *Aegypius monachus*, is the largest bird of prey and plays a key role in the ecosystem by removing carcasses, thus preventing the spread of diseases. Its feeding habits force it to cope with constant exposure to pathogens, making this species an interesting target for discovering functionally selected genetic variants. Furthermore, the presence of two independently evolved vulture groups, Old World and New World vultures, provides a natural experiment in which to investigate convergent evolution due to obligate scavenging.

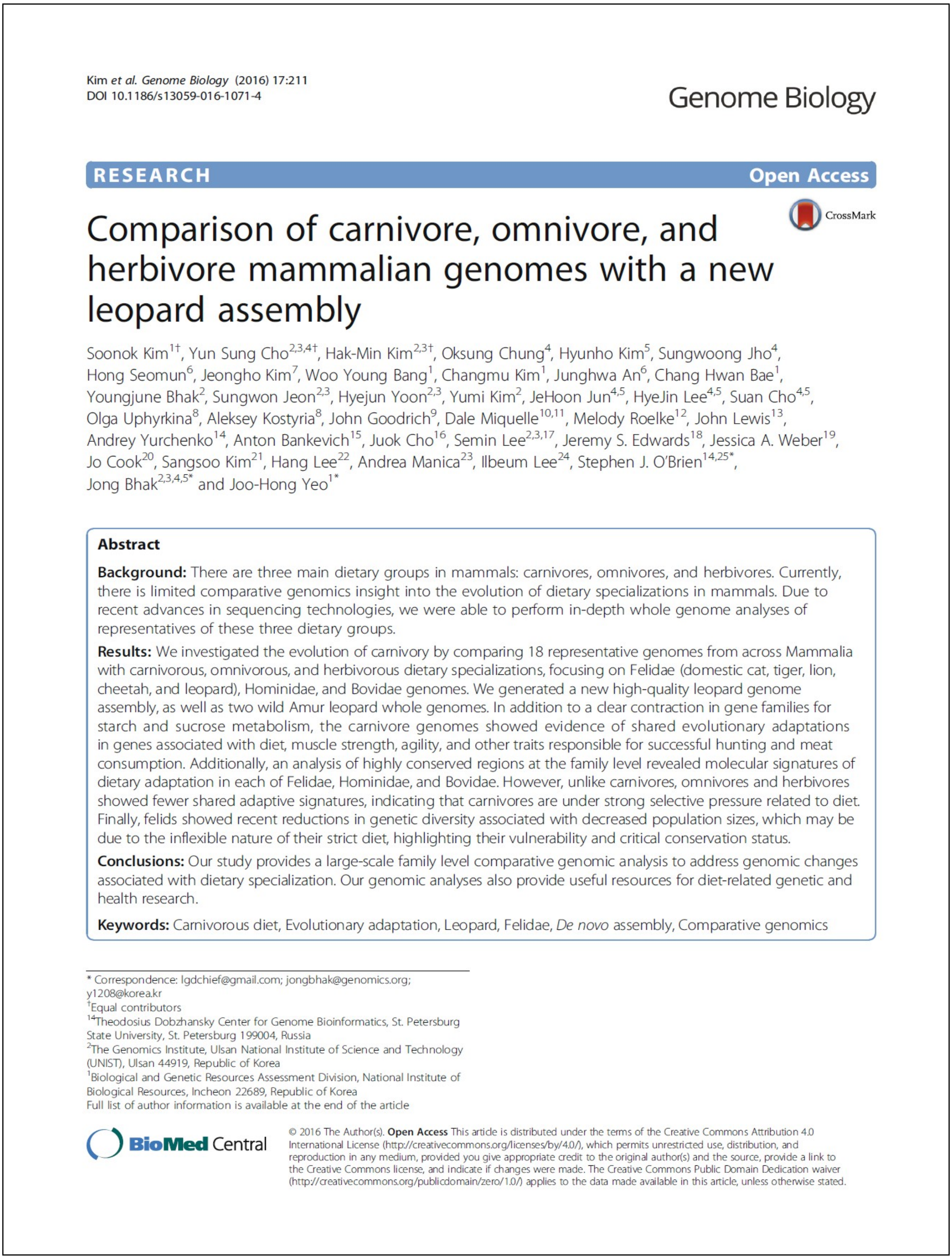
Results: We sequenced the genome of a cinereous vulture, and mapped it to the bald eagle reference genome, a close relative with a divergence time of 18 million years. By comparing the cinereous vulture to other avian genomes, we find positively selected genetic variations in this species associated with respiration, likely linked to their ability of immune defense responses and gastric acid secretion, consistent with their ability to digest carcasses. Comparisons between the Old World and New World vulture groups suggest convergent gene evolution. We assemble the cinereous vulture blood transcriptome from a second individual, and annotate genes. Finally, we infer the demographic history of the cinereous vulture which shows marked fluctuations in effective population size during the late Pleistocene.

Conclusions: We present the first genome and transcriptome analyses of the cinereous vulture compared to other avian genomes and transcriptomes, revealing genetic signatures of dietary and environmental adaptations accompanied by possible convergent evolution between the Old World and New World vultures.

Keywords: Cinereous vulture, Old world vulture, New world vulture, Transcriptome, Genome, Next-generation sequencing

* Correspondence: jongbhak@genomics.org; paekwk@naver.com
†Equal contributors
[1]Personal Genomics Institute, Genome Research Foundation, Osong 361-951, Republic of Korea
[4]National Science Museum, Daejeon 305-705, Republic of Korea
Full list of author information is available at the end of the article

The leopard reference genome



Kim *et al. Genome Biology* (2016) 17:211
DOI 10.1186/s13059-016-1071-4

Genome Biology

RESEARCH Open Access

Comparison of carnivore, omnivore, and herbivore mammalian genomes with a new leopard assembly

Soonok Kim[1†], Yun Sung Cho[2,3,4†], Hak-Min Kim[2,3†], Oksung Chung[4], Hyunho Kim[5], Sungwoong Jho[4], Hong Seomun[6], Jeongho Kim[7], Woo Young Bang[1], Changmu Kim[1], Junghwa An[6], Chang Hwan Bae[1], Youngjune Bhak[2], Sungwon Jeon[2,3], Hyejun Yoon[2,3], Yumi Kim[2], JeHoon Jun[4,5], HyeJin Lee[4,5], Suan Cho[4,5], Olga Uphyrkina[8], Aleksey Kostyria[8], John Goodrich[9], Dale Miquelle[10,11], Melody Roelke[12], John Lewis[13], Andrey Yurchenko[14], Anton Bankevich[15], Juok Cho[16], Semin Lee[2,3,17], Jeremy S. Edwards[18], Jessica A. Weber[19], Jo Cook[20], Sangsoo Kim[21], Hang Lee[22], Andrea Manica[23], Ilbeum Lee[24], Stephen J. O'Brien[14,25*], Jong Bhak[2,3,4,5*] and Joo-Hong Yeo[1*]

Abstract

Background: There are three main dietary groups in mammals: carnivores, omnivores, and herbivores. Currently, there is limited comparative genomics insight into the evolution of dietary specializations in mammals. Due to recent advances in sequencing technologies, we were able to perform in-depth whole genome analyses of representatives of these three dietary groups.

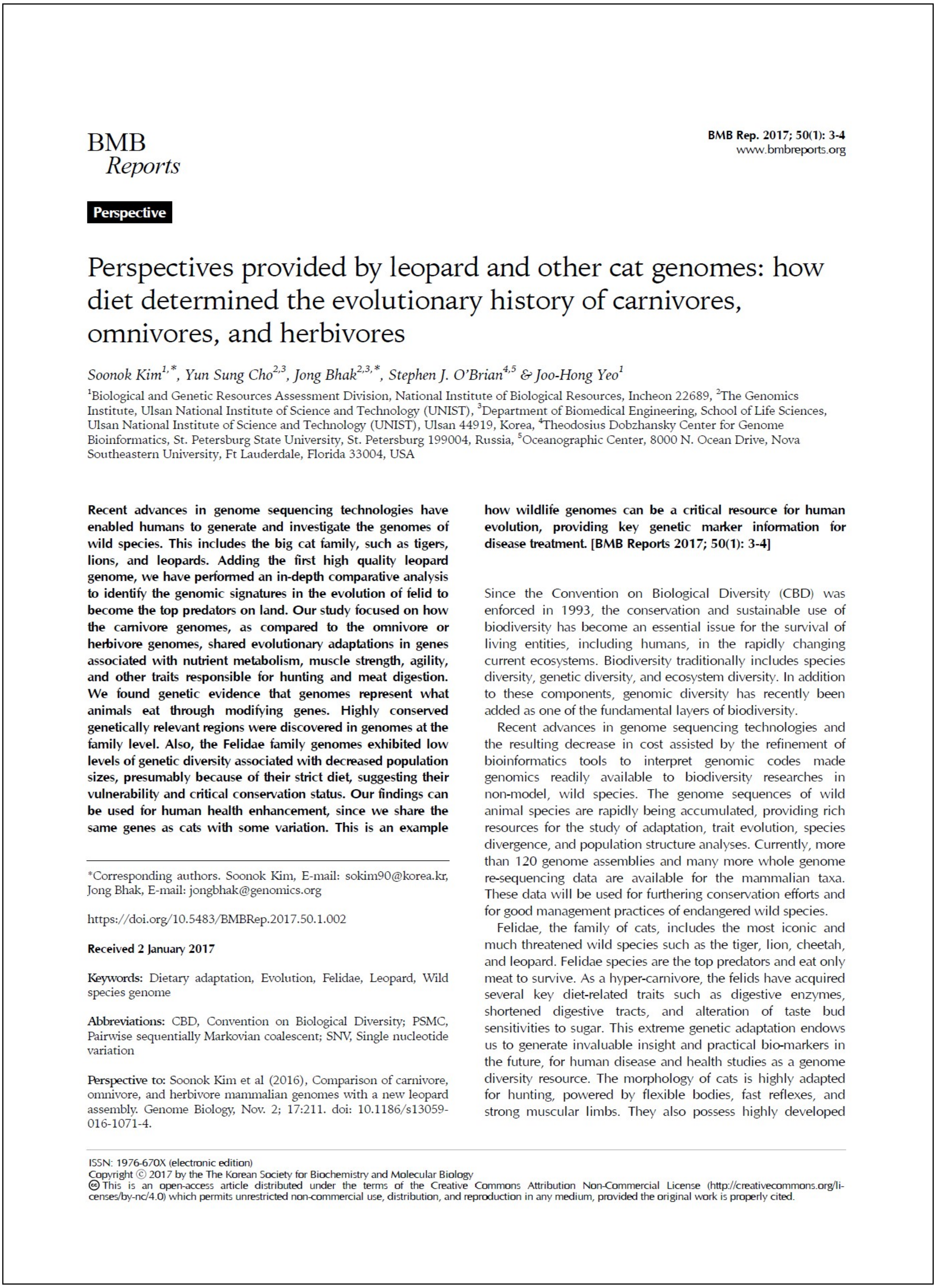
Results: We investigated the evolution of carnivory by comparing 18 representative genomes from across Mammalia with carnivorous, omnivorous, and herbivorous dietary specializations, focusing on Felidae (domestic cat, tiger, lion, cheetah, and leopard), Hominidae, and Bovidae genomes. We generated a new high-quality leopard genome assembly, as well as two wild Amur leopard whole genomes. In addition to a clear contraction in gene families for starch and sucrose metabolism, the carnivore genomes showed evidence of shared evolutionary adaptations in genes associated with diet, muscle strength, agility, and other traits responsible for successful hunting and meat consumption. Additionally, an analysis of highly conserved regions at the family level revealed molecular signatures of dietary adaptation in each of Felidae, Hominidae, and Bovidae. However, unlike carnivores, omnivores and herbivores showed fewer shared adaptive signatures, indicating that carnivores are under strong selective pressure related to diet. Finally, felids showed recent reductions in genetic diversity associated with decreased population sizes, which may be due to the inflexible nature of their strict diet, highlighting their vulnerability and critical conservation status.

Conclusions: Our study provides a large-scale family level comparative genomic analysis to address genomic changes associated with dietary specialization. Our genomic analyses also provide useful resources for diet-related genetic and health research.

Keywords: Carnivorous diet, Evolutionary adaptation, Leopard, Felidae, *De novo* assembly, Comparative genomics

* Correspondence: lgdchief@gmail.com; jongbhak@genomics.org; y1208@korea.kr
[†]Equal contributors
[14]Theodosius Dobzhansky Center for Genome Bioinformatics, St. Petersburg State University, St. Petersburg 199004, Russia
[2]The Genomics Institute, Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea
[1]Biological and Genetic Resources Assessment Division, National Institute of Biological Resources, Incheon 22689, Republic of Korea
Full list of author information is available at the end of the article

Perspectives provided by leopard reference genome



BMB
Reports

BMB Rep. 2017; 50(1): 3-4
www.bmbreports.org

Perspective

Perspectives provided by leopard and other cat genomes: how diet determined the evolutionary history of carnivores, omnivores, and herbivores

Soonok Kim[1,*], *Yun Sung Cho*[2,3], *Jong Bhak*[2,3,*], *Stephen J. O'Brian*[4,5] *& Joo-Hong Yeo*[1]

[1]Biological and Genetic Resources Assessment Division, National Institute of Biological Resources, Incheon 22689, [2]The Genomics Institute, Ulsan National Institute of Science and Technology (UNIST), [3]Department of Biomedical Engineering, School of Life Sciences, Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Korea, [4]Theodosius Dobzhansky Center for Genome Bioinformatics, St. Petersburg State University, St. Petersburg 199004, Russia, [5]Oceanographic Center, 8000 N. Ocean Drive, Nova Southeastern University, Ft Lauderdale, Florida 33004, USA

Recent advances in genome sequencing technologies have enabled humans to generate and investigate the genomes of wild species. This includes the big cat family, such as tigers, lions, and leopards. Adding the first high quality leopard genome, we have performed an in-depth comparative analysis to identify the genomic signatures in the evolution of felid to become the top predators on land. Our study focused on how the carnivore genomes, as compared to the omnivore or herbivore genomes, shared evolutionary adaptations in genes associated with nutrient metabolism, muscle strength, agility, and other traits responsible for hunting and meat digestion. We found genetic evidence that genomes represent what animals eat through modifying genes. Highly conserved genetically relevant regions were discovered in genomes at the family level. Also, the Felidae family genomes exhibited low levels of genetic diversity associated with decreased population sizes, presumably because of their strict diet, suggesting their vulnerability and critical conservation status. Our findings can be used for human health enhancement, since we share the same genes as cats with some variation. This is an example how wildlife genomes can be a critical resource for human evolution, providing key genetic marker information for disease treatment. [BMB Reports 2017; 50(1): 3-4]

Since the Convention on Biological Diversity (CBD) was enforced in 1993, the conservation and sustainable use of biodiversity has become an essential issue for the survival of living entities, including humans, in the rapidly changing current ecosystems. Biodiversity traditionally includes species diversity, genetic diversity, and ecosystem diversity. In addition to these components, genomic diversity has recently been added as one of the fundamental layers of biodiversity.

Recent advances in genome sequencing technologies and the resulting decrease in cost assisted by the refinement of bioinformatics tools to interpret genomic codes made genomics readily available to biodiversity researches in non-model, wild species. The genome sequences of wild animal species are rapidly being accumulated, providing rich resources for the study of adaptation, trait evolution, species divergence, and population structure analyses. Currently, more than 120 genome assemblies and many more whole genome re-sequencing data are available for the mammalian taxa. These data will be used for furthering conservation efforts and for good management practices of endangered wild species.

Felidae, the family of cats, includes the most iconic and much threatened wild species such as the tiger, lion, cheetah, and leopard. Felidae species are the top predators and eat only meat to survive. As a hyper-carnivore, the felids have acquired several key diet-related traits such as digestive enzymes, shortened digestive tracts, and alteration of taste bud sensitivities to sugar. This extreme genetic adaptation endows us to generate invaluable insight and practical bio-markers in the future, for human disease and health studies as a genome diversity resource. The morphology of cats is highly adapted for hunting, powered by flexible bodies, fast reflexes, and strong muscular limbs. They also possess highly developed

*Corresponding authors. Soonok Kim, E-mail: sokim90@korea.kr, Jong Bhak, E-mail: jongbhak@genomics.org

https://doi.org/10.5483/BMBRep.2017.50.1.002

Received 2 January 2017

Keywords: Dietary adaptation, Evolution, Felidae, Leopard, Wild species genome

Abbreviations: CBD, Convention on Biological Diversity; PSMC, Pairwise sequentially Markovian coalescent; SNV, Single nucleotide variation

Perspective to: Soonok Kim et al (2016), Comparison of carnivore, omnivore, and herbivore mammalian genomes with a new leopard assembly. Genome Biology, Nov. 2; 17:211. doi: 10.1186/s13059-016-1071-4.

ISSN: 1976-670X (electronic edition)

Neolithic pastoralist genome

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

The genetics of an early Neolithic pastoralist from the Zagros, Iran

M. Gallego-Llorente[1], S. Connell[2], E. R. Jones[1], D. C. Merrett[3], Y. Jeon[4,5], A. Eriksson[1,6], V. Siska[1], C. Gamba[2,7], C. Meiklejohn[8], R. Beyer[9], S. Jeon[4,5], Y. S. Cho[4,5], M. Hofreiter[10], J. Bhak[4], A. Manica[1,*] & R. Pinhasi[2,*]

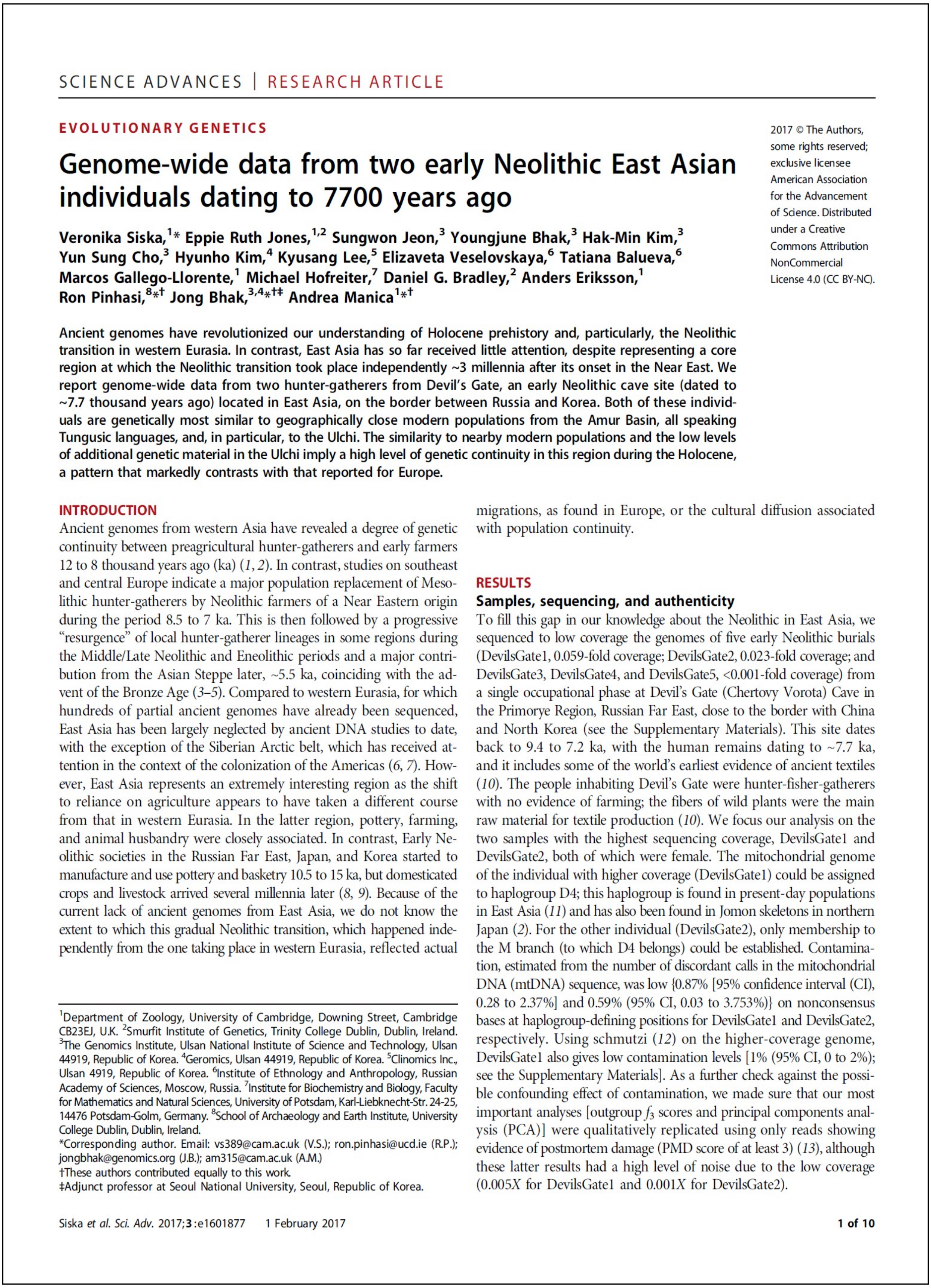
Received: 28 May 2016
Accepted: 15 July 2016
Published: 09 August 2016

The agricultural transition profoundly changed human societies. We sequenced and analysed the first genome (1.39x) of an early Neolithic woman from Ganj Dareh, in the Zagros Mountains of Iran, a site with early evidence for an economy based on goat herding, ca. 10,000 BP. We show that Western Iran was inhabited by a population genetically most similar to hunter-gatherers from the Caucasus, but distinct from the Neolithic Anatolian people who later brought food production into Europe. The inhabitants of Ganj Dareh made little direct genetic contribution to modern European populations, suggesting those of the Central Zagros were somewhat isolated from other populations of the Fertile Crescent. Runs of homozygosity are of a similar length to those from Neolithic farmers, and shorter than those of Caucasus and Western Hunter-Gatherers, suggesting that the inhabitants of Ganj Dareh did not undergo the large population bottleneck suffered by their northern neighbours. While some degree of cultural diffusion between Anatolia, Western Iran and other neighbouring regions is possible, the genetic dissimilarity between early Anatolian farmers and the inhabitants of Ganj Dareh supports a model in which Neolithic societies in these areas were distinct.

The agricultural transition started in a region comprising the Ancient Near East and Anatolia ~12,000 years ago with the first Pre-Pottery Neolithic villages and the first domestication of cereals and legumes[1,2]. Archaeological evidence suggests a complex scenario of multiple domestications in a number of areas[3], coupled with examples of trade[4]. Ancient DNA (aDNA) has revealed that this cultural package was later brought into Europe by dispersing farmers from Anatolia (so called 'demic' diffusion, as opposed to non-demic cultural diffusion[5,6]) ~8,400 years ago. However a lack of aDNA from early Neolithic individuals from the Near East leaves a key question unanswered: was the agricultural transition developed by one major population group spanning the Near East, including Anatolia and the Central Zagros Mountains; or was the region inhabited by genetically diverse populations, as is suggested by the heterogeneous mode and timing of the appearance of early domesticates at different localities?

To answer this question, we sequenced the genome of an early Neolithic female from Ganj Dareh, GD13a, from the Central Zagros (Western Iran), dated to 10000-9700 cal BP[7], a region located at the eastern edge of the Near East. Ganj Dareh is well known for providing the earliest evidence of herd management of goats beginning at 9,900 BP[7–9]. It is a classic mound site at an altitude of ~1400 m in the Gamas-Ab Valley of the High Zagros zone in Kermanshah Province, Western Iran. It was discovered in the 1960s during survey work and excavated over four seasons between 1967 and 1974. The mound, ~40 m in diameter, shows 7 to 8 m of early Neolithic cultural deposits. Five major levels were found, labelled A through E from top to bottom. Extended evidence showed a

[1]Department of Zoology, University of Cambridge, Cambridge, CB2 3EJ, UK. [2]School of Archaeology and Earth Institute, University College Dublin, Belfield, Dublin 4, Ireland. [3]Department of Archaeology, Simon Fraser University, Burnaby, BC V5A 1S6, Canada. [4]The Genomics Institute, Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea. [5]Department of Biomedical Engineering, School of Life Sciences, Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea. [6]Integrative Systems Biology Laboratory, Division of Biological and Environmental Sciences & Engineering, King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia. [7]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5–7, Copenhagen 1350, Denmark. [8]Department of Anthropology, University of Winnipeg, Winnipeg, MB R3B 2E9, Canada. [9]McDonald Institute for Archaeological Research, University of Cambridge, Cambridge CB2 3ER, UK. [10]Evolutionary Adaptive Genomics, Institute for Biochemistry and Biology, Department of Mathematics and Natural Sciences, University of Potsdam, Karl-Liebknechtstraße 24-25, Potsdam, 14476, Germany. [*]These authors jointly supervised this work. Correspondence and requests for materials should be addressed to M.G.-L. (email: mg632@cam.ac.uk) or A.M. (email: am315@cam.ac.uk) or R.P. (email: ron.pinhasi@ucd.ie)



SCIENCE ADVANCES | RESEARCH ARTICLE

EVOLUTIONARY GENETICS

Genome-wide data from two early Neolithic East Asian individuals dating to 7700 years ago

Veronika Siska,[1]* Eppie Ruth Jones,[1,2] Sungwon Jeon,[3] Youngjune Bhak,[3] Hak-Min Kim,[3] Yun Sung Cho,[3] Hyunho Kim,[4] Kyusang Lee,[5] Elizaveta Veselovskaya,[6] Tatiana Balueva,[6] Marcos Gallego-Llorente,[1] Michael Hofreiter,[7] Daniel G. Bradley,[2] Anders Eriksson,[1] Ron Pinhasi,[8]*† Jong Bhak,[3,4]*†‡ Andrea Manica[1]*†

Ancient genomes have revolutionized our understanding of Holocene prehistory and, particularly, the Neolithic transition in western Eurasia. In contrast, East Asia has so far received little attention, despite representing a core region at which the Neolithic transition took place independently ~3 millennia after its onset in the Near East. We report genome-wide data from two hunter-gatherers from Devil's Gate, an early Neolithic cave site (dated to ~7.7 thousand years ago) located in East Asia, on the border between Russia and Korea. Both of these individuals are genetically most similar to geographically close modern populations from the Amur Basin, all speaking Tungusic languages, and, in particular, to the Ulchi. The similarity to nearby modern populations and the low levels of additional genetic material in the Ulchi imply a high level of genetic continuity in this region during the Holocene, a pattern that markedly contrasts with that reported for Europe.

INTRODUCTION

Ancient genomes from western Asia have revealed a degree of genetic continuity between preagricultural hunter-gatherers and early farmers 12 to 8 thousand years ago (ka) (*1*, *2*). In contrast, studies on southeast and central Europe indicate a major population replacement of Mesolithic hunter-gatherers by Neolithic farmers of a Near Eastern origin during the period 8.5 to 7 ka. This is then followed by a progressive "resurgence" of local hunter-gatherer lineages in some regions during the Middle/Late Neolithic and Eneolithic periods and a major contribution from the Asian Steppe later, ~5.5 ka, coinciding with the advent of the Bronze Age (*3*–*5*). Compared to western Eurasia, for which hundreds of partial ancient genomes have already been sequenced, East Asia has been largely neglected by ancient DNA studies to date, with the exception of the Siberian Arctic belt, which has received attention in the context of the colonization of the Americas (*6*, *7*). However, East Asia represents an extremely interesting region as the shift to reliance on agriculture appears to have taken a different course from that in western Eurasia. In the latter region, pottery, farming, and animal husbandry were closely associated. In contrast, Early Neolithic societies in the Russian Far East, Japan, and Korea started to manufacture and use pottery and basketry 10.5 to 15 ka, but domesticated crops and livestock arrived several millennia later (*8*, *9*). Because of the current lack of ancient genomes from East Asia, we do not know the extent to which this gradual Neolithic transition, which happened independently from the one taking place in western Eurasia, reflected actual migrations, as found in Europe, or the cultural diffusion associated with population continuity.

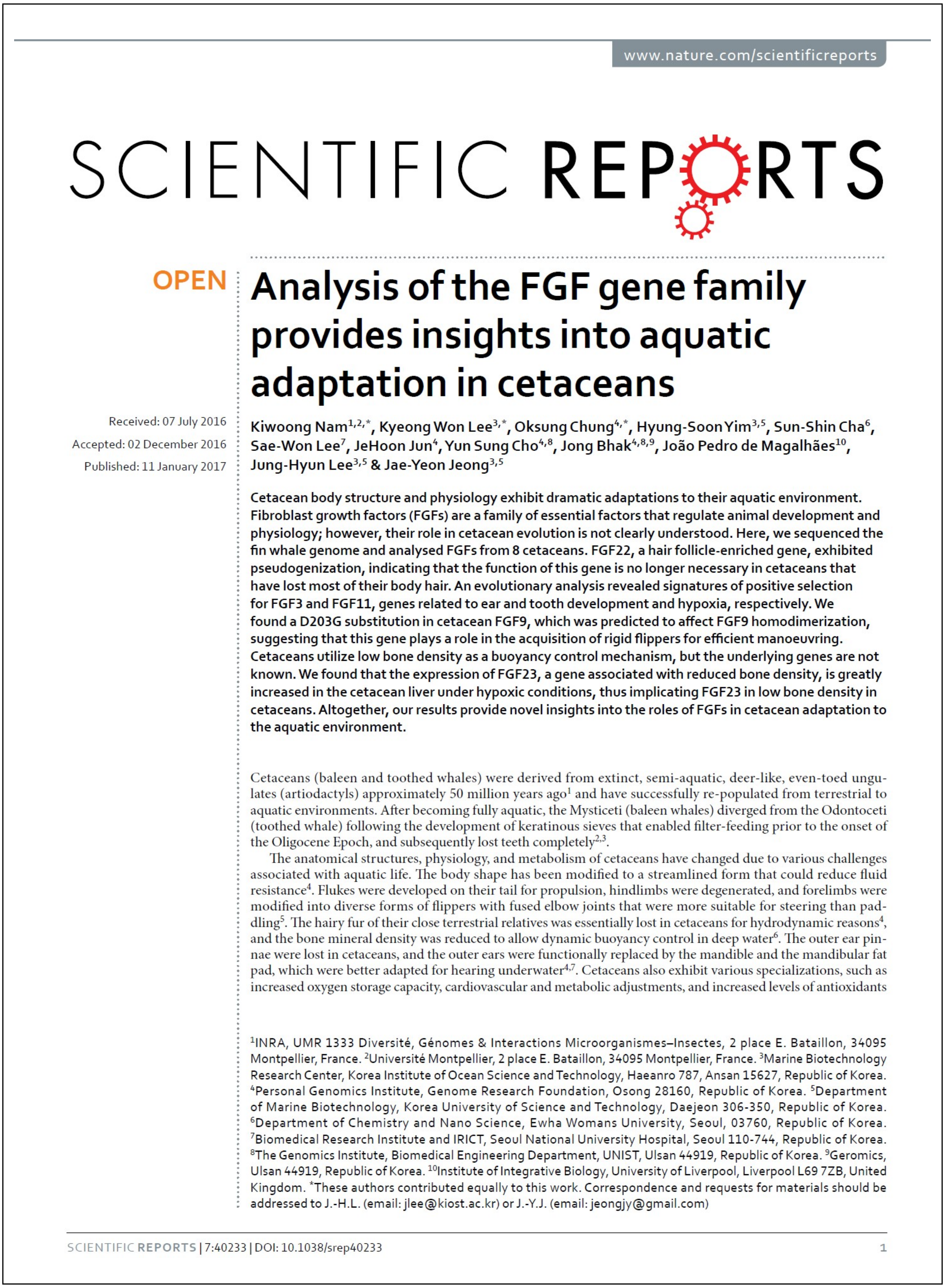
RESULTS

Samples, sequencing, and authenticity

To fill this gap in our knowledge about the Neolithic in East Asia, we sequenced to low coverage the genomes of five early Neolithic burials (DevilsGate1, 0.059-fold coverage; DevilsGate2, 0.023-fold coverage; and DevilsGate3, DevilsGate4, and DevilsGate5, <0.001-fold coverage) from a single occupational phase at Devil's Gate (Chertovy Vorota) Cave in the Primorye Region, Russian Far East, close to the border with China and North Korea (see the Supplementary Materials). This site dates back to 9.4 to 7.2 ka, with the human remains dating to ~7.7 ka, and it includes some of the world's earliest evidence of ancient textiles (*10*). The people inhabiting Devil's Gate were hunter-fisher-gatherers with no evidence of farming; the fibers of wild plants were the main raw material for textile production (*10*). We focus our analysis on the two samples with the highest sequencing coverage, DevilsGate1 and DevilsGate2, both of which were female. The mitochondrial genome of the individual with higher coverage (DevilsGate1) could be assigned to haplogroup D4; this haplogroup is found in present-day populations in East Asia (*11*) and has also been found in Jomon skeletons in northern Japan (*2*). For the other individual (DevilsGate2), only membership to the M branch (to which D4 belongs) could be established. Contamination, estimated from the number of discordant calls in the mitochondrial DNA (mtDNA) sequence, was low {0.87% [95% confidence interval (CI), 0.28 to 2.37%] and 0.59% (95% CI, 0.03 to 3.753%)} on nonconsensus bases at haplogroup-defining positions for DevilsGate1 and DevilsGate2, respectively. Using schmutzi (*12*) on the higher-coverage genome, DevilsGate1 also gives low contamination levels [1% (95% CI, 0 to 2%); see the Supplementary Materials]. As a further check against the possible confounding effect of contamination, we made sure that our most important analyses [outgroup f_3 scores and principal components analysis (PCA)] were qualitatively replicated using only reads showing evidence of postmortem damage (PMD score of at least 3) (*13*), although these latter results had a high level of noise due to the low coverage (0.005*X* for DevilsGate1 and 0.001*X* for DevilsGate2).

[1]Department of Zoology, University of Cambridge, Downing Street, Cambridge CB23EJ, U.K. [2]Smurfit Institute of Genetics, Trinity College Dublin, Dublin, Ireland. [3]The Genomics Institute, Ulsan National Institute of Science and Technology, Ulsan 44919, Republic of Korea. [4]Geromics, Ulsan 44919, Republic of Korea. [5]Clinomics Inc., Ulsan 4919, Republic of Korea. [6]Institute of Ethnology and Anthropology, Russian Academy of Sciences, Moscow, Russia. [7]Institute for Biochemistry and Biology, Faculty for Mathematics and Natural Sciences, University of Potsdam, Karl-Liebknecht-Str. 24-25, 14476 Potsdam-Golm, Germany. [8]School of Archaeology and Earth Institute, University College Dublin, Dublin, Ireland.
*Corresponding author. Email: vs389@cam.ac.uk (V.S.); ron.pinhasi@ucd.ie (R.P.); jongbhak@genomics.org (J.B.); am315@cam.ac.uk (A.M.)
†These authors contributed equally to this work.
‡Adjunct professor at Seoul National University, Seoul, Republic of Korea.

Siska *et al. Sci. Adv.* 2017;**3**:e1601877 1 February 2017 **1 of 10**



www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Analysis of the FGF gene family provides insights into aquatic adaptation in cetaceans

Received: 07 July 2016

Accepted: 02 December 2016

Published: 11 January 2017

Kiwoong Nam[1,2,*], Kyeong Won Lee[3,*], Oksung Chung[4,*], Hyung-Soon Yim[3,5], Sun-Shin Cha[6], Sae-Won Lee[7], JeHoon Jun[4], Yun Sung Cho[4,8], Jong Bhak[4,8,9], João Pedro de Magalhães[10], Jung-Hyun Lee[3,5] & Jae-Yeon Jeong[3,5]

Cetacean body structure and physiology exhibit dramatic adaptations to their aquatic environment. Fibroblast growth factors (FGFs) are a family of essential factors that regulate animal development and physiology; however, their role in cetacean evolution is not clearly understood. Here, we sequenced the fin whale genome and analysed FGFs from 8 cetaceans. FGF22, a hair follicle-enriched gene, exhibited pseudogenization, indicating that the function of this gene is no longer necessary in cetaceans that have lost most of their body hair. An evolutionary analysis revealed signatures of positive selection for FGF3 and FGF11, genes related to ear and tooth development and hypoxia, respectively. We found a D203G substitution in cetacean FGF9, which was predicted to affect FGF9 homodimerization, suggesting that this gene plays a role in the acquisition of rigid flippers for efficient manoeuvring. Cetaceans utilize low bone density as a buoyancy control mechanism, but the underlying genes are not known. We found that the expression of FGF23, a gene associated with reduced bone density, is greatly increased in the cetacean liver under hypoxic conditions, thus implicating FGF23 in low bone density in cetaceans. Altogether, our results provide novel insights into the roles of FGFs in cetacean adaptation to the aquatic environment.

Cetaceans (baleen and toothed whales) were derived from extinct, semi-aquatic, deer-like, even-toed ungulates (artiodactyls) approximately 50 million years ago[1] and have successfully re-populated from terrestrial to aquatic environments. After becoming fully aquatic, the Mysticeti (baleen whales) diverged from the Odontoceti (toothed whale) following the development of keratinous sieves that enabled filter-feeding prior to the onset of the Oligocene Epoch, and subsequently lost teeth completely[2,3].

The anatomical structures, physiology, and metabolism of cetaceans have changed due to various challenges associated with aquatic life. The body shape has been modified to a streamlined form that could reduce fluid resistance[4]. Flukes were developed on their tail for propulsion, hindlimbs were degenerated, and forelimbs were modified into diverse forms of flippers with fused elbow joints that were more suitable for steering than paddling[5]. The hairy fur of their close terrestrial relatives was essentially lost in cetaceans for hydrodynamic reasons[4], and the bone mineral density was reduced to allow dynamic buoyancy control in deep water[6]. The outer ear pinnae were lost in cetaceans, and the outer ears were functionally replaced by the mandible and the mandibular fat pad, which were better adapted for hearing underwater[4,7]. Cetaceans also exhibit various specializations, such as increased oxygen storage capacity, cardiovascular and metabolic adjustments, and increased levels of antioxidants

[1]INRA, UMR 1333 Diversité, Génomes & Interactions Microorganismes–Insectes, 2 place E. Bataillon, 34095 Montpellier, France. [2]Université Montpellier, 2 place E. Bataillon, 34095 Montpellier, France. [3]Marine Biotechnology Research Center, Korea Institute of Ocean Science and Technology, Haeanro 787, Ansan 15627, Republic of Korea. [4]Personal Genomics Institute, Genome Research Foundation, Osong 28160, Republic of Korea. [5]Department of Marine Biotechnology, Korea University of Science and Technology, Daejeon 306-350, Republic of Korea. [6]Department of Chemistry and Nano Science, Ewha Womans University, Seoul, 03760, Republic of Korea. [7]Biomedical Research Institute and IRICT, Seoul National University Hospital, Seoul 110-744, Republic of Korea. [8]The Genomics Institute, Biomedical Engineering Department, UNIST, Ulsan 44919, Republic of Korea. [9]Geromics, Ulsan 44919, Republic of Korea. [10]Institute of Integrative Biology, University of Liverpool, Liverpool L69 7ZB, United Kingdom. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to J.-H.L. (email: jlee@kiost.ac.kr) or J.-Y.J. (email: jeongjy@gmail.com)